The Dissertation Committee for Johnny Chung Wu Certifies that this is the approved version of the following dissertation:

Development of Accurate and Efficient Models for Biological Molecules

Committee:

Pengyu Ren, Supervisor

Kevin Dalby

Ron Elber

Robin Gutell

Mia Markey

Development of Accurate and Efficient Models for Biological Molecules

by

Johnny Chung Wu, B.A.; M.S.E.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December 2011

Dedication

To my parents, Tony and Jenny, and my brother, Ben.

And, of course, to Su.

Acknowledgements

I thank my advisor, Dr. Pengyu Ren, for his mentorship and for unselfishly sharing his knowledge and experiences with me. It has been an exciting journey to be a part of the growth of his research group. Despite the challenges you face, you do all that you can to absorb them so that your students can enjoy doing research without distractions and I appreciate that. I would like to thank my committee members. Dr. Kevin Dalby has been collaborating with our lab as the *de facto* medicinal chemist and his presence reminds me that models should be developed with a purpose and should be useful to somebody other than myself. Dr. Ron Elber is an accomplished scientist who I will continue to learn from after my graduate studies. It has been a pleasure to work with Dr. Robin Gutell during my initial years as a graduate student. I find inspiration from Dr. Mia Markey's discipline and drive.

I thank all of the members of the Ren lab for the help you have given me over the years. Dr. Michael Schnieders has always played a role in the lab, but officially joined recently and has made our work so exciting with the technology he develops, his infectious energy, and enthusiasm. Jiajing Zhang joined the lab the same semester as me and I have relied on her on many occasions for help. I am confident she will be spectacular at her dissertation defense. Yue Shi has been prolific in her contributions to the lab and I have benefitted greatly from them. I have worked closely with Zhen Xia and I thank him for the hard work he has put in to our project. I thank David Gardner from the Gutell lab for making our collaboration as smooth as possible. I thank the undergraduate researchers that I have worked with, Gaurav Chattree and Yujan Shrestha. I really have

been fortunate to have bright minds with fresh perspectives offer to work with me and make substantial contributions.

I thank my parents, brother, and my family for their support and love. Their care and concern for me extends beyond the years of my graduate studies and I am endlessly grateful for them. Lastly, I thank Su for reminding me what matters most. My most challenging moments are tremendously more bearable and my joyful moments are even more gratifying because you are here.

Development of Accurate and Efficient Models for Biological Molecules

Johnny Chung Wu, Ph.D.

The University of Texas at Austin, 2011

Supervisor: Pengyu Ren

The abnormal expression or function of biological molecules, such as nucleic acids, proteins, or other small organic molecules, lead to the majority of diseases. Consequently, understanding the structure and function of these molecules through modeling can provide insight and perhaps suggest treatment for diseases. However, biologically relevant molecular phenomenon can vary vastly in the nature of their interactions and different classes of models are required to accommodate for this diversity. The objective of this thesis is to develop models for small molecules, amino acid peptides, and nucleic acids. A physical polarizable molecular mechanics model is described to accurately represent small molecules and single atom ions and applied to predict experimentally measurable thermodynamic properties such as hydration and binding free energies. A novel physical coarse-grain model based on Gay-Berne potentials and electrostatic multipoles has been developed for short peptides. The fraction of residues that adopt the alpha-helix conformation agrees with all-atom molecule dynamics results.. Finally, a statistically-derived model based on sequence comparative sequence alignments is developed and applied to improve folding accuracy of RNA molecules.

Table of Contents

List of Tables

List of Figures

1 Introduction

Models of biological molecules are powerful tools that elucidate biological mechanisms at microscopic levels. Such modeling, which can be physics-based or knowledge-based, is used for a variety of studies from small molecules to biological macromolecules such as protein and nucleic acids [1, 2]. The scale of these models ranges from the accurate representation of electron distribution around a molecule to the coarse residue-level representation of polymers such as amino acids and nucleic acids. In this work, we focus on the development of biologically-relevant models at three scales.

The scale of highest granularity is the atomic scale in which a single atom is the smallest particle. Based on the AMOEBA polarizable force field, a model for the zinc divalent cation is developed based on comparison with quantum mechanics calculations. It is validated with experimental thermodynamic results such as hydration free energy. A protocol and automated tool to model small molecules with AMOEBA is also developed. Similarly, this protocol parameterizes small molecules based on quantum mechanics calculations and comparisons with experimental hydration and ligand-protein binding free energies are accurate to within 1 kcal/mol.

At the united-atom scale, groups of atoms are agglomerated in to a single particle. This method adopts the Gay-Berne potential to represent repulsion-dispersion interactions and higher order moments for electrostatic interactions. The dialanine peptide model has been developed and parameters are obtained by matching with all-atom physical interactions. Conformational energy of the coarse-grain peptide shows qualitative agreement with that of all-atom. Furthermore, the coarse-grain peptide has been polymerized and studied with replica exchange molecular dynamics. The per-

residue conformational distribution of alpha-helices falls within the range of various all-atom models.

Finally, at the scale of nucleic acids up to the ribosomal RNA, the statistical model is based on secondary structures obtained from comparative sequence analysis. The statistical potentials, developed from the statistics of secondary structures such as consecutive base-pair stacks, hairpin flanks, and internal loops, are applied to an RNA folding algorithm. Incorporation of statistical potentials markedly improved folding accuracy of tRNA, 5S rRNA, and 16S rRNA macromolecules.

1.1 ATOMIC REPRESENTATION OF SINGLE ATOM IONS AND SMALL MOLECULES

Small molecules are often modeled with classical molecular mechanics (MM) models. Force fields are composed of the energy functions used to describe atomic interactions and their corresponding parameters. Energy functions have explicit terms for interactions such as electrostatics, van der Waals, and bond lengths. Parameters are typically obtained by comparing with quantum mechanics (QM) calculations or experimental results [3, 4]. Examples of atomic force fields are GAFF[4], GROMOS[5], CHARMM[6], and AMOEBA[7]. All-atom force fields are models of the highest granularity that are capable of calculating various condensed-phase properties, such as density and acid dissociation constants, which can also be measured experimentally.

Although several varieties of atomic force fields have been developed, fixed-charge force fields are currently most widely adopted. Electrostatic interactions in the fixed-charge force field occur between sites of partial charges. Electrostatic sites are associated with atom or bonds. Although electron densities are redistributed as the environment changes, charges remain static in these types of force fields. However, we have realized as early as 1967 through Buckingham that electrostatic intermolecular

forces can be accurately represented with higher order moments and induction [8]. Despite investigations of various polarizable force fields [9-15], few large-scale macromolecular simulations have taken advantage of these models. Furthermore, the need for a polarizable force field has been widely acknowledged [16, 17].

1.1.1 AMOEBA polarizable force field

The AMOEBA (Atomic Multipole Optimized Energetics for Biomolecular Applications) force field is an effort to achieve chemical accuracy of conformational and interaction energies to quantum mechanical models[7]. Moreover, AMOEBA addresses differences in systems where assuming an averaged polarization is not adequate. Its permanent point multipole up to quadrupole is capable of describing intricate electrostatic potential surfaces such as those created by electron lone pairs. Polarization is represented with polarizable point dipoles that can fully describe the directionality of electron redistribution without the need for fictitious particles.

The representation of single atom ions poses an interesting problem for MM models. Single ionic atoms such as potassium, chlorine, calcium, magnesium, zinc, and iron are known to participate in various biological functions. Zinc, in particular, has been identified to be a critical component of specific enzymatic processes[18]. The Zn^{2+} divalent cation may act directly as a structural element in proteins such as Zn-fingers[19] or as a cofactor to facilitate the activity of metalloenzymes, such as carbonic anhydrase[20].

Quantum mechanics (QM) is usually the primary methodology for studying Zn^{2+}-metalloproteins[21-23] due to the ion's soft character and subtle nature of its interactions in the biological environment[24]. As a result of the high computational demands of accurate *ab initio* methods, studies are limited to "static" structures of relatively small

biomimetic models. Hybrid methods that combine QM and molecular mechanics models (QM/MM) [25-28] offer the possibility to treat the whole protein on longer time scales. However, a robust molecular mechanics model continues to be in need to study the dynamical behavior of Zn^{2+} complexes. Traditional fixed-charge force fields are unable to capture the interactions between Zn^{2+} and its ligands. To maintain structural stability of Zn^{2+}-protein complexes with these force fields, the introduction of artificial bonds[29] or extra charge sites[30] are required.

Moreover, studies using quasi-chemical theory [31, 32] have demonstrated the importance of polarization to ion hydration. The AMOEBA (Atomic Multipole Optimized Energetics for Biomolecular Applications) force field provides a computationally efficient potential energy functions relative to QM calculations and is able to capture the polarization effects that fix charged force fields lack. Accordingly, AMEOBA has enjoyed a series of successes with modeling monovalent ions such as K^+, Na^+, and Cl^- [33], as well as divalent ions such as Ca^{2+} and Mg^{2+} [34]. Additional studies of ion hydration with mixed quantum mechanics/molecular mechanics (QM/MM) calculations have concluded that AMOEBA is sufficient in generating ensemble configurations for post-analysis, but that fixed charge force fields have difficulties with mimicking local charge rearrangements [35]. Similar to the development of previously derived ions for AMOEBA, obtaining parameters for the zinc divalent cation involves fitting to Zn^{2+}-water dimer gas-phase *ab initio* calculations. The Thole-based [36] dipole polarization is determined by comparison with theoretically calculated interactions, such as the Constrained Space Orbital Variations (CSOV) [37]. The dimerization energy of Zn^{2+}-water with results from several theory levels and basis sets over a range of distances are able to provide the comparisons required to obtain parameters for van der Waals interactions. Subsequently, the Zn^{2+} model can be substantiated with molecular dynamics

simulations of Zn^{2+} solvation in bulk water. Calculations of the first-shell water coordination number, water residence time and free energy of hydration and establishing agreement with experimental and theoretical values can further support the model.

Additionally, studies identify that fixed charge force fields have difficulties with calculating the solvation free energy of polar small molecules, particular those containing hydroxyl groups[38], which are common in carbohydrates. Patel *et al*.[39] studied polarization in further detail and proposed the TIP4P-QDP charge-dependent polarizability water model. This work revealed the effects of polarization variability on the enhanced structure at liquid-vapor interface. The importance of polarization in protein-ligand recognition [40-42] has been identified. Although other classical force fields have been extended to include polarization interactions, such as PIPF-CHARMM[43] , most polarizable force fields lack higher order electronic moments.

1.2 COARSE-GRAIN MODEL OF SHORT PEPTIDE

The ambition to understand molecular systems of increasing length and time scales drives the pursuit and development of coarse grain computational models. It continues to be prohibitively expensive for all-atom molecular mechanics models to collect statistically converged measurements of molecular phenomena that involve large conformational rearrangements, such as protein folding, protein-protein interaction, and allosteric regulation [44]. Although there has been much development in the areas of enhanced sampling, the need to study the dynamics of large biomolecular systems over long time scales remains. Consequently, various coarse-graining strategies have been endeavored to model the systems of interest. Much effort has been made to develop coarse-grained models by matching the intermolecular interaction energy and force at the functional group or molecular level with all-atom simulations of specific systems. Klein

and co-workers reported coarse-grained models of membrane lipids and proposed various coarse-graining strategies based on previous studies of polymer melts [45, 46]. DeVane and coworkers have recently embarked on a method that employs the Lennard-Jones 9-6 and 12-4 forms to model non-bonded interactions of coarse-grain sites and have thus far validated the model on various amino acid side-chain analogs [47]. Hills et al. has demonstrated that a physics-based, isotropic site, solvent-free method is able to maintain the native structures of Trpzip, Trp-cage, and the open/close conformations of adenylate kinase [48]. Moreover, the united-residue force field developed by Scheraga et al. has matured significantly and used to study the folding mechanism of specific domains of the staphylococcal protein A and the formin-binding protein [49-56].

In this work, a general and transferable coarse-grain (CG) framework based on the Gay-Berne potential and electrostatic point multipole expansion is presented for polypeptide simulations. The solvent effect is described by the Generalized Kirkwood theory. The CG model is calibrated using the results of all-atom simulations of model compounds in solution. Instead of matching the overall effective forces produced by atomic models, the fundamental intermolecular forces such as electrostatic, repulsion-dispersion and solvation are represented explicitly at a CG level. We demonstrate that the CG alanine dipeptide model is able to reproduce quantitatively the conformational energy of all-atom force fields in both gas and solution phases, including the electrostatic and solvation components. Replica exchange molecular dynamics (REMD) and microsecond dynamic simulations of polyalanine of 5 and 12 residues reveal that the CG polyalanines fold into “alpha helix” and “beta sheet” structures. The 5-residue polyalanine display a substantial increase in the “beta strand” fraction relative to the 12-residue polyalanine. The detailed conformational distribution is compared with those reported from recent all-atom simulations and experiments. The results suggest that the new coarse-graining

approach presented in this study has the potential to offer both accuracy and efficiency for biomolecular modeling.

1.3 STATISTICALLY-DERIVED ENERGY FUNCTIONS FOR RNA FOLDING

The accurate prediction of the secondary and tertiary structure of RNA with different folding algorithms is dependent on several factors, including the energy functions. However, an RNA higher-order structure cannot be accurately predicted from its sequence based on a limited set of energy parameters. The inter- and intra-molecular forces between this RNA and other small molecules and macromolecules, in addition to other factors in the cell such as pH, ionic strength, and temperature influence the complex dynamics associated with a single stranded RNA's transitioning to its secondary and tertiary structure. Since all of the factors that affect the formation of an RNAs three-dimensional structure cannot be determined experimentally, statistically-derived potential energy has been used in the prediction of protein structure. In the current work, we evaluate the statistical free energy of various secondary structure motifs, including base-pair stacks, hairpin flanks, and internal loops, using their statistical frequencies obtained from the comparative analysis of more than 50,000 RNA sequences stored in the RNA Comparative Analysis Database (rCAD). Statistical energies were computed from the structural statistics for several datasets. While the statistical energies for base-pair stacks correlate with experimentally derived free energy values, suggesting a Boltzmann-like distribution, variation is observed between different molecules and their location on the phylogenetic tree of life. Our statistical energies for several structural elements were utilized in the Mfold RNA folding algorithm. The combined statistical energies for base-pair stacks, hairpins and internal loop flanks results in a significant improvement in the

accuracy of secondary structure prediction; however, the hairpin flanks contribute the most.

The stabilization of G:C, A:U, and G:U base pairs by hydrogen bonding and base stacking has been identified since these canonical base-pairs were first observed to form regular helices when arranged in an anti-parallel manner. With experimental calorimetric measurements of simple base-paring oligonucleotides, the majority of the known free energy values were determined for consecutive base-pairs by Turner and his collaborators[57]. The relative contribution to the overall stability of base stacking and the hydrogen bonding, however, is not well understood [58, 59]. While the full extent of the types of RNA structural elements and helices have not been identified and characterized, the energetic contribution for only a small percentage of the characterized RNA structural elements have been determined. It is known that approximately 66% of the nucleotides in larger RNAs like the 16S and 23S rRNA form regular secondary structure helix[60]. The remaining third of the nucleotides are involved in more complex secondary and tertiary structures[61]. A partial list of these includes: U-turns[62], lone pair tri loops[63], a very high percentage of unpaired A's in the secondary structure[64] that are involved in several motifs, including - A minor motifs[65, 66], E and E-like motifs[64], UAA/GAN internal loop motif[67], GNRA tetraloops[67, 68], and a high percentage of A:A and A:G juxtapositions at the ends of regular helices[69]. The majority of the base pairs in these structural motifs form unusual non-canonical base pair types and their base pair conformations [70-72].

With an incomplete knowledge of all possible RNA structural motifs and their energetic stabilities in different structural environments, an alternative approach to the simplified energy model dominated by base-pair stacks in regular secondary structure helices is needed. While experimental approaches have been essential to our

understanding of macromolecular structure and their energetic stabilities, it is not feasible to determine the energetic stabilities for all possible structural motifs. In contrast, an analysis of high resolution crystal structures in parallel with the statistical analysis of different sets of comparative macromolecular sequences that form identical or very similar secondary and tertiary structure has been utilized to determine knowledge-based potentials or scoring functions. This approach has been used frequently in protein structure prediction[73, 74] following the work by Scheraga[75]. An important assumption involved in the conversion of structural statistics into (pseudo) free energy is the Boltzmann-like distribution of the structure, which has been substantiated for proteins[76, 77].

Due to the emerging technologies that rapidly determine nucleic acid sequences for entire genomes, we are obtaining nucleotide sequences for an increasing number of RNA families with significant increases in the number of sequences per family. To facilitate the analysis of these increasing comparative RNA datasets and to enhance the analysis of these datasets, we have collaborated with Microsoft Research to develop an RNA Comparative Analysis Database (rCAD) that cross-indexes sequence, structure, and phylogenetic information (Ozer, Doshi, Cannone, Xu, and Gutell, manuscript in preparation). In this work, we have utilized rCAD to obtain structural statistics from the comparative analysis of these sequences, and derived statistical energies that can be used for RNA structure prediction. We demonstrate that structural motifs beyond base-pair stacking in helices are also important in determining the RNA structure. The statistical energies derived from sequence information have the potential to significantly improve the accuracy of RNA secondary structure prediction. The results of this work motivate further investigation of statistical potentials of a broader range of motifs, including a

diversity of scale. The statistical energy approach has been applied to physical models as well [78].

1.4 REFERENCES

1. Gutell, R.R.; Noller, H.F.; Woese, C.R., *Higher order structure in ribosomal RNA*. EMBO J., 1986. **5**(5): p. 1111-3.
2. Jonikas, M.A.; Radmer, R.J.; Laederach, A.; Das, R.; Pearlman, S.; Herschlag, D.; Altman, R.B., *Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters.* RNA, 2009. **15**(2): p. 189-99.
3. Ren, P.Y.; Ponder, J.W., *Polarizable atomic multipole water model for molecular mechanics simulation.* J. Phys. Chem. B, 2003. **107**(24): p. 5933-5947.
4. Wang, J.M.; Wolf, R.M.; Caldwell, J.W.; Kollman, P.A.; Case, D.A., *Development and testing of a general amber force field.* J. Comput. Chem., 2004. **25**(9): p. 1157-1174.
5. Oostenbrink, C.; Villa, A.; Mark, A.E.; van Gunsteren, W.F., *A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6.* J. Comput. Chem., 2004. **25**(13): p. 1656-76.
6. MacKerell, A.D.; Brooks, B.; Brooks, C.L.; Nilsson, L.; Roux, B.; Won, Y.; Karplus, M., *CHARMM: The Energy Function and Its Parameterization*, in *Encyclopedia of Computational Chemistry*. 2002, John Wiley & Sons, Ltd.
7. Ren, P.; Ponder, J.W., *Consistent treatment of inter- and intramolecular polarization in molecular mechanics calculations.* J. Comput. Chem., 2002. **23**(16): p. 1497-506.
8. Buckingham, A.D., *Permanent and Induced Molecular Moments and Long-Range Intermolecular Forces*. Adv. Chem. Phys. 1967: John Wiley & Sons, Inc. 107-142.
9. Jensen, L.; Astrand, P.O.; Osted, A.; Kongsted, J.; Mikkelsen, K.V., *Polarizability of molecular clusters as calculated by a dipole interaction model.* J. Chem. Phys., 2002. **116**(10): p. 4001-4010.
10. Gresh, N.; Cisneros, G.A.; Darden, T.A.; Piquemal, J.P., *Anisotropic, polarizable molecular mechanics studies of inter- and intramoecular interactions and ligand-macromolecule complexes. A bottom-up strategy.* J. Chem. Theory Comput., 2007. **3**(6): p. 1960-1986.
11. Piquemal, J.P.; Williams-Hubbard, B.; Fey, N.; Deeth, R.J.; Gresh, N.; Giessner-Prettre, C., *Inclusion of the ligand field contribution in a polarizable molecular mechanics: SIBFA-LF*. J. Comput. Chem., 2003. **24**(16): p. 1963-1970.
12. Rick, S.W.; Stuart, S.J.; Berne, B.J., *Dynamical Fluctuating Charge Force-Fields - Application to Liquid Water.* J. Chem. Phys., 1994. **101**(7): p. 6141-6156.

13. Xie, W.; Orozco, M.; Truhlar, D.G.; Gao, J., *X-Pol Potential: An Electronic Structure-Based Force Field for Molecular Dynamics Simulation of a Solvated Protein in Water.* J. Chem. Theory Comput., 2009. **5**(3): p. 459-467.
14. Senn, H.M.; Thiel, W., *QM/MM Methods for Biomolecular Systems.* Angewandte Chemie-International Edition, 2009. **48**(7): p. 1198-1229.
15. Jakalian, A.; Jack, D.B.; Bayly, C.I., *Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation.* J. Comput. Chem., 2002. **23**(16): p. 1623-1641.
16. Cieplak, P.; Dupradeau, F.Y.; Duan, Y.; Wang, J.M., *Polarization effects in molecular mechanical force fields.* Journal of Physics-Condensed Matter, 2009. **21**(33): p. 333102.
17. Lopes, P.E.M.; Roux, B.; MacKerell, A.D., *Molecular modeling and dynamics studies with explicit inclusion of electronic polarizability: theory and applications.* Theor. Chem. Acc., 2009. **124**(1-2): p. 11-28.
18. Keilin, D.; Mann, T., *Carbonic anhydrase. Purification and nature of the enzyme.* Biochem. J., 1940. **34**(8-9): p. 1163-1176.
19. Maynard, A.T.; Covell, D.G., *Reactivity of zinc finger cores: Analysis of protein packing and electrostatic screening.* J. Am. Chem. Soc., 2001. **123**(6): p. 1047-1058.
20. Lipscomb, W.N.; Strater, N., *Recent advances in zinc enzymology.* Chem. Rev. (Washington, DC, U. S.), 1996. **96**(7): p. 2375-2433.
21. Gresh, N.; Garmer, D.R., *Comparative binding energetics of Mg2+, Ca2+, Zn2+, and Cd2+ to biologically relevant ligands: Combined ab initio SCF supermolecule and molecular mechanics investigation.* J. Comput. Chem., 1996. **17**(12): p. 1481-1495.
22. Rayon, V.M.; Valdes, H.; Diaz, N.; Suarez, D., *Monoligand Zn(II) complexes: Ab initio benchmark calculations and comparison with density functional theory methodologies.* J. Chem. Theory Comput., 2008. **4**(2): p. 243-256.
23. Amin, E.A.; Truhlar, D.G., *Zn coordination chemistry: Development of benchmark suites for geometries, dipole moments, and bond dissociation energies and their use to test and validate density functionals and molecular orbital theory.* J. Chem. Theory Comput., 2008. **4**(1): p. 75-85.
24. De Courcy, B.; Gresh, N.; Piquemal, J.P., *Importance of lone pair interactions/redistribution in hard and soft ligands within the active site of alcohol dehydrogenase Zn-metalloenzyme: Insights from electron localization function.* Interdisciplinary Sciences: Computational Life Sciences, 2009. **1**(1): p. 55-60.
25. Warshel, A.; Levitt, M., *Theoretical studies of enzymic reactions - dielectric, electrostatic and steric stabilization of carbonium-ion in reaction of lysozyme.* J. Mol. Biol., 1976. **103**(2): p. 227-249.
26. Estiu, G.; Suarez, D.; Merz, K.M., *Quantum mechanical and molecular dynamics simulations of ureases and Zn beta-lactamases.* J. Comput. Chem., 2006. **27**(12): p. 1240-1262.

27. Friesner, R.A.; Guallar, V., *Ab initio quantum chemical and mixed quantum mechanics/molecular mechanics (QM/MM) methods for studying enzymatic catalysis.* Annu. Rev. Phys. Chem., 2005. **56**: p. 389-427.
28. Ryde, U., *Combined quantum and molecular mechanics calculations on metalloproteins.* Curr. Opin. Chem. Biol., 2003. **7**(1): p. 136-142.
29. Tuccinardi, T.; Martinelli, A.; Nuti, E.; Carelli, P.; Balzano, F.; Uccello-Barretta, G.; Murphy, G.; Rossello, A., *Amber force field implementation, molecular modelling study, synthesis and MMP-1/MMP-2 inhibition profile of (R)and (S)-N-hydroxy-2-(N-isopropoxybiphenyl-4-ylsulfonamido)-3-methylbutanamides.* Bioorg. Med. Chem., 2006. **14**(12): p. 4260-4276.
30. Yuan-Ping, P., *Successful molecular dynamics simulation of two zinc complexes bridged by a hydroxide in phosphotriesterase using the cationic dummy atom method.* Proteins: Struct., Funct., Genet., 2001. **45**(3): p. 183-189.
31. Asthagiri, D.; Pratt, L.R.; Paulaitis, M.E.; Rempe, S.B., *Hydration structure and free energy of biomolecularly specific aqueous dications, including Zn2+ and first transition row metals.* J. Am. Chem. Soc., 2004. **126**(4): p. 1285-1289.
32. Rogers, D.M.; Beck, T.L., *Quasichemical and structural analysis of polarizable anion hydration.* J. Chem. Phys., 2010. **132**(1): p. 014505.
33. Grossfield, A.; Ren, P.Y.; Ponder, J.W., *Ion solvation thermodynamics from simulation with a polarizable force field.* J. Am. Chem. Soc., 2003. **125**(50): p. 15671-15682.
34. Jiao, D.; King, C.; Grossfield, A.; Darden, T.A.; Ren, P.Y., *Simulation of Ca2+ and Mg2+ solvation using polarizable atomic multipole potential.* J. Phys. Chem. B, 2006. **110**(37): p. 18553-18559.
35. Beck, T., *Hydration Free Energies by Energetic Partitioning of the Potential Distribution Theorem.* Journal of Statistical Physics, 2011: p. 1-20.
36. Thole, B.T., *Molecular Polarizabilities Calculated with a Modified Dipole Interaction.* Chem. Phys., 1981. **59**(3): p. 341-350.
37. Bagus, P.S.; Illas, F., *Decomposition of the chemisorption bond by constrained variations - Order of the variations and construction of the variational spaces.* J. Chem. Phys., 1992. **96**(12): p. 8962-8970.
38. Klimovich, P.V.; Mobley, D.L., *Predicting hydration free energies using all-atom molecular dynamics simulations and multiple starting conformations.* J. Comput-Aided. Mol. Des., 2010. **24**(4): p. 307-316.
39. Bauer, B.A.; Warren, G.L.; Patel, S., *Incorporating Phase-Dependent Polarizability in Nonadditive Electrostatic Models for Molecular Dynamics Simulations of the Aqueous Liquid-Vapor Interface.* J. Chem. Theory Comput., 2009. **5**(2): p. 359-373.
40. Jiao, D.; Golubkov, P.A.; Darden, T.A.; Ren, P., *Calculation of protein-ligand binding free energy by using a polarizable potential.* Proc. Natl. Acad. Sci. U. S. A., 2008. **105**(17): p. 6290-6295.

41. Jiao, D.; Zhang, J.; Duke, R.E.; Li, G.; Schnieders, M.J.; Ren, P., *Trypsin-ligand binding free energies from explicit and implicit solvent simulations with polarizable potential.* J. Comput. Chem., 2009. **30**(11): p. 1701-11.
42. Shi, Y.; Jiao, D.A.; Schnieders, M.J.; Ren, P.Y., *Trypsin-Ligand Binding Free Energy Calculation with AMOEBA.* Embc: 2009 Annual International Conference of the Ieee Engineering in Medicine and Biology Society, Vols 1-20, 2009: p. 2328-2331.
43. Xie, W.S.; Pu, J.Z.; MacKerell, A.D.; Gao, J.L., *Development of a polarizable intermolecular potential function (PIPF) for liquid amides and alkanes.* J. Chem. Theory Comput., 2007. **3**(6): p. 1878-1889.
44. Kern, D.; Zuiderweg, E.R.P., *The role of dynamics in allosteric regulation.* Curr. Opin. Struct. Biol., 2003. **13**(6): p. 748-757.
45. Shelley, J.C.; Shelley, M.Y.; Reeder, R.C.; Bandyopadhyay, S.; Klein, M.L., *A coarse grain model for phospholipid simulations.* J. Phys. Chem. B, 2001. **105**(19): p. 4464-4470.
46. Nielsen, S.O.; Lopez, C.F.; Srinivas, G.; Klein, M.L., *Coarse grain models and the computer simulation of soft materials.* Journal Of Physics-condensed Matter, 2004. **16**(15): p. R481-R512.
47. DeVane, R.; Klein, M.L.; Chiu, C.C.; Nielsen, S.O.; Shinoda, W.; Moore, P.B., *Coarse-Grained Potential Models for Phenyl-Based Molecules: I. Parametrization Using Experimental Data.* J. Phys. Chem. B, 2010. **114**(19): p. 6386-6393.
48. Hills, R.D.; Lu, L.Y.; Voth, G.A., *Multiscale Coarse-Graining of the Protein Energy Landscape.* PLoS Comput. Biol., 2010. **6**(6).
49. Makowski, M.; Sobolewski, E.; Czaplewski, C.; Oldziej, S.; Liwo, A.; Scheraga, H.A., *Simple physics-based analytical formulas for the potentials of mean force for the interaction of amino acid side chains in water. IV. Pairs of different hydrophobic side chains.* J. Phys. Chem. B, 2008. **112**(36): p. 11385-11395.
50. Maisuradze, G.G.; Liwo, A.; Oldziej, S.; Scheraga, H.A., *Evidence, from simulations, of a single state with residual native structure at the thermal denaturation midpoint of a small globular protein.* J. Am. Chem. Soc., 2010. **132**(27): p. 9444-52.
51. Liwo, A.; Czaplewski, C.; Pillardy, J.; Scheraga, H.A., *Cumulant-based expressions for the multibody terms for the correlation between local and electrostatic interactions in the united-residue force field.* J. Chem. Phys., 2001. **115**(5): p. 2323-2347.
52. Wu, C.; Shea, J.E., *Coarse-grained models for protein aggregation.* Curr. Opin. Struct. Biol., 2005. **21**(2): p. 209-220.
53. Liwo, A.; Khalili, M.; Czaplewski, C.; Kalinowski, S.; Oldziej, S.; Wachucik, K.; Scheraga, H.A., *Modification and optimization of the united-residue (UNRES) potential energy function for canonical simulations. I. Temperature dependence of the effective energy function and tests of the optimization method with single training proteins.* J. Phys. Chem. B, 2007. **111**(1): p. 260-285.

54. Liwo, A.; Czaplewski, C.; Oldziej, S.; Rojas, A.V.; Kazmierkiewicz, R.; Makowski, M.; Murarka, R.K.; Scheraga, H.A., *Simulation of protein structure and dynamics with the coarse-grained UNRES force field* in *Coarse-Graining of Condensed Phase and Biomolecular Systems*, G. Voth, Editor. 2008, CRC Press, Taylor & Francis Group: Farmington, CT. p. 107-122.
55. C. Czaplewski; A. Liwo; M. Makowski; S. Oldziej; H.A. Scheraga, *Coarse-Grained Models of Proteins: Theory and Applications* in *Multiscale Approaches to Protein Modeling*, A. Kolinski, Editor. 2010, Springer.
56. Voth, G., ed. *Coarse-Graining of Condensed Phase and Biomolecular Systems*. 2008, CRC Press, Taylor & Francis Group: Farmington, CT.
57. Zuker, M.; Jaeger, J.A.; Turner, D.H., *A comparison of optimal and suboptimal RNA secondary structures predicted by free energy minimization with structures determined by phylogenetic comparison.* Nucleic Acids Res., 1991. **19**(10): p. 2707-14.
58. Yakovchuk, P.; Protozanova, E.; Frank-Kamenetskii, M.D., *Base-stacking and base-pairing contributions into thermal stability of the DNA double helix.* Nucleic Acids Res., 2006. **34**(2): p. 564-574.
59. Guckian, K.M.; Schweitzer, B.A.; Ren, R.X.F.; Sheils, C.J.; Tahmassebi, D.C.; Kool, E.T., *Factors contributing to aromatic stacking in water: Evaluation in the context of DNA.* J. Am. Chem. Soc., 2000. **122**(10): p. 2213-2222.
60. Gutell, R.R.; Weiser, B.; Woese, C.R.; Noller, H.F., *Comparative Anatomy of 16-S-Like Ribosomal-RNA.* Prog. Nucleic Acid Res. Mol. Biol., 1985. **32**: p. 155-216.
61. Gutell, R.R.; Lee, J.C.; Cannone, J.J., *The accuracy of ribosomal RNA comparative structure models.* Curr. Opin. Struct. Biol., 2002. **12**(3): p. 301-310.
62. Gutell, R.R.; Cannone, J.J.; Konings, D.; Gautheret, D., *Predicting U-turns in ribosomal RNA with comparative sequence analysis.* J. Mol. Biol., 2000. **300**(4): p. 791-803.
63. Lee, J.C.; Cannone, J.J.; Gutell, R.R., *The lonepair triloop: A new motif in RNA structure.* J. Mol. Biol., 2003. **325**(1): p. 65-83.
64. Gutell, R.R.; Cannone, J.J.; Shang, Z.; Du, Y.; Serra, M.J., *A story: Unpaired adenosine bases in ribosomal RNAs.* J. Mol. Biol., 2000. **304**(3): p. 335-354.
65. Nissen, P.; Ippolito, J.A.; Ban, N.; Moore, P.B.; Steitz, T.A., *RNA tertiary interactions in the large ribosomal subunit: The A-minor motif.* Proc. Natl. Acad. Sci. U. S. A., 2001. **98**(9): p. 4899-4903.
66. Battle, D.J.; Doudna, J.A., *Specificity of RNA-RNA helix recognition.* Proc. Natl. Acad. Sci. U. S. A., 2002. **99**(18): p. 11676-11681.
67. Woese, C.R.; Winker, S.; Gutell, R.R., *Architecture of Ribosomal-Rna - Constraints on the Sequence of Tetra-Loops.* Proc. Natl. Acad. Sci. U. S. A., 1990. **87**(21): p. 8467-8471.
68. Michel, F.; Westhof, E., *Modeling of the 3-Dimensional Architecture of Group-I Catalytic Introns Based on Comparative Sequence-Analysis.* J. Mol. Biol., 1990. **216**(3): p. 585-610.

69. Elgavish, T.; Cannone, J.J.; Lee, J.C.; Harvey, S.C.; Gutell, R.R., *AA.AG@Helix.Ends: A : A and A : G base-pairs at the ends of 16 S and 23 S rRNA helices.* J. Mol. Biol., 2001. **310**(4): p. 735-753.
70. Lee, J.C.; Gutell, R.R., *Diversity of base-pair conformations and their occurrence in rRNA structure and RNA structural motifs.* J. Mol. Biol., 2004. **344**(5): p. 1225-1249.
71. Leontis, N.B.; Stombaugh, J.; Westhof, E., *The non-Watson-Crick base pairs and their associated isostericity matrices.* Nucleic Acids Res., 2002. **30**(16): p. 3497-3531.
72. Xin, Y.R.; Olson, W.K., *BPS: a database of RNA base-pair structures.* Nucleic Acids Res., 2009. **37**: p. D83-D88.
73. Floudas, C.A.; Fung, H.K.; McAllister, S.R.; Monnigmann, M.; Rajgaria, R., *Advances in protein structure prediction and de novo protein design: A review.* Chem. Eng. Sci., 2006. **61**(3): p. 966-988.
74. Shen, M.Y.; Sali, A., *Statistical potential for assessment and prediction of protein structures.* Protein Sci., 2006. **15**(11): p. 2507-2524.
75. Tanaka, S.; Scheraga, H.A., *Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins.* Macromolecules, 1976. **9**(6): p. 945-50.
76. Bryant, S.H.; Lawrence, C.E., *The Frequency Of Ion-Pair Substructures In Proteins Is Quantitatively Related To Electrostatic Potential - A Statistical-Model For Nonbonded Interactions.* Proteins-Structure Function and Genetics, 1991. **9**(2): p. 108-119.
77. Finkelstein, A.V.; Badretdinov, A.Y.; Gutin, A.M., *Why Do Protein Architectures Have Boltzmann-Like Statistics.* Proteins-Structure Function and Genetics, 1995. **23**(2): p. 142-150.
78. Xia, Z.; Gardner, D.P.; Gutell, R.R.; Ren, P., *Coarse-grained model for simulation of RNA three-dimensional structures.* J. Phys. Chem. B, 2010. **114**(42): p. 13497-506.

