Development of Methods for the Investigation of RNA-Ligand Interactions

by

Joseph D. Yesselman

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biophysics)
in the University of Michigan
2013

Doctoral Committee:

Professor Charles L. Brooks III, Co-Chair
Professor Hashim M. Al-Hashimi, Co-Chair
Professor Heather A. Carlson
Professor Jens-Christian Meiners
Professor David Sept

2013

To my friends and family.

Acknowledgments

I would first like to thank my advisors Prof. Charles L. Brooks III and Prof. Hashim M. Al-Hashimi. Their encouragement and support were alone enough to keep me going but they also changed my entire outlook on scientific research and how to approach questions. In addition I would also like to thank Hashim for the interview we had when I was applying to University of Michigan. His enthusiasm and his explanation of his research absolutely sold me on this field when I was still having doubts on what I wanted to do, I truly would not be here today if not for that interview. I would like to thank all former students and post-docs from both the Al-Hashimi group and the Brooks group, everyone who I have had a chance to interact with during my work here has shaped how I think about research. I Specifically would like to thank, Dr. Andrew Stelzer, who initially trained me on the basics of RNA, and RNA-small molecule binding. Not only was he critical to getting me situated at the beginning of graduate school but also a great friend and playing racquetball together was a ton of fun. I would also like to thank Dr. Jennifer Knight who helped mentor me in the Brooks group on CHARMM as well as on approaches to computation in general. Dr. Scott Horowitz also contributed greatly to my success here, not only did he include me in his own projects but he also contributed greatly to my own work through thoughtful work-related discussions. I am also thankful for his continued friendship and willingness to collaborate. Tony Mustoe, Dr. Jeet Chugh and Dr. Anette Casiano have also helped me repeatedly throughout my time here. I would finally like to thank my wife to be Dr. Catherine Eichhorn. I am not sure there are words to describe the appreciation I have for all that you have done for me, especially during the writing and editing here. All I can say is thanks for everything and I love you dearly.

Table of Contents

List of Figures

Figure

List of Tables

Table

List of Appendices

Appendix

Abstract

Three critical features of RNA make it a unique challenge for drug discovery: a) it is highly negatively charged, increasing non-specific binding and b) it can be highly dynamic, adopting different conformations upon binding varying ligands, c) it has solvent exposed shallow binding pockets. All these properties represent distinct problems in the advancement of RNA-drug discovery. To address this first problem, MATCH was developed to rapidly, accurately, and universally parameterizes small molecules for docking. MATCH accomplishes this by deconstructing a force field into a set of fundamental rules which best replicates existing parameters and permits extension to new molecules, yielding near instantaneous parameterization of novel molecules. Using a set of 400 small molecules, the MATCH algorithm was rigorously validated by computing their hydration free energies. This benchmark also discovered specific chemical groups that require additional refinement in the underlying small molecule force field. MATCH is not only necessary to study RNA-ligand interactions *en masse* but will also contribute to understanding the charge-charge consequences of ligand binding.

To address RNA flexibility, a method to combine NMR chemical shifts and Molecular Dynamics (MD) was developed to generate ensembles representing the dynamic landscape of RNA targets. NMR chemical shifts are simple to measure and contain a wealth of structural information. To benchmark this technique, a set of 26 RNA structures with internal bulges and experimentally determined chemical shift was selected. For each system reproduction of chemical shifts was used to select an ensemble of conformations from a pool of MD generated structures. Each produced ensemble was consistent with conformations determined by traditional NMR structural restraints of NOE and RDCs. To demonstrate the utility of this method a large pool of structures (~350,000) was used to generate an ensemble that best agreed with the experimental chemical shifts for a prominent RNA target – the ribosomal decoding site. The conformations within this

ensemble were found on favorable areas of the free energy landscape, independently indicating the validity of these structures. In addition, the selected conformations were similar to many experimentally solved structures.

Finally to address the solvent exposed binding pocket of RNA and its flexible ligands, a new docking approach for RNA was developed, which performs an enhanced sampling technique by fragmenting the ligand and independently optimizing the conformation of each fragment. To properly benchmark this novel algorithm, a large set of 230 nucleic acid-ligand complexes was compiled. Compared to previous sets this one has 89 complexes with experimental binding affinities, which is 50% more then previous sets with at most 60 complexes. Utilizing this large set of this enhanced sampling technique was compared to ICM – a leading docking program. ICM produced native-like conformations 45% of the time, while our approach yields native-like conformations 55% of the time. Demonstrating the effectiveness of this novel sampling procedure.

CHAPTER 1

Introduction

1.1 RNA as a Therapeutic Target

1.1.1 The Role of RNA in Cellular Function

For most of the past century RNA was largely viewed a simple carrier of information between stable DNA and functional proteins. The past couple of decades have revealed a paradigm shift in the importance and function of RNA since the discovery of catalytic and functional RNA(1,2). Although only 2-3% of RNA is translated into protein up to 96% is transcribed into so called non-coding RNAs (ncRNAs) (3-5). Like their protein counterparts, RNA is capable of folding into complex tertiary structures to perform a variety of functions. It is now understood that ncRNAs carry out a wide range of roles such as protein synthesis(6-8), self-splicing intron removal(1,9-12), pre-mRNA splicing(13-15) and telomere maintenance(16,17), illustrating the breadth of RNA functions beyond the original singular role of messenger initially given to RNA. Given the ubiquity and importance of RNA in cellular function, research in drug design and discovery has begun to consider RNA a possible drug target.

1.1.2 RNA Targeted Therapies

Identifying small molecules which will inhibit disease-state enzymes has long been a problem in traditional drug design: a study by the FDA on approved small molecule drugs found only 207 proteins, 50% of which are G-protein coupled receptors(18-24). Although this may be an underestimate of the possible total percentage of the proteome that is be druggable, the numbers are striking especially given that there are over 1500 proteins currently known to be directly linked to genetic disease and hundreds of thousands of proteins are translated in human cells(24). There are two common ways that

a protein may be considered ‘undruggable’. First, it can be part of a large family of proteins with very similar substrates and binding pockets. A classic case of this is kinases, which are involved in many cancers: only recently has there been any success in targeting them(25). Second, many proteins perform their function by binding to other proteins, lacking catalytic function. Developing inhibitors to prevent two proteins from binding at the protein-protein interface by small molecules has not yet been shown to be feasible(26). New avenues need to be developed to treat the wide variety of diseases with known biological mechanisms yet contain ‘undruggable’ protein targets. One such avenue is the use of potential RNA targets.

Drugs have already developed which target RNA, the most established of which targets the ribosomal decoding site (A-site), which is located on the 30S ribosome and is responsible for correctly matching a cognate tRNA to the current mRNA codon in the acceptor site(6-8,27). Many FDA-approved antibiotics bind to the A-site, disrupting the translational fidelity and resulting in the incorporation of incorrect amino acid residues and subsequent cell death(28,29). In addition A-site has served as a paradigm for developing novel RNA inhibitors and has possibly the largest number of known binders out of any RNA(30-41). In addition to A-site, antiviral therapies have been in development to treat HIV-1 by targeting the transactivation response (TAR) element (42,43), which permits transcription of the HIV-1 genome. Additional HIV-1 RNA targets are the rev response element (RRE)(44), which is the recognition site for the Rev protein responsible for transporting the HIV genome to the cytoplasm of the cell, and the dimerization initiation site (DIS)(45-47), which is the key recognition site for HIV genome dimerization prior to viral maturation.

1.1.3 Structural Basis of RNA-ligand interactions

Unlike proteins, RNA has a high degree of flexibility, adopting radically different states upon binding proteins or ligands. Some of these changes are quite drastic: for example HIV-1 TAR adopts different conformations depending on the small molecule it is bound to. Its interhelical bend can change from 5º to 47º: a massive helix orientation change(**Figure 1.1**) (48-53). A-site consists of a S2S1 bulge with two adenines (A1492 and A1493) on one strand and A1408 on the other. The free state is thought to have either

one or both A1492 and A1493 in a flipped in conformation but upon binding of aminoglyocides such as paromomycin, both adenines flip out with paromomycin

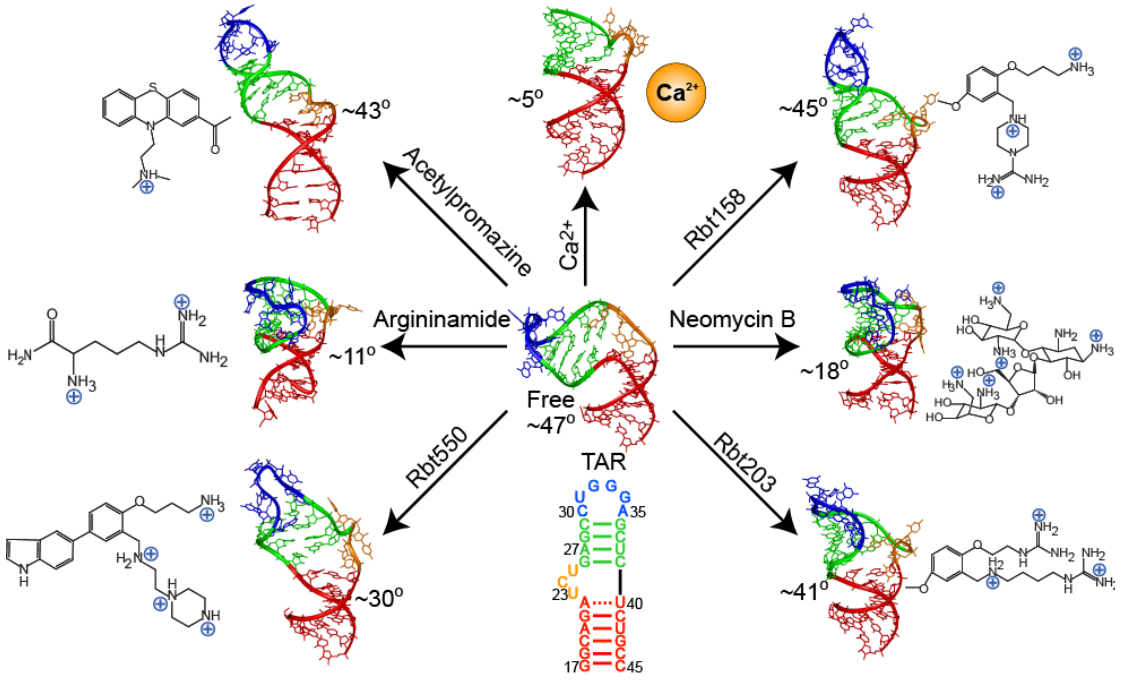

Figure 1.1: TAR structures vary significantly in RNA conformation (inter-helical bend angle ranges from 5° to 47°). The large structural rearrangements observed for TAR are likely governed by a conformational selection mechanism(61).

occupying the space previously by A1492 and A1493(29,54-60).

Two possible mechanisms can describe a binding event: induced fit or conformational capture. Induced fit occurs when a conformational change takes place in a receptor on coming into contact with a ligand, allowing the ligand to bind to the modified receptor. In contrast, conformational capture occurs when the receptor samples a wide number of states and the ligand can bind to one or a subsection of the total number of states. Induced fit and conformational capture mechanisms are not mutually exclusive: each binding event may have some aspects of induced fit and some of conformational capture. It is critical for drug design development to try and understand which type of binding event is dominant to tailor drugs to a given system. For example if the mechanism is believed to be induced fit then when performing computational docking it would be beneficial to use one conformation of the receptor and then allow flexibility in the receptor to find ligands

that bind the receptor. On the other hand, if the mechanism were conformational capture then docking against an ensemble of rigid structures would be favorable. There is great interest in identifying whether RNA-small molecule binding is primarily induced fit or conformational capture. In the most well-studied systems: Asite, HIV-1 TAR, and HIV-1 RRE, it is thought that the mechanism is primarily conformational capture(30,62-65).

For HIV-1 TAR, recent work performed by Al-Hashimi and colleagues have further verified that RNA has a conformational capture ligand binding mechanism. Using a combination of NMR and MD, the authors generated a dynamic ensemble of structures for HIV-1 TAR and docked a small molecule library against it(66-68). Multiple binders were identified which later were confirmed using florescence, NMR titrations and *in vivo* inhibition of live HIV-1 viral replication (68). The general method for generating these dynamic ensembles using a combination of NMR data and MD simulations will be discussed in detail in section (**1.4.1**).

1.1.4 RNA-Small Molecule Interactions

RNA like proteins can fold into complex tertiary structures, which contain features that allow them to be targeted with small molecules. Common features of RNA tertiary structure is double helices, bulges, internal loops, and junctions that can generate accessible binding pockets where it is possible to form interactions with other RNAs(69), proteins(70) and small molecules(71). Many RNA-binding small molecules contain elements of nucleobases themselves, with pyrimidine- and purine-like rings and sugars. Because of the likeness to nucleic acids, binding interactions share many of the traits of RNA tertiary structure. There are three interactions that are thought to play a key role in the binding of small molecules to RNA: hydrogen bonds, electrostatics and ring stacking(72,73) (**Figure 1.2**).

Hydrogen Bonding of Ligands to RNA

Hydrogen bonds can be formed on either face of the nucelobase or with the phosphate backbone. For nucleobase hydrogen bonding, riboswitches employ Watson-Crick and non-canonical base pairing interactions to produce extremely high affinity with their respective ligands. It is typical for a riboswitch ligand to have every hydrogen bond donor and acceptor occupied with a hydrogen bond with the riboswitch(74-76). For example the C74U mutation of the guanine riboswitch binds 2,6-diaminopurine with extremely high affinity (-9 kcal/mol binding energy)(76,77). Upon binding, eight out the nine hydrogen bond donors and acceptors are occupied in hydrogen bonds with the riboswitch. On the other hand a U47C mutation allows the binding of hypoxanthine,

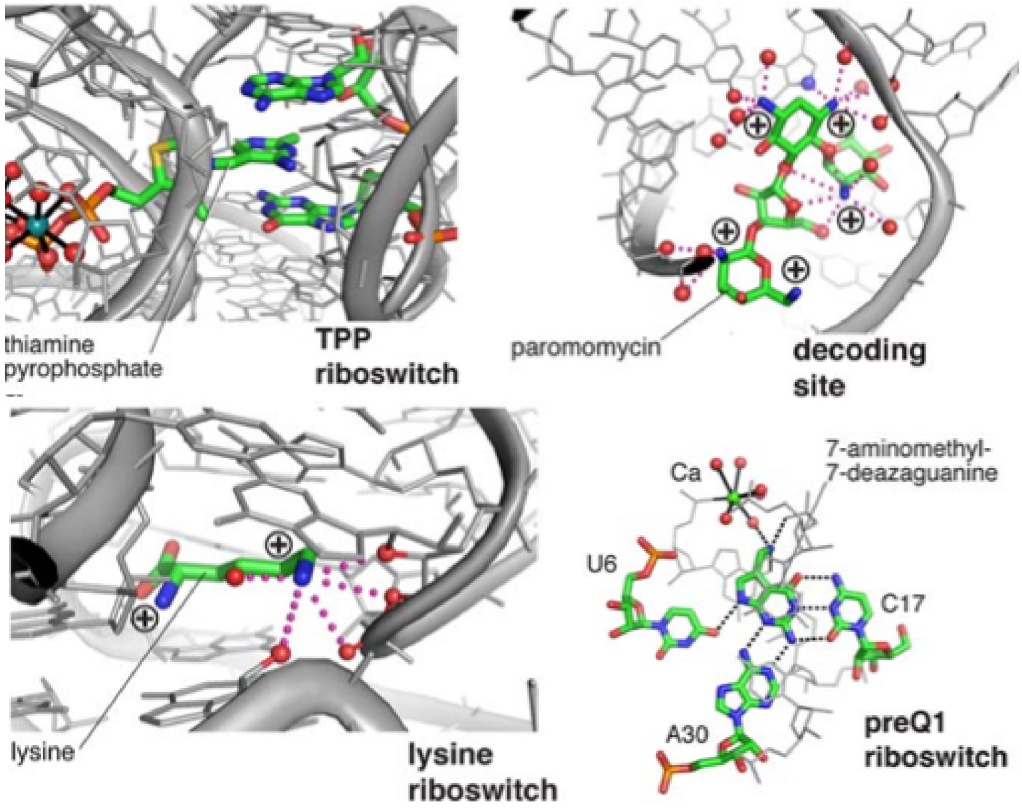

Figure 1.2: Examples of the different interactions between RNA and small molecules. Top-left, stacking or π-π interactions in the TPP riboswitch. Top-right and Bottom left, electrostatic charge-charge interactions in both decoding site and the lysine riboswitch. Bottom-right, hydrogen bonding in the preQ1 riboswitch(72).

which has five out of its six hydrogen bonds occupied which is 3 less than 2,6-diaminopurine with a much weaker affinity (-6 kcal/mol)(78). Lastly, benzimidazole

cannot bind to the C74U mutation at all as it only forms two hydrogen bonds with the riboswitch and lacks specificity(76). This demonstrates that binding is highly dependent on the number of hydrogen bonds generated between the ligand and receptor.

Electrostatic Interactions in RNA-Ligand Binding

Due the large negative charge present the RNA phosphate backbone, RNA ligands tend to have an abundance of positive charge, thus electrostatic interactions play a critical role in small molecule-RNA binding. Aminoglycosides, which are sugars covered in NH_3 groups tend to form a large number of charge-charge interactions with the RNA backbone phosphates and surrounding water molecules(58,79-81). In addition, like in proteins that are ATP dependent, the TPP riboswitch uses Mg^{2+} ions to coordinate binding(82). Furthermore, as Mg^{2+} ion binding sites are common in RNA, it has been demonstrated that some aminoglycosides compete for Mg^{2+} binding sites(83-85).

Ring Stacking in Ligand-RNA Binding

Stacking or π-π interactions play a critical role in the stability of RNA and DNA helices(86-89). Because of the abundance of aromatic groups in nucleic acids, strong small molecule binding often forms stacking interactions with unpaired residues or a full base pair. This is extremely common in riboswitches. All the classes of the SAM riboswitches employ a pocket where the adenine base can perfectly stack with two other bases riboswitches(90-92). There is also significant π-π stacking of acetylpromazine upon binding to TAR with anthracene-like stacking with the U25, G26 and U40 in the S3S0 bulge of its binding pocket(50).

1.1.5 Prokaryotic Ribosomal Decoding Site

The ribosomal decoding site, located in the 16S rRNA of the 30S subunit, maintains translational fidelity by ensuring that the correct aminoacyl-tRNA anticodon base pairs to its associated mRNA codon. Residues A1492 and A493 which are universally conserved, enable the decoding site to recognize canonical pairing with cognate tRNA by coming in close contact with the first two anticodon-codon base pairs, determining the overall shape of the minor groove in addition to the pattern of hydrogen bond acceptors. During the process of proofreading both adenosines, located in an asymmetric internal loop, flip out

and remain so if contacts to the mRNA-tRNA minihelix are properly formed indicating correct base pairing(7,93,94). In the presence of the antibiotics such as paromomycin A1492 and A1493 adopt a stable flipped out state, significantly reducing the energetic cost of forming contacts with the anticodon-codon complex leading to recognition of mismatched tRNA-codon pairs, and high mutation rates during translation(27,54)(**Figure 1.3**). The prokaryotic ribosomal decoding site is an ideal model system to study RNA-ligand interactions. In addition to the large number of both structural information from NMR and X-ray crystallography of both the free and bound states, the prokaryotic ribomal decoding site is the only RNA target that has FDA approved drugs targeting it.

Although the bound pose of inhibitors to decoding site has been thoroughly studied and there is consensus that both A1492 and A1493 are flipped out, it is less clear what the ground state of the free state is. There are multiple structures available with conflicting conformations: the NMR structure 1A3M has both A1492 and A1493 flipped in, with A1493 forming hydrogen bonds with A1408. The three X-ray constructs contain a very diverse set of conformations: 3BNL has A1492 base paired to A1408 and A1493 flipped in, 1TOE has A1493 base paired to A1408 and A1492 flipped out, and lastly 3LOA has

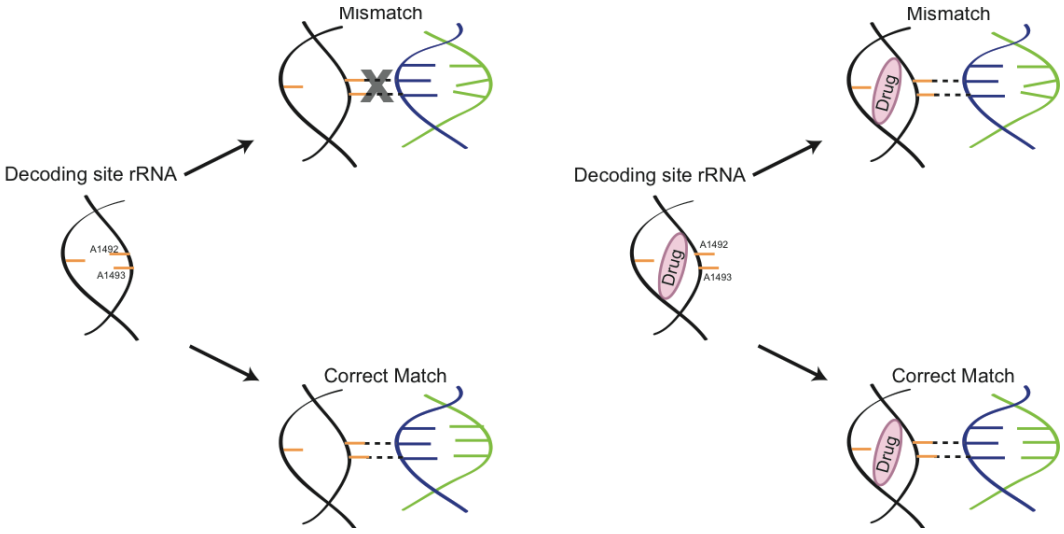

Figure 1.3: Left panel: The proposed mechanism for decoding site. Right panel, upon drug binding decoding mechanism no longer functions properly do to A1492 and A1493 force into the solvent.

A1493 based paired to A1408 and A1492 flipped out. Likewise, conformations found in

the ribosome also differ but generally follow a trend of both A1492 and A1493 flipped in with different hydrogen bond patterns with A1408. While nearly all structures have both A1492 and A1493 in an anti conformation, 3ZVO places A1493 in a syn conformation. In addition biophysical studies show that in the free state, A1492 is looped inside, likely base-pairing with A1408, while A1493 is partially flipped out and flexible(95). Additional work is required to examine which of these states is the dominant state in solution if any. It is possible that some of these structures are artifacts of either crystal packing or lack of density. Investigation of the dominant state of the ribosomal decoding site will be performed in **Chapter 4** using a combination of NMR data and MD simulations.

1.2 Methods of Discovering RNA-Ligand Binding

The most common techniques for discovering small molecules that bind to proteins use high throughput screening methods, which typically rely on biomolecule enzymatic activity. However, most RNAs that are interesting as drug targets are not enzymatically active(96-99). Other methods such as monitoring 2-aminopurine fluorescence can be used, which, when solvent exposed or not forming a stacking interaction, greatly increases its fluorescent properties. Strategies using 2-aminopurine will commonly substitute 2-aminopurine for a single stranded residue and, upon binding of a small molecule, if the nucleotide is extruded, a sharp increase in fluorescence should be observed(95,100-103). In addition, it is also possible to attach fluorescein or pyrene to the 2' position of the ribose sugar(104-108). So far few studies have utilized high-throughput screening technology for indentifying novel small molecule RNA binders. The only reported high-throughput screens have focused on well-characterized RNA targets such as TAR(109,110) and A-site(111,112) with rather limited numbers of molecules tested.

In contrast to experimental high-throughput screens, computational docking provides a means to rapidly screen millions of small molecules against RNA targets. The major constraint for computational docking is the use of 3D structures of the RNA target, which can greatly limit which RNAs can be targeted. Docking a potential small molecule drug into the binding pocket of a RNA target is effectively two separate steps. The first step is

sampling both the different possible binding location on the RNA and the internal degrees of freedom of the small molecule. There are multiple methods to accomplishing this step, including Monte Carlo(113-115), Genetic Algorithms(116-118) and fragment based approaches(119,120). The second step is scoring, or approximation of binding affinity between the small molecule and receptor. There are many approaches to this: either a) force field-based using a molecule mechanics energy function(119,121), b) empirically-based using a system of equations to approximate proper binding(122,123), or c) knowledge-based, building a set of parameters from training on known complexes that can be applied to new systems(124,125). Generally determining intermolecular interactions such as hydrogen bonding, electrostatic interactions, Van der Waals contacts, and solvent effects are used, in addition to small molecule properties.

Computational docking has yielded quite a few successes in protein-ligand drug discovery(126,127). There has been notable drug leads discovered for a host of different targets such as DNA gyrase(128), tyrosine phosphatase(129), dihydrodipicolinate reductase(130) and HIV-1 integrase(131,132). In the docking studies for HIV-1 integrase, the lead compounds discovered led to the development of FDA approved drug Isentress. Jorgensen and colleagues, in addition to discovering novel inhibitors to HIV-1 reverse transcriptase using computational docking, were able to iteratively refine them into sub-nanomolar affinity using free energy perturbation calculations(133-136). Computational docking in combination with large virtual libraries(137) have radically altered how drug design is performed for generating novel protein-small molecule inhibitors.

Compared to the success of computational docking for proteins, RNA has had much less success(138). There are two general approaches to designing RNA docking programs, first is to take existing docking software for proteins and modulate its scoring function to better target RNA. Programs such as Autodock(113), ICM(115), GOLD(121), Glide(139-141), DOCK 6.0(142) and Surflex-dock(143) have been benchmarked and found to be able to place ligand generally in native-like conformation ranging from 35% of the time to 75%. (144-146). In addition, there are newer programs that are built solely for RNA such as Ribodock(147) and MORDOR(148). For all of these programs, the benchmarking

complex sets are usually less than 60 complexes, which is very small compared to the excellent protein-ligand sets out there such as Database (LPDB)(149), Mother of all Databases (MOAD)(150) or The Community Structure−Activity Resource (CSAR)(151-153) which can have over 200 high quality complexes. Generally as a whole the validation and success rates for RNA-ligand interactions are inferior to those found for proteins, of which is due to the insufficient amount of experimental data available. Regardless, there have still been some successes by computational docking. Howes and colleagues docked over one million compounds to the ribosomal A-site using Ribodock and found 34 compounds that bound, confirmed using FRET and some validated using NMR(154). James and colleagues, using the program DOCK and ICM, screened HIV-1 TAR against 181,000 small molecules where they identified acetylpromazine that binds TAR with a Kd of 270 μM and inhibits TAR-Tat mediated HIV transcription *in-vitro*(97). In addition MORDOR was used to screen 5750 small molecules against the human telomerase RNA and determined 48 compounds that bound, confirming binding using saturation transfer NMR experiments(155).

Although there has been some success using computational docking to determine novel RNA binders, there is much room for improvement. To successfully model RNA-ligand binding it is necessary to properly account for the uniqueness of RNA compared to proteins. First, RNA has a high degree of flexibility, adopting radically different states on protein or ligand binding and, as such, it is unreasonable to represent an RNA drug target as a single structure in computational drug discovery. Second, RNA has shallow and solvent-exposed binding pockets where potential ligands can bind. Third, RNA ligands tend to be flexible themselves, sometimes with greater than ten rotatable bonds(73,142,147). The high degree of flexibility of both the RNA and their ligands, in addition to the shallow binding pockets makes it very difficult to find global minimum energy poses(73). Dealing with the high degree of flexibility in RNA is investigated in **Chapter 4** while both the shallow binding and flexibility of the small molecules are dealt with in **Chapter 5**.

1.3 Nuclear Magnetic Resonance

1.3.1 NMR Chemical Shifts

Chemical shifts are sensitive to a wide variety of structural and electronic properties of biopolymers. The chemical shift-structure relationship has been well described in proteins and is routinely used to determine secondary as well as tertiary structure(156-159). Recent advances in the prediction of chemical shifts based on large databases of protein structures have permitted the determination of full 3D structures utilizing exclusively chemical shifts as an experimental restraint(160,161). For nucleic acids, this relationship is still in the early stages of development, with ^{1}H chemical shifts used to predict secondary and tertiary RNA structure. Due to the inherent nature of nucleic acids, proton chemical shifts are largely affected by ring current and magnetic anisotropy effects(162,163). These effects have been implemented in programs such as SHIFTS(164) and NUCHEMICS(165). These programs have been shown to discriminate between native and non-native structures(166,167). Additionally, Summers and colleagues found that ^{1}H chemical shifts were able to discriminate between different base-pair triplets, non-canonical base-pairs and terminal base pairs(167). This work has demonstrated that it is possible to use NMR chemical shifts to inform tertiary RNA structure and this will be explored further in **Chapter 4**.

1.3.2 NMR Residual Dipolar Couplings

Theory

A dipolar coupling occurs between two permanent dipoles, where the local magnetic field of one nucleus perturbs the magnetic field of a neighboring nucleus. Considering a carbon and hydrogen bond, the dipolar interaction between the carbon and hydrogen modulates the effective magnetic field at the carbon nucleus. The effective magnetic field experienced by the carbon nucleus is a summation of the static external magnetic field and the much smaller magnetic field generated by the hydrogen nuclei it is bound to. Because the hydrogen can be either parallel or anti-parallel to the magnetic field, the proton-induced magnetic field experienced by the carbon nucleus varies with the orientation relative to the static external magnetic field. This is due either to internal

dynamics or overall molecular tumbling, and the average strength over all the molecules in a NMR.

$$D_{ij} = -\left(\frac{\mu_0}{4\pi}\right)\frac{\gamma_i\gamma_j h}{2\pi^2 r_{ij,eff}^3}\left\langle\frac{3\cos^2\theta - 1}{2}\right\rangle \tag{1.1}$$

Equation 1.1 describes the magnitude of a dipolar coupling between nuclei *i* and *j*, where μ_0 is the magnetic permittivity of a vacuum, *h* is Planck's constant, r_{ij}(168-170) is the distance between the two nuclei, and γ is the gyromagnetic ratio. The angular brackets denote a time average over all angles sampled by the inter-nuclear bond vector.

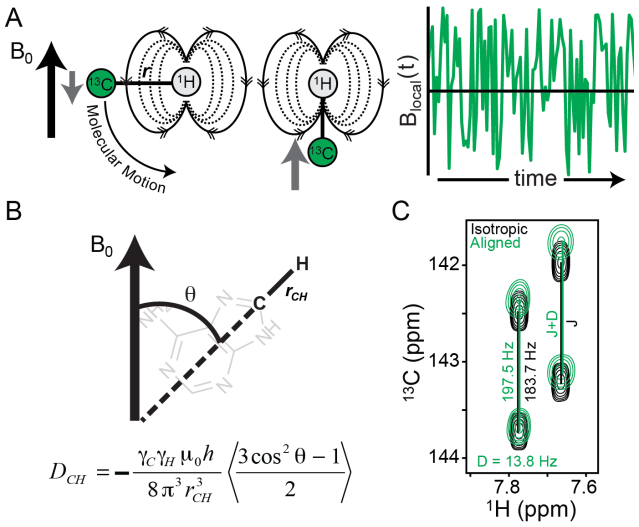

Figure 1.4: Description of RDCs**.** A) Changes in orientation alter the local magnetic field. B) The orientation of the internuclear bond vector relative to the external magnetic field. C) RDCs are calculated by measuring the splitting in peaks (black resonances) observed upon partial alignment (green resonances) (171).

Given that the sample behaves isotropically, random molecule tumbling will reduce the angular term to zero. Thus, dipolar couplings cannot be observed under normal solution conditions. However, if an alignment media is introduced, permitting the molecule to be partially aligned to the magnetic field, the angular value will no longer result in an average of zero, and the carbon nucleus will experience a net magnetic field from the hydrogen it is bonded to in addition to the external magnetic field(172). As a result, the carbon resonance frequency will be split, one adding to the external magnetic field and one subtracting from it. The magnitude of the split is the 'residual dipolar coupling' (RDC)(173,174). The splitting in peaks generated through dipolar coupling (D) and bond scalar couplings (J) both effectively increase or decrease the magnetic field. By subtracting the J component the dipolar coupling term can be measured. Although RDCs do not provide information about motional timescales beyond

being within the sub-millisecond regime, they are exquisitely sensitive to the orientational distribution sampled by the bond vector and, therefore, the 3D choreography of the motion(175-177)(**Figure 1.4C**).

Alignment Methods

The magnitude of RDCs depends on the degree of alignment achieved in solution, accomplished by dissolving an alignment media into the NMR sample. A specific level of alignment must be achieved in order to detect the RDC: alignment levels of less then 1 in 10^5 molecules is insufficient to accurately measure compared to NMR linewidths. Conversely, it is possible to have too much alignment: alignment levels where 1 in 100 molecules are aligned give rise to extensive dipolar couplings, compromising the spectral resolution required for large molecule. The ideal amount of alignment typically occurs when 1 in 1000 molecules is completely aligned(172,178). Many alignment medias are commercially available that can be used to align nucleic acids. Since nucleic acids are highly negatively charged, it is beneficial for the alignment media to take advantage of this property. In addition nucleic acids also require high ionic strength conditions to remain stable. The most common ordering medium that can satisfy both of these constraints is filamentous *Pf1* bacteriophage, which is negatively charged thus allowing alignment via electrostatics and sterics(179-181). The typical concentration of phage used to align an NMR sample is typically 10-20 mg/mL but can vary based on the shape and length of the target nucleic acid.

1.4 Combining NMR and Computational Methods to Identify RNA Therapeutics

1.4.1 RNA Dynamic Ensembles for Computational Docking

RNA is inherently flexible, and can adopt radically different conformations upon binding of proteins or small molecules. For drug discovery protocols RNA's dynamic nature can be modeled by generating an ensemble of structures to represent highly populated states. For HIV-1 TAR(**Figure 1.5**) this has shown to be useful in both modeling the bound state RNA but also discovering new inhibitors(66,67). These dynamic ensembles can be generated by measuring experimental observables with NMR and then using a pool of conformations produced by MD to back-predict the observables. This approach has been

successful because NMR and MD are complementary on both spatial and temporal scales, which can overcome the shortcomings of either. MD can fill in the shortage of structural information from NMR data and NMR can provide validation for the MD conformations. There have been several studies that have used both NMR and MD in the study of protein dynamics(182-185) with great success. Recently, there has also been an effort to apply a similar technique to RNA; however, RNA is more challenging due to the tendency of RNA dynamics to couple overall tumbling and the internal motions. Nevertheless there has been some success combining RNA and MD for the study of nucleic acid dynamics(66,186-190).

The basic premise of the algorithm now known as sample and select (SAS) is to generate a set of conformations for an RNA of interest by running an extended MD simulation. To initiate the SAS selection, an N-membered subset of structures is randomly selected from a total pool of M structures. One then searches to increase the agreement of the N-membered subset of structures with the NMR observable data. This is done using simulated annealing stages following the Metropolis criteria for acceptance of new subsets.

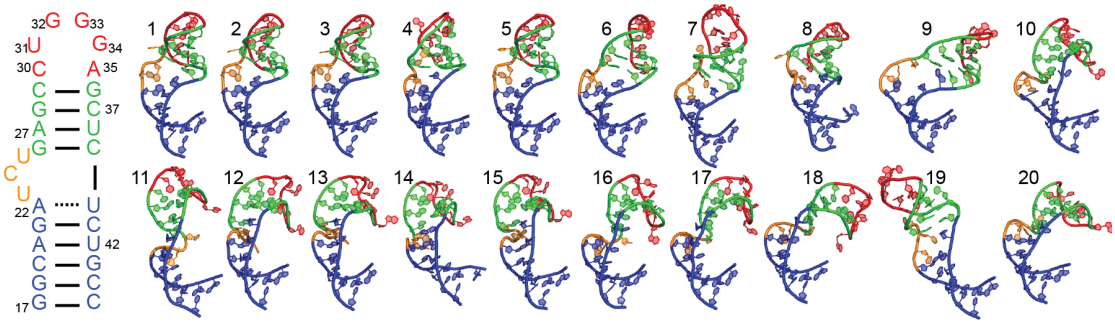

Figure 1.5: The dynamic ensemble generated for HIV-1 TAR using a combination of NMR and MD and was used to dock against (66).

1.4.2 Automated Ligand Parameterization

A key feature for performing large-scale computational docking screens is the parameterization of small molecules. Unlike proteins or nucleic acids, which are biopolymers composed of a set number of amino acids or nucleic acids with set atomic

parameters and charges, each small molecule needs to be independently parameterized. For this reason many docking protocols have adopted parameterization schemes built on organic molecule force fields that were optimized independently of the biomolecular force fields. Each force field has adopted a different strategy to optimize the bonded and nonbonded parameters and attempted to reproduce experimental data or quantum mechanical properties of model compounds. Therefore, it is unlikely that the combination of a particular biomolecular force field with an arbitrary ligand force field to yield properly balanced intermolecular interactions(191). Rather, it is crucial that the small molecule parameters follow a parameterization similar scheme to that which was used to develop the biomolecular force field. Methods for deconstructing a force field into a set of fundamental rules which best replicates existing parameters and permits extension to new molecules yielding near instantaneous parameterization of novel molecules will be discussed in **Chapter 2**. In addition, extensive validation of this method and how it compares to other procedures will be examined in **Chapter 3**.

1.5 References

1. Kruger K, Grabowski PJ, Zaug AJ, Sands J, Gottschling DE, Cech TR. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. Cell. Elsevier. 1982;31(1):147–57.

2. Fletcher S, Hamilton AD. Protein surface recognition and proteomimetics: mimics of protein surface structure and function. Current Opinion in Chemical Biology. 2005 Dec;9(6):632–8.

3. Guerrier-Takada C, Gardiner K, Marsh T, Pace N, Altman S. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. Cell. Elsevier. 1983;35(3):849–57.

4. Frith MC, Pheasant M, Mattick JS. Genomics: The amazing complexity of the human transcriptome. Eur J Hum Genet. 2005 Jun 22;13(8):894–7.

5. Wimberly BT, Brodersen DE, Clemons WM, Morgan-Warren RJ, Carter AP, Vonrhein C, et al. Structure of the 30S ribosomal subunit. Nature. 2000 Sep 21;407(6802):327–39.

6. Mattick JS. Non-coding RNAs: the architects of eukaryotic complexity. EMBO Rep. 2001 Nov;2(11):986–91.

7. Rodnina MV, Daviter T, Gromadski K, Wintermeyer W. Structural dynamics of ribosomal RNA during decoding on the ribosome. Biochimie. 2002 Aug;84(8):745–54.

8. Mattick JS. RNA regulation: a new genetics? Nat Rev Genet. 2004 Apr;5(4):316–23.

9. Noller HF. RNA structure: reading the ribosome. Science. American Association for the Advancement of Science. American Association for the Advancement of Science; 2005;309(5740):1508–14.

10. Noller HF. Biochemical characterization of the ribosomal decoding site. Biochimie. 2006 Aug;88(8):935–41.

11. Thomas CL, Gregory RJ, Winslow G, Muto A, Zimmermann RA. Mutations within the decoding site of Escherichia coli 16S rRNA: growth rate impairment, lethality and intragenic suppression. Nucleic Acids Research. 1988 Aug 25;16(16):8129–46.

12. Vicens Q, Westhof E. Crystal Structure of Paromomycin Docked into the Eubacterial Ribosomal Decoding A Site. 2001 Jul 31;:1–12.

13. Hermann T. Rational ligand design for RNA: the role of static structure and conformational flexibility in target recognition. Biochimie. 2002 Sep;84(9):869–75.

14. Guo F, Gooding AR, Cech TR. Structure of the Tetrahymena ribozyme: base triple sandwich and metal ion at the active site. Mol. Cell. 2004 Nov 5;16(3):351–62.

15. Vourloumis D, Takahashi M, Winters GC, Simonsen KB, Ayida BK, Barluenga S, et al. Novel 2, 5-dideoxystreptamine derivatives targeting the ribosomal decoding site RNA. Bioorganic & Medicinal Chemistry Letters. Elsevier; 2002;12(23):3367–72.

16. Cate JH, Hanna RL, Doudna JA. A magnesium ion core at the heart of a ribozyme domain. Nat. Struct. Biol. 1997 Jul;4(7):553–8.

17. Connell SR, Trieber CA, Stelzl U, Einfeldt E, Taylor DE, Nierhaus KH. The tetracycline resistance protein Tet(o) perturbs the conformation of the ribosomal decoding centre. Mol. Microbiol. 2002 Sep;45(6):1463–72.

18. Bevilacqua PY, Nucleobase catalysis in ribozyme mechanism. Current Opinion in Chemical Biology. 2006 Oct;10(5):455–64.

19. Simonsen KB, Ayida BK, Vourloumis D, Takahashi M, Winters GC, Barluenga S, et al. Novel paromamine derivatives exploring shallow-groove recognition of ribosomal-decoding-site RNA. ChemBioChem. 2002 Dec 2;3(12):1223–8.

20. Weeks KM, Cech TR. Protein facilitation of group I intron splicing by assembly of the catalytic core and the 5′ splice site domain. Cell. Elsevier; 1995;82(2):221–30.

21. Vourloumis D, Winters GC, Takahashi M, Simonsen KB, Ayida BK, Shandrick S, et al. Novel Acyclic Deoxystreptamine Mimetics Targeting the Ribosomal Decoding Site. ChemBioChem. 2003 Sep 4;4(9):879–85.

22. Simonsen KB, Ayida BK, Vourloumis D, Winters GC, Takahashi M, Shandrick S, et al. Piperidine Glycosides Targeting the Ribosomal Decoding Site. ChemBioChem. 2003 Sep 4;4(9):886–90.

23. Vourloumis D, Winters GC, Simonsen KB, Takahashi M, Ayida BK, Shandrick S, et al. Aminoglycoside-Hybrid Ligands Targeting the Ribosomal Decoding Site. ChemBioChem. 2004 Nov 29;6(1):58–65.

24. Fran ois B, Szychowski J, Adhikari SS, Pachamuthu K, Swayze EE, Griffey RH, et al. Antibacterial Aminoglycosides with a Modified Mode of Binding to the Ribosomal-RNA Decoding Site. Angew. Chem. Int. Ed. 2004 Dec 10;43(48):6735–8.

25. Hermann T. Drugs targeting the ribosome. Current Opinion in Structural Biology. 2005 Jun;15(3):355–66.

26. Dibrov SM, Parsons J, Hermann T. A model for the study of ligand binding to the ribosomal RNA helix h44. Nucleic Acids Research. 2010 Jul 26;38(13):4458–65.