2 Modeling Zn(II) in Water with the AMOEBA Polarizable Force Field

2.1 INTRODUCTION

Understanding the role of the zinc divalent cation in enzymatic processes[1, 2] and structural motifs such as zinc-fingers[3] motivates an accurate molecular mechanics model for the ion. Furthermore, AMOEBA has been shown to be particularly suited for the computation of dynamical properties of metal cations [4-10]. The AMOEBA parameterization process is based on gas phase *ab initio* calculations[11]. The Constrained Space Orbital Variations (CSOV)[12] energy decomposition analysis technique is applied to determine the contribution polarization to the overall interaction energy. The importance of charge transfer contribution will be elucidated with energy decomposition analysis and the nature of the interaction of Zn^{2+} with water will be investigated using the Electron Localization Function (ELF) topological analysis [13, 14]. Furthermore, Zn^{2+} solvation properties such as the ion-water radial distribution function, water residence times, coordination number, and the solvation free energy we will be calculated from AMOEBA condensed-phase simulations. These properties are compared with experimental results for Zn^{2+} and other divalent cations studied using AMOEBA and the results of this work is based on our previous reports[15].

2.2 METHODS

2.2.1 AMOEBA polarizable force field energy terms

The AMOEBA (Atomic Multipole Optimized Energetics for Biomolecular Applications) polarizable force field is a molecular mechanics model that represents electrostatic interactions with charge, dipole, and quadrupoles and had been described previously[16-18]. Bonded interactions are bond-stretching, angle-bending, bond-angle

stretch-bending, out-of-plane bending, and rotation about torsion. Non-bonded interactions are van der Waals, permanent and induced electrostatics.

$$U = U_{bond} + U_{angle} + U_{b\theta} + U_{oop} + U_{torsion} + U_{vdW} + U_{ele}^{perm} + U_{ele}^{ind}$$

Bond stretch energies utilize the fourth-order Taylor expansion of the Morse potential. Bond angle bend and torsion energies utilize a sixth-order potential and a six-term Fourier series expansion, respectively. These valence functional forms are the same as those used by the MM3[19]classical molecular mechanics potential. An out-of-plane bending was restrained at sp^2-hybridized trigonal centers with a Wilson-Decius-Cross function[20].

$$U_{bond} = K_b(b-b_0)^2[1-2.55(b-b_0)+3.793125(b-b_0)^2$$

$$U_{angle} = K_\theta(\theta-\theta_0)^2[1-0.014(\theta-\theta_0)+5.6\times10^{-5}(\theta-\theta_0)^2 - 7.0\times10^{-7}(\theta-\theta_0)^3+2.2\times10^{-8}(\theta-\theta_0)^4]$$

$$U_{b\theta} = K_{b\theta}[(b-b_0)+(b'-b'_\theta)](\theta-\theta_0)$$

$$U_{torsion} = \sum_n K_{n\phi}[1+\cos(n\phi\pm\delta)]$$

$$U_{oop} = K_\chi\chi^2$$

Bond lengths, bond/torsion phase angles, and energies are in units of Å, degrees and kcal/mol, respectively. The repulsion-dispersion interactions are represented with a buffered 14-7 potential [21].

$$U_{vdw}(ij) = \varepsilon_{ij}\left(\frac{1.07}{\rho_{ij}+0.07}\right)^7\left(\frac{1.12}{\rho_{ij}^7+0.12}-2\right)$$

The potential is a function of separation distance, R_{ij}, between atoms *i* and *j* $\rho_{ij} = R_{ij}/R_{ij}^0$ where R_{ij}^0 is the minimum energy distance and is combined for heterogeneous atom pairs $R_{ij}^0 = \frac{(R_{ii}^0)^3+(R_{jj}^0)^3}{(R_{ii}^0)^2+(R_{jj}^0)^2}$. In addition, the potential minimum in kcal/mol is combined for heterogeneous atom pairs $\varepsilon_{ij} = \frac{4\varepsilon_{ii}\varepsilon_{jj}}{(\varepsilon_{ii}^{1/2}+\varepsilon_{jj}^{1/2})^2}$.

Permanent electrostatic interactions are computed with higher order moments where

$$M_i = [q_i, d_{ix}, d_{iy}, d_{iz}, Q_{ixx}, Q_{ixy}, Q_{ixz}, Q_{iyx}, Q_{iyy}, Q_{iyz}, Q_{izx}, Q_{izy}, Q_{izz}]^T$$

is a multipole composed of charge, q_i, dipoles, $d_{i\alpha}$, and quadrupoles, $Q_{i\beta\gamma}$. The interaction energy between two multipole sites is

$$U_{emp}^{perm} = \begin{bmatrix} q_i \\ d_{ix} \\ d_{iy} \\ d_{iz} \\ Q_{ixx} \\ \vdots \end{bmatrix}^T \begin{bmatrix} 1 & \frac{\partial}{\partial x_j} & \frac{\partial}{\partial y_j} & \frac{\partial}{\partial z_j} & \cdots \\ \frac{\partial}{\partial x_i} & \frac{\partial^2}{\partial x_i \partial x_j} & \frac{\partial^2}{\partial x_i \partial y_j} & \frac{\partial^2}{\partial x_i \partial x_j} & \cdots \\ \frac{\partial}{\partial y_i} & \frac{\partial^2}{\partial y_i \partial x_j} & \frac{\partial^2}{\partial y_i \partial y_j} & \frac{\partial^2}{\partial y_i \partial y_j} & \cdots \\ \frac{\partial}{\partial z_i} & \frac{\partial^2}{\partial z_i \partial x_j} & \frac{\partial^2}{\partial z_i \partial y_j} & \frac{\partial^2}{\partial z_i \partial z_j} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \frac{1}{R_{ij}} \begin{bmatrix} q_j \\ d_{jx} \\ d_{jy} \\ d_{jz} \\ Q_{jxx} \\ \vdots \end{bmatrix}.$$

Multipoles are defined at atomic centers in relation to a local frame defined by other atoms that are bonded to it. A triplet of atoms is used to specify a local frame following the z-then-x convention. **Figure 2.1a** illustrates an example of an asymmetric local frame for atom A defined by atoms A, B, and C. The vector created by AB is the direction of the positive z-axis. The positive x-axis lies on the plane created by ABC and creates an acute angle with AC. The positive y-axis is defined to create a cubic right-handed coordinate system. **Figure 2.1b** illustrates an example of a local frame in which B and C are symmetric with respect to A, such as a water molecule. The z-axis is defined as the bisector of $\angle BAC$. The x-axis is defined as the vector along the aforementioned plane that creates an acute angle with AB and is orthogonal to the z-axis. As with the former case the y-axis is defined to create a right-handed coordinate system.

Electronic polarization describes the redistribution of electron density due to an external field. Polarizable point dipoles are utilized by AMOEBA at atomic centers to describe this effect. The induced dipole, $\mu_{i,\alpha}^{\text{ind}}$, at site *i* is

$$\mu_{i,\alpha}^{\text{ind}} = \alpha_i \left(\sum_{\{j\}} T_{\alpha}^{ij} M_j + \sum_{\{j'\}} T_{\alpha\beta}^{ij'} \mu_{j',\beta}^{\text{ind}} \right) \text{ for } \alpha, \beta = 1,2,3$$

where $M_j = [q_j, \mu_{j,1}, \mu_{j,2}, \mu_{j,3}, \dots]^T$ are the permanent charge, dipole, and quadrupole moments, and $T_{\alpha}^{ij} = [T_{\alpha}, T_{\alpha 1}, T_{\alpha 2}, T_{\alpha 3}, \dots]$ is the interaction matrix between atoms *i* and *j*. The Einstein convention is used to sum over indices α and β. The atomic polarizability, α_i, is parameterized for the zinc cation in this work. Note that the first term within the parenthesis corresponds to the polarization field due to permanent multipoles, while the second term corresponds to the polarization field due to induced dipoles produced at the other atoms.

An iterative induction approach originally developed by Thole [22] is adopted in which an induced dipole at site *i* continues to polarize all other sites until convergence is achieved at all induced dipole sites. This method imposes a damped polarization interaction at very short range in order to avoid a well-known artifact of point polarizability models by smearing one of the atomic multipole moments in each pair of interaction sites[23]. The damping functions for charge, dipole and quadrupole interactions have been derived previously [17]. The smearing function of a charge has the form

$$\rho = \frac{3a}{4\pi} \exp(-au^3)$$

and $u = r_{ij}/(\alpha_i \alpha_j)^{1/6}$ where r_{ij} is the linear separation between sites *i*, *j* and α_i, α_j are their corresponding atomic polarizabilities. The factor "*a*" is a dimensionless width parameter that determines the damping strength.

2.2.2 Gas phase *ab initio* calculations

The intermolecular interaction energies of Zn^{2+}-H_2O at various separation distances were calculated using GAUSSIAN 03[24] at the MP2(full) level. Basis Set Superposition Error (BSSE) correction was included in the interaction energy. The geometry of the water was fixed to the AMOEBA minimized structure [17, 25]. The aug-cc-pVTZ basis set[26] was employed for water and the 6-31G(2d,2p) basis set for the Zn^{2+} cation. Constrained Space Orbital Variations (CSOV) polarization energy calculations were performed using a modified version[27] of HONDO95.3[27] with the B3LYP methods[28, 29] using the above basis sets. The Zn^{2+} atomic polarizability was computed using GAUSSIAN 03 at the MP2(full)/6-31G** level.

Additional energy decomposition analysis was performed on zinc hydrated cluster with the Reduced Variational Space (RVS) scheme as implemented in the GAMESS [30, 31] software. The RVS energy decomposition computations were performed at the Hartree-Fock (HF) level using the CEP 4-31G(2d) basis set[32] augmented with two diffuse 3*d* polarization functions on heavy atoms (double zeta quality pseudopotential) and at the aug-cc-pVTZ basis set level (6-31G** for Zn(II)).

2.2.3 Electron Localization Function analysis (ELF)

In the framework of the ELF topological analysis [13, 14, 33], the molecular space is divided into a set of molecular volumes (basins) localized around maxima (attractors) of the vector field of the scalar ELF function. The ELF function can be interpreted as a signature of the electronic-pair distribution and ELF is defined to have values restricted between 0 and 1 to facilitate its computation on a 3D grid and its interpretation. The core regions can be determined (if $Z > 2$) for any atom A. Regions associated with lone pairs are referred to as V(A) and regions of chemical bonds are denoted V(A,B). The approach offers an evaluation of the basin electronic population as

well as an evaluation of local electrostatic moments. It is also important to point out that metal cations exhibit a specific topological signature in the electron localization of their density interacting with ligands according to their "soft" or "hard" character. A metal cation can split its outer-shell density (the so-called subvalent domains or basins) according to its capability to form a partly covalent bond involving charge transfer [34]. More details about the ELF function and its application to biology can be found in a recent review[35]. All computations have been performed using a modified version [36] of the Top-Mod package[37].

2.2.4 Parameterization of the AMOEBA Zn^{2+} model

The Zn^{2+} cation is parameterized, with a procedure similar to that previously used for Ca^{2+} and Mg^{2+} [6], by first matching the distance-dependence of AMOEBA polarization energies of the ion-water dimer in gas-phase obtained with *ab initio* Constrained Space Orbital Variations (CSOV) polarization energy results. With water geometry fixed, the Zn^{2+}-O distance were varied between 1.5 and 5 Å. The damping factor "*a*" was adjusted so that the AMOEBA polarization energy matched the CSOV values as much as possible. The remaining contribution (namely charge transfer) will be included in the van der Waals term as a result of matching the total binding energy of AMOEBA to that of QM. Parameters for the van der Waals interaction, R^0 (radius) and ϵ (well-depth), were derived by comparing the total ion-water binding energy computed by AMOEBA to the *ab initio* values at various distances. In the absence of an explicit charge transfer term, such strategy is justified as the charge transfer contribution is notably smaller in magnitude compared to polarization[10, 11, 38, 39] and a good percentage of it (dominated by the 2-body interactions) could be accurately included within AMOEBA's van der Waals term assuming that many-body charge transfer is not the driving force of

Zn(II) solvation dynamics. The validity of such assumption and the applicability of the present parameterization scheme to Zn^{2+} will be discussed in the first section of the discussion. For interactions between different types of atoms, these parameters undergo combination rules as described by Ponder [18]. The binding energies were computed as the total energy less the isolated water and ion energies at infinite separation distance.

2.2.5 Molecular dynamics simulations with AMOEBA

The AMOEBA polarizable force field [5, 17, 25] is used to study the solvation dynamics of Zn(II). Molecular dynamics simulations were performed via the TINKER 5 package[40] to compute the solvation free energy of Zn^{2+}. Independent simulations were first performed to "grow" the Zn vdW particle by gradually varying $R(\lambda) = \lambda(R_{final})$ and $\varepsilon(\lambda) = \lambda(\varepsilon_{final})$ where λ = (0.0, 0.0001, 0.001, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0). Subsequently, simulations were performed to "grow" the (+2) charge of Zn^{2+} along with its polarizability such that $q(\lambda') = \lambda'(q_{final})$ and $\alpha(\lambda') = \lambda'(\alpha_{final})$ where λ' = (0.0, 0.1, 0.2, 0.3, 0.325, 0.350, 0.375, 0.400, 0.425, 0.450, 0.475, 0.500, … , 1.0). The long-range electrostatics is modeled with particle-mesh Ewald summation for atomic multipoles with a cutoff of 7 Å in real space and 0.5 Å spacing and a 5th-order spline in reciprocal space[41]. The convergence criterion is 0.01 D for induced dipole computation. Molecular dynamics simulations were performed with a 1 fs timestep for 500 ps at each perturbation step. Trajectories were saved every 0.1 ps after the first 50 ps equilibration period. Temperature was maintained using the Berendsen weak coupling method at 298K[42, 43]. The system contained 512 water molecules with one Zn^{2+} ion and 24.857 Å is the length of each side of the cube.

The absolute free energy was computed from the perturbation steps by using the Bennett acceptance ratio (BAR), a free energy calculation method that utilizes forward

and reverse perturbations to minimize variance[43, 44]. MD simulations were extended for 2.2ns (total 2.7 ns) with the final Zn^{2+} parameters and the resulting trajectory was used in the analysis of the structure and dynamics of water molecules in the first solvation shell. Water molecules separated by a distance less than the first minimum of the Zn^{2+}-O RDF were considered to be in the first solvation shell. The averaged residence time of the first shell water molecules was directly measured by monitoring the entering and exit events.

2.3 RESULTS AND DISCUSSION

2.3.1 Contribution of charge transfer in Zn^{2+}-water complexes

Consistent with most molecular mechanics models, AMOEBA does not explicitly represent charge transfer. However, when the charge transfer contribution is significant, despite its limited magnitude in many-body complexes, it may be difficult to capture the overall many-body effect by only considering polarization. Therefore, it is important to investigate the charge transfer contribution to the Zn^{2+}-water interaction energy and its dependence on the system size. To estimate the magnitude of charge transfer, we performed several Reduced Variational Space (RVS) energy decomposition analysis on complexes up to $[Zn(H_2O)_6]^{2+}$.

We report here complexes that were initially studied by Gresh *et al.*[10, 45, 46]: monoligated $[Zn(H_2O)]^{2+}$ complex and polyligated $[Zn(H_2O)_6]^{2+}$, $[Zn(H_2O)_5(H_2O)]^{2+}$, and $[Zn(H_2O)_4(H_2O)_2]^{2+}$ arrangements (octahedral-> pyramidal -> tetrahedral first-shell). As we can see in **Table 2.1**, the importance of charge transfer relative to polarization varies with the size of the Zn^{2+}-(H_2O)n complex and depends on the basis set. It makes up a significant portion of induction for a monoligated $[Zn(H_2O)]^{2+}$ and its contribution decreases as number of ligating water molecules increase to 6. The charge transfer effect

appears to be diluted within the entire induction energy (polarization and charge transfer) as the number of water molecules grows in agreement with previous observation of anti-cooperative effects[10, 45, 46]. Note that basis set superposition error (BSSE) is not taken into account. As indicated by Stone[47], such systematic error can be clearly associated with the charge transfer effect. In contrast to the inverse relationship between CT and water ligation expressed by the zinc cation, the CT contribution associated with anions, such as Cl^-, has been observed to increase as ligation increases[48]. These phenomena may be due to the asymmetric solvation environment for the anions as well as their modes of water ligation. However, analyses of CT effects are not apparent as they are found in both induction energy and basis set superposition error[47]. For the largest complex $[Zn(H_2O)_6]^{2+}$), the BSSE amounts to 3.3 kcal/mol. After BSSE correction, the relative weight of charge transfer to the total induction reduces from 16.6% (**Table 2.1**) to 15.3% at the CEP-31G(2d) level. Using the large aug-cc-PVTZ for water coupled to the 6-31G** basis set for Zn(II), the observed trends are even more pronounced as the relative importance of charge transfer strongly diminishes from 6.4% of the whole induction for $[Zn(H_2O)_4]^{2+}$ to less than 4% for the $[Zn(H_2O)_6]^{2+}$ complex while polarization becomes more dominant. Thus the magnitude of the CT estimated by *ab initio* methods is greatly dependent on the basis set used. While our results have been obtained at the Hartree-Fock level, recent studies clearly show that correlation acts on induction and leads to greater charge transfer energy[27, 49]. For this reason, we computed the induction energies on selected water clusters at both HF and DFT level using a recently introduced energy decomposition analysis (EDA) technique based on single configuration-interaction (CI) localized fragment orbitals[50]. We indeed find that the CT contribution increases slightly with DFT, however, overall it accounts for less

than 20% of the total induction energy for monoligated complexes and presumably would be even less in the bulk water environment.

To gain further insight into the interaction of Zn^{2+} with water, we performed the Electron Localization Function (ELF) analysis. An important asset of the ELF topological analysis is that is provides a clear description of a covalent bond between two atoms as it exhibits a basin between atoms to indicate electron sharing. Here, we have considered several Zn^{2+}-$(water)_n$ complexes, n =1 to 6. An important discovery from ELF analysis is that a covalent V(Zn,O) is only observed in the monoligated Zn^{2+}-water complex (Figure 2.2). In that case, we observe a net concentration of electrons between the zinc cation and the water oxygen, a clear sign of covalent bonding (1.9 e- on the bond). As n increases, the covalent V(Zn,O) feature disappears despite a residual mixing of Zn^{2+} contributions in the oxygen basin. Indeed, as the Zn-O distances increase with n (Figure 2.3 and Figure 2.4), the Zn-O bond becomes more ionic as the charge transfer quickly diminishes. Such behavior could be then understood using the subvalence concept[34]. As shown by de Courcy *et al.*[34, 38], the cation density is split into several "subvalent" domains as its outer shells appear strongly polarized, which explains why covalency is not achieved. If the cation electron density is strongly delocalized towards the oxygen atoms, the center of the basin remains closer to Zn^{2+} (covalent bonding would implicate a polarized bond with a covalent V(Zn,O) basin localized closer to the more electronegative oxygen). ELF results thus suggest that although the induction in the Zn^{2+}-water monoligated complex is dominated by charge transfer, this is not to the case for n from 2 to 6. In the latter case, the many-body effects are driven by the Zn^{2+} outer shells' plasticity that accommodates the strongly polarized water molecules. The Atoms in Molecules (AIM) population analysis confirms that such behavior is present in DFT as well as at the MP2 level. As expected (see [27, 51] for example), DFT tends to slightly over-bind the complexes as

compared to MP2 which clearly gives a better description of the bonding over Hartree-Fock.

To conclude on these various results, we expect that AMOEBA will improve in accuracy with increase in system size as the charge transfer effect becomes less important and the total induction will be dominated by polarization. In other words, we anticipate the discrepancy between AMOEBA and QM observed in the monoligated water-Zn^{2+} complex to disappear in the condensed-phase. This also suggests that an "ad-hoc" inclusion of the charge transfer into the polarization contribution by adjusting the polarization damping scheme (see the Thole model in the Computational Details) is probably not a suitable strategy. Indeed, charge transfer can rapidly vanish, and "polarization only" models over-fitted on monoligated complexes to include charge transfer will lead to an overestimated many-body effect in bulk-phase simulation as the polarization would still contain the unphysical charge transfer. Charge transfer should be treated explicitly or included in the van der Waals to certain extent. In this study, we adopt the latter approach to effectively incorporate the charge transfer in the bulk environment into the vdW interactions.

2.3.2 Accuracy of the AMOEBA parameterization

The distance dependent dimer binding energies were used to adjust vdW parameters (R and ε) and the damping factor of polarizability (*a*) for Zn^{2+} was adjusted to match the CSOV polarization energy. **Table 2.2** lists the final parameters of the Zn^{2+} cation as well as the Mg^{2+} and Ca^{2+} cations parameterized by Jiao *et al.*[5] that are optimized for the Tinker implementation of AMOEBA. Meanwhile, parameters optimized for a slightly modified implementation of the AMOEBA force field present in Amber which embodies a modified periodic boundary condition treatment of long range

van der Waals are available as well[6]. It should be noted that although the previously reported parameters for Mg^{2+} and Ca^{2+} contained typographic inconsistencies[5], results from that work (thermodynamic energy, structural analysis, etc.) are obtained from parameters consistent with **Table 2.2**. **Figure 2.5** compares CSOV polarization energy calculations with the AMOEBA polarizable force field as a function of distance between the cation and water. The difference between the two methods is mainly found at distances between 2 – 3 Å, where the charge transfer effect in the two-body system is strong. However, such discrepancy is expected to diminish in bulk water as the charge transfer effect is expected to be less important as explained above. Comparison between total binding energies of the AMOEBA polarizable model and *ab initio* calculations are shown in **Figure 2.6**. As expected, the interaction energy between 2 and 3.5 Å appears to be underestimated (less negative) compared to *ab initio* result. The strategy here is, however, not to over-fit the AMOEBA model to the monoligated Zn^{2+} complex as the polarization energy and total interaction energy are already very reasonable considering the relatively simple force field functional form. The AMOEBA association energies for $[Zn(H_2O)_6]^{2+}$, $[Zn(H_2O)_5(H_2O)]^{2+}$ and $[Zn(H_2O)_4(H_2O)_2]_2$ complexes are -334.4, -333.4, and -331.9/-333.7 kcal/mol, respectively. Given that AMOEBA is mainly targeting condensed phase, the trend observed here is in reasonable agreement with the previous *ab initio* results (-345.3, -341.3, -337.4/-337.8 kcal/mol using CEP 4-31G (2d) basis set; -365.9, -363.3, -360.0, -362.4 kcal/mol using 6-311G** basis set)[46]. Our approach is further validated in the condensed-phase hydration properties calculation next.

2.3.3 Evaluation of Zn^{2+} Solvation in Water Using AMOEBA

The hydration free energy is the key quantity describing the thermodynamic stability of an ion in solution. The solvation free energy of zinc in water has been

computed from molecular dynamics simulations using free energy perturbation (FEP). **Table** 2.3 lists the free energy of hydration for Zn^{2+}, Mg^{2+}, and Ca^{2+} compared with experiment-derived values[52, 53] and the results from the quasi-chemical approximation method[54]. The free energy values computed from AMOEBA are closer to those from quasi-chemical approximation (QCA) than to the data interpreted from experimental measurement. In the QCA method, the region around the solute of interest is partitioned into inner and outer shell domains. The inner shell is treated quantum mechanically while the outer shell was evaluated using a dielectric continuum model. Note that to decompose the hydration free energy of a neutral ion-pair, tetraphenylarsonium tetraphenylborate (TATB) has been most widely chosen as a reference salt, based on the extra thermodynamic assumption that the large and hydrophobic ions do not produce charge-specific solvent ordering effects[52, 55]. Our results show better agreement with "experimental values" for Ca^{2+} and Mg^{2+} ions by Schmid who derived the single ion hydration free energy by using the theoretically determined proton hydration free energy as a reference[53]. The hydration free energy for Zn^{2+} ion computed using AMOEBA is in good agreement with values given by Marcus[52] and Asthagiri et al.[54], with deviations less than 1.9% and 0.2%, respectively.

2.3.4 Solvent Structure and Dynamics

To characterize the structure of water molecules around the ion, the radial distribution function (RDF) between the Zn^{2+} and oxygen atom of water molecule has been obtained from the 2.7-ns molecular dynamics simulation (**Figure 2.7**). The running integration of Zn-O, which imparts water-ion coordination information, is also plotted. The first minimum in the ion–O RDF is at a distance of 2.85 Å, which can be interpreted as the effective "size" of the complex composed of the ion and first water solvent shell.

The running integration indicates a water-coordination number of 6 in the first solvation shell, which is consistent with experimental observations [56-61]. The zinc cation expectedly binds to the first water shell more tightly than other ions, as evident in the more pronounced, narrow first peak as well as the shortest separation as shown in the ion-O RDFs in **Figure 2.8**. Overall the zinc solvation structure show greater similarity to Mg^{2+} than Ca^{2+}.

The Born theory of ion solvation[62] states that there exists an effective solvation radius, R_B, for each ion such that the solvation free energy of the ion in a dielectric medium is given by

$$\Delta A = -\frac{q^2}{2R_B}\left(1-\frac{1}{\epsilon_d}\right)$$

where q is the charge of the ion and ϵ_d is the dielectric constant of the medium (80 for water). We have calculated the effective radius of zinc based on the Born equation from the solvation free energy obtained from our simulations. **Table 2.4** gives a detailed comparison among Zn^{2+}, Mg^{2+} and Ca^{2+}. It should be noted, however, that previous studies have shown ion hydration energy is not symmetric with respect to electronegativity[4, 63, 64] as is implied by the Born theory. The first peak of the Zn^{2+}-O RDF is at 1.98 Å and the effective Born radius of the cation is calculated to be 1.47 Å. A difference of ~0.5 Å between the two quantities is consistent with the results of other mono- and divalent metal ions[4, 5, 65-67]. The difference between the first minimum in the Zn^{2+}-O RDF and the Born radius is 1.38 Å and is consistent with studies of other ions as well[4, 5].

In addition to the RDF, the solvation structure has been analyzed from the distribution of the angles formed by O–ion–O in the first water shell. **Figure 2.9** compares the distribution of angles for Zn^{2+}, Mg^{2+}, and Ca^{2+} cations. With sharp peaks

located near 90° and 180°, the distribution of O-Zn^{2+}-O angle suggests rigid octahedron geometry with the Zn^{2+} surrounded by six water molecules. Mg^{2+} shares a similar but slightly more flexible geometry, while results for Ca^{2+} suggest a more amorphous structure. **Figure 2.11** is a sample frame from the molecular dynamics simulation to illustrate the octahedron arrangement between the zinc and the first shell water molecules.

2.3.5 Dipole Moment

The average dipole moment of water as a function of distance away from the zinc cation is computed. At the closest distance of 1.9 – 2.5 Å water experiences dipole moment from 3.0 – 3.9D. Due to the highly organized structure of the first water shell, a "vacuum" space free of water molecules is observed between 2.6 – 3.2 Å away from the cation, also evident in the Zn^{2+}-O RDF. The higher dipole moment of Zn^{2+} relative to bulk water (2.77D[17]) within the first water shell is consistent with previous observation of other divalent cations[5]. The dipole moment of water in the first solvation shell of monovalent cations such as K^+ and Na^+, however, is lower than that of bulk water[55].

2.3.6 Residence Time

We have investigated the lifetime of ion-water coordination by directly examining the average amount of time that a water molecule resides within the first solvation shell. The first solvation shell is determined by position of the first minimum of the Zn-O RDF. If an oxygen atom is less than 2.85 Å away from the Zn^{2+}, the water is considered to be in the first solvation shell. Cutoff distances used for the first solvation shells of Mg^{2+} and Ca^{2+} are 2.95 Å and 3.23 Å, respectively. In **Table 2.5**, coordination numbers and residence times from AMOEBA simulations are compared with experimental values for Zn^{2+}, Mg^{2+} and Ca^{2+}[8, 68-74]. The Zn^{2+} to water-proton dynamics are studied with

quasi-elastic neutron scattering methods (QENS) as described by Salmon[68]. The water residence times directly sampled from the MD simulations are in better agreement with experimental results than those previously inferred from the time correlation function of the instantaneous first shell coordination number[5]. According to AMOEBA simulations, the residence time in the first solvation shell around Zn^{2+} is at least 2 ns and the water molecules around Ca^{2+} have a life time on the order of several ps, both of which are within the experimental ranges. For Mg^{2+}, experiment suggests that water molecules could live up to a few μs while the simulations using AMOEBA indicates a residence time similar to that of Zn^{2+}. Classical fixed-charge molecular mechanic methods suggest a residence time of 146 ps[75] for water around Zn^{2+}, while quantum mechanical methods have not attained simulation times long enough to observe the exchange of water molecules in the first shell[56, 76]. The calculated water residence times are consistent with the analyses of radial distribution function and water angle distribution. A longer residence time is accompanied by a more ordered and closely packed water structure near the cation.

2.4 CONCLUSIONS

We showed that AMOEBA was able to provide a reasonably accurate description of Zn^{2+} interaction with water, especially in the bulk water environment. We explained in detail one of the reasons for such good performance - the *ab initio* calculations demonstrated that the relative importance of charge transfer diminishes as the number of water molecule increases, a sign of anti-cooperativity. We have established a fitting strategy for induction: charge transfer can be included into the pair-wise dispersion in the van der Waals contribution; incorporation of charge transfer into polarization would lead to an overestimation of the many-body effects. Despite the difficulty of the AMOEBA

model to reproduce the binding energy of the monoligated Zn^{2+}-water complex, which exhibits non-classical covalent bonding as shown by ELF topological analysis, AMOEBA is able to afford robust estimation of the hydration free energy along with reasonable solvation structure and dynamics. The current and previous studies suggest that the classical polarizable multipole-based AMOEBA is an effective tool to model ion in bulk solution as good relative solvation free energies, structure and dynamic properties have been obtained for a range of mono- and divalent cations. The work clearly demonstrates the need of "interpretative" *ab initio* techniques (ELF, EDA methods) in order to follow a bottom-up approach going from the gas phase *ab initio* calculations to condensed-phase MD simulations.

Table 2.1: Polarization energy and charge transfer energy from restricted variational space (RVS) energy decomposition of Zn^{2+} in the presence of water clusters of sizes 1, 4, 5, and 6 at the HF/CEP-41G(2d) level (or HF/aug-cc-PVTZ/6-31G**, results in parentheses). Percentage of induction energy due to charge transfer is presented in the last row. All are in units of kcal/mol.

Complex	**$Zn(H_2O)$**	**$[Zn(H_2O)_4]^{2+}$**	**$[Zn(H_2O)_5]^{2+}$**	**$[Zn(H_2O)_6]^{2+}$**
E_{pol}(RVS)	-37.6	-118.7 (-135.3)	-110.8 (-127.5)	-104.3 (-117.5)
E_{CT}(RVS)	-10.9	-28.7 (-9.3)	-24.5 (-6.7)	-21.8 (-4.51)
$(E_{CT}/(E_{pol}{}^{+}E_{CT}))*100$	22.5	19.4 (6.4)	18.1 (5.0)	16.6 (3.7)

Table 2.2: Ion parameters are shown: diameter, well depth, polarizability and dimensionless damping coefficient.

Ion	R (Å)	ϵ (kcal/mol)	α (Å^3)	a[a]
Zn^{2+}	2.68	0.222	0.260	0.2096
Mg^{2+}	2.94	0.300	0.080	0.0952
Ca^{2+}	3.63	0.350	0.550	0.1585

[a] *a* is the dimensionless damping coefficient.

Table 2.3: Solvation Free Energy of Zinc in Water[a]

Ion	ΔG (kcal/mol)	Experimental	Quasi-chemical[d]
Zn^{2+}	-458.9 (4.4)	-467.7[b]	-460.0
Mg^{2+}	-431.1 (2.9)	-435.4[c]	-435.2
Ca^{2+}	-354.9 (1.7)	-357.2[c]	-356.6

a 1 mol L^{-1} solution is chosen as the standard state.
b Reference [52]
c Reference [53]
d Reference [54]

Table 2.4: Radii results for Zn^{2+}, Mg^{2+}, and Ca^{2+} cations. Born radii, first peak in ion-O RDF with AMOEBA polarizable force field, experimental first peak in ion-O RDF, and first minimum in ion-O RDF are all indicated in Å.

Ion	Born Radius (Å)	First peak in ion-O RDF	Experimental first peak in ion-O RDF	QM/MM first peak	First minimum in ion-O RDF
Zn^{2+}	1.47	1.98	2.07[a]	2.11-2.18 [a]	2.85
Mg^{2+}	1.56[b]	2.07	2.09[c]	2.13 [d]	2.95
Ca^{2+}	1.89 [b]	2.41	2.41-2.44; 2.437; 2.46[c]	2.43 – 2.44 [d]	3.23

a Reference [57]
b Reference [5]
c References [65, 66] and [67]
d References [65, 66]

Table 2.5: The coordination number, experimental coordination number, residence time, experimental residence time, and QM/MM residence times for each type of divalent cations.

Ion	Coordination Number	Exp. Coordination Number	Residence Time (s)	Exp Residence Time (s)
Zn^{2+}	6	6[a]	2.2×10^{-9}	$10^{-10} – 10^{-9}$ [d]
Mg^{2+}	6	6[b]	1.9×10^{-9}	$2\times10^{-6} – 10^{-5}$ [e,f]
Ca^{2+}	7.3	7.2 +- 1.2[c]	1.33×10^{-10}	$<10^{-10} – 10^{-7}$ [f]

a Reference [56-61]
b Reference [73]
c Reference [74]
d Reference [68] and [8]
e References (Neely, 1970)[69]
f Reference (Helm, 1999)[72], (Friedman, 1985)[70] and references within (Ohtaki, 1993)[71]

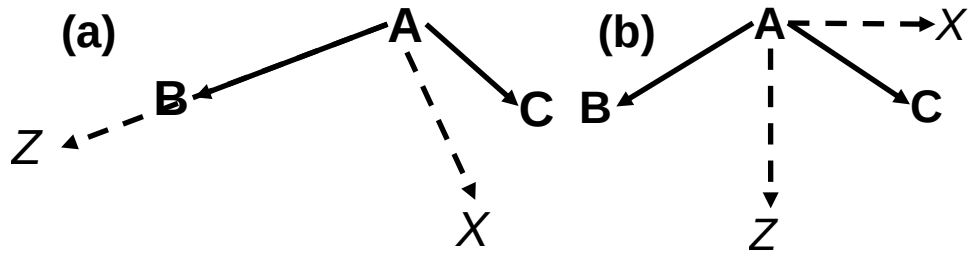


Figure 2.1: Given atoms A, B, and C, the local frame is identified by the z-axis and x-axis as shown. The y-axis is defined to create a right-handed coordinate system with the existing axes.

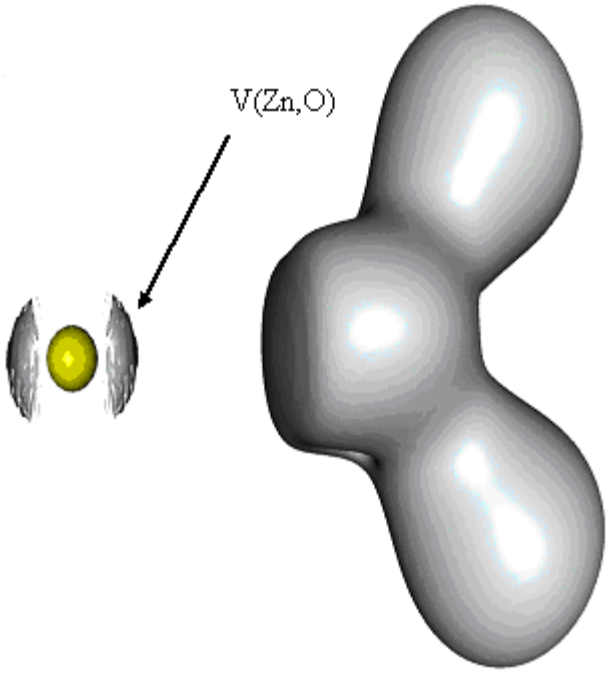


Figure 2.2: ELF localization domains (basins) for the Zn^{2+}-H_2O complex. A covalent V(Zn,O) basin reflecting electron sharing is observed and reveals the covalent nature of the Zn-O interaction.

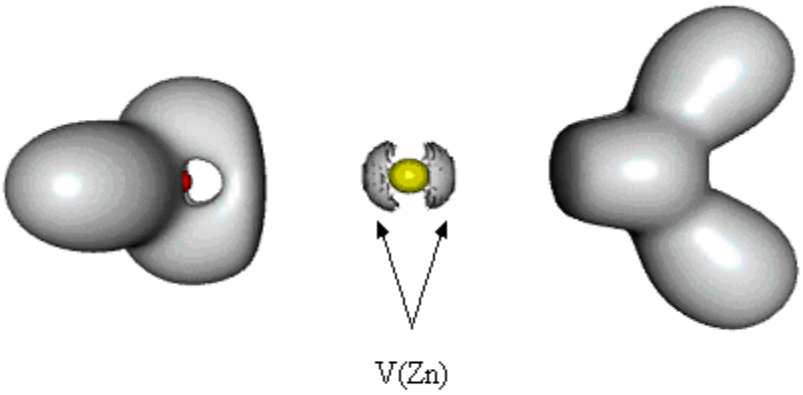


Figure 2.3: ELF localization domains (basins) for the Zn^{2+}-$(H_20)_2$ complex. Non-covalent V(Zn) basin are observed describing the deformation of Zn^{2+} outer-shells density within the fields of the water molecules.

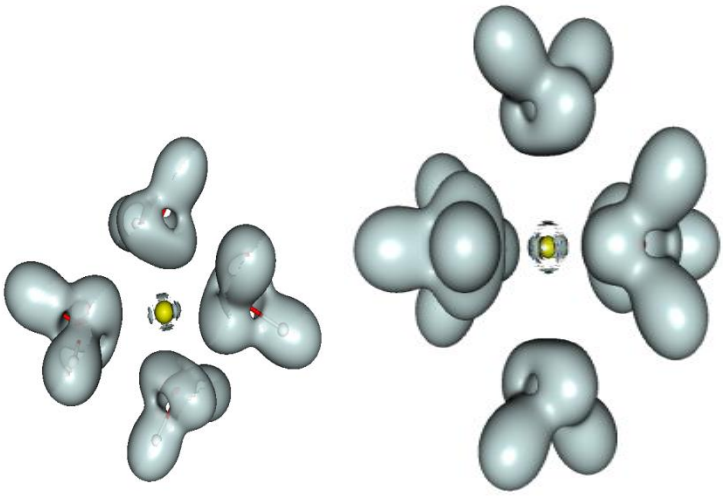


Figure 2.4: ELF localization domains (basins) for the Zn^{2+}-$(H_20)_4$ and Zn^{2+}-$(H_20)_6$ complexes. Again, non covalent V(Zn) basin are observed.

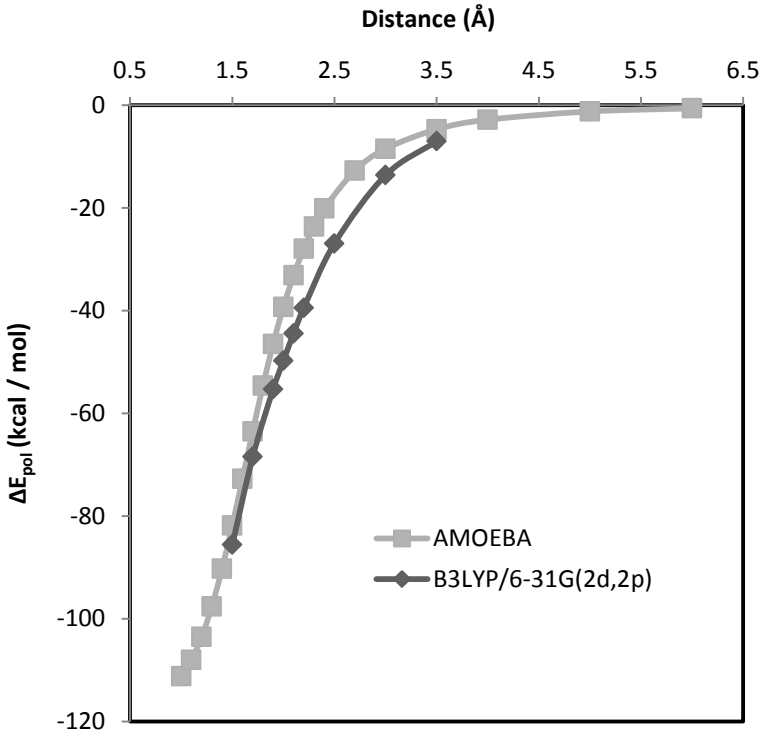


Figure 2.5: Polarization energy of zinc and water dimer in gas phase as a function of separation distance.

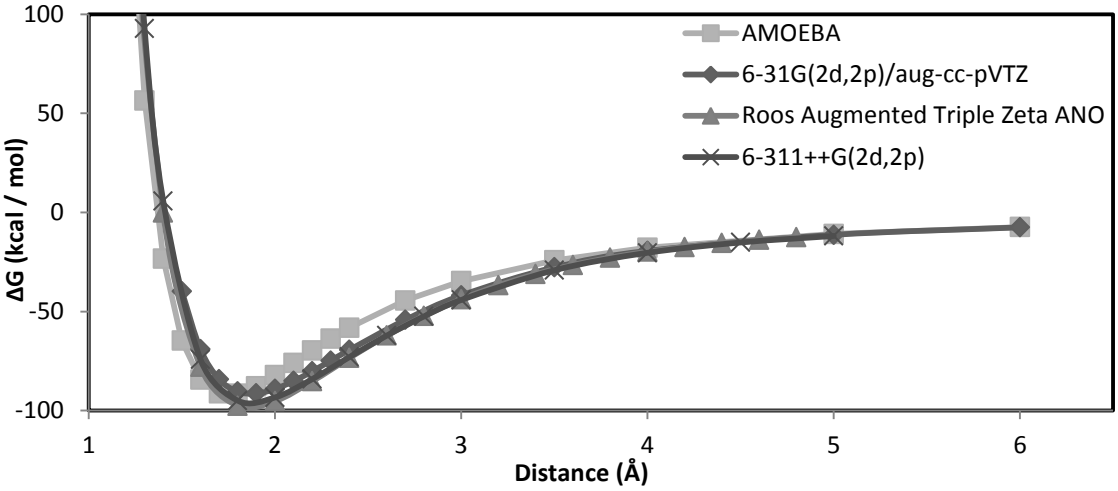


Figure 2.6: Binding energy of zinc and water dimer in gas phase as a function of separation distance. The 6-31G(2d,2p)/aug-cc-pVTZ indicates that 6-31G(2d,2p) was used to represent the Zn^{2+} cation and aug-cc-pVTZ was used to represent the water molecule. Binding energy obtained from the last two basis sets used the same basis sets for both ion and water.

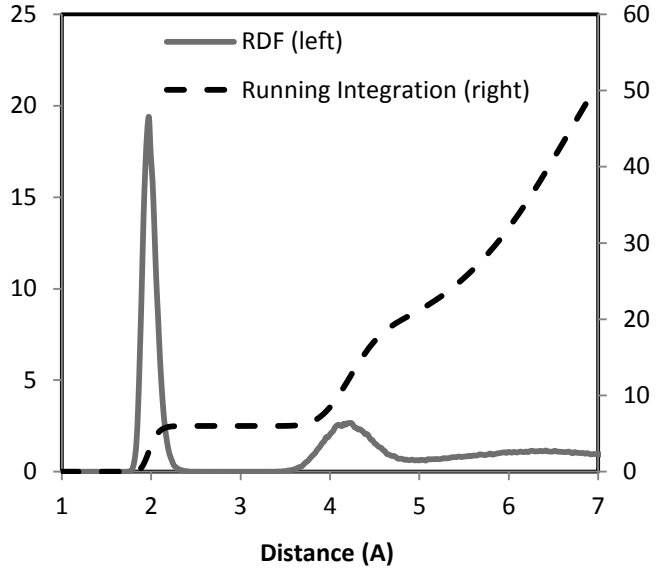


Figure 2.7: Radial distribution function of Zn^{2+}-O (left axis) and water coordination number (right axis).

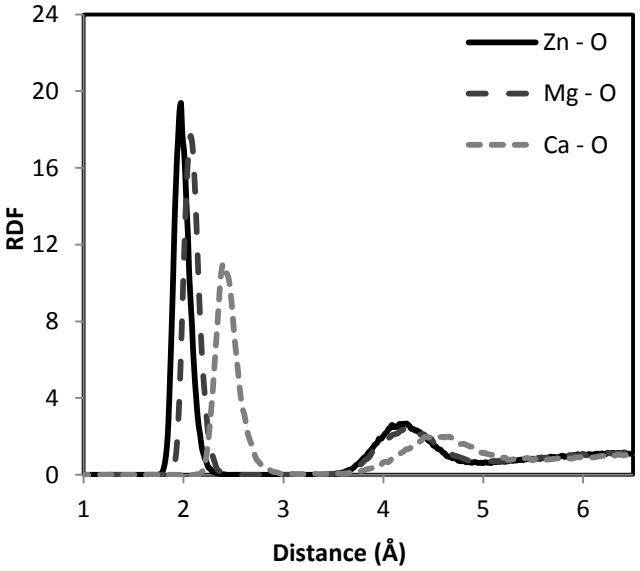


Figure 2.8: Radial distribution function of divalent cations (Zn^{2+}, Mg^{2+}, and Ca^{2+}) and oxygen atom in water.

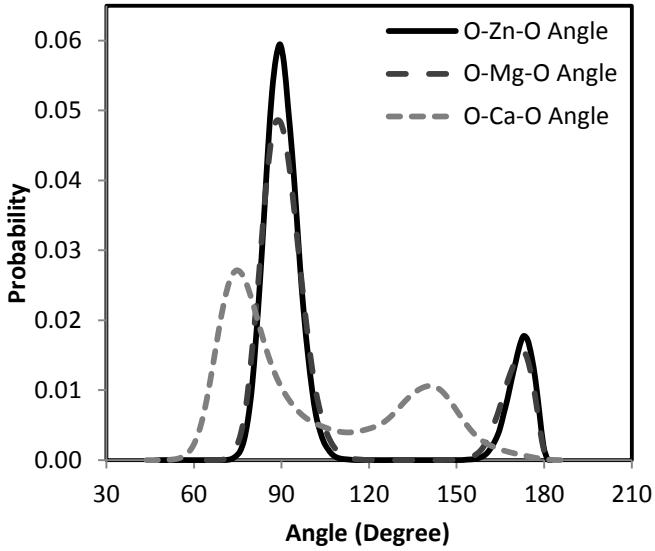


Figure 2.9: Water-Ion-Water Angle distribution of divalent cations (Zn^{2+}, Mg^{2+}, and Ca^{2+}) and oxygen atom in water.

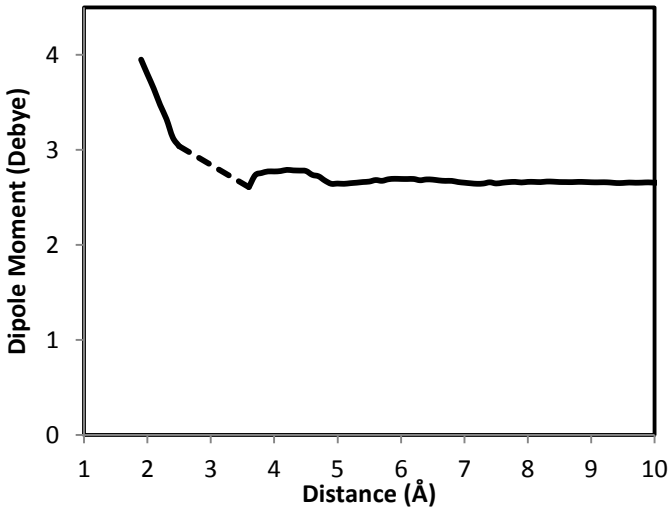


Figure 2.10: Dipole moment at each distance (Å) around ion. The dashed line is the interpolated dipole moment since water molecules were not sampled for the duration of the molecular dynamics simulation.

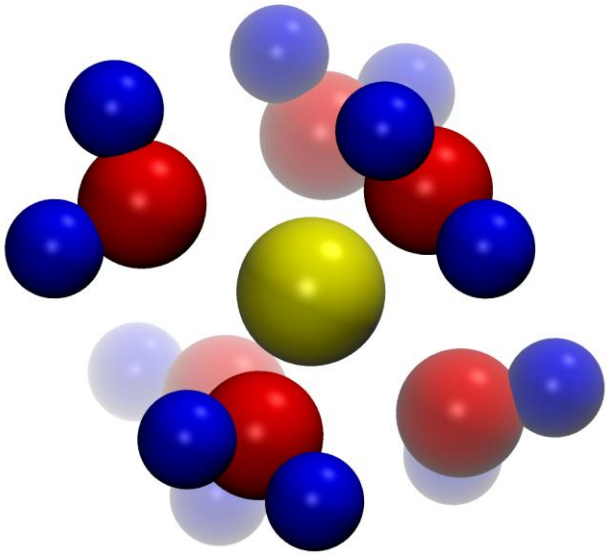


Figure 2.11: First solvation shell around Zn^{2+} ion.

2.5 REFERENCES

1. Keilin, D.; Mann, T., *Carbonic anhydrase. Purification and nature of the enzyme.* Biochem. J., 1940. **34**(8-9): p. 1163-1176.
2. Lipscomb, W.N.; Strater, N., *Recent advances in zinc enzymology.* Chem. Rev. (Washington, DC, U. S.), 1996. **96**(7): p. 2375-2433.
3. Maynard, A.T.; Covell, D.G., *Reactivity of zinc finger cores: Analysis of protein packing and electrostatic screening.* J. Am. Chem. Soc., 2001. **123**(6): p. 1047-1058.
4. Grossfield, A., *Dependence of ion hydration on the sign of the ion's charge.* J. Chem. Phys., 2005. **122**(2): p. 024506.
5. Jiao, D.; King, C.; Grossfield, A.; Darden, T.A.; Ren, P.Y., *Simulation of Ca2+ and Mg2+ solvation using polarizable atomic multipole potential.* J. Phys. Chem. B, 2006. **110**(37): p. 18553-18559.
6. Piquemal, J.P.; Perera, L.; Cisneros, G.A.; Ren, P.Y.; Pedersen, L.G.; Darden, T.A., *Towards accurate solvation dynamics of divalent cations in water using the polarizable amoeba force field: From energetics to structure.* Journal of Chemical Physics, 2006. **125**(5): p. 054511.
7. de Courcy, B.; Piquemal, J.P.; Gresh, N., *Energy Analysis of Zn Polycoordination in a Metalloprotein Environment and of the Role of a Neighboring Aromatic Residue. What Is the Impact of Polarization?* J. Chem. Theory Comput., 2008. **4**(10): p. 1659-1668.
8. Roux, C.; Gresh, N.; Perera, L.; Piquemal, J.; Salmon, L., *Binding of 5-phospho-D-arabinonohydroxamate and 5-phospho-D-arabinonate inhibitors to zinc*

phosphomannose isomerase from Candida albicans studied by polarizable molecular mechanics and quantum mechanics. J. Comput. Chem., 2007. **28**(5): p. 938-957.

9. Jenkins, L.; Hara, T.; Durell, S.; Hayashi, R.; Inman, J.; Piquemal, J.; Gresh, N.; Appella, E., *Specificity of acyl transfer from 2-mercaptobenzamide thioesters to the HIV-1 nucleocapsid protein.* J. Am. Chem. Soc., 2007. **129**(36): p. 11067-11078.
10. Gresh, N.; Piquemal, J.; Krauss, M., *Representation of Zn(II) complexes in polarizable molecular mechanics. Further refinements of the electrostatic and short-range contributions. Comparisons with parallel ab initio computations.* J. Comput. Chem., 2005. **26**(11): p. 1113-1130.
11. Gresh, N.; Cisneros, G.A.; Darden, T.A.; Piquemal, J.P., *Anisotropic, polarizable molecular mechanics studies of inter- and intramoecular interactions and ligand-macromolecule complexes. A bottom-up strategy.* J. Chem. Theory Comput., 2007. **3**(6): p. 1960-1986.
12. Bagus, P.S.; Illas, F., *Decomposition of the chemisorption bond by constrained variations - Order of the variations and construction of the variational spaces.* J. Chem. Phys., 1992. **96**(12): p. 8962-8970.
13. Becke, A.D.; Edgecombe, K.E., *A simple measure of electron localization in atomic and molecular-systems.* J. Chem. Phys., 1990. **92**(9): p. 5397-5403.
14. Silvi, B.; Savin, A., *Classification of chemical-bonds based on topological analysis of electron localization functions.* Nature, 1994. **371**(6499): p. 683-686.
15. Wu, J.C.; Piquemal, J.P.; Chaudret, R.; Reinhardt, P.; Ren, P., *Polarizable molecular dynamics simulation of Zn(II) in water using the AMOEBA force field.* J. Chem. Theory Comput., 2010. **6**(7): p. 2059-2070.
16. Ren, P.Y.; Ponder, J.W., *Consistent treatment of inter- and intramolecular polarization in molecular mechanics calculations.* Journal of Computational Chemistry, 2002. **23**(16): p. 1497-1506.
17. Ren, P.Y.; Ponder, J.W., *Polarizable atomic multipole water model for molecular mechanics simulation.* J. Phys. Chem. B, 2003. **107**(24): p. 5933-5947.
18. Ponder, J.W.; Wu, C.J.; Ren, P.Y.; Pande, V.S.; Chodera, J.D.; Schnieders, M.J.; Haque, I.; Mobley, D.L.; Lambrecht, D.S.; DiStasio, R.A.; Head-Gordon, M.; Clark, G.N.I.; Johnson, M.E.; Head-Gordon, T., *Current Status of the AMOEBA Polarizable Force Field.* Journal of Physical Chemistry B, 2010. **114**(8): p. 2549-2564.
19. Allinger, N.L.; Yuh, Y.H.; Lii, J.H., *Molecular Mechanics - the MM3 Force-Field for Hydrocarbons .1.* J. Am. Chem. Soc., 1989. **111**(23): p. 8551-8566.
20. Wilson, E.B.; Decius, J.C.; Cross, P.C., *Molecular Vibrations: TheTheory of Infrared and Raman Vibrational Spectra*. 1955, New York: McGraw-Hill.
21. Halgren, T.A., *Representation of Vanderwaals (Vdw) Interactions in Molecular Mechanics Force-Fields - Potential Form, Combination Rules, and Vdw Parameters.* J. Am. Chem. Soc., 1992. **114**(20): p. 7827-7843.

22. Thole, B.T., *Molecular Polarizabilities Calculated with a Modified Dipole Interaction.* Chem. Phys., 1981. **59**(3): p. 341-350.
23. Burnham, C.J.; Li, J.C.; Xantheas, S.S.; Leslie, M., *The parametrization of a Thole-type all-atom polarizable water model from first principles and its application to the study of water clusters (n=2-21) and the phonon spectrum of ice Ih.* J. Chem. Phys., 1999. **110**(9): p. 4566-4581.
24. Frisch, M.J.; Trucks, G.W.; Schlegel, H.B.; Scuseria, G.E.; Robb, M.A.; Cheeseman, J.R.; J. A. Montgomery, J.; Vreven, T.; Kudin, K.N.; Burant, J.C.; Millam, J.M.; Iyengar, S.S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G.A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J.E.; Hratchian, H.P.; Cross, J.B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R.E.; Yazyev, O.; Austin, A.J.; Cammi, R.; Pomelli, C.; Ochterski, J.W.; Ayala, P.Y.; Morokuma, K.; Voth, G.A.; Salvador, P.; Dannenberg, J.J.; Zakrzewski, V.G.; Dapprich, S.; Daniels, A.D.; Strain, M.C.; Farkas, O.; Malick, D.K.; Rabuck, A.D.; Raghavachari, K.; Foresman, J.B.; Ortiz, J.V.; Cui, Q.; Baboul, A.G.; Clifford, S.; Cioslowski, J.; Stefanov, B.B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R.L.; Fox, D.J.; Keith, T.; Al-Laham, M.A.; Peng, C.Y.; Nanayakkara, A.; Challacombe, M.; Gill, P.M.W.; Johnson, B.; Chen, W.; Wong, M.W.; Gonzalez, C.; Pople, J.A., *Gaussian 03*. 2004, Gaussian, Inc.: Wallingford, CT.
25. Ren, P.Y.; Ponder, J.W., *Temperature and pressure dependence of the AMOEBA water model.* Journal of Physical Chemistry B, 2004. **108**(35): p. 13427-13437.
26. Dunning, T.H., *Gaussian-Basis Sets for Use in Correlated Molecular Calculations .1. The Atoms Boron through Neon and Hydrogen.* J. Chem. Phys., 1989. **90**(2): p. 1007-1023.
27. Piquemal, J.; Marquez, A.; Parisel, O.; Giessner-Prettre, C., *A CSOV study of the difference between HF and DFT intermolecular interaction energy values: The importance of the charge transfer contribution.* J. Comput. Chem., 2005. **26**(10): p. 1052-1062.
28. Becke, A.D., *Density-Functional Exchange-Energy Approximation with Correct Asymptotic-Behavior.* Phys. Rev. A, 1988. **38**(6): p. 3098-3100.
29. Lee, C.T.; Yang, W.T.; Parr, R.G., *Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron-Density.* Physical Review B, 1988. **37**(2): p. 785-789.
30. Stevens, W.J.; Fink, W.H., *Frozen fragment reduced variational space analysis of hydrogen-bonding interactions - Application to the water dimer.* Chem. Phys. Lett., 1987. **139**(1): p. 15-22.
31. Gordon, M.S.; Schmidt, M.W., *Advances in electronic structure theory: GAMESS a decade later*, in *Theory and Applications of Computational Chemistry, the first forty years*, C.E. Dykstra, Frenking, G., Kim, K. S., Scuseria, G. E., Editor. 2005, Elsevier: Amsterdam.

32. Stevens, W.J.; Basch, H.; Krauss, M., *Compact Effective Potentials and Efficient Shared-Exponent Basis-Sets for the 1st-Row and 2nd-Row Atoms.* J. Chem. Phys., 1984. **81**(12): p. 6026-6033.
33. Savin, A.; Nesper, R.; Wengert, S.; Fassler, T.F., *ELF: The electron localization function.* Angewandte Chemie-International Edition in English, 1997. **36**(17): p. 1809-1832.
34. de Courcy, B.; Pedersen, L.G.; Parisel, O.; Gresh, N.; Silvi, B.; Pilme, J.; Piquemal, J.P., *Understanding Selectivity of Hard and Soft Metal Cations within Biological Systems Using the Subvalence Concept. 1. Application to Blood Coagulation: Direct Cation-Protein Electronic Effects versus Indirect Interactions through Water Networks.* J. Chem. Theory Comput., 2010. **6**(4): p. 1048-1063.
35. Piquemal, J.P.; Pilme, J.; Parisel, O.; Gerard, H.; Fourre, I.; Berges, J.; Gourlaouen, C.; De La Lande, A.; Van Severen, M.C.; Silvi, B., *What can be learnt on biologically relevant systems from the topological analysis of the electron localization function?* Int. J. Quantum Chem., 2008. **108**(11): p. 1951-1969.
36. Pilme, J.; Piquemal, J.P., *Advancing beyond charge analysis using the electronic localization function: Chemically intuitive distribution of electrostatic moments.* J. Comput. Chem., 2008. **29**(9): p. 1440-1449.
37. Noury, S.; Krokidis, X.; Fuster, F.; Silvi, B., *Computational tools for the electron localization function topological analysis.* Comput. Chem., 1999. **23**(6): p. 597-604.
38. De Courcy, B.; Gresh, N.; Piquemal, J.P., *Importance of lone pair interactions/redistribution in hard and soft ligands within the active site of alcohol dehydrogenase Zn-metalloenzyme: Insights from electron localization function.* Interdisciplinary Sciences: Computational Life Sciences, 2009. **1**(1): p. 55-60.
39. Antony, J.; Piquemal, J.P.; Gresh, N., *Complexes of thiomandelate and captopril mercaptocarboxylate inhibitors to metallo-beta-lactamase by polarizable molecular mechanics. Validation on model binding sites by quantum chemistry.* J. Comput. Chem., 2005. **26**(11): p. 1131-1147.
40. Ponder, J., *TINKER: Software Tools for Molecular Design*. 2009: Saint Louis, MO.
41. Sagui, C.; Pedersen, L.G.; Darden, T.A., *Towards an accurate representation of electrostatics in classical force fields: Efficient implementation of multipolar interactions in biomolecular simulations.* J. Chem. Phys., 2004. **120**(1): p. 73-87.
42. Berendsen, H.J.C.; Postma, J.P.M.; Vangunsteren, W.F.; Dinola, A.; Haak, J.R., *Molecular-Dynamics with Coupling to an External Bath.* J. Chem. Phys., 1984. **81**(8): p. 3684-3690.
43. Bennett, C.H., *Efficient Estimation of Free-Energy Differences from Monte-Carlo Data.* J. Comput. Phys., 1976. **22**(2): p. 245-268.

44. Shirts, M.R.; Bair, E.; Hooker, G.; Pande, V.S., *Equilibrium free energies from nonequilibrium measurements using maximum-likelihood methods.* Phys. Rev. Lett., 2003. **91**(14): p. 140601-1-4.
45. Gresh, N., *Energetics of Zn2+ Binding to a Series of Biologically Relevant Ligands - a Molecular Mechanics Investigation Grounded on Ab-Initio Self-Consistent-Field Supermolecular Computations.* J. Comput. Chem., 1995. **16**(7): p. 856-882.
46. Tiraboschi, G.; Gresh, N.; Giessner-Prettre, C.; Pedersen, L.G.; Deerfield, D.W., *Parallel ab initio and molecular mechanics investigation of polycoordinated Zn(II) complexes with model hard and soft ligands: Variations of binding energy and of its components with number and charges of ligands.* J. Comput. Chem., 2000. **21**(12): p. 1011-1039.
47. Stone, A.J., *The Theory of Intermolecular Forces*. 1997, USA: Oxford University Press.
48. Zhao, Z.; Rogers, D.M.; Beck, T.L., *Polarization and charge transfer in the hydration of chloride ions.* J. Chem. Phys., 2010. **132**(1): p. 014502.
49. Cisneros, G.A.; Darden, T.A.; Gresh, N.; Pilmé, J.; Reinhardt, P.; Parisel, O.; Piquemal, J.P., *Design Of Next Generation Force Fields From AB Initio Computations: Beyond Point Charges Electrostatics*, in *Multi-scale Quantum Models for Biocatalysis*. 2009, Springer Netherlands. p. 137-172.
50. Reinhardt, P.; Piquemal, J.; Savin, A., *Fragment-Localized Kohn-Sham Orbitals via a Singles Configuration-interaction Procedure and Application to Local Properties and Intermolecular Energy Decomposition Analysis.* J. Chem. Theory Comput., 2008. **4**(12): p. 2020-2029.
51. Rayon, V.M.; Valdes, H.; Diaz, N.; Suarez, D., *Monoligand Zn(II) complexes: Ab initio benchmark calculations and comparison with density functional theory methodologies.* J. Chem. Theory Comput., 2008. **4**(2): p. 243-256.
52. Marcus, Y., *A Simple Empirical-Model Describing the Thermodynamics of Hydration of Ions of Widely Varying Charges, Sizes, and Shapes.* Biophys. Chem., 1994. **51**(2-3): p. 111-127.
53. Schmid, R.; Miah, A.M.; Sapunov, V.N., *A new table of the thermodynamic quantities of ionic hydration: values and some applications (enthalpy-entropy compensation and Born radii).* Phys. Chem. Chem. Phys., 2000. **2**(1): p. 97-102.
54. Asthagiri, D.; Pratt, L.R.; Paulaitis, M.E.; Rempe, S.B., *Hydration structure and free energy of biomolecularly specific aqueous dications, including Zn2+ and first transition row metals.* J. Am. Chem. Soc., 2004. **126**(4): p. 1285-1289.
55. Grossfield, A.; Ren, P.Y.; Ponder, J.W., *Ion solvation thermodynamics from simulation with a polarizable force field.* J. Am. Chem. Soc., 2003. **125**(50): p. 15671-15682.
56. Mohammed, A.M.; Loeffler, H.H.; Inada, Y.; Tanada, K.-i.; Funahashi, S., *Quantum mechanical/molecular mechanical molecular dynamic simulation of zinc(II) ion in water.* J. Mol. Liq., 2005. **119**(1-3): p. 55-62.

57. D'Angelo, P.; Barone, V.; Chillemi, G.; Sanna, N.; Meyer-Klaucke, W.; Pavel, N.V., *Hydrogen and Higher Shell Contributions in Zn2+, Ni2+, and Co2+ Aqueous Solutions: An X-ray Absorption Fine Structure and Molecular Dynamics Study*. J. Am. Chem. Soc., 2002. **124**(9): p. 1958-1967.
58. Obst, S.; Bradaczek, H., *Molecular dynamics simulations of zinc ions in water using CHARMM.* Journal of Molecular Modeling, 1997. **3**(6): p. 224-232.
59. Kuzmin, A.; Obst, S.; Purans, J., *X-ray absorption spectroscopy and molecular dynamics studies of Zn2+ hydration in aqueous solutions.* J. Phys.: Condens. Matter, 1997. **9**(46): p. 10065-10078.
60. Marini, G.W.; Texler, N.R.; Rode, B.M., *Monte Carlo simulations of Zn(II) in water including three-body effects.* J. Phys. Chem., 1996. **100**(16): p. 6808-6813.
61. Yongyai, Y.P.; Kokpol, S.; Rode, B.M., *Zinc Ion in Water - Intermolecular Potential with Approximate 3-Body Correction and Monte-Carlo Simulation.* Chem. Phys., 1991. **156**(3): p. 403-412.
62. Born, M., *Volumen und Hydratationswärme der Ionen.* Zeitschrift für Physik A Hadrons and Nuclei, 1920. **1**(1): p. 45-48.
63. Garde, S.; Hummer, G.; Paulaitis, M.E., *Free energy of hydration of a molecular ionic solute: Tetramethylammonium ion.* J. Chem. Phys., 1998. **108**(4): p. 1552-1561.
64. Rajamani, S.; Ghosh, T.; Garde, S., *Size dependent ion hydration, its asymmetry, and convergence to macroscopic behavior.* J. Chem. Phys., 2004. **120**(9): p. 4457-4466.
65. Naor, M.M.; Van Nostrand, K.; Dellago, C., *Car-Parrinello molecular dynamics simulation of the calcium ion in liquid water.* Chem. Phys. Lett., 2003. **369**(1-2): p. 159-164.
66. Badyal, Y.S.; Barnes, A.C.; Cuello, G.J.; Simonson, J.M., *Understanding the effects of concentration on the solvation structure of Ca2+ in aqueous solutions. II: Insights into longer range order from neutron diffraction isotope substitution.* J. Phys. Chem. A, 2004. **108**(52): p. 11819-11827.
67. Jalilehvand, F.; Spangberg, D.; Lindqvist-Reis, P.; Hermansson, K.; Persson, I.; Sandstrom, M., *Hydration of the calcium ion. An EXAFS, large-angle X-ray scattering, and molecular dynamics simulation study.* J. Am. Chem. Soc., 2001. **123**(3): p. 431-441.
68. Salmon, P.S.; Bellissentfunel, M.C.; Herdman, G.J., *The Dynamics of Aqueous Zn-2+ Solutions - a Study Using Incoherent Quasi-Elastic Neutron-Scattering.* Journal of Physics-Condensed Matter, 1990. **2**(18): p. 4297-4309.
69. Neely, J.; Connick, R., *Rate of water exchange from hydrated magnesium ion.* J. Am. Chem. Soc., 1970. **92**(11): p. 3476-3478.
70. Friedman, H., *Hydration complexes - some firm results and some pressing questions.* Chemica Scripta, 1985. **25**(1): p. 42-48.
71. Ohtaki, H.; Radnai, T., *Structure and dynamics of hydrated ions.* Chem. Rev. (Washington, DC, U. S.), 1993. **93**(3): p. 1157-1204.

72. Helm, L.; Merbach, A.E., *Water exchange on metal ions: experiments and simulations.* Coord. Chem. Rev., 1999. **187**: p. 151-181.
73. Caminiti, R.; Licheri, G.; Piccaluga, G.; Pinna, G., *X-ray-diffraction study of a 3-ion aqueous-solution.* Chem. Phys. Lett., 1977. **47**(2): p. 275-278.
74. Lightstone, F.C.; Schwegler, E.; Allesch, M.; Gygi, F.; Galli, G., *A first-principles molecular dynamics study of calcium in water.* ChemPhysChem, 2005. **6**(9): p. 1745-1749.
75. Fatmi, M.Q.; Hofer, T.S.; Randolf, B.R.; Rode, B.M., *An extended ab initio QM/MM MD approach to structure and dynamics of Zn(II) in aqueous solution.* J. Chem. Phys., 2005. **123**(5): p. 054514-8.
76. Fatmi, M.Q.; Hofer, T.S.; Randolf, B.R.; Rode, B.M., *Temperature Effects on the Structural and Dynamical Properties of the Zn(II)−Water Complex in Aqueous Solution: A QM/MM Molecular Dynamics Study.* J. Phys. Chem. B, 2006. **110**(1): p. 616-621.

3 Automation of AMOEBA Polarizable Force Field Parameterization for Small Molecules

3.1 INTRODUCTION

While small molecule parameters can be obtained relatively easily for fixed charge force fields using tools such as ANTECHAMBER and eLBOW[1, 2], the assignment of higher order multipole moments has traditionally been performed manually and in an ad-hoc manner. The lack of automation of the parameterization process has hindered the common adoption of multipole-based force fields. In addition, the choice of a local frame is required when an anisotropic electrostatic term is used. The assignment of atomic polarizabilities is necessary as well. This chapter articulates a procedure to generate the AMOEBA force field parameters for small molecules. A utility, POLTYPE, has been implemented to fully automate this procedure and is available at http://water.bme.utexas.edu/wiki/index.php/Software:Poltype as previously reported [3]. The parameters obtained from this procedure are substantiated via comparisons with

quantum mechanics calculations, other molecular mechanics simulations, and experimental measurements for a range of properties.

3.2 METHODS

3.2.1 Protocol

Given the structure, net charge, and multiplicity of a molecule, all parameters can be systematically determined. The AMOEBA force field requires parameters for atomic charges, dipoles, quadrupoles, polarizabilities, damping coefficients (for high valence ions only), van der Waals diameters, and well-depths. Valence parameters include force constants and equilibrium values for bond lengths, angles, and torsion force constants of up to 6-fold. **Figure 3.1** depicts an overview of the parameterization process. This procedure is implemented by the POLTYPE polarizable atomic typing utility.

Prior to parameterization, the *molecule with coordinates* (**Figure 3.1**) are needed. Rotatable bonds are identified about four heavy atoms in which the second and third atoms share a single bond. Bond types can be provided in the structure given to POLTYPE. If not assigned, atom and bond perception is performed by taking advantage of the mechanism offered by The Open Babel Package, version 2.3.0. http://openbabel.sourceforge.net. Symmetric *multipoles are classified* (**Figure 3.1**) based on an iterative algorithm to identify graph invariant indices[4, 5] based on the maximum graph theoretical distance, heavy valence, aromaticity, ring atom, atomic number, heavy bond sum, and formal charge of an atom.