27. Han Q, Zhao Q, Fish S, Simonsen KB, Vourloumis D, Froelich JM, et al. Molecular Recognition by Glycoside Pseudo Base Pairs and Triples in an Apramycin-RNA Complex. Angew. Chem. Int. Ed. 2005 Apr 29;44(18):2694–700.

28. Sutcliffe JA. Improving on nature: antibiotics that target the ribosome. Current Opinion in Microbiology. 2005 Oct;8(5):534–42.

29. Guthrie C. Messenger RNA splicing in yeast: clues to why the spliceosome is a ribonucleoprotein. Science. American Association for the Advancement of Science. American Association for the Advancement of Science; 1991;253(5016):157–63.

30. Krebs A, Ludwig V, Boden O, Göbel MW. Targeting the HIV Trans-Activation Responsive Region-Approaches Towards RNA-Binding Drugs. ChemBioChem. 2003 Sep 26;4(10):972–8.

31. Brody E, Abelson J. The“ spliceosome”: yeast pre-messenger RNA associates with a 40S complex in a splicing-dependent reaction. Science. American Association for the Advancement of Science. American Association for the

Advancement of Science; 1985;228(4702):963–7.

32. Mei H-Y, Galan AA, Halim NS, Mack DP, Moreland DW, Sanders KB, et al. Inhibition of an HIV-1 Tat-derived peptide binding to TAR RNA by aminoglycoside antibiotics. Bioorganic & Medicinal Chemistry Letters. Elsevier. 1995;5(22):2755–60.

33. Wahl MC, Will CL, Lührmann R. The Spliceosome: Design Principles of a Dynamic RNP Machine. Cell. 2009 Feb;136(4):701–18.

34. Shippen-Lentz D, Blackburn EH. Functional evidence for an RNA template in telomerase. Science. American Association for the Advancement of Science. 1990;247(4942):546–52.

35. Zapp ML, Stern S, Green MR. Small molecules that selectively block RNA binding of HIV-1 Rev protein inhibit Rev function and viral production. Cell. Elsevier; 1993;74(6):969–78.

36. Blackburn EH. Structure and function of telomeres. Nature. 1991;350(6319):569–73.

37. Zheng CJ. Therapeutic Targets: Progress of Their Exploration and Investigation of Their Characteristics. Pharmacological Reviews. 2006 Jun 1;58(2):259–79.

38. Laughrea M, Jette L. A 19-nucleotide sequence upstream of the 5'major splice donor is part of the dimerization domain of human immunodeficiency virus 1 genomic RNA. Biochemistry. ACS Publications; 1994;33(45):13464–74.

39. Wishart DS. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Research. 2006 Jan 1;34(90001):D668–72.

40. Paillart JC, Marquet R, Skripkin E, Ehresmann C, Ehresmann B. Dimerization of retroviral genomic RNAs: structural and functional implications. Biochimie. 1996;78(7):639–53.

41. Hopkins AL, Groom CR. The druggable genome. Nat Rev Drug Discov. Nature Publishing Group; 2002;1(9):727–30.

42. Skripkin E, Paillart J-C, Marquet R, Ehresmann B, Ehresmann C. Identification of the primary site of the human immunodeficiency virus type 1 RNA dimerization in vitro. Proc. Natl. Acad. Sci. U.S.A. National Acad Sciences; 1994;91(11):4945–9.

43. Drews J, Ryser S. The role of innovation in drug development. Nat Biotechnol. 1997 Dec;15(13):1318–9.

44. Drews J. Genomic sciences and the medicine of tomorrow. Nat Biotechnol. 1996 Nov;14(11):1516–8.

45. Imming P, Sinning C, Meyer A. Drugs, their targets and the nature and number of drug targets. Nat Rev Drug Discov. Nature Publishing Group; 2007;6(2).

46. Overington JP, Al-Lazikani B, Hopkins AL. How many drug targets are there? Nat Rev Drug Discov. Nature Publishing Group; 2006;5(12):993–6.

47. Puglisi JD, Tan R, Calnan BJ, Frankel AD, Williamson JR. Conformation of the TAR RNA-arginine complex by NMR spectroscopy. Science. American Association for the Advancement of Science; 1992;257(5066):76–80.

48. Davis B, Afshar M, Varani G, Murchie AIH, Karn J, Lentzen G, et al. Rational Design of Inhibitors of HIV-1 TAR RNA through the Stabilisation of Electrostatic "Hot Spots." Journal of Molecular Biology. 2004 Feb;336(2):343–56.

49. Du Z, Lind KE, James TL. Structure of TAR RNA complexed with a Tat-TAR interaction nanomolar inhibitor that was identified by computational screening. Chemistry & Biology. Elsevier; 2002;9(6):707–12.

50. Faber C. Structural Rearrangements of HIV-1 Tat-responsive RNA upon Binding of Neomycin B. Journal of Biological Chemistry. 2000 Mar 27;275(27):20660–6.

51. Murchie AIH, Davis B, Isel C, Afshar M, Drysdale MJ, Bower J, et al. Structure-based Drug Design Targeting an Inactive RNA Conformation: Exploiting the Flexibility of HIV-1 TAR RNA. Journal of Molecular Biology. 2004 Feb;336(3):625–38.

52. Aboul-ela F, Karn J, Varani G. The structure of the human immunodeficiency virus type-1 TAR RNA reveals principles of RNA recognition by Tat protein. Journal of Molecular Biology. 1995 Oct 20;253(2):313–32.

53. Kamb A, Wee S, Lengauer C. Why is cancer drug discovery so difficult? Nat Rev Drug Discov. 2007 Feb;6(2):115–20.

54. Vicens Q, Westhof E. Crystal structure of a complex between the aminoglycoside tobramycin and an oligonucleotide containing the ribosomal decoding a site. Chemistry & Biology. 2002 Jun;9(6):747–55.

55. Kondo J, Francois B, Russell R, Murray J, Westhof E. Crystal structure of the bacterial ribosomal decoding site complexed with amikacin containing the γ-amino-α-hydroxybutyryl (haba) group. Biochimie. 2006 Aug;88(8):1027–31.

56. Francois B. Crystal structures of complexes between aminoglycosides and decoding A site oligonucleotides: role of the number of rings and positive charges in the specific binding leading to miscoding. Nucleic Acids Research. 2005 Sep 25;33(17):5677–90.

57. Hermann T. A-site model RNAs. Biochimie. 2006 Aug;88(8):1021–6.

58. Vicens Q, Westhof E. Crystal Structure of Geneticin Bound to a Bacterial 16S Ribosomal RNA A Site Oligonucleotide. Journal of Molecular Biology. 2003 Feb;326(4):1175–88.

59. Kondo J, Westhof E. The bacterial and mitochondrial ribosomal A-site molecular switches possess different conformational substates. Nucleic Acids Research. 2008 Feb 29;36(8):2654–66.

60. Varani L, Spillantini MG, Goedert M, Varani G. Structural basis for recognition of the RNA major groove in the tau exon 10 splicing regulatory element by aminoglycoside antibiotics. Nucleic Acids Research. 2000 Feb 1;28(3):710–9.

61. Michael K, Tor Y. Designing novel RNA binders. Chem. Eur. J. Wiley Online Library; 1998;4(11):2091–8.

62. Leulliot N, Varani G. Current Topics in RNA−Protein Recognition: Control of Specificity and Biological Function through Induced Fit and Conformational Capture †. Biochemistry. 2001 Jul;40(27):7947–56.

63. Hermann T, Westhof E. Saccharide–RNA Recognition. 1999 May 14;:1–11.

64. Schroeder. Modulation of RNA function by aminoglycoside antibiotics. 1999 Dec 8;:1–9.

65. Zhang Q, Stelzer AC, Fisher CK, Al-Hashimi HM. Visualizing spatially correlated dynamics that directs RNA conformational transitions. Nature. 2007 Dec 20;450(7173):1263–7.

66. Frank AT, Stelzer AC, Al-Hashimi HM, Andricioaei I. Constructing RNA dynamical ensembles by combining MD and motionally decoupled NMR RDCs: new insights into RNA dynamics and adaptive ligand recognition. Nucleic Acids Research. 2009 Jun 21;37(11):3670–9.

67. Stelzer AC, Frank AT, Kratz JD, Swanson MD, Gonzalez-Hernandez MJ, Lee J, Al-Hashimi HM,. Discovery of selective bioactive small molecules by targeting an RNA dynamic ensemble. Nature Publishing Group. 2011 Jun 26;7(8):553–9.

68. Sevignani C, Calin GA, Siracusa LD, Croce CM. Mammalian microRNAs: a small world for fine-tuning gene expression. Mamm Genome. 2006 Mar 3;17(3):189–202.

69. Hermann T. Strategies for the Design of Drugs Targeting RNA and RNA-Protein Complexes. Angew. Chem. Int. Ed. 2000 Jun 2;39(11):1890–904.

70. Winkler WC, Breaker RR. Genetic Control by Metabolite-Binding Riboswitches. ChemBioChem. 2003 Sep 26;4(10):1024–32.

71. Vicens Q. RNA's coming of age as a drug target. J Incl Phenom Macrocycl Chem. 2009 Jun 27;65(1-2):171–88.

72. Thomas JR, Hergenrother PJ. Targeting RNA with Small Molecules. Chem. Rev. 2008 Apr;108(4):1171–224.

73. Mandal M, Breaker RR. Adenine riboswitches and gene activation by disruption of a transcription terminator. Nat Struct Mol Biol. 2003 Dec 29;11(1):29–35.

74. Serganov A, Yuan Y, Pikovskay AO, Polonskaia A, Malinina L, Phan A. Structural Basis for Discriminative Regulation of Gene Expression by Adenine- and Guanine-Sensing mRNAs. Chemistry & Biology. 2004 Dec;11(12):1729–41.

75. Gilbert SD, Stoddard CD, Wise SJ, Batey RT. Thermodynamic and Kinetic Characterization of Ligand Binding to the Purine Riboswitch Aptamer Domain. Journal of Molecular Biology. 2006 Jun;359(3):754–68.

76. Gilbert SD, Mediatore SJ, Batey RT. Modified Pyrimidines Specifically Bind the Purine Riboswitch. J. Am. Chem. Soc. 2006 Nov;128(44):14214–5.

77. Gilbert SD, Reyes FE, Edwards AL, Batey RT. Adaptive Ligand Binding by the Purine Riboswitch in the Recognition of Guanine and Adenine Analogs. Structure/Folding and Design. Elsevier Ltd; 2009 Jun 10;17(6):857–68.

78. Ban N. The Complete Atomic Structure of the Large Ribosomal Subunit at 2.4 A Resolution. Science. 2000 Aug 11;289(5481):905–20.

79. Böttger EC, Springer B, Prammananan T, Kidan Y, Sander P. Structural basis for selectivity and toxicity of ribosomal antibiotics. EMBO Rep. 2001 Apr;2(4):318–23.

80. Prammananan T, Sander P, Brown BA, Frischkorn K, Onyi GO, Zhang Y, et al. A single 16S ribosomal RNA substitution is responsible for resistance to amikacin and other 2-deoxystreptamine aminoglycosides in Mycobacterium abscessus and Mycobacterium chelonae. J. Infect. Dis. 1998 Jun;177(6):1573–81.

81. Winkler W, Nahvi A, Breaker RR. Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. Nature. 2002 Oct 31;419(6910):952–6.

82. Walter F, Pütz J, Giegé R, Westhof E. Binding of tobramycin leads to conformational changes in yeast tRNA(Asp) and inhibition of aminoacylation. EMBO J. 2002 Feb 15;21(4):760–8.

83. Hoch I, Berens C, WESTHOF E, Schroeder R. Antibiotic inhibition of RNA catalysis: neomycin B binds to the catalytic core of the td group I intron displacing essential metal ions. Journal of Molecular Biology. 1998 Sep

25;282(3):557–69.

84. Zhang Y, Li Z, Pilch DS, Leibowitz MJ. Pentamidine inhibits catalytic activity of group I intron Ca.LSU by altering RNA folding. Nucleic Acids Research. 2002 Jul 1;30(13):2961–71.

85. Watson JD, Crick FH. Genetical implications of the structure of deoxyribonucleic acid. 1953. Nature. 1953. pp. 1967–9.

86. Crick F, Watson J. Molecular structure of nucleic acids. Nature. 1953;171(4356):737–8.

87. Rosenberg JM, Seeman NC, Kim JJP, Suddath FL, Nicholas HB, Rich A. Double helix at atomic resolution. Nature. 1973;243:150–4.

88. Heinemann U, Alings C, Hahn M. Crystallographic studies of DNA helix structure. Biophysical chemistry. Elsevier; 1994;50(1):157–67.

89. Gilbert SD, Rambo RP, Van Tyne D, Batey RT. Structure of the SAM-II riboswitch bound to S-adenosylmethionine. Nat Struct Mol Biol. 2008 Jan 20;15(2):177–82.

90. Montange RK, Batey RT. Structure of the S-adenosylmethionine riboswitch regulatory mRNA element. Nature. 2006 Jun 29;441(7097):1172–5.

91. Lu C, Smith AM, Fuchs RT, Ding F, Rajashankar K, Henkin TM, et al. Crystal structures of the SAM-III/SMK riboswitch reveal the SAM-dependent translation inhibition mechanism. Nat Struct Mol Biol. 2008 Sep 21;15(10):1076–83.

92. Westhof E, Leontis N. Atomic Glimpses on a Billion-Year-Old Molecular Machine. 2000 Apr 19;:1–5.

93. Ogle JM, Carter AP, Ramakrishnan V. Insights into the decoding mechanism from recent ribosome structures. Trends in Biochemical Sciences. 2003 May;28(5):259–66.

94. Shandrick S, Zhao Q, Han Q, Ayida BK, Takahashi M, Winters GC, et al. Monitoring Molecular Recognition of the Ribosomal Decoding Site. Angew. Chem. Int. Ed. 2004 Jun 14;43(24):3177–82.

95. Parsons J, Castaldi MP, Dutta S, Dibrov SM, Wyles DL, Hermann T. Conformational inhibition of the hepatitis C virus internal ribosome entry site RNA. Nat Meth. 2009 Sep 20;5(11):823–5.

96. Lind KE, Du Z, Fujinaga K, Peterlin BM, James TL. Structure-based computational database screening, in vitro assay, and NMR assessment of compounds that target TAR RNA. Chemistry & Biology. Elsevier; 2002;9(2):185–93.

97. Blount KF, Breaker RR. Riboswitches as antibacterial drug targets. Nat Biotechnol. 2006 Dec;24(12):1558–64.

98. Tuccinardi T. Binding-interaction prediction of RNA-binding ligands. Future Medicinal Chemistry. 2011 Apr;3(6):723–33.

99. Bradrick TD. Ligand-induced changes in 2-aminopurine fluorescence as a probe for small molecule binding to HIV-1 TAR RNA. RNA. 2004 Sep 1;10(9):1459–68.

100. Kirk SR, Luedtke NW, Tor Y. 2-Aminopurine as a real-time probe of enzymatic cleavage and inhibition of hammerhead ribozymes. Bioorganic & Medicinal Chemistry. 2001 Sep;9(9):2295–301.

101. Lacourciere KA, Stivers JT, Marino JP. Mechanism of Neomycin and Rev Peptide Binding to the Rev Responsive Element of HIV-1 As Determined by Fluorescence and NMR Spectroscopy. Biochemistry. 2000 May;39(19):5630–41.

102. Yan Z, Baranger AM. Binding of an aminoacridine derivative to a GAAA RNA tetraloop. Bioorganic & Medicinal Chemistry Letters. 2004 Dec;14(23):5889–93.

103. Blount KF, Zhao F, Hermann T, Tor Y. Conformational Constraint as a Means for Understanding RNA-Aminoglycoside Specificity. J. Am. Chem. Soc. 2005 Jul;127(27):9818–29.

104. Thomas JR, Liu X, Hergenrother PJ. Size-Specific Ligands for RNA Hairpin Loops. J. Am. Chem. Soc. 2005 Sep;127(36):12434–5.

105. Liu X, Thomas JR, Hergenrother PJ. Deoxystreptamine Dimers Bind to RNA Hairpin Loops. J. Am. Chem. Soc. 2004 Aug;126(30):9196–7.

106. Thomas JR, DeNap JCB, Wong ML, Hergenrother PJ. The Relationship between Aminoglycosides' RNA Binding Proclivity and Their Antiplasmid Effect on an IncB Plasmid. Biochemistry. 2005 May;44(18):6800–8.

107. DeNap JCB, Thomas JR, Musk DJ, Hergenrother PJ. Combating Drug-Resistant Bacteria: Small Molecule Mimics of Plasmid Incompatibility as Antiplasmid Compounds. J. Am. Chem. Soc. 2004 Dec;126(47):15402–4.

108. Hwang S. Discovery of a Small Molecule Tat-trans-Activation-responsive RNA Antagonist That Potently Inhibits Human Immunodeficiency Virus-1 Replication. Journal of Biological Chemistry. 2003 Jul 30;278(40):39092–103.

109. Mei H-Y, Mack DP, Galan AA, Halim NS, Heldsinger A, Loo JA, et al. Discovery of selective, small-molecule inhibitors of RNA complexes—1. The tat protein/TAR RNA complexes required for HIV-1 transcription. Bioorganic & Medicinal Chemistry. Elsevier; 1997;5(6):1173–84.

110. Disney MD, Seeberger PH. Aminoglycoside Microarrays To Explore Interactions of Antibiotics with RNAs and Proteins. Chem. Eur. J. 2004 Jul 5;10(13):3308–14.

111. Hofstadler SA, Sannes-Lowery KA, Crooke ST, Ecker DJ, Sasmor H, Manalili S, et al. Multiplexed Screening of Neutral Mass-Tagged RNA Targets against Ligand Libraries with Electrospray Ionization FTICR MS: A Paradigm for High-Throughput Affinity Screening. Anal. Chem. 1999 Aug;71(16):3436–40.

112. Goodsell DS, Lauble H, Stout CD, Olson AJ. Automated docking in crystallography: analysis of the substrates of aconitase. Proteins. 1993 Sep;17(1):1–10.

113. Hart TN, Read RJ. A multiple-start Monte Carlo docking method. Proteins. Wiley Online Library; 1992;13(3):206–22.

114. Totrov M, Abagyan R. Flexible protein-ligand docking by global energy optimization in internal coordinates. Proteins. 1997;Suppl 1:215–20.

115. Oshiro CM, Kuntz ID, Dixon JS. Flexible ligand docking using a genetic algorithm. J. Comput. Aided Mol. Des. 1995 Apr;9(2):113–30.

116. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, et al. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J. Comput. Chem. 1998;19(14):1639–62.

117. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. Journal of Molecular Biology. 1997 Apr 4;267(3):727–48.

118. Ewing TJ, Makino S, Skillman AG, Kuntz ID. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. J. Comput. Aided Mol. Des. 2001 May;15(5):411–28.

119. Kramer B, Rarey M, Lengauer T. Evaluation of the FLEXX incremental construction algorithm for protein–ligand docking. Proteins. Wiley Online Library; 1999;37(2):228–41.

120. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD. Improved protein-ligand docking using GOLD. Proteins. 2003 Sep 1;52(4):609–23.

121. Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. J. Comput. Aided Mol. Des. 1997 Sep;11(5):425–45.

122. Sheridan RP, Holloway MK, McGaughey G, Mosley RT, Singh SB. A simple method for visualizing the differences between related receptor sites. Journal of

Molecular Graphics and Modelling. 2002 Aug;21(1):71–9.

123. DeWitte RS, Shakhnovich EI. SMoG: de novo design method based on simple, fast, and accurate free energy estimates. 1. Methodology and supporting evidence. J. Am. Chem. Soc. ACS Publications; 1996;118(47):11733–44.

124. Gohlke H, Hendlich M, Klebe G. Knowledge-based scoring function to predict protein-ligand interactions. Journal of Molecular Biology. 2000 Jan 14;295(2):337–56.

125. Shoichet BK, McGovern SL, Wei B, Irwin JJ. Lead discovery using molecular docking. Current Opinion in Chemical Biology. 2002 Aug;6(4):439–46.

126. Blake JF, Laird ER. Chppter 30. Recent advances in virtual ligand screening. Annual Reports in Medicinal Chemistry. Elsevier; 2003;38:305–14.

127. Boehm H-J, Boehringer M, Bur D, Gmuender H, Huber W, Klaus W, et al. Novel Inhibitors of DNA Gyrase: 3D Structure Based Biased Needle Screening, Hit Validation by Biophysical Methods, and 3D Guided Optimization. A Promising Alternative to Random Screening. J. Med. Chem. 2000 Jul;43(14):2664–74.

128. Doman TN, McGovern SL, Witherbee BJ, Kasten TP, Kurumbail R, Stallings WC, et al. Molecular Docking and High-Throughput Screening for Novel Inhibitors of Protein Tyrosine Phosphatase-1B. J. Med. Chem. 2002 May;45(11):2213–21.

129. Paiva AM, Vanderwall DE, Blanchard JS, Kozarich JW, Williamson JM, Kelly TM. Inhibitors of dihydrodipicolinate reductase, a key enzyme of the diaminopimelate pathway of Mycobacterium tuberculosis. Biochim. Biophys. Acta. 2001 Feb 9;1545(1-2):67–77.

130. Schames JR, Henchman RH, Siegel JS, Sotriffer CA, Ni H, McCammon JA. Discovery of a Novel Binding Trench in HIV Integrase. J. Med. Chem. 2004 Apr;47(8):1879–81.

131. Sotriffer CA, Ni, McCammon JA. Active Site Binding Modes of HIV-1 Integrase Inhibitors. J. Med. Chem. 2000 Nov;43(22):4109–17.

132. Ruiz-Caro J, Basavapathruni A, Kim JT, Bailey CM, Wang L, Anderson KS, et al. Optimization of diarylamines as non-nucleoside inhibitors of HIV-1 reverse transcriptase. Bioorganic & Medicinal Chemistry Letters. 2006 Feb;16(3):668–71.

133. Jorgensen WL, Ruiz-Caro J, Tirado-Rives J, Basavapathruni A, Anderson KS, Hamilton AD. Computer-aided design of non-nucleoside inhibitors of HIV-1 reverse transcriptase. Bioorganic & Medicinal Chemistry Letters. 2006 Feb;16(3):663–7.

134. Thakur VV, Kim JT, Hamilton AD, Bailey CM, Domaoal RA, Wang L, et al. Optimization of pyrimidinyl- and triazinyl-amines as non-nucleoside inhibitors of HIV-1 reverse transcriptase. Bioorganic & Medicinal Chemistry Letters. 2006 Nov;16(21):5664–7.

135. Kim JT, Hamilton AD, Bailey CM, Domoal RA, Wang L, Anderson KS, et al. FEP-Guided Selection of Bicyclic Heterocycles in Lead Optimization for Non-Nucleoside Inhibitors of HIV-1 Reverse Transcriptase. J. Am. Chem. Soc. 2006 Dec;128(48):15372–3.

136. Irwin JJ, Shoichet BK. ZINC--a free database of commercially available compounds for virtual screening. J. Chem. Inf. Model. 2005 Jan;45(1):177–82.

137. Foloppe N, Matassova N, Aboul-ela F. Towards the discovery of drug-like RNA ligands? Drug Discovery Today. 2006 Nov;11(21-22):1019–27.

138. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. J. Med. Chem. 2004 Mar;47(7):1739–49.

139. Halgren TA, Murphy RB, Friesner RA, Beard HS, Frye LL, Pollard WT, et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. J. Med. Chem. 2004 Mar;47(7):1750–9.

140. Friesner RA, Murphy RB, Repasky MP, Frye LL, Greenwood JR, Halgren TA, et al. Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein−Ligand Complexes. J. Med. Chem. 2006 Oct;49(21):6177–96.

141. Lang PT, Brozell SR, Mukherjee S, Pettersen EF, Meng EC, Thomas V, et al. DOCK 6: Combining techniques to model RNA-small molecule complexes. RNA. 2009 May 20;15(6):1219–30.

142. Jain AN. Surflex-Dock 2.1: Robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. J. Comput. Aided Mol. Des. 2007 Mar 27;21(5):281–306.

143. Detering C, Varani G. Validation of Automated Docking Programs for Docking and Database Screening against RNA Drug Targets. J. Med. Chem. 2004 Aug;47(17):4188–201.

144. Li Y, Shen J, Sun X, Li W, Liu G, Tang Y. Accuracy Assessment of Protein-Based Docking Programs against RNA Targets. J. Chem. Inf. Model. 2010 Jun 28;50(6):1134–46.

145. Chen L, Calin GA, Zhang S. Novel Insights of Structure-Based Modeling for RNA-Targeted Drug Discovery. J. Chem. Inf. Model. 2012 Oct 22;52(10):2741–

53.

146. Morley SD, Afshar M. Validation of an empirical RNA-ligand scoring function for fast flexible docking using Ribodock. J. Comput. Aided Mol. Des. 2004 Mar;18(3):189–208.

147. Guilbert C, James TL. Docking to RNA via Root-Mean-Square-Deviation-Driven Energy Minimization with Flexible Ligands and Flexible Targets. J. Chem. Inf. Model. 2008 Jun;48(6):1257–68.

148. Roche O, Kiyama R, Brooks CL. Ligand−Protein DataBase: Linking Protein−Ligand Complex Structures to Binding Data. J. Med. Chem. 2001 Oct;44(22):3592–8.

149. Hu L, Benson ML, Smith RD, Lerner MG, Carlson HA. Binding MOAD (Mother Of All Databases). Proteins. 2005 Jun 21;60(3):333–40.

150. Smith RD, Dunbar JB Jr., Ung PM-U, Esposito EX, Yang C-Y, Wang S, et al. CSAR Benchmark Exercise of 2010: Combined Evaluation Across All Submitted Scoring Functions. J. Chem. Inf. Model. 2011 Sep 26;51(9):2115–31.

151. Damm-Ganamet KL, Smith RD, Dunbar JB Jr., Stuckey JA, Carlson HA. CSAR Benchmark Exercise 2011–2012: Evaluation of Results from Docking and Relative Ranking of Blinded Congeneric Series. J. Chem. Inf. Model. 2013 May 10;:130510104547004.

152. Dunbar JB Jr., Smith RD, Yang C-Y, Ung PM-U, Lexa KW, Khazanov NA, et al. CSAR Benchmark Exercise of 2010: Selection of the Protein–Ligand Complexes. J. Chem. Inf. Model. 2011 Sep 26;51(9):2036–46.

153. Foloppe N, Chen I-J, Davis B, Hold A, Morley D, Howes R. A structure-based strategy to identify new molecular scaffolds targeting the bacterial ribosomal A-site. Bioorganic & Medicinal Chemistry. 2004 Mar;12(5):935–47.

154. Gómez Pinto I, Guilbert C, Ulyanov NB, Stearns J, James TL. Discovery of Ligands for a Novel Target, the Human Telomerase RNA, Based on Flexible-Target Virtual Screening and NMR. J. Med. Chem. 2008 Nov 27;51(22):7205–15.

155. Cavalli A, Salvatella X, Dobson CM, Vendruscolo M. Protein structure determination from NMR chemical shifts. Proc. Natl. Acad. Sci. U.S.A. 2007 Jun 5;104(23):9615–20.

156. Wishart DS, Arndt D, Berjanskii M, Tang P, Zhou J, Lin G. CS23D: a web server for rapid protein structure generation using NMR chemical shifts and sequence data. Nucleic Acids Research. 2008 May 19;36:W496–W502.

157. Shen Y, Vernon R, Baker D, Bax A. De novo protein structure generation from

incomplete chemical shift assignments. J Biomol NMR. 2008 Nov 26;43(2):63–78.

158. Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, Liu G, et al. Consistent blind protein structure generation from NMR chemical shift data. Proc. Natl. Acad. Sci. U.S.A. 2008 Mar 25;105(12):4685–90.

159. Grzesiek S, Sass H-J. From biomolecular structure to functional understanding: new NMR developments narrow the gap. Current Opinion in Structural Biology. 2009 Oct;19(5):585–95.

160. Mielke SP, Krishnan VV. Characterization of protein secondary structure from NMR chemical shifts. Progress in Nuclear Magnetic Resonance Spectroscopy. 2009 Apr;54(3-4):141–65.

161. Prado FR, Giessner-Prettre C. Parameters for the calculation of the ring current and atomic magnetic anisotropy contributions to magnetic shielding constants: Nucleic acid bases and intercalating agents. Journal of Molecular Structure: THEOCHEM. Elsevier; 1981;76(1):81–92.

162. Wijmenga SS, Kruithof M, Hilbers CW. Analysis of (1)H chemical shifts in DNA: Assessment of the reliability of (1)H chemical shift calculations for use in structure refinement. J Biomol NMR. 1997 Dec;10(4):337–50.

163. Wishart DS, Case DA. Use of chemical shifts in macromolecular structure determination. Meth. Enzymol. Elsevier; 2002;338:3–34.

164. Cromsigt JA, Hilbers CW, Wijmenga SS. Prediction of proton chemical shifts in RNA–their use in structure refinement and validation. J Biomol NMR. Springer; 2001;21(1):11–29.

165. Frank AT, Horowitz S, Andricioaei I, Al-Hashimi HM. Utility of 1H NMR Chemical Shifts in Determining RNA Structure and Dynamics. J. Phys. Chem. B. 2013 Feb 21;117(7):2045–52.

166. Werf RM, Tessari M, Wijmenga SS. Nucleic acid helix structure determination from NMR proton chemical shifts. J Biomol NMR. 2013 Apr 6;56(2):95–112.

167. Barton S, Heng X, Johnson BA, Summers MF. Database proton NMR chemical shifts for RNA signal assignment and validation. J Biomol NMR. 2012 Nov 23.

168. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, et al. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. J. Am. Chem. Soc. 1995 May;117(19):5179–97.

169. Gelbin A, Schneider B, Clowney L, Hsieh S-H, Olson WK, Berman HM. Geometric parameters in nucleic acids: sugar and phosphate constituents. J. Am. Chem. Soc. ACS Publications; 1996;118(3):519–29.

170. Xu X-P, Case DA. Automated prediction of 15N, 13Cα, 13Cβ and 13C′ chemical shifts in proteins using a density functional database. J Biomol NMR. Springer; 2001;21(4):321–33.

171. Eichhorn CD, Yang S, Al-Hashimi HM. Characterising RNA Dynamics using NMR Residual Dipolar Couplings. Recent Developments in Biomolecular NMR. Royal Society of Chemistry; 2012;(25):184.

172. Tjandra N. Establishing a degree of order: obtaining high-resolution NMR structures from molecular alignment. Structure/Folding and Design. 1999 Sep 15;7(9):R205–11.

173. Tolman JR, Flanagan JM, Kennedy MA, Prestegard JH. Nuclear magnetic dipole interactions in field-oriented proteins: information for structure determination in solution. Proc. Natl. Acad. Sci. U.S.A. 1995 Sep 26;92(20):9279–83.

174. Tjandra N, Bax A. Measurement of dipolar contributions to 1JCH splittings from magnetic-field dependence of J modulation in two-dimensional NMR spectra. J. Magn. Reson. 1997 Feb;124(2):512–5.

175. Salmon L, Bouvignies G, Markwick P, Blackledge M. Nuclear Magnetic Resonance Provides a Quantitative Description of Protein Conformational Flexibility on Physiologically Important Time Scales. Biochemistry. 2011 Apr 12;50(14):2735–47.

176. Tolman JR, Ruan K. NMR Residual Dipolar Couplings as Probes of Biomolecular Dynamics. Chem. Rev. 2006 May;106(5):1720–36.

177. Getz M, Sun X, Casiano-Negroni A, Zhang Q, Al-Hashimi HM. Review NMR studies of RNA dynamics and structural plasticity using NMR residual dipolar couplings. Biopolymers. 2007;86(5-6):384–402.

178. Tjandra N, Bax A. Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. Science. 1997 Nov 7;278(5340):1111–4.

179. Clore GM, Starich MR, Gronenborn AM. Measurement of residual dipolar couplings of macromolecules aligned in the nematic phase of a colloidal suspension of rod-shaped viruses. J. Am. Chem. Soc. ACS Publications; 1998;120(40):10571–2.

180. Hansen MR, Mueller L, Pardi A. Tunable alignment of macromolecules by filamentous phage yields dipolar coupling interactions. Nat. Struct. Biol. 1998 Dec;5(12):1065–74.

181. Hansen MR, Hanson P, Pardi A. Filamentous bacteriophage for aligning RNA, DNA, and proteins for measurement of nuclear magnetic resonance dipolar coupling interactions. Meth. Enzymol. 2000;317:220–40.

182. Lindorff-Larsen K, Best RB, Depristo MA, Dobson CM, Vendruscolo M. Simultaneous determination of protein structure and dynamics. Nature. 2005 Jan 13;433(7022):128–32.

183. Chen Y, Campbell SL, Dokholyan NV. Deciphering Protein Dynamics from NMR Data Using Explicit Structure Sampling and Selection. Biophysical Journal. Elsevier; 2007 Oct 1;93(7):2300–6.

184. Markwick PRL, Bouvignies G, Blackledge M. Exploring Multiple Timescale Motions in Protein GB3 Using Accelerated Molecular Dynamics and NMR Spectroscopy. J. Am. Chem. Soc. 2007 Apr;129(15):4724–30.

185. Showalter SA, Brüschweiler R. Quantitative Molecular Ensemble Interpretation of NMR Dipolar Couplings without Restraints. J. Am. Chem. Soc. 2007 Apr;129(14):4158–9.

186. Showalter SA, Hall KB. Isotropic reorientational eigenmode dynamics complements NMR relaxation measurements for RNA. Meth. Enzymol. Elsevier; 2005;394:465–80.

187. Duchardt E, Nilsson L, Schleucher J. Cytosine ribose flexibility in DNA: a combined NMR 13C spin relaxation and molecular dynamics simulation study. Nucleic Acids Research. 2008 May 23;36(12):4211–9.

188. Ferner J, Villa A, Duchardt E, Widjajakusuma E, Wohnert J, Stock G, et al. NMR and MD studies of the temperature-dependent dynamics of RNA YNMG-tetraloops. Nucleic Acids Research. 2008 Feb 5;36(6):1928–40.

189. Hall KB. RNA in motion. Current Opinion in Chemical Biology. 2008 Dec;12(6):612–8.

190. Hall KB, Tang C. 13C relaxation and dynamics of the purine bases in the iron responsive element RNA hairpin. Biochemistry. ACS Publications; 1998;37(26):9323–32.

191. Mackerell AD. Empirical force fields for biological macromolecules: Overview and issues. J. Comput. Chem. 2004;25(13):1584–604.

CHAPTER 2

MATCH: An Atom-Typing Toolset for Molecular Mechanics Force Fields

2.1 Introduction

The increasing availability of computing resources is reshaping the researcher's approach in the utilization of molecular simulations for the modeling of proteins, nucleic acids, their ligands and inhibitors. It is now feasible, with the growth of large libraries of drug-like compounds, to investigate receptor-ligand binding poses and other properties using molecular mechanics force fields in a high throughput manner(1). A significant barrier in this process, however, is the accurate generation of explicit inter- and intra-molecular parameters for novel potential drugs that are consistent with the biomolecular force field utilized in modeling the other components of the system(2, 3). Modern force fields such as CHARMM(4, 5), AMBER(6, 7) and OPLS(8) rely on empirical parameters and have been developed to yield accurate modeling of conformational changes and non-covalent interaction energies for protein and nucleic acids(9, 10). However, these force fields do not contain all the required parameters to represent drug-like molecules for studying receptor-ligand interactions.

Developers of the biomolecular force fields have adopted different strategies to optimize the bonded and nonbonded parameters and attempt to reproduce experimental data or quantum mechanical properties of model compounds. Therefore, it is unlikely that the combination of a particular biomolecular force field with an arbitrary ligand force field would yield properly balanced intermolecular interactions. Rather, it is crucial that the small molecule parameters follow a similar parameterization scheme to that which was used to develop the biomolecular force field(11). The most straightforward parameters that can be generalized to novel compounds are those associated with the intra-molecular energy terms (e.g. equilibrium values and force constants for bond lengths and angles as

well as optimal torsion angles and the respective barrier heights). Similarly, van der Waals parameters, the atomic radii r_i and energy well-depth ε_i, are often successfully transferred among analogous atom types(12). While small changes in these parameters might significantly impact the energy, they are not particularly sensitive to the bonded environment of the molecule. In contrast, the partial charges that are associated with each atom are the primary components in the electrostatic energy terms and are significantly challenging to transfer from one molecule to another due to their dependence on both bonded and nonbonded chemical environment.

Two main strategies have been suggested for generating partial charge assignments that are compatible with current biomolecular force fields. In one fixed-charge strategy, charges are adopted for an entire molecule, often based on *ab initio* calculations or parameterized methods that mimic these charge distributions. A restrained electrostatic potential (RESP) charge fitting procedure is advised for assigning partial charges to novel ligands in a manner that is consistent with the Generalized AMBER force field (GAFF)(3). Antechamber(13), an auxiliary program in the AMBER molecular modeling suite of programs that uses coordinate and connectivity information to assign atom and bond types for ligands based on atom and bond type definition tables, can generate charges using RESP, AM1 Mulliken, AM1-BCC, CM2 or Gasteiger charge methods.

In an alternative other fixed-charge strategy, used by the CHARMM and OPLS family of force fields, charge distributions of a molecule are built-up from charges assigned to the component fragments of the molecule. Halgren(14), in developing the MMFF94 force field, proposed bond charge increment “rules” in which optimal charges are determined for fragments of molecules and these fragments are then pieced together to construct charge distributions for novel compounds. Several programs exist to assign atom types and atomic partial charges based on the bonded environment of the atom. These automated assignment programs convert a three-dimensional structure file into a representation of the bonded environment, such as a connectivity table of atoms and bonds. Patterns within this connectivity table are identified as fragments for which atom-types and partial charges are associated. These programs differ in how the bonded

environment is determined, how the specific "rules" are defined for matching the fragments in the new molecule with those of "known" fragments, and how the partial charges are distributed throughout the molecule. For example, the molecular modeling package IMPACT(15) accepts a PDB file format and automatically assigns atom types and parameters for a wide range of organic molecules that are consistent with the OPLS_2003 force field. For the entire molecule, the partial atomic charges are assigned by distributing any formal ionic charges over one or more atoms and then adding contributions from the bond charge increment (BCI) parameters associated with the chemical bonds. PRODRG(16, 17), through its web interface, generates molecular topologies from a coordinate file and assigns partial charge distributions from a molecule's constitutive fragments for use with the GROMOS force field(18).

Recent developments in the CHARMM community have led to the generation of small molecule parameters in a novel force field denoted CHARMM General Force Field (CGENFF). Although a notable step in the right direction, the chemical space covered by the ~400 molecules in CGENFF is limited and still requires manual efforts to extend it to new molecules. Our preliminary goal with this work was to develop a publicly available solution for generating parameters for novel molecules that are consistent with the CHARMM parameterization scheme. We were interested not only in creating a way to process molecules en masse within CHARMM, but also to develop a tool to investigate the merits of different types of parameterization choices and strategies. We developed a general approach that extracts rules for both charging and parameterization based on a library of topology and parameter files for an existing biomolecular force field. This scheme then allows for the fragments comprising existing parameters to be applied or extrapolated to novel molecules in a fashion consistent with the parameterization strategy or philosophy within a given biomolecular force field. We have focused our efforts on CHARMM; however, this approach and the MATCH toolset that we have developed can be used to extract rules for charging and parameterization based on any biomolecular force field.

A fundamental feature of the MATCH algorithm is the representation of molecular structures as mathematical graphs. Chemoinformatics has benefited greatly from the representation of chemical structure as graphs, such as in ring identification and characterizing chemical connectivity(19). In particular, structure and substructure searching of chemical databases, such as those performed on inventory or patent databases, automated retrosynthetic analyses, property prediction, quantitative structure-activity/toxicity/property relationship analyses, visualization, and similarity/diversity analyses are applications with chemical pattern recognition solutions(20-23)The unique chemical environment defining an atom type can also be depicted in graph form, enabling a chemical characteristic comparison between a library of known atom type definitions and atoms in a novel compound. Furthermore, additional operations are vastly simplified while functioning within a graph reference frame, including: quantification of the similarity of chemical environment between atom types within a given force field, atomic ring identification, and identification of atoms requiring improper angles to enforce accurate geometry. Much of the next section will be devoted to discussing the implementation of mathematical graphs within MATCH.

At present, we will demonstrate the utility of MATCH and discuss the primary components comprising the software package. MATCH supports every force field currently implemented in CHARMM36 (i.e., force fields specific for proteins, nucleic acids, lipids, carbohydrates, ethers and model small molecules). Here, we demonstrate two primary functions of the MATCH toolset. First, we show how MATCH is used to extract fragment-based atom types and associated bond charge increment rules. More specifically, we discuss how MATCH constructs libraries that contain definitions for the chemical environments described by the force field topology files for a given force field as well as the schemes for assigning partial charges to these atom type definitions. Second, we show how MATCH is used to generate force-field specific MATCH libraries. These libraries are shown to be self-consistent with existing CHARMM force fields by their ability to reproduce atomic charges contained in the force field that was used to infer the rules. In addition, the viability of parameter substitution to determine missing parameters and thereby enabling complete parameterization will be demonstrated through

a leave-one-out substitution study. Our benchmark for the transferability of rules learned from a force field will be the charging and parameterization of the molecules in other existing force fields. We then directly compare computed values to existing ones and measure how well the results are correlated. This exhaustive exercise, using each existing CHARMM force field to charge and parameterize the other, demonstrates the ability of the methods implemented in MATCH in generalizing a force field representation. Finally, the parameterization of one million small molecules from the PubChem database(24) with the implementation of the CGENFF-based libraries within MATCH illustrates the scope and potential of MATCH in real world usage.

2.2 MATCH Strategies and Components

MATCH is a suite of tools that has been developed for constructing molecular fragment-based libraries and BCI rules to be utilized for the extension of a given biomolecular force field. There are two distinct applications of the MATCH toolkit: i) the utilization of atom-type molecular fragment and BCI rule libraries in the charging and parameterization of novel molecules and ii) the tools required to assemble these libraries as well as the generation of rules to allow substitution of parameters to assist in the parameterization of new molecules. The procedure in which MATCH extends a force field to a novel molecule is displayed in **Figure 2.1**. Development of the MATCH libraries of fragments for atom typing and bond increment rules are illustrated in **Figures 2.2 and 2.3** respectively. Here, we explore the ability of MATCH, with some expert intervention, to effectively construct force field specific MATCH libraries, which is to "learn" atom type definitions and bond charge increment rules from multiple CHARMM force fields.