Multipoles are obtained with Stone's distributed multipole analysis[6] and then refined via electrostatic potential fitting. All quantum mechanics (QM) calculations are performed with Gaussian 09[7]. The structure was first optimized at the HF/6-31G* level. The *initial QM single point calculation* (**Figure 3.1**) then computes the electron density

matrix using the MP2/6-311G** level of theory and basis set. Additionally, a grid of electrostatic potentials are populated from *high-level basis set single point calculations* (**Figure 3.1**), which may be either the MP2/6-311++G(2d,2p) or MP2/aug-cc-pVTZ basis sets. The electrostatic potential is computed for a grid around each molecule. Grid points of four shells of increasing distance around a molecule with an offset of 1 Å and 0.35 Å apart was generated. The GDMA program[8] implements distributed multipole analysis[6]. It arranges multipole sites at atomic centers and analytically *assigns initial multipoles* (**Figure 3.1**) based on the density matrix. In GMDA v2.2, ‘‘Switch 0’’ and ‘‘Radius H 0.65’’ were set to access the original DMA procedure. Atomic polarizabilities are assigned based solely on the element type of each atom. Polarization groups are partitioned between rotatable bonds. The *final multipole parameters* (**Figure 3.1**) are further optimized by fitting to electrostatic potentials with a 0.1 kcal mol^{-1} $electron^{-2}$ gradient convergence criteria. When there is intramolecular polarization, the electrostatic potential around the molecule is calculated from the permanent multipoles with full induced-dipoles added. Potential fitting and multipole assignment are currently based on modified utilities available in TINKER 5.1 and standalone versions of POLTYPE will be developed within the Force Field X (FFX) platform available at http://ffx.kenai.com/. In accordance with the solvation study conducted by Shi *et al.*[9], quadrupoles of hydroxyl groups are scaled by 60% after electrostatic potential fitting.

Diameter and well-depth values for van der Waals are assigned based on elements and their valence orbitals. A SMARTS string pattern was used to search for bond orders with its neighbors and assigned after a *database lookup* (**Figure 3.1**). Hydrogen atoms also have a reduction factor that is based on the valence orbital of the atom to which it is bonded. Force constants for bond-length, angle-bend, stretch-bend, out-of-plane bend,

and torsions about non-rotatable bonds are similarly obtained from a *database lookup* (**Figure 3.1**). Equilibrium values are taken from the QM optimized geometry.

Torsional parameters about rotatable bonds (**Figure 3.1**) are obtained by comparing the conformational energy profile calculated from QM with the AMOEBA model that includes electrostatics, vdW, bonds, angles, etc. parameters. A set of torsion parameters is identified by 4 atom classes that surround the rotatable bond and are composed of force constants for each periodicity (1-6). The dihedral angle was scanned by minimizing all torsions about the rotatable bond of interest at 30° intervals with restraints. The 6th order Fourier series is then fit to the difference between the QM conformational energy and AMOEBA's energy without the rotatable –bond torsion term. The QM conformational energy was obtained at the M06L/6-31G** level. Since torsion scanning gives 12 data points, no more than 8 parameters may be used to fit to the conformational energy profile. Torsions about the same central bond that are also in-phase are collapsed in to one set of parameters for the fitting and the contributions are distributed evenly among the parameters. Additionally, if a torsional parameter greater than the difference between the maximum and minimum energy, then that parameter was omitted and the rest of the parameters were fit again. However, if all parameters are removed after the magnitude test, the torsion parameters of only the atoms used to restrain the torsions was fitted. If more than one rotatable bond contains the same classes, the force constants of all classes are averaged.

3.3 RESULTS AND DISCUSSION

3.3.1 Monomeric Comparisons

Quantum mechanics calculations provide molecular properties such as dipole moments, optimized structures, conformational energies, and electrostatic potentials of a

grid around a molecule. The parameters of a diverse set of small organic molecules have been obtained using the POLTYPE polarizable atomic typing utility. A representative set of amino acid side model compounds obtained from the Atlas of Protein Side-Chain Interactions[10, 11] was parameterized. Additionally, the parameters were obtained for a subset of small molecules, which have corresponding experimental hydration free energies[12, 13]. The full listing of dipole moments, electrostatic potential room mean square deviation, optimized structure, and conformations energies computed with POLTYPE/AMOEBA parameters and quantum mechanics has been reported[3]. The molecular dipole moment of the optimized geometry computed using the POLTYPE/AMOEBA parameters compared with quantum mechanics calculations is shown in **Figure 3.2**. The RMS error of all molecular dipoles is 0.16 Debye and the correlation coefficient is 0.998. Molecules with particularly large dipole moment errors are anionic molecules such as CH3S- and C6H5S- with errors of 0.62 and 0.42 Debye, respectively. The molecular dipole moment from quantum mechanics calculations are 6.71 and 3.40 Debye, respectively. Although C6H5S- may not have a large relative error, it poses one of the largest absolute errors. The electrostatic potential (ESP) RMS difference of a grid of point charges around a molecule is 0.16 kcal/mol and the molecules with the largest errors follow the same trend as that for dipoles. The average RMS distance between optimized geometries from POLTYPE/AMOEBA molecules and those optimized from quantum mechanics is 0.08 Å. For conformational energies, the correlation of all conformations prior to torsional fitting about rotatable bonds yielded a 3.6 kcal/mol RMS deviation from QM and a 0.13 correlation coefficient. After fitting, the RMS deviation decreased to 1.24 kcal/mol with a 0.91 correlation coefficient.

3.3.2 Dimer Calculations

The packing of side chains makes significant contribution to protein stability. Investigation of interactions between side chain model compounds is commonly used for evaluating the potential energy models. Typically, fixed charge potential energy models are not able to both gas-phase and solution-phase properties. However, AMOEBA aims to capture the energetics in different environments by including explicit polarization effects. The aforementioned Atlas of Protein Side-Chain Interactions[10, 11] was compiled by clustering interacting side chain pair conformations in 2548 nonhomologous protein structures from the Protein Data Bank. The geometry of the top cluster from the side chain pairs were used for dimer calculations. As the atlas only contains the conformation of heavy atoms, the systems were prepared [14] by adding hydrogen atoms to each model compound and then each pair was optimized at the DFT/TZVP level. Heavy atoms were held fixed during optimization. **Table 3.1** shows the interaction energy of the most common dimer configurations calculated with AMOEBA compared with other QM and molecular mechanics methods. Amino acids are identified by their standard abbreviations. Charged residues that are neutralized have an "(N)" designation. Interactions calculated with CCSD(T)/CBS[15-17] are considered reference energies. The OPLS-AA/L[18] and Amber parm03[19] are energies computed with fixed charge force fields and were taken directly from a study by Berka and coworkers[14]. Note that typically fixed charge force fields use "enhanced" atomic charges for condensed-phase modeling such that comparison with gas-phase QM is not entirely useful. The DFT/TZVP and RI-MP2/aVTZ are quantum mechanics calculations used for comparison.

Overall, the mean relative error (MRE) and maximal relative error (MRX) of interactions energies computed with parameters from POLTYPE for the AMOEBA force field are lower than errors of other force fields as well as DFT, but comparable to RI-

MP2 results. The mean absolute error (MAE), maximal absolute error (MAX), and root mean square error (RMS) are also lower than other molecular mechanics methods but worse than DFT. Interestingly, the MRX of all molecular mechanical methods perform better than DFT, as the latter shows significant "relative" errors for weak associating dimers. Similarly, AMOEBA and the fixed charge force fields yield a lower MAE compared to DFT. However, when an empirical dispersion function was incorporated in the DFT method[20, 21], the interaction energy prediction improves significantly [14]. RI-MP2 performs remarkably well in producing accurate interaction energies when compared to CCSD(T)/CBS.

The charged pairs arginine-aspartate and lysine-glutamate (RD and KE) seems to be the source of the largest absolute error for all molecular mechanics methods. However, the relative error of the RD pair was less than 10% for POLTYPE/AMOEBA and OPLS-AA/L with a larger error for Amber parm03. The relative error for the KE pair was less than 5% for AMOEBA and OPLS force fields. Additionally, a Symmetry Adapted Perturbation Theory (SAPT) decomposition of the KE interaction reveals that higher order energy beyond first-order electrostatics and repulsion and second-order induction and dispersion stabilizes the pair by about 3kcal/mol. Conversely, higher order energy stabilizes the RD pair by more than 6kcal/mol and suggests that the difficulty with this pair may be due to interactions not captured by the energy function of molecular mechanics models.

Pairs with polar residues yield lower absolute error for POLTYPE/AMOEBA. However, it should be noted that the conformation of residues such as aspartic acid may be artificial due to the system preparation described above. Since all geometries chosen for the aspartic acid have C–O bonds lengths in a narrow range between 1.24 – 1.25 Å, this geometry only corresponds to the COO^- carboxylate ion[22, 23]. Typically,

protonated carboxylic acid exhibits assymetric bond-lengths of 1.31 Å and 1.2 Å. Since geometries in the current test set are obtained from PDB structures and then minimized with heavy atoms fixed and hydrogen atoms added, pairs with artificially protonated carboxylic acid such as D(N)H(N) do not accurately describe electron distributions of charged carboxylate nor neutral carboxylic acid.. Reassignment of multipoles with the artificial structure indeed yields an interaction energy of the D(N)H(N) pair that more closely matches the reference energy of the structure. The R(N)D(N) pair exhibits a similar sensitivity to geometry in which an assignment of multipoles with the given structure yields the error in **Table 3.1**, but the assignment of multipoles with a monomer-optimized structure further underestimated the interaction energy by ~3 kcal/mol. These examples suggest that the electron distribution of unphysical structures, particularly protonation states that are incompatible with its heavy atom conformation, cannot be captured by molecular mechanical models including AMOEBA. The parameterization of molecular mechanics models is based on minimum energy structures as it is unlikely for simple classic mechanical model to capture the complete potential energy surface especially when the structures deviate significantly from the local minima and "chemical" changes are involved. Nonetheless, dimer interaction energy calculations provide insight in to the non-covalent interactions of a system and are conducive to the development of a force field. This is particularly true for AMOEBA since polarization allows parameters to be transferable between gas- and condensed-phase without the need to "pre-polarize" and scale up partial charges. Moreover, other workers[24] support the proficiency of AMOEBA in predicting interaction energies of fragment pairs decomposed from the HIV−II protease crystal structure and show improvement over other classical molecular mechanics models.

3.3.3 Solvation

The thermodynamic properties of molecules developed with POLTYPE are studied and compared with experimental values. The parameters of several families of small molecules containing functional groups in drug-like molecules were obtained for AMOEBA using POLTYPE and their hydration free energies (HFE) are computed with the Bennett acceptance Ratio (BAR)[25]. In a similar procedure as a previous AMOEBA HFE study[9], perturbations of the solute required the decoupling of electrostatic and van der Waals interactions. The perturbation of electrostatic atomic multipoles and polarizabilities was scaled down the linearly with λ = (1.0, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, and 0.0). We also scale down the radius and well-depth of vdW interactions linearly with λ = (1.0, 0.9, 0.8, 0.75, 0.7, 0.65, 0.6, 0.5, 0.4, 0.2, and 0.0). Molecular dynamics in solvent were carried out by placing the solute molecule at the origin of a pre-equilibrated a cubic 28.78 Å periodic box containing 800 water molecules. The system was then equilibrated for 50 ps at 298 K. For each perturbation step, 500 ps molecular dynamics simulations were performed with 1 fs time steps and vdW cutoff of 12 Å at 298 K constant temperature using the Berendsen thermostat[26]. The long-range electrostatics for all the systems were treated using Particle Mesh Ewald (PME) summation [27-29]. The atomic coordinates at every 500 fs were used for post-analysis except for first 100 ps simulation. Gas phase simulations were run on the single solute molecule for 50 ps with a time step of 0.1 ps at 298 K using a stochastic thermostat. Atomic coordinates at every 100 fs were used for post-analysis. Previously, Mobley *et al.*[30] conducted a study to compute HFE for a larger set of molecules with the fixed charge general Amber force field (GAFF)[31] by assigning AM1-BCC partial charges[32, 33]. Hydration free energies with the AMOEBA force field, GAFF and experimental results[12, 13] are listed in **Table 3.2**. Included in the table is also the free energy difference observed while

electrostatics and van der Waals interactions are perturbed. The RMS error of HFE with POLTYPE/AMOEBA for the set of molecules in this study is 0.75 kcal/mol. Although previous work with a larger set of molecules with GAFF yielded a lower RMS error, the error for the set in this study is 1.56 kcal/mol. When families of molecules are considered, alkenes have errors of ~0.6 kcal/mol while the errors for GAFF are ~1 kcal/mol. Similarly, HFE predicted by POLTYPE/AMOEBA of nitro-containing molecules consistently yield lower errors than GAFF.

However, GAFF had errors of lower magnitude than POLTYPE/AMOEBA for two alkanes (22-dimethylbutane and n-octane). It should be noted that the free energy differences observed for these molecules due to van der Waals perturbations are consistent between AMOEBA and GAFF. It seems that electrostatics is too "attractive" in AMOEBA. The source of error in AMOEBA electrostatics is likely due to the ESP optimization procedure. For example, in the fitting process of n-octane there are significantly changes in the quadrupoles of the terminal hydrogen atoms, which are shared by 6 atoms. Large deviations from those obtained from DMA may result in unphysical electrostatic parameters. Relaxing the convergence criteria of ESP fitting from 0.1 to 0.5 kcal mol^{-1} $electron^{-2}$ gradient convergence criteria or moving the grid points away from the vdW surface may prevent unphysical multipoles resulting from the optimization. We are currently investigating this procedure especially for large linear molecules. Additionally, care must be taken when defining polarization groups. Some families such as aldehydes cannot be partitioned between carbonyl C=O and its neighboring heaving atom. When the groups are inappropriately partitioned across the bond, errors in HFE prediction increase to 1.53 and 1.55 kcal/mol for isobutyraldehyde and propionaldehyde, respectively.

Additionally, the hydration of ionic molecules is studied. For this preliminary study, a generic scaling factor for formally charged atoms was applied. For hydrogen atoms bonded to atoms with positive formal charge, their vdW diameters were scaled down by 10% from their original parameters. Conversely, the vdW diameters of atoms with a formal negative charge were scaled up by 10%. These scaling methods will be further investigated and refined. Simulations details are the same as those of neutral solutes. It should be noted, though, that difficulties arise in the comparison with experiments of homogenous ions as they are not directly accessible and must be conducted in salt solutions. The contributions of anions and cations then must be determined through various schemes such as self-consistent thermodynamic analysis[34], the TATB assumption[35], or the cluster-pair approximation[36]. The experimental hydration data that we compare with here apply the cluster-pair approximation, which is based on the correlation between ion–water clustering data and aqueous solvation free energies of neutral ion pairs. However, when comparing ion solvation quantities, differences between experimental methods and measurements should be taken in to consideration. Therefore the comparison of an anion-cation salt pair is more appropriate than single ions alone, if the HFE of the pair has been determined in a consistent manner. Experimental hydration free energy of single ions and salt and corresponding energies using POLTYPE/AMOEBA are shown in **Table 3.3**. The salt HFE of an anionic molecule is taken here to be the sum of HFE of the molecule and the sodium cation[37, 38]. Similarly, the salt HFE of a cationic molecule is taken to be the sum of the HFE of the molecule and the chlorine anion [37, 38]. The correlation coefficient between salt HFE obtained from experimental and POLTYPE/AMOEBA is 0.95 and the unsigned mean relative error is less than 3%. Phosphor and sulfur containing cationic molecules produced salt HFE that agreed with well experiment. Some of the largest errors come

from the oxonium cations. As mentioned previously, a simple scaling has been applied to all atoms with a formal negative charge or hydrogen atoms bonded to atoms with a positive charge. Further investigation of accurate ion parameters is required.

3.3.4 Ligand-protein Binding

We also performed free energy calculations with implicit solvent to elucidate the effects of configurational change due to binding that are not addressed properly. The free energy difference of binding between two ligands (L1 and L2) to a common protein is depicted by a thermodynamic cycle [39] and is based on previously reported results[40]. The relative binding energy then is $\Delta\Delta G = \Delta G_{bind,L2} - \Delta G_{bind,L1} = \Delta G_{complex,solv} - \Delta G_{ligand,solv} - \Delta G_{protein,solv}$. The Bennett Acceptance Ratio (BAR) [25, 41], a free energy calculation method that minimizes variance by utilizing forward and reverse perturbations was applied to perturb and calculate the free energy difference between states. Snapshots for free energy calculations were generated with molecular dynamics via the AMOEBA polarizable force field [42]. The electrostatic contribution to solvation energy was calculated implicitly with the polarizable Generalized-Kirkwood model developed by Schnieders *et al* [43]. The nonpolar contributions to solvation energy are composed of cavitation and dispersion terms [44-47]. The former involves the work required to displace the solvent and is a function of the solvent-accessible surface area (SASA). The latter describes the nonpolar dispersion interactions between solvent and solute. Unlike the MM-PBSA method [48], which calculates free energy from end-state simulations, we apply BAR/GK to perturb the system in implicit solvent.

The soft-core buffered 14-7 potential [49] was used to prevent energetic instabilities as annihilated atoms are penetrated by other atoms. This interaction replaces Halgren's buffered 14-7 interaction only for annihilated and non-annihilated atoms.

Although the reaction coordinate in the Halgren potential is sufficient to perturb the vdW interactions, the radii of annihilated atoms still need to be scaled for proper treatment of the effective Born radii. This radii is used for implicit polar and nonpolar implicit solvation. The RATTLE algorithm [50] was used to constrain ideal bond lengths and angles of hydrogen bonds. The time step was 1.5 fs and each simulation was run for 150 ps. Since the configurational degrees of freedom are already incorporated in the nonpolar implicit solvation contribution, less simulation time is required.

Additional studies [39] examined the binding free energy of 4 ligands to benzamidine relative to the binding of trypsin-benzamidine and decomposed the entropy into translational, rotational, and vibrational contributions [51, 52]. The entropy calculated from these terms was added to the relative binding energy obtained from the PMPB/SA (Polarizable Multipole Possion Boltzmann/Surface Area) method. The binding free energy calculated via PMPB/SA with and without entropic terms compared to experimental values [53-58] yielded correlations of 0.667 and 0.708, respectively, and do not have a significant difference. Meanwhile, a correlation of 0.933 and root mean square error of 0.88kcal/mol were achieved using the GK/BAR method when compared with experiment (see **Figure 3.3**). The GK/BAR binding free energies offset by the explicit water BAR calculations of trypsin-benzamidine[59] and MM-PBSA calculated energies are shown in **Figure 3.3**. This suggests that the "slower" perturbation between end-states in the GK/BAR method may be necessary to capture the configurational degrees of, freedeom.

3.4 CONCLUSIONS

A protocol to develop AMOEBA models for small molecules has been established. In this work, we have described a standard approach to generate the

AMOEBA force field for small and drug-like molecules. Although the parameterization process for the AMOEBA polarizable force field requires more sophistication compared to process for fixed-charge force fields, a straightforward procedure is described here. The POLTYPE utility allows one to generate the AMOEBA polarizable force field for a small molecule in a fully automated manner. We have shown good agreement with quantum mechanics calculations in gas phase for monomers and dimers for neutral as well as charged molecules. Although electrostatic parameters for traditional fixed-charge force fields need to be "pre-polarized" due to their lack of ability to modify the electron distribution of a molecule, this precludes an accurate model where the electric field undergoes significant changes. Alternatively, POLTYPE obtains AMOEBA parameters in gas-phase, which are transferred directly to liquid-phase systems without modification. The hydration free energy (HFE) of neutral and charged molecules have been calculated with the Bennett Acceptance Ratio and compared with experimental values. The RMS error for the HFE of neutral molecules is less than 1 kcal/mol, while the unsigned mean relative error is less than 3% and a correlation coefficient of 0.95 for the HFE of salts containing charged molecules. Although some assignments such as the van der Waals diameters of atoms with formal charge of ionic molecules need further investigation, POLTYPE readily facilitates the systematic study of these chemical functional groups. We believe the advantage of polarizable force fields such as AMOEBA will be further illustrated in processes where significant environmental changes (i.e. electric field) are involved. Additionally, as more powerful computing paradigms develop[60], including those that may seem inaccessible in the near future[61], the applicability of existing models will expand. An implementation of the AMOEBA force field, for example, has taken advantage of such developments [62] as the model becomes more widely adopted and a strong supporting community grows.

Table 3.1: Interaction energies (kcal/mol) for amino acid pairs calculated using several approaches in gas phase. Interaction energies computed with the CCSD(T) complete basis set (CBS) is used as the reference method. Interaction energies calculated with the AMOEBA force field parameterized with POLTYPE are performed as a part of this work. The DFT method was carried out with the TPSS functional and TZVP basis. The aug-cc-pVTZ basis set and resolution of identity approximation was used for the MP2 method. MRE is the unsigned mean relative error (%), MRX is the signed maximal relative error (%), MAE is the unsigned mean absolute error, MAX is the signed maximal absolute error, and RMS is the signed root mean square error.

Dimer[a]	CCSD(T) CBS[b]	POLTYPE AMOEBA	OPLS AA/L[c]	parm03[d]	DFT TZVP	RI-MP2 aVTZ
RD	-110.80	-100.33	-105.71	-90.37	-110.60	-110.21
KE	-108.40	-104.86	-106.02	-103.57	-108.27	-107.75
DH(N)	-30.64	-28.15	-12.20	-22.36	-28.83	-30.91
D(N)H(N)	-17.97	-15.10	-10.90	-7.80	-16.26	-17.94
R(N)D(N)	-16.32	-12.38	-8.94	---	-14.71	-15.92
K(N)E(N)	-10.76	-9.54	-8.80	-9.11	-9.81	-10.65
QN	-7.37	-4.86	-8.61	-8.84	-5.66	-6.92
TT	-6.50	-7.99	-7.96	-6.83	-4.81	-6.28
YY	-4.66	-5.41	-3.84	-3.62	1.35	-5.51
TS	-4.50	-5.15	-4.38	-4.40	-3.36	-4.30
LW	-4.04	-4.16	-3.46	-3.46	1.00	-4.74
YP	-3.79	-3.83	-3.05	-3.09	0.44	-4.11
FF	-2.33	-2.41	-1.97	-2.26	1.11	-3.04
MM	-2.03	-1.95	-3.14	-2.35	1.22	-2.01
LY	-1.72	-1.72	-1.86	-1.52	0.96	-1.66
LL	-1.62	-1.60	-1.40	-1.66	0.00	-1.60
MC	-1.46	-1.28	-2.01	-1.20	0.25	-1.43
VV	-1.39	-1.52	-1.36	-1.43	0.44	-1.28
IL	-1.39	-1.36	-1.19	-1.41	0.06	-1.35
II	-1.24	-1.22	-1.13	-1.20	0.62	-1.11
LT	-1.09	-1.11	-0.91	-1.05	0.02	-1.02
VL	-1.08	-1.06	-0.81	-1.11	0.11	-1.01
AL	-1.07	-1.11	-1.00	-0.94	0.71	-0.93
LG	-0.77	-0.75	-0.75	-0.53	-0.09	-0.71
	MRE [%]	8.69	19.54	13.55	83.61	6.52
	MRX [%]	34.01	60.19	56.58	166.28	-30.62
	MAE	1.28	2.11	2.22	2.03	0.26
	MAX	10.47	18.44	20.43	6.01	-0.85
	RMS	2.61	4.16	4.78	1.4	0.36

[a] Other than POLTYPE/AMOEBA, interaction energies were calculated by (Berka, 2009)
[b] Reference calculation (Tsuzuki, 2005; Sinnokrot, 2004; Hobza, 1999)
[c] Interaction energy computed using OPLS-AA/L force field (Kaminski, 2001)
[d] Interaction energy computed using parm03 force field (Duan, 2003)

Table 3.2: Hydration free energies (kcal/mol) of small molecules obtained from experiment, POLTYPE/AMOEBA, and general Amber force field (GAFF).

	Exp[a]	POLTYPE/AMOEBA				GAFF[b]			
Molecule Name	ΔG_{exp}	ΔG_{ele}	ΔG_{vdw}	ΔG_{AMOEBA}	Error $_{AMOEBA}$	ΔG_{ele}	ΔG_{vdw}	ΔG_{GAFF}	Error $_{GAFF}$
2 methylbut 2 ene	**1.31**	-1.78	2.51	**0.72**	-0.59	-0.55	2.83	**2.28**	0.97
but 1 ene	**1.38**	-1.65	2.48	**0.83**	-0.55	-0.37	2.85	**2.48**	1.10
1 nitrobutane	**-3.09**	-4.50	1.95	**-2.55**	0.54	-2.43	0.92	**-1.51**	1.58
2 nitrophenol	**-4.58**	-5.43	1.24	**-4.19**	0.39	-5.40	0.06	**-5.34**	-0.76
4 nitrophenol	**-10.64**	-11.09	1.49	**-9.61**	1.03	-8.04	-0.18	**-8.22**	2.42
22 dimethylbutane	**2.51**	-0.75	2.44	**1.70**	-0.81	0.01	2.52	**2.53**	0.02
n octane	**2.88**	-1.37	3.11	**1.74**	-1.14	0.01	3.12	**3.13**	0.25
23 dimethylphenol	**-6.16**	-7.24	2.28	**-4.96**	1.20	-6.49	1.82	**-4.67**	1.49
tert butylbenzene	**-0.44**	-3.54	2.21	**-1.33**	-0.89	-2.98	2.56	**-0.42**	0.02
3 chloropyridine	**-4.01**	-5.42	1.40	**-4.02**	-0.01	-3.78	1.28	**-2.50**	1.51
di n butylamine	**-3.24**	-6.86	3.28	**-3.58**	-0.34	-4.71	3.08	**-1.63**	1.61
di n propyl ether	**-1.16**	-5.41	2.71	**-2.70**	-1.54	-2.67	2.88	**0.21**	1.37
methyl isopropyl ether	**-2.01**	-5.37	2.62	**-2.75**	-0.74	-2.89	2.14	**-0.75**	1.26
di n propyl sulfide	**-1.28**	-4.16	2.35	**-1.81**	-0.53	-2.15	2.64	**0.49**	1.77
dimethyl disulfide	**-1.83**	-3.22	1.99	**-1.23**	0.60	-0.72	2.20	**1.48**	3.31
isobutyraldehyde	**-2.86**	-5.29	2.17	**-3.12**	-0.26	-4.98	2.05	**-2.93**	-0.07
propionaldehyde	**-3.43**	-5.26	1.85	**-3.41**	0.02	-5.06	1.98	**-3.08**	0.35
methanethiol	**-1.24**	-2.85	2.02	**-0.83**	0.41	-2.25	1.99	**-0.26**	0.98
n butanethiol	**-0.99**	-3.77	2.10	**-1.68**	-0.69	-2.39	2.27	**-0.12**	0.87
methyl acetate	**-3.13**	-6.20	2.01	**-4.19**	-1.06	-5.44	1.71	**-3.73**	-0.60
oct 1 yne	**0.71**	-2.40	2.94	**0.54**	-0.17	-0.83	3.29	**2.46**	1.75
pent 1 yne	**0.01**	-2.37	2.39	**0.02**	0.01	-0.81	2.74	**1.93**	1.92
octan 1 ol	**-4.09**	-7.87	2.68	**-5.20**	-1.11	-5.13	2.48	**-2.65**	1.44
p dibromobenzene	**-2.30**	-3.7	2.22	**-1.48**	0.82	-1.70	1.69	**-0.01**	2.29
tribromomethane	**-2.13**	-3.82	2.45	**-1.37**	0.76	-0.70	1.58	**0.88**	3.01

[a] Experimental HFE (Abraham *et al.*, 1990 and Chambers *et al.*, 1996)
[b] HFE calculated from General Amber force field (Mobley, 2009 and Wang, 2004)

Table 3.3 Hydration free energy (kcal/mol) of ionic molecules and their corresponding salt.

Molecule	ΔG_{exp}	$\Delta G_{salt\ exp}$	ΔG_{AMOEBA}	$\Delta G_{salt\ AMOEBA}$
$(CH3)2PH^{2+}$	-57	-131.5	-47.87	-134.37
CH3PH3+	-63	-137.5	-51.80	-138.30
(CH3)2SH+	-64.5	-139	-51.53	-138.03
$CH3SH^{2+}$	-74	-148.5	-57.54	-144.04
CH3NH3+	-76.4	-150.9	-68.16	-154.66
$(CH3)2NH^{2+}$	-68.6	-143.1	-63.01	-149.51
$HC(OH)NH^{2+}$	-78	-152.5	-67.31	-153.81
C6H5NH3+	-72.4	-146.9	-63.15	-149.65
C5H5NH+	-58	-132.5	-52.19	-138.69
imidazoleH+	-64	-138.5	-54.45	-140.95
CH3C(OH)CH3+	-77.1	-151.6	-53.57	-140.07
(CH3)2OH+	-79.7	-154.2	-59.01	-145.51
$CH3CH2OH^{2+}$	-88.4	-162.9	-68.32	-154.82
$CH3OH^{2+}$	-93	-167.5	-73.75	-160.25
CH3O-	-95	-198.2	-105.67	-197.47
HCO2-	-76.2	-179.4	-91.27	-183.07
CH3S-	-73.8	-177	-85.13	-176.93
C6H5S-	-63.4	-166.6	-67.25	-159.05

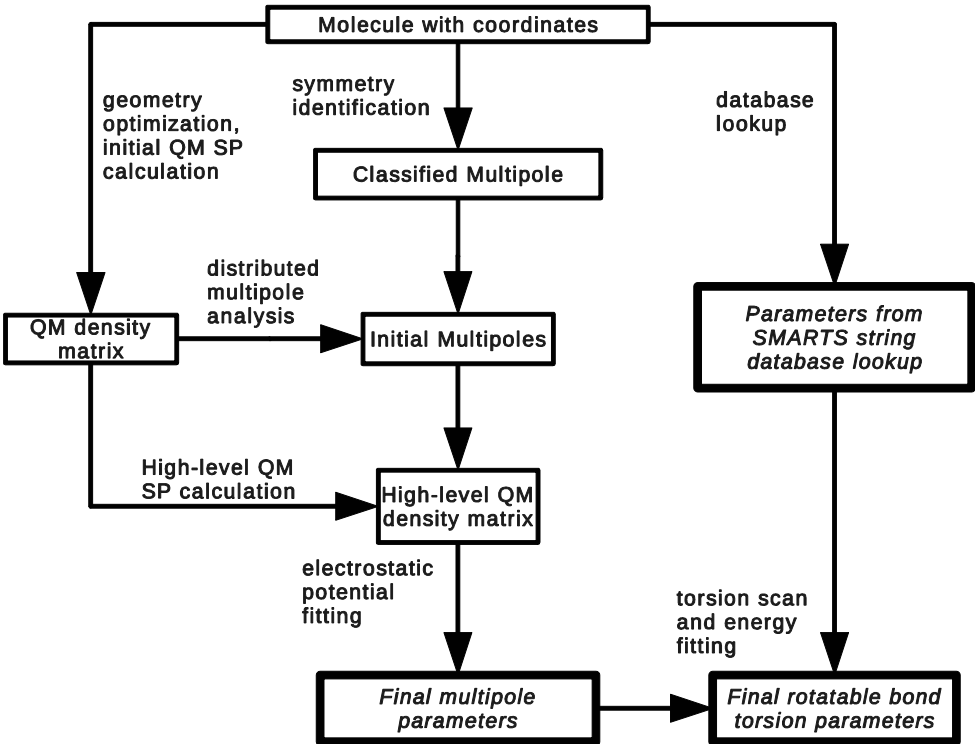


Figure 3.1: Overview of the parameterization procedure for POLTYPE.

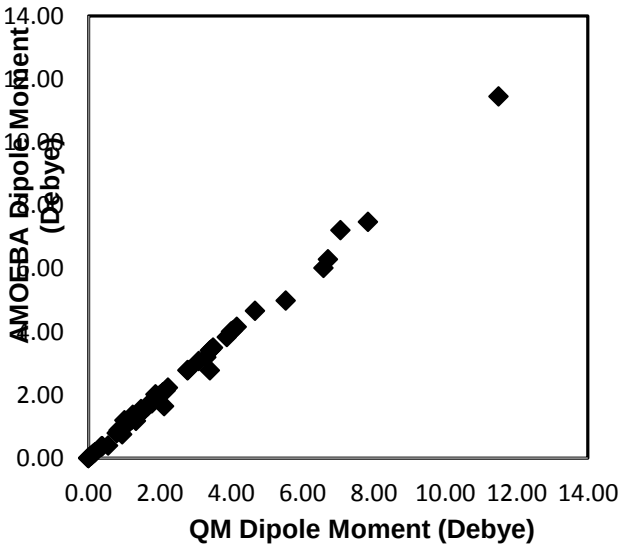


Figure 3.2: Molecular dipole moment computed from AMOEBA parameters and quantum mechanical calculations.

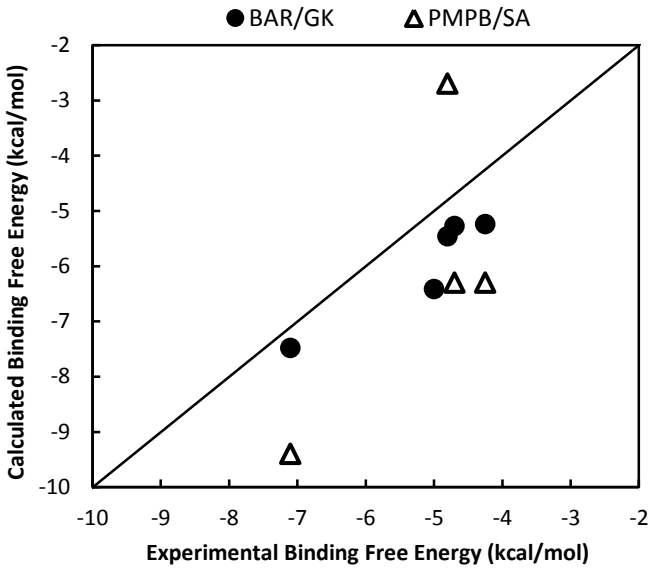


Figure 3.3: Comparison of experimental and calculated binding free energies from BAR/GK and PMPB/SA calculations.

3.5 REFERENCES

1. Wang, J.M.; Wang, W.; Kollman, P.A.; Case, D.A., *Automatic atom type and bond type perception in molecular mechanical calculations.* J. Mol. Graph. Model., 2006. **25**(2): p. 247-260.
2. Moriarty, N.W.; Grosse-Kunstleve, R.W.; Adams, P.D., *electronic Ligand Builder and Optimization Workbench (eLBOW): a tool for ligand coordinate and restraint generation.* Acta Crystallographica Section D-Biological Crystallography, 2009. **65**: p. 1074-1080.
3. Wu, J.C.; Chattree, G.; Ren, P., *Automation of AMOEBA polarizable force field parameterization for small molecules.* Theoretical Chemistry Accounts, In press.
4. Weininger, D., *Smiles, a Chemical Language and Information-System .1. Introduction to Methodology and Encoding Rules.* J. Chem. Inf. Comput. Sci., 1988. **28**(1): p. 31-36.
5. Morgan, H.L., *The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service.* Journal of Chemical Documentation, 1965. **5**(2): p. 107-113.
6. Stone, A.J.; Alderton, M., *Distributed Multipole Analysis - Methods and Applications.* Mol. Phys., 1985. **56**(5): p. 1047-1064.
7. Frisch, M.J.; Trucks, G.W.; Schlegel, H.B.; Scuseria, G.E.; Robb, M.A.; Cheeseman, J.R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G.A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H.P.; Izmaylov, A.F.; Bloino, J.; G.

Zheng; Sonnenberg, J.L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; M. Ishida, T.N.; Y. Honda; O. Kitao, H.N.; Vreven, T.; J. A. Montgomery, J.; Peralta, J.E.; Ogliaro, F.; Bearpark, M.; J. J. Heyd, E.B.; Kudin, K.N.; Staroverov, V.N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J.C.; Iyengar, S.S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J.M.; Klene, M.; Knox, J.E.; Cross, J.B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R.E.; Yazyev, O.; Austin, A.J.; Cammi, R.; Pomelli, C.; Ochterski, J.W.; Martin, R.L.; Morokuma, K.; Zakrzewski, V.G.; Voth, G.A.; Salvador, P.; Dannenberg, J.J.; Dapprich, S.; A. D. Daniels; Farkas, Ö.; Foresman, J.B.; Ortiz, J.V.; Cioslowski, J.; Fox, D.J., *Gaussian 09*. 2009, Gaussian, Inc.: Wallingford CT.

8. Stone, A.J., *Distributed multipole analysis: Stability for large basis sets.* J. Chem. Theory Comput., 2005. **1**(6): p. 1128-1132.
9. Shi, Y.; Wu, C.J.; Ponder, J.W.; Ren, P.Y., *Multipole Electrostatics in Hydration Free Energy Calculations.* J. Comput. Chem., 2011. **32**(5): p. 967-977.
10. Singh, J.; Thornton, J.M., *Atlas of Protein Side-Chain Interactions*. Vol. I & II. 1992, Oxford: IRL press.
11. Thornton, J.M. 2009; Available from: http://www.ebi.ac.uk/thornton-srv/databases/sidechains/.
12. Abraham, M.H.; Whiting, G.S.; Fuchs, R.; Chambers, E.J., *Thermodynamics of Solute Transfer from Water to Hexadecane.* J. Chem. Soc., Perkin Trans. 2, 1990(2): p. 291-300.
13. Chambers, C.C.; Hawkins, G.D.; Cramer, C.J.; Truhlar, D.G., *Model for aqueous solvation based on class IV atomic charges and first solvation shell effects.* J. Phys. Chem., 1996. **100**(40): p. 16385-16398.
14. Berka, K.; Laskowski, R.; Riley, K.E.; Hobza, P.; Vondrasek, J., *Representative Amino Acid Side Chain Interactions in Proteins. A Comparison of Highly Accurate Correlated ab Initio Quantum Chemical and Empirical Potential Procedures.* J. Chem. Theory Comput., 2009. **5**(4): p. 982-992.
15. Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M., *Ab initio calculations of structures and interaction energies of toluene dimers including CCSD(T) level electron correlation correction.* J. Chem. Phys., 2005. **122**(14): p. 144323.
16. Sinnokrot, M.O.; Sherrill, C.D., *Highly accurate coupled cluster potential energy curves for the benzene dimer: Sandwich, T-shaped, and parallel-displaced configurations.* J. Phys. Chem. A, 2004. **108**(46): p. 10200-10207.
17. Hobza, P.; Sponer, J., *Structure, energetics, and dynamics of the nucleic Acid base pairs: nonempirical ab initio calculations.* Chem Rev, 1999. **99**(11): p. 3247-76.
18. Kaminski, G.A.; Friesner, R.A.; Tirado-Rives, J.; Jorgensen, W.L., *Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides.* J. Phys. Chem. B, 2001. **105**(28): p. 6474-6487.

19. Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M.C.; Xiong, G.M.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.M.; Kollman, P., *A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations.* J. Comput. Chem., 2003. **24**(16): p. 1999-2012.
20. Rapcewicz, K.; Ashcroft, N.W., *Fluctuation Attraction in Condensed Matter - a Nonlocal Functional-Approach.* Physical Review B, 1991. **44**(8): p. 4032-4035.
21. Andersson, Y.; Hult, E.; Apell, P.; Langreth, D.C.; Lundqvist, B.I., *Density-functional account of van der Waals forces between parallel surfaces.* Solid State Commun., 1998. **106**(5): p. 235-238.
22. Ahmed, H.U.; Blakeley, M.P.; Cianci, M.; Cruickshank, D.W.J.; Hubbard, J.A.; Helliwell, J.R., *The determination of protonation states in proteins.* Acta Crystallographica Section D-Biological Crystallography, 2007. **63**: p. 906-922.
23. Fenn, T.D.; Schnieders, M.J.; Brunger, A.T.; Pande, V.S., *Polarizable Atomic Multipole X-Ray Refinement: Hydration Geometry and Application to Macromolecules.* Biophys. J., 2010. **98**(12): p. 2984-2992.
24. Faver, J.C.; Benson, M.L.; He, X.A.; Roberts, B.P.; Wang, B.; Marshall, M.S.; Kennedy, M.R.; Sherrill, C.D.; Merz, K.M., *Formal Estimation of Errors in Computed Absolute Interaction Energies of Protein-Ligand Complexes.* J. Chem. Theory Comput., 2011. **7**(3): p. 790-797.
25. Bennett, C.H., *Efficient Estimation of Free-Energy Differences from Monte-Carlo Data.* J. Comput. Phys., 1976. **22**(2): p. 245-268.
26. Berendsen, H.J.C.; Postma, J.P.M.; Vangunsteren, W.F.; Dinola, A.; Haak, J.R., *Molecular-Dynamics with Coupling to an External Bath.* J. Chem. Phys., 1984. **81**(8): p. 3684-3690.
27. Essmann, U.; Perera, L.; Berkowitz, M.L.; Darden, T.; Lee, H.; Pedersen, L.G., *A Smooth Particle Mesh Ewald Method.* J. Chem. Phys., 1995. **103**(19): p. 8577-8593.
28. Darden, T.; York, D.; Pedersen, L., *Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems.* J. Chem. Phys., 1993. **98**(12): p. 10089-10092.
29. Sagui, C.; Darden, T.A., *Molecular dynamics simulations of biomolecules: Long-range electrostatic effects.* Annu. Rev. Biophys. Biomol. Struct., 1999. **28**: p. 155-179.
30. Mobley, D.L.; Bayly, C.I.; Cooper, M.D.; Shirts, M.R.; Dill, K.A., *Small Molecule Hydration Free Energies in Explicit Solvent: An Extensive Test of Fixed-Charge Atomistic Simulations.* J. Chem. Theory Comput., 2009. **5**(2): p. 350-358.
31. Wang, J.M.; Wolf, R.M.; Caldwell, J.W.; Kollman, P.A.; Case, D.A., *Development and testing of a general amber force field.* J. Comput. Chem., 2004. **25**(9): p. 1157-1174.
32. Jakalian, A.; Bush, B.L.; Jack, D.B.; Bayly, C.I., *Fast, efficient generation of high-quality atomic Charges. AM1-BCC model: I. Method.* J. Comput. Chem., 2000. **21**(2): p. 132-146.

33. Jakalian, A.; Jack, D.B.; Bayly, C.I., *Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation.* J. Comput. Chem., 2002. **23**(16): p. 1623-1641.
34. Schmid, R.; Miah, A.M.; Sapunov, V.N., *A new table of the thermodynamic quantities of ionic hydration: values and some applications (enthalpy-entropy compensation and Born radii).* Phys. Chem. Chem. Phys., 2000. **2**(1): p. 97-102.
35. Krishnan, C.V.; Friedman, H.L., *Solvation Enthalpies of Various Ions in Water and Heavy Water.* J. Phys. Chem., 1970. **74**(11): p. 2356.
36. Tissandier, M.D.; Cowen, K.A.; Feng, W.Y.; Gundlach, E.; Cohen, M.H.; Earhart, A.D.; Tuttle, T.R.; Coe, J.V., *The proton's absolute aqueous enthalpy and Gibbs free energy of solvation from cluster ion solvation data (vol 102A, pg 7791, 1998).* J. Phys. Chem. A, 1998. **102**(46): p. 9308-9308.
37. Grossfield, A.; Ren, P.Y.; Ponder, J.W., *Ion solvation thermodynamics from simulation with a polarizable force field.* J. Am. Chem. Soc., 2003. **125**(50): p. 15671-15682.
38. Kelly, C.P.; Cramer, C.J.; Truhlar, D.G., *Aqueous solvation free energies of ions and ion-water clusters based on an accurate value for the absolute aqueous solvation free energy of the proton.* J. Phys. Chem. B, 2006. **110**(32): p. 16066-16081.
39. Jiao, D.; Zhang, J.; Duke, R.E.; Li, G.; Schnieders, M.J.; Ren, P., *Trypsin-ligand binding free energies from explicit and implicit solvent simulations with polarizable potential.* J. Comput. Chem., 2009. **30**(11): p. 1701-11.
40. Yang, T.Y.; Wu, J.C.; Yan, C.L.; Wang, Y.F.; Luo, R.; Gonzales, M.B.; Dalby, K.N.; Ren, P.Y., *Virtual screening using molecular simulations.* Proteins-Structure Function and Bioinformatics, 2011. **79**(6): p. 1940-1951.
41. Shirts, M.R.; Bair, E.; Hooker, G.; Pande, V.S., *Equilibrium Free Energies from Nonequilibrium Measurements Using Maximum-Likelihood Methods.* Phys. Rev. Lett., 2003. **91**: p. 140601.
42. Ren, P.; Ponder, J.W., *Consistent treatment of inter- and intramolecular polarization in molecular mechanics calculations.* J. Comput. Chem., 2002. **23**(16): p. 1497-506.
43. Schnieders, M.J.; Ponder, J.W., *Polarizable Atomic Multipole Solutes in a Generalized Kirkwood Continuum.* J. Chem. Theory Comput., 2007. **3**(6): p. 2083-2097.
44. Gallicchio, E.; Zhang, L.Y.; Levy, R.M., *The SGB/NP hydration free energy model based on the surface generalized born solvent reaction field and novel nonpolar hydration free energy estimators.* J. Comput. Chem., 2002. **23**(5): p. 517-29.
45. Qiu, D.; Shenkin, P.S.; Hollinger, F.P.; Still, W.C., *The GB/SA Continuum Model for Solvation. A Fast Analytical Method for the Calculation of Approximate Born Radii.* J. Phys. Chem. A, 1997. **101**(16): p. 3005-3014.
46. Roux, B.; Simonson, T., *Implicit solvent models.* Biophys. Chem., 1999. **78**(1-2): p. 1-20.

47. Wagoner, J.A.; Baker, N.A., *Assessing implicit models for nonpolar mean solvation forces: the importance of dispersion and volume terms.* Proc. Natl. Acad. Sci. U. S. A., 2006. **103**(22): p. 8331-6.
48. Kollman, P.A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D.A.; Cheatham, T.E., 3rd, *Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models.* Acc. Chem. Res., 2000. **33**(12): p. 889-97.
49. Beutlera, T.C.; Marka, A.E.; van Schaikb, R.C.; Gerberc, P.R.; van Gunsteren, W.F., *Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations* Chem. Phys. Lett., 1994. **222**(6): p. 529-539.
50. Andersen, H.C., *Rattle - A Velocity Version of the Shake Algorithm for Molecular-Dynamics Calculations.* J. Comput. Phys., 1983. **52**(1): p. 24-34.
51. McQuarrie, D.A., *Statistical Mechanics*. 2000: University Science Books.
52. Tidor, B.; Karplus, M., *The contribution of vibrational entropy to molecular association. The dimerization of insulin.* J. Mol. Biol., 1994. **238**(3): p. 405-14.
53. Grater, F.; Schwarzl, S.M.; Dejaegere, A.; Fischer, S.; Smith, J.C., *Protein/ligand binding free energies calculated with quantum mechanics/molecular mechanics.* J. Phys. Chem. B, 2005. **109**(20): p. 10474-83.
54. Katz, B.A.; Elrod, K.; Luong, C.; Rice, M.J.; Mackman, R.L.; Sprengeler, P.A.; Spencer, J.; Hataye, J.; Janc, J.; Link, J.; Litvak, J.; Rai, R.; Rice, K.; Sideris, S.; Verner, E.; Young, W., *A novel serine protease inhibition motif involving a multi-centered short hydrogen bonding network at the active site.* J. Mol. Biol., 2001. **307**(5): p. 1451-86.
55. Leiros, H.K.; Brandsdal, B.O.; Andersen, O.A.; Os, V.; Leiros, I.; Helland, R.; Otlewski, J.; Willassen, N.P.; Smalas, A.O., *Trypsin specificity as elucidated by LIE calculations, X-ray structures, and association constant measurements.* Protein Sci., 2004. **13**(4): p. 1056-70.
56. Ota, N.; Stroupe, C.; Ferreira-da-Silva, J.M.; Shah, S.A.; Mares-Guia, M.; Brunger, A.T., *Non-Boltzmann thermodynamic integration (NBTI) for macromolecular systems: relative free energy of binding of trypsin to benzamidine and benzylamine.* Proteins, 1999. **37**(4): p. 641-53.
57. Schwarzl, S.M.; Tschopp, T.B.; Smith, J.C.; Fischer, S., *Can the calculation of ligand binding free energies be improved with continuum solvent electrostatics and an ideal-gas entropy correction?* J. Comput. Chem., 2002. **23**(12): p. 1143-9.
58. Talhout, R.; Engberts, J.B., *Thermodynamic analysis of binding of p-substituted benzamidines to trypsin.* Eur. J. Biochem., 2001. **268**(6): p. 1554-60.
59. Jiao, D.; Golubkov, P.A.; Darden, T.A.; Ren, P., *Calculation of protein-ligand binding free energy by using a polarizable potential.* Proc. Natl. Acad. Sci. U. S. A., 2008. **105**(17): p. 6290-6295.
60. Stone, J.E.; Gohara, D.; Shi, G., *OpenCL: A Parallel Programming Standard for Heterogeneous Computing Systems.* Comput Sci Eng, 2010. **12**(3): p. 66-72.

61. Wesenberg, J.H.; Ardavan, A.; Briggs, G.A.; Morton, J.J.; Schoelkopf, R.J.; Schuster, D.I.; Molmer, K., *Quantum computing with an electron spin ensemble.* Phys. Rev. Lett., 2009. **103**(7): p. 070502.
62. Schnieders, M.J.; Fenn, T.D.; Pande, V.S., *Polarizable Atomic Multipole X-Ray Refinement: Particle Mesh Ewald Electrostatics for Macromolecular Crystals.* J. Chem. Theory Comput., 2011. **7**(4): p. 1141-1156.

4 Gay-Berne and Electrostatic Multipole-based Coarse-grain Potential in Implicit Solvent

4.1 INTRODUCTION

A general coarse-grain model, consisting of rigid bodies of anisotropic Gay-Berne particles and point multipoles, has been developed[1]. The Generalized Kirkwood method is applied to account for the solvation effects [2]. While the current coarse-grain (CG) model is constructed from atomic force fields as with other coarse-grained models, our focus is on representing the general components of intermolecular forces such as electrostatic and repulsion-dispersion at a CG level, rather than matching the overall effective forces produced by atomic models. The strategy is much similar to that of developing empirical atomic potential energy model from quantum mechanical principles. The resulting CG model is transferable and not limited to specific systems or environments. Another distinct feature is that the model adopts the common functional forms that are supersets of all-atom model, which will facilitate future multi-scale applications.

4.2 METHODS

4.2.1 Gay-Berne Potential

The coarse-grain repulsion-dispersion interactions are represented with anisotropic Gay-Berne (GB) potentials. A full description of the Gay-Berne potential is available in our previous work [3, 4]. Based on Gaussian-overlap potential, the potential energy between two particles *i* and *j* has the form

$$U_{GB}(\hat{\mathrm{u}}_i, \hat{\mathrm{u}}_j, \mathrm{r}_{ij}) = 4\varepsilon(\hat{\mathrm{u}}_i, \hat{\mathrm{u}}_j, \hat{\mathrm{r}}_{ij})\left[\left(\frac{d_w\sigma_0}{r_{ij} - \sigma(\hat{\mathrm{u}}_i, \hat{\mathrm{u}}_j, \hat{\mathrm{r}}_{ij}) + d_w\sigma_0}\right)^{12} - \left(\frac{d_w\sigma_0}{r_{ij} - \sigma(\hat{\mathrm{u}}_i, \hat{\mathrm{u}}_j, \hat{\mathrm{r}}_{ij}) + d_w\sigma_0}\right)^{6}\right] \quad (1)$$

Where the range parameter $\sigma(\hat{\mathrm{u}}_i, \hat{\mathrm{u}}_j, \hat{\mathrm{r}}_{ij})$ has the generalized form as

$$\sigma(\hat{u}_i,\hat{u}_j,\hat{r}_{ij})=\sigma_0\left[1-\left\{\frac{\chi\alpha^2(\hat{u}_i\cdot\hat{r}_{ij})^2+\chi\alpha^{-2}(\hat{u}_j\cdot\hat{r}_{ij})^2-2\chi^2(\hat{u}_i\cdot\hat{r}_{ij})(\hat{u}_j\cdot\hat{r}_{ij})(\hat{u}_i\cdot\hat{u}_j)}{1-\chi^2(\hat{u}_i\cdot\hat{u}_j)^2}\right\}\right]^{-\frac{1}{2}} \quad (2)$$

and

$$\sigma_0=\sqrt{d_i^{\,2}+d_j^{\,2}} \quad (3)$$

$$\chi=\left[\frac{\left(l_i^{\,2}-d_i^{\,2}\right)\left(l_j^{\,2}-d_j^{\,2}\right)}{\left(l_j^{\,2}+d_i^{\,2}\right)\left(l_i^{\,2}+d_j^{\,2}\right)}\right]^{1/2} \quad (4)$$

$$\alpha^2=\left[\frac{\left(l_i^{\,2}-d_i^{\,2}\right)\left(l_j^{\,2}+d_i^{\,2}\right)}{\left(l_j^{\,2}-d_j^{\,2}\right)\left(l_i^{\,2}+d_j^{\,2}\right)}\right]^{1/2} \quad (5)$$

where l and d are the length and breadth of each particle, respectively.

The terms $\chi\alpha^2$, $\chi\alpha^{-2}$ and χ^2 can be calculated as:

$$\chi\alpha^2=\frac{l_i^{\,2}-d_i^{\,2}}{l_i^{\,2}+d_j^{\,2}} \quad (6)$$

$$\chi\alpha^{-2}=\frac{l_j^{\,2}-d_j^{\,2}}{l_j^{\,2}+d_i^{\,2}} \quad (7)$$

$$\chi^2=\frac{\left(l_i^{\,2}-d_i^{\,2}\right)\left(l_j^{\,2}-d_j^{\,2}\right)}{\left(l_j^{\,2}+d_i^{\,2}\right)\left(l_i^{\,2}+d_j^{\,2}\right)} \quad (8)$$

The total well-depth parameter is presented as

$$\varepsilon(\hat{u}_i,\hat{u}_j,\hat{r}_{ij})=\varepsilon_0\varepsilon_1^{\nu}(\hat{u}_i,\hat{u}_j)\varepsilon_2^{\mu}(\hat{u}_i,\hat{u}_j,\hat{r}_{ij}) \quad (9)$$

The orientation-dependent strength terms are calculated in the following manner

$$\varepsilon_1(\hat{u}_i,\hat{u}_j)=\left[1-\chi^2(\hat{u}_i\cdot\hat{u}_j)^2\right]^{-1/2} \quad (10)$$

$$\varepsilon_2(\hat{u}_i,\hat{u}_j,\hat{r}_{ij})=1-\left\{\frac{\chi'\alpha'^2(\hat{u}_i\cdot\hat{r}_{ij})^2+\chi'\alpha'^{-2}(\hat{u}_j\cdot\hat{r}_{ij})^2-2\chi'^2(\hat{u}_i\cdot\hat{r}_{ij})(\hat{u}_j\cdot\hat{r}_{ij})(\hat{u}_i\cdot\hat{u}_j)}{1-\chi'^2(\hat{u}_i\cdot\hat{u}_j)^2}\right\} \quad (11)$$

Where

$$\chi'=\left[\frac{(\varepsilon_{Si}^{\,1/\mu}-\varepsilon_{Ei}^{\,1/\mu})\times(\varepsilon_{Sj}^{\,1/\mu}-\varepsilon_{Ej}^{\,1/\mu})}{(\varepsilon_{Sj}^{\,1/\mu}+\varepsilon_{Ei}^{\,1/\mu})\times(\varepsilon_{Si}^{\,1/\mu}+\varepsilon_{Ej}^{\,1/\mu})}\right]^{1/2} \quad (12)$$

$$\alpha'^2 = \left[\frac{(\varepsilon_{Si}^{1/\mu} - \varepsilon_{Ei}^{1/\mu}) \times (\varepsilon_{Sj}^{1/\mu} + \varepsilon_{Ei}^{1/\mu})}{(\varepsilon_{Sj}^{1/\mu} - \varepsilon_{Ej}^{1/\mu}) \times (\varepsilon_{Si}^{1/\mu} + \varepsilon_{Ej}^{1/\mu})} \right]^{1/2} \tag{13}$$

The well depth of the cross configuration is denoted by ε_0, the well depth of the end-to-end/face-to-face configuration is presented as ε_E, and ε_S denotes the well depth of the side-by-side configuration[5]. Here we improved the accuracy of the Gay-Berne model by separating the ratio of $\varepsilon_E / \varepsilon_S$ to two independent variables, ε_E and ε_S.

The new representations of χ' and α'^2 allow the consistent result for a pair of Gay-Berne particles of arbitrary types. Between unlike pairs, all ε_0 values and their ε_S and ε_E are specified explicitly or computed using a combining rule [6]. The d_W parameter describes the "softness" of the potential to allow better correlation with the all-atom energy profile. The parameters μ and ν were set to canonical values of 2.0 and 1.0, respectively. The current Gay-Berne potential with electrostatic multipole (GBEMP) model is implemented based on the TINKER molecular dynamics package [7].

The terms χ'^2, $\chi'\alpha'^2$, and $\chi'\alpha'^{-2}$ were treated as inseparable and computed directly as:

$$\chi'^2 = \frac{(\varepsilon_{Si}^{1/\mu} - \varepsilon_{Ei}^{1/\mu}) \times (\varepsilon_{Sj}^{1/\mu} - \varepsilon_{Ej}^{1/\mu})}{(\varepsilon_{Sj}^{1/\mu} + \varepsilon_{Ei}^{1/\mu}) \times (\varepsilon_{Si}^{1/\mu} + \varepsilon_{Ej}^{1/\mu})} \tag{14}$$

$$\chi'\alpha'^2 = \frac{(\varepsilon_{Si}^{1/\mu} - \varepsilon_{Ei}^{1/\mu})}{(\varepsilon_{Si}^{1/\mu} + \varepsilon_{Ej}^{1/\mu})} \tag{15}$$

$$\chi' \alpha'^{-2} = \frac{\left(\varepsilon_{Sj}^{1/\mu} - \varepsilon_{Ej}^{1/\mu}\right)}{\left(\varepsilon_{Sj}^{1/\mu} + \varepsilon_{Ei}^{1/\mu}\right)} \tag{16}$$

Electrostatic potentials are represented with pairwise interactions of point multipole sites up to quadrupole. Each rigid body may contain zero or more off-center multipole sites where the local frame of the site is aligned with the principle axis of the rigid body. Electrostatic interactions of the GBEMP model is analogous to the AMOEBA's permanent electrostatic terms described previously and also provided in previous work [3, 4].

4.3 RESULTS AND DISCUSSION

4.3.1 Benzene and Methanol Model

The improved Gay-Berne functional form has been validated on benzene and methanol molecules, which were represented by disk-like and rod-like particles, respectively. As with the previous studies [3, 4], the Gay-Berne parameters were derived by first fitting to the gas-phase homodimer intermolecular interaction energy and then refined in the liquid simulations. All-atom homodimer interactions energy for cross, end-end, face-face, and side-by-side configurations was obtained at various separations up to 12 Å apart. At each separation, the dimer interaction energy was calculated as a Boltzmann average over configurations generated by rotation about the primary axis of each Gay-Berne particle. Molecular electrostatic multipole (EMP) moments of benzene and methanol in liquid environments were obtained from atomic multipoles, including induced dipoles, given by the all-atom AMOEBA polarizable force field [8, 9].

In coarse-grained liquid simulations, the initial structures of benzene and methanol particles were created by mapping from all-atom structures. After rigid-body energy minimization, MD simulations of a box of ~300 molecules were performed with

an NPT ensemble at 298 K and 1 atm. The periodic boundary condition was applied with a cutoff of 12 Å. Different time steps (up to 20 fs) were tested in the CG simulations.

A comparison of dimer interaction energies between the all-atom and GBEMP models shows that the new functions for combining the Gay-Berne well-depth parameters, ε_E and ε_S, produce a better agreement than the previous Gay-Berne function (see Appendix). The well-depth for benzene in the T shape configuration has increased to 0.91 Kcal/mol from 0.52 Kcal/mol using the previous model) and more closely matches that of all-atom result (1.60 Kcal/mol) (See **Figure S1**). Liquid simulations for benzene and methanol yield bulk properties, such as internal potential energy and density, that are in excellent agreement with the experimental values (error < 2%) (see Appendix). The GBEMP model is next extended to polyalanine peptides that consist of bonded coarse-grained particles.

4.3.2 Alanine Model

In our CG model, a peptide is composed of covalently bonded rigid bodies, with Gay-Berne and/or electrostatic multipole sites. Bonding occurs between the Gay-Berne or EMP sites on different rigid bodies. Bond stretch energies adopt the fourth-order Taylor expansion of the Morse potential. Bond angle bend energies utilize a sixth-order potential. A three-term Fourier series expansion is calculated with the torsion energy. These valence functional forms are similar to those used by classical molecular mechanics potential such as MM3 [10]. To use large time-step in MD simulations, the bond and angle terms can be restrained using rattle algorithm [11].

In a previous work [4], we have devised a general rigid-body representation containing an arbitrary number of off-centered Gay-Berne and multipole interaction sites that share the same local frame. Gay-Berne interactions are computed using orientation

and site location vectors in Cartesian coordinates, relative to the local frame of the rigid body, as variables. Likewise, multipole interactions are computed via positions given by Cartesian coordinates relative to the local frame of the rigid body. The dialanine model consists of 5 rigid bodies (I through V) as depicted in **Figure 4.1**. Gay-Berne parameters of amide and methyl groups were obtained with the same procedure as described above, by fitting to AMOEBA atomic force field. As in **Figure 4.1**, the rigid body that represents the amide group consists of one Gay-Berne particle and two EMP sites. Gay-Berne sites 1, 5, and 10 are spherical methyl groups while sites 3 and 8 are equivalent ellipsoid amide groups. Similarly, sites 2 and 7 share the same EMP type, as do sites 4 and 9. Site 6 is used to compute bonded interactions only. Bonds exist between sites (1, 3), (4, 6), (6, 8), and (9, 10). An example of an angle is composed of sites (1, 3, 2) and a torsion angle is composed of sites (3, 4, 6, 8). The 12-mer alanine model polymerizes rigid bodies II and III from **Figure 4.1** as a repeating unit 12 times, thus, requiring 27 rigid bodies. For each rigid body type, the coordinates of the corresponding atoms are recorded in the local frame, which allow us to map the coarse grain molecules back to all-atom structures. Note that although the Gay-Berne particle is symmetric about the primary axis, the rigid body is not necessarily symmetric due to the presence of off-center site and/or multipoles.

Solvation is represented implicitly and is composed of polar and nonpolar contributions. Polar solvation employs the Generalized Kirkwood (GK) method [2], a multipolar extension of the Generalized Born approach [12, 13] and is computed for all the multipole sites. The Grycuk effective radius [14] is used in the polar solvation calculations. Nonpolar solvation is evaluated for all Gay-Berne sites with the ACE surface area method [15] and Still method[12, 16] to estimate the effective radius of each particle. All solvation methods as well as effective radii estimation methods are

implemented in the TINKER 5[7] molecular modeling package and adapted to the current GBEMP suite. Particle radii used for effective radii estimation are taken from the maximum of the Gay-Berne l or d parameters. Rigid bodies with more than one multipole site, like the amide groups in **Figure 4.1** (II and IV), uniformly divide the Gay-Berne radius value among all sites.

Parameters for the alanine model were obtained for the non-bonded terms, such as Gay-Berne and electrostatic multipole potentials, as well as the bonded terms, such as bond stretching, angle bending, and torsion energies. Applying the same procedure used to parameterize benzene and methanol, Gay-Berne and EMP parameters for each rigid body in an alanine residue were fit to all-atom homodimer energy and monomer multipole (in solution environments), respectively. Bond stretch and angle bend parameters were parameterized via Boltzmann inversion with atomic configurations generated from molecular dynamics of alanine dipeptide using AMOEBA. Molecular dynamics were executed in an NVT ensemble with explicit solvent (209 water molecules) in a 19.7 Å box with a 1 fs time step at 298 K. Torsional energy parameters were fit to the all-atom conformational energy map generated with fixed-charge OPLSAA with Generalized Born Surface Area implicit solvation [12, 15]. OPLSAA is chosen as it is a commonly used atomic force field and uses the similar torsional energy function as in the current coarse-grain model. Nonetheless, the torsional parameters will be refined in the future by comparing directly to experimental data [17]. As we discuss below, the torsional term only contributes to a fraction of the conformational energy along with the intramolecular nonbonded electrostatic and van der Waals interactions.

4.3.3 Dialanine Energy Components from CG Model

The conformational energy of dialanine as a function of backbone dihedral angles, ϕ and ψ, is investigated in solution and gas phases. Conformations are generated at 30-degree intervals starting at the origin of the energy map by minimization with restraints. Conformational energies for the GBEMP model in solution- and gas-phase are shown in **Figure 4.2**, compared with corresponding energies from all-atom model using the OPLSAA field [18]. The energy surface of the GBEMP model is smoother than that of the all-atom model as a consequence of coarse-graining. Nonetheless, the overall features of the CG gas phase energy maps are in fair agreement with the corresponding map of the atomic OPLSAA force field. Moreover, solution phase energy maps are in excellent qualitative agreement between the GBEMP and atomic force field. The agreement between solution phase energy maps is better than that of the gas phase maps and is expected since both are designed to describe solution phase properties. This is encouraging as the CG torsional parameters were only fit to the OPLSAA energy in solution. In addition, the solution-phase minima for alpha-helix, beta-sheet, as well as the less stable left-handed alpha-helix conformations are well manifested in the energy map.

When compared to the gas-phase electrostatic energy (**Figure 4.3 b and e**), the solvation energy contribution (**Figure 4.3 c and f**) clearly compensates the electrostatic interactions in gas-phase. This observation, true for both all-atom (OPLSAA) and the current CG potentials, is consistent with the physical interpretation that when secondary structure forms, intramolecular hydrogen bonds replace the hydrogen bonds between peptide and surrounding water.

We further compared the energy components of the coarse-grained GBEMP model with OPLSAA. A decomposition of the non-bonded interactions indicates that steric interaction given by the Gay-Berne function in the GBEMP model resemble that

given the atomic vdW interaction energy of the OPLSAA force field over the Ramachandran map (**Figure 4.3 a and d**), including the scale. Likewise, contour maps of the gas-phase electrostatic energy (**Figure 4.3 b and e**), as well as the implicit solvation energy (**Figure 4.3 c and f**), show good agreement between the coarse grain and the all-atom results. Although the overall scales are different, the two components seem to mostly cancel each other as discussed above. As a result the total energy minimum at the alpha-helix conformation mostly arises from the vdW contribution (**Figure 4.3 a and d**). A comparison of the torsional energy contribution (see Appendix) between the CG and all-atom models also expresses a consistent behavior. The gas-phase conformational energy captures the C5 local minimum well[19]. However, the C7eq and C7ax minima have drifted slightly from the all-atom conformations. This may be due to the torsional energy contributions since their parameters were fit to the condensed-phase energy map. However, as with other all-atom fixed-charge models, transferability between gas- and solution-phase requires the inclusion of polarization effect.

4.3.4 Simulation of Polyalanine

The conformation of polyalanine with various lengths has been investigated with both experimental and computational approaches [17, 20-33]. To compare the GBEMP model with experiments and all-atom MD simulations, we investigated the blocked 5-mer polyalanine using GBEMP model in MD simulations. The aforementioned Generalized Kirkwood implicit solvent was utilized. The replica exchange molecular dynamics (REMD) [34] was performed to elucidate the conformational distribution of the 5-mer polyalanine. Thirty replicas were used between 298 and 800K and the simulation time for each replica was 200 ns. The distribution of ϕ and ψ angles for all residues is shown in (**Figure 4.4 a**). Three dominant populations were observed: alpha-helix ($-160^{\circ} \le \phi \le -20^{\circ}$

and $-120^\circ \le \psi \le 50^\circ$), beta-strand ($-180^\circ \le \phi \le -90^\circ$ and $50^\circ \le \psi \le 240^\circ$; or $160^\circ \le \phi \le 180^\circ$ and $110^\circ \le \psi \le 180^\circ$), and left-handed helix ($20^\circ \le \phi \le 160^\circ$ and $-50^\circ \le \psi \le 120^\circ$). The 5-mer polyalanine conformations observed are comparable with all atom simulation results (**Table 4.1**). Although circular dichroism (CD) spectroscopy and Fourier-transform infrared (FTIR) experiments reported somewhat less alpha-helix conformation [33], the distributions sampled from MD simulations using all-atom force fields seem to be in qualitative agreement with what we obtained from the GBEMP simulations. Moreover, since the GBEMP model was developed based on interactions of all-atom force fields, it is reasonable for the model to behave consistently with all-atom simulation. Additionally, the population of full alpha helices, in which ϕ and ψ angles of all five residues adopt the alpha-helical conformation, occurs at 4.62%, in comparison with 8% and 1% observed in all-atom simulations using CHARMM and Amber03 force field, respectively [33].

To study the effects of chain-length, a 12-residue polyalanine system was simulated using REMD with 30 replicas and 500 ns for each replica. Residue-level conformational distributions observed were 42%, 4.3%, and 21%, for alpha-helix, beta-strand, and left-handed helix conformations, respectively. Although the beta-strand conformation exhibits a minima in the conformational energy landscape (**Figure 4.2 a, c**), a substantial (5-fold) decrease in the beta-strand distribution compared to the 5-mer polyalanine suggests that the hydrogen bonding scheme provided by the alpha-helix conformation stabilizes the 12-mer polyalanine. Additionally, simulated annealing MD simulations were performed to inspect the minimum-energy structure of the peptide after an initial rigid-body energy minimization. The systems were heated to 1,000 K within the first 50 ps and then cooled linearly to less than 1 K over 60 ns. Final polyalanine structures after simulated annealing all adopt the alpha-helical conformation at low

temperatures and a comparison of the RMSD between structures obtained from the simulated annealing trajectory and a canonical alpha-helix suggests that the accessible area of phase-space noticeably increases as the temperature rises above 500 K (See Appendix).

Furthermore, MD simulations of a few microseconds were performed at room temperature to verify the convergence of the conformational space determined by the GBEMP/REMD. These simulations started with different initial structures, including the extended conformation, alpha helix, and partial alpha-helical and beta-strand conformations. The torsional distribution sampled from the GB-EMP MD simulation (6 μs for 12-mer and 2 μs for 5-mer) at 298K is in agreement with the REMD conformational map (See Appendix).

4.3.5 Computational Efficiency of the GBEMP Model

The GBEMP model provides a great improvement in the performance of molecular modeling. Due to the reduction of particle numbers and larger time-steps, the computational efficiency is enhanced by a factor of 50 – 800 compared to all-atom models tested with implicit and explicit solvent in this study (Table 4.2). Furthermore, the absence of high frequency motions, as required by all-atom models, allows time steps of up to 5 fs in MD simulations. Therefore, the CG model can achieve an improvement of about three orders of magnitude in the simulation speed and enable studies of large systems or extended simulation times from nanoseconds to microseconds.

4.4 CONCLUSIONS

A unique coarse-grained GBEMP (Gay-Berne potential with electrostatic multipole) model has been developed based on the general physical principles of molecular interactions. The GBEMP potential explicitly represents the fundamental

components of intermolecular forces. The van der Waals interaction is described by treating molecules as soft uniaxial ellipsoids interacting via a generalized anisotropic Gay-Berne function. The charge distribution is represented by off-center multipoles, which are composed of point charge, dipole, and quadrupole moments. The Generalized Kirkwood method and the ACE surface area method are used to calculate the polar and nonpolar solvation energy, respectively [2, 15]. The coarse-grained GBEMP model has been implemented in the TINKER modeling package capable of rigid-body molecular dynamics simulation. The replica-exchange method is implemented to enhance sampling. The CG parameters are calibrated using all-atom force field (AMOEBA and OPLS-AA) and extension to other molecular systems is straightforward. Most importantly, there is no need for constant re-parameterization when applied to different environments. We tested the CG model on the alanine peptides of various lengths. The results show that the model and parameters can be directly transferred from gas phase to solution (with implicit solvent model), and from dialanine to polyalanine of different lengths. For the first time, we show that the individual energy components in the coarse-grained model, including vdW, electrostatics, solvation and torsional energy contributions, match closely with those of all-atom force fields, in both gas-phase and solution. REMD and room-temperature MD simulations of 5-residue and 12-residue polyalanines predict reasonable alpha-helix and beta-sheet distributions in comparison with all-atom simulations and experiments. Due to the reduction of particle numbers and larger time-steps, the computational efficiency is enhanced by a factor of up to 1,000 compared with all-atom simulations. The coarse-graining potential presented in this study can be extended to various biomolecular systems and even combined with all-atom potential in multiscale applications.

Table 4.1: Per-residue conformational distributions of 5-mer polyalanine from experiments and all-atom simulations.