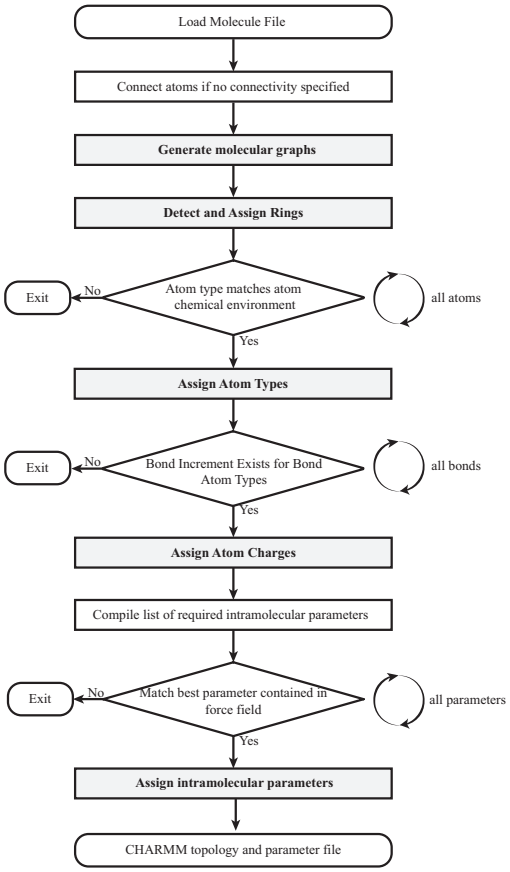

Figure 2.1. Overview of the MATCH algorithm. All major algorithm components discussed in the paper appear in bold.

We also investigate the ability of MATCH to use these libraries and the substitution rules to parameterize molecules in different force fields.

2.2.1 Molecular graphs

Molecular graphs are assembled using the supplied connectivity information (CONECT lines in a PDB file, a CHARMM RTF file, a bond list for MOL2, MOL, and SDF, etc) or predicted using the atomic coordinates and bonding rules based on atomic radii. As described by Downs and coworkers(25) , molecular graphs are constructed as labeled, directed, connected graphs, where each atom is represented by a vertex and stores information about itself including element, number of bonds, ring membership, and pointers to neighboring atoms. For small molecules (less then 10 atoms) a graph represents the entire molecule; however, due to the computational expense of constructing these graphs for larger molecules, a limitation was imposed such that atoms greater than 10 bonds away are not included in the definitions of the chemical environment of a given atom. This limitation was arbitrarily set to 10 as a compromise between accuracy and efficiency. In fact, the chemical space characterization of all atom types in this study does not extend further than 3 to 4 bonds away (see atom type declarations in MATCH). The molecular graph is then expanded following a breadth-first algorithm(26). Starting from one atom, each atom to which it is bonded in then added to the tree; atoms to which they are bonded and are not yet included in the tree are then added and so on until all atoms in the molecule are either represented in the tree or 10 bonds away. The end result is a branched data structure that allows for tree comparisons and other operations, which are crucial to the workings of MATCH. The process is

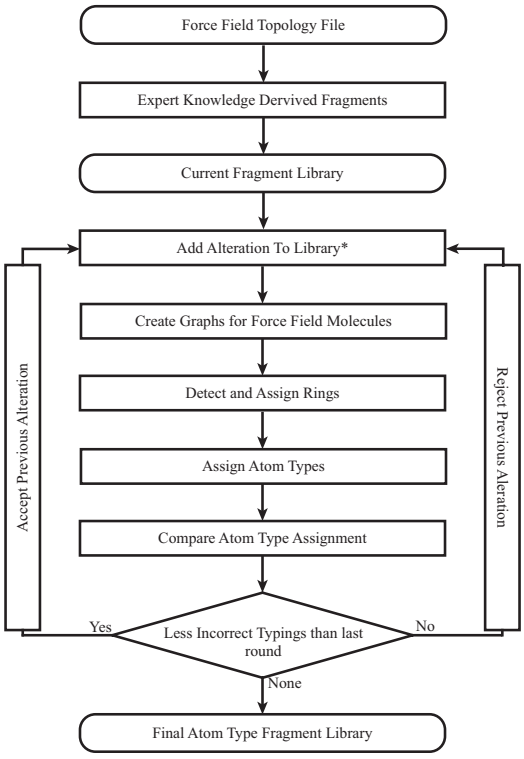

Figure 2.2. Overview of the process of developing atom type molecular fragments for a given force field, which is the basis of MATCH's atom typing engine.

repeated for each atom as the head vertex. While a bond is not normally considered directional, an artificial directionality is imposed by this representation and is harnessed in algorithms that will be discussed later. For clarity, atoms occurring higher in the depth of the tree are considered the parent, while bonded atoms that are added beneath are considered to be children. The only exception to this definition is in cyclic compounds, where two connected atoms may be positioned such that they are at the same depth. In this case the first atom to be traversed is considered to be the parent of the other.

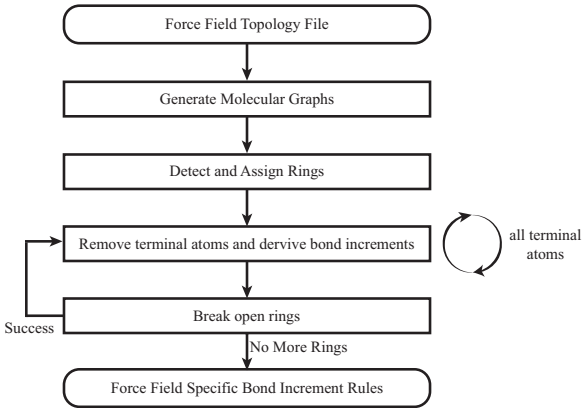

Figure 2.3. Overview of the process of extracting the bond charge increment rules for a given force field.

Calculating whether one molecular graph is similar enough to another to be considered a "match" is a fundamental process in atom typing. The procedure to do this is straightforward: For two graphs to be considered a match, each node of the smaller graph (i.e., the atom type fragment) must exist within the larger graph (i.e., the molecular graph within the new molecule with the current atom being the head node) with the same connectivity. The proceduFre is analogous to a typical tree data-structure comparison in which the comparison is initiated at the head nodes. Confirming a match is a two-step process: first, features such as ring membership, aromaticity, etc, of the nodes of the smaller graph must be contained in the nodes of the larger graph. Second, the element and number of bonds of each node must be consistent. This process continues until the smaller graph has all of its nodes matched or until one node is unable to be matched to a node in the other graph. Occasionally, there are two possible matches for a node; when this occurs, the children of both potential matches are compared to the node's children in a recursive manner until a difference is identified or the graphs are found to be identical.

2.2.2 Ring Detection

Identification of ring membership is crucial in the atom typing process of MATCH due to the specificity of atom types that are only found in rings. Ring discovery has received considerable attention in the literature because of its computational demands(19). Much

of the algorithmic development in this area has focused on the identification of subsets of rings that have particular meaning in some applications, for example in the analysis of synthetic pathways(27). In this situation, exhaustive enumeration of rings is required for accurate atom typing due to the fact that atom types are ring specific. The algorithm developed here relies heavily on the use of molecular graphs discussed earlier and is based on the works described by Tiernan(28) on mathematical graph circuit detection. The elements of our ring detection algorithm are as follows: each heavy atom with more than one bond is considered in turn unless the atom has already been detected as being part of a ring. The ring detection algorithm is a breadth-first search(26) that traverses the molecular graph of the atom being considered. During each iteration of the search, each current path is extended to new heavy atoms, the path that is currently the closest in level to the starting point will always be selected to be followed first (see **Figure 2.4**). Upon reaching the start atom with the path containing more than two atoms, successful termination is reached and all atoms that were traversed along the path are marked with ring membership. Failed termination is reached when each path covers more than 50% of the molecule and is more than 50% away (by depth) from the start, making it impossible to successfully return to the starting point.

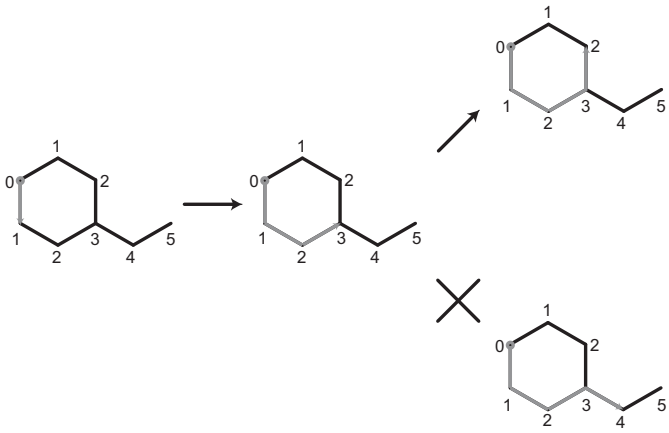

Figure 2.4. Depiction of the usage of the molecule graphs to our advantage in determining rings.

This algorithm is very fast, requiring only one atom in a ring to be searched. It also prevents duplicate identification of rings. While other ring detection algorithms have been shown to be more efficient(19), this algorithm was selected because of its reliance on molecular graphs. In fact, most of the computational efforts in MATCH are for the construction of the molecular graph and the typing, charging and parameterization of novel molecules. Therefore, ring detection is not the computational bottleneck of

MATCH and, thus, the decision to implement this ring detection algorithm does not hinder the performance.

2.2.3 Molecular Fragment-based Atom Typing

Converting molecules into mathematical graph form enables the direct comparison of the local chemical environment of one atom to another and the quantitative evaluation of the similarity between the two structures. Continuing with this ideology, the chemical space that defines an atom type can be represented as a molecular fragment. We define a molecular fragment as a group of connected atomic nodes that contain the required atomic features that describe the chemical space of the atom type (i.e., atomic element, number of bonds, ring membership, etc). These molecular fragments have the same properties as the graphs that are built for actual molecules and thus can be compared in a similar procedure. Adopting this philosophy of representing distinct chemical space as a molecular fragment reduces the atom typing process to one of tree comparison, in which the largest molecular fragment that completely matches an atom's molecular graph is assigned to that atom. The library of atom type molecular fragments is preserved in super smiles string format with similarities to the implementation by Bone et al(29).

In super smiles string notation, each atom is represented by its chemical element plus its number of bonds. More specific information is appended to the end of the string. Examples of super smiles format are displayed in Table 1. The "!" attribute denotes no ring membership whereas the "%" attribute indicates ring membership and is followed by the ring size and aromaticity. Connectivity between atoms is denoted by parentheses, where atoms within a parenthesis are bound and considered to be children to the one outside. The decision to describe atom type molecular fragments in this representation is to allow for effortless management of MATCH atom type force field libraries. It is a straightforward process to modify or add new atom types to an existing library or to create entirely new ones. In this study, we demonstrate how atom types in the CHARMM force fields can be represented by molecular fragments. Certainly, this strategy can be extended to represent other biomolecular force fields.

Smiles String	Atoms Matched
*	Any Atom
C.4	Carbon atoms with 4 bonds
!N.3	Aliphatic Nitrogens atoms with 3 bonds
O	Any Oxygen atom
[^C]	Not a Carbon atom
S.2%	Sulfur ring atoms with 2 bonds
C.3%6,6	Carbons atoms with 3 bonds in 2 6
N.3%6A	Nitrogen atoms with 3 bonds in a 6 membered aromatic ring
C.4%5N	Carbon atoms with 4 bonds in a 5 membered non-aromatic ring

Table 2.1. Examples of the syntax of the super smiles strings used to represent atoms within MATCH encoded molecular fragments.

2.2.4 Bond Charge Increment Rules

Inspection of atomic charges in commonly used force fields suggest that they often follow bond charge increment rules(5). Bond increments are a description of the magnitude and direction of charge of the covalent bonding of two atoms. Decomposing the atomic charges of a molecule into these increments yields a set of generalized rules based on the type of the atoms in the bond. Once these rules are identified they can be extended and applied to new molecules.

Development of bond charge increment rules has been implemented in the past for other force fields such as in MMFF94. In their approach they globally optimized a set of rules through an iterative process that best fit the training set(14). While this is a valid approach and was considered when investigating charging rules in MATCH, it was discarded due to the inability to precisely reproduce the training charges in force fields such as CGENFF (data not shown). In addition, our goal is consistency: to preserve the charging rules found in the protein and nucleic acid force fields as much as possible. Our

approach is consistent with our fragment-based atom typing procedure and accurately reproduces partial charge assignments in all CHARMM force fields.

Empirical force fields generally reuse atom types for bonded parameter assignments despite slightly different charge distributions. For example, in the CHARMM protein force field most methylenes transfer -0.09 electron units of charge from each of the two aliphatic hydrogens onto the adjacent carbon; however, for methylenes adjacent to a primary ammonium this increment is -0.05 electron units, despite the fact that identical atom types are used to describe the respective carbon and hydrogen atoms. This discrepancy is dealt with in MATCH by conducting a secondary level of atom typing. During the development of BCI from a force field, if there are multiple solutions exist for a given pair of bonded atom types, the most frequently used BCI is stored as the default increment rule and the infrequent ones are stored separately as refining increment rules. These refining increments are associated with a molecular fragment as found in the atom typing process. This fragment is the description of the chemical environment that is correlated with the divergent increment rule. During the charging procedure MATCH considers the refining increment rules and checks to see that the corresponding chemical environment matches the current local connectivity of the molecule: if so, the refining increment rule is applied instead of the default rule.

The simplest case to describe the magnitude and direction of charge in the covalent bonding of two atoms involves a terminal atom, that is, one of the atoms has only one covalent bond. In a neutral system, it is straightforward to determine the bond increment between a terminal atom and its bonding partner: it must be the terminal atom's assigned atomic charge balanced by an equal and opposite signed charge assigned to its bonding partner. For example, most aliphatic hydrogen atoms bound to aliphatic carbons have a charge of 0.09 in the CHARMM force field, and this yields charge increments of +0.09 and -0.09 to the hydrogen and carbon atoms, respectively. Unfortunately, as the number of bonds increase it becomes increasingly difficult to deconvolute the charge relationship in each bond. The solution we adopted is to disassemble a molecule by removing terminal atoms and subtracting their respective bond charge increments. Returning to the

example of the aliphatic hydrogen atoms, the procedure would involve removing the terminal hydrogen atoms and subtracting the charge from the BCI (i.e., subtracting -0.09 e^-) from the remaining carbon atom. This process effectively nullifies the bond between the hydrogen and carbon and reduces the number of bonds in which the carbon atom participates. If the carbon atom is a methyl carbon, it would now be rendered a terminal atom with a charge reduced by -0.27 e^- charge. This iterative procedure of nullifying bonds and adjusting the charges allows the parameterization of a large portion of the BCIs. However, atom types that are exclusively contained in rings fail to yield BCIs through this method. Thus, two follow-up algorithms have been implemented to deal with ring atom types. The first algorithm is used for rings with "symmetry" points in which one atom is bonded to two atoms with the same type. In this case it is possible to break the ring at this "symmetry" point and establish the bond charge increment rules by assuming that each of the bonds contributes exactly half of the charge of the atom bonded to both of them. The second algorithm is used for rings in which no symmetry exists for any ring atom; in this case, a previously determined BCIs is used to break the ring. If all the charge is accounted for by existing rules then it is accepted as the correct increment. Using the methods described above it is possible to delineate the vast majority of bond charge increment rules for a given force field. The minority of compounds in a force field whose BCIs can not be deconvoluted with these methods can be examined on a case-by-case basis. In these situations, it is usually possible to take existing increment rules and apply them to these molecules checking to see if all the charge is accounted for.

2.2.5 Parameter Generation

For a molecule to be successfully represented by a force field, it requires intramolecular parameters for the bond, angle and dihedral energy contributions and intermolecular parameters: atomic partial charges and van der Waals parameters describing the nonbonded energy contributions. Assignment of atomic charges was covered in the previous section; we will now discuss generation of the remaining parameters. Producing all required bond parameters for a novel compound in MATCH is trivially accomplished by removing duplicates from the list of bonded atom types that was already acquired during the process of assigning atomic partial charges and by identifying the corresponding parameters for these bonded atom types in the parameter file. To produce

the required angle parameters, each bond is traversed and the neighboring bonded atom is added to each side, growing out a bond into an angle. The same procedure is repeated with angles to obtain all required dihedral angles. The required parameters are then added to the new compound's own parameter file. Parameters that do not exist in the parent force field parameter file are generated via substitution of the best-fit parameter.

No force field contains parameters of all chemical space; therefore, means to "interpolate" within or "extrapolate" beyond the parameterized chemical space are necessary. Our solution is to identify existing parameters that "best fit" the required parameter through a form of parameter substitution. Upon examination of atom types in a given force field it is apparent that some types are more similar than others. For example, investigating the CHARMM36 protein force field, it is evident that a correlation exists between atom types. CT1, an aliphatic carbon bonded to one hydrogen, shares many of the same bond, angle, and dihedral parameters as CT2, also an aliphatic carbon but bonded to two hydrogen atoms. This is unlikely coincidental: atom types that share a similar chemical space should also have similar bonded parameters. From careful analysis of the chemical space of atom types basic rules can be derived. First, aliphatic types behave more similarly to other aliphatic types then they do to atoms that have ring membership. Second, the number of bonds a type has also affects how similar parameters are to each other: types that share the same number of bonds have more similar angle and dihedral parameters. Keeping these basic observations in mind, it is possible to create a score describing how one type is related to another based on comparison of the molecular fragment representations used for typing. The use of this substitution method vastly increases the number of molecules that can be assimilated into the working force field. A brief overview of how the relatedness between types is built is: the molecular fragment of each atom type is compared and the overlap between the two is computed. Special penalties are put in place to distinguish the score of atom types only found in rings from atom types that are only found in aliphatic chains and the reverse case. These scores are preserved in text format that may also be altered by users if desired.

Substitution is available during both the atom charging and parameter generation stages in MATCH. In both cases the procedure is equivalent. For example, if there is a bond between atom type A and B, but the corresponding bonded parameters do not exist in the force field parameter file, the relation matrix will be queried. Each existing bond parameter is scored in the simple fashion of how closely its first atom type is related to A and how closely its second is related to B; the reverse is also considered. The relatedness is 1 if the atom types are the same and is 0 if the two atom types have neither the same element nor the same bond number; the summation of the relatedness of each pair is the score. The bond parameters with the highest score are selected as the substitution parameters for this new pair of bonded atom types.

2.2.6 Program Organization

MATCH supports a wide variety of molecular formats, (PDB, MOL2, MOL, SDF, RTF) and exports CHARMM formatted PDBs for files supplied in non-PDB formats. The core algorithms of the MATCH toolset have been implemented in Perl. Perl was chosen to maximize portability. MATCH is a small package that utilizes PerlChemistry, another package for more general applications of identifying similar connectivity between molecules, renaming atoms given a chemical environment (e.g., the naming convention used in RESP(7)), or averaging charges over atoms with identical connectivity. PerlChemistry is a set of Perl packages that provides object representations such as atom, bond and molecule. This distinction was planned so PerlChemistry can exist independently from MATCH and allows users to have access to the molecular graph API. Several examples of applications of molecular graphs are provided as part of the distribution.

The key properties of the MATCH package are contained in a single Perl package called MATCHer.pm, which allows users to write additional scripts to facilitate any of the algorithms discussed in this paper in the context of other force fields. The default script, MATCH.pl, provides the core functions of atom typing, charging and parameter determination that are associated with the processes depicted in **Figure 2.1**. All the MATCH force field libraries discussed in this paper are included in the current version of

the MATCH package. The MATCH package is supplied together with basic usage instructions and can be downloaded at brooks.chem.lsa.umich.edu/software.

2.3 Methods

2.3.1 Constructing Force Field-Specific MATCH Libraries via MATCH

Force field-specific MATCH libraries were constructed via MATCH based on the CHARMM36 topology files: top_all22_prot, top_all27_na, top_all35_carb, top_all35_ethers, top_all36_cgenff and top_all36_lipid. For each force field the molecular fragments for each atom type were constructed through an iterative optimization procedure. Using a given force field the goal is to correctly assign types for all the atoms within the force field. The main concern in this process is to avoid mistyping by incorrectly making one type cover the space of another. To avoid this, atom types were grouped together by the atom element and bond number and were developed simultaneously. That is, each time there was a modification of a fragment, each atom that was of the group's element and number of bonds was typed and if there were fewer mistypings this change was accepted. This was repeated until there were no mistypings. Most aliphatic atom types have rather distinct chemical space and, thus, required a few rounds of optimization. On the other hand, it was more difficult to create the optimal set of fragments for atom types that are exclusively based in rings and, thus, these atom types required multiple rounds of optimization. The Perl script TestBuildTypeStrings.t that is required for this optimization is provided in the MATCH package distribution for future optimizations and development of atom-type fragments for new force fields. Another challenge in this optimization scheme is keeping the atom-type fragments as general as possible while preserving their unique chemical environment.

For each force field that contained residue patches, each patch was applied if it increased the chemical space of the set (i.e., added new atom types or bond increment rules) or was necessary to correct polymer connectivity. By default, the NTER and CTER patches were applied to the protein force field residues and the 5TER and 3TER patches were applied to the nucleic acid force field residues. With the exception of CGENFF, all molecules in the topology files were included in the process of constructing the force field-specific

MATCH libraries. In total, 53 of the 415 molecules in the CGENFF topology file were eventually excluded because the required number of refinement increments was too large (i.e., did not obey any of the default bond increment rules), of the typing of atoms could not be performed by fragments (atom types CG2DC1 and CG2DC2). A list of the CGENFF excluded molecules appears in the supplementary material.

Bond increments were extracted from each force field topology file in an automated fashion as discussed in the previous section and can be run in MATCH using GenerateBondIncrementRules.pl. Refinement bond increments were added to fix obvious exceptions to the BCIs, e.g., where the default BCIs could not reproduce the charge distributions in the molecules, and were usually small in number, with exception of CGENFF. In addition to the compounds that were excluded when constructing the CGENFF-specific MATCH libraries, several other compounds in the CGENFF topology file do not obey clear bond increment rules. With additional refinement rules, however, it was possible to reliably reproduce charges for these compounds.

2.3.2 Extrapolating and Interpolating Force Field Parameters via MATCH Libraries

Both the self-validation and cross-validation of atomic charge was conducted with the same procedure (TestIncrements.pl). To assess the ability of the MATCH libraries to extrapolate and interpolate to new contexts, MATCH libraries of force field A were used to assign charges to the atoms of each molecule in force field B. Molecular graphs of each molecule in B were constructed and each molecule object was duplicated, but with all atom types and charges removed. Each molecule copy was then typed using the MATCH libraries based on force field A's atom type molecular fragments. If any of the atoms could not be typed, the algorithm proceeded to the next molecule. Upon successful completion of the atom type assignments, BCIs were applied to assign atomic charges. The differences between the original and assigned charges for atoms in molecules that were successfully charged were computed. For the self-validation analyses, A and B were the same force field. A similar procedure is in place for comparing the atomic parameters of one force field compared to another and is performed using TestParameters.pl. Analysis was also completed on atoms that could be completely charged/parameterized regardless of whether its entire molecule could be (**Table S1** and **S2**).

2.3.3 High-Throughput Small Molecule Parameterization

PubChem small molecules were obtained from the PubChem database in MOL2 format. Since submission of molecules is random and we are interested in the chemical space that can be covered using MATCH we took the first million of the ~26 million molecules that met the following criteria: molecular weight <600 Da and contained exclusively elements H, C, N, O, F, P, S, Cl, Br, and/or I. Molecules that fit these criteria were processed by MATCH using the CGENFF force field-based MATCH libraries. If force field generation was successful, the molecule was minimized in CHARMM using steepest descent minimization for 100 steps with nonbonded cutoffs that are defined in the protein force field. Finally, the RMSD between the minimized structure and original structure was calculated.

2.4 Results and Discussion

2.4.1 Recapitulating Bond Charge Increment Rules

This novel suite of MATCH tools includes facilities to aid in developing force field-specific MATCH libraries that are learned from a given biomolecular force field, and to generate sets of parameters for novel compounds that are consistent with this force field in a short amount of time. Here, we explore the ability of MATCH with some expert intervention to effectively construct force field specific MATCH libraries, which is to "learn" atom type definitions and bond charge increment rules from multiple CHARMM force fields. We also investigate the ability of MATCH to use these libraries and the substitution rules to parameterize molecules in different force fields. **Figure 2.5** summarizes the results for the MATCH-assigned partial charges for each atom compared to that in the original CHARMM topology file. The plots along the diagonal in **Figure 2.6** represent the results for the self-consistency study in which MATCH re-predicts the properties of the force field on which the MATCH libraries were based. The off-diagonal plots represent the results from the cross validation study in which MATCH interpolates and extrapolates parameters and atom type assignments from a different force field.

First, the results from the self-consistency study demonstrate that MATCH successfully recapitulates the atomic charges in the CHARMM topology files. It would be difficult to get such excellent agreement without also perfectly capturing the correct atom types. Most of the correlations have an R^2 ~1.00 with the exception of the carbohydrate force field and CGENFF. In the carbohydrate force field there is a small discrepancy for the increments between atom types: CC2O3 and OC2D3, which exists in both D-Psicose and ketose. In the original topology file, the assigned charges for these atoms are quite

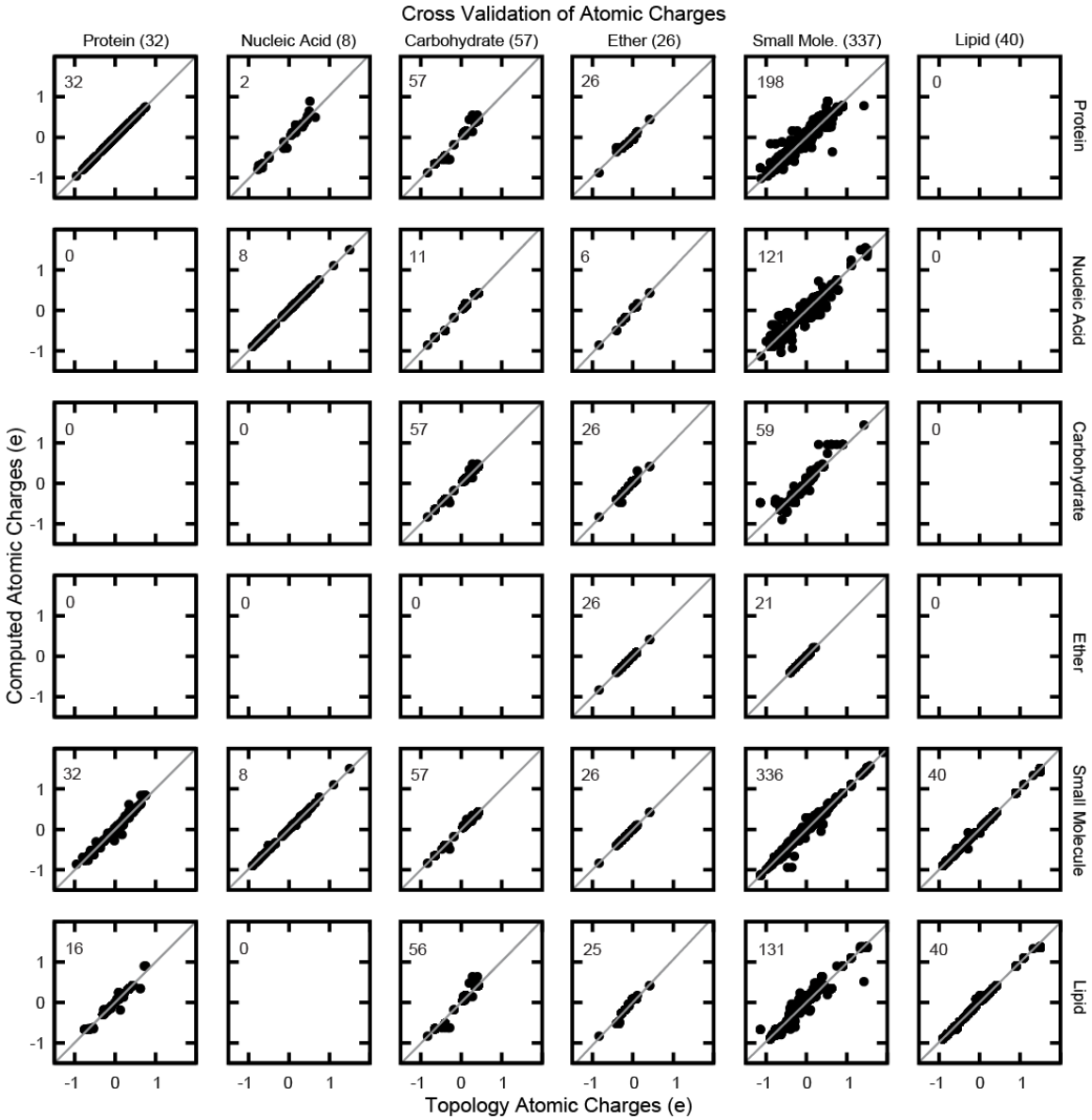

Figure 2.5. The x-axis denotes the reference force fields while the y-axis is the force field libraries within MATCH. The numbers in the top left corners of each graph indicate the number of molecules that were successfully charged using a given MATCH library.

different from each other even though the chemical environment appears identical, leading MATCH to learn radically different bond increment rules. This makes it impossible to create a refining bond increment rule to assign different increments in each case as the application requires a unique chemical environment to discriminate. As mentioned in the Methods section, CGENFF molecules whose partial charge assignments were clearly not consistent with bond increment rules were omitted from the "learning" phase in which the CGENFF-based MATCH libraries were constructed. Consequently, charges for these molecules are not reproduced exactly, but are still of very high quality such that the overall R^2 is 0.997.

Second, the cross-validation study illustrates how MATCH libraries can be extended to generate parameters for novel compounds. Starting with the protein force field, which contains only atom types and bond increment rules designed for the amino acids, the rules are successfully generalized out to a significantly larger chemical space. This is illustrated primarily in typing and charging CGENFF molecules with the top_all22_prot force field, in which over 59% of the molecules in CGENFF were successfully processed in MATCH. The R^2 correlation between the MATCH-assigned atomic charges and those found in the CGENFF topology file is 0.941. The average unsigned error is 0.024 electron units while the percentage unsigned error is 15.99%. While the ability of the MATCH libraries based on the CHARMM protein topology file to be successfully extended to other small molecules is very promising, it is worth considering why certain CGENFF molecules were unable to be processed in MATCH. The most significant contributor is the lack of necessary atom types for atoms of elements that are not included in the protein force field. Almost 19% of CGENFF, that is 64 molecules, contain elements P, F, Cl, Br, I, and these elements are not present in the protein force field. The remaining CGENFF molecules that could not be processed with MATCH failed because it was not possible to construct a substitution for a necessary bond increment rule to complete the atomic charges. While the substitution rules can be further generalized, the quality of both atomic charges and parameters will suffer.

The MATCH libraries based on the protein force field successfully processed all of the molecules contained in the carbohydrate and ether force fields. The R^2 correlation between the MATCH-assigned and original atomic charges was 0.998 and 0.985 for the carbohydrate and ether force fields respectively. The average percentage error for molecules in the carbohydrate force field was 2.58% while that for molecules in the ether force field was 14.21%. The high quality of these results is not surprising since both of these force fields represent a narrow chemical space and it is mostly covered by the chemical space found in the side chains of the amino acids. However, the MATCH libraries based on the protein force field had more limited success in extending coverage to the nucleic acid and the lipid force fields. Only adenine and cytidine in the nucleic acid force field were successfully parameterized with MATCH: though the R^2 correlation and average unsigned error between the computed and existing charges is excellent at 0.970 and 0.035 electron unit respectively. None of the molecules in the lipid force field were successfully parameterized with MATCH because every lipid molecule has a phosphate head group and the protein force field does not contain any atom types for phosphorus.

CGENFF is the most chemically diverse force field; this is partly due to the inclusion of the model compounds from each of the other force fields. This diversity suggests that a large portion of the chemical space of the other force fields could be covered when generating molecular fragments for the atom types and the bond increment rules based on CGENFF. This hypothesis is supported by the results of typing and charging molecules from the other force fields with the MATCH libraries developed from CGENFF. MATCH was able to successfully parameterize all compounds from the other force fields and reproduced the partial atomic charges very reliably. In fact, for the charges assigned to the 7570 atoms there is an average unsigned error of 0.0013 charge units, an average percentage error of 1.02%, and an R^2 correlation with the existing charges of 0.999 This level of agreement has profound implications for further development of MATCH. As mentioned earlier, there are model compounds from each of the other force fields within CGENFF. These compounds have lead to accurate generation of the bond charge increment rules that are shared with all the other force fields, suggesting that extending the chemical space can be accomplished by adding few model compounds that represent

the desired novel chemical connectivity. For example, a huge library of novel scaffolds can be parameterized by performing quantum chemical optimizations on the simplest representations of new connectivity and then extracting the necessary bond increments to develop the rules necessary to parameterize the entire set.

The lipid force field is the second largest in terms of the number of atoms next to CGENFF and is the only other force field that has the atom types and bond increments that are necessary to type and parameterize some of the protein force field. Using the lipid force field libraries within MATCH, 50% of the protein force field could be parameterized with an average percentage error for atomic charges of 6.19% and an R^2 of 0.985. The majority of the error comes from attempting to parameterize phenylalanine without aromatic atom types. It is interesting that there is an overlap of chemical space between the head groups of lipids and amino acid backbone and side chains. For the carbohydrate force field near complete parameterization was possible, with an average percentage error of 2.98% and an R^2 of 0.994. This agreement is excellent and further illustrates the power of extrapolating the bond increment rules. Lipids contain no ring-specific atom types and yet are still able to correctly recapitulate the atomic charges of the carbohydrate force field, which are primarily sugar rings. With the lipid-force field-based MATCH libraries, all but one molecule in the ether force field could be parameterized. The error of 27% and R^2 of 0.97 indicates that the atomic charges were not computed flawlessly. Most of the error stems from the lack of an atom type that is specific for ether chemical space; the closest one that exists is for an ester oxygen. Similarly, only 40.4% of the CGENFF molecules were successfully parameterized with the lipid-based MATCH libraries, with an average percentage error of 27% and an R^2 of 0.963 for atomic charges. As with the protein force field, attempts to parameterize aromatic rings lead to error. However, the largest errors come for ribose atoms. C3' has some of the largest error with an original charge of 0.01 and a computed charge of 0.16 a percentage error of 1600%. This further reiterates the need for a quality control method, or cut off where parameter substitutions may be too unreliable.

The CHARMM nucleic acid, carbohydrate and ether force fields cover significantly less chemical space than the protein and CGENFF force fields and, thus, far fewer compounds were successfully processed with the MATCH libraries based on these force fields. For example, the nucleic acid force field-based MATCH libraries were able to type and parameterize ~36% of the molecules within the CGENFF force field and 11 molecules in the carbohydrate force field. However, the partial charge assignments for the successfully processed molecules is very high; with R^2 of 0.996 and 0.934 for the carbohydrate and CGENFF molecules respectively and with an average percentage error of 5.02% and 28.53% respectively. In addition the nucleic acid force field was also able to parameterize 6 of the ether molecules with an average percentage error of 11.7% and R^2 of 0.990. The decrease in coverage as compared to the protein and CGENFF force fields is not surprising as the nucleic acid force field does not supply as many aliphatic atom types. In fact, all of the aliphatic groups in the nucleic acid based MATCH libraries come from select patches that modulate the purine and pyrimidine groups. Similarly, the carbohydrate force field although interesting for the parameterization of 5 and 6 membered sugar rings has a very narrow chemical space. The MATCH libraries based on the carbohydrate force field could only parameterize 59 out of the 336 CGENFF

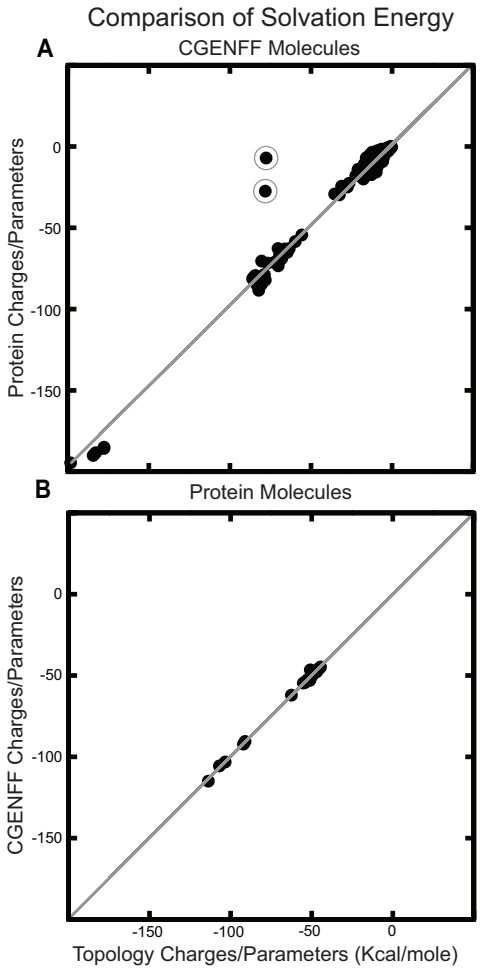

Figure 2.6. A) The correlation between the solvation energy calculated using the charges and parameters found in the CGENFF topology and parameter files compared to the solvation energy calculated using the MATCH computed protein charges and parameters. There are two distinct outliers, for which MATCH computed the incorrect formal charge. Removing these outliers yields an average error of 2.2 kcal/mole. B) The correlation between the solvation energy calculated using the charges and parameters found in the protein topology and parameter files compared to the MATCH computed CGENFF charges and parameters. Excellent agreement is achieved in this test: an average unsigned error of 0.6 kcal/mole.

molecules and all the ether molecules. Similarly, the ether force field much like the carbohydrate force field is very specific and contains simple aliphatic and ring ether molecules and the associated MATCH libraries were only able to type and parameterize 21 of the CGENFF molecules. The R^2 correlation for the atomic charges is 0.999 and the average percentage error is 0.52%, though all but one molecule had charges that were exactly reproduced.

To quantify the relationship between the percentage unsigned error in charge with physical properties of a molecule we calculated the solvation energy using the implicit solvent model: Generalized Born using Molecular Volume (GBMV)(30) of both the CGENFF and protein molecules with their original charge/parameters and their MATCH generated parameters. **Figure 2.6A** displays the correlation of solvation energy for CGENFF molecules using their topological charges and parameters compared to the ones calculated using the protein MATCH libraries. As mentioned prior there was a 15.99% percentage unsigned error when calculating the atomic charges of the molecules found in CGENFF using the protein MATCH libraries, yet there remains a very strong correlation (R^2 = 0.993) of between the solvation energy calculated using both charging/parameterization schemes. It should be noted that there were two outliners GUAN and PHEO, in both cases the protein force field lacked the necessary chemical space to correctly calculate the formal charge. These were the first instances of observing this type of malfunction and appear to be very rare. When removing these the average difference in solvation energy is 2.2 Kcal/mole. In **Figure 2.6B** the reverse case is examined which is the calculation of the solvation energy of protein molecules with their native CHARMM charge and parameters compared to the CGENFF MATCH scheme of parameters. In this case the cross validation study yielded and 6%, as expected the average difference in solvation energy is around 0.6 kcal/mole with and R^2 of 0.998, which is very promising.

2.4.2 Cross Validation of Parameters

The quality of the parameters that are generated through this cross validation study are further investigated for the case in which the MATCH libraries from the protein force field are used to parameterize the CGENFF molecules. This combination requires the

most extensive extrapolation of parameters from the MATCH libraries and, thus, should give a realistic scenario of our parameterization procedure (**Figure 2.7**). The R^2 correlation between the predicted and actual van der Waals well-depth (ε) and radius (R_{min}) parameters are 0.788 and 0.956 respectively. It is not entirely surprising that there is a minor decrease in the correlation for the well-depth parameters due to the fact that a large proportion of CGENFF molecules use the new aliphatic carbon van der Waals parameters (top_all22_prot_aliphatic_c27.str) that are not found in the older protein force field. The quality of the parameters is quite high for the equilibrium bond length and force constants with an R^2 correlation of 0.977 and 0.870 respectively and average unsigned error of 0.723% and 5.43% respectively. The quality of the angle parameters show a deterioration with an R^2 for equilibrium angle and force constant of 0.568 and 0.410 respectively and average unsigned errors of 1.87% and 20.9% respectively. The low error for equilibrium angles with the much higher error in the force constant suggest that the geometry is being reproduced but the rigidity of the angles is not reproduced with the same level of accuracy. For dihedral parameters it is more difficult to evaluate how similar two sets of dihedrals are to each other due to the possibility of multiple declarations of the same dihedral with different multiplicity values. Thus, we investigated how often there was the same number of declarations as this would be the major contributor to differences in behavior of the dihedrals. 94.6% of the dihedrals shared the

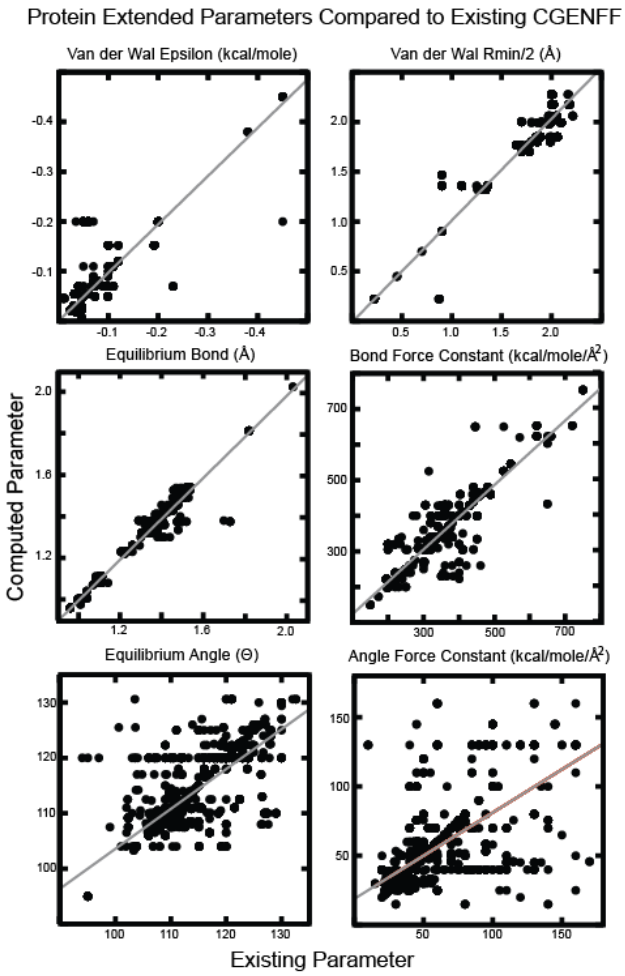

Figure 2.7. Quality of the CGENFF force field parameters that were extrapolated from the protein force field libraries in MATCH.

same number of declarations and of these 87% had the same multiplicity and 85% had the same multiplicity and identical optimum angles. These results demonstrate that a large number of dihedral angles are correctly represented as taking part in the same number of declarations with the same multiplicity impart the shape of the energy landscape with the force constants giving the height the energy barriers.