Conformation	CD[a]	FTIR[a]	All-atom[a,b]	CHARMM 27/cmap[b]	OPLSAA/L[b]	GBEMP
alpha-helix	13± 3%	13±5%	4% - 60% [a]	57.5%	32.8%	46%
beta-strand	N/A	N/A	9.8% - 55.5% [a]	19.8%	32.0%	28%

[a] Hegefeld, 2010 distributions from experiment and various force fields.
[b] Best, 2008 distributions of various force fields

Table 4.2: Comparison of computational efficiency of GBEMP model. All simulations were performed in TINKER package. The time step is 1 fs for all-atom simulations and 5fs for GBEMP simulations.

	GMEMP	Amber 99SB with implicit water	Amber 99SB with TIP3P water
5-mer polyalaine	~290 ns/day	~5.3 ns/day	~0.34 ns/day
12-mer polyalaine	~122 ns/day	~2.4 ns/day	~0.15 ns/day

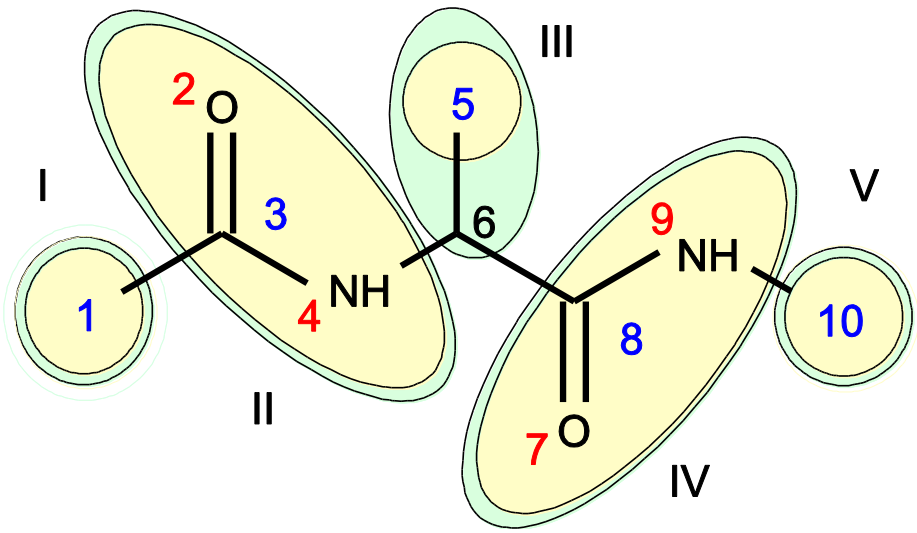


Figure 4.1: Representation of dialanine coarse-grained GBEMP model. Ellipsoids encompass the rigid bodies (green) that contains Gay-Berne (blue) and multipole (red) interaction sites. The Gay-Berne particles are located are at the center of the mass of the corresponding atoms.

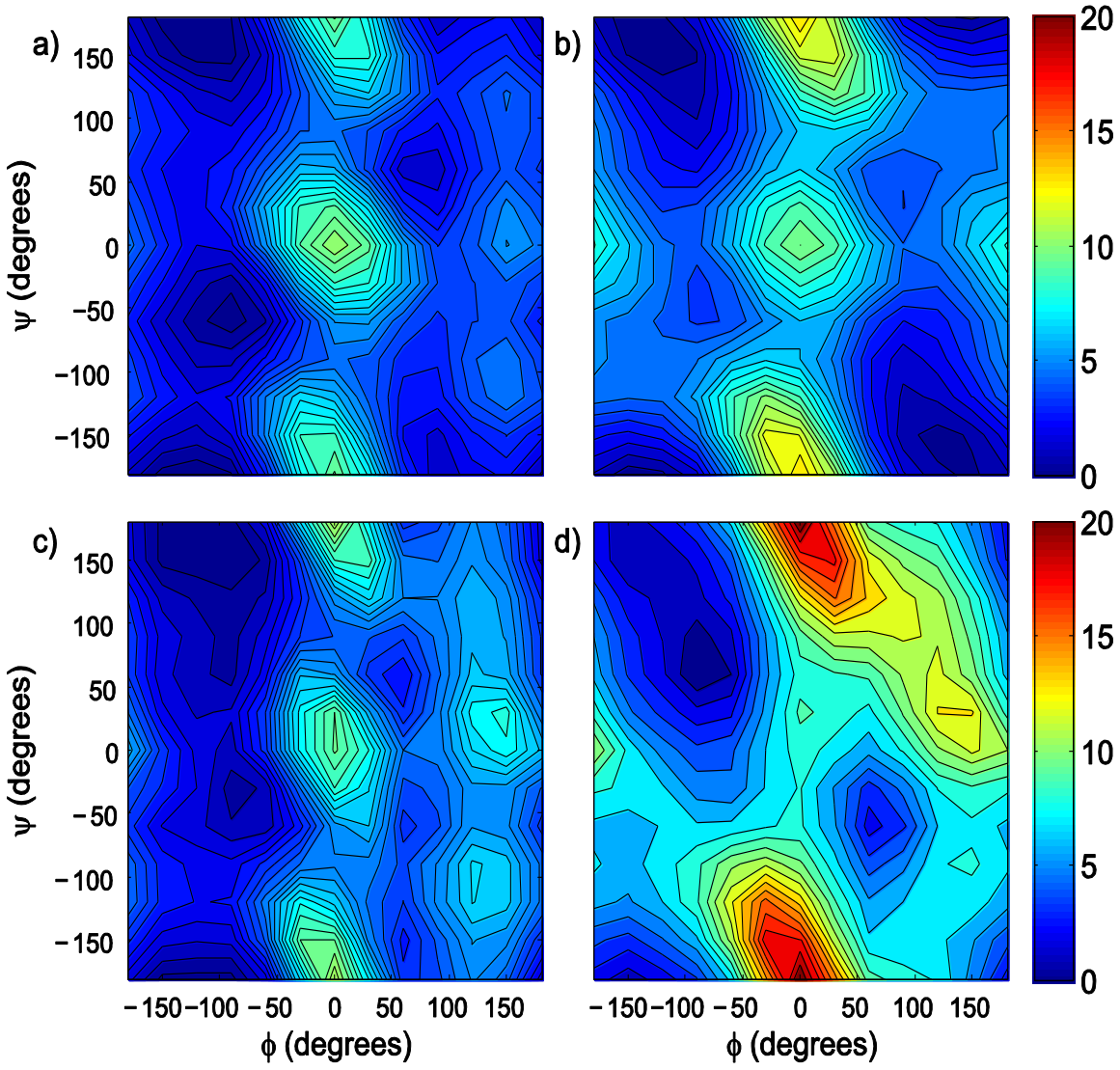


Figure 4.2: Total conformational energy (kcal/mol) of alanine dipeptide: (a) CG model in solution, (b) CG model in gas-phase, (c) all-atom model (OPLSAA) in solution, (d) all-atom model (OPLSAA) in gas-phase.

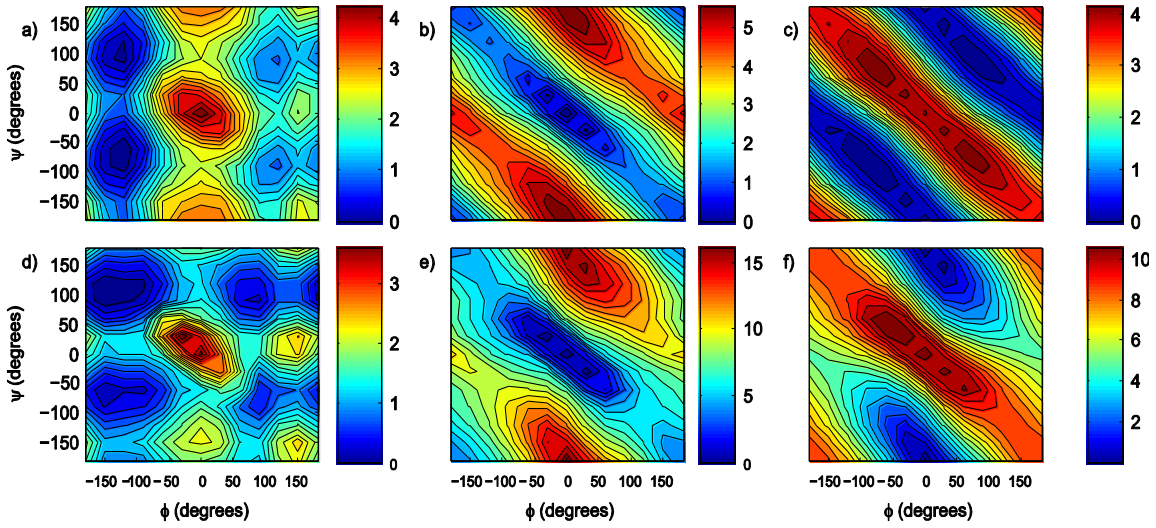


Figure 4.3: Decomposition of alanine dipeptide energy (kcal/mol). **Coarse-grain:** (a) Gay-Berne energy (b) Gas-phase electrostatic energy (c) implicit solvation energy from GK/SA. **All-atom**: (d) vdW energy (e) Gas-phase electrostatic energy (f) implicit solvation energy from GB/SA.

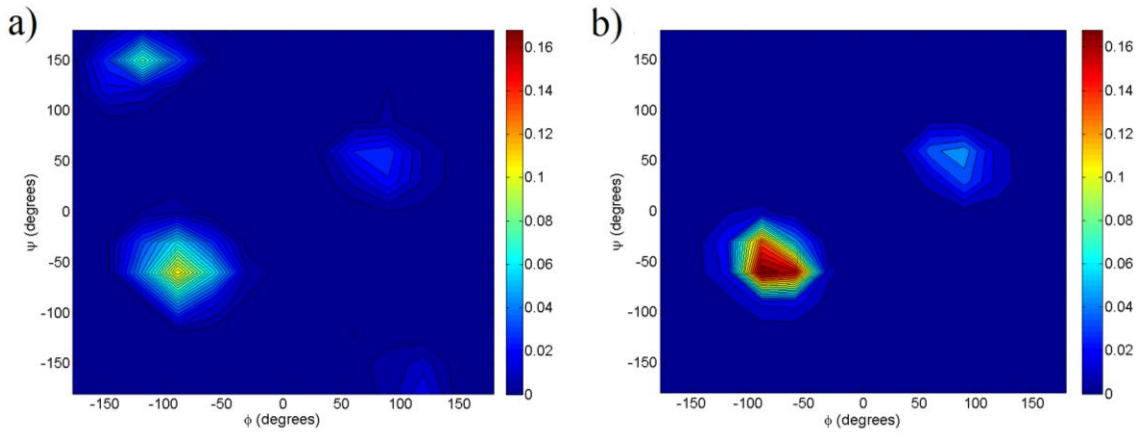


Figure 4.4: Conformational distribution of 5-mer (a) and 12-mer (b) polyalanine from CG REMD simulations.

4.5 REFERENCES

1. Wu, J.C.; Xia, Z.; Shen, H.; Li, G.; Ren, P.Y., *Gay-Berne and electrostatic multipole based coarse-grain potential in implicit solvent.* J. Chem. Phys., 2011. **135**(15): p. 155104.
2. Schnieders, M.J.; Ponder, J.W., *Polarizable atomic multipole solutes in a generalized Kirkwood continuum.* Journal of Chemical Theory and Computation, 2007. **3**(6): p. 2083-2097.
3. Golubkov, P.A.; Ren, P.Y., *Generalized coarse-grained model based on point multipole and Gay-Berne potentials.* J. Chem. Phys., 2006. **125**(6): p. 64103.
4. Golubkov, P.A.; Wu, J.C.; Ren, P.Y., *A transferable coarse-grained model for hydrogen-bonding liquids.* Phys. Chem. Chem. Phys., 2008. **10**(15): p. 2050-2057.
5. Cleaver, D.J.; Care, C.M.; Allen, M.P.; Neal, M.P., *Extension and generalization of the Gay-Berne potential.* Physical Review E, 1996. **54**(1): p. 559-567.
6. Halgren, T.A., *Representation of Vanderwaals (Vdw) Interactions in Molecular Mechanics Force-Fields - Potential Form, Combination Rules, and Vdw Parameters.* J. Am. Chem. Soc., 1992. **114**(20): p. 7827-7843.
7. Ponder, J.W., *TINKER molecular modeling package.* Washington University Medical School, 2010.
8. Ren, P.; Ponder, J.W., *Polarizable Atomic Multipole Water Model for Molecular Mechanics Simulation.* Journal of Physical Chemistry B, 2003. **107**: p. 5933-5947.
9. Ponder, J.W.; Wu, C.J.; Ren, P.Y.; Pande, V.S.; Chodera, J.D.; Schnieders, M.J.; Haque, I.; Mobley, D.L.; Lambrecht, D.S.; DiStasio, R.A.; Head-Gordon, M.;

Clark, G.N.I.; Johnson, M.E.; Head-Gordon, T., *Current Status of the AMOEBA Polarizable Force Field.* Journal of Physical Chemistry B, 2010. **114**(8): p. 2549-2564.

10. Allinger, N.L.; Yuh, Y.H.; Lii, J.H., *Molecular Mechanics - the MM3 Force-Field for Hydrocarbons .1.* J. Am. Chem. Soc., 1989. **111**(23): p. 8551-8566.
11. Andersen, H.C., *Rattle - A Velocity Version of the Shake Algorithm for Molecular-Dynamics Calculations.* J. Comput. Phys., 1983. **52**(1): p. 24-34.
12. Still, W.C.; Tempczyk, A.; Hawley, R.C.; Hendrickson, T., *Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics.* J. Am. Chem. Soc., 1990. **112**(16): p. 6127-6129.
13. Constanciel, R.; Contreras, R., *Self-Consistent Field-Theory of Solvent Effects Representation by Continuum Models - Introduction of Desolvation Contribution.* Theor. Chim. Acta, 1984. **65**(1): p. 1-11.
14. Grycuk, T., *Deficiency of the Coulomb-field approximation in the generalized Born model: An improved formula for Born radii evaluation.* J. Chem. Phys., 2003. **119**(9): p. 4817-4826.
15. Schaefer, M.; Karplus, M., *A comprehensive analytical treatment of continuum electrostatics.* J. Phys. Chem., 1996. **100**(5): p. 1578-1599.
16. Qiu, D.; Shenkin, P.S.; Hollinger, F.P.; Still, W.C., *The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii.* J. Phys. Chem. A, 1997. **101**(16): p. 3005-3014.
17. Best, R.B.; Buchete, N.V.; Hummer, G., *Are current molecular dynamics force fields too helical?* Biophys. J., 2008. **95**(1): p. L7-L9.
18. Jorgensen, W.L.; Maxwell, D.S.; 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): p. 11225-11236.
19. Headgordon, T.; Headgordon, M.; Frisch, M.J.; Brooks, C.L.; Pople, J.A., *Theoretical-Study of Blocked Glycine and Alanine Peptide Analogs.* J. Am. Chem. Soc., 1991. **113**(16): p. 5989-5997.
20. Chou, P.Y.; Fasman, G.D., *Conformational Parameters for Amino-Acids in Helical, Beta-Sheet, and Random Coil Regions Calculated from Proteins.* Biochemistry, 1974. **13**(2): p. 211-222.
21. Richardson, J.S.; Richardson, D.C., *Amino-Acid Preferences for Specific Locations at the Ends of Alpha-Helices.* Science, 1988. **240**(4859): p. 1648-1652.
22. Hudgins, R.R.; Ratner, M.A.; Jarrold, M.F., *Design of helices that are stable in vacuo.* J. Am. Chem. Soc., 1998. **120**(49): p. 12974-12975.
23. Levy, Y.; Jortner, J.; Becker, O.M., *Solvent effects on the energy landscapes and folding kinetics of polyalanine.* Proc. Natl. Acad. Sci. U. S. A., 2001. **98**(5): p. 2188-2193.
24. Counterman, A.E.; Clemmer, D.E., *Large anhydrous polyalanine ions: Evidence for extended helices and onset of a more compact state.* J. Am. Chem. Soc., 2001. **123**(7): p. 1490-1498.

25. Henzler, K.A.; Lee, D.K.; Ramamoorthy, A., *Conformational stability of solid-state poly(l-alanine)*. Biophys. J., 2001. **80**(1): p. 187a-187a.
26. Nguyen, H.D.; Marchut, A.J.; Hall, C.K., *Solvent effects on the conformational transition of a model polyalanine peptide.* Protein Sci., 2004. **13**(11): p. 2909-2924.
27. Soto, P.; Baumketner, A.; Shea, J.E., *Aggregation of polyalanine in a hydrophobic environment.* J. Chem. Phys., 2006. **124**(13): p. 134904.
28. Zhou, J.; Thorpe, I.F.; Izvekov, S.; Voth, G.A., *Coarse-grained peptide modeling using a systematic multiscale approach.* Biophys. J., 2007. **92**(12): p. 4289-4303.
29. Chu, J.W.; Izvekov, S.; Voth, G.A., *The multiscale challenge for biomolecular systems: coarse-grained modeling.* Mol. Simul., 2006. **32**(3-4): p. 211-218.
30. Jarrold, M.F., *Helices and sheets in vacuo.* Phys. Chem. Chem. Phys., 2007. **9**(14): p. 1659-1671.
31. Graf, J.; Nguyen, P.H.; Stock, G.; Schwalbe, H., *Structure and dynamics of the homologous series of alanine peptides: A joint molecular dynamics/NMR study.* J. Am. Chem. Soc., 2007. **129**(5): p. 1179-1189.
32. Albrieux, F.; Calvo, F.; Chirot, F.; Vorobyev, A.; Tsybin, Y.O.; Lepere, V.; Antoine, R.; Lemoine, J.; Dugourd, P., *Conformation of Polyalanine and Polyglycine Dications in the Gas Phase: Insight from Ion Mobility Spectrometry and Replica-Exchange Molecular Dynamics.* J. Phys. Chem. A, 2010. **114**(25): p. 6888-6896.
33. Hegefeld, W.A.; Chen, S.E.; DeLeon, K.Y.; Kuczera, K.; Jas, G.S., *Helix Formation in a Pentapeptide Experiment and Force-field Dependent Dynamics.* J. Phys. Chem. A, 2010. **114**(47): p. 12391-12402.
34. Penev, E.S.; Lampoudi, S.; Shea, J.E., *TiREX: Replica-exchange molecular dynamics using TINKER.* Comput. Phys. Commun., 2009. **180**(10): p. 2013-2019.

5 Correlation of RNA Secondary Structure Statistics with Thermodynamic Stability and Applications to Folding

5.1 INTRODUCTION

Stochastic grammar-based models[1-4] have been used as a non-physical-based method to model RNA. Do *et al.* developed, a method that maximizes the expectation of the objective function related to the accuracy of the prediction[5]. Although this knowledge-based approach has only recently been applied to RNA, there has been an increased effort to apply statistically derived potentials to the prediction of RNA structure. A few years ago, Dima and coworkers[6] extracted RNA structural statistics from experimentally determined high-resolution structures from the Protein Data Bank to attain PDB-derived potentials for consecutive base-pairs in helices[7]. Statistics of the tertiary interactions in the three-dimensional structures of RNAs in PDB were not determined due to the complexity and uncertainty of the sets of nucleotides to study. They determined, to a first approximation, that the structural potentials derived from the base pairings in the secondary structure of experimentally determined 3D structures are similar to those energy values determined experimentally by Turner and collaborators.

Recently, Das et al. developed a Rosetta-like scheme to predict tertiary structure of small RNA sequences of ~30 nucleotides[8]. In their scheme, a statistical potential is inferred from the distance and angle distributions of base-pairs in the ribosome crystal structure, following the method by Sykes and Levitt[9]. Parisien[10] et al. predicts the secondary and tertiary structures of short RNA molecules with the statistics of nucleotide cyclic motifs using the dynamic programming Waterman-Byers algorithm. Recently, Jonikas et al developed a coarse-grain nucleotide based model to effectively predict structure[11]It has been appreciated since the first few tRNA sequences were determined, that different RNA sequences with similar function can form a similar secondary and

tertiary structure[12, 13] Comparative analysis of tRNA sequences revealed the classic cloverleaf secondary structure[14] and parts of the three-dimensional structure[15]. Comparative analysis has also revealed other RNA secondary structures that are each similar to different sets of RNA sequences with similar functions. This list includes 5S rRNA[16], 16S rRNA[17-19], 23S rRNA[20-22], RNase P[23], and group I and II introns[24-26]. Covariation and other comparative analysis have the potential to be extremely accurate when there are a sufficient number and diversity of properly aligned sequences. For the rRNA, 97-98% of the base-pairs predicted with covariation analysis were present in the high-resolution crystal structure of the ribosome[27].

5.2 METHODS

5.2.1 RNA Comparative Structure

The RNA molecules - 5S rRNA, 16S rRNA, and 23S rRNA are available for different phylogenetic groups from the rCAD database. The rCAD database is implemented in the Microsoft SQL Server, a relational database management system. Analysis performed on rCAD can be accessed online at the Comparative RNA Web Site CRW) at http://www.rna.ccbb.utexas.edu. The rCAD system stores over 50 000 RNA sequences, 1746521 nucleotides, and comparative structures. We have rich information on base-pair stack statistics of 319 Bacterial 16S rRNA sequences, 650 tRNA sequences, 263 Eukaryotic 5S rRNA sequences, and 96 Bacterial 5S rRNA. Sequences of the above molecules are available from all phylogenetic domains including Eukaryote, Bacteria, and Archaea. Sequences with similarity of greater than 97% have been pruned to eliminate duplicates. We assume an ensemble of good distribution[28]. The ribosomal RNA is studied for its rich structural diversity, its functional abilities and for its well-conserved qualities within phylogenetic domains. The secondary structures of all

sequences are evaluated from comparative analysis[12, 13, 20, 29] and statistical potentials based on these structures have been reported[30]. Structural motifs such as base-pair stacks, hairpin flanks, hairpin loops, internal loops, multi-stem loops, and bulges are stored in tables within the SQL Server.

5.2.2 Base-pair Stack Statistical Energy

Statistics of base-pair stacks found via comparative analysis have been collected to calculate statistical energies. A base-pair stack is denoted AB/CD where AB and CD are base-pairs, and A and D are on the 5' ends of the stack. For example, the UA/GC base-pair stack is identified in Figure 5.1A. The sample size of base-pair stacks obtained from sequence analysis is three orders of magnitude greater that those obtained from crystal structures[6]. In addition to containing a larger sample size than the set of crystal structures, sequence data contains a greater diversity spanning a larger portion from the phylogenetic tree of life and, hence, a more complete ensemble. Statistics were obtained by counting base-pair stacks composed of canonical Watson-Crick (CG and AU), and GU base-pairs and have been previously reported [30].

Statistical energies are calculated from base-pair stack statistics:

$$\Delta G_{BP-ST}(ij,kl) = -\lambda k_B T \ln\left(\frac{P_{BP-ST}(ij,kl)}{P_{BP-ST}^{(rand)}(ij,kl)}\right) \tag{6}$$

where

$$P_{BP-ST}(ij,kl) = \frac{N_{BP-ST}(ij,kl) + N_{BP-ST}(kl,ij)}{N_{BP-ST}} \tag{7}$$

and

$$P_{BP-ST}^{(rand)}(ij,kl) = (1+\delta_{ij,kl})P_iP_jP_kP_l \tag{8}$$

The indices *i, j, k, l* represent any of the nucleotides A, C, G, or U. We define $N_{ST}(ij,kl)$ as the number of base-pair stacks composed of *ij* and *kl*. The total number of

base-pair stacks is N_{ST}. The delta function $\delta_{ij,kl}$ is equal to 1 if $ij=kl$ and 0 otherwise. P_i is the probability of occurrence of nucleotide i. The scaling factor, λ, is determined by setting $\lambda = \min\{G_{ST}^{(Turner)}(ij,kl)\}/\min\{G_{ST}(pq,rs)\}$, where the numerator is the minimum experimental base-pair stack energy[31] and the denominator is the minimum base-pair stack statistical energy. Any statistical energies that are higher than the corresponding maximum experimental value are set equal to the maximum experimental value.

Equations (6) –(8) are used as a common treatment of base-pair stacks where symmetric base-pair stacks such as UA/CG (Figure 5.1A) and CG/UA (Figure 5.1B) are considered to be equivalent. For example, the base-pair stack indicated in Figure 5.1A and Figure 5.1B are degenerate since the same nucleotides are on the 5' end of the strands (namely U and C) while A and G are on the 3' end of the strands in both figures. These two rotated configurations are commonly considered to be equivalent and their statistics would be the average of the two conformations. However, such treatment may not accurately represent the distribution of base-pair stack configurations and asymmetry of directionality of RNA structures, such as the consideration of individual nucleotides that are immediately 5' and 3' to a hairpin loop. Hence, we will investigate the asymmetric statistical energy as well. Only slight modifications are needed to equation (7) by eliminating the sum and equation (8) by eliminating the delta function. The statistical energy is normalized by the reference state, $P^{(rand)}$, calculated from the probability of finding each individual nucleotide in the sequences of a given molecule.

We have developed molecule-specific energies by sampling sequences that are specific to particular molecules. For example, to evaluate tRNA-specific energies, we sample from the set of tRNA sequences.

5.2.3 Additional Statistical Energy Terms

In addition to base-pair stacks, the energetics of other secondary structural motifs, although limited in Mfold, may be modified to further refine our free energy calculations and potentially improve folding results. Statistical potentials can be developed by identifying nucleotides at the ends of hairpin loops and nucleotides that flank those loops as shown in Figure 5.1C. We let *i* be the nucleotide in a base-pair that surround the 5' end of the hairpin loop, *j* be the nucleotide on the 5' end of the loop, *k* be the nucleotide on the 3' end of the loop, and *l* be nucleotide in a base-pair (with nucleotide *i*) that surround the 3' end of the loop. Then the free energy of hairpin flanks can be evaluated as follows:

$$\Delta G_{HF}(ij,kl) = -\lambda k_B T \ln\left(\frac{P_{HF}(ij,kl)}{P_{HF}^{(rand)}(ij,kl)}\right) \tag{9}$$

where

$$P_{HF}(ij,kl) = \frac{N_{HF}(ij,kl)}{N_{HF}} \tag{10}$$

and

$$P_{HF}^{(rand)}(ij,kl) = P_i P_j P_k P_l \,. \tag{11}$$

The indices *i* and *j* represent any of the nucleotides A, C, G, or U. We define $N_{HF}(i, j)$ as the number of hairpins composed of nucleotides, *i* and *j*, surrounding the motifs. The total number of hairpin flanks is N_{HF}. P_i is the probability of occurrence of nucleotide *i*. The scaling factor, λ, is estimated in a similar manner to the previous scaling factors by comparing the minimum value of experimental data[31] with the minimum value of $\Delta G_{HF}(i, j)$. Furthermore, hairpin flank statistical potentials are limited to the highest energy found in experimental results.

Internal loops are unpaired nucleotides surrounded by helices as shown in Figure 5.1D. Energy terms for internal loops are derived using nucleotides of the loops on the 5' and 3' strands as well as the base-pairs surrounding those nucleotides. Loops of different

lengths on each strand use a different energy function. For the example of internal loops with 2 nucleotides on the 5' and 3' ends, we evaluate the free energy of internal loops as:

$$\Delta G_{IL}(i1,i2,i3,i4,\ j1,\ j2,\ j3,\ j4) = -\lambda k_B T \ln\left(\frac{P_{IL}(i1,i2,i3,i4,\ j1,\ j2,\ j3,\ j4)}{P_{IL}^{(rand)}(i1,i2,i3,i4,\ j1,\ j2,\ j3,\ j4)}\right)$$

where

$$P_{IL}(i1,i2,i3,i4,\ j1,\ j2,\ j3,\ j4) = \frac{N_{IL}(i1,i2,i3,i4,\ j1,\ j2,\ j3,\ j4)}{N_{IL\_2-2}} \tag{13}$$

and

$$P_{IL}^{(rand)}(i1,i2,i3,i4,\ j1,\ j2,\ j3,\ j4) = P_{i1}P_{i2}P_{i3}P_{i4}P_{j1}P_{j2}P_{j3}P_{j4} \tag{14}$$

.

The indices *i1* and *i4* are the base-pairs around the internal loop on the 5' end of the molecule. Indices *i2*, and *i3* are the nucleotides consisting of the internal loop on the 5' end of the molecule. Indices *j1*, *j2*, *j3*, and *j4* are the corresponding nucleotides on the 3' end of the molecule. We define N_{IL}(*i1*, *i2*, *i3*, *i4*, *j1*, *j2*, *j3*, *j4*) as the number of internal loops composed of nucleotides, *i1*, *i2*, *i3*, *i4*, *j1*, *j2*, *j3*, and *j4*. Similarly, internal loops with 1 nucleotide on the 5' and 3' ends are evaluated as:

$$\Delta G_{IL}(i1,i2,i3,\ j1,\ j2,\ j3) = -\lambda k_B T \ln\left(\frac{P_{IL}(i1,i2,i3,\ j1,\ j2,\ j3)}{P_{IL}^{(rand)}(i1,i2,i3,\ j1,\ j2,\ j3)}\right)$$ **Disp**

.

In this case, *i1*, *i3*, *j1*, and *j3* are the surrounding base-pairs while *i2* and *j2* are the nucleotides of the internal loops. Finally, internal loops with 1 nucleotide on one strand and 2 nucleotides on another strand are evaluated as:

$$\Delta G_{IL}(i1,i2,i3,\ j1,\ j2,\ j3,\ j4) = -\lambda k_B T \ln\left(\frac{P_{IL}(i1,i2,i3,\ j1,\ j2,\ j3,\ j4)}{P_{IL}^{(rand)}(i1,i2,i3,\ j1,\ j2,\ j3,\ j4)}\right)$$ **D**

.

The nucleotides *i1*, *i3*, *j1*, and *j4* are the surrounding base-pairs and the rest are the internal loops. The scaling factor, λ, was estimated similarly to the previous scaling factors by comparing the minimum value of experimental data[31] with the minimum value of $\Delta G_{IL}(i1, i2, i3, j1, j2, j3, j4)$. Note that λ is uniquely calculated for each equation. Furthermore, hairpin flank statistical potentials are limited to the highest energy found in experimental results. All statistically derived energies can be downloaded and applied to the original Mfold program at http://biomol.bme.utexas.edu/~wuch/statistical_energies.

5.2.4 Statistical Energy Derived from All-Sequence Dataset

In addition to molecule-specific potentials, sequences from all molecules are combined to derive the all-sequence statistical energy that can potentially be applied to prediction. For base-pair stacks, we have directly combined all sequences from all molecules to derive the statistical energy. For hairpin flanks and internal loops, however, due to the limited occurrences in smaller molecules such as the 5S rRNA and tRNA, a different approach was used to combine the statistics from individual molecules. Hence, the all-sequence hairpin flank statistical energy and all-sequence internal loop statistical energy are derived by averaging the molecule-specific statistical energies from all molecules using Boltzmann weights:

$$\Delta G^{P}_{ALL-SEQ} = \frac{\sum_{i \in I} \exp(-\Delta G^{P}_{i} / k_b T) \Delta G^{P}_{i}}{\sum_{i \in I} \exp(-\Delta G^{P}_{i} / k_b T)} \tag{16}$$

For each structural element P, such as the base-pair stack and internal loops discussed above, the all-sequence statistical energy is a weighted sum of energies from each molecule *i*.

5.2.5 Evaluation of Statistical Potentials

We applied different combinations of base-pair stack (BP-ST), hairpin flank (HF), and internal loop (IL) statistical energies (SE) to the Mfold program[31] developed by Zuker et al to test its ability to predict RNA folding. The Mfold program was also modified to utilize asymmetric BP-ST SE in energy calculations. Five sets of asymmetric SE were derived from tRNA, Eukaryotic 5S rRNA, Bacterial 5S rRNA, Bacterial 16S rRNA, and all sequences combined. Each set of SE utilizing BP-ST and BP-ST-HF-IL terms were used to fold each of the four molecules.

In order to make performance comparisons of energy values, the number of base-pairs of a structure predicted by Mfold that are in agreement with comparative sequence analysis[27] is divided by all base-pairs determined by comparative sequence analysis. Only canonical (Watson-Crick G-C, A-U, and G-U) base-pairs are counted. Although this performance measurement does not count erroneously predicted base-pairs, false positives affect the predicted structure and may preclude correct base-pairs from forming. This metric is simple and sufficient in evaluating the accuracy of folding.

5.3 RESULTS AND DISCUSSION

5.3.1 Correlation of Base-Pair Stack Statistical Energies and Experimental Thermodynamic Stability

With comparative structural information, we have examined the base-pair stack statistics and their correlation with thermodynamic stability extracted from RNA duplex melting experiments**[32]**. Since statistically derived energies require a reference value to set the absolute scale, we normalized the lowest base-pair stack (BP-ST) statistical energies (SE) with the lowest BP-ST experimental value (See Material and Methods). We have computed molecule-specific BP-ST SE for tRNA (for amino acids A, D, E, G, I, L, M, F), Eukaryotic 5S rRNA, Bacterial 5S rRNA, and Bacterial 16S rRNA. A BP-ST SE

was also determined by combining all the sequences from the four datasets. In Figure 5.2, we plot the statistical energy against experimental data. The correlation coefficient and standard deviation of difference between SE and experimental values for each BP-ST are reported in Table 5.1. The highest correlation coefficient (0.870) between the statistical and experimental free energies is found for Bacterial 16S rRNA, followed closely by Eukaryotic 5S rRNA (0.857) and tRNA (0.833). The statistical energies for Bacterial 5S rRNA is the least correlated with experimental stability (0.477). More than 80% of the nucleotides in the all-sequence dataset are from Bacterial 16S rRNA, thus the calculated BP-ST SE is biased by the Bacterial 16S rRNA.(Table 5.1). As a result, the correlation of the all-sequence base-pair SE with experimental stability is also high (0.86). The standard deviations for all-sequence base-pair stack SE, Bacterial 16S rRNA, Bacterial 5S rRNA, Eukaryotic 5S rRNA, and tRNA are 0.53, 0.51, 1.19, 0.68, and 0.99, respectively. In comparison, the PDB-derived base-pair stack energies have a correlation coefficient of 0.78 with experimental stability (Figure 5.2F) and a standard deviation of 0.81 [6].

For comparison with experimental base-pair stack free energy we have used symmetric SE for correlation calculations. Hence, 5' and 3' directionality are equivalent, e.g. GC/CG and CG/GC are considered to be the same structure. However, asymmetric SE, in which 5' and 3' directions are differentiated, will be used to study the folding applications of BP-ST SE. Overall the statistics of base-pair stack frequency show a good correlation with experiment. However a few, base-pair stacks differ more significantly. One of them involves consecutive UG base pairs. The UG/GU base-pair stack statistical energy (SE) from the Bacterial 16S rRNA(-1.0 kcal/mol) and all molecules (-0.82 kcal/mol)(See Appendix Tables S 2.4 and S2.5) datasets are much lower than that used by Mfold (-0.5 kcal/mol). However, the value for that base-pair stack in Mfold has been adjusted from the experimental value of 0.47 kcal/mol[31]. Additionally, the

experimental stability of UG/GU has the largest error (0.96 kcal.mol) among all the base-pair stacks since only one duplex containing the UG/GU was measured which was insufficient for linear regression[31]. The statistical potentials for the UG/GU base-pair stack derived from PDB structures also varied significantly from the experimental energy values[6].

The largest discrepancy between the all-sequence SE and experimental stability is for the UG/UG (-1.23 kcal/mol) and GU/GU (-0.12 kcal/mol) base-pair stacks (Appendix Table S 2.5). The corresponding experimental energy for UG/UG is 0.30 kcal/mol and GU/GU is 1.30 kcal/mol (Appendix Table S 2.12). As noted earlier, the statistical energy for these are significantly different from the experimental data[31]. The UA/AU base-pair stack statistical energies (Appendix Tables S 2.1 – S 2.5) are all consistently below -2.3 kcal/mol and differ from the experimental value by more than 1 kcal/mol. This indicates that UA/AU base-pair stacks occur more frequently in these RNA molecules than suggested by the stability derived from duplex melting experiment. The GC/GC and GC/CG base-pair stacks determined from calorimetric experiments are the most stable structures with the stability of the former being slightly lower than the latter by 0.1 kcal/mol. However, the lowest (i.e. most stable) base-pair stack SE is GC/CG for the all-sequence, Bacterial 16S rRNA, and Bacterial 5S rRNA datasets. The difference between the SE of the GC/CG base-pair stack for each aforementioned datasets (Appendix Tables S 2.3 – S 2.5) and the corresponding experimental value (Appendix Table S 2.12) is 0.9 kcal/mol on average. The GC/GC BP-ST SEs do not immediately follow GC/CG in terms of stability. The PDB-derived potential[6] GC/CG base-pair stack minimum agrees with that of our BP-ST SE.

Overall, base-pair stack SE derived from comparative sequences analysis show a better correlation with experimental free energy[33] than the PDB-derived potentials.

However, noticeable differences exist between SEs of different phylogenetic domains, which concur with experimental observations. For example, although the Escherichia coli loop E of Bacterial 5S rRNA and *Spinacia oleracia* of Eukaryotic 5S rRNA are found to be isosteric and are able to bind to the same L25 protein, they show substantial differences of stability in various ionic conditions[34]. In addition, sequences of RNA molecules have been confirmed by Kiparisov *et al* to be highly specific and optimized through evolution as only 7 alleles in *Saccharomyces cerevisiae* 5S rRNA are found to be viable[35]. The high correlation indicates that the distribution of base-pair stacks in three out of the four molecule specific datasets, or when combined, is Boltzmann-like. The agreement is rather remarkable especially given that the experiments were performed on isolated duplexes while the statistics were collected from biological sequences, which also represents tertiary contacts and many other factors in the cell. For example, structural studies have supported that hairpin loops can be stabilized by tertiary interactions that are not considered in the oligomer experiments[36-38].

5.3.2 Application of Statistical Energies to Folding

We investigated the ability of the statistically derived energies of BP-ST, hairpin flanks, and internal loops to improve the prediction of an RNA's secondary structure. We incorporated our statistical-energies in place of those energy parameters determined experimentally in the Mfold program. Statistical energy (SE) derived from the molecule-specific and the all-sequence datasets are utilized within Mfold to predict the secondary structure for tRNA, Eukaryotic rRNA, Bacterial 5S rRNA, and Bacterial 16S rRNA.

Similar to previous studies[31], the folding accuracy is evaluated by comparing base-pairs that are accurately predicted by Mfold with base-pairs determined by comparative sequence analysis[27] (details in Methods Section 5.2.5). Although Mfold

determines the optimal (most stable) structure and a set of sub-optimal secondary structure models, the structure with the lowest energy was used in our analysis for folding accuracy. Furthermore, the experimental energy values were obtained from oligonucleotide duplexes where base-pair stacks have no directionality. However, the anisotropic nature of a base-pair stack is apparent when, for example, it is adjacent to a hairpin loop. Thus, equation (6) and its corresponding modifications were used to evaluate symmetric and asymmetric BP-ST energies for folding accuracy. As shown in Appendix Tables S 2.6 – S 2.10, asymmetric SE offered slightly better folding results overall and will be used in the folding evaluation. Figure 5.4 provides the accuracy of folding tRNA, Eukaryotic 5S rRNA, Bacterial 5S rRNA, Bacterial 16S rRNA, using BP-ST SE derived from each molecular and phylogenetic dataset and the all-sequence dataset. The accuracy in secondary structure prediction is either improved or remains the same when the SE derived from each dataset is applied to the sequences within the same dataset. For example, when tRNA-specific SEs are used in place of the experimental BP-ST energies to fold tRNA sequences, the accuracy improved from 0.70 to 0.79. The improvement in the prediction accuracy is more significant for Bacterial 5S rRNA; 0.74 vs. 0.63 for SE and experimentally derived energy values. However, for other datasets (e.g., Bacterial 16S rRNA and Eukaryotic 5S rRNA) the prediction accuracy was similar between the molecule-specific SEs and the experimentally derived energy values. The prediction accuracy usually decreases when the BP-ST SE determined from one molecule/phylogenetic dataset is utilized to predict the secondary structure for another dataset. And last, prediction accuracy for the SE derived from all-sequence dataset or the Bacterial 16S rRNA is similar with the experimentally determined energy values. These results indicated that, the statistics of base-pair stacking in these two datasets is indeed

Boltzmann like, and the statistically derived energy is as reliable as the experimental measurements.

5.3.3 Extension of Statistical Energies to Hairpin Flanks and Internal Loops

Our analysis has shown some improvement in the prediction of RNA secondary structure helices from the structural statistics of consecutive base pairs within a helix. About 66% of the nucleotides in an RNA structure predicted with covariation analysis form a base pair. And the vast majority of these are G:C, A:U, and G:U pairings that occur within a regular helix. The remaining third of the RNA secondary structure form hairpin, internal, and multistem loops. However, an analysis of the three-dimensional structure and our growing knowledge about the variety of structural motifs that occur in RNA structure provides a foundation to improve the accuracy in the prediction of an RNAs secondary and even tertiary structure.

The majority of these unpaired nucleotides in the secondary structure are base-paired with non-canonical pairing types and conformation[39] . And nearly all of the nucleotides in the rRNA high-resolution structure are stacked onto another nucleotide. While the majority of these stackings are formed between adjacent nucleotides that are base-paired, a significant number of nucleotides that are stacked are not base-paired or consecutive.

Towards that end, Lee, Gardner, and Gutell (manuscript in preparation) have identified and characterized a set of different types of base stackings at the ends of helices that add stability to the helix and potentially protect the ends of the helix while bridging the ends of regular helices with different structural motifs that occur in the hairpin, internal, and multistem loops in an RNA secondary structure. Many of these helix cappings are associated with numerous structural motifs that have been identified

and characterized for their chemical structure and energetic properties. One example, the UAA/GAN motif[40] has several non-canonical base pairs and unpaired nucleotides that form a longer co-axial stack that bridges the two flanking helices. Another example, the E loop and the E-like loop have been well characterized[41, 42] and also contain several non-canonical base pairs that form a contiguous stack onto the regular secondary structure helix.

While the current thermodynamic based folding algorithms consider these unpaired regions of the secondary structure to be destabilizing, as stated earlier, it is known that base stacking contributes more to the overall stability of the RNA structure. While our longer term objectives are to determine the structural potentials associated with all of the structural motifs in the unpaired regions of the secondary structure, for this paper we only determine those structural potentials that can be utilized within the Mfold program.

The statistical energies for hairpin flanks (HF) and internal loops (IL) have been evaluated and utilized in the prediction of an RNA secondary structure with Mfold. HF and IL SE are computed according to equations (9), (12), (14) and (15). The statistical energies for HF, IL (1x1, 1x2, 2x2, and flanks) are available in Appendix Tables S 2.16 – 2.20, respectively. The minimum and maximum SE values were calibrated with those determined in experiments[31]. Statistical energies are derived with comparative structures from several RNA datasets: tRNA, Eukaryotic 5S rRNA, Bacterial 5S rRNA, and Bacterial 16S rRNA. All-sequence SE is obtained from the Boltzmann-average of the individual datasets since the occurrences of HF and IL are sparse in structures. We then replaced the energy values for HF-IL in Mfold with the corresponding SE, and set the rest to maximum experimental energy values.

We compare the SE of those motifs from the all-sequence dataset with the corresponding free energy obtained from experiment. Hairpin flanks yield a low correlation with experiment (0.5132) which is computed from SE in Appendix Table S 2.16 and experimental values from Mathews work[31]. We analyze the differences in SE values between each of the datasets. SE values are only shown for the Base-pair/hairpin flanks that occur within each dataset. The arrangement of the nucleotides is described in the following example. The hairpin flank denoted GC-AC, as illustrated in Figure 5.1C, is composed of a GC base-pair at the end of a helix and the A is an unpaired nucleotide flanking the paired G and the C is an unpaired nucleotide flanking the paired C. These nucleotides correspond to the first two - columns of the Appendix Table S 2.16. With six types of helix ending base pairs (GC, CG, etc.) and 16 possible unpaired flanking pairs (AA, AC, AG, AU, etc.), 96 possible BP/HF flanks are possible. Of the 55 observed, 27 occur in only one dataset (excluding the all-sequence dataset), 16 occur in two datasets, nine occur in three datasets, and three occur in all four datasets.

The most pronounced patterns are:

- The hairpin flanking nucleotides that are associated with the most helix ending base pair types is GA, occurring with six different base pairs, followed by CA, UC and UU that both have four different helix ending base pairs.
- The vast majority of the most stable helix flanking nucleotides in hairpin loops are GA, UC, UU, and CA with most helix ending base pair types.
- The other very stable BP/HF nucleotide sets that do not have a GA, UC, UU, or CA hairpin flank are: CG/AA; CG/AC; GC/AC; GC/AG; UA/UA; CG/UG.

Several of these BP/HF sets are associated with known tetraloops, although not exclusively. For example, CG/UG is associated with one of the most stable statistical energies occurs frequently in numerous RNA molecules with the sequence C(UUCG)G;

This motif has been determined to be very stable[43, 44] . Many but not all of the family of BP/HF nucleotide sets with a GA hairpin flank are associated with the GNRA tetraloop[43, 45]. The closing base pairs of these tetraloops were investigated experimentally and generally consistent with our statistical energies for the BP/HF[46]. Some of the other GA BP/HP are associated with hairpin loops with six nucleotides. Two experimental studies revealed that the G and A in the hairpin flank form a base pair with a sheared conformation[47, 48].

A previous study revealed that many helices in the ribosomal RNAs have an AA or AG flanking the end of the helix in the unpaired region of the secondary structure[36]. This analysis revealed that all of the GA flanks associated with a hairpin loop form a base pair. Two of the three most stable BP/HF statistical energies have a helix ending CG base pair and a AA (-2.81) and GA (-2.8) flank (the statistical energy for the most stable BP/HF - AU/UC is -2.87). An experimental study revealed that the hairpin flank GA is more stable than AA for a hairpin loop with the same intervening four nucleotides[49]. We anticipate that the accuracy of the prediction of an RNA secondary structure will be further enhanced once we associate the BP/HF statistical energies with the size and sequence of the hairpin loop.

The 1x1 IL-SE has a low correlation with experiment (0.18) as computed from Appendix Table S 2.17. Two of the most stable BP/IL/BP 1x1 internal loops has a UU juxtaposition. The next most stable BP/IL/BP 1x1 internal loop has a GA juxtaposition. The UU juxtaposition in the 1x1 internal loops occurs with eight different sets of base pair types that flank both sides of the non-canonical set of nucleotides. Of these, five are considered more stable with statistical energy values > -1.0. The GA and AG juxtapositions in the 1x1 internal loops are both associated with seven different sets of flanking basepairs. However, while six of the seven GA's have a stability greater than -

1.0, only two of the AG's have an energy value greater than -1.0. Five of the six UC juxtapositions have an energy value greater than -1.0 and all three of the CU juxtapositions have energy values greater than -1.0. While some of our statistical energies are similar to those determined with experiments, others are very different. For example the UG/U-U/AU (BP/IL/BP) loop statistical energy is underestimated (-1.59 kcal/mol) compared to the 1.7 kcal/mol obtained from experiment[31]. The AU/G-G/GC loop SE (-0.45 kcal/mol) is also underestimated, compared with the -1.4 kcal/mol of experiment. No 1x1 IL (Appendix Tables S 2.17), were observed in the tRNA datasets we analyzed. The Eukaryotic 5S rRNA contains six and the Bacterial 5S rRNA has eight. All of these are considered stable (energy value greater than -1.0 and all have a different set of BP/IL/BP sequences).

The number of occurrences of internal loops with one nucleotide on one side of the helix and two on the other is low. None were present in the tRNA and two 5S rRNA datasets analyzed. Seven were present in the Bacterial 16S rRNA. Of the nine 1x2 internal loops, 24 of the 27 nucleotides are a purine. The correlation between the 1x2 IL-SE (0.11) with experiment is low; The largest underestimation of SE is for the GC/A-AA/AU loop (0.31 kcal/mol) compared with experiment (3.2 kcal/mol).

Twenty five different 2x2 internal loops with unique sets of nucleotides within the internal loop and the two base pairs that flank the internal loop occur in the Bacterial 16S rRNA dataset. These 2x2 internal loops are not in the tRNA and 5S rRNA datasets analyzed here. Of the 25 different 2x2 internal loops, nine of them have tandem GA/AG internal loop, four have the tandem AA/AG IL, four more have the GA/AA IL, and two have the AA/AA IL. In total, 19 have one of the 2x2 internal loops that form a unique structural motif family[50]. The G in the GA juxtaposition can be replaced with an A and still maintain the same sheared conformation in many of the tandem GA motifs.

Experimental studies revealed an association between the base pairs that flank the tandem GA 2x2 internal loop[51] The 2x2 internal loop in three of the 27 BP/2x2 IL/BP is UU/UU. Tandem UU mismatches have been studied experimentally and are stable in some structural environments[52]. The 2x2 IL-SE has a low correlation with experiment as well (0.08). The AU/AA-AG/GC loop is underestimated with -4.70 kcal/mol while experiment has 1.00 kcal/mol.

While the SE values determined from the nucleotide frequencies of consecutive base pairs are similar to and sometimes slightly better than the experimentally determined energy values for the same consecutive base pairs, the incorporation of hairpin flank and internal loop SE terms significantly increase the accuracy of the prediction (Figure 5.5). Using all statistical energies (base-pair stack, hairpin flank, and internal loop), we attain an increase in folding accuracy from 0.70 to 0.89 in tRNA, from 0.72 to 0.84 in Eukaryote 5S rRNA, from 0.63 to 0.88 in Bacterial 5S rRNA, and from 0.49 to 0.56 in Bacterial 16S rRNA. The substantial improvement suggests that the structural elements, by themselves and/or coordinated with the ends of the secondary structure helix stabilizes the higher-order structure of the RNA molecule. The statistical energies for the hairpin flank contributed more to the improvement in the prediction of an RNA secondary structure than for the internal loops and base-pair stacks. In particular, while the folding accuracy increased from 0.70 (for experimental energy values) to 0.89 (for all statististical energies), statistical energies for only internal loops increased the accuracy to 0.78, while statistical energies for only hairpin flanks increased the accuracy to 0.85. Accordingly, the folding accuracy of Eukaryotic 5S rRNA when only hairpin and only internal loop statistical energies are used are 0.79 and 0.74, respectively, in constrast with 0.72 (experimental) and 0.84 (for all statistical energies). However, statistical energies for hairpin flanks do not always contribute more than internal loop flanks to the

accuracy of the prediction of an RNA secondary structure. The internal loop flank statistical energy increases the accuracy of RNA folding more than hairpin flanks for Bacterial 5S rRNA. This may be due to the small sampling of the molecule. While the improvement in the folding accuracy with HF SE is not as significant for Bacterial 16S rRNA, the overall improvement does occur. This observation emphasizes the significance of hairpin flanks to RNA structures.

While internal loops within 5S rRNA ranges from 1 – 8 nucleotides in length and 16S rRNA range from 1 – 12 nucleotides in length, only energies for internal loops of lengths with less than four nucleotides can be utilized by Mfold. As a result, the statistical energy contributions for the longer internal loops are not used. Similarly, Mfold has energy functions based on the length and nucleotide composition of hairpins and yet cannot assign an energy contribution for specific hairpin loops with more than 4 nucleotides. This is particularly important since tRNA are mainly composed of loops of lengths 7 and 8, while 5S rRNA are composed of hairpin loops that are longer than 4 nucleotides as well. Hence, the inclusion of energy parameters for larger hairpin loops in to the folding algorithm should increase the prediction accuracy. Although bulge loops are known to be important to the stability of RNA, the simple model that only accounts for the length of the loop to be used as input to evaluate energy does not lend itself to much improvement when statistical methods are applied. Additional development of Mfold would be needed to incorporate statistical energies of bulge loops. Furthermore, many unpaired nucleotides in an RNA's secondary structure participate in tertiary interactions to further stabilize the structure[53-56]. Information on such interactions may be needed to determine loop energies. In addition, a special bonus is used by Mfold when empirical results deviate from the general rules. For example, hairpin loops are checked for a special GU closure and given a bonus[31]. Due to these limitations, a new

framework beyond the current Mold is necessary to utilize the statistical energies that can be evaluated from comparative structures.

While the statistical energies derived from one molecular/phylogenetic dataset have the potential to significantly improve the prediction accuracy for sequences in the same dataset (i.e. self-folding), we expect the same set of statistical energies derived from numerous molecular RNA sequences that span the phylogenetic tree of life can be determined and utilized by an RNA folding algorithm to accurately predict the secondary structure for any RNA molecule.

The accuracy of folding each molecule using statistical energies derived from molecule-specific and all-sequence datasets is shown in Figure 5.6. The statistical energies for BP-ST-HF-IF structural parameters for the all-sequence dataset did predict the secondary structures more accurately (~0.1) with Mfold than the experimentally derived energy values for tRNA, Eukaryotic 5S rRNA, and Bacterial 5S rRNA (tables or figures that substantiate this claim). The improved accuracy for the Bacterial 16S rRNA dataset was moderate (~0.05). However, as noted earlier, Bacterial 16S rRNA has a larger number of HF and IL structural elements that cannot be utilized by the current Mfold program.

The use of hairpin flank and internal loop energies with base-pair stack energies will enhance the prediction accuracy for some self and cross folding, and decrease the prediction accuracy for other folding. For example, tRNA-specific BP-ST-HF-IL statistical energies increase the accuracy to 89% of the base-pairs in tRNA. However, the statistical energies derived from Eukaryotic and Bacterial 5S rRNA decrease the accuracy for sequences in the tRNA dataset to 0.62 and 0.65, respectively. This corresponds to a decrease by almost 0.3 when the SE values for a different molecule/phylogenetic dataset are used. However, when only BP-ST SE is examined (Figure 5.5), the difference in

folding accuracy is about 0.1 between the tRNA-specific SE and SE from other molecules. Similar trends are observed for other molecules. This is again due to the infrequent occurrences of secondary structures such as hairpin flanks and internal loops as opposed to BP-ST. It is possible that additional comparative structures from other molecules in the future would be helpful. Currently, it appears that Boltzmann average between molecular contributions is an effective approach to retain the "signals" and to derive a general SE for these motifs.

5.4 CONCLUSIONS

Statistical energies generated from structural statistics of sequences are found to agree with experimental. The statistical potentials of base-pair stacks achieve a correlation coefficient of ~0.9 between the energies derived structural statistics and the free energy values extracted from experiments. Statistics from individual molecules, such as Bacterial 5S rRNA and tRNA, express specificity, and to an extent, rigidity as Smith et al[40] have found nucleotides in Loop E, Helix I, and Helix IV that are lethal to *Saccharomyces cerevisiae*. However, statistics from a single molecule may not sufficiently represent the complete Boltzmann distribution of base-pair stacks due to biologically-driven biases as well as the small sample size of each small molecule as shown by the number of nucleotides in Table 5.1, although the sampling issue will improve as more sequences are aligned. Conversely, the dataset for Bacterial 16S rRNA has a vast sample size (slightly less than 1.5 million nucleotides) and, subsequently, represents a Boltzmann-like distribution.

We further evaluated the SE for base-pair stack, internal loop and hairpin-flank by applying them in Mfold to predict the secondary structures. The statistical energies have been derived from sequences of individual molecules (molecule-specific) and all

combined (all-sequence). Molecule-specific base-pair stack energies improve folding accuracy for some molecules and have little effect on others, as compared to the experimental values used by original Mfold. Using all-sequence base-pair stack SE, we observe the accuracy of folding prediction to be comparable to that of free energy obtained experimentally. When SE for hairpin flanks and internal loops are included, we see dramatic improvements in the folding accuracy to 0.80, 0.79, and 0.77, and 0.53 for tRNA, Eukaryotic 5S rRNA, Bacterial 5S rRNA, and Bacterial 16S rRNA, respectively. Much of the improvement in accuracy is due to the application of hairpin flank statistical energies. Since Mfold does not utilize energy parameters for structures such as internal loops that have more than 3 nucleotides, not all statistical energies can be employed.

Overall, the prediction accuracy when utilizing the SE values from one dataset to the sequences in another dataset is worse than the prediction accuracy for the experimentally determined energy values. Consistent with our previous analysis, the results suggest that individual RNA molecules have insufficient HF and IL statistics for a Boltzmann-like distribution and cares need to be taken when combining the statistics from different molecules into general statistical energies. More importantly, we have demonstrated here that motifs beyond BP-ST are critical to a more complete understanding of RNA folding and to the refinement of folding algorithms.

Table 5.1: For statistical energies derived from sequences of each specified molecule, the number of nucleotides in base-pair stacks, correlation coefficient and standard deviation compared with experimental base-pair stack energies[31] are presented. The last row lists equivalent information for PDB-derived statistical potentials.

	Number of Nucleotides	Number of Sequences	Corr Coef	Std Dev
Bacterial 16S rRNA	1 468 052	319	0.870	0.507
Bacterial 5S rRNA	34 443	96	0.477	1.194
Eukaryotic 5S rRNA	95 778	263	0.857	0.677
tRNA	148 248	650	0.833	0.988
All-sequence	1 746 521	1328	0.862	0.525
PDB-derived[b]	7424[c]	N/A	0.7764	0.8138

[a]Mathews, 1999
[b]Dima, 2005
[c]Number of bases in stacks

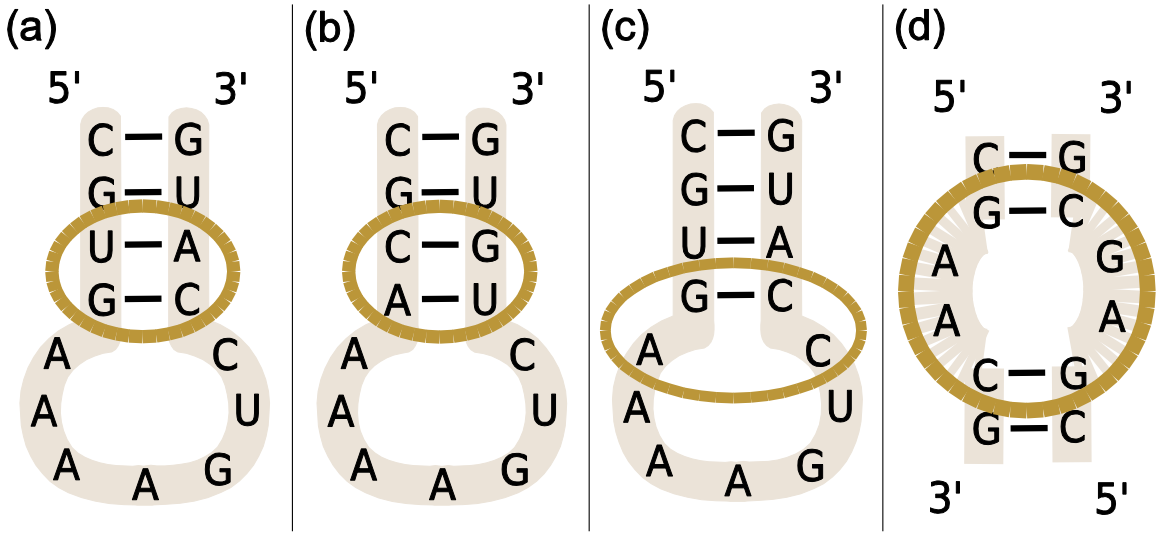


Figure 5.1: Depictions of 4 secondary structures. (a) An example of a base-pair stack that is denoted UA/CG. (b) An example of a base-pair stack that is denoted CG/UA. (c) An example of a hairpin flank that is denoted GC/CA. (d) An example of an internal loop that is denoted GC/CG/AG/AA.

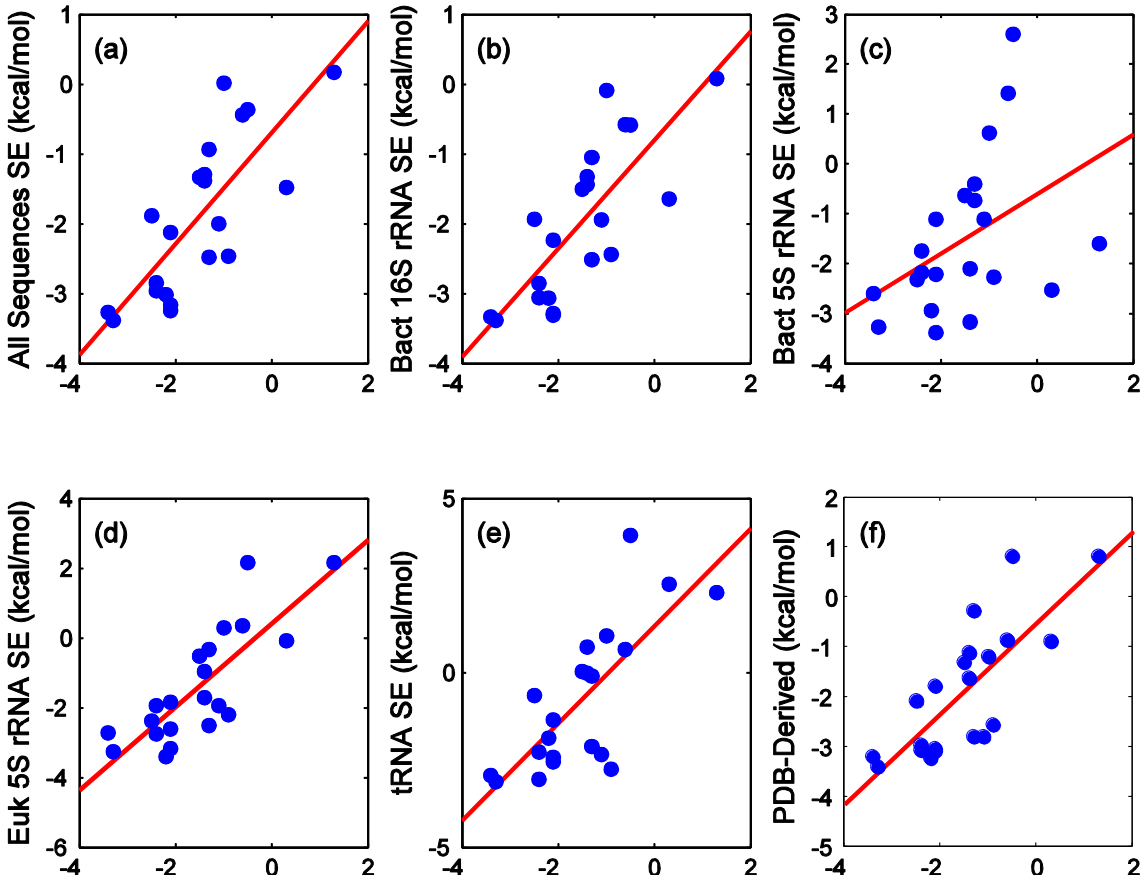


Figure 5.2: (a) Base-pair stack statistical energy (SE) derived from all-sequence dataset, (b) Bacterial 16S rRNA, (c) Bacterial 5S rRNA, (d) Eukaryotic 5S rRNA, and (e) tRNA versus free energy obtained experimentally for a given base-pair stack. (f) PDB-derived statistical potentials versus free energy obtained experimentally. Experimental values are in kcal/mol.

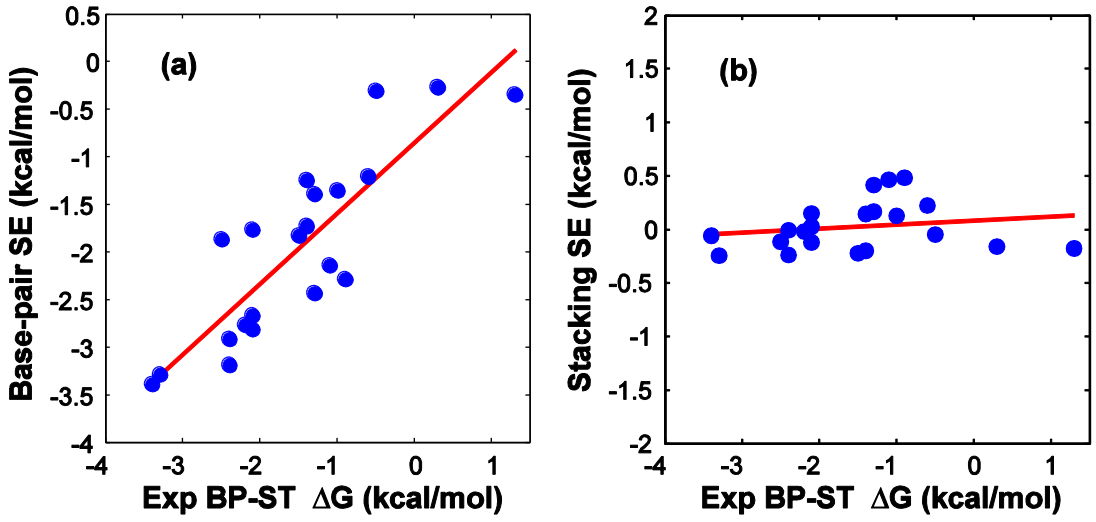


Figure 5.3: (a) Base-pair statistical energy versus base-pair stack statistical energy with a correlation coefficient of 0.8743. (b) Stacking statistical energy versus base-pair stack statistical energy with a correlation coefficient of 0.0071.

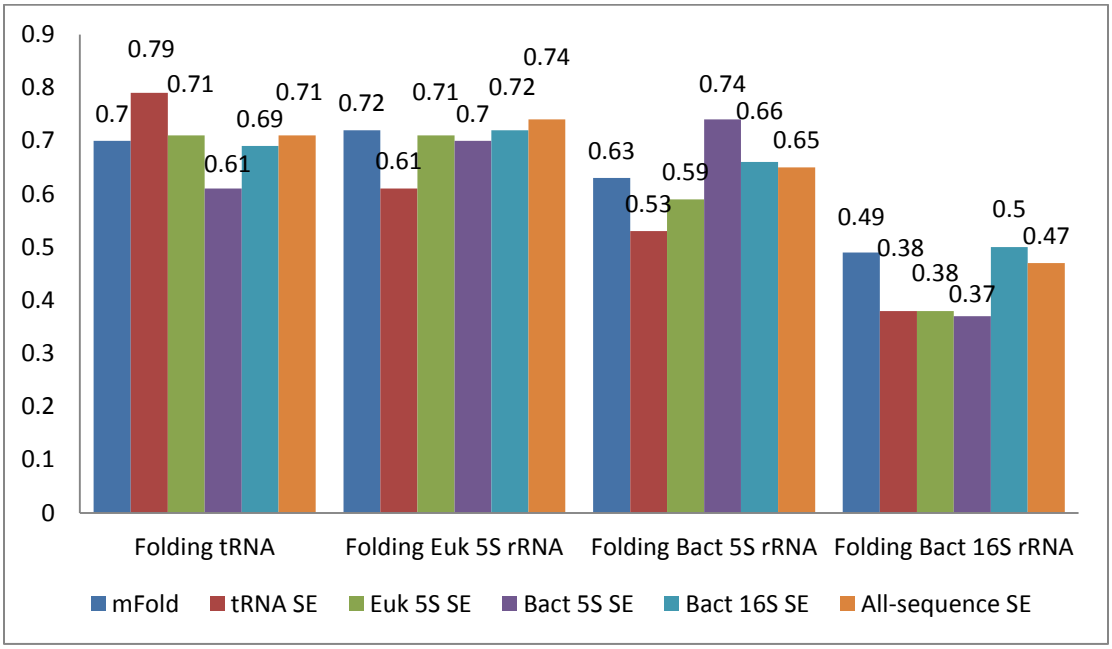


Figure 5.4: Each group of bars represents folding accuracy of (from left to right) tRNA, Eukaryotic 5S rRNA, Bacterial 5S rRNA, and Bacterial 16S rRNA. Within each group, each bar represents (from left to right) unmodified Mfold, base-pair stack SE derived using tRNA, Eukaryotic 5S rRNA, Bacterial 5S rRNA, Bacterial 16S rRNA, and all-sequence dataset.

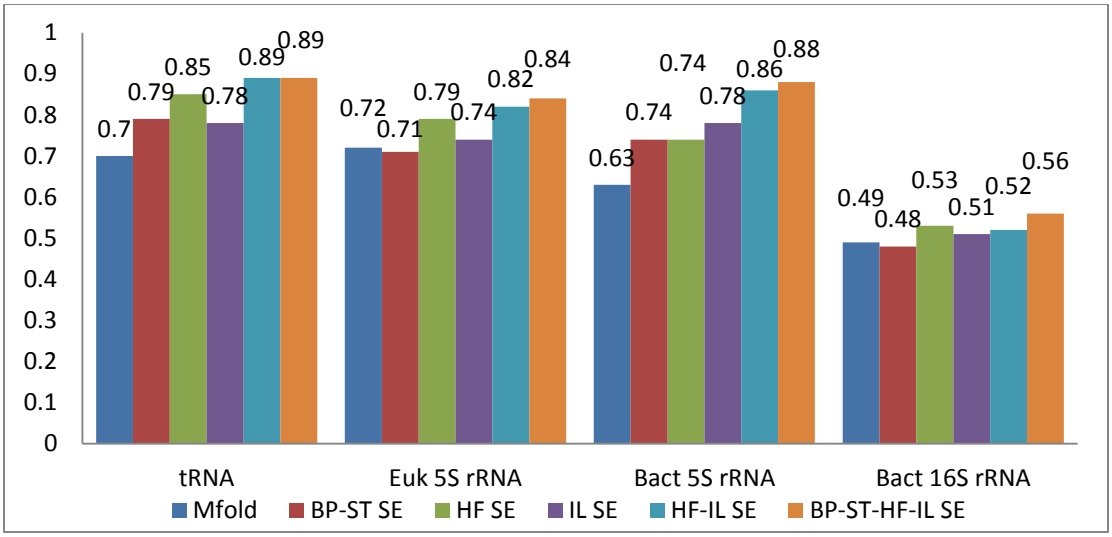


Figure 5.5: Each group of bars represents the folding accuracy of (from left to right) tRNA, Eukaryotic 5S rRNA, Bacterial 5S rRNA, and Bacterial 16S rRNA. Within each group, each bar represents (from left to right) unmodified Mfold, base-pair stack SE, hairpin flank SE, internal loop SE, and all available SE (base-pair stack, hairpin flank, and internal loops).

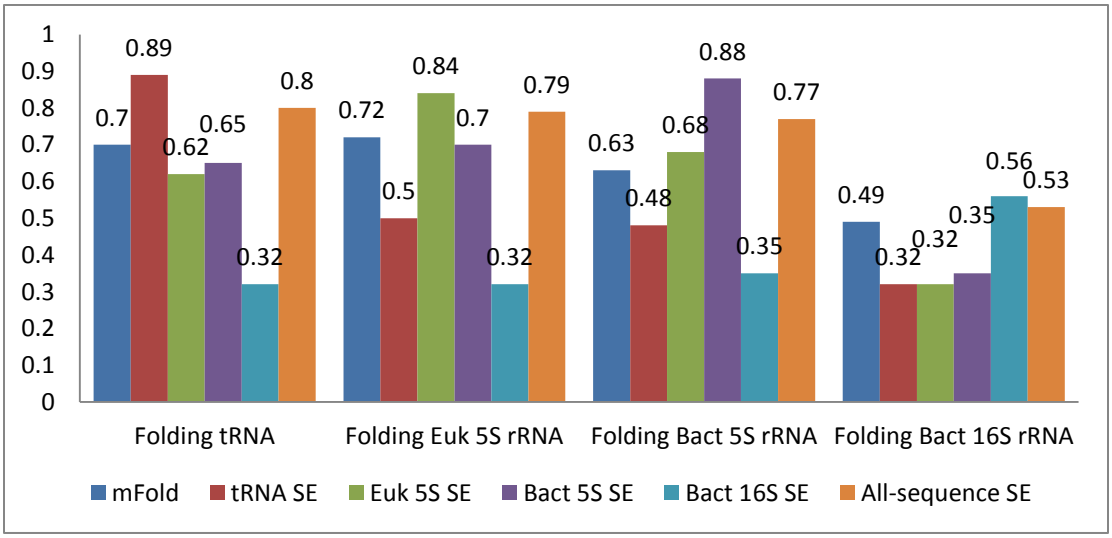


Figure 5.6: Each group of bars represents folding accuracy of (from left to right) tRNA, Eukaryotic 5S rRNA, Bacterial 5S rRNA, and Bacterial 16S rRNA. Within each group, each bar represents (from left to right) unmodified Mfold, base-pair stack, hairpin flank, and internal loop SEs derived using tRNA, Eukaryotic 5S rRNA, Bacterial 5S rRNA, Bacterial 16S rRNA, and all-sequence dataset.