2.4.3 Parameter Substitution

To demonstrate that our atom-type substitution procedure is able to yield accurate results, we systematically removed one of the bond, angle or dihedral parameters within the CGENFF parameter file or a bond increment rule from the CGENFF-based MATCH libraries and identified the "best fit" or "nearest neighbor" parameter given the remaining parameters. **Figure 2.8** summarizes the results for this leave-one-out substitution study on bond, angle, dihedral parameters and bond charge increment rules. Substitutions are not applicable to van der Waals parameters since CGENFF includes van der Waals parameters for each atom type. For equilibrium bond lengths and angles there is good agreement between the original and the substituted parameters with an R^2 of 0.945 and 0.559, respectively, and average percentage errors of 1.62% and 2.37%, respectively. This ability to accurately preserve the geometry of a novel molecule without prior knowledge of the parameters is critical. The results for substitutions involving the bond and angle force constants display a decrease in accuracy with an R^2 of 0.508 and 0.385, respectively. However, correctly computing the force constants is of less importance than

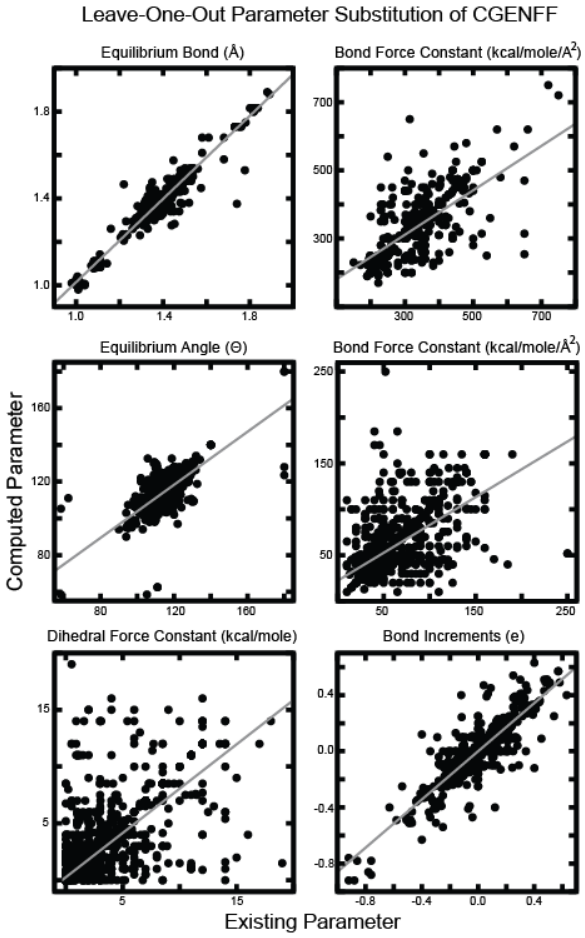

Figure 2.8. Quality of the parameter that was predicted from the "best fit" to the remaining parameters in the leave one out substitution study.

the equilibrium values as they impact the flexibility of the molecule, but not its lowest energy conformation. In many high-throughput drug design scenarios, a rough estimate of the force constants suffice in producing a reasonable parameter set for modeling the structure of a novel molecule. As mentioned in the previous section, there is a difficulty in assessing similarity in dihedral parameters. We examined whether the substituted dihedral had the same number of declarations as the original value: 88% of the time the substituted dihedral shares the same number of declarations as the original. Of these 88%, 95% had the same multiplicity value and 92% had the same multiplicity and identical optimum angles. Lastly, the bond-increment substitution is promising with an R^2 of 0.760 and average unsigned error of 0.055 electron units. Not all parameters will be generated by substitutions, rather just the minority that are not explicitly defined in a given force field. This leave-one-out study along with the results from our cross-validation studies demonstrate that the atom type substitution strategy greatly increases the chemical space that a force field can cover by extrapolation or interpolation without a significant sacrifice in accuracy.

2.4.4 PubChem Database Screen

CGENFF is the most diverse CHARMM force field and provides the most extensive coverage of chemical space. As observed in the previous section it encompasses the chemical space that is spanned by all of the other CHARMM force fields. Here, we assess the ability of the CGENFF-specific MATCH libraries to generate topology and parameter files for one million drug-like molecules within the PubChem database and estimate the upper limit of the extensibility of these MATCH libraries. The overall success rate is 84.14%, where success is defined as MATCH's ability to generate CHARMM rtf and param files and CHARMM's ability to minimize the energy of the molecule. MATCH required only ~2 seconds to process each compound and so ensures that MATCH can be incorporated into high-throughput drug design strategies. Additionally, each molecule has only to be processed once to be included in any number of molecular mechanics simulations. Furthermore, as illustrated by the results summarized in **Figure 2.9**, the quality of the final minimized structures is very high. After energy minimization, 88% of the molecules are within 0.149 Å RMSD and 99% of the molecules are within 0.249 Å RMSD of their initial pdb structure, suggesting that the

MATCH parameterization based on the CGENFF libraries yield similar lowest energy conformations. Although these are encouraging results, it is important to investigate why some molecules were not successfully processed by MATCH. First, there is an innate hierarchy for the CGENFF atom types such that a molecular fragment exists for each element containing the element with a set number of bonds and thus atom typing for each of the compounds taken from PubChem is guaranteed. However, due to the large number of types, there are many combinations that do not have known bond charge increment rules and no satisfactory substitution increment rule is identified to estimate the partial charge distributions. It is possible to vastly increase the combination of allowed bond increments by allowing a less strict substitution routine of unknown increments. However, this generally leads to overall lower quality results (results not shown). A more practical approach is to increase the chemical space of the bond increment rules by indentifying distinct model compounds or fragments that lie outside of the chemical space encompassed by the CGENFF libraries, determine the associated charge distributions from quantum chemical calculations(2) and then construct the additional bond charge increment rules. Future goals are to census the entire PubChem database for drug-like molecules and look for the chemical space that is prevalent yet is not included in the CGENFF-specific MATCH libraries. Learning the BCIs and parameters for a select number of new chemical groups will greatly expand the total chemical space covered by MATCH.

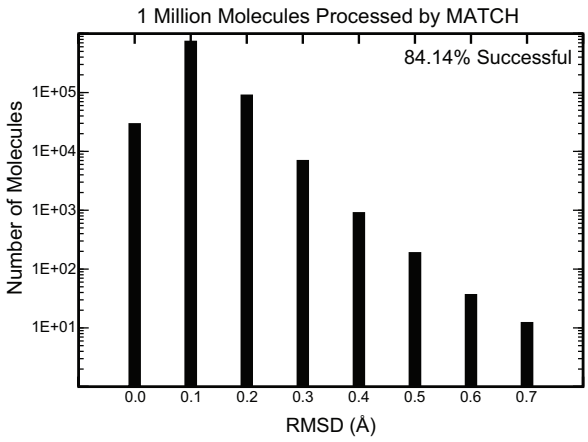

Figure 2.9. Quality of the minimized structures for the PubChem drug-like molecules that were successfully processed using the CGENFF libraries within MATCH to generate their respective topology and parameter files.

2.5 Conclusion

We have presented a library of functions and data structures, collectively called MATCH that is designed to facilitate the automated selection of appropriate atom types, partial charges, and molecular parameters for common molecular mechanics force fields. The

toolset is customizable and extensible, such that it can act both as a solution for extrapolating and interpolating from the known chemical space to novel molecules and as a tool to study the effects of specific parameter choices and parameterization strategies. Through cross validation studies we have shown that it is possible to accurately replicate atomic charges and parameters using rules derived from another force field. This strategy has significant potential; however, the ability of MATCH to successfully generate the new parameter and topology files and the quality of the results are directly dependent on the chemical diversity that exists in the original force field topology file that is used to generate the force-field specific MATCH libraries. Given the ability of the CGENFF-derived MATCH libraries to construct physically meaningful parameters and partial charge assignments for 84% of the randomly selected drug-like compounds in the PubChem Database, MATCH with its current CHARMM-based libraries is a promising tool for high-throughput drug design applications based on the biomolecular CHARMM force field.

Future work will focus on the development of an automated procedure for generating the molecular fragments of atom types and the development of a measure of the quality of both the atomic charges and parameters to understand when a substitution of a parameter or a bond increment is likely to be too detrimental to be included. In this study we actively participated in defining the molecular fragments to ensure that the simplest representation of an atom type was generated. Automated procedures were investigated, but ultimately they produced suboptimal results compared with strategies that incorporated expert knowledge. With further research, the automated fragment generation feature will enable the MATCH strategy to be even more generalizable and facilitate the seamless integration of additional force field topology files into force-field specific MATCH libraries.

This work has been published in the Journal of Computational Chemistry. The idea was conceived by Yesselman, J.D., Brooks, C.L., III and Price D.J. The analysis was performed by Yesselman, J.D. and was assisted by Knight, J.L.

2.6 References

1. Reddy AS, Pati SP, Kumar PP, Pradeep HN, Sastry GN. Virtual screening in drug discovery - A computational perspective. Curr Protein Pept Sc. 2007;8(4):329-51.

2. Vanommeslaeghe K, Hatcher E, Acharya C, Kundu S, Zhong S, Shim J, et al. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. Journal of Computational Chemistry. 2010;31(4):671-90.

3. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and testing of a general amber force field. Journal of Computational Chemistry. 2004;25(9):1157-74.

4. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. Charmm - a Program for Macromolecular Energy, Minimization, and Dynamics Calculations. Journal of Computational Chemistry. 1983;4(2):187-217.

5. Brooks BR, Brooks CL, III, Mackerell AD, Jr., Nilsson L, Petrella RJ, Roux B, et al. CHARMM: the biomolecular simulation program. Journal of Computational Chemistry. 2009;30(10):1545-614.

6. Weiner SJ, Kollman PA, Case DA, Singh UC, Ghio C, Alagona G, et al. A New Force-Field for Molecular Mechanical Simulation of Nucleic-Acids and Proteins. J Am Chem Soc. 1984;106(3):765-84.

7. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, et al. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules (vol 117, pg 5179, 1995). J Am Chem Soc. 1996;118(9):2309-.

8. Jorgensen WL, Maxwell DS, TiradoRives J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J Am Chem Soc. 1996;118(45):11225-36.

9. Klepeis JL, Lindorff-Larsen K, Dror RO, Shaw DE. Long-timescale molecular dynamics simulations of protein structure and function. Curr Opin Struc Biol. 2009;19(2):120-7.

10. MacKerell AD, Nilsson L. Molecular dynamics simulations of nucleic acid-protein complexes. Curr Opin Struc Biol. 2008;18(2):194-9.

11. Mackerell AD. Empirical force fields for biological macromolecules: Overview and issues. Journal of Computational Chemistry. 2004;25(13):1584-604.

12. Halgren TA. Merck molecular force field .5. Extension of MMFF94 using experimental data, additional computational data, and empirical rules. Journal of Computational Chemistry. 1996;17(5-6):616-41.

13. Wang JM, Wang W, Kollman PA, Case DA. Automatic atom type and bond type perception in molecular mechanical calculations. J Mol Graph Model. 2006;25(2):247-60.

14. Halgren TA, Bush BL. The Merck molecular force field (MMFF94). Extension and application. Abstr Pap Am Chem S. 1996;212:2-COMP.

15. Murphy RB, Philipp DM, Friesner RA. A mixed quantum mechanics/molecular mechanics (QM/MM) method for large-scale modeling of chemistry in protein environments. Journal of Computational Chemistry. 2000;21(16):1442-57.

16. Schuttelkopf AW, van Aalten DMF. PRODRG: a tool for high-throughput crystallography of protein-ligand complexes. Acta Crystallogr D. 2004;60:1355-63.

17. Kleywegt GJ. Crystallographic refinement of ligand complexes. Acta Crystallogr D. 2007;63:94-100.

18. Oostenbrink C, Villa A, Mark AE, Van Gunsteren WF. A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. Journal of Computational Chemistry. 2004;25(13):1656-76.

19. Downs GM, Gillet VJ, Holliday JD, Lynch MF. Review of Ring Perception Algorithms for Chemical Graphs. J Chem Inf Comp Sci. 1989;29(3):172-87.

20. Holliday JD, Downs GM, Gillet VJ, Lynch MF. Computer-Storage and Retrieval of Generic Chemical Structures in Patents .14. Fragment Generation from Generic Structures. J Chem Inf Comp Sci. 1992;32(5):453-62.

21. Welford SM, Lynch MF, Barnard JM. Computer-Storage and Retrieval of Generic Chemical Structures in Patents .5. Algorithmic Generation of Fragment Descriptors for Generic Structure Screening. J Chem Inf Comp Sci. 1984;24(2):57-66.

22. Gillet VJ, Welford SM, Lynch MF, Willett P, Barnard JM, Downs GM, et al. Computer-Storage and Retrieval of Generic Chemical Structures in Patents .7. Parallel Simulation of a Relaxation Algorithm for Chemical Substructure Search. J Chem Inf Comp Sci. 1986;26(3):118-26.

23. Holliday JD, Downs GM, Gillet VJ, Lynch MF. Computer-Storage and Retrieval of Generic Chemical Structures in Patents .15. Generation of Topological Fragment Descriptors from Nontopological Representations of Generic Structure Components. J Chem Inf Comp Sci. 1993;33(3):369-77.

24. Bolton EE, Wang Y, Thiessen PA, Bryant SH. Chapter 12 PubChem: Integrated Platform of Small Molecules and Biological Activities. In: Ralph AW, David CS, editors. Annual Reports in Computational Chemistry. Volume 4: Elsevier; 2008. p. 217-41.

25. Downs GM, Gillet VJ, Holliday JD, Lynch MF. Computer-Storage and Retrieval of Generic Chemical Structures in Patents .10. Assignment and Logical Bubble-up of Ring Screens for Structurally Explicit Generics. J Chem Inf Comp Sci. 1989;29(3):215-24.

26. Moore EF. The shortest path through a maze1959.

27. Schreiber SL. Target-oriented and diversity-oriented organic synthesis in drug discovery. Science. 2000;287(5460):1964-9.

28. Tiernan JC. Efficient Search Algorithm to Find Elementary Circuits of a Graph. Commun Acm. 1970;13(12):722-&.

29. Bone RGA, Firth MA, Sykes RA. SMILES extensions for pattern matching and molecular transformations: Applications in chemoinformatics. J Chem Inf Comp Sci. 1999;39(5):846-60.

30. Lee MS, Salsbury FR, Brooks CL, III. Novel generalized Born methods. J Chem Phys. 2002;116(24):10606-14.

CHAPTER 3

Assessing The Quality Of Absolute Hydration Free Energies Among CHARMM-Compatible Ligand Parameterization Schemes

3.1 Introduction

In molecular mechanics simulations, ligand parameterization procedures are traditionally computationally intensive and can represent a bottleneck in structure-based drug design. Thus, it is imperative that information about well-parameterized compounds be leveraged to describe novel compounds under investigation and that rapid optimization strategies be developed that are transferable across a wide variety of functional groups. Several publically-available resources exist that generate topology and parameter files for a molecule of interest so that further molecular modeling may be performed in combination with established macromolecular force fields.

Automated ligand parameterization tools assume that the bonded parameters (i.e. force constants and equilibrium bond lengths, angles and torsions) and van der Waals parameters are relatively independent of the environment and so it is straightforward to assign these parameters for a novel compound given an extensive database of parameters for known compounds. To devise partial charges for each atom in a novel molecule there are two distinct strategies used. The first strategy, employed by the ligand parameterization program Antechamber, uses a restrained electrostatic potential to generate charges for the entire molecule concurrently, often based on ab initio calculations or parameterized methods that mimic these charge distributions. In contrast, tools such as MATCH, ParamChem and SwissParam use a fragment-based approach, where charge distributions of a molecule are built-up from charges that are assigned to the component fragments of the molecule. Halgren, in developing the MMFF94 force field, first proposed bond charge increment "rules" in which optimal charges are

determined for fragments of molecules and these fragments are then pieced together to construct charge distributions for novel compounds(1).

All three fragment-based approaches mentioned above have become available recently for generating CHARMM-compatible ligand parameters and charge distributions. In our lab, the toolset of program libraries collectively titled Multipurpose Atom-Typer for CHARMM (MATCH) has been released.(2) The MATCH program itself was developed to learn atom-type definitions and bond charge increment rules from an arbitrary force field and MATCH libraries have been constructed by inferring atom-type definitions, parameters and bond charge increment rules from the CHARMM Generalized Force Field (CGENFF) topology and parameter files.(3) MATCH parameters and a topology file for a given ligand can be obtained by uploading a small molecule PDB, mol, mol2 or sdf files via a web-interface (http://brooks.chem.lsa.umich.edu/software) or, alternatively, the MATCH source code and libraries can be downloaded and further customized for local use. In their on-going work to develop CGENFF, the Mackerell lab devised ParamChem, using a strategy similar to ours, which generates topology and parameter files for novel molecules given general rules based on CGENFF. These ParamChem topology and parameter files can be obtained by uploading a small molecule mol2 file to the ParamChem web-based facility (http://www.paramchem.org). The molecular modeling group at the Swiss Institute of Bioinformatics recently released SwissParam, a web interface (http://www.swissparam.ch) that generates CHARMM or GROMACS-compatible parameter and topology files in which the van der Waals parameters are assigned from the closest atom type in CHARMM22 and the remaining parameters and partial charges are derived from the Merck Molecular Force Field (MMFF).(1, 4) While it is assumed that there may be some noise present by combining information from CHARMM22 and MMFF, this strategy takes advantage of the breadth of the chemical space covered by MMFF that is not explicitly represented in CGENFF.

Several studies have investigated the quality of automated parameterization tools by generating parameters for a diverse set of small organic molecules and computing their hydration free energies.(5) Mobley et al. used Antechamber, the AMBER facility that

generates ligand parameter and topology files using the General Amber Force Field (GAFF). Given the GAFF parameters and utilizing implicit solvent simulations, Mobley et al. computed the absolute hydration free energies for 499 small organic molecules and found that they agreed with those obtained from experiment to within ~2 kcal/mol(6). In a subsequent study, Mobley et al. found improved agreement between the calculated and experimental hydration free energies using the TIP3P water model in explicit solvent simulations for the same database of compounds, with RMS errors of 1.2 kcal/mol.(7) In our recent study, where we optimized the surface tension coefficients for scaling the surface area term in the nonpolar contribution, most implicit solvent models demonstrated reasonable agreement with experimental hydration free energies with an average unsigned errors=1.1-1.4 kcal/mol and R^2=0.66-0.81.

Shivakumar et al. recently investigated a database of 239 small molecules; all but 18 of which were contained in the database that was studied by Mobley et al.(6, 7) In their study, they evaluated the quality of hydration free energies that were computed for different force field parameters combined with implicit and explicit solvent.(8, 9) Originally, calculated hydration free energy estimates for these 239 compounds were obtained using GAFF and CHARMm-MSI ligand parameters combined with charge assignments from ChelpG, RESP or AM1-BCC protocols. Overall, ligands modeled using the GAFF charge strategy in explicit TIP3P solvent environment provided the best agreement for the calculated hydration free energies compared with experimental values; specifically, GAFF parameters yielded an R^2 of 0.87 while the CHARMm-MSI/AM1-BCC parameters resulted in an R^2 of 0.76.(8) In a more recent study, Shivakumar et al. computed hydration free energies from explicit solvent simulations using the OPLS-AA force field and charge parameterization scheme and achieved even better agreement with experiment (R^2=0.94).(9)

In this work, we compare the ability of MATCH, ParamChem, SwissParam and GAFF, to generate parameters for a diverse set of small molecules and to reproduce their respective experimental absolute hydration free energies. Given MATCH's ability to learn atom-type definitions and bond charge increment rules, we also evaluate the quality

of alternative MATCH libraries that are constructed from non-CGENFF CHARMM topology and parameter files. This analysis allows us to assess the value that is associated with enhancing the breadth and quality of the parameters that are already included in a given force field in terms of its ability to be used to extend to novel chemical contexts.

3.2 Theory

3.2.1 Overview Of Implicit Solvent Models

The specifics of each implicit solvent model are already fully documented in the original papers and, in our recent study, we have highlighted the fundamental differences among the implicit solvent models that are investigated here.(10) GBMV2 and GBSW models decompose the total hydration free energy into an electrostatic component and a nonpolar component and they employ variations of the Generalized Born model to approximate the electrostatic contribution to the solvation free energy. The GB formalism originally proposed by Still and coworkers is described by the equation(11):

$$\Delta G_{elec}^{GB} = -\frac{1}{2}\left(\frac{1}{\varepsilon_m} - \frac{1}{\varepsilon_{solv}}\right)\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{q_i q_j}{\sqrt{r_{ij}^2 + \alpha_i\alpha_j \exp\left(r_{ij}^2 / \kappa\alpha_i\alpha_j\right)}} \quad (1)$$

where r_{ij} is the distance between the charges q_i and q_j, e_m and e_{solv} are the dielectric constants assigned to the solute molecule and solvent respectively, N is the number of solute atoms, a_i is the effective Born radius for atom i and k has a value of 4 in the work of Still et al.(11) and typically is set between 2 and 10.(12) The effective Born radius of each solute atom reflects the degree of its burial within the molecule and becomes the key parameter for the calculation of the electrostatic contribution to the solvation free energy. The effective Born radius for atom i can be calculated from the atomic electrostatic self-solvation energy in the Born equation(13) (**Eq 1**):

$$\alpha_i = -\frac{1}{2}\left(\frac{1}{\varepsilon_m} - \frac{1}{\varepsilon_{solv}}\right)\frac{q_i^2}{G_{elec,i}^{GB}} \quad (2)$$

The primary advantage of GB models lies in their ability to estimate the Born radii by alternative, computationally-efficient means. Here, we focus primarily on volume-based GB models where the Coulomb Field Approximation (CFA), which approximates the electric displacement around an atom by the Coulomb field, is used to estimate the magnitude of the Born radius:

$$\frac{1}{\alpha_i} = \frac{1}{R_i} - \frac{1}{4\pi} \int_{solute} \frac{1}{r^4} dV \tag{3}$$

where R_i is the intrinsic radius of atom i (the Born radius in the absence of all other atoms) which is often set equal to the van der Waals radius and where the second term is the Coulomb field integral which is computed over the volume of the solute excluding the sphere of radius R_i around atom i. Different flavors of GB models employ alternative approaches to calculating and scaling this integral and some include higher order correction terms to account for limitations in the CFA that arise from off-center charges and non-spherical volumes of many systems.

GBMV2(14, 15) is a five-parameter analytical Generalized Born Molecular Volume model in which the molecular volume is constructed from a superposition of atomic functions. GBMV2 includes an empirical correction term, DG^1_{elec}, to the Coulomb field approximation, DG^0elec, based on a measure for the deviation from the ideal spherical shape such that:

$$\Delta G_{elec,i} = \Delta G^0_{elec,i} + \Delta G^1_{elec,i} \tag{4}$$

where the effective Born radii are estimated from:

$$\alpha_i = \frac{S}{C_0 A_4 + C_1 A_7} + D \tag{5}$$

In this formalism, A_4 is related to the Coulomb Field term in Eq. 3 and A_7 to the correction term, such that:

$$A_4 = \left(\frac{1}{R_i} - \frac{1}{4\pi} \int_{solute} \frac{1}{r^4} dV \right) \tag{6}$$

and

$$A_7 = \left(\frac{1}{4R_i^4} - \frac{1}{4\pi} \int_{solute} \frac{1}{r^7} dV \right)^{1/4} \tag{7}$$

The fundamental advantage of this analytical approach over the grid representation is that forces are readily expressed.

Generalized Born with a smooth SWitching function model, or GBSW(16), alleviates the numerical instability of solvent force calculations arising from discontinuities in the

dielectric boundary by using a simple polynomial switching function to smooth the dielectric boundary. In the original GBSW formalism, a van der Waals surface representation replaces the more expensive molecular surface representation in GBMV. In GBSW, the two parameters C_0 and C_1 in Eq. 5 (with S=1 and D=0) are obtained for various smoothing lengths, 2w, to reproduce the exact self-solvation free energies from Poisson theory using a van der Waals definition of the dielectric boundary. With the smooth switching function, the Coulomb term is described by:

$$A_4 = \left(\frac{1}{R_i} - \frac{1}{4\pi} \int_{solute} \frac{V\left(r;\{r_\alpha\}\right)}{\left|r - r_i\right|^4} dV \right) \tag{8}$$

and the correction term is described by:

$$A_7 = \left(\frac{1}{4R_i^4} - \frac{1}{4\pi} \int_{solute} \frac{V\left(r;\{r_\alpha\}\right)}{\left|r - r_i\right|^7} dV \right)^{1/4} \tag{9}$$

where $V(r,\{r_\alpha\})$ is the solute interior volume and is defined by:

$$V\left(r;\{r_\alpha\}\right) = 1 - \prod_\alpha \mathrm{H}\left(\left|r - r_\alpha\right|\right) \tag{10}$$

and where the atomic volume exclusion function, $H_i(r)$, is given by:

$$\mathrm{H}(r) = \begin{cases} 0, & r \le R_i^{PB} - w \\ \dfrac{1}{2} + \dfrac{3}{4w}\left(r - R_i^{PB}\right) - \dfrac{1}{4w^3}\left(r - R_i^{PB}\right)^3, & R_i^{PB} - w < r < R_i^{PB} + w \\ 1, & r \ge R_i^{PB} + w \end{cases} \tag{11}$$

where $\{R^{PB}\}$ are the set of atomic radii that are used to define the dielectric boundary in the PB calculations.

The GBMV2 and GBSW implicit solvent models approximate nonpolar contributions to the total hydration free energy using a solvent-accessible surface area term. In traditional MM-PBSA and MM-GBSA methods, the total molecular solvent-accessible surface area, *SASA*, is used and the nonpolar contribution is described by:

$$\Delta G_{np} = \gamma SASA + \beta \tag{12}$$

where g and b are the surface tension parameter and off-set values respectively.

In this study, we also consider the Fast Analytical Continuum Treatment of Solvation model, FACTS(17), that was recently developed by Caflisch and coworkers.(17, 18) This empirical strategy is significantly different from the above GB models in that it does not assume the Coulomb Field approximation and does not require the dielectric boundary between the solvent and solute to be defined. Instead FACTS is based on the analytical evaluation of the volume, A_i, and spatial symmetry, B_i, of the solvent that is displaced from around solute atom i. These two measures are combined in empirically parameterized equations to approximate the self-electrostatic energies:

$$\Delta G_{elec,i}^{FACTS} = a_0 + \frac{a_1}{1 + e^{-a_2(A_i + b_1 B_i + b_2 A_i B_i - a_3)}} \tag{13}$$

where a_o and a_1 are determined by using the limiting cases of a fully buried and fully exposed atom respectively. The other parameters: b_1, b_2, a_2, a_3 and R^{sphere} (which defines the solute volume considered in calculating A_i and B_i) are optimized for each van der Waals radius. The self-electrostatic energies then provide the effective Born radii via Eq. 2. Similarly, the solvent-accessible surface area is approximated by:

$$SASA_i^{FACTS} = c_0 + \frac{c_1}{1 + e^{-c_2(A_i + d_1 B_i + d_2 A_i B_i - c_3)}} \tag{14}$$

and its corresponding parameters are optimized to reproduce exact SASA values. Since the FACTS model only requires the vectors between neighboring atom centers it is significantly faster than the corresponding families of GBMV and GBSW calculations and has been documented to be only four times slower than vacuum calculations.(17)

3.3 Methods

3.3.1 Small Molecule Database

A large database of 499 small neutral organic compounds has been studied previously.(10) The original database was made available from Mobley et al.(7) which in turn was compiled from molecules from Rizzo et al.(19), Guthrie(20) and their earlier studies.(21, 22) Five duplicate compounds were identified in the original database of 504 compounds and were removed. This database contains a wide variety of chemical environments that are commonly encountered in drug design applications, including saturated and unsaturated hydrocarbons, aromatic and heterocyclic rings, halides and

polar functional groups. Checkmol(23) was used to classify the functional groups that are represented in each molecule.

3.3.2 Small Molecule Parameterization

AMBER GAFF(5)/AM1-BCC(24, 25) parameters and partial charges for all compounds in the database were obtained directly from the supplementary materials provided by Mobley et al.(7) and the AMBER *prmtop* files were converted to the corresponding CHARMM topology and parameter files using the conversion tool AMBER2CHARMM as described previously.(10) Sets of ParamChem and SwissParam parameters and partial charges were obtained by uploading the 499 mol2 files to the ParamChem (http://www.paramchem.org) and SwissParam (http://www.swissparam.ch) interactive websites, respectively. A MATCH library designated MATCH(cgenff_c36a) was constructed based on the CGENFF topology and parameter files in the c36a release of CHARMM (toppar/all36_cgenff.rtf and toppar/all36_cgenff.prm respectively). Another MATCH library designated MATCH(cgenff) was constructed from the CGENFF topology and parameter files in the c36b release which included updated parameters for several compounds. A third MATCH library designated MATCH(combined) was constructed from the union of five non-CGENFF CHARMM force field topology and parameter files, specifically, the force fields for proteins (toppar/all22_prot), nucleic acids (toppar/all27_na), carbohydrates (toppar/all35_carb), ethers (toppar/all35_ethers) and lipids (toppar/all36_lipid). To construct this MATCH(combined) library, a consistent atom type convention had to be developed in order to incorporate information from the individual CHARMM topology and parameter files. In most cases, individual force fields had the same parameter assignments for a given atom type definition. However, in the few cases in which two force fields assigned different parameters for a given atom type definition information from the more recently developed force field was incorporated into the MATCH(combined) library. Sets of MATCH parameters and partial charges for the ligands in the small molecule dataset were subsequently obtained based on these MATCH libraries.

3.3.3 Molecular Dynamics Simulations And Analysis

Simulation trajectories were generated for each MATCH(cgenff_c36a) molecule in both vacuum and the GBMV2 implicit solvent environment. No cutoffs were used; covalent bonds involving hydrogen atoms were constrained using the SHAKE(26) algorithm and the time step was 1.5 fs. The temperature was maintained near 298 K by coupling all heavy atoms to a Langevin heat bath using a frictional coefficient of 10 ps^{-1}. Simulation trajectories were 10.5 ns in length. Snapshots were saved every 5 ps throughout the last 10 ns for subsequent free energy analysis with each combination of parameterization scheme and implicit solvent model. Simulation trajectories were generated and energy evaluations associated with the GBSW and FACTS implicit solvent models were obtained using the CHARMM molecular dynamics package c36b6.(27, 28) Simulations were analyzed by the Bennett Acceptance Ratio method (BAR)(29) using a modified version of pyMBAR.(30) All simulations and calculations were performed on dual 2.66 GHz Intel Quad Core Xeon CPUs.

The GBMV2 model used a Lebedev angular integration grid with grid size of 38, geometric cross-term in the Still equation and k=8 in Eq. 1; the multiplicative factor, S, and shift, D, of a_i in Eq. 5 were 0.9085 and -0.102 respectively. For the GBSW calculations, the half smoothing length, w, was 0.3 Å; the grid spacing in the lookup table was 1.5 Å and the optimized default values for the coefficients for the Coulomb Field approximation and correction terms were used (i.e. C_o and C_1 in Eq. 5). The GBMV2 and GBSW intrinsic radii were assigned from the van der Waals radii. Default FACTS parameters were employed with infinite nonbonded cutoffs. FACTS parameters were used that had been optimized for a solute dielectric constant of 1. van der Waals radii which had not been investigated in the original FACTS study had their FACTS parameters estimated by interpolation or extrapolation from the optimized FACTS parameters using the "tavw" option in CHARMM. To be consistent with the FACTS parameterization strategy, polar hydrogens were assigned van der Waals radii of 1.0 Å.

For each implicit solvent model the nonpolar surface tension coefficient, g, was systematically varied between 0.0 and 0.07 kcal/(mol·$Å^2$). The optimal surface tension

coefficient was identified for each combination of parameterization scheme and implicit solvent model to be the value of *g* that minimized the average unsigned error for the compounds that were included in the CHARMM CGENFF topology file.

3.4 Results & Discussion

3.4.1 Coverage Of Automated Parameter Generation Schemes

Of the 499 compounds in the full dataset for which GAFF parameters and AM1-BCC charges were already available, parameters and atomic charges were successfully generated for 491 and 468 compounds by MATCH(cgenff_c36a) and ParamChem, respectively. Parameter and topology files for an additional 22 compounds were generated by ParamChem, but with error messages, so these compounds were omitted from further consideration. ParamChem successfully generated parameters for five compounds for which MATCH(cgenff_c36a) failed while MATCH(cgenff_c36a) successfully generated parameter files for eight compounds for which ParamChem failed. In total, 460 compounds were successfully processed by both parameterization schemes. SwissParam parameter and topology files were generated for all 460 compounds, except for ammonia and methane. GAFF parameter and topology files were available for all 460 compounds, but trajectory analyses failed for *N,N*-dimethyl-*p*-nitrobenzamide. For ease of comparison across the parameterization schemes, this study focuses on the 457 compounds that were successfully processed by these four parameterization schemes. This dataset encompasses 82 compounds that are explicitly included in the CHARMM CGENFF topology file and 375 compounds for which parameters and atomic charges needed to be extrapolated and interpolated from known parameters. In essence, the 82 compounds were part of the training set for developing the MATCH libraries and ParamChem rules while the 375 compounds can be considered to be a test set. During this course of this analysis, an updated version of CGENFF was released (CHARMM version c36b), so results are reported for the MATCH libraries constructed from the latest version of CGENFF (MATCH(cgenff)).

For each parameterization scheme and implicit solvent model, the optimal nonpolar surface tension parameter, *g*, was identified as the value that yielded the lowest average

unsigned error (AUE) in the absolute hydration free energies among the 82 compounds that are included in the CHARMM CGENFF topology file. Given these optimal values for *g*, the measures of model quality are summarized in **Table 3.1**.

3.4.2 Recapitulating Charge Distributions for CGENFF Compounds

The set of 82 molecules found in CGENFF were included in the training sets used by both MATCH(cgenff) and ParamChem to devise the underlying bond charge increment (BCI) rules in their respective parameterization strategies. Comparing the predicted partial charges that are based on these BCI rules with the original charges in the CGENFF topology file provides an estimate of the error that is specifically associated with the process of learning and re-applying the rules. Of the 1038 atoms in the 82 CGENFF molecules in this dataset, MATCH(cgenff) and ParamChem reproduce the CGENFF partial charge assignments within 0.005 e^- for 1022 and 997 atoms, respectively. For the remaining atoms, the partial charge differences are quite small and are less than 0.03 e^- and 0.06 e^- for MATCH(cgenff) and ParamChem, respectively. Deviations in the MATCH(cgenff) parameters are primarily due to the decision to keep the learned rules more general rather than permit highly specific definitions that while they would exactly reproduce the CGENFF charges they would likely be less transferable. In most cases, the local environment for atom type definitions was 1-2 bonds while the refinement rules for assigning bond-charge increments was 2-3 bonds from a given atom. The largest deviations between MATCH(cgenff) and the original CGENFF topology file arises from the inability of MATCH(cgenff) to reproduce partial charge distribution in CGENFF amines. For example, the H41 and H42 atoms in cytosine derivatives modeled in CGENFF have identical bond connectivity but different chemical environments due to the three-dimensional shape of the molecule. In this case, in CGENFF the H41 and H42 atoms are assigned partial charges of 0.37 e^- and 0.32 e^-, respectively, whereas MATCH(cgenff), which is based on bond connectivity alone, assigns partial charges of 0.345 e^- to both hydrogen atoms, *i.e.,* the average of 0.37 e^- and 0.32 e^-.

However, while the changes in the partial charge assignments are relatively small, they do affect the estimated hydration energies. **Figure 3.1** depicts the partial charge

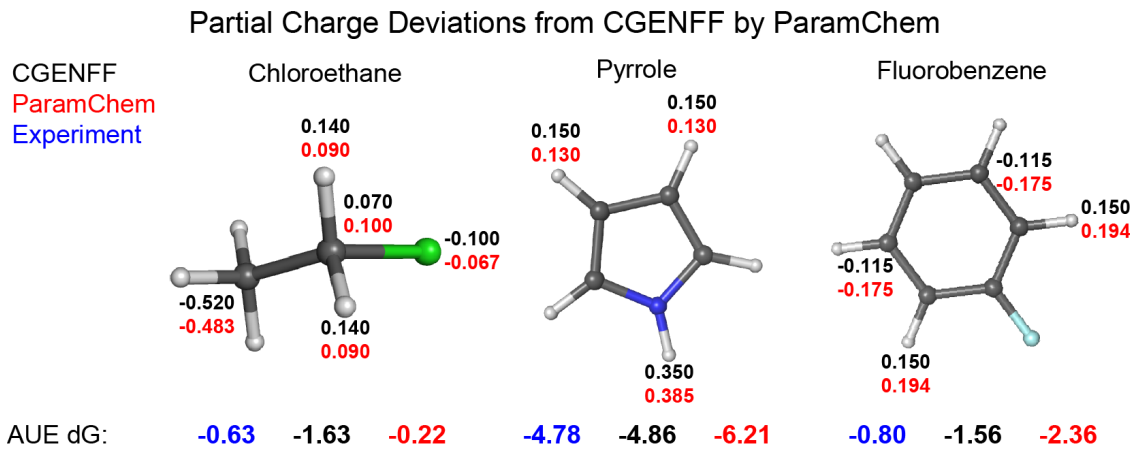

Figure 3.1. Schematic of the compounds whose partial charge distributions in ParamChem resulted in a molecular dipole difference of more than 0.01 Debye compared to the partial charge assignments in CGENFF. For clarity, only atoms whose ParamChem charges were more than 0.01 e^- from CGENFF are labeled. Note: MATCH(cgenff) charges essentially reproduce the CGENFF charges for these compounds so are not labeled.

assignments and estimated hydration free energies for the three compounds that have deviations in molecular dipoles for ParamChem relative to the CGENFF compounds that are greater than 0.1 D. Note: there were no MATCH(cgenff) compounds whose dipoles differed from CGENFF by more than 0.1 D. The ParamChem partial charge distribution for chloroethane improves the quality of the estimated hydration free energy relative to the corresponding CGENFF estimate value while the partial charge distributions for pyrrole and fluorobenzene degrade the estimate.

3.4.3 Overall Quality of Absolute Hydration Free Energy Estimates for Different Parameterization Schemes

The quality of the hydration free energies of the compounds in the small molecule dataset is summarized in **Table 3.1** and provides a direct measure of the ability of the automated parameterization schemes to characterize the chemical space of a given compound as well as the quality of the parameters in the CGENFF topology file. MATCH(cgenff) and ParamChem parameters modeled with GBMV2 and GBSW implicit solvent models demonstrate good agreement with experimental hydration free energies across the 82 CGENFF compounds with AUEs of 0.94 to 0.99 kcal/mol and R^2 values between 0.81 and 0.85. Over half of the CGENFF compounds (57-62%) have hydration free energies that are correctly predicted within 1 kcal/mol of their experimental values. Most of the

compounds (90-93%) have hydration free energies that are correctly predicted within 2 kcal/mol and almost all of the compounds (99-100%) have hydration free energies that are correctly predicted within 3 kcal/mol. Given that these compounds are the ones from which the libraries and databases of atom-typing definitions and bond-charge increment rules are derived, these results can be seen as the upper bound of the quality that can currently be expected from either MATCH(cgenff) or ParamChem automated parameterization strategies.

Parameterization scheme:	MATCH(cgenff)		ParamChem		GAFF		SwissParam	
Implicit solvent model:	GBMV2	GBSW	GBMV2	GBSW	GBMV2	GBSW	GBMV2	GBSW
CGENFF								
Opt g	0.0075	0.01	0.0075	0.01	0.0075	0.02	0.01	0.015
<|Error|>	0.97	0.94	0.99	0.96	0.88	0.95	1.12	0.99
<Error>	-0.10	-0.06	-0.11	-0.07	0.15	-0.04	0.32	0.10
R^2	0.846	0.816	0.841	0.808	0.870	0.841	0.815	0.801
% |Error|:								
<3 kcal/mol	100	99	100	99	99	98	93	96
<2 kcal/mol	90	93	90	93	90	90	77	84
<1 kcal/mol	60	62	57	60	65	61	55	61
non-CGENFF								
<|Error|>	1.47	1.43	1.51	1.49	1.24	1.33	1.49	1.16
<Error>	0.31	0.23	0.03	0.01	0.35	0.07	0.24	0.05
R^2	0.688	0.669	0.634	0.633	0.758	0.701	0.721	0.744
% |Error|:								
<3 kcal/mol	92	90	90	89	95	97	88	94
<2 kcal/mol	74	74	73	71	84	78	72	79
<1 kcal/mol	41	43	45	41	48	43	48	54

Table 3.1. Overall measures of model quality (in kcal/mol) for absolute hydration free energy predictions for trajectories analyzed using the GBMV2 and GBSW implicit solvent models and different parameterization schemes.

The overall quality of hydration free energy estimates using the MATCH and ParamChem parameterization schemes are comparable to those obtained when the small molecules are modeled with AMBER/GAFF parameters and AM1-BCC charges (AUEs of 0.88 to 0.95 kcal/mol and R^2 values of 0.84-0.87) and SwissParam (AUEs of 0.99-1.12 kcal/mol and R^2 values of 0.80-0.82). The percentage of compounds whose hydration

free energies were correctly predicted within 1 kcal/mol of the experimental values by SwissParam is comparable to the other parameterization strategies. However, the results for correct predictions within 2 and 3 kcal/mol were slightly degraded to 77-84% and 93-96% respectively.

3.4.4 Extending Parameterization Schemes to Novel Contexts

For the remaining 375 compounds that are not included in the CHARMM CGENFF topology file, the quality of the hydration free energies of these compounds is a more direct measure of the ability of MATCH or ParamChem to extend their respective atom-typing and parameterization schemes to novel contexts. MATCH(cgenff) and ParamChem parameters modeled with GBMV2 and GBSW implicit solvent models demonstrate reasonable agreement with experimental hydration free energies across these 375 compounds with AUEs between 1.4 and 1.5 kcal/mol and R^2 values between 0.63 and 0.69. For this dataset, slightly less than half (41-45%) of the compounds have hydration free energies that are correctly predicted within 1 kcal/mol of their experimental values. About three-quarters of the compounds (71-74%) have hydration free energies that are correctly predicted within 2 kcal/mol and about ninety percent (89-92%) have hydration free energies that are correctly predicted within 3 kcal/mol.

Interestingly, just as the quality of the MATCH(cgenff) and ParamChem estimates of the hydration free energies of the compounds in the CGENFF training set was higher by ~0.5 kcal/mol compared with estimates for the non-CGENFF test set compounds, the quality of the estimates based on the GAFF/AM1-BCC parameterization scheme was ~0.4 kcal/mol higher for the CGENFF compounds than the non-CGENFF compounds. The AUEs for the 375 test compounds modeled by GAFF/AM1-BCC are 1.2-1.3 kcal/mol while the R^2 values are 0.70-0.76. Since AM1 charges are assigned *de novo* for each molecule and BCC corrections were parameterized with an extensive training set of 2775 compounds that spanned the functional space represented in the CGENFF and non-CGENFF sets, the CGENFF training set/test set designations should not be applicable for AM1-BCC parameterization scheme. Thus, the poorer estimates of the hydration free energies for the test compounds over the training set compounds suggest that the compounds in the test set are inherently more challenging to model than those in

CGENFF. The SwissParam parameters also yielded a slight degradation (~0.2-0.4 kcal/mol) in the quality of the hydration free estimates for the test set relative to the training set of compounds. The AUEs for the 375 compounds modeled with SwissParam are 1.2-1.5 kcal/mol and had the largest R^2 values of any parameterization scheme of 0.72-0.74.

3.4.5 Targeting Chemical Classes for Further Parameter Optimization

Across the 82 CGENFF compounds, a subset of the full CGENFF training set, as well as the test set of 375 compounds, the AUEs for the majority of the chemical classes are less than 1.5 kcal/mol. **Figure 3.2** summarizes the AUEs of the compounds within each chemical class designation for each of the parameterization schemes with the GBMV2 implicit solvent model. **Figure 3.3** focuses on specific chemical classes that may be targeted in for further parameterization efforts. These parameterization efforts can be viewed as increasing the breadth of compounds that are reliably covered by these automated rules or increasing the depth of the meaningful coverage of a particular region of chemical space.

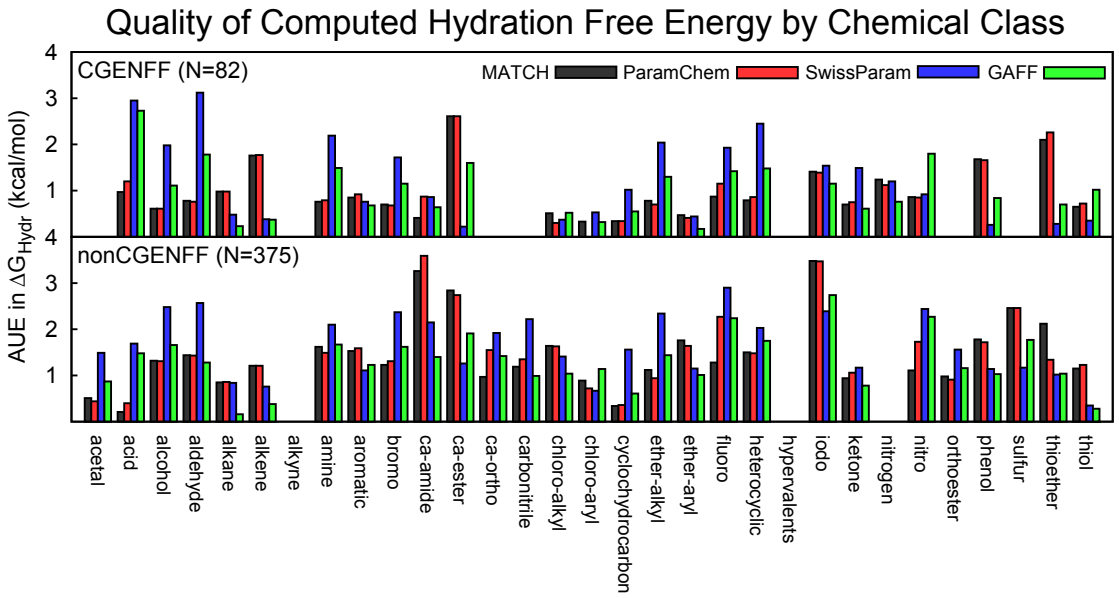

Figure 3.2. Average unsigned errors of hydration free energies by chemical class for four different parameterization schemes in the GBMV2 implicit solvent model for the A) 82 molecules that are in CGENFF and B) the 375 compounds that are not included in CGENFF.

First, **Figure 3.3A** highlights the AUEs for the four chemical classes of compounds that have errors in their respective hydration free energy estimates that are more than 1

kcal/mol larger for the non-CGENFF compounds relative to the CGENFF compounds: iodo-, carboxylic acid amides (ca_amide), chloro-alkyl, ether-aryl compounds. The low AUEs in the context of CGENFF compounds and high AUEs in the context of non-CGENFF compounds for MATCH(cgenff) and ParamChem suggest that the learned rules in MATCH(cgenff) and ParamChem for these contexts are not sufficiently transferable to accurately model the chemical space associated with these groups. For example, the rules for iodine-containing compounds are severely limited in MATCH(cgenff) because the CGENFF topology file only contains iodobenzene. Thus, it is not surprising that the AUE for the iodo- compounds is so large when there are exclusively aliphatic iodo-compounds in the non-CGENFF test set. While there is extensive coverage of the carboxylic acid amide chemical class in CGENFF topology file with examples of primary, secondary and tertiary amides, the three compounds in the ca_amide group that perform particularly poorly are ones in which the amide is a substituent on a ring and there are no examples of this type in the CGENFF dataset.

The chloro_alkyl group in the CGENFF dataset is limited to three compounds: 111_trichloroethane, chloroethane and 11_dichloroethane. This coverage is insufficient to characterize the bond charge increments of the wide variety of aliphatic halide compounds in the nonCGENFF dataset. Unlike compounds in the iodo- and ca_amide groups in which both MATCH(cgenff) and ParamChem yield similar errors in their respective hydration free energies, several compounds in the chloro_alkyl group are modeled differently by MATCH(cgenff) compared with ParamChem. For example, MATCH(cgenff) yields poor hydration free energies for molecules that contain chloro groups on opposite sides of the molecule (e.g., 1,4-dichlorobutane, bis-2-chloroethylether and 1,1,2,2,-tetrachloroethane) whereas the high AUE for the chloro_alkyl group for ParamChem results from molecules that have three fluorine atoms bound to the same aliphatic carbon (e.g., isoflurane, halothane, 1_chloro_222_trifluoroethane). Thus, the degradation in hydration free energies for these latter molecules likely results from a less than ideal bond charge increment for fluoride rather than there inherently being a problem with modeling compounds containing chlorine. Finally, the degradation observed in the ether-aryl class of compounds is dominated by the error in modeling *N,N*-

dimethyl-*p*-methoxybenzamide which suggests that the issue lies in the poor parameterization of the amide (as observed with the ca_amide group) rather than the ether functionality itself. Thus, while no specific parameterization efforts are required to improve the quality of the ether-aryl group, the larger errors for compounds in the iodo-, ca_amide and chloro_alkyl groups clearly suggest that subsequent generations of MATCH libraries and ParamChem rules would benefit from a broader template of well-parameterized training compounds for these chemical classes.

Second, examining the classes of compounds for which the AUEs differ significantly between force fields can be informative for identifying possible strategies for improving the parameterization of a particular functional group. **Figure 3.3B** highlights the AUEs in the chemical classes whose errors deviate by more than 1 kcal/mol between the MATCH(cgenff) and the ParamChem parameterization schemes. The AUE for the fluoro compounds in the CGENFF set are similar for MATCH(cgenff) and ParamChem. However, as depicted in **Figure 3.1**, the partial charge distribution for fluorobenzene modeled by ParamChem is significantly different from that modeled by MATCH(cgenff) and CGENFF itself. The underlying differences in the bond charge increment rules leads to larger differences when modeling the non-CGENFF compounds and the average error for ParamChem is about 1 kcal/mol larger than for MATCH(cgenff). In fact, ParamChem has the most difficulty producing accurate hydration free energies for molecules with multiple fluorine atoms bound to the same aliphatic carbon. Thus, it is likely that additional refinement rules within the ParamChem parameterization scheme could ameliorate the hydration free energies for this class.

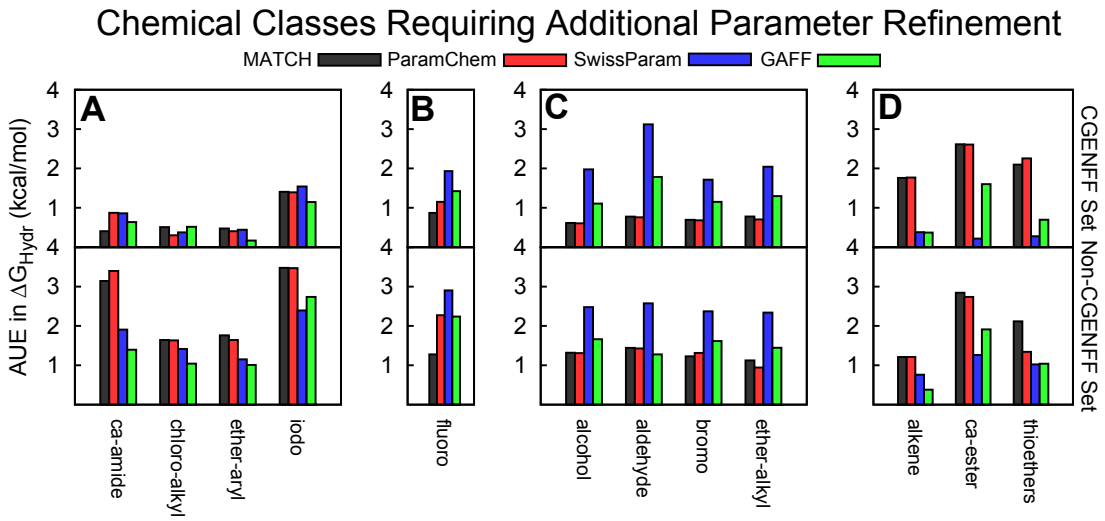

Figure 3.3. Average unsigned errors of hydration free energies for specific chemical classes for (top panel) CGENFF molecules and (bottom panel) non-CGENFF compounds. Classes in which A) both MATCH(cgenff) and ParamChem have AUEs for the non-CGENFF set more than 1 kcal/mol worse than the CGENFF set; B) MATCH(cgenff) performs 1 kcal/mol better or worse than ParamChem; C) SwissParam performs more than 1 kcal/mol poorer than the other force fields; and D) both MATCH(cgenff) and ParamChem perform more than 1 kcal/mol poorer than either SwissParam or GAFF/AM1-BCC.

Next, given the systematically poorer results for the alcohols, aldehydes, bromo- and ether alkyl groups modeled by SwissParam compared to the other force fields (see **Figure 3.3C**), further parameterization of these specific groups by the SwissParam developers would likely further strengthen SwissParam's performance. In fact, the aldehyde compounds yield the highest error for any group modeled by SwissParam with AUEs of 3.12 kcal/mol for CGENFF molecules and 2.56 kcal/mol for non-CGENFF molecules. In general, in these compounds, the partial charges assigned by SwissParam to the functional groups are systematically larger in magnitude than the corresponding charges modeled by MATCH(cgenff), ParamChem and GAFF. Schematics for compounds with the largest differences and the partial charge assignments across the force fields are presented in the Supplementary Materials.

Finally, the ca_ester, alkene and thioether classes of compounds are the only three classes that demonstrate a systematic degradation in the AUE for MATCH/ParamChem models compared to SwissParam and GAFF/AM1-BCC for the CGENFF molecules (see **Figure 3.3D**). These groups have AUEs of 2.6, 2.2 and 1.8 kcal/mol respectively in MATCH(cgenff) and ParamChem. The decrease in the quality of the esters (ca_ester) and

alkenes for both MATCH(cgenff) and ParamChem is correlated with a systematic increase in the magnitude of the CGENFF partial charges of the respective functional groups compared with those assigned by SwissParam and GAFF parameterization schemes. For example, MATCH(cgenff) and ParamChem assign an average of 0.90 e^- to the carboxyl carbon of the esters which is ~50% larger than the corresponding partial charges assigned by SwissParam and GAFF. Similarly, the hydrogen atoms at the end of conjugated alkenes have partial charges of 0.21 e^- in MATCH(cgenff) and ParamChem compared with 0.10-0.15 e^- in SwissParam and GAFF. The thioether class in CGENFF only has one member: methylethylsulfide. In this case, the CGENFF assigned partial charge of the sulfur atom is -0.1 e^- while SwissParam and GAFF assign partial charges of -0.46 and -0.30 e^-, respectively, which contributes to the increase in the molecular dipole from 0.24 to 0.43 D. Thus, the CGENFF parameters for these three chemical classes could be targeted for further improvement to more reliably reproduce experimental hydration free energies. Of course, given the differences in the parameterization philosophies across these force fields, simply adopting the GAFF or SwissParam partial charges for these compounds in order to reproduce hydration free energies estimated with an implicit solvent model cannot guarantee that these parameters will transfer appropriately to simulations in more realistic biomolecular contexts.

3.4.6 FACTS Implicit Solvent Model

FACTS is a recently developed implicit solvent model in which the Born radii are parameterized so that the electrostatic component of the hydration free energy is estimated from pairwise interactions alone. Specifically, in the FACTS parameterization, the DG_{elec} is estimated from the density of neighboring atoms and their symmetrical arrangement around the atom in question. This parameterization scheme greatly increases the computational efficiency of the calculations; in fact, the original study reported that the computational expense was only four times that of the corresponding vacuum calculations. However, this strategy requires a higher degree of parameterization than other Generalized Born implicit solvent models.

Table 3.2 summarizes the measures of model quality for the four parameterization schemes when the solvent environment is represented by FACTS. In this study, it is clear

that regardless of the ligand parameterization scheme, the FACTS implicit solvent model exhibits a slight, but systematic, degradation in the quality of the hydration free energies relative to either GBMV2 or GBSW implicit solvent models. The AUEs tend to be about 0.2-0.3 kcal/mol higher for the FACTS models than either GBMV2 or GBSW models while the R^2 values tend to be lower by 0.1 to 0.15. Thus, these results suggest that modeling with FACTS, especially in contexts where computational resources are limited, is a viable alternative to the more costly, though more accurate, implicit solvent models. Furthermore, many atom types in this work rely on interpolations and extrapolations from the values for FACTS parameterized radii; thus, the quality of the FACTS model will also likely improve as more van der Waals radii are specifically parameterized and made available to the community. These results also suggest that the FACTS implicit solvent model is transferable across these CHARMM-compatible force fields.

Parameterization scheme:	MATCH(cgenff)	ParamChem	GAFF	SwissParam
CGENFF				
Opt g	0.0025	0.0025	0.0025	0.005
<|Error|>	1.22	1.25	1.20	1.20
<Error>	0.04	0.01	0.30	0.17
R^2	0.680	0.672	0.757	0.694
% |Error|:				
<3 kcal/mol	96	96	94	94
<2 kcal/mol	85	85	82	77
<1 kcal/mol	49	49	55	62
non-CGENFF				
<|Error|>	1.59	1.74	1.42	1.50
<Error>	0.29	0.08	0.52	0.00
R^2	0.566	0.482	0.628	0.508
% |Error|:				
<3 kcal/mol	87	86	90	86
<2 kcal/mol	70	67	74	72
<1 kcal/mol	42	38	51	57

Table 3.2. Overall measures of model quality (in kcal/mol) for absolute hydration free energy predictions for trajectories analyzed using the FACTS implicit solvent model for different parameterization schemes.

3.4.7 Combining Force Fields in CHARMM

The development of CGENFF in CHARMM attempts to create general atom types and parameters for model compounds and fragments that may be important in biomolecular

simulations. This philosophy stands in contrast to that for previous CHARMM force fields where atom types and parameters were optimized for very specific chemical space within the biomolecules that were being simulated, *e.g.,* proteins, nucleic acids, lipids. Using the automated approach of MATCH, we explored the ability of the union of the non-CGENFF "context-specific" CHARMM force fields to extrapolate their parameters to model the chemical diversity in the small molecule dataset.

Parameterization scheme:	MATCH(cgenff)			MATCH(combined)		
Implicit solvent model:	GBMV2	GBSW	FACTS	GBMV2	GBSW	FACTS
CGENFF						
Opt γ	0.0075	0.0075	0.0025	0.005	0.005	0.000
<|Error|>	1.00	0.94	1.24	1.00	1.00	1.24
<Error>	-0.06	-0.21	0.14	-0.25	-0.06	-0.11
R^2	0.835	0.805	0.660	0.820	0.752	0.645
% |Error|:						
<3 kcal/mol	100	99	96	95	97	90
<2 kcal/mol	89	89	85	90	88	85
<1 kcal/mol	56	62	49	59	60	47
non-CGENFF						
<|Error|>	1.52	1.46	1.71	1.34	1.29	1.57
<Error>	0.17	-0.19	0.36	-0.39	-0.15	-0.16
R^2	0.665	0.671	0.547	0.715	0.730	0.593
% |Error|:						
<3 kcal/mol	92	90	86	91	95	91
<2 kcal/mol	73	74	67	81	82	72
<1 kcal/mol	40	35	38	49	43	37

Table 3.3: Overall measures of model quality (in kcal/mol) for absolute hydration free energy predictions for trajectories analyzed using the GBMV2, GBSW and FACTS implicit solvent models for the MATCH(cgenff) and MATCH(combined) libraries for the 73 CGENFF, and 277 non-CGENFF compounds for which MATCH(combined) libraries successfully generated topology and parameter files.

From the resulting MATCH(combined) libraries, topology and parameter files were successfully generated for 73 of the CGENFF compounds and 277 of the non-CGENFF compounds in the dataset. It was not clear, though, how meaningful subsequent hydration free energy calculations would be for this parameterization scheme. Since each of these CHARMM force fields was optimized individually, there was the potential that

combining them might produce non-physical results, particularly for compounds that would encompass chemical space that overlapped with two or more CHARMM force fields, *i.e.,* where rules were learned from different force fields. **Table 3.3** summarizes the measures of model quality obtained for these compounds in each of the implicit solvent models and shows that, for the compounds that it could parameterize from the MATCH(combined) libraries, that the compounds are modeled at a comparable level of quality to that observed from the MATCH(cgenff) library. For the set of 73 CGENFF molecules the combined force field achieved virtually the same quality as MATCH(cgenff) and, interestingly, for the more challenging test set of 277 compounds, the MATCH(combined) parameters exhibited a slight but systematic improvement over the MATCH(cgenff) parameters with AUEs of 1.3-1.7 kcal/mol and R^2 values of 0.55 to 0.73. Thus, even though the non-CGENFF CHARMM force field parameters are optimized for specific chemical environments, the high quality of results is likely a product of the consistency of the overall philosophy governing the developing of CHARMM force fields and the coherence of the optimization procedures.

Examining the differences in hydration free energy estimates by chemical class may again be useful to determine if there is any chemical space that could be optimized in CGENFF. The comparison between AUEs by chemical class for MATCH(cgenff) and MATCH(combined) is summarized in **Figure 3.4**. The three chemical groups with the largest improvement in hydration free energy estimates compared to those produced using the charges from MATCH(cgenff) are the amines, aldehydes and thiols with improvements on average of 1.1, 0.8 and 0.7 kcal/mol respectively. For the amines, increases in the partial charge on the nitrogen were responsible for the cases where the combined MATCH force field significantly outperformed the MATCH(cgenff) force field. For example, in triethylamine the partial charge on the amine nitrogen atom changes from -0.63 e^- in MATCH(cgenff) to -0.84 e^- in MATCH(combined) and is compensated by increases in the partial charges assigned to the adjacent carbon atoms

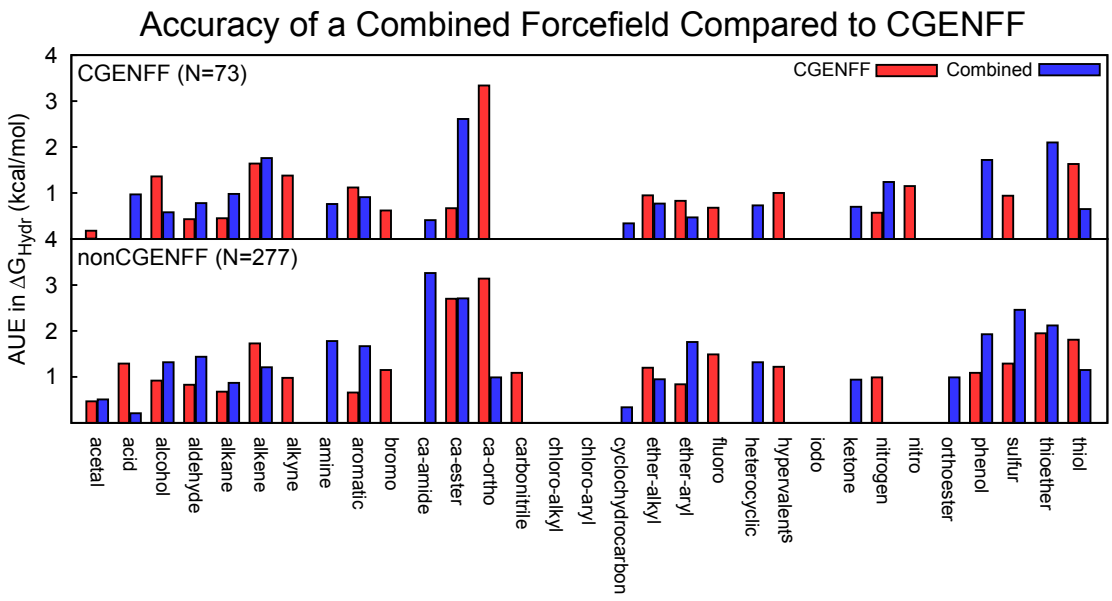

Figure 3.4. Average unsigned errors of hydration free energies by chemical class for the MATCH(cgenff) and MATCH(combined) parameterization schemes in the GBMV2 implicit solvent model for the A) 73 molecules that are in CGENFF and B) the 277 compounds that are not included in CGENFF.

from 0.03 e^- to 0.10 e^- and, thus, an increase in the N-C dipole. By contrast, the differences in the partial charge distributions of the thiol compounds results in a reduction in the S-C dipole and an improvement in the hydration free energy estimates. Similarly, the difference in the performance for the aldehyde group is dominated by three compounds, *i.e.,* E-but-2-enal, E-hex-2-enal, and E-oct-2-enal in which the C=O dipole is systematically smaller in the MATCH(combined) parameterization. Thus, these three chemical classes that could be revisited in the CGENFF force field development and/or the MATCH(cgenff) libraries could be modified to incorporate the amine, aldehyde and thiol parameters and charge assignment rules.

3.5 Conclusion

We have recently developed MATCH, an Atom-Typing Toolset for Molecular Mechanics Force Fields, in our lab. This toolset is designed to construct force field-specific libraries containing parameters and bond charge increment rules that can be learned from the topology and parameter file for a given force field. Once constructed, the MATCH library can be used to assign parameters for an arbitrary compound provided

that the chemical space represented in the compound was covered in the original force field.

We present a comparison of absolute hydration free energies that have been calculated for an extensive database of small neutral molecules using MATCH libraries constructed from CGENFF (MATCH(cgenff)) and a variety of CHARMM-compatible force fields in GBMV2, GBSW, and FACTS implicit solvent models. Of the 499 small molecules, topology and parameter files for 460 compounds were successfully generated from the ParamChem webserver and from the MATCH toolset libraries MATCH(cgenff), which were constructed from CGENFF. MATCH(cgenff) and ParamChem reproduce the partial charge distributions for most of the compounds in the dataset that were part of CGENFF.

Given optimized surface tension coefficients for scaling the surface area term in the nonpolar contribution, these automated parameterization schemes and GBMV2 and GBSW demonstrate reasonable agreement with experimental hydration free energies (average unsigned errors=0.9-1.5 kcal/mol and R^2=0.63-0.87). The FACTS parameterization yielded hydration free energies that were slightly poorer than the GBMV2 and GBSW estimates, though at a fraction of the computational expense. Antechamber parameters (GAFF with AM1-BCC partial charges) resulted in marginally more accurate estimates than the current generation of MATCH, ParamChem and SwissParam parameterization strategies.

This study highlights the importance of having sufficient coverage of chemical space within the underlying databases of these automated schemes and the benefit of targeting specific functional groups for parameterization efforts in order to maximize both the breadth and depth of the parameterized space. By analyzing the quality of hydration free energies associated with different chemical classes, it was clear that (i) MATCH(cgenff) and ParamChem would benefit from further specificity in their learned rules associated with the iodo-, amides attached to rings, and chloro-alkyl groups; (ii) ParamChem accuracy would improve with additional refinement rules for modeling fluorine-containing compounds; (iii) SwissParam could leverage parameters from other force

fields to improve how alcohols, aldehydes, bromo- and ether alkyls are modeled to better reproduce experimental hydration free energies; (iv) and parameters in CGENFF for esters, thioethers and alkenes would need to be revisited to reproduce the quality of hydration free energy estimates that are observed with GAFF/AM1-BCC and SwissParam. Finally, modeling with MATCH libraries that were derived from the non-CGENFF CHARMM topology and parameter files indicates that amine, aldehyde, and thiol parameters in MATCH(cgenff) could be improved by incorporating parameters from the context-specific force fields in CHARMM.

The overall success of these automated strategies for parameterizing arbitrary compounds indicates that a critical step forward has been taken towards making biomolecular simulations more readily accessible for a wide range of applications involving small molecules. The quality of the hydration free energies given these CHARMM-compatible force fields and implicit solvent models is promising and sets the stage for a systematic evaluation of the quality of protein-ligand binding affinities.

This work has been published in the Journal of Computational Chemistry. The idea was conceived by Yesselman, J.D., Knight, J.L, Brooks, C.L., III. Parameterization of small molecules by MATCH and ParamChem were performed by Yesselman, J.D. Parameterization of small molecules by SwissParam and AnteChamber, along with the free energy perturbation simulations. The analysis of the hydration free energy computations was done by both by Yesselman, J.D. and Knight, J.L.

3.6 References

1. Halgren TA. Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. J Comput Chem. 1998;17(5-6):520-52.

2. Yesselman JD, Price DJ, Knight JL, Brooks CL, III. MATCH: An Atom-Typing Toolset for Molecular Mechanics Force Fields. J Comput Chem. 2011;33(2):189-202.

3. Vanommeslaeghe K, Hatcher E, Acharya C, Kundu S, Zhong S, Shim J, et al. CHARMM General Force Field: A Force Field for Drug-Like Molecules

Compatible with the CHARMM All-Atom Additive Biological Force Fields. J Comput Chem. 2010;31(4):671-90.

4. Halgren TA. MMFF VII. Characterization of MMFF94, MMFF94s, and other widely available force fields for conformational energies and for intermolecular-interaction energies and geometries. J Comput Chem. 1999;20(7):730-48.

5. Wang J, Wolf R, Caldwell J, Kollman P, Case D. Development and testing of a general amber force field. J Comput Chem. 2004;25(9):1157-74.

6. Mobley DL, Dill KA, Chodera JD. Treating entropy and conformational changes in implicit solvent Simulations of small molecules. J Phys Chem B. 2008;112(3):938-46.

7. Mobley DL, Bayly CI, Cooper MD, Shirts MR, Dill KA. Small Molecule Hydration Free Energies in Explicit Solvent: An Extensive Test of Fixed-Charge Atomistic Simulations. J Chem Theory Comput. 2009;5(2):350-8.

8. Shivakumar D, Deng Y, Roux B. Computations of Absolute Solvation Free Energies of Small Molecules Using Explicit and Implicit Solvent Model. J Chem Theory Comput. 2009;5(4):919-30.

9. Shivakumar D, Williams J, Wu Y, Damm W, Shelley J, Sherman W. Prediction of Absolute Solvation Free Energies using Molecular Dynamics Free Energy Perturbation and the OPLS Force Field. J Chem Theory Comput. 2010;6(5):1509-19.

10. Knight JL, Brooks CL, III. Surveying implicit solvent models for estimating small molecule absolute hydration free energies. J Comput Chem. 2011;32(13):2909-23.

11. Still WC, Tempczyk A, Hawley RC, Hendrickson T. Semianalytical treatment of solvation for molecular mechanics and dynamics. J Am Chem Soc. 1990;112:6127-9.

12. Feig M, Brooks CL, III. Recent advances in the development and application of implicit solvent models in biomolecule simulations. Curr Opin Struct Biol. 2004;14(2):217-24.

13. Born M. Volumes and hydration warmth of ions. Z Phys. 1920;1:45-8.

14. Lee M, Feig M, Salsbury F, Brooks CL, III. New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations. J Comput Chem. 2003;24(11):1348-56.

15. Lee MS, Salsbury F, Brooks CL, III. Novel generalized Born methods. J Chem Phys. 2002;116(24):10606-14.

16. Im W, Lee M, Brooks CL, III. Generalized born model with a simple smoothing function. J Comput Chem. 2003;24(14):1691-702.

17. Haberthuer U, Caflisch A. FACTS: Fast analytical continuum treatment of solvation. J Comput Chem. 2008;29(5):701-15.

18. Haberthur U, Majeux N, Werner P, Caflisch A. Efficient evaluation of the effective dielectric function of a macromolecule in aqueous solution. J Comput Chem. 2003;24(15):1936-49.

19. Rizzo R, Aynechi T, Case D, Kuntz I. Estimation of absolute free energies of hydration using continuum methods: Accuracy of partial, charge models and optimization of nonpolar contributions. J Chem Theory Comput. 2006;2(1):128-39.

20. Guthrie JP. A Blind Challenge for Computational Solvation Free Energies: Introduction and Overview. J Phys Chem B. 2009;113(14):4501-7.

21. Mobley D, Dumont E, Chodera J, Dill K. Comparison of charge models for fixed-charge force fields: small-molecule hydration free energies in explicit solvent. J Phys Chem B. 2007;111(9):2242-54.

22. Nicholls A, Mobley DL, Guthrie JP, Chodera JD, Bayly CI, Cooper MD, et al. Predicting small-molecule solvation free energies: An informal blind test for computational chemistry. J Med Chem. 2008;51(4):769-79.

23. Haider N. Functionality Pattern Matching as an Efficient Complementary Structure/Reaction Search Tool: an Open-Source Approach. Molecules. 2010;15(8):5079-92.

24. Jakalian A, Bush B, Jack D, Bayly C. Fast, efficient generation of high-quality atomic Charges. AM1-BCC model: I. Method. J Comput Chem. 2000;21(2):132-46.

25. Jakalian A, Jack D, Bayly C. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. J Comput Chem. 2002;23(16):1623-41.

26. van Gunsteren WF, Berendsen, H.J.C. Algorithms for macromolecular dynamics and constrained dynamics. Mol Phys. 1977;34:1311-27.

27. Brooks BR, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. J Comput Chem. 1983;4:187-217.

28. Brooks BR, Brooks CL, III, Mackerell AD, Jr., Nilsson L, Petrella RJ, Roux B, et al. CHARMM: the biomolecular simulation program. J Comput Chem. 2009;30(10):1545-614.

29. Bennett CH. Efficient Estimation of Free-Energy Differences from Monte-Carlo Data. J Comput Phys. 1976;22(2):245-68.

30. Shirts MR, Chodera JD. Statistically optimal analysis of samples from multiple equilibrium states. J Chem Phys. 2008;129(12):124105.

CHAPTER 4

Using NMR Data to Generate Dynamic RNA Ensembles

4.1 Introduction

Dynamics play a critical role in functional RNAs(1-5). Unlike proteins, RNA has a high degree of flexibility, adopting radically different states upon binding of proteins or ligands(6-8). Due to RNA's structural plasticity, it is unreasonable to represent an RNA drug target as a single structure in computational drug discovery applications. Difficulty in representing extensive RNA dynamics has contributed to less optimal results in virtual screening efforts as compared to protein targets. What is required is the determination of conformational ensembles that capture highly-populated states *in vitro*, and such dynamic representations of RNA have already yielded better binders in virtual screening efforts(1)

In the past several years, techniques to create dynamic structural ensembles to use in computational docking have been developed using a combination of NMR data and molecular dynamics (MD)(9-13). Using MD to generate a pool of potential conformations, NMR observables such as spin relaxation order parameters or residual dipolar couplings (RDCs) can be back-predicted using ensemble averaging to optimize a set of conformers from the pool. Recently, Al-Hashimi and colleagues applied the sample and select (SAS) method using RDCs to produce an ensemble for the HIV-1 transactivation RNA (TAR) element. This ensemble was then combined with computational docking to find multiple novel binders to TAR that include compounds that inhibit HIV-1 replication *in vivo*(1,13). This is a great leap forward for computational RNA drug design. Understanding the success of SAS and applying it to other systems is critical to improving our success in RNA drug design.

Two major questions remain to be answered: first, can this approach be applied to other RNA systems and second, can other NMR observables be used? Although RDCs can be sensitive reporters of structure and dynamics, RDCs primarily report on global aspects of structure, and may not provide adequate information about local structure. In contrast. NMR chemical shifts report on the local electronic environment and are sensitive to stacking, hydrogen bonding, dihedral angles and other electrostatic effects. In proteins, chemical shifts are widely used in 3D structure determination and more recently in determining ensembles.

Recently, ^{1}H chemical shifts in nucleic acids have been shown to discriminate between native and non-native structures using programs such as SHIFTS(14) and NUCHEMICS(15,16). Additionally, Summers and colleagues showed that ^{1}H chemical shifts can be used to discriminate between different base-pair triplets, non-canonical base-pairs and terminal base pairs(17). Chemical shifts presents many potential benefits in ensemble determination: first, chemical shifts are significantly easier to measure then RDCs and second, chemical shifts have the ability to report of more local effects such as base orientation in internal bulges and junctions: the common locations for protein and ligand binding sites in RNA(18-20).

To investigate the utility of chemical shifts in the production of dynamic structural ensembles for the internal bulges of RNA, we ran molecular dynamics on 26 NMR structures with known chemical shifts from the Biological Magnetic Resonance Bank (BMRB)(21) to produce structure pools, and then used SAS selections employing only chemical shift data to construct ensembles. We test the CS generated ensembles using independent data, including Nuclear Overhauser Effect (NOEs) and RDCs. We then use this CS SAS approach to determine ensembles for the ribosomal decoding site, or A-site – the most well studied RNA drug target.

A-site is located in the small 30S subunit and is responsible for verifying that the mRNA codon and tRNA anticodon are perfectly complementary to ensure correct amino acid incorporation and error free translation (22,23). Upon formation of a canonical mini-helix

between mRNA and its cognate tRNA, two internal loop adenines A1492 and A1493 flip out of the A-site and form tertiary interactions with the phosphate backbone of the codon-anticodon helix, which in turn signals acceptance of the tRNA through mechanisms that are not fully understood(24,25). However, for near or non-cognate tRNAs, the non-canonical helix forms a distorted structure, and A1492 and A1493 do not flip out, thus rejecting the tRNA. Many antibiotics such as paramomycin bind to the A-site, and induce the flipping out of A1492 and A1493(26-30). This leads to loss of proofreading and incorporation of incorrect amino acids and bacterial death.

Despite being one of the most intensely studied RNAs there are still conflicting reports regarding the conformation of the adenine residues in the unbound A-site. In particular, it remains unclear whether in the unbound state, A1492 and A1493 are both flipped in, possibly hydrogen bonding to A1408, or if one or the two residues adopt a partially flipped out conformation. Clearly, a detailed understanding of the A-site binding site is critical for applying virtual screening in the development of antibiotics.

4.2 Materials and Methods

4.2.1 Simulation Protocol for BMRB structures

In total, MD simulations were performed on 26 NMR structures from the BRMB to build conformational pools to perform SAS. Each construct was subjected to 100 steps of steepest descent minimization and subsequently solvated with TIP3 and charge-neutralized using sodium counter ions. Simulations were run using CHARMM36 with the newly updated nucleic acid force field(31,32). Short dynamics runs were utilized to equilibrate the system to the final temperature of 300K. A Nosé-Hoover thermostat(33,34) was utilized to maintain a constant temperature of 300K. In addition a 2 fs timestep was employed using the shake algorithm (cite). Atom positions were saved every 2 ps. Simulation length varied from 5 – 190 ns, see **Appendix A1.1** for full details.

4.2.2 NMR Sample Preparation and Data Analysis

We employed a modified domain elongation strategy, appending the well-characterized Stem-I helix from HIV-1 TAR(35-38) to the ribosomal decoding site construct to remove coupling of internal and global dynamics, improve alignment for RDC measurement, and

increase stability of the bulge and UU base pair (**Figure 4.5**). The $^{13}C/^{15}N$ labeled A-site construct was prepared by in vitro transcription using T7 RNA polymerase as described previously(39). Briefly, transcription buffer conditions were optimized in small scale 50 μL reactions and the final conditions were scaled to a 10 mL reaction and run for 20 hours at 37 °C. Inorganic phosphate precipitates were removed by centrifugation and filtration and the supernatant concentrated to approximately 1 mL and run on a 20% denaturing PAGE gel. The desired product was cut from the band and the RNA electroeluted using the Elutrap (Whatman). Residual polyacrylamide and other trace contaminants were removed by ethanol precipitation. The RNA pellet was resuspended in ddH_2O and annealed at 95 °C for 5 minutes, followed by repeated buffer exchange into NMR buffer (25 mM NaCl, 15 mM Sodium Phosphate pH 6.4, 0.1 mM EDTA) using an Ultra-4 amicon (Millipore Corp.). The final RNA concentration was ~0.3 mM with 10% D2O (v/v).

Residue	Bond	RDC (Hz)	Residue	Bond	RDC (Hz)
G17T	*C8H8*	*28.35*	*U15*	*C5H5*	*10.70*
G17T	*C1'H1'*	*-0.56*	*U15*	*C1'H1'*	*18.39*
G18T	C8H8	27.70	*C16*	*C6H6*	*12.08*
A20T	C8H8	32.03	*C16*	*C5H5*	*-10.13*
A20T	C2H2	28.02	*C16*	*C1'H1'*	*-26.79*
G21T	C8H8	30.93	*G17*	*C8H8*	*23.37*
G21T	C1'H1'	-27.25	*G17*	*C1'H1'*	*14.71*
C04	C6H6	21.53	G87	C8H8	28.17
G05	C8H8	27.78	G87	C1'H1'	-29.39
G05	C1'H1'	-37.96	G89	C8H8	23.44
U06	*C6H6*	*16.01*	*A92*	*C8H8*	*23.53*
C07	C5H5	12.58	*A92*	*C2H2*	*27.07*
A08	*C2H2*	*21.25*	*A92*	*C1'H1'*	*5.55*
A08	*C1'H1'*	*-26.40*	*A93*	*C8H8*	*14.21*
C09	C6H6	4.63	*A93*	*C2H2*	*18.29*
C09	C5H5	-7.41	*A93*	*C1'H1'*	*-1.74*
C09	C1'H1'	-31.16	G94	C8H8	25.34
A10	C8H8	22.44	U95	C6H6	14.37
A10	C2H2	32.19	C96	C5H5	13.70
C11	C6H6	23.60	G97	C8H8	28.29
C11	*C5H5*	*-31.16*	C41T	C6H6	12.68
C11	C1'H1'	-8.03	U42T	C6H6	26.52
G12	C8H8	23.21	U42T	C5H5	24.63
C13	C1'H1'	-14.47	G43T	C8H8	21.53
U14	*C6H6*	*4.63*	C44T	C6H6	19.65
U14	*C5H5*	*19.17*	*C45T*	*C6H6*	*23.53*
U15	*C6H6*	*19.21*	*C45T*	*C5H5*	*16.97*

Table 4.1: A table of measured RDCs. Flexible residues, shown in italics, were excluded from order tensor analysis.

Chemical shifts were measured using 2D C-H or N-H HSQC experiments. For RDC measurements, nucleobase and sugar ^{1}H-^{13}C splittings were measured from the difference between the upfield and downfield components of the ^{1}H-^{13}C doublet along the ^{1}H component using the narrow transverse relaxation-optimized spectroscopy (TROSY) component in the ^{13}C dimension as implemented in the 2D ^{1}H-^{13}C S^3 CT-heteronuclear single quantum correlation (HSQC) experiments. For RDC measurements, 15 mg/ml *Pf1* phage solution (Asla Biotech) (40-43) in NMR buffer with 10% D_2O was added to RNA. RNA concentration in *Pf1* phage was 0.2 mM. Idealized A-form structures were constructed using Insight II (Molecular Simulations, Inc.) correcting the propeller twist angles from +15° to -15° using an in-house program, as previously described.

4.2.3 Preparation of Ribosomal Decoding Site Database

Utilizing the RNA Frabase 2.0 web-server(44), all structures containing the minimal secondary structure of the ribosomal decoding site were gathered. Although over 150 structures were found, a large percentage of these structures were co-crystallized with antibiotic or with mRNA and tRNA, all of which can radically alter the conformation of A1492 and A1493(24,45). For example, on addition of mRNA and tRNA, A1492 and A1493 flip out of the helix to form tertiary interactions if there is complementary pairing between the mRNA-tRNA minihelix. Likewise, antibiotics that bind to the ribosomal decoding site can also displace A1492 and A1493, forcing them into an extruded conformation. In addition, some ribosome structures also contained A1913 from the 50S ribosome, which forms interactions with the decoding site after tRNA acceptance**(46)**. In total, 10 unique conformations for the decoding site were identified and analyzed.

4.2.4 Simulation Protocol for the Ribosomal Decoding Site Construct

The 2QBF pdb was used for the starting structure for the MD simulation as it had a high resolution (3.30Å) and had a conformation of A1493 flipped out and A1492 base paired to A1408, which represents our best understanding of the decoding site ground state. The pdb was modified to include the cUUCGg tetraloop for increased stability and Stem-I of

HIV-1 TAR was added below the bulge to recapitulate the experimentally used construct in Section 4.2.3. The UUCG structure was taken from pdb 1A3M and the Stem-I of TAR was generated through the Nucleic Acid Builder server (http://structure.usc.edu/make-na/server.html). Monte Carlo simulations were performed to maximize structural alignment when attaching pdbs together. The construct was then subjected to 100 steps of steepest descent minimization and subsequently solvated with TIP3 and charge neutralized using sodium counterions. Simulations were run using CHARMM36 with the newly updated nucleic acid forcefield(31,32). As the 3D construct was a hybridization of three pdbs, 0.5 ns of equilibrium dynamics were performed with harmonic restraints to allow the structure to relax. A Nosé-Hoover(33,34) thermostat was utilized to maintain a constant temperature of 300K. In addition a 2 fs timestep was employed, using the shake algorithm. Two separate 100 ns simulations starting with different initial velocities. Atom positions were saved every 2 ps producing 100,000 total structures.

4.2.5 SAS Protocol

To initiate the SAS selection, an N-membered subset of structures was randomly selected from a total pool of M structures and an initial χ^2-value was generated using Equation 1, where L is the total number of data points used in the selection. NUCHEMICS was used to compute ^{1}H chemical shifts for each structure in the pool. The prediction alignment software PALES(48) was used to compute RDCs for each structure in the pool. χ^2 minimization was used to identify potential structures: one of the N-membered structures was randomly chosen and replaced by a random structure from the conformational pool. The 'move' from step k to $k+1$ was then accepted if $\chi^2_{(k+1)} < \chi^2_{(k)}$. If $\chi^2_{(k+1)} > \chi^2_{(k)}$, the move was accepted with a probability $P = \exp((\chi^2_{(k)} - \chi^2_{(k+1)})/T_{\text{eff}})$, where T_{eff} is an effective temperature that is linearly decreased in a simulated annealing scheme. For the RDC SAS selection, before the χ^2 is calculated a scaling factor is applied that best correlates the current N-member computed RDCs to the measured RDCs in order to alleviate PALES' shortcoming in computing the precise degree of alignment at a given concentration of *Pf1* phage.

$$\chi^2 = L^{-1}\sum_{i}^{L}(x_i^{calc} - x_i^{meas})^2 \tag{4.1}$$

4.3 Results and Discussion

4.3.1 BMRB Analysis of Bulged Residues

Analysis of SAS Selections for BMRB Structures

Recently there have been multiple attempts to use ^{1}H chemical shifts to aid in structure prediction in RNA. Al-Hashimi and colleagues demonstrated that NUCHEMICS predicted ^{1}H chemical shifts were able to discriminate between native like conformations and conformations that were deformed in high temperature windows of replica exchange simulations for four different structures(40). Additionally, Wijmenga and colleagues in a similar study were able to generate high quality 3D structures of helical RNA with chemical shifts alone(16). These studies indicate that NUCEHMICS predicted ^{1}H chemical shifts have some resolving power for both secondary and tertiary structure, however it remains unclear how well ^{1}H chemical shifts can be used to define the local conformation of bulges. Understanding the dynamics and dominant conformations of internal bulges is critical to illuminating their potential binding modes for both small molecules and proteins. In addition, it can be very difficult to determine whether a bulge in an RNA is flipped in our out based on NOE methods, which are biased to distances of closer approach, and which are complicated by potential dynamics. Thus for 26 NMR structures with BMRB deposited chemical shifts, we generated a pool of conformations using molecular dynamics and back predicted ^{1}H chemical shifts for each structure using NUCHEMICS. For each structure pool, we generated SAS selected ensembles that best reproduced the experimental ^{1}H chemical shifts of only the internal bulge residues. For detailed analysis of the distribution of bulged residues, how well chemical shifts are correlated with physical properties and the correlation of NUCHEMICS computed chemical shifts compared to experiment see **Appendix A1.2-1.5**.

Starting at an ensemble size of N=1, the ensemble size would be increased until the SAS selected ensemble had an RMSD to the experimentally determined chemical shifts equal

to or lower then 0.16 ppm or a size of N=20 was reached. This value was selected, as it is the error calculated by Wijmenga and colleagues when optimizing NUCHEMICS on their training data set(15). As the error in the measurement of experimental chemical shift is quite low, the error in prediction is the dominant source of error and increasing the size of the ensemble past the N that satisfies this limit would only be adding noise to the generated ensemble. All but five systems reached the NUCHEMICS threshold of error (**Table 4.2**). Other then 1LPW, which contains a modified pseudouridine nucleotide which was modeled as uridine in the MD simulations, all the other structures are all large bulges with multiple residues on each strand and extensive non-canonical pseudo base pairs. It is possible that the conformation that would have satisfied the chemical shifts was never sampled due to the large possible conformational space.

Of the remaining systems that were satisfied below the NUCHEMICS error threshold the majority of the systems were satisfied with an ensemble size of N=1. While this could imply that the structures are mostly static, dynamics are still present in the degeneracy of the ^{1}H chemical shifts. In most systems there were multiple conformations that were below 0.16 ppm RMSD from the experimental chemical shifts. For systems that contained S1S0 bulges, the distribution of degenerate conformations was consistent with fluctuations around a dominant single state. In contrast, 1N8X, 1OW9, 1ZC5 and 2L3E contain larger internal bulges but still fit to N=1 ensembles. Here the degeneracy in selection cluster into a few dominant states instead of a single like the S1S0 bulges.

Calculating the predicted ^{1}H chemical shifts from the NMR structure ensembles refined from NOE and RDC constraints has a substantially higher RMSD to the experimentally

Structure Infomation			SAS Selected Ensemble					
PDB	Bulge Type	CS RMSD (ppm)	N	CS RMSD (ppm)	RMSD (Å)	#NOEs	#NOE satisfied	Percent
1K8S	S1S0	0.199	1	0.033	0.440	N/A	N/A	N/A
1LC6	S1S0	0.221	1	0.045	1.468	14	11	78.6
1LMV	S1S0	0.951	1	0.142	0.813	10	10	100
1LPW	S1S0	0.500	20	0.275	1.254	11	10	90.9
1M82	S1S0	0.322	1	0.033	0.533	8	6	75.0
1MFK	S1S0	0.323	1	0.058	1.387	N/A	N/A	N/A
1N8X	S2S0	0.330	1	0.146	1.443	9	8	88.8
1NC0	S1S0	0.174	1	0.139	0.757	N/A	N/A	N/A
1OW9	S2S1	0.248	1	0.133	1.786	N/A	N/A	N/A
1R7W	S6S0	0.213	5	0.161	1.490	10	9	90.0
1R7Z	S6S0	0.280	2	0.145	4.282	N/A	N/A	N/A
1S34	S1S0	0.200	1	0.105	1.020	N/A	N/A	N/A
1XHP	S1S0	0.460	1	0.108	0.517	7	7	100
1Z2J	S3S0	0.367	3	0.122	1.635	26	24	92.3
1ZC5	S3S0	0.296	1	0.161	1.336	N/A	N/A	N/A
28SR	S3S3	0.674	20	0.353	1.676	N/A	N/A	N/A
2FDT	S1S0	0.561	1	0.103	2.690	N/A	N/A	N/A
2JWV	S4S3	0.216	20	0.207	0.996	38	34	89.3
2JXS	S1S0	0.431	1	0.076	0.396	N/A	N/A	N/A
2JYM	S1S0	0.228	1	0.134	0.817	19	16	84.2
2K41	S1S0	0.212	1	0.135	0.610	3	3	100
2L3E	S6S0	0.284	1	0.118	3.251	94	74	78.7
2L5Z	S2S0	0.352	1	0.159	2.800	28	26	92.9
2LDT	S3S3	0.300	20	0.358	1.522	19	16	84.2
2LDZ	S3S1	0.389	20	0.187	1.619	N/A	N/A	N/A
2QH3	S1S0	0.317	1	0.179	5.626	11	2	18.2

Table 4.2. A summary of the results of performing SAS selections of all 26 NMR structures with internal bulges from the BMRB.

determined chemical shifts compared to the SAS selected ensemble. One possible reason for this discrepancy is SAS selected structures differ greatly from the NOE/RDC refined structures. If these structures due differ it is critical to assess the differences between them, especially do they still satisfy the known NOE constraints? Thus it is critical to access the similarity between the structures submitted to the Protein DataBank(47) and the structures selected by SAS.

One of the most surprising findings from this survey of chemical shift SAS selections is that structures selected have high similarity to the NOE/RDC refined structures, with RMSDs ranging from 0.396 to 5.626 Å (**Table 4.2**). This result may seem counter intuitive as the NOE/RDC refined ensembles have substantially higher chemical shift RMSD to their respective experimental chemical shifts. Structures such as 1LMV and 2FDT have chemical shift RMSDs of 0.951 and 0.561 ppm respectively, for their internal bulges; upon performing a SAS selection RMSDs improve significantly to 0.142 and

0.103 ppm, respectively. Of course it is important to confirm that the SAS-selected structures resemble the experimentally determined NMR structures. With a few exceptions, SAS selection produces conformations that are surprisingly similar to the original NMR structure. The clear outlier is 2QH3, which has a single flipped out uridine. The SAS-selected conformation also is flipped out, but since there are few contacts with other residues the uridine is free to move unrestrained yielding an abnormally high RMSD. The other structures that have high spatial RMSD are 1R7Z, and 3L3E, which are both S6S0 bulges and have large degrees of freedom. These results highlight a clear issue in the SAS selection of large internal loops: if the experimentally determined structures do not agree with the experimental chemical shifts but the SAS selected ones do but are very similar to the NOE/RDC refined structures where is the difference? In each case it is not possible to give the exact reason but the general trend seems to be that slight changes in stacking and base tilt seem to be predominately the changes that are observed. This likely relates to the local effects that chemical shifts are subject to and the imperfections in the NUCHEMICS prediction method.

Lastly, for structures that have submitted NOE constraints, we examined whether the SAS selected structures satisfied the inter-residue constraints of the internal bulge (**Table 4.2**). This is an additional way to validate the ensemble of structures selected: although some of the structures match the NOE/RDC refined structure, these ensembles represent a range of dynamics and some differ greatly from the NOE/RDC refined structure. For structures where N=1 satisfied, all degenerate structures were weighted equally for the contribution to average distance for NOE. For structures with N > 1, a SAS selection was run 1000 times to generate a distribution of structures and the frequency that a specific conformation was selected was used as a weight. **Table 4.2** gives a summary of the results of the NOE analysis. Generally good agreement was observed despite not including any NOE information in the SAS selection. It should be noted that the experimentally solved structures for 1R7W, 2JWV and 2L3E did not meet all the observed NOEs with 3, 2 and 1 violations respectively. In the SAS selection 2QH3 performed quite poorly compared to the other structures with only 2/11 NOEs satisfied. 2QH3 is a curious case as only one structure was below the NUCHEMICS error

threshold and the SAS-selected structure has a single-bulged Uridine completely flipped out into solution whereas the experimentally determined structure has the uridine flipped out but still forming contacts with the minor groove. These contacts are quickly lost in the MD simulation and thus do not satisfy the NOE constraints. It seems odd that the solution NMR structure could be stable in that conformation. No RDCs corroborating this conformation have been reported.

Surveying SAS-selected ensembles for 26 NMR structures from the BMRB reveals the inherent strength of the use of chemical shifts to select realistic dynamic ensembles to understand the dynamic nature of a given system. Although the dataset is limited to only proton chemical shifts this sparse information proved sufficient to exhibit discrimination power: once carbon chemical shift prediction methods are developed, techniques such as SAS will be able to produce even higher quality ensembles. In addition, improving the ability to predict 1H chemical shifts and lowering the 0.16 ppm threshold will also enable higher quality structures to be generated. Using these structures as a benchmark, it is clear that using chemical shifts as input for SAS methods to produce dynamic ensembles is possible. In the following section, we apply this technique to the ribosomal decoding site to attempt to give insights into the inherent dynamics of this system.

4.3.2 Analysis of Experimental Ribosomal Decoding Site Conformations

Survey of Bulge Conformations of Experimentally Solved Structures

A wealth of structural information exists for the ribosomal decoding site in both isolated constructs studied by NMR and X-ray crystallography and within crystallographic structures of the entire ribosome. To understand the breath of structure diversity of conformations we performed a comprehensive survey of all structures of the A-site and characterized their conformations. **Figure 4.1** summarizes all non-redundant conformations of the ribosomal decoding site produced from this survey. Surprisingly, the isolated constructs (1A3M, 1TOE, 3LOA, 3BNL) have bulge conformations that are radically different from one other: for example, the NMR structure 1A3M has both A1492 and A1493 flipped in, with A1493 forming hydrogen bonds with A1408. The three X-ray constructs contain a very diverse set of conformations: 3BNL has A1492 base paired to A1408 and A1493 flipped in, 1TOE has A1493 base paired to A1408 and

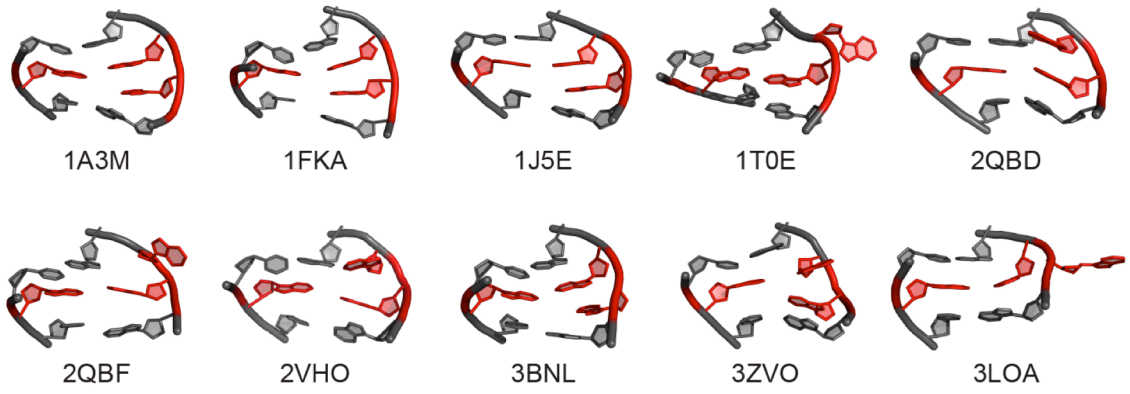

Figure 4.1: A summary of the diverse conformations of A92 and A93 that have been experimentally determined by NMR and X-ray crystallography. A93 always appears below A92.

A1492 flipped out, and lastly 3LOA has A1493 based paired to A1408 and A1492 flipped out. Likewise, conformations found in the ribosome also differ but generally follow a trend of both A1492 and A1493 flipped in with different hydrogen bond patterns with A1408. While nearly all structures have both A1492 and A1493 in an anti conformation, 3ZVO places A1493 in a syn conformation. Lastly, in both 1J5E and 1FKA, it appears that A1493 is partially interacting with C1407, which is consistent with the transition to the excited state discussed later.

NMR Construct Design and Assignment

An elongation strategy was employed similar to successful constructs for HIV-1 TAR previously studied(2-5,35). This strategy is useful to help decouple the overall tumbling of the system from the internal interhelical motions. Unfortunately yields were not high enough to perform RDCs, thus we attempted a new strategy, taking the Stem-I helix from HIV-1 TAR and using it to extend our A-site construct. HIV-1 TAR's Stem-I helix was selected as it has been extensively studied in the Al-Hashimi group by NMR, allowing for comparison to previous spectra for ease in assignment (**Figure 4.2**). In addition to elongating A-site for better alignment for RDCs it will further stabilize the bulge and UU base pair. The chemical shifts had excellent agreement to A-

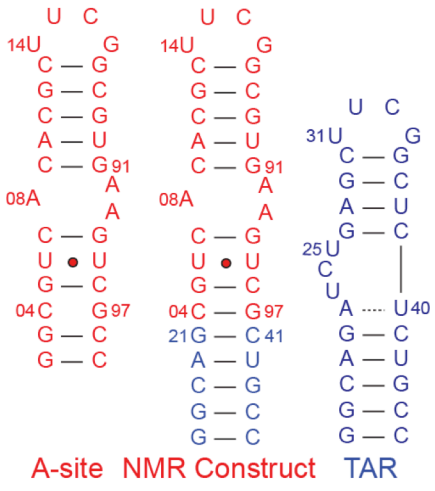

Figure 4.2: A-site construct design

site and HIV-1 TAR constructs, demonstrating that the structure had the expected secondary structure and did not perturb the structure or dynamics of the bulge region (Appendix for chemical shift overlays).

RDC Analysis

RDCs report on the orientation of a bond vector with respect to the magnetic field and report on sub-millisecond motions, providing an alternate dataset to use for dynamic ensemble construction. RDCs were measured for nucleobase C5H5, C6H6, C8H8, C2H2, and ribose C1'H1' bond vectors. Assuming the helical elements form idealized A-form helices, the measured RDCs were used to determined the best-fit order tensors for both helices using singular value decomposition, implemented by the in-house written program RAMAH. The final RNA conformer was assembled by rotating each domain into the principal axis system (PAS) of each best-fit order tensor and assembling the two helices. The order tensor has inherent 4^{n-1} degeneracy, where n is equal to the number of helical domains. The degeneracy, conserving the chirality of the molecule, is 180° about each axis (S_{xx}+180°, S_{yy}+180°, S_{ZZ}+180°), resulting in four potential orientations of the single stranded domain with respect to the reference helix. Non-rotated, Syy+180°, and Sxx+180° were ruled out due to steric clashes with Stem I, leaving only the Szz+180 (**Figure 4.3A**). The order tensor was validated by correlating the predicted

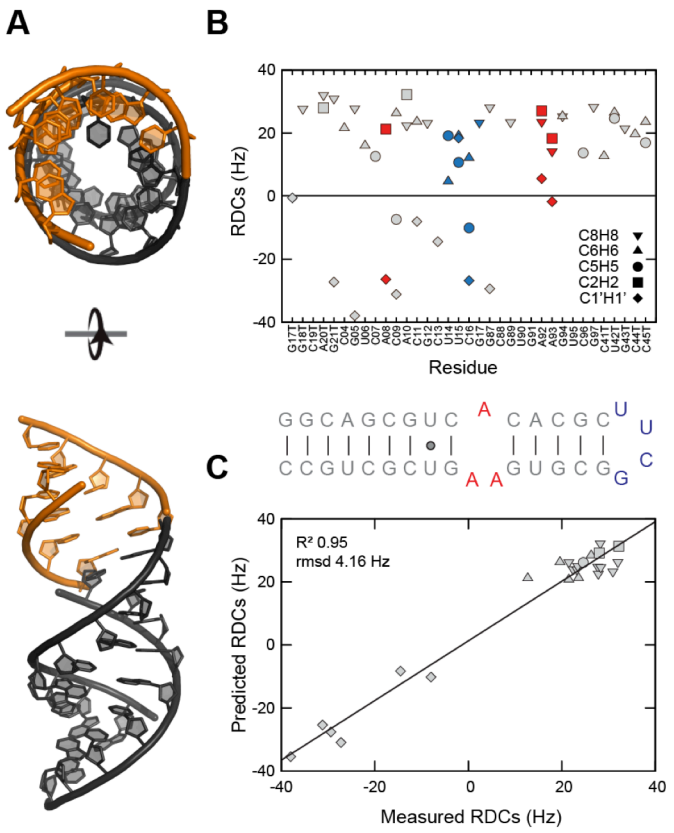

Figure 4.3: A-site RDCs and order tensor analysis A) Helical orientation determined from order tensor analysis. B) RDCs as a function of secondary structure. Colors correspond to secondary structure bellow. C) Back-calculated RDC values computed compared to measured RDCs shows excellent agreement.

RDCs with the measured, which showed excellent agreement (**Figure 4.3C**), with an R^2 of 0.95 and a RMSD of 4.16 Hz.

Due to the fact that RDCs are time-averaged over sub-ms timescales, they are sensitive reporters of dynamics as well as orientation and generally will be attenuated in a manner dependent on the motional amplitudes(49-51). The analysis of the RDCs agrees with our previous model of the ribosomal decoding site's dynamics(52). Both U1406 and U1495, which form a UU base pair have reduced C6H6 values of 16.0 Hz and 14.4 Hz respectively, indicating that the base pair is not as stable as a Watson-Crick base pair which have RDC values between 25-30 Hz. More importantly A1493 also has reduced C2H2 and C8H8 RDCs of 18.3 Hz and 14.2 Hz, suggesting that A1493 is more dynamic and does not form a stable base pair to A1408. These data are consistent with fluorescence experiments showing A1493 is less stacked than A1492. In comparison, A1492 has a C8H8 RDC value of 23.5 Hz and a C2H2 value of 27.1 Hz, which is similar to RDC values for base paired adenines. These measured RDCs present additional evidence to the model of A1492 base pairing with A1408 while A1493 transiently enters and leaves the helix. It should be noted that A1493 RDCs values are not consistent with it being highly dynamic, where one might expect near-zero RDCs. In addition both A1492 and A1493 have significantly reduced C1'H1' RDCS, with 5.6 Hz and -1.7 Hz respectively compared to base-paired C1'H1' RDCs, which are ~25 Hz. Although the significance of these reduced values is unclear, it may imply that the sugar pucker conformation does not have a stable C3'-endo conformation, typical of A-form RNA.

Domain	*N*	*CN*	RMSD (Hz)	R^2	η	ϑ x 10	ϑint	θ	ξ
Stem I	13	4.25	5.00	0.95	0.18	1.82	0.74	11.61	-6.94
Stem II	11	3.26	3.30	0.99	0.28	1.35			

Figure 4.4: Summary of order tensor parameters

The interhelical bend (θ) and twist (ξ) angles were computed using an in-house program. The interhelical bend was found to be ~11°, which is to be expected since previous structures have shown it to be coaxially stacked. The interhelical twist is also quite small

at ~7°. The relatively high ϑ_{int} and low interhelical bend angle suggest that interhelical motions are rather limited.

4.3.3 Analysis of Umbrella Sampled Conformations

In a recent study in order to study the free energy changes associated with the motions of A1492/3, an exhaustive umbrella sampling procedure was conducted using a progress variable called the center-of-mass pseudo-dihedral angle (CPD) for both A1492 and A1493. Umbrella sampling has been found to be an efficient progress variable for examining base flipping in other nucleic acid systems. A two-dimensional umbrella sampling protocol was performed to explore the conformational space of both A1492 and A1493 base flipping conformations and generate a 2D free energy landscape of A-site models with respect to these progress variables. A major feature determined from this study was that the energetic barrier for the A1493 nucleobase to flip out is greatly reduced relative to A1492, consistent with previous fluorescence data and experimentally determined structures(52). In all, four 2D free energy landscapes were generated producing approximately 322,000 structures. The exhaustive sampling within this dataset generates an ideal pool of structures to determine which structures best fit the NMR chemical shifts and RDCs.

Sample and Select A-site Structures by ^{1}H Chemical Shifts

Using NUCHEMICS, proton chemical shifts were predicted for each of the structures in the pool. Since only A1492, A1493, and A1408 conformations were significantly altered in the umbrella sampling simulations, the chemical shifts for these residues alone were used as experimental data. The procedure used to generate the pdbs initially moves both A1492 and A1493 into their desired initial conformations using a biasing potential, which occasionally breaks the G1491-C1408 or the G1494-C1407 base pair due to steric clashes. Because only chemical shifts of A1492, A1493, and A1408 were used in the SAS selection, structures in which this occurs were removed prior to the selection of structures that best matches the chemical shifts.

An ensemble size of N=1 was able to satisfy the threshold of 0.16 ppm (the expected error in NUCHEMICS prediction). All structures that met this requirement or were below

this cutoff were examined. 277 structures out of 322,000, or 0.08% of the total pool, were below this cutoff. Analysis of the distribution of structures is displayed in **Figure 4.5B** where surprisingly one can see matches the favorable area found on the free energy in **Figure 4.5A**. To further investigate how well the chemical shift selected distribution coincides with the free energy, the distribution was compared in Figure 4.8C. The majority of the conformational space examined in the free energy distribution is located in 6 – 11 kcal/mol which can be seen clearly in **Figure 4.5A** represented as yellow to red. However, the distribution generated from the conformations with a low RMSD to the experimental chemical shifts are primarily found in lower free energy areas 2 – 6 kcal/mol. This was despite the fact that the chemical shift prediction has no knowledge of what is energetically favorable to a molecular mechanics force field.

Figure 4.5D and **E** display the composition of the conformers for both A1492 and A1493 flip and chi angles respectively. Although approximately 50% of the conformations in the pool have A1493 in a syn conformation only 47 of the conformations selected by the chemical shifts are in the syn conformation or about 17%. This is interesting as all but one structure (3ZVO) has A1493 in an anti conformation, but since it understood that A1493 is able to completely flip out of the helix it is possible that a fraction of the time it rotates about its chi bond and reenters the helix in a syn conformation. Since chemical shifts report on the local interactions of the nuclei environments, which is largely influenced by stacking interactions, it may be difficult to distinguish syn vs anti conformations using only proton chemical shifts. In contrast to A1493, A1492 has one

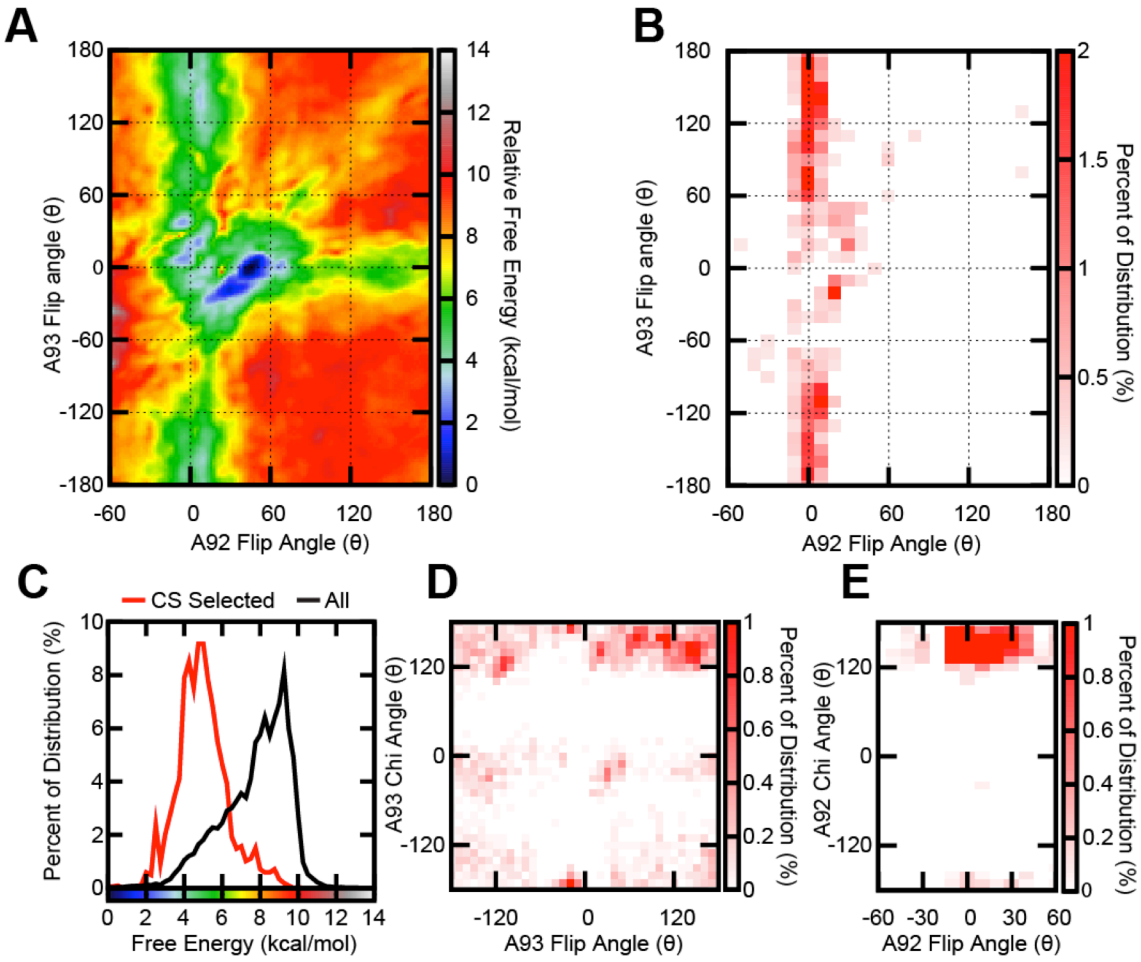

Figure 4.5. Chemical Shift SAS for A-site A) Relative free energy as a function of flip angle of A1492 and A1493. B) Distribution of the flip angles of A1492 and A1493 that were selected by chemical shift SAS. C) Distribution of free energy of conformations selected by chemical shift SAS compared to the total distribution. D) Distribution of flip angle and chi angle of A1493 of selected conformations. E) Distribution of flip angle and chi angle of A1492 of selected conformations.

relatively dominant state. Chemical shift selected conformations show A1492 in an anti and flipped in state nearly 99% of the time.

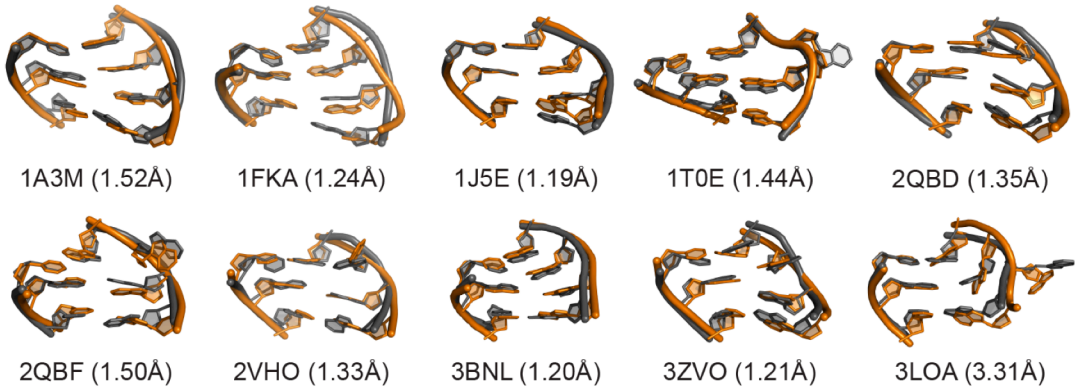

Figure 4.6. Closest structure by RMSD to each of the experimental A-site structures. Black is the experimental structures, while orange is the closest structure selected by chemical shifts. A93 always appears below A92.

Conformations of the experimentally determined decoding sites and from the chemical shift selected pool with the closest RMSD were compared (Figure 4.9). For each experimental structure excluding 3LOA, a conformation from the chemical shift selected pool matched with an RMSD equal or less then 1.5Å, corroborating that selected conformations agree with existing experimental data. Only 3LOA shows significant deviation from the selected distribution of conformations with the closest having an RMSD of 3.31Å. This finding is not entirely surprising: 3LOA diverges significantly from the model of the decoding site, with A1492 in a flipped out conformation with A1493 flipped in and base paired to A1408. It should also be noted that A1492 in 3LOA forms crystal-packing interactions with another unit cell in the crystal, so its conformation may be dependent on how it was crystallized and could explain why its conformation is not well represented in the selected conformations and the other experimental structures. To see how closely they reproduce the proton chemical shifts, the chemical shifts with NUCHEMICS were predicted from each of the experimentally determined structures (**Figure 4.6**). If each of the structures already produced the chemical shifts that are within 0.16 ppm, RMSD this analysis would not be very interesting. There appears to be no correlation between the prediction quality of the

PDB	RMSD (ppm)
1A3M	0.30
1FKA	0.43
1J5E	0.52
1TOE	0.48
2QBD	0.78
2QBF	0.63
2VHO	0.71
3BNL	0.40
3LOA	0.42
3ZVO	0.79

Table 4.3 Table of RMSD to experimental chemical shifts of known experimental structures using NUCHEMICS.

chemical shifts produced by the experimentally determined structure and how close a structure that does match the chemical shifts is to each structure. This could be because chemical shifts are governed by local conformations, so slight changes in base orientation may have dramatic consequences in chemical shifts predicted. This work highlights the importance of sampling. While initially it would seem each experimental structure is a poor match to the chemical shifts, structures that have similar orientations may still agree with the chemical shifts.

Sample and Select by RDCs

To compare directly to the chemical shift SAS, only the RDCs of A1492/3 and A1408 were initially used in the RDC SAS selection. It quickly became clear that this approach was insufficient to give statistically significant for two critical reasons. First, because PALES cannot correctly model the degree of alignment of a given structure to the Pfl alignment media, a scaling factor is introduced during the SAS selection to scale the predicted RDCs to best match the experimental RDCs. Without a frame of reference of some helical RDCs, the dynamical residues in the bulge lose their context of being dynamically attenuated and the scaling factor will improperly scale predicted RDCs, permitting the inclusion of a variety of inaccurate solutions. To account for this discrepancy, RDCs from the C07-G94 base pair as well as the UU base pair below were used in the SAS selection for reference points. Secondly, the C1'H1' RDC values of both A1492/3 are inconsistent with stable 3'-endo sugar puckers which are typically observed in A-form RNA. Because of the nature of the umbrella sampled conformations, only 2ns were run at each point in the conformational space, they were unlikely to sample any sugar pucker other then their starting conformation which was 3'-endo. Thus both C1'H1' RDCs were also removed since the sugar pucker dynamics had insufficient sampling.

Because the proper ensemble size during RDC SAS selections can vary, selecting an ensemble size that yields an RMSD comparable to the experimental RDCs is preferred. Unlike the chemical shift analysis which used an ensemble size of N=1, using RDCs an ensemble size of N=2 was sufficient to satisfy the RDCs (**Figure 4.7**). The behavior of the relationship between ensemble size and RMSD to the measured RDCs is interesting. Initially N=1 and N=2 yield result extremely low RMSDs while N=3 onward yield relatively higher RMSDs. This differs from previous work using TAR HIV-1 where increasing the ensemble size kept improving the RMSD until N=20 where increasing N no longer yielded any additional improvement. The only clear difference between the two systems is that A-site is much more rigid then TAR HIV-1 which has a S3S0 bulge compared to the S2S1 bulge of A-site. It should also be noted that for N=2 for A-site the resulting ensemble is consistent upon repeat runs with more detailed analysis later, while N=3 onward the results appear more chaotic following less of a clear trend, this is a characteristic of over fitting data and may be responsible for the aberrant results.

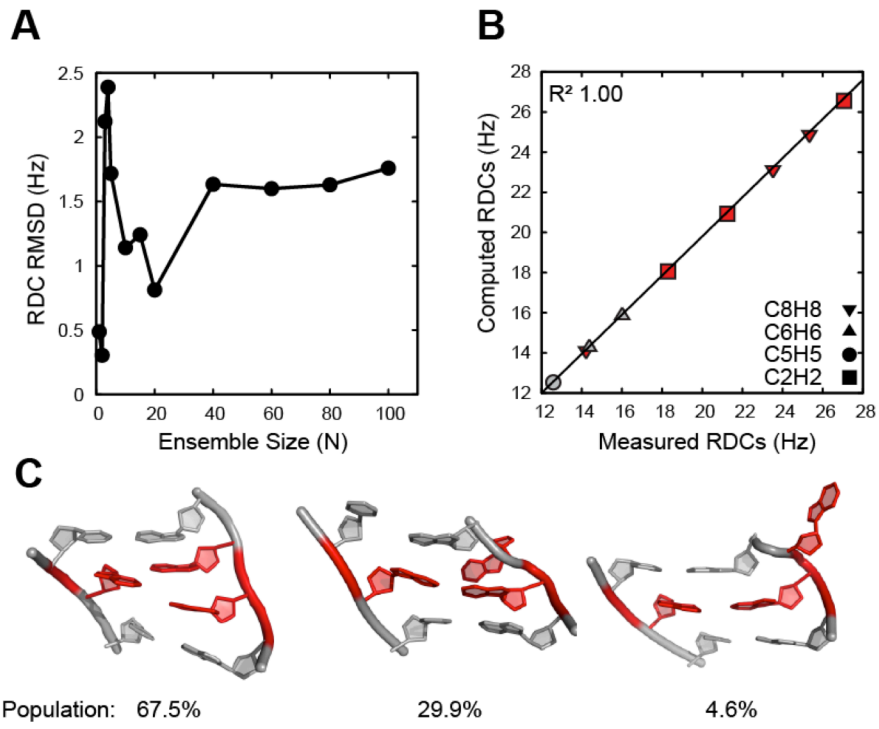

Figure 4.7: RDC SAS results for A-site. A) SAS selection RMSD to measured RDCs as a function of ensemble size. B) Correlation between computed RDCs from SAS ensemble of 2 to measured RDCs. C) Population exhibited from 1000 SAS selections with ensemble size 2.

Understanding this behavior will further our understanding of using RDCs to generate dynamic ensembles of RNA systems using SAS.

To build up a distribution of conformations that agreed with the measured RDCs, 1000 SAS selections were run at N=2 yielding a total of 2000 structures in total. Interestingly only 17 unique structures were selected in total. These structures can be characterized into three clusters, which are displayed in **Figure 4.7**. These states can describe as A1492/3 both flipped in, A1492 flipped in with A1493 partially flipped out and finally A1492 flipped in with A1493 completely flipped out. These states have a population of 67.5%, 29.9% and 4.6% respectively. These 3 states and population reiterate the model of having a dominant model of having both A1492 and A1493 in with A1493 transiently leaving and reentering the helix.

Comparison between Chemical Shift and RDC Selections

Ensembles generated from chemical shifts and RDCs SAS selections are highly overlapping. For conformation cluster three also appears in the chemical shift selection, while cluster one and two have an RMSD of 1.41 and 0.95 Å respectively with conformations from the chemical shift selection. This is another independent line of evidence of the proposed model for A-site in which A1493 samples flipped in and flipped out states. The only real discrepancy between the two selections is the population between the flipped in and out states. In the CS SAS ensemble, the flipped in and out conformations are 25% and 75% populated, respectively whereas in the RDC SAS, they are 95% and 5%, respectively. It should be noted that the RDC sensitivity to motional amplitudes is generally weak for low motional amplitudes and only becomes significant for extensive motions characterized by order parameters on the order of $S^2 < 0.6$. Thus it is possible that the RDCs are more heavily weighted by the more rigid flipped in state and that the motions between the flipped in and flipped out state lead to motional contributions comparable to measurement uncertainly. By contrast, chemical shifts are strongly sensitive to the flipped out state which both alter stacking interactions as well as the sugar pucker and glycosidic bond angle χ. It should be reiterated that as displayed in the free energy plot in Figure 4.8A the relative free energy between A1493 flipped out in

flipped in is around 2 kcal/mol between there respective minimum. This is not a large barrier and compared the possible structures in the pool this is still good agreement between the two ensembles.

4.3.4 Analysis of MD Simulation for A-site

Although the RNA ensembles constructed using umbrella sampling data were consistent with experimental data and the generalized model of A-site bulge dynamics, the construct used differed from that used experimentally. Straight MD simulations were performed to determine if using all measured RDCs and chemical shifts altered the generated ensembles, although unlikely as the modification occurs two base pairs below the UU base pair, far removed from the bulge site. This procedure is in line with previous studies, which also used straight MD simulations to produce the conformation pool, to allow direct comparison to prior results on HIV-1 TAR. Comparing the ensembles generated using either umbrella sampling or the MD simulation also provides a platform to determine the effects of sampling within a conformational pool, as the umbrella sampled pool is nearly exhaustive while the simulations yield a much smaller amount of sampled space, representing the typical range of accessible conformations.

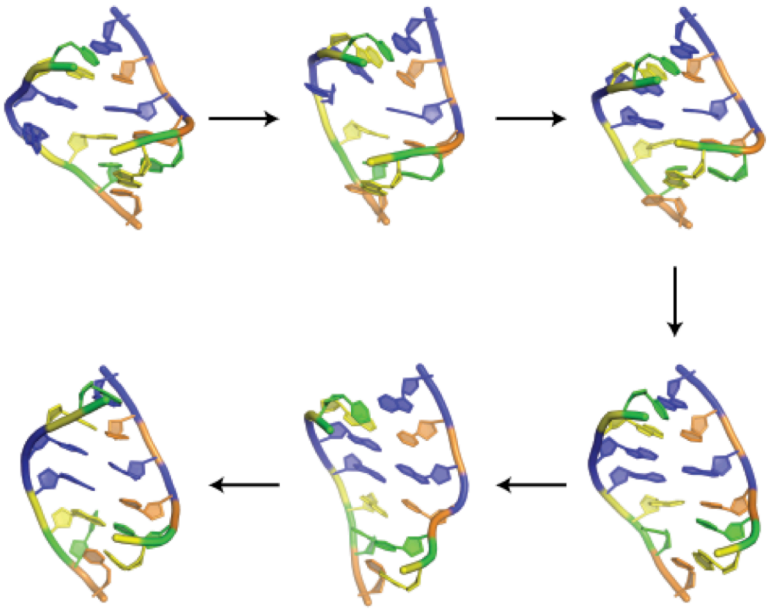

Figure 4.8: Transition from initial secondary structure into excited state secondary structure, which has U1495, flipped out with A1492 and A1493 base-paired to A1408 and C1407 respectively.

MD simulation reproduce experimentally observed A-site Excited State

Two 100 ns MD simulations were run using different initial velocities in order to remove condition bias from the starting structure. In both simulations, a transition occurred to the proposed excited state which was determined by Al-Hashimi and coworkers (53) in which U1495 is flipped out and both A1492 and A1493 form base pairs within the helix (**Figure 4.8)**. Although this transition was already observed in a previous MD simulation, the transition from ground state to excited state was not documented. Starting from a low free energy conformation where A1493 is flipped out and A1492 is base paired to A1408, A1492 exhibits stochastic movement up and down within the helix as seen in the second panel of the **Figure 4.8**. At one point A1493 was in the process of flipping in and A1492's movement made space for A1493 to insert into the helix. This arrangement is stable for around ~1 nanosecond when U1495 becomes destabilized and loses its hydrogen bonding to U1406. UU base pairs are not particularly strong (~2 kcal/mol) and may reflect a physiologically relevant event. Afterwards, G1494 points downward and allows A1493 to base pair with C1407. This step seems to be counterintuitive, as GC base pairs are more stable then AC: it is likely that the G1494-C1407 hydrogen bonds are insufficiently strong without stacking interactions from above and below. Lastly, G1494 forms a relatively stable base pair with U1495. Experimental evidence supporting this transition has been observed by NMR by Al-Hashimi and coworkers(53).

The presence of this alternative secondary structure posed serious problems for performing both the RDC and chemical shift SAS selections. As noted in the section of performing SAS selections with the umbrella sampling data, structures in which base pairing broke had to be removed since only a subset of the chemical shifts were being used, which were unable to discriminate between the two states. In that case utilizing all the chemical shifts did not have this problem but yet the sensitivity of the selection was reduced as the majority of the chemical shifts are from base paired regions. Here a similar situation existed, with the inclusion of the excited state structures into the conformation pool, these conformations were preferred in both the RDC and chemical shift selections. When using all the chemical shifts, conformations with the original secondary structure were selected, but surprisingly when all the RDCs were used the excited structure was

still selected. This may be related to the fact the RDCs report on interhelical motions and as discussed earlier A-site is largely coaxially stacked, the excited state is even more stacked then the wild type. This work demonstrates the importance of understanding the limitations of what the experimental data is capable of discriminating. For this reason all structures with the excited state secondary structure were removed.

SAS Selections with MD Pool by Chemical Shifts

Unlike the in the umbrella sampled pool, N=2 was the minimal ensemble size that satisfied an RMSD to the experimental chemical shifts less then 0.16 ppm. In the umbrella sampled pool only N=1 was required: this increase in required conformations may be due to the quality of structures produced in the MD simulation compared to the umbrella sampling. This result suggests that a major factor in the ensemble size may not be completely dependent on the system but also the pool of structures that is being drawn from and analysis of dynamics dependent solely the number of conformations required to satisfy a system may not be appropriate. The fraction of space sampled by the MD is displayed in **Figure 4.9**. Only a fraction of the space was sampled compared to the umbrella sampling. In addition, this plot displays areas sampled if at least one conformation had that given flip angles for both A1492 and A1493. In the umbrella sampling pool there are at least 100 conformations per 10° by 10° area, since this pool is a combination of 4 separate runs there is likely much more. This exhaustive numeration gives many more possible favorable states that may be rarely sampled in the course of MD simulations and there may be energetic barriers that keep a structure in a given state for extended periods.

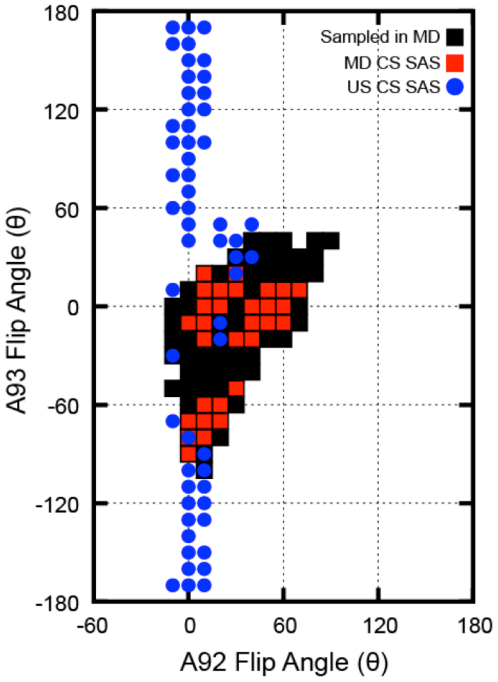

Figure 4.9. Coverage of MD simulations compared to the umbrella-sampled pool.

Given the discrepancy in conformation pool characteristics, performing chemical shift SAS using the MD pool highlights the importance of sampling. The distribution produced by performing 1000 SAS selections is shown in red in Figure 4.8. This result is very interesting since the highest populated distribution features A1493 with a flip angle of ~0° and A1492 between 0°-60°, which is the most favorable area on the free energy surface. Unfortunately, this result differs from the result produced from the umbrella sampling ensemble which has A1493 flipped out, and also satisfies the chemical shifts with a RMSD much under 0.16 ppm with only one conformation instead of the two required using the MD pool. This analysis stresses the importance of generating a pool with proper sampling.

SAS Selections with MD Pool by RDCs

Unlike the chemical shift SAS where the most favorable conformations have A1493 flipped out in the umbrella sampled distribution, RDC SAS selections built a distribution with a predominately flipped in conformations. Like the distribution built from the umbrella sampled pool, the conformations selected using the MD pool also were all flipped in. N=2 was also sufficient and N=3 onwards did not improve the RMSD to the measured RDCs. The only minor difference was the lack of any flipped out conformations in the MD selected conformers compared to the 4.6% selected in the umbrella sampled pool.

4.4 Conclusion

Here we have shown the utility of chemical shifts in the generation of dynamic ensembles for the bulges of RNAs. We first demonstrated that proton chemical shifts are capable on reporting on stacking and the motion of single stranded nucleobases. The utility of chemical shifts were demonstrated in the survey of chemical shift SAS selections of 26 NMR structures that selected out structures that were similar to the NOE/RDC refined structures. With larger data sets and the ability to predict carbon and nitrogen shifts, the future of RNA structure and dynamics predictions by chemical shifts looks very promising.

The application of chemical shift SAS to the decoding site generated an ensemble of states that was corroborated by both relative free energy calculations and the existing experimental structures. This is a proof of principle in generating a dynamic ensemble using chemical shifts without a known ensemble to reference to such as in previous work. In addition the comparison between RDC SAS and chemical shift SAS illustrates the utility of chemical shifts over RDCs in cases where there is little helical motion for determining the dynamics of the bulge. The comparison of the SAS distributions generated between the umbrella sampled pool and the MD pool of conformations highlights the importance of proper sampling over critical degrees of freedom. It is quite possible that without proper sampling that skewed distribution could be incorrectly accepted. It is imperative that more research be conducted to further understand the significance of the different ensemble sizes and what they represent.

4.5 References

1. Stelzer AC, Frank AT, Kratz JD, Swanson MD, Gonzalez-Hernandez MJ, Lee J, et al. Discovery of selective bioactive small molecules by targeting an RNA dynamic ensemble. Nature Publishing Group. 2011 Jun 26;7(8):553–9.

2. Rinnenthal J, Buck J, Ferner J, Wacker A, FÜrtig B, Schwalbe H. Mapping the Landscape of RNA Dynamics with NMR Spectroscopy. Acc. Chem. Res. 2011 Dec 20;44(12):1292–301.

3. Dethoff EA, Chugh J, Mustoe AM, Al-Hashimi HM. Functional complexity and regulation through RNA dynamics. Nature. 2012 Feb 15;482(7385):322–30.

4. Al-Hashimi HM, Walter NG. RNA dynamics: it is about time. Current Opinion in Structural Biology. 2008 Jun;18(3):321–9.

5. Stelzer AC, Kratz JD, Zhang Q, Al-Hashimi HM. RNA Dynamics by Design: Biasing Ensembles Towards the Ligand-Bound State. Angew. Chem. Int. Ed. 2010 Jun 25;49(33):5731–3.

6. Puglisi JD, Tan R, Calnan BJ, Frankel AD, Williamson JR. Conformation of the TAR RNA-arginine complex by NMR spectroscopy. Science. American Association for the Advancement of Science; 1992;257(5066):76–80.

7. Fourmy D, Recht MI, Puglisi JD. Binding of neomycin-class aminoglycoside antibiotics to the A-site of 16 S rRNA. Journal of Molecular Biology. Elsevier;

1998;277(2):347–62.

8. Fourmy D, Recht MI, Blanchard SC, Puglisi JD. Structure of the A site of Escherichia coli 16S ribosomal RNA complexed with an aminoglycoside antibiotic. Science. American Association for the Advancement of Science; 1996;274(5291):1367–71.

9. Duchardt E, Nilsson L, Schleucher J. Cytosine ribose flexibility in DNA: a combined NMR 13C spin relaxation and molecular dynamics simulation study. Nucleic Acids Research. 2008 May 23;36(12):4211–9.

10. Ferner J, Villa A, Duchardt E, Widjajakusuma E, Wohnert J, Stock G, et al. NMR and MD studies of the temperature-dependent dynamics of RNA YNMG-tetraloops. Nucleic Acids Research. 2008 Feb 5;36(6):1928–40.

11. Hall KB, Tang C. 13C relaxation and dynamics of the purine bases in the iron responsive element RNA hairpin. Biochemistry. ACS Publications; 1998;37(26):9323–32.

12. Hall KB. RNA in motion. Current Opinion in Chemical Biology. 2008 Dec;12(6):612–8.

13. Frank AT, Stelzer AC, Al-Hashimi HM, Andricioaei I. Constructing RNA dynamical ensembles by combining MD and motionally decoupled NMR RDCs: new insights into RNA dynamics and adaptive ligand recognition. Nucleic Acids Research. 2009 Jun 21;37(11):3670–9.

14. Wishart DS, Case DA. Use of chemical shifts in macromolecular structure determination. Meth. Enzymol. Elsevier; 2002;338:3–34.

15. Cromsigt JA, Hilbers CW, Wijmenga SS. Prediction of proton chemical shifts in RNA–their use in structure refinement and validation. J Biomol NMR. Springer; 2001;21(1):11–29.

16. Werf RM, Tessari M, Wijmenga SS. Nucleic acid helix structure determination from NMR proton chemical shifts. J Biomol NMR. 2013 Apr 6;56(2):95–112.

17. Barton S, Heng X, Johnson BA, Summers MF. Database proton NMR chemical shifts for RNA signal assignment and validation. J Biomol NMR. 2012 Nov 23.

18. Thomas JR, Hergenrother PJ. Targeting RNA with Small Molecules. Chem. Rev. 2008 Apr;108(4):1171–224.

19. Vicens Q. RNA's coming of age as a drug target. J Incl Phenom Macrocycl Chem. 2009 Jun 27;65(1-2):171–88.

20. Foloppe N, Matassova N, Aboul-ela F. Towards the discovery of drug-like RNA ligands? Drug Discovery Today. 2006 Nov;11(21-22):1019–27.

21. Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, et al. BioMagResBank. Nucleic Acids Research. 2007 Dec 23;36

22. Noller HF. RNA structure: reading the ribosome. Science; 2005;309(5740):1508–14.

23. Westhof E, Leontis N. Atomic Glimpses on a Billion-Year-Old Molecular Machine. 2000 Apr 19;:1–5.

24. Wimberly BT, Brodersen DE, Clemons WM, Morgan-Warren RJ, Carter AP, Vonrhein C, et al. Structure of the 30S ribosomal subunit. Nature. 2000 Sep 21;407(6802):327–39.

25. Noller HF. Biochemical characterization of the ribosomal decoding site. Biochimie. 2006 Aug;88(8):935–41.

26. Vicens Q, Westhof E. Crystal structure of a complex between the aminoglycoside tobramycin and an oligonucleotide containing the ribosomal decoding a site. Chemistry & Biology. 2002 Jun;9(6):747–55.

27. Vicens Q, Westhof E. Crystal Structure of Paromomycin Docked into the Eubacterial Ribosomal Decoding A Site. 2001 Jul 31;:1–12.

28. Kondo J, Francois B, Russell R, Murray J, Westhof E. Crystal structure of the bacterial ribosomal decoding site complexed with amikacin containing the γ-amino-α-hydroxybutyryl (haba) group. Biochimie. 2006 Aug;88(8):1027–31.

29. Hermann T. A-site model RNAs. Biochimie. 2006 Aug;88(8):1021–6.

30. Francois B. Crystal structures of complexes between aminoglycosides and decoding A site oligonucleotides: role of the number of rings and positive charges in the specific binding leading to miscoding. Nucleic Acids Research. 2005 Sep 25;33(17):5677–90.

31. Foloppe N, Mackerell AD Jr. All-atom empirical force field for nucleic acids: I. Parameter optimization based on small molecule and condensed phase macromolecular target data. J. Comput. Chem. Wiley Online Library; 2000;21(2):86–104.

32. Denning EJ, Priyakumar UD, Nilsson L, Mackerell AD Jr. 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. 2011 Apr 5;32(9):1929–43.

33. Nosé S. A unified formulation of the constant temperature molecular dynamics methods. J. Chem. Phys. 1984;81(1):511.

34. Hoover W. Canonical dynamics: Equilibrium phase-space distributions. Phys. Rev., A. 1985 Mar;31(3):1695–7.

35. Zhang Q, Stelzer AC, Fisher CK, Al-Hashimi HM. Visualizing spatially correlated dynamics that directs RNA conformational transitions. Nature. 2007 Dec 20;450(7173):1263–7.

36. Al-Hashimi HM, Gosser Y, Gorin A, Hu W, Majumdar A, Patel DJ. Concerted motions in HIV-1 TAR RNA may allow access to bound state conformations: RNA dynamics from NMR residual dipolar couplings. Journal of Molecular Biology. 2002 Jan;315(2):95–102.

37. Al-Hashimi HM, Pitt SW, Majumdar A, Xu W, Patel DJ. Mg2+-induced Variations in the Conformation and Dynamics of HIV-1 TAR RNA Probed Using NMR Residual Dipolar Couplings. Journal of Molecular Biology. 2003 Jun;329(5):867–73.

38. Casiano-Negroni A, Sun X, Al-Hashimi HM. Probing Na +-Induced Changes in the HIV-1 TAR Conformational Dynamics Using NMR Residual Dipolar Couplings: New Insights into the Role of Counterions and Electrostatic Interactions in Adaptive Recognition. Biochemistry. 2007 Jun;46(22):6525–35.

39. Zhang Q. Resolving the Motional Modes That Code for RNA Adaptation. Science. 2006 Feb 3;311(5761):653–6.

40. Frank AT, Horowitz S, Andricioaei I, Al-Hashimi HM. Utility of 1H NMR Chemical Shifts in Determining RNA Structure and Dynamics. J. Phys. Chem. B. 2013 Feb 21;117(7):2045–52.

41. Clore GM, Starich MR, Gronenborn AM. Measurement of residual dipolar couplings of macromolecules aligned in the nematic phase of a colloidal suspension of rod-shaped viruses. J. Am. Chem. Soc. ACS Publications; 1998;120(40):10571–2.

42. Hansen MR, Hanson P, Pardi A. Filamentous bacteriophage for aligning RNA, DNA, and proteins for measurement of nuclear magnetic resonance dipolar coupling interactions. Meth. Enzymol. 2000;317:220–40.

43. Hansen MR, Mueller L, Pardi A. Tunable alignment of macromolecules by filamentous phage yields dipolar coupling interactions. Nat. Struct. Biol. 1998 Dec;5(12):1065–74.

44. Popenda M, Szachniuk M, Blazewicz M, Wasik S, Burke EK, Blazewicz J, et al. RNA FRABASE 2.0: an advanced web-accessible database with the capacity to search the three-dimensional fragments within RNA structures. BMC Bioinformatics. 2010;11(1):231.

45. Rodnina MV, Daviter T, Gromadski K, Wintermeyer W. Structural dynamics of ribosomal RNA during decoding on the ribosome. Biochimie. 2002 Aug;84(8):745–54.

46. Selmer M. Structure of the 70S Ribosome Complexed with mRNA and tRNA. Science. 2006 Sep 29;313(5795):1935–42.

47. Berman HM. The Protein Data Bank. Nucleic Acids Research. 2000 Jan 1;28(1):235–42.

48. Zweckstetter M, Bax A. Prediction of Sterically Induced Alignment in a Dilute Liquid Crystalline Phase: Aid to Protein Structure Determination by NMR. J. Am. Chem. Soc. 2000 Apr;122(15):3791–2.

49. Salmon L, Bouvignies G, Markwick P, Blackledge M. Nuclear Magnetic Resonance Provides a Quantitative Description of Protein Conformational Flexibility on Physiologically Important Time Scales. Biochemistry. 2011 Apr 12;50(14):2735–47.

50. Tolman JR, Ruan K. NMR Residual Dipolar Couplings as Probes of Biomolecular Dynamics. Chem. Rev. 2006 May;106(5):1720–36.

51. Getz M, Sun X, Casiano-Negroni A, Zhang Q, Al-Hashimi HM. Review NMR studies of RNA dynamics and structural plasticity using NMR residual dipolar couplings. Biopolymers. 2007;86(5-6):384–402.

52. Shandrick S, Zhao Q, Han Q, Ayida BK, Takahashi M, Winters GC, et al. Monitoring Molecular Recognition of the Ribosomal Decoding Site. Angew. Chem. Int. Ed. 2004 Jun 14;43(24):3177–82.

53. Dethoff EA, Petzold K, Chugh J, Casiano-Negroni A, Al-Hashimi HM. Visualizing transient low-populated structures of RNA. Nature. 2012 Oct 7.

CHAPTER 5

Enhanced Sampling Procedure Improves Success Rate for Nucleic Acid–Small Molecule Docking

5.1 Introduction

In recent years, the proposed role of RNA as a simple carrier of information from the nucleus to the cytoplasm for translation has been radically overturned. It has been shown that non-coding RNAs (ncRNAs) perform a wide range of roles such as protein synthesis(1-3), self-splicing intron removal(4,5), pre-mRNA splicing(6-9) and telomere maintenance(6,10). All of these cellular processes now have additional potential drug targets due to RNAs involvement. In addition RNAs, such as the transactivation response element (TAR) from the human immunodeficiency type 1 virus (HIV-1), recruit and bind proteins that allow RNA polymerase to copy the HIV-1 genome and thus presents a potential drug target as well for the treatment of HIV-1(11,12).

With the emergence of RNA as an important drug target, techniques to discover novel ligands against RNAs are critical. Current approaches utilized in protein-ligand small molecule discovery are incompatible for RNA. The majority of conventional high-throughput screening methods rely on biomolecule enzymatic activity – RNA, with few exceptions, lacks enzymatic activity(13-18). Computational docking can, in principle, provide an alternative avenue to discover novel binders to RNA targets without the constraints of enzymatic activity. There have been several attempts to generate novel small molecule binders to RNA targets with various levels of success(13,15,17,18). RNA poses many new difficulties compared to proteins for computational docking. First, RNA has a high degree of flexibility, adopting radically different states on protein or ligand binding and, as such, it is unreasonable to represent an RNA drug target as a single

structure in computational drug discovery. Second, RNA has shallow and solvent-exposed binding pockets where potential ligands can bind. Third, RNA ligands tend to be flexible themselves, sometimes with greater than ten rotatable bonds(18-20). The high degree of flexibility of both the RNA and their ligands, in addition to the shallow binding pockets, make it very difficult to find global minimum energy poses. Another difficulty in applying computational drug design to RNA is the development of a scoring function that can correctly score native-like ligand poses. Multiple attempts have been made to update scoring functions unique to RNA or modify functions originally developed for proteins(19-21). These updated RNA-specific scoring functions appear to improve the success rates of finding native poses; however, the success rate is low compared to protein-ligand computational docking(22). In contrast to some other programs, here we investigate the use of a novel sampling protocol that uses a simplistic scoring function, to rigorously explore conformation space.

There have been many attempts to improve the performances of sampling. Glide(23-25), one of the premiere docking programs for proteins, fragments ligands into molecular fragments to remove internal degrees of freedom. Glide and other programs(19,26) that also fragment their small molecules, have a build up procedure where the largest fragment is placed first and each successive fragment is placed so it can connected to the previous until the entire molecule is built into a docked pose. These techniques can employ clustering to reduce the number of possible conformations considered. Our sampling technique follows this concept, also employing ligand fragmentation to deal with the difficulties of the internal degrees of freedom of small molecules. What differs from previous approaches is that we initially sample all conformations of each fragment independently, and with clustering, find hotspots for each fragment. Afterwards, there is a reconnection algorithm that attempts to reconnect the fragments back together preserving each fragments independent favorable conformation into a favorable pose. This approach has similarities to the multiple copy simultaneous search method (MCSS)(27), which placed functional groups over the surface of proteins finding hotspot locations for different functional groups. This algorithm was later employed in HOOK(28), which attempted to find novel protein-ligand binders by connecting the functional groups found

by MCSS with a skeleton scaffold into a full small molecule. In addition the Small Molecule Growth (SMoG)(29) algorithm also utilized a similar ideal of independently indentifying functional group hotspots and later joining them together with a scaffold. The fundamental difference between both HOOK and SMoG and our approach is that we are docking previous existing small molecules not generating novel ones. But, the concept of independent optimization followed by reconnection has been shown to be successful and we employ it here for dealing with the large number of rotatable bonds of RNA and DNA small molecule binders.

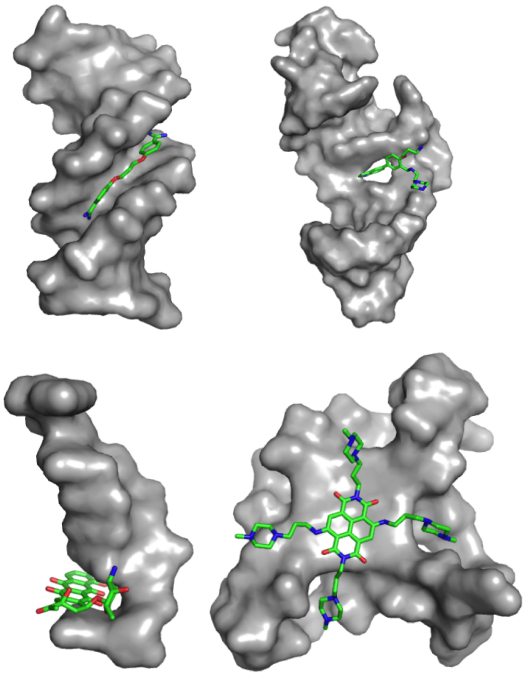

Figure 5.1: Examples of the various types of receptors in the nucleic acid database. Top-left, minor groove DNA binder. Top-right, RNA bulge binder. Bottom-left, single stranded DNA binder. Bottom-right G- quadruplex binder.

Previous RNA computational docking strategies have typically used varying benchmarking sets of RNA-ligand complexes due to the lack of experimentally determined structures. Here, we have compiled a much larger set of complexes to benchmark and validate our new docking algorithm: 230 complexes in total. By including DNA-ligand complexes, which share many of the same characteristics of RNA-ligand complexes, we were able to nearly double the dataset. Compared to previous sets, which have at most 50 complexes with experimental binding affinities, our dataset has 89. In addition to demonstrating that our new sampling algorithm is effective and efficient compared to previous algorithms, we also plan to release this dataset to be used to benchmark RNA docking algorithms, similar to either the Ligand Protein Database (LPDB)(30), Mother of all Databases (MOAD)(31) or The Community Structure−Activity Resource (CSAR)(32-34) for protein-ligand docking.

5.2 Strategies and Components

5.2.1 Reduced Ligand Topology

Dividing ligands into fragments that do not contain rotatable bonds, removes the requirement to sample the internal degrees of freedom of a small molecule during the search phase of computational docking. Fragmenting the ligand also allows for the more efficient placement of key chemical groups for favorable ring stacking and hydrogen bond interactions, which play a large role in RNA-ligand interactions(18). However, not all fragments will contribute equally to the total interaction energy, in previous docking algorithms, usually the largest fragment is placed first and the rest of the molecule is built onto it using the internal coordinates to rebuild the molecule. This approach can unfortunately still lead to suboptimal solutions, particularly when a ligand contains two fragments of equal size (number of heavy atoms). Instead of building up a molecule by adding each consecutive bonded fragments, in the sampling scheme we explore below, all fragments with five or greater heavy atoms are independently scored and are connected to produce the core portion of the molecule. The connectivity information for each fragment is stored to later build the rest of the molecule.

5.2.2 Enhanced Sampling Procedure

We devised a simple yet effective solution based on exhaustive sampling to ensure that the most favorable energy minimum pose is identified. Exhaustive docking procedures have been attempted before, but are prohibitively slow and only exhibit minor pose improvements when implemented in protein-ligand docking(35,36). The difference for RNA-ligand docking is the search space is much more vast then the usual deep and defined pockets found in proteins. The critical quality of the success of an exhaustive algorithm is to minimize the computational cost and only evaluate poses that are likely to be the top conformations. Our algorithm takes full advantage of this feature and has an average runtime of ~10 minutes to find the maximum score of a ligand with a given RNA target. A detailed explanation of our sampling procedure follows.

Generating Fragment Hotspot Clusters

Initially, a grid with 0.5 Å resolution is generated using a given defined center or the center of a known ligand. This grid resolution has been shown to be ideal from previous studies(37). At each grid point three scores are calculated: a contact score, a ring stacking score and a hydrogen bond score, details of each term are described at great length in Section 5.2.3. The contact score is used to determine whether or not to sample at a given grid point. If a contact score is less favorable then a cutoff then there is most likely a steric clash at that point and it would be computationally wasteful to sample there. Typically between 2000 and 4000 points are sampled depending on the size of the binding pocket. Each fragment is rotated about its center of mass in 18º increments generating 5832 rotations in all. This increment size, 18º, was chosen as a compromise between speed and accuracy: previous studies using exhaustive sampling used 10º increments, arguing that it is the minimum required to accurately resolve proper hydrogen bond geometry(35,36). In our own benchmarks this appears to be a bit over cautious and 18º performs just as well as 10º (data not shown).

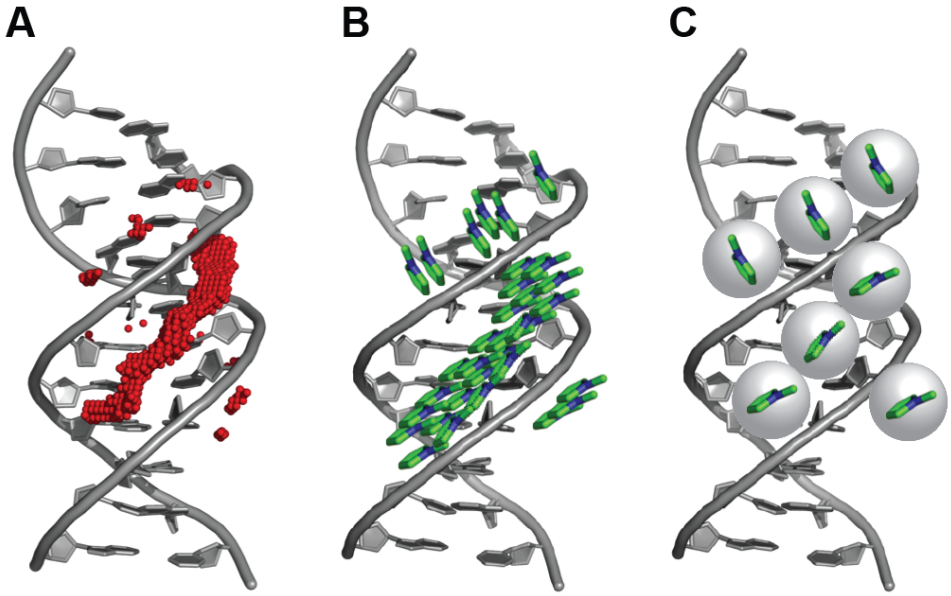

Figure 5.2 Initial exhaustive sampling procedure. A) Red points are all the grid points where each fragment is sampled. B) Visualization of the all the possible solutions for a single fragment at each grid point. C) After clustering the centroids of each cluster shown, which are the top poses for the sampling volume.

Following the production of the rotations, the center of mass of each rotated fragment is then moved to each grid point and is scored. Rotating the fragment prior to moving it to

each grid point greatly speeds up the algorithm, as rotations are computationally costly in comparison to translations. At each grid point the rotation that yields the maximum fragment score (MFS) is recorded. MFS is the theoretical maximum score that a fragment can achieve at a specific grid point with the resolution of 0.5 Å and a rotational increment of 18º. While saving only the MFS requires successive score evaluations later, storing all the scores for each rotation requires enormous amounts of data (several gigabytes) per fragment. The MFS from each allowable grid point is then sorted from most favorable to least favorable, followed by centroid clustering procedure, where the grid point with the most favorable MFS is the first centroid and all grid points within 3 Å are considered part of this cluster, the next most favorable grid point outside this cutoff becomes the second centroid, and all grid points within 3 Å not already assigned to a previous clusters become part of this cluster, continuing until every grid point is assigned a cluster. Clustering this way creates a hierarchical grouping of poses to efficiently searching poses, as the MFS of the centroid represents the effective maximum score that a given volume on the binding surface can produce. This clustering step is critical to the success of this approach to reduce search time when fragments are reconnected later. The computational cost up to this point has been rather low; the initial sampling for each fragment takes approximately 30 seconds to one minute on a 3.2 GHz i7 Intel CPU. This observation was key to the further development of our algorithm; it is computationally inexpensive to exhaustively determine the most favorable hotspots for fragments to bind.

Building Fragment Connection Graphs

With all fragment poses organized into clusters it is now possible to apply basic distance constraints that will eliminate a large percentage of possible combination of fragment poses. The centroid of each fragment cluster is assessed to determine whether cluster A is close in space to the centroid of cluster B, to which it should be covalently attached. Since the cluster radius is 3 Å, it is impossible for two clusters to be connected if their centroid distance is separated by greater than 6 Å plus their bond distance. If fragment A was initially bonded to fragment B, then clusters from fragment A are paired with the clusters from fragment B within this distance constraint. Once connected, each pair of clusters can now be connected to another pair of clusters that share a common fragment

between them. Assuming fragment A is also connected to fragment C, then a pair of clusters from AC and AB can be joined if the same fragment A cluster is common between them. The possibility of steric collisions is assessed for combinations of fragment clusters of three or more: if the centroids of two fragments that are not covalently bonded to each other are within 4 Å, than that group of clusters is rejected. Once one cluster for each fragment are connected, the center of mass of the entire molecule is calculated and if it is greater then 5 Å from the center of the grid this combination of fragment clusters is rejected in order to prevent generating poses outside the selected binding site. This part of the algorithm is the only portion that is highly memory intensive and if more then one million partial molecules are generated with more fragments to be connected, a pruning stage is included in which partial molecules are removed if their total scores are below the average. When all possible combinations of fragment clusters are generated, clusters are sorted by the summation of each fragment's MFS. In the next section, the pose will be built back up using the clusters that have the potential to yield the highest score first.

Reconnection of Fragments

Starting with the pair of clusters that together have the highest total MFS, each of the poses within each cluster is paired together and is sorted by it's combined MFS. Again, this score represents the upper limit of the potential score, and sorting them in this fashion allows the pair to be selected and no further search is required. A repeated theme of this algorithm involves taking advantage of the combined MFS of fragments to determine whether it is possible to beat the existing best score. At this stage of the algorithm, each pair of fragments currently exist in random orientations with respect to one other and must be reoriented in order to reassemble the molecule with intact bonding geometry. The alignment is performed with a center of mass rotation such that the connection atom on each fragment - the atom that bonds to an atom on the other fragment – is pointed at the center of mass of the other fragment. Once the two fragments are aligned, the distance between the two connecting atoms is calculated and the algorithm continues only if the distance is within ±10% of the original distance. A bond vector is now defined between the two fragments and each fragment is rotated about this axis in

4.5° increments. Only rotations that produce favorable scores are saved and are clustered based on the root-mean square deviation (RMSD) of the position of their connection points to fragments yet to be connected. These clusters are then paired between the two fragments with their scores added and they are sorted from most favorable to least favorable. The starting core of the pose, which is the two fragments with the best potential score, is now connected. This core could be thought of as the head node in a tree data structure, where the edges leading to the next level nodes are fragments that were originally bonded to the core two fragments before the molecule was split. In the next level of nodes there is a collection of clusters that are close enough to the position of the core where it is possible to connect them based on center of mass distance. Again, similar to generating the molecular core, each pose in a cluster is aligned using a rotation to best optimize the distance between the connecting atom between the new fragment and the existing fragment in the core that it is bonded too. This is continued down the fragment connection tree until all fragments are connected and a total pose score is determined.

MFS Optimization of Fragment Connection Space

Summing each fragment's MFS is a powerful metric to judge whether a partial pose has the ability to have an actual score better then current best finished pose. Because exhaustively searching the connection space is very computationally intensive, exploiting this cutoff makes it feasible to finish a docking run in a realistic amount of time. The summation of each fragment's MFS in a partial pose is unrealistic as each fragment's conformation is optimized to yield the highest score without knowledge that it needs to be reconnected. Because of this, when the fragments are connected the score is always equal or less favorable. Before a finished pose is generated, no cutoff is used as cluster combinations that have the highest summed MFS may not be successfully connected based on the rules discussed in the last section. Once the first finished pose is generated and the poses are scored with the connected conformation, we now have an actual score to compare too. These are the connected fragment scores (CFS) and they can be equal to or less favorable than the MFS. The summation of the CFSs for all the fragments becomes the cutoff score. Any combination of fragments with a summed MFS less then the known most favorable CFS is discarded.

Optimization of Flexible Linkers and Chains

As mentioned previously, this search procedure is a two-step process. First initial placement of the core of a molecule is performed, ignoring long flexible chains of carbon and/or nitrogen single bonds. These chains are quite common in nucleic acid small molecule binders and greatly increase the internal coordinate search space. Here we first place the core of the ligand that acts as an anchor to build the flexible chains of a given small molecule.

All the possible conformations for flexible chains are built at the initiation of the program. At each rotatable bond in the chain, a set number of possible rotations is sampled. For example, if there were three fragments with two rotatable bonds, with an increment of 20°, then there would be one conformation of the first fragment, 18 of the second and 324 for the third. The conformations are stored in a tree data structure and when a given chain needs to be built the head node is aligned so it properly connects to the ligand core with the correct bond and angles that are required from its parameter file. This alignment is then stored in a homogenous transform and is propagated down the tree when the next fragment in the chain is being optimized. Generating all the conformations beforehand trades memory for speed, greatly accelerating the procedure for a moderate use of memory. This optimization procedure is similar to the exhaustive methods developed for protein loop modeling where sampling over small changes in ϕ and ψ space can greatly improve results(38,39).

Optimization of the chains is done in two discrete steps: initially a depth first search is performed to find a possible solution, followed by a more thorough solution. This second search attempts to find the conformation that yields the most favorable score in the fewest visits to nodes in the conformation tree. Using the current best solution as a reference, any partial conformation that could not produce a more favorable score, assuming that the most favorable possible scores for the remaining fragments is not visited. This greatly reduces computational time.

5.2.3 Scoring Function

The focus of many of the leading RNA-ligand docking programs has been the development of cutting edge scoring functions to better discriminate native poses from non-native ones. Whether it is building better generalized Born models for implicit solvation and neutralization of charge, or improving electrostatic models, as both RNA and its ligands are highly charged, these are certainly advantageous and bring the field closer to using models that better represent the underlying atomic level interactions(19). Because of the number of poses that are evaluated in our exhaustive sampling procedure, any energy term that could not be represented as a grid-based potential would be too slow to be useful. Thus our scoring function has to be as simple as possible while still having a high discrimination potential.

	Aromatic	Aliphatic	Hbond Donor	Hbond Acceptor
Aromatic	3.4	3.6	3.4	3.4
Aliphatic	3.6	3.6	3.5	3.5
Hbond Donor	3.4	3.5	3.4	3.0
Hbond Acceptor	3.4	3.5	3.0	3.4

Figure 5.3: The ideal distance between each atom type in the contact scoring term, all distances are in Angstroms.

Our scoring function has three terms: a contact score, a ring stacking score and a hydrogen bond score. The contact score is a simplistic van der Waals substitute using a Lennard-Jones potential (**Equation 5.1**) with three atom types, aliphatic, aromatic and hydrogen bonder, only heavy atoms are considered and the contact distance reflects the distance between heavy atoms. Each atom type combination has a different optimal distance between their centers (**Figure 5.3**). Ideal distances were generated by measuring all the contacts in our nucleic acid docking set and fitting each pairing of atom types to a Gaussian taking the center as the ideal distance. Compared to the aromatic and aliphatic atom types, hydrogen bond donors and acceptor types have a shorter ideal distance of 3 Å. This is compared to the 3.4 Å and 3.6 Å for aromatic and aliphatic pairings. This is because around 3 Å is the ideal distance for hydrogen bonds, this is why there is a separate type for hydrogen bond donors and acceptors. If they were lumped in with the others the ideal distance will not permit them to be close enough to form a hydrogen bond.

$$S_{contact} = \left(\frac{r_{\min}}{r}\right)^{12} - 2\left(\frac{r_{\min}}{r}\right)^{6} \tag{5.1}$$

The ring stacking is similar to the term found in Ribodock(20). In Ribodock, both how perpendicular a ligand ring is with respect to a receptor ring and the angle separating them. **Equation 5.2** outlines the equation used in our approach, where θ_{plane} is the angle between the plane of the ring and the grid point, and r is the distance. 3.6 Å is the average distance between a stacking ligand and receptor ring. The first part of the term is a weight based on the how parallel the rings are to each other, and the second part is a Gaussian centered on the ideal distance between two stacking rings.

$$S_{stack} = \exp\left(-\left(\frac{\pi}{4} - \theta_{plane}\right)\right)\exp\left(\frac{(r-3.6)^2}{0.4^2}\right) \tag{5.2}$$

Lastly, the hydrogen bond potential is modeled as an 8-6 potential based on the distance between the acceptor and heavy atom connected to the donor (**Equation 5.3**). In addition, there are angular cutoffs to verify proper geometric conditions for hydrogen bonding. For all acceptors, a plane angle is utilized with a cutoff of 40° out of the plane. For specifically sp^2 oxygen acceptors an angle between the carboxyl carbon, the oxygen atom and the heavy atom of the hydrogen donor is also utilized with a 30° cutoff around both the 45° and 135° position. These are the optimal angles for hydrogen bonding for Sp^2 oxygen. Similarly donors have a plane angle with a cutoff of 40° and an angle cutoff of 145° for donor heavy atom, donor and acceptor.

$$S_{hbond} = \left(\frac{r_{\min}}{r}\right)^{8} - 2\left(\frac{r_{\min}}{r}\right)^{6} \tag{5.3}$$

5.3 Materials and Methods

5.3.1 Preparation of Complexes

Preparation began with querying the protein databank for all nucleic acid structures containing a ligand. Filtering criteria were then applied to remove structures that were not appropriate for docking. Criteria for rejection of a complex included: ligands smaller than 5 heavy atoms or greater than 100; complexes with multiple copies of the same ligand in a single binding pocket, i.e. stoichiometry > 1, ligands with elements not standard such as platinum or zinc in organic molecules. Each receptor was isolated and minimized for 100 steps using steepest descents with a harmonic restraint of 10 Newtons on all heavy atoms. Three-dimensional coordinates of each ligand were independently downloaded from the PDB's ligand summary resource and were protonated and minimized using the ChemAxon software package (http://www.chemaxon.com). MATCH was then employed to generate topology and parameter files for each ligand(40,41). Complexes in which it was not possible to generate either the topology or parameter files were ultimately removed, yielding a total of 230 complexes in total.

5.4 Results and Discussion

5.4.1 Analysis of Nucleic Acid Training Dataset

Previous studies developing docking protocols have also collected a set of RNA-small molecule complexes to benchmark and validate their respective algorithms. These sets have been small, ranging from 30 to 50 complexes with experimentally determined binding affinities. The limited dataset is a direct consequence of the paucity of structural data available in the Protein Databank for RNA. In this study we take a different approach: in addition to benchmarking and validating the algorithm on complexes with known binding affinities, we also prepared all nucleic acid structures including DNA-ligand complexes, which amounts to 230 complexes, and targeted the ligand poses. The addition of DNA into the set complements the RNA, adding almost 100 more complexes. The major difference between DNA-binding compounds and RNA-binding compounds is that about half bind to the minor groove of DNA helices, which is an uncommon binding mode for RNA. However, the overall characteristics of the ligand-receptor complex are identical in that both exhibit shallow and solvent exposed binding pockets, have highly charged ligands that possess high rotatable bond counts and an abundance of hydrogen bond interactions. The set represents a much more rigorous test of ligand placement algorithms for researchers. In addition the set also contains 89 complexes with

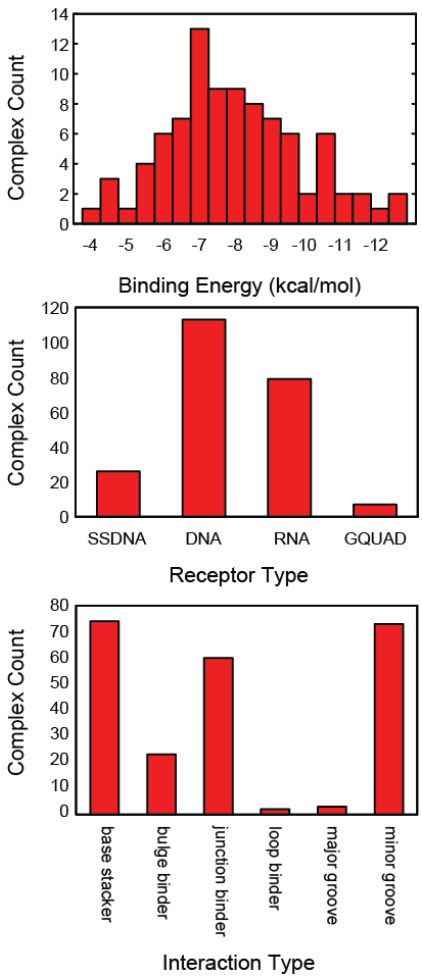

Figure 5.4: Analysis of nucleic acid docking set. A) Distribution of Gibbs free energy of binding of the subset of complexes with experimentally determined binding affinities, 89 in total. B) Breakdown of receptor types, single stranded DNA (SSDNA), DNA, RNA and G-quadruplex DNA (GQUAD). C) Distribution of binding mode of ligand, not mutually exclusive.

experimentally determined binding affinities, almost doubling the next largest set for nucleic acid docking validation(19,21,42) (**Figure 5.4**).

Sets such as the LPDB(30) and MOAD(31) are high quality datasets for protein-ligand docking that greatly helped to standardize current benchmarks. In addition to containing large number of complexes with known binding affinities, they are both easy to use, easy to retrieve via a web interface and contain exhaustive characterization of the ligand, receptor and complex. Following this model, in addition for the use in benchmarking of the current algorithm discussed here, we plan to release this set to create the first standard for benchmarking current and future nucleic acid docking algorithms. This dataset contains the same format as LPDB(30) and MOAD(31) in the sense that there is classification of binding, and of ligand, receptor and complex characteristics, and the set will be accessible online. We expect this set to foster rigor and consistency in benchmarking and validation of nucleic acid docking programs so that the comparison between differing docking programs will become easier and will lead to improvements in design of docking algorithms.

5.4.2 Validation of Placement Procedure

Fragment Size Preference

One of the key concepts employed in our approach is that molecular fragments can be independently docked into a receptor and then later optimized to restore their original connectivity. So here we investigate whether sampling with a spatial resolution of 0.5 Å and a rotation increment of 18° yield favorable poses that are near the native bounds conformations. Furthermore, how does the size of the fragment affect its specificity to native-like conformations? To examine specificity of fragments as a function of size (# of heavy atoms), the results of the exhaustive sampling procedure's clusters were analyzed. For each fragment, the cluster containing the native conformation was recorded and the average cluster number per fragment size is shown in **Figure 5.5**. A clear trend emerges: as the size of the fragment increases, the number of clusters required to find the native conformation decreases. For example, cluster zero has the lowest score and each cluster contains, on average, 100 to 200 conformations. The fewer the clusters, the exponentially faster the procedure will yield a native-like pose, assuming that the native pose is the lowest energy. For this reason, fragments below five atoms are not initially searched and are refined in the second stage of pose generation.

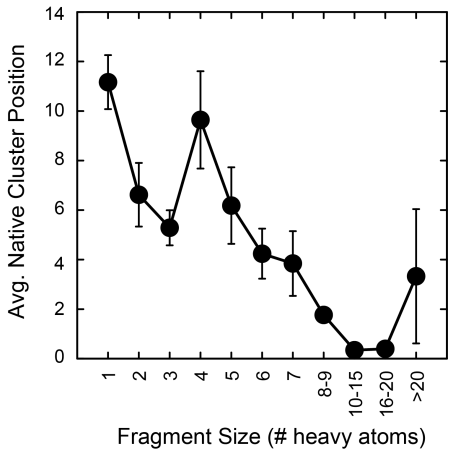

Figure 5.5: The average position of a native conformation in the spatial clusters determined by the exhaustive sampling procedure compared to the number of heavy atoms in a molecule fragment

Validation of Flexible Chain Optimization

The second distinct part of our procedure is to reattach flexible chains after the core or rigid portion of the molecule is placed. Using a bond rotation increment of 40° (nine rotations per bond), each flexible chain in each complex was considered in this validation. To validate only the flexible chain portion of the placement procedure, the remaining part of the molecule was kept in its native conformation. The number of rotations about each bond is low compared to other algorithms, which could have been insufficient to accurately capture native conformation. This was a major concern, due to the exhaustive nature of the sampling; even slightly increasing the number of rotations per bond adds an

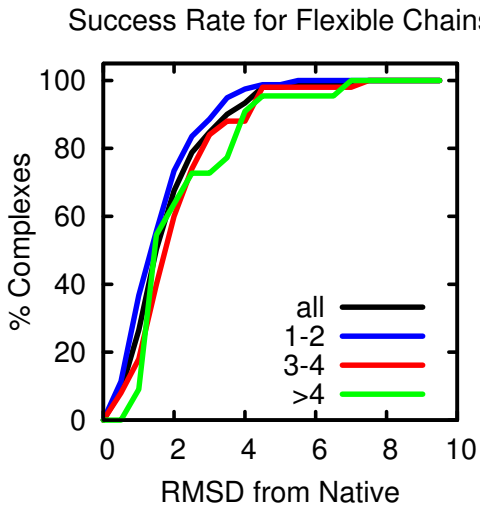

Figure 5.6: Success rate of refining flexible segments of ligands when non-flexible segments are in their native conformation. Black is all flexible chains, blue contains only chains with one or two rotatable bonds, red contains chains with thee or four and green are all chains with more then four rotatable bonds, the max is seven.

exponential amount of time to the search procedure. This in the end turned out not be a significant problem as a high level of accuracy was achieved (**Figure 5.6**). The percentage of flexible chains that were within 2.5 Å of their native conformation was 79% for all chains. For success rates based on number of rotatable bonds, chains with 1-2 rotatable bonds was 84%, 3-4 rotatable bonds was 74% and >4 rotatable bonds was 73%. The high level of accuracy of the 1-2 rotatable bond set was expected since there are only two degrees of freedom; however, a success rate of 73% for flexible chains with >4 rotatable bonds is quite surprising. The least successful ligand in the >4 category was 2BEE, which has seven long flexible chain atoms that form no contacts with its receptor and exposed entirely into solvent.

Validation of Initial Core Placement

Validation of the first stage of the placement algorithm is critical to benchmark independently. The second validation stage, which builds the flexible chains until the core of the molecule has a high degree of accuracy, assuming the core of the molecule is already in its native conformation. In this stage of validation we investigate whether it is possible to generate the core or rigid segments of a pose independent of the flexible chains. We used our exhaustive docking procedure to reconnect the fragments of the rigid portion of each molecule in the dataset (**Figure 5.7**). It should be noted that the entire molecule is not being rebuilt – just the two highest scoring fragments – with the assumption that connecting fragments that are in favorable binding poses will yield a complete molecule pose that is also favorable. In all, 63% of the complexes generated poses within 2.5 Å of the native conformation out of the top five best poses. This is

extremely good for a set of 230 and is quite striking given that the entire molecule is not being used during this stage. The high success rate gives confidence that it is possible to dock each fragment independently, rank the most favorable clusters for the fragment to bind and then connect the fragments to each other.

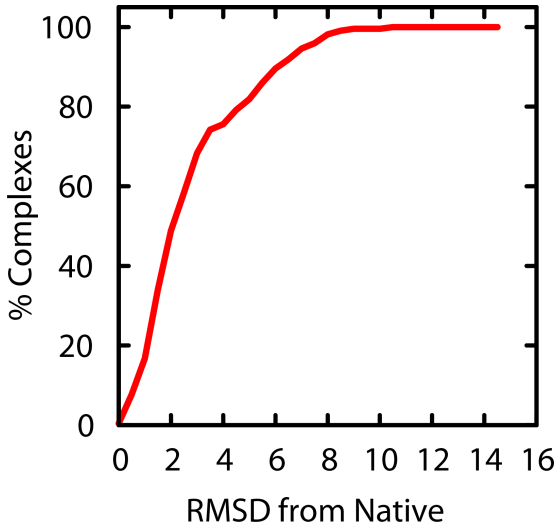

Figure 5.7: Success rate of refining the rigid segments of each ligand compared to their native conformation.

Because this set is by far the largest and most difficult validation set, it is difficult to judge the quality of this success. For this reason we also used the ICM docking program as a point of comparison, which has shown success in the Al-Hashimi lab in the past, finding multiple novel small molecules that bind to HIV-1 TAR, including one that inhibiting HIV-1 replication *in vivo*(13). Using ICM we generated ten docked poses for each complex in the set. Unfortunately only 197 could be processed by ICM, it was unclear why specific complexes could not be processed, so we examined only this subset. Fortunately, the overall success rate for our own procedure did not change from 63%. To make a one-to-one comparison only the atoms that were docked in this stage by our algorithm were considered, since ICM's docking algorithm is not a fragment based approach, all atoms were docked at once. This should give ICM an edge in this docking stage as it is possible that some poses require the entire molecule to be present to correctly position the rigid section of the ligand. ICM successfully docked only 55% below 2.5 Å RMSD from the native pose, considering the top five poses produced (**Figure 5.8**). This result should be expected as the extra degrees of freedom from the flexible portions of each ligand hinder the success of correctly placing the rigid core. This results further validates our approach to independently place the rigid core of the molecule before optimizing the flexible chains and linkers.

5.4.3 Success of Full Pose Placement

Putting all the refinement sets together yields full ligand docked poses. For the entire set of 230 complexes, our procedure produced docked poses with 67% below 3 Å RMSD using the top five poses generated. To compare to our previous success rate of only docking the core of the molecule with a success rate of 63% below 2.5 Å. This is compared to the 56% when one considers the entire molecule (**Figure 5.9A)**. This drop is expected; docking in the flexible parts is by far the hardest part. When docking the flexible chains starting with the native pose, on average we successfully placed 79% within 2.5 Å, thus some drop off in total accuracy makes sense. To again attempt to put these results in context, we can compare to the success rate of ICM on the subset of 179 structures. Here ICM had a success rate of 45% within 2.5 Å, which experienced an even greater dip in success rate compared to considering only the core of the molecule, which was at 55% discussed in the previous section. So again it appears that our novel exhaustive strategy is comparable to the results of current leading RNA docking programs. If we consider the success for ICM at 3 Å to compare to the recent benchmarking study, it achieves 52%. This would put it right under Glide's 55%. So it is definitely still having comparable success rates to other leading programs.

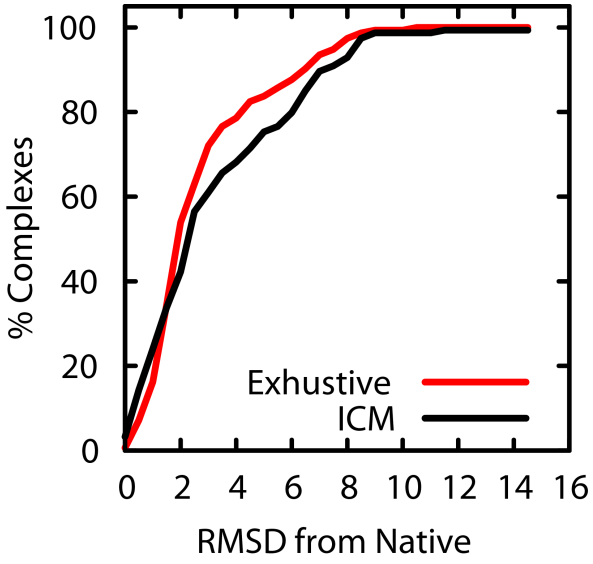

Figure 5.8: Success rate of refining the rigid segments of each ligand compared to their native conformation for both our novel algorithm compared to ICM.

There were a couple of classes of ligands that showed differences using our approach compared to ICM. The first class is acridine ligands that bind to G-quadruplex. These are fused aromatic ring compounds with long flexible linkers to sugar like rings. An example is 3ERU. While our approach was able to successfully dock these ligands bellow 2.5 Å RMSD from the native, ICM generated poses that were 9-10 Å from the native. The long flexible linkers that connect the sugar rings to the aromatic rings are difficult to correctly

dock, splitting up the ligand and independently placing each fragment yields a huge advantage in this case as the linker is trivial to optimize once it is clear where all the rings need to be placed. Another class where our approach was yields superior results is with berenil-based derivatives that bind to the minor groove of DNA. These ligands consist of a series of five or six membered aromatic rings connected by single inflexible bonds. In addition to fragmenting rotatable bonds, our algorithm also fragments very large fragments where it is possible to segment them by a single bond such as in the case of these ligands. This is key to being able to get native like poses as they are very bulky and difficult to correctly fit into the minor groove of DNA. Once split, they easily form native like contacts with the minor groove generating poses bellow 2.5 Å RMSD, an example of this type of binder is 269D. ICM appears to do better then our approach is aminoglycosides, these are sugar like compounds with lots of NH3 groups on them. These ligands are absolutely covered in hydrogen bond donors and acceptors. Since we do not include a solvation term, all hydrogen bonds are attempted to be satisfied and situations where part of an aminoglycosides is sitting in solvent does not score well with our scoring function. Our scoring function favors burying these rings as deep as possible into the RNA leading to incorrectly docked poses. This is an area that we are still looking into and considering adding a simple solvation term to counter act situations like this.

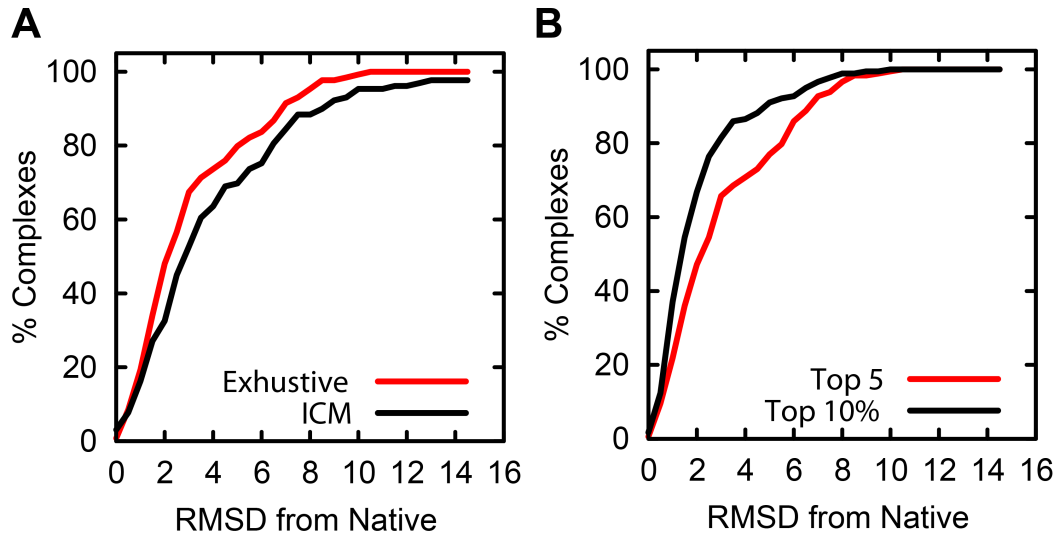

Figure 5.9 A) A comparison between the success rate between our exhaustive docking procedure and ICM on the subset of 179 complexes. B) A comparison in success rate between using the top five poses produced by our procedure and the top 10% of structures produced.

Due to its exhaustive nature it is possible to keep generating poses that are of more favorable score then the most favorable poses. Instead of setting the score cutoff to be higher then the current highest pose; it can be set to an arbitrary value. To investigate whether native like poses were being sampled for some of the complexes that did not yield top near-native poses, a score cutoff was set for each complex that was 10% below the most favorable pose. If we consider the pose with the best RMSD from these poses the success rate within 2.5 Å goes up to 76%, which is a noticeable increase (**Figure 5.9B)**. This demonstrates that more native-like poses exist in the top 10% of favorable poses but are not within the top 5 most favorable poses. A possible explanation for this is that the scoring function used is rather simplistic and is not capturing some of the finer interactions required to score native-like poses more favorably then near native poses. Future work will explore implementing a second level of scoring using a more complex scoring function which could select out more native-like poses from the top 10% of poses yielding a success rate more similar to 76%.

5.4.4 Comparison to Other Leading Programs

In a recent paper(42), GOLD, Ribodock, Glide, Autodock4.1 and Surflex-dock were benchmarked on a set 56 complexes. The highest success rate (<3.0 Å) using the top five poses of 73% was generated by GOLD and Ribodock. None of the other programs did better then our 67%. The next highest was Glide with 55%. This suggests that our exhaustive sampling strategy already performs as well as some of the best RNA docking programs out there, but to directly compare with these results, we ran both our exhaustive sampling procedure and ICM on the same 56 structures. The exhaustive sampling achieved a success rate of 66%, while ICM yielded a success rate of 57%. Given the same set of complexes the exhaustive sampling procedure was able to out perform all but GOLD and Ribodock. This set has a much larger fraction of aminoglycosides (26/56) then the set of 230 (31/230) complexes used earlier. As discussed earlier, complexes with aminoglycosides have a lower success rate with the exhaustive sampler. Out of the 26 complexes with aminoglycosides only 12 were successively docked below 3.0 Å rmsd compared to the 18 and 13 of GOLD and Ribodock respectively. Thus it appears taking

steps to improve docking aminoglycosides could radically improve the overall success rate, as for other ligand the exhaustive sampler out performed GOLD and was on par with Ribodock.

5.5 Conclusion

We have demonstrated the ability of using a simplistic scoring function with an exhaustive sampling procedure to generate high quality docked poses using a newly developed nucleic acid / small molecule set of 230 complexes. This procedure out performs ICM, a leading docking program, in a one to one comparison on this complex set. Furthermore, the computational cost of running our program, which only needs to be done once, was faster then it took ICM to generate 10 poses. The average run time is approximately ~10 minutes per ligand once the receptor is prepared. In addition, the success rate of 67% within 3 Å RMSD of the native pose using the top five poses produced for each complex is comparable to that of the leading RNA docking programs benchmarked in a smaller set of 56 complexes. Using this sampling technique in conjunction with a second level of scoring could possibly produce even higher success rates as investigating the 10% most favorable scoring conformations contained poses that were below 2.5 Å 76% of the time. Our approach demonstrates that in addition to improving scoring functions, there is still additional way to improve success rates with novel enhanced sampling techniques.

5.6 References

1. Noller HF. RNA structure: reading the ribosome. Science [Internet]. American Association for the Advancement of Science. American Association for the Advancement of Science; 2005;309(5740):1508–14.

2. Noller HF. Biochemical characterization of the ribosomal decoding site. Biochimie. 2006 Aug;88(8):935–41.

3. Wimberly BT, Brodersen DE, Clemons WM, Morgan-Warren RJ, Carter AP, Vonrhein C, et al. Structure of the 30S ribosomal subunit. Nature. 2000 Sep 21;407(6802):327–39.

4. Kruger K, Grabowski PJ, Zaug AJ, Sands J, Gottschling DE, Cech TR. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening

sequence of Tetrahymena. Cell. Elsevier. 1982;31(1):147–57.

5. Guerrier-Takada C, Gardiner K, Marsh T, Pace N, Altman S. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. Cell. Elsevier. 1983;35(3):849–57.

6. Blackburn EH. Structure and function of telomeres. Nature. 1991;350(6319):569–73.

7. Brody E, Abelson J. The“ spliceosome”: yeast pre-messenger RNA associates with a 40S complex in a splicing-dependent reaction. Science. 1985;228(4702):963–7.

8. Guthrie C. Messenger RNA splicing in yeast: clues to why the spliceosome is a ribonucleoprotein. Science. 1991;253(5016):157–63

9. Wahl MC, Will CL, Lührmann R. The Spliceosome: Design Principles of a Dynamic RNP Machine. Cell. 2009 Feb;136(4):701–18.

10. Shippen-Lentz D, Blackburn EH. Functional evidence for an RNA template in telomerase. Science. American Association for the Advancement of Science. 1990;247(4942):546–52.

11. Krebs A, Ludwig V, Boden O, Göbel MW. Targeting the HIV Trans-Activation Responsive Region-Approaches Towards RNA-Binding Drugs. ChemBioChem. 2003 Sep 26;4(10):972–8.

12. Mei H-Y, Galan AA, Halim NS, Mack DP, Moreland DW, Sanders KB, et al. Inhibition of an HIV-1 Tat-derived peptide binding to TAR RNA by aminoglycoside antibiotics. Bioorganic & Medicinal Chemistry Letters. Elsevier. 1995;5(22):2755–60.

13. Stelzer AC, Frank AT, Kratz JD, Swanson MD, Gonzalez-Hernandez MJ, Lee J, et al. Discovery of selective bioactive small molecules by targeting an RNA dynamic ensemble. Nature Publishing Group. 2011 Jun 26;7(8):553–9.

14. Parsons J, Castaldi MP, Dutta S, Dibrov SM, Wyles DL, Hermann T. Conformational inhibition of the hepatitis C virus internal ribosome entry site RNA. Nat Meth. 2009 Sep 20;5(11):823–5.

15. Lind KE, Du Z, Fujinaga K, Peterlin BM, James TL. Structure-based computational database screening, in vitro assay, and NMR assessment of compounds that target TAR RNA. Chemistry & Biology. Elsevier; 2002;9(2):185–93.

16. Blount KF, Breaker RR. Riboswitches as antibacterial drug targets. Nat Biotechnol. 2006 Dec;24(12):1558–64.

17. Tuccinardi T. Binding-interaction prediction of RNA-binding ligands. Future

Medicinal Chemistry. 2011 Apr;3(6):723–33.

18. Thomas JR, Hergenrother PJ. Targeting RNA with Small Molecules. Chem. Rev. 2008 Apr;108(4):1171–224.

19. Lang PT, Brozell SR, Mukherjee S, Pettersen EF, Meng EC, Thomas V, et al. DOCK 6: Combining techniques to model RNA-small molecule complexes. RNA. 2009 May 20;15(6):1219–30.

20. Morley SD, Afshar M. Validation of an empirical RNA-ligand scoring function for fast flexible docking using Ribodock. J. Comput. Aided Mol. Des. 2004 Mar;18(3):189–208.

21. Guilbert C, James TL. Docking to RNA via Root-Mean-Square-Deviation-Driven Energy Minimization with Flexible Ligands and Flexible Targets. J. Chem. Inf. Model. 2008 Jun;48(6):1257–68.

22. Foloppe N, Matassova N, Aboul-ela F. Towards the discovery of drug-like RNA ligands? Drug Discovery Today. 2006 Nov;11(21-22):1019–27.

23. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. J. Med. Chem. 2004 Mar;47(7):1739–49.

24. Halgren TA, Murphy RB, Friesner RA, Beard HS, Frye LL, Pollard WT, et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. J. Med. Chem. 2004 Mar;47(7):1750–9.

25. Friesner RA, Murphy RB, Repasky MP, Frye LL, Greenwood JR, Halgren TA, et al. Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein−Ligand Complexes. J. Med. Chem. 2006 Oct;49(21):6177–96.

26. Kramer B, Rarey M, Lengauer T. Evaluation of the FLEXX incremental construction algorithm for protein–ligand docking. Proteins. Wiley Online Library; 1999;37(2):228–41.

27. Miranker A, Karplus M. Functionality maps of binding sites: a multiple copy simultaneous search method. Proteins. Wiley Online Library; 1991;11(1):29–34.

28. Eisen MB, Wiley DC, Karplus M, Hubbard RE. HOOK: a program for finding novel molecular architectures that satisfy the chemical and steric requirements of a macromolecule binding site. Proteins. Wiley Online Library; 1994;19(3):199–221.

29. DeWitte RS, Shakhnovich EI. SMoG: de novo design method based on simple, fast, and accurate free energy estimates. 1. Methodology and supporting evidence. J. Am. Chem. Soc. ACS Publications; 1996;118(47):11733–44.

30. Roche O, Kiyama R, Brooks CL. Ligand−Protein DataBase: Linking Protein−Ligand Complex Structures to Binding Data. J. Med. Chem. 2001 Oct;44(22):3592–8.

31. Hu L, Benson ML, Smith RD, Lerner MG, Carlson HA. Binding MOAD (Mother Of All Databases). Proteins. 2005 Jun 21;60(3):333–40.

32. Smith RD, Dunbar JB Jr., Ung PM-U, Esposito EX, Yang C-Y, Wang S, et al. CSAR Benchmark Exercise of 2010: Combined Evaluation Across All Submitted Scoring Functions. J. Chem. Inf. Model. 2011 Sep 26;51(9):2115–31.

33. Damm-Ganamet KL, Smith RD, Dunbar JB Jr., Stuckey JA, Carlson HA. CSAR Benchmark Exercise 2011–2012: Evaluation of Results from Docking and Relative Ranking of Blinded Congeneric Series. J. Chem. Inf. Model. 2013 May 10;:130510104547004.

34. Dunbar JB Jr., Smith RD, Yang C-Y, Ung PM-U, Lexa KW, Khazanov NA, et al. CSAR Benchmark Exercise of 2010: Selection of the Protein–Ligand Complexes. J. Chem. Inf. Model. 2011 Sep 26;51(9):2036–46.

35. Majeux N, Scarsi M, Apostolakis J, Ehrhardt C, Caflisch A. Exhaustive docking of molecular fragments with electrostatic solvation. Proteins. 1999;37(1):88–105.

36. Zsoldos Z, Reid D, Simon A, Sadjad SB, Johnson AP. eHiTS: A new fast, exhaustive flexible ligand docking system. Journal of Molecular Graphics and Modelling. 2007 Jul;26(1):198–212. 4

37. Wu G, Robertson DH, Brooks CL, Vieth M. Detailed analysis of grid-based molecular docking: A case study of CDOCKER-A CHARMm-based MD docking algorithm. J. Comput. Chem. 2003 Oct;24(13):1549–62.

38. Fiser A, Do RKG, Sali A. Modeling of loops in protein structures. Protein science. Wiley Online Library; 2000;9(9):1753–73.

39. Jacobson MP, Pincus DL, Rapp CS, Day TJF, Honig B, Shaw DE, et al. A hierarchical approach to all-atom protein loop prediction. Proteins. 2004 Mar 5;55(2):351–67.

40. Yesselman JD, Price DJ, Knight JL, Brooks CL. MATCH: an atom-typing toolset for molecular mechanics force fields. J. Comput. Chem. 2012 Jan 15;33(2):189–202.

41. Knight JL, Yesselman JD, Brooks CL III. Assessing the quality of absolute hydration free energies among CHARMM-compatible ligand parameterization schemes. J. Comput. Chem. 2013 Jan 7;34(11):893–903.

42. Chen L, Calin GA, Zhang S. Novel Insights of Structure-Based Modeling for RNA-Targeted Drug Discovery. J. Chem. Inf. Model. 2012 Oct 22;52(10):274

CHAPTER 6

Conclusion and Future Directions

6.1 MATCH: An Atom-Typing Toolset for Molecular Mechanics Force Fields

Here we have constructed a framework called MATCH that can facilitate the automated parameterization of atom types, partial charges and molecular parameters for command molecular mechanics force fields. This toolset has the ability to deconstruct a force field into a set of fundamental rules which best replicates existing parameters and permits extension to new molecules yielding near instantaneous parameterization of novel molecules. It is customizable and extensible, such that it can act both as a solution for extrapolating and interpolating from the known chemical space to novel molecules and as a tool to study the effects of specific parameter choices and parameterization strategies. Through cross validation studies we have demonstrated that rules derived from one force field may be applied to other or to a set of novel molecules. Development of MATCH was critical in our ability to study RNA-ligand interactions as it allowed us to seamlessly parameterize ligands quickly and efficiently for molecular dynamics simulations and computational docking protocols.

Future work will focus on the development of automated procedure for generating molecular fragments of atom types. Automated procedures were investigated attempting to use genetic programs to evolve an ideal solution for each atom type simultaneously, but unfortunately they produced suboptimal solutions compared with the strategies that incorporated expert knowledge. In addition instead of substituting in known bond charge increments and atomic parameters when they are missing, it would be preferable to have some type of model built that would generate new parameters based on similarity in atom type. This was also attempted but failed to be significantly better then our substitution procedure already outlined. Implementation of automated procedures will allow the

MATCH strategy to be even more generalizable and facilitate the seamless integration of additional force field topology files into force-field specific MATCH libraries.

6.2 Assessing The Quality Of Absolute Hydration Free Energies Among CHARMM-Compatible Ligand Parameterization Schemes

This study highlights the importance of having sufficient coverage of chemical space within the underlying databases of these automated schemes and the benefit of targeting specific functional groups for parameterization efforts in order to maximize both the breadth and depth of the parameterized space. This work demonstrates that MATCH and other automated parameterization schemes generate atomic partial charges and parameters that are sufficient in quality. This was a critical step in determining whether automated parameterization was of high enough quality to be used in molecular dynamics simulations and computational docking techniques.

6.3 Using NMR Data to Generate Dynamic RNA Ensembles

RNA flexibility plays a critical role in its dynamics and interaction with proteins and small molecules. To account for its structural placidity we have employed the use of ^{1}H chemical shifts and molecular dynamics to generate dynamic structural ensembles. We first demonstrated that chemical shift SAS selections were valid using the set of 26 NMR structures which generated conformations similar to the NOE/RDC refined structures. It was critical to confirm that using only ^{1}H chemical shift data still generated conformations that were physically realistic compared to those generated using NOE/RDC. There were some subtle differences, which is related to the resolution of both ^{1}H chemical shift and the prediction method of NUCHEMICS. With larger data sets and the ability to predict carbon and nitrogen shifts, the future of RNA structure and dynamics predictions by chemical shifts looks very promising.

The comparison between the chemical shift SAS and the RDC SAS illustrates the utility of chemical shifts over RDCs in cases where there is little helical motion for determining the dynamics of the bulge. Additional work could be in combining both chemical shifts and RDC SAS in try and gain both helical information from the RDCs and local dynamical information from the chemical shifts. This is a proof of principle in generating

a dynamic ensemble using chemical shifts without a known ensemble to reference for RNA. The ensemble generated by the chemical shift SAS for the ribosomal decoding site was corroborated by both relative free energy calculations and the existing experimental structures which differs from previous work(1). It would also be interesting to investigate the structure agreement between the SAS ensemble and the bound conformations for A-site. In previous studies generating a dynamic ensemble for HIV-1 TAR it was shown that some of the conformations resembled bound conformations to different small molecules(2). This ensemble was also later shown to select out binders from a large small molecule database that were later shown to inhibit HIV-1 replication(3).

This work also investigated the effects of using different conformation pools for SAS selections. The pool generated by umbrella sampling produced an ensemble that differed from the one generated from the MD pool of conformations. It is quite possible that without proper sampling that skewed distribution could be incorrectly accepted. It is imperative that more research be conducted to further understand the significance of the different ensemble sizes and what they represent.

6.4 Enhanced Sampling Procedure Improves Success Rate for Nucleic Acid–Small Molecule Docking

RNA binding pockets are solvent exposed and shallow. In addition small molecules that bind RNA tend to be flexible with a large number of rotatable bonds(4,5). To address this issues we have developed an enhanced sampling procedure that uses a simplistic scoring function to generate high quality docked poses using our newly developed nucleic acid / small molecule set of 230 complexes. This procedure out performs ICM, a leading docking program, in a one to one comparison on this complex set. Furthermore, the computational cost is of running our program, which only needs to be done once, was faster then it took ICM to generate 10 poses. The average run time is approximately ~10 minutes per ligand once the receptor is prepared.

In addition, the success rate of 67% within 3 Å RMSD of the native pose using the top five poses produced for each complex is comparable to the success rates of the leading RNA docking programs benchmarked in smaller set of 56 complexes. Using this

sampling technique in conjunction with a second level of scoring could possibly produce even higher success rates as investigating the 10% most favorable scoring conformations contained poses that were below 2.5 Å 76% of the time. Our approach demonstrates that in addition to improving scoring functions, there is still additional way to improve success rates with novel enhanced sampling techniques.

6.5 References

1. Frank AT, Horowitz S, Andricioaei I, Al-Hashimi HM. Utility of 1H NMR Chemical Shifts in Determining RNA Structure and Dynamics. J. Phys. Chem. B. 2013 Feb 21;117(7):2045–52.

2. Frank AT, Stelzer AC, Al-Hashimi HM, Andricioaei I. Constructing RNA dynamical ensembles by combining MD and motionally decoupled NMR RDCs: new insights into RNA dynamics and adaptive ligand recognition. Nucleic Acids Research. 2009 Jun 21;37(11):3670–9.

3. Stelzer AC, Frank AT, Kratz JD, Swanson MD, Gonzalez-Hernandez MJ, Lee J, et al. Discovery of selective bioactive small molecules by targeting an RNA dynamic ensemble. Nature Publishing Group. 2011 Jun 26;7(8):553–9.

4. Thomas JR, Hergenrother PJ. Targeting RNA with Small Molecules. Chem. Rev. 2008 Apr;108(4):1171–224.

5. Vicens Q. RNA's coming of age as a drug target. J Incl Phenom Macrocycl Chem. 2009 Jun 27;65(1-2):171–88.

Appendix 1

A-site NMR Chemical Shift Spectra

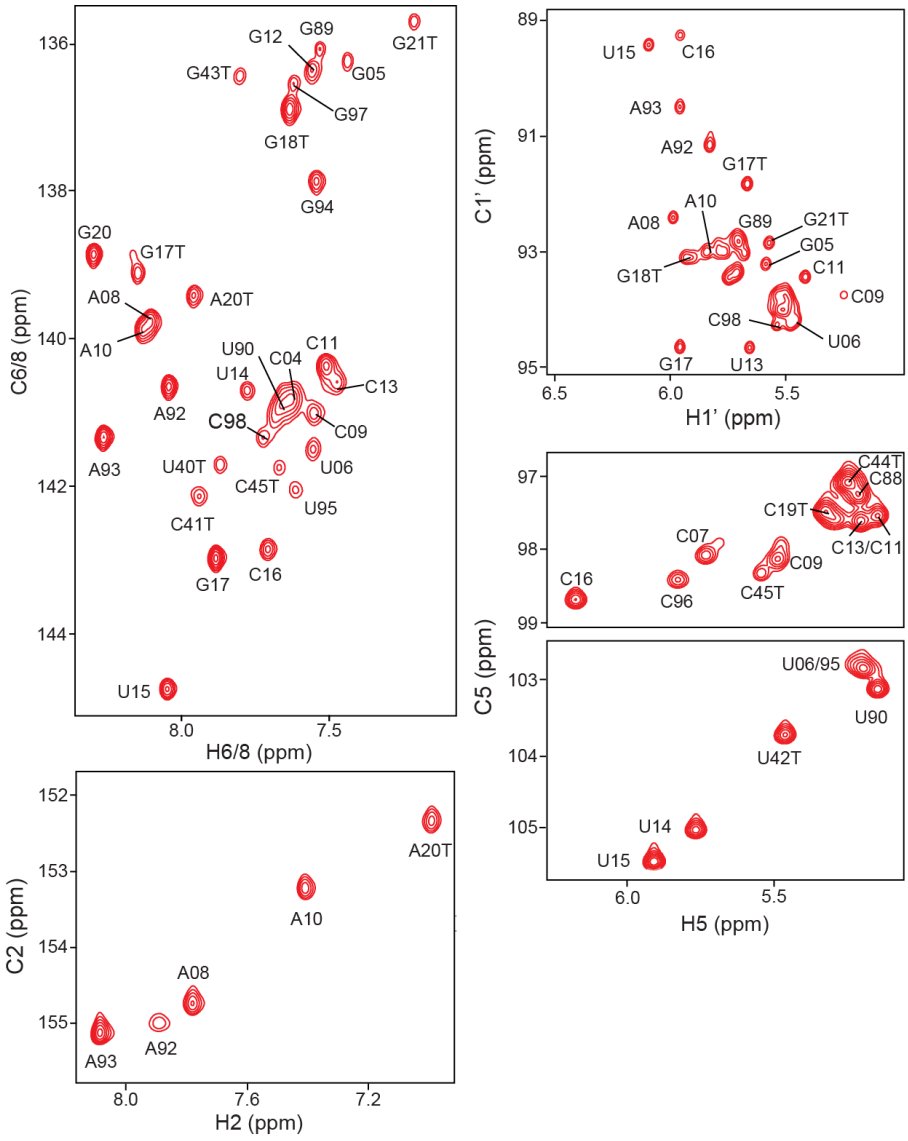

Figure A1.1: NMR resonance assignments of ribosomal decoding site construct (15 mM $NaPO_4$ pH 6.4, 25 mM NaCl, 0.1 mM EDTA, 298 K).

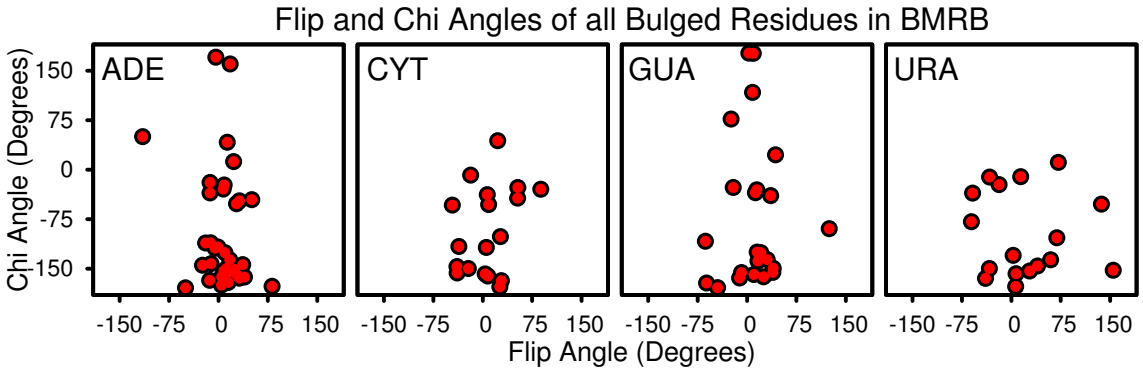

Figure A1.2: An overview of all the residues that in bulges with the BMRB broken up by residue type. It is clear that there is a much more broad distribution over flip angle for both Uridine and Cytosine

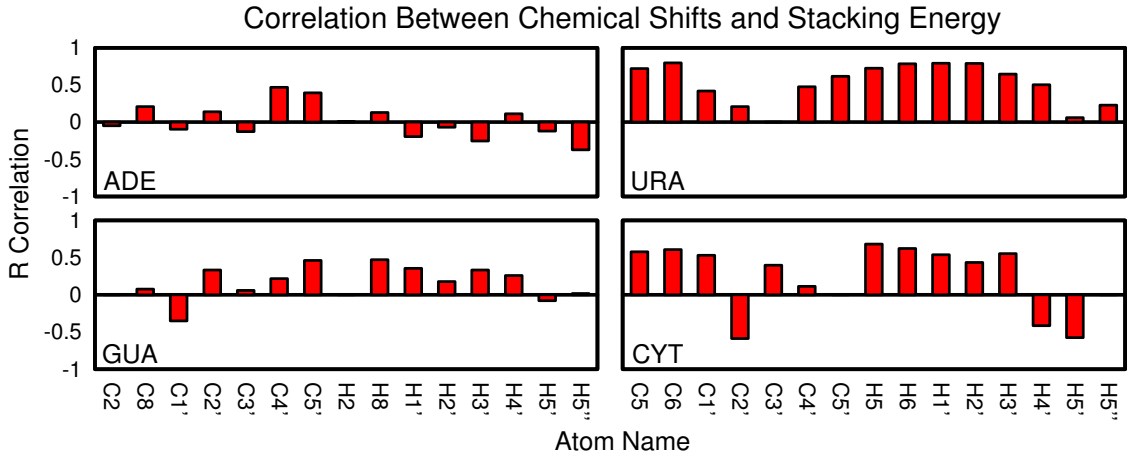

Figure A1.3: The R correlation of each nuclei per residue to calculated stacking energy. Only nuclei with 10 or more data points were considered. Uridine and Cytosine present the highest correlations, which may be due the more uniform sampling of flipped out states

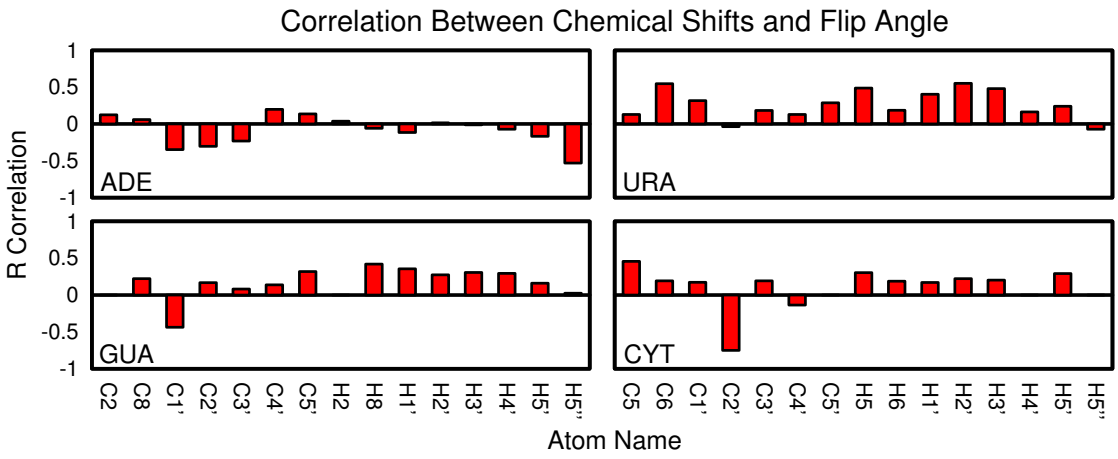

Figure A1.4: The R correlation of each nuclei per residue to calculated flip angle. Only nuclei with 10 or more data points were considered.

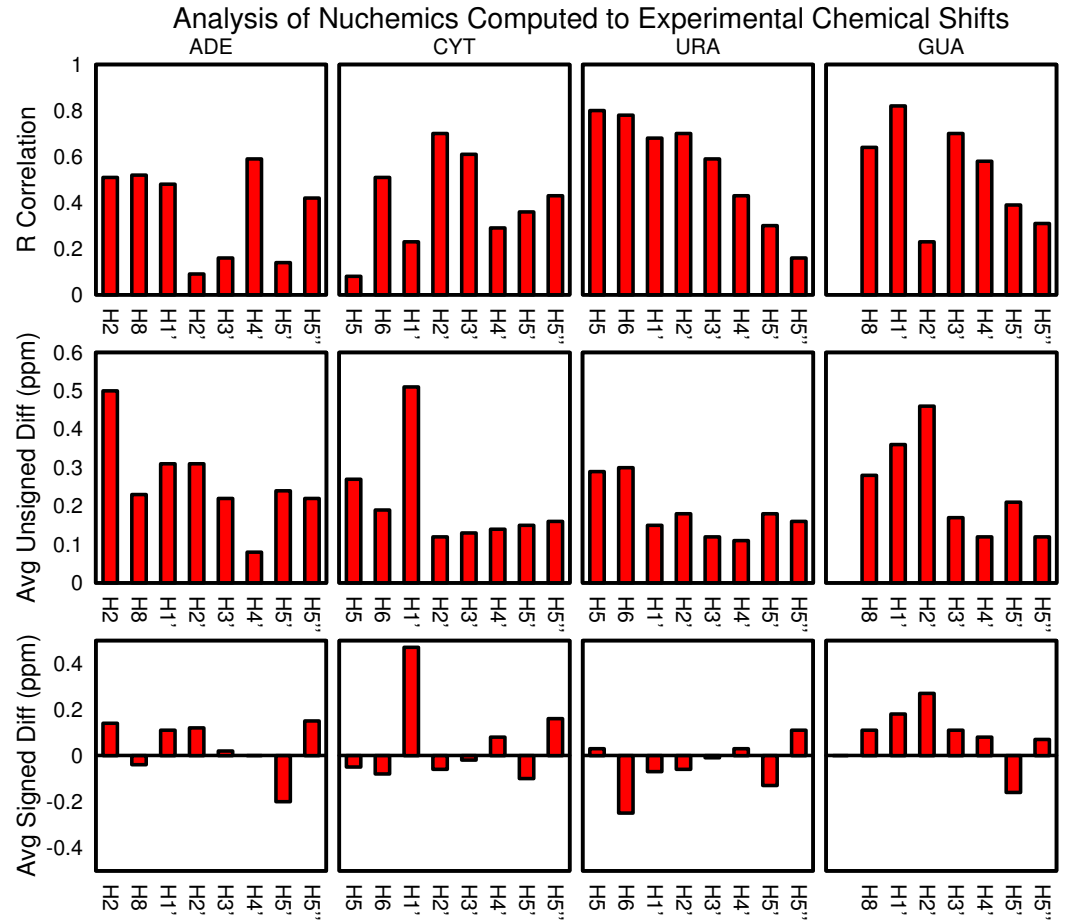

Figure A1.5: First panel depicts the R correlation between the experimental proton chemical shifts and those computed by NUCHEMICS. Second panel compares the average deviation between predicted and experimental by nuclei. Lastly, the third panel displays the average signed difference between the two to see if there are any noticeable trends in error, which none appear