5.5 REFERENCES

1. Sakakibara, Y.; Brown, M.; Hughey, R.; Mian, I.S.; Sjolander, K.; Underwood, R.C.; Haussler, D., *Stochastic Context-Free Grammars for Transfer-RNA Modeling*. Nucleic Acids Res., 1994. **22**(23): p. 5112-5120.
2. Durbin, R.; Eddy, S.R.; Krogh, A.; Mitchison, G., *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* 1998: Cambridge University Press.
3. Knudsen, B.; Hein, J., *RNA secondary structure prediction using stochastic context-free grammars and evolutionary history.* Bioinformatics, 1999. **15**(6): p. 446-454.
4. Knudsen, B.; Hein, J., *Pfold: RNA secondary structure prediction using stochastic context-free grammars.* Nucleic Acids Res., 2003. **31**(13): p. 3423-3428.
5. Do, C.B.; Woods, D.A.; Batzoglou, S., *CONTRAfold: RNA secondary structure prediction without physics-based models.* Bioinformatics, 2006. **22**(14): p. E90-E98.
6. Dima, R.I.; Hyeon, C.; Thirumalai, D., *Extracting stacking interaction parameters for RNA from the data set of native structures.* J. Mol. Biol., 2005. **347**(1): p. 53-69.
7. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E., *The Protein Data Bank.* Nucleic Acids Res., 2000. **28**(1): p. 235-242.
8. Das, R.; Baker, D., *Automated de novo prediction of native-like RNA tertiary structures.* Proc. Natl. Acad. Sci. U. S. A., 2007. **104**(37): p. 14664-14669.
9. Sykes, M.T.; Levitt, M., *Describing RNA structure by libraries of clustered nucleotide doublets.* J. Mol. Biol., 2005. **351**(1): p. 26-38.
10. Parisien, M.; Major, F., *The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data.* Nature, 2008. **452**(7183): p. 51-55.
11. Jonikas, M.A.; Radmer, R.J.; Laederach, A.; Das, R.; Pearlman, S.; Herschlag, D.; Altman, R.B., *Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters.* RNA-Publ. RNA Soc., 2009. **15**(2): p. 189-199.
12. Woese, C.R.; Gutell, R.; Gupta, R.; Noller, H.F., *Detailed Analysis of the Higher-Order Structure of 16s-Like Ribosomal Ribonucleic-Acids.* Microbiological Reviews, 1983. **47**(4): p. 621.
13. Gutell, R.R.; Weiser, B.; Woese, C.R.; Noller, H.F., *Comparative Anatomy of 16-S-Like Ribosomal-RNA.* Prog. Nucleic Acid Res. Mol. Biol., 1985. **32**: p. 155-216.
14. Holley, R.W.; Apgar, J.; Everett, G.A.; Madison, J.T.; Marquise.M; Merrill, S.H.; Penswick, J.R.; Zamir, A., *Structure Of A Ribonucleic Acid.* Science, 1965. **147**(3664): p. 1462.
15. Levitt, M., *Detailed Molecular Model For Transfer Ribonucleic Acid.* Nature, 1969. **224**(5221): p. 759.
16. Fox, G.E.; Woese, C.R., *5S-RNA Secondary Structure.* Nature, 1975. **256**(5517): p. 505-507.

17. Woese, C.R.; Magrum, L.J.; Gupta, R.; Siegel, R.B.; Stahl, D.A.; Kop, J.; Crawford, N.; Brosius, J.; Gutell, R.; Hogan, J.J.; Noller, H.F., *Secondary Structure Model For Bacterial 16S Ribosomal-RNA - Phylogenetic, Enzymatic And Chemical Evidence.* Nucleic Acids Res., 1980. **8**(10): p. 2275-2293.
18. Zwieb, C.; Glotz, C.; Brimacombe, R., *Secondary Structure Comparisons Between Small Subunit Ribosomal-RNA Molecules From 6 Different Species.* Nucleic Acids Res., 1981. **9**(15): p. 3621-3640.
19. Stiegler, P.; Carbon, P.; Zuker, M.; Ebel, J.P.; Ehresmann, C., *Secondary Structure And Topography Of 16S-Ribosomal RNA From Escherichia-Coli.* Comptes Rendus Hebdomadaires Des Seances De L Academie Des Sciences Serie D, 1980. **291**(12): p. 937-940.
20. Noller, H.F.; Kop, J.; Wheaton, V.; Brosius, J.; Gutell, R.R.; Kopylov, A.M.; Dohme, F.; Herr, W.; Stahl, D.A.; Gupta, R.; Woese, C.R., *Secondary Structure Model for 23s Ribosomal-Rna.* Nucleic Acids Res., 1981. **9**(22): p. 6167-6189.
21. Glotz, C.; Zwieb, C.; Brimacombe, R.; Edwards, K.; Kossel, H., *Secondary Structure of the Large Subunit Ribosomal-Rna from Escherichia-Coli, Zea-Mays Chloroplast, and Human and Mouse Mitochondrial Ribosomes.* Nucleic Acids Res., 1981. **9**(14): p. 3287-3306.
22. Branlant, C.; Krol, A.; Machatt, M.A.; Pouyet, J.; Ebel, J.P.; Edwards, K.; Kossel, H., *Primary and Secondary Structures of Escherichia-Coli Mre-600-23s Ribosomal-Rna - Comparison with Models of Secondary Structure for Maize Chloroplast 23s Ribosomal-Rna and for Large Portions of Mouse and Human 16s Mitochondrial Ribosomal-Rnas.* Nucleic Acids Res., 1981. **9**(17): p. 4303-4324.
23. James, B.D.; Olsen, G.J.; Liu, J.S.; Pace, N.R., *The Secondary Structure of Ribonuclease-P Rna, the Catalytic Element of a Ribonucleoprotein Enzyme.* Cell, 1988. **52**(1): p. 19-26.
24. Michel, F.; Jacquier, A.; Dujon, B., *Comparison of Fungal Mitochondrial Introns Reveals Extensive Homologies in Rna Secondary Structure.* Biochimie, 1982. **64**(10): p. 867-881.
25. Cech, T.R., *Conserved Sequences and Structures of Group-I Introns - Building an Active-Site for Rna Catalysis - a Review.* Gene, 1988. **73**(2): p. 259-271.
26. Michel, F.; Umesono, K.; Ozeki, H., *Comparative and Functional-Anatomy of Group-Ii Catalytic Introns - a Review.* Gene, 1989. **82**(1): p. 5-30.
27. Gutell, R.R.; Lee, J.C.; Cannone, J.J., *The accuracy of ribosomal RNA comparative structure models.* Curr. Opin. Struct. Biol., 2002. **12**(3): p. 301-310.
28. Bloch, F., *Fundamentals of statistical mechanics : manuscript and notes of Felix Bloch*. 2000, London : Singapore :: Imperial College Press ; World Scientific. xii, 302 p. :.
29. Gutell, R.R., *COLLECTION OF SMALL-SUBUNIT (16S- AND 16S-LIKE) RIBOSOMAL-RNA STRUCTURES - 1994.* Nucleic Acids Res., 1994. **22**(17): p. 3502-3507.

30. Wu, J.C.; Gardner, D.P.; Ozer, S.; Gutell, R.R.; Ren, P.Y., *Correlation of RNA Secondary Structure Statistics with Thermodynamic Stability and Applications to Folding.* J. Mol. Biol., 2009. **391**(4): p. 769-783.
31. Mathews, D.H.; Sabina, J.; Zuker, M.; Turner, D.H., *Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure.* J. Mol. Biol., 1999. **288**(5): p. 911-940.
32. Xia, T.B.; SantaLucia, J.; Burkard, M.E.; Kierzek, R.; Schroeder, S.J.; Jiao, X.Q.; Cox, C.; Turner, D.H., *Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs.* Biochemistry, 1998. **37**(42): p. 14719-14735.
33. Zuker, M.; Jaeger, J.A.; Turner, D.H., *A comparison of optimal and suboptimal RNA secondary structures predicted by free energy minimization with structures determined by phylogenetic comparison.* Nucleic Acids Res., 1991. **19**(10): p. 2707-14.
34. Vallurupalli, P.; Moore, P.B., *The solution structure of the loop E region of the 5 S rRNA from spinach chloroplasts.* J. Mol. Biol., 2003. **325**(5): p. 843-856.
35. Kiparisov, S.; Petrov, A.; Meskauskas, A.; Sergiev, P.V.; Dontsova, O.A.; Dinman, J.D., *Structural and functional analysis of 5S rRNA in Saccharomyces cerevisiae.* Mol. Genet. Genomics, 2005. **274**(3): p. 235-247.
36. Elgavish, T.; Cannone, J.J.; Lee, J.C.; Harvey, S.C.; Gutell, R.R., *AA.AG@Helix.Ends: A : A and A : G base-pairs at the ends of 16 S and 23 S rRNA helices.* J. Mol. Biol., 2001. **310**(4): p. 735-753.
37. Gutell, R.R.; Woese, C.R., *Higher-Order Structural Elements in Ribosomal-RNAs - Pseudo-Knots and the Use of Noncanonical Pairs.* Proc. Natl. Acad. Sci. U. S. A., 1990. **87**(2): p. 663-667.
38. Liang, X.G.; Kuhn, H.; Frank-Kamenetskii, M.D., *Monitoring single-stranded DNA secondary structure formation by determining the topological state of DNA catenanes.* Biophys. J., 2006. **90**(8): p. 2877-2889.
39. Lee, J.C.; Gutell, R.R., *Diversity of base-pair conformations and their occurrence in rRNA structure and RNA structural motifs.* J. Mol. Biol., 2004. **344**(5): p. 1225-1249.
40. Smith, M.W.; Meskauskas, A.; Wang, P.; Sergiev, P.V.; Dinman, J.D., *Saturation Mutagenesis of 5S rRNA in Saccharomyces cerevisiae.* Mol. Cell. Biol., 2001. **21**(24): p. 8264-8275.
41. Wimberly, B.; Varani, G.; Tinoco, I., *The Conformation Of Loop-E Of Eukaryotic 5S-Ribosomal RNA.* Biochemistry, 1993. **32**(4): p. 1078-1087.
42. Gutell, R.R.; Cannone, J.J.; Shang, Z.; Du, Y.; Serra, M.J., *A story: Unpaired adenosine bases in ribosomal RNAs.* J. Mol. Biol., 2000. **304**(3): p. 335-354.
43. Woese, C.R.; Winker, S.; Gutell, R.R., *Architecture of Ribosomal-Rna - Constraints on the Sequence of Tetra-Loops.* Proc. Natl. Acad. Sci. U. S. A., 1990. **87**(21): p. 8467-8471.
44. Tuerk, C.; Gauss, P.; Thermes, C.; Groebe, D.R.; Gayle, M.; Guild, N.; Stormo, G.; Daubentoncarafa, Y.; Uhlenbeck, O.C.; Tinoco, I.; Brody, E.N.; Gold, L.,

CUUCGG Hairpins - Extraordinarily Stable RNA Secondary Structures Associated With Various Biochemical Processes. Proc. Natl. Acad. Sci. U. S. A., 1988. **85**(5): p. 1364-1368.
45. Michel, F.; Westhof, E., *Modeling of the 3-Dimensional Architecture of Group-I Catalytic Introns Based on Comparative Sequence-Analysis.* J. Mol. Biol., 1990. **216**(3): p. 585-610.
46. Antao, V.P.; Tinoco, I., *Thermodynamic Parameters For Loop Formation In RNA And DNA Hairpin Tetraloops.* Nucleic Acids Res., 1992. **20**(4): p. 819-824.
47. Huang, S.G.; Wang, Y.X.; Draper, D.E., *Structure of a hexanucleotide RNA hairpin loop conserved in ribosomal RNAs.* J. Mol. Biol., 1996. **258**(2): p. 308-321.
48. Fountain, M.A.; Serra, M.J.; Krugh, T.R.; Turner, D.H., *Structural features of a six-nucleotide RNA hairpin loop found in ribosomal RNA.* Biochemistry, 1996. **35**(21): p. 6539-6548.
49. Serra, M.J.; Lyttle, M.H.; Axenson, T.J.; Schadt, C.A.; Turner, D.H., *RNA Hairpin Loop Stability Depends On Closing Base-Pair.* Nucleic Acids Res., 1993. **21**(16): p. 3845-3849.
50. Gautheret, D.; Konings, D.; Gutell, R.R., *A Major Family Of Motifs Involving G-Center-Dot-A Mismatches In Ribosomal-RNA.* J. Mol. Biol., 1994. **242**(1): p. 1-8.
51. Walter, A.E.; Wu, M.; Turner, D.H., *The Stability and Structure of Tandem Ga Mismatches in Rna Depend on Closing Base-Pairs.* Biochemistry, 1994. **33**(37): p. 11349-11354.
52. Santalucia, J.; Kierzek, R.; Turner, D.H., *Stabilities Of Consecutive A.C, C.C, G.G, U.C, And U.U Mismatches In RNA Internal Loops - Evidence For Stable Hydrogen-Bonded U.U And C.C+ Pairs.* Biochemistry, 1991. **30**(33): p. 8242-8251.
53. Gate, J.H.; Gooding, A.R.; Podell, E.; Zhou, K.H.; Golden, B.L.; Szewczak, A.A.; Kundrot, C.E.; Cech, T.R.; Doudna, J.A., *RNA tertiary structure mediation by adenosine platforms.* Science, 1996. **273**(5282): p. 1696-1699.
54. Costa, M.; Michel, F., *Frequent Use Of The Same Tertiary Motif By Self-Folding RNAs.* EMBO J., 1995. **14**(6): p. 1276-1285.
55. Jaeger, L.; Michel, F.; Westhof, E., *Involvement Of A GNRA Tetraloop In Long-Range Tertiary Interactions.* J. Mol. Biol., 1994. **236**(5): p. 1271-1276.
56. Lee, J.C.; Gutell, R.R.; Russell, R., *The UAA/GAN internal loop motif: A new RNA structural element that forms a cross-strand AAA stack and long-range tertiary interactions.* J. Mol. Biol., 2006. **360**(5): p. 978-988.

6 Conclusions and Future Work

At the scale of single atoms, the AMOEBA polarizable force field is capable of describing ions in gas-phase as well as liquid-phase without re-parameterization. Although a difference exists in the polarization energy of the monoligated Zn^{2+}-water complex obtained by the AMOEBA model and QM calculations, which exhibits non-classical covalent bonding as shown by ELF topological analysis, AMOEBA is able to compute the hydration free energy with less than a 2% error. Investigation of the interaction of a negatively charged species with Zn^{2+} would determine the requirement of an explicit charge transfer contribution. Nonetheless, calculations of hydration free energies, structure, and dynamic properties for a variety of single atom ions, such as Na^{+}, K^{+}, Cl^{-}, Ca^{2+}, and Mg^{2+} using AMOEBA agree exquisitely with experiment[1, 2]. The zinc model developed in this work enables the studies of zinc-containing metalloproteins.

For small molecules, we have described a standard protocol to generate the AMOEBA force field parameters and have implemented the POLTYPE utility to automate this process. A clear and straightforward procedure is described for the parameterization process for the AMOEBA polarizable force field, although more sophistication is required than the parameterization of fixed-charge force fields. Properties calculated with the AMOEBA force field such as dipole moments, conformational energies, and interactions energies and show good agreement with quantum mechanics calculations for neutral as well as charged molecules. Parameters obtained in gas-phase are transferred directly to liquid-phase systems without modification. Hydration free energies (HFE) of neutral calculated with AMOEBA yield an RMS error of less than 1 kcal/mol for neutral molecules when compared to experimental values. Although the absolute error of HFE for salts containing charged

molecules are larger, the unsigned mean relative error is less than 3% the correlation coefficient is 0.95. Although the HFE RMS error of the neutral molecules in this study demonstrate improved solvation energies over fixed charge force fields, confirmation with a more extensive dataset continues to be necessary. Moreover, combined with enhanced sampling methods[3] using AMOEBA, POLTYPE enables a host of molecular mechanics applications for small molecules such as polymorph prediction, solubility calculations, crystal structure refinement, and protein-ligand binding energy calculations.

The consistency of parameterizing small molecules for AMOEBA has been improved and the amount of time spent by a researcher has been drastically reduced as a result of POLTYPE. Yet it should be noted that the entirety of chemical is expansive and that the database definitions for parameters such as vdW, out-of-plane bend, and polarization will need to be added as fundamentally new functional groups are encountered. Localized bonds that are not in rings currently require specific input by the operator and implementation of an automatic identification would be convenient. Additionally, the requirement to perform highly accurate quantum mechanical calculations may be too computationally. Future work of parameterization for the AMOEBA force field would be to develop a "semi-empirical" method that would provide a more efficient method to assign multipoles. Precedence for such work has been in the application of bond-charge increment method[4] and the development of the AM1-BCC semi-empirical method[5, 6]. The AM1-BCC method has the advantage that it is able to maintain the molecular charge for ionic molecules. Essentially, AM1-BCC performs a least squares fit to a bond connectivity matrix that "corrects" AM1 charges by matching to electrostatic potentials around a training set of molecules. Although this method was used for assignment of partial charges, they can be generalized to assign dipoles and higher order moments. However, a method that corrects multipoles based on bonds

between two atoms may not yield appropriate multipole assignments. The bond order of all the bonds of an atom and perhaps the hybridization of all its neighboring atoms may be needed as input to the optimization. Various strategies will need to be explored, but fitting to electrostatic potential is a promising approach. As with the partial charge methods, care must be taken to maintain the correct charge of the molecule throughout the fitting process. Optimization methods with constraints such as Langrange multipliers may be applied to maintain the molecular charge.

For the efficient modeling of macromolecules, a coarse-grained model based on fundamental intermolecular forces represented with Gay-Berne potential with electrostatic multipole (GBEMP) potentials has been developed. The GBEMP model has been substantiated with alanine peptides of various lengths and can be transferred from gas phase to an implicit solvent without the need for re-parameterization. Furthermore, predictions of alpha-helix and beta-sheet distributions via replica-exchange molecular dynamics (REMD) and room-temperature MD simulations of 5-residue and 12-residue polyalanines are compared with all-atom simulations and experiments. The computational efficiency of the GBEMP model is up to 1,000 times faster than all-atom molecular mechanics models as a result of the reduction of particle numbers and larger time-steps. Further speedup is possible if the bonds and angles are restrained [7] to eliminate the higher frequency motions. Removing high frequency motions allows for larger time-steps, which increases simulation speed. Additionally, the parameterization of the rest of the 19 amino acid side-chains should subsequently be developed in order to represent whole proteins. Assignment of rigid bodies by partitioning each side chain requires consideration and should match conformational energy as well as interaction energy between side-chains. The backbone of proline requires parameterization as well. Although much more development is required, frameworks such as MSCALE[8] would

be able to combine the GBEMP model with all-atom molecular mechanics models to achieve multi-scale simulations.

Lastly, comparative sequence analysis was used to generate statistical energies (SE) that agree with experimental measurements. For base-pair stack, a correlation coefficient of ~0.9 has been achieved between the energies derived structural statistics and the free energy values extracted from experiments. We substantiated the base-pair stack, internal loop, and hairpin-flank SE by applying them in Mfold to predict secondary structures of RNA molecules of various lengths. We observe the accuracy of folding to be comparable to that of free energy obtained experimentally when using only base-pair stack SE. However, the application of hairpin flank (HF) and internal loop (IL) SEs to the Mfold algorithm yields dramatic improvements in the folding accuracy. Much of the improvement in accuracy is due to the hairpin flank statistical energies in particular. However, the accuracy of predicting one molecule (i.e. Bacterial 16S rRNA) using SE derived from sequences of another molecule yields accuracies worse than those of experimentally determined energy values. This suggests that the statistics obtained from individual RNA molecules are insufficient in sampling Boltzmann-like distribution of HF and IL and care needs to be taken when combining the statistics from different molecules into general statistical energies. More importantly, we have identified that a comprehensive understanding of RNA folding requires the converged characterization of motifs beyond BP-ST. Consequently, the development and validation of statistical potentials based on other motifs such as multi-stem loops and coaxial stacks will further substantiate the method. Additionally, statistical potentials can be generalized to incorporate many-body interactions and should be straight-forward to implement since these potentials have already been applied to 3-D protein structure prediction studies [9, 10]. Although the Mfold dynamic programming algorithm is efficient, it is not able to

fold sequences in to structures containing pseudo-knots. A more biologically relevant folding algorithm currently is being developed by Gardner and Gutell based on nucleation sites to address issues that challenge the Mfold algorithm. Moreover, the algorithm will employ statistical potentials from comparative sequence alignments.

6.1 REFERENCES

1. Grossfield, A.; Ren, P.Y.; Ponder, J.W., *Ion solvation thermodynamics from simulation with a polarizable force field.* J. Am. Chem. Soc., 2003. **125**(50): p. 15671-15682.
2. Jiao, D.; King, C.; Grossfield, A.; Darden, T.A.; Ren, P.Y., *Simulation of Ca2+ and Mg2+ solvation using polarizable atomic multipole potential.* J. Phys. Chem. B, 2006. **110**(37): p. 18553-18559.
3. Zheng, L.; Chen, M.; Yang, W., *Random walk in orthogonal space to achieve efficient free-energy simulation of complex systems.* Proc. Natl. Acad. Sci. U. S. A., 2008. **105**(51): p. 20227-32.
4. Bush, B.L.; Bayly, C.I.; Halgren, T.A., *Consensus bond-charge increments fitted to electrostatic potential or field of many compounds: Application to MMFF94 training set.* J. Comput. Chem., 1999. **20**(14): p. 1495-1516.
5. Jakalian, A.; Bush, B.L.; Jack, D.B.; Bayly, C.I., *Fast, efficient generation of high-quality atomic Charges. AM1-BCC model: I. Method.* J. Comput. Chem., 2000. **21**(2): p. 132-146.
6. Jakalian, A.; Jack, D.B.; Bayly, C.I., *Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation.* J. Comput. Chem., 2002. **23**(16): p. 1623-1641.
7. Miller, T.F.; Eleftheriou, M.; Pattnaik, P.; Ndirango, A.; Newns, D.; Martyna, G.J., *Symplectic quaternion scheme for biophysical molecular dynamics.* J. Chem. Phys., 2002. **116**(20): p. 8649-8659.
8. Woodcock, H.L.; Miller, B.T.; Hodoscek, M.; Okur, A.; Larkin, J.D.; Ponder, J.W.; Brooks, B.R., *MSCALE: A General Utility for Multiscale Modeling.* J. Chem. Theory Comput., 2011. **7**(4): p. 1208-1219.
9. Singh, R.K.; Tropsha, A.; Vaisman, II, *Delaunay tessellation of proteins: four body nearest-neighbor propensities of amino acid residues.* J. Comput. Biol., 1996. **3**(2): p. 213-21.
10. Krishnamoorthy, B.; Tropsha, A., *Development of a four-body statistical pseudo-potential to discriminate native from non-native protein conformations.* Bioinformatics, 2003. **19**(12): p. 1540-8.

Appendices

1 Supplemental data for Gay-Berne and Electrostatic Multipole-based Coarse-grain Potential in Implicit Solvent

Table S 1.1: Gay-Berne parameters of benzene, methanol, and water GBEMP models.

	Benzene	Methanol	Water
d_w	0.74	1.0	1.0
l (Å)	2.01	3.20	2.27
d (Å)	4.53	2.52	2.27
ε_0 (kcal/mol)	0.56	0.43	0.14
ε_E	4.08	0.43	1.0
ε_S	0.58	0.58	1.0

Table S 1.2: MD simulation results for benzene.

	GBEMP Model[a]		All-atom	Experiment
	Old	New	AMOEBA	
Potential energy (kcal/mol)	-7.48	-7.42	-7.38[b]	-7.50[c]
Density (NPT) (g cm^{-3})	0.874	0.884	0.868	0.870

[a] Using 20 fs time step
[b] The potential energy of all-atom model is calculated from the difference between the potential energies in gas and liquid phases.
[c] From the enthalpy of vaporization in reference [1]

Table S 1.3: MD simulation results for methanol.

	GBEMP Model[a]		All-atom[b]	Experiment
	Old	New	OPLS	
Potential energy (kcal/mol)	-8.29	-8.31	-7.933	-8.36[c]
Density (NPT) (g cm^{-3})	0.794	0.782	0.773	0.787[d]

[a] Using 20 fs time step
[b] From reference [2]
[c] From the enthalpy of vaporization in reference [3]
[d] From reference [4]

Table S 1.4: Bond, bond-angle, torsional, and multiple parameters for GBEMP model.

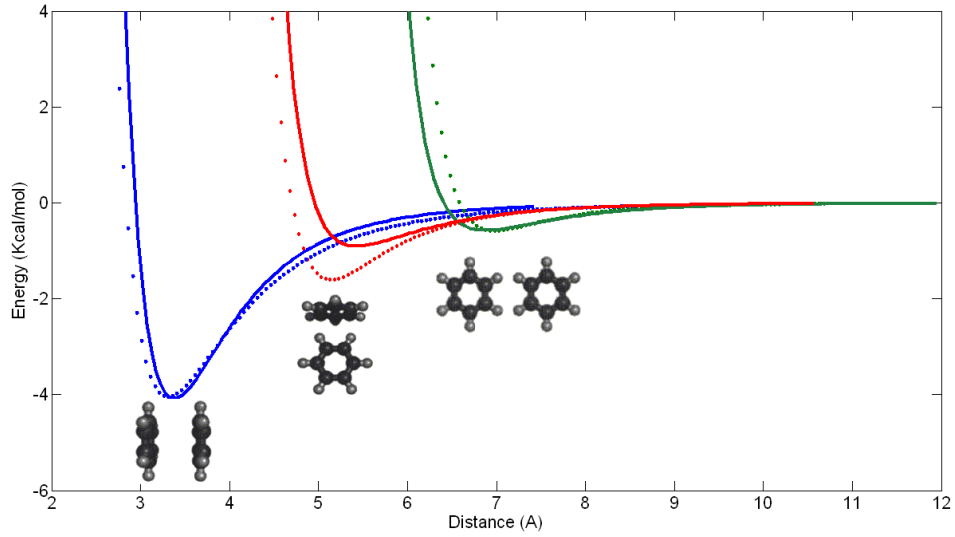
bond	b0	Kb
1 - 3	2.0358	235.6993
4 - 6	1.4678	245.377
6 - 8	1.9831	200.6274
9 -10	1.5294	99.7949

angle	θ0	Ka
1-3-2	98.6732	212.2176
2-4-6	99.9832	72.8651
4-6-5	112.0933	81.3831
4-6-8	112.1952	76.6092
5-6-8	113.3026	64.9691
6-8-7	98.7199	234.9351
8-9-10	108.0353	82.3129

torsion	V1	δ1	V2	δ2	V3	δ3
1-3-4-6	1.008	0	1.698	180	0.002895	0
3-4-6-5	1.182	0	-0.689	180	0.793	0
5-6-8-7	1.049	0	-0.3331	180	-0.01305	0
3-4-6-8	-2.088	0	0.2062	180	0.4669	0
6-8-9-10	1.008	0	1.698	180	0.002895	0

EMP site	Oxygen			Nitrogen		
Charge (e)	0.0			0.0		
Dipole (D)	-2.763	1.381	-0.003	-1.279	-0.637	-0.024
Quadrupole (D e)	3.001	-2.148	-0.014	1.067	-1.148	0.138
	-2.148	-1.427	0.009	-1.148	1.258	-0.128
	-0.014	0.009	-1.574	0.138	-0.128	-2.325

A



B

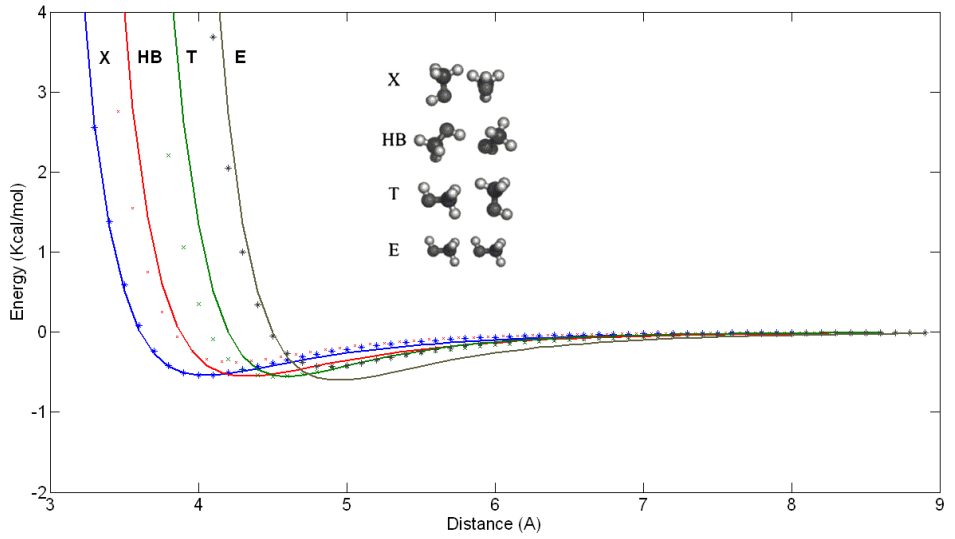


Figure S 1.1: Comparison of homodimer interaction energy given by the Gay-Berne model and all-atom model. All atom value are shown as data point, and GB as line in different colors. A. The interaction energy of benzene, the conformations that shown from left to right are: face to face, T shape, side by side. B. The interaction energy of methanol, the conformations that shown from left to right are: cross, hydrogen bonding, T shape, and end to end.

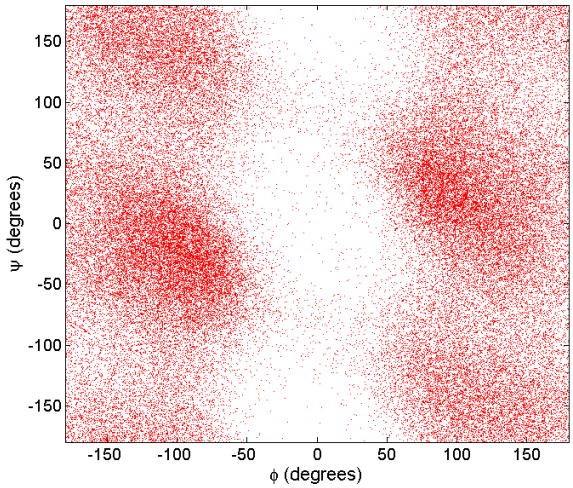


Figure S1.2: Phi and Psi torsion angle distribution of 12-mer polyalanine at temperature of 800 K to 900 K in the simulated annealing simulation. All possible conformations of polyalanine were sampled at the high temperature.

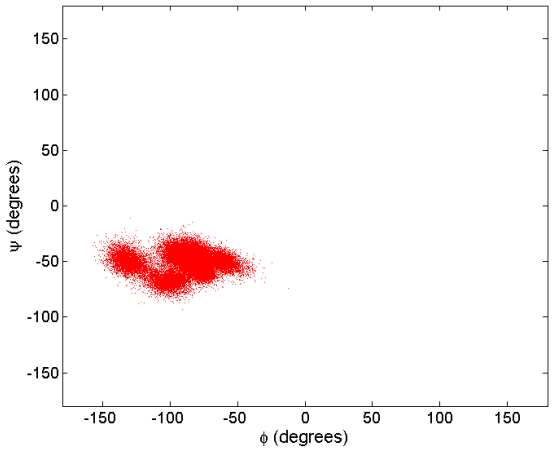


Figure S1.3: Phi and Psi torsion angle distribution of 12-mer polyalanine at temperature of 1 K to 100 K in the simulated annealing simulation. Alpha-helix become the only structure at low temperature for polyalanine.

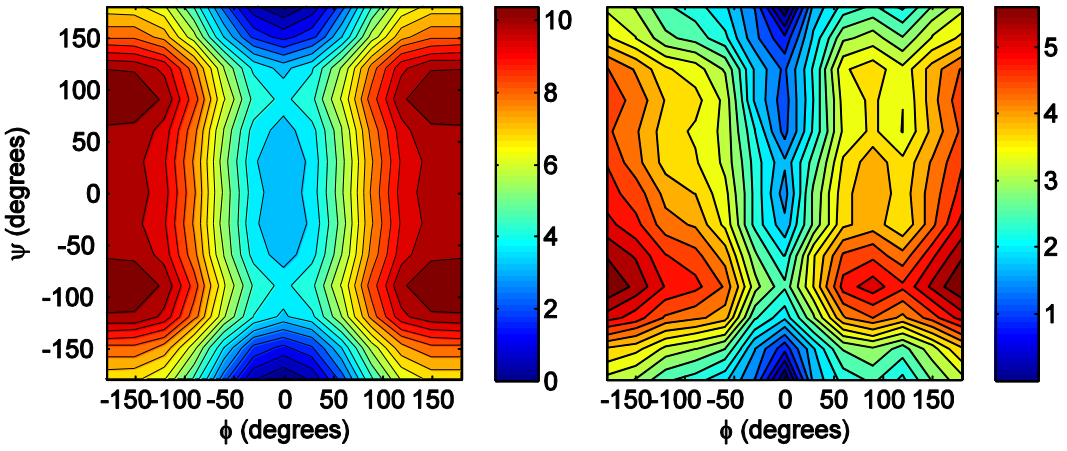


Figure S1.4: Contribution of torsional energy to total conformational energy of dialanine (a) GBEMP model (b) All-atom OPLSAA.

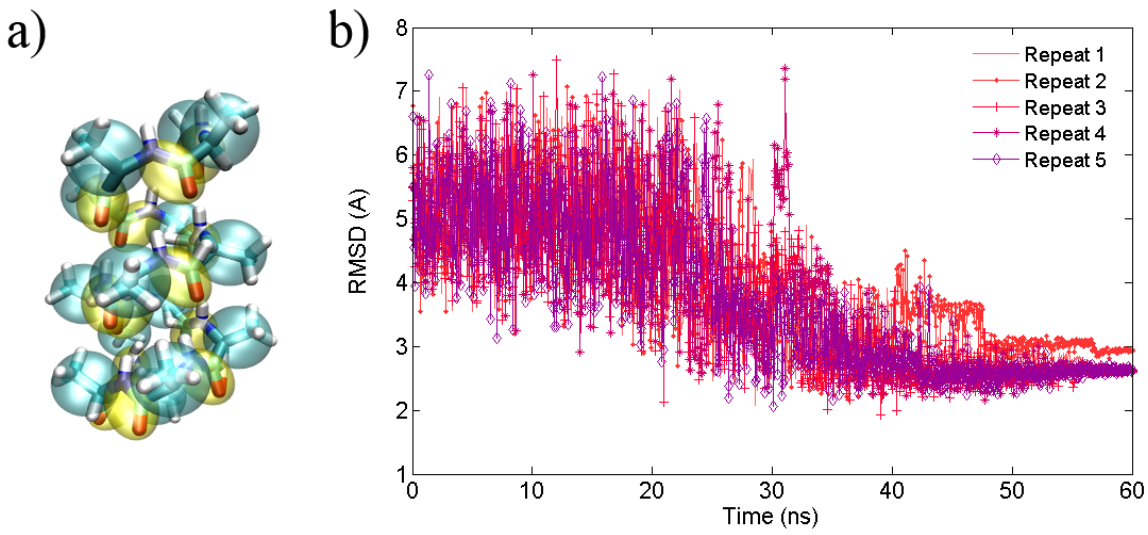


Figure S 1.5: (a) A final snapshot of polyalanine from a 60-ns simulated annealing simulation using GBEMP potential. (b) Heavy-atom RMSD of the 12-residue polyalanine from 5 simulated annealing simulations to inspect the minimum-energy structure. The systems were heat to 1,000 K and cooled linearly to less than 1 K over 60 ns. The final polyalanine structures after simulated annealing all adopt the alpha-helical conformation at low temperatures (100 K). The RMSD was calculated by mapping the CG trajectories to all-atom structures and then compared to atomic canonical alpha-helical structure.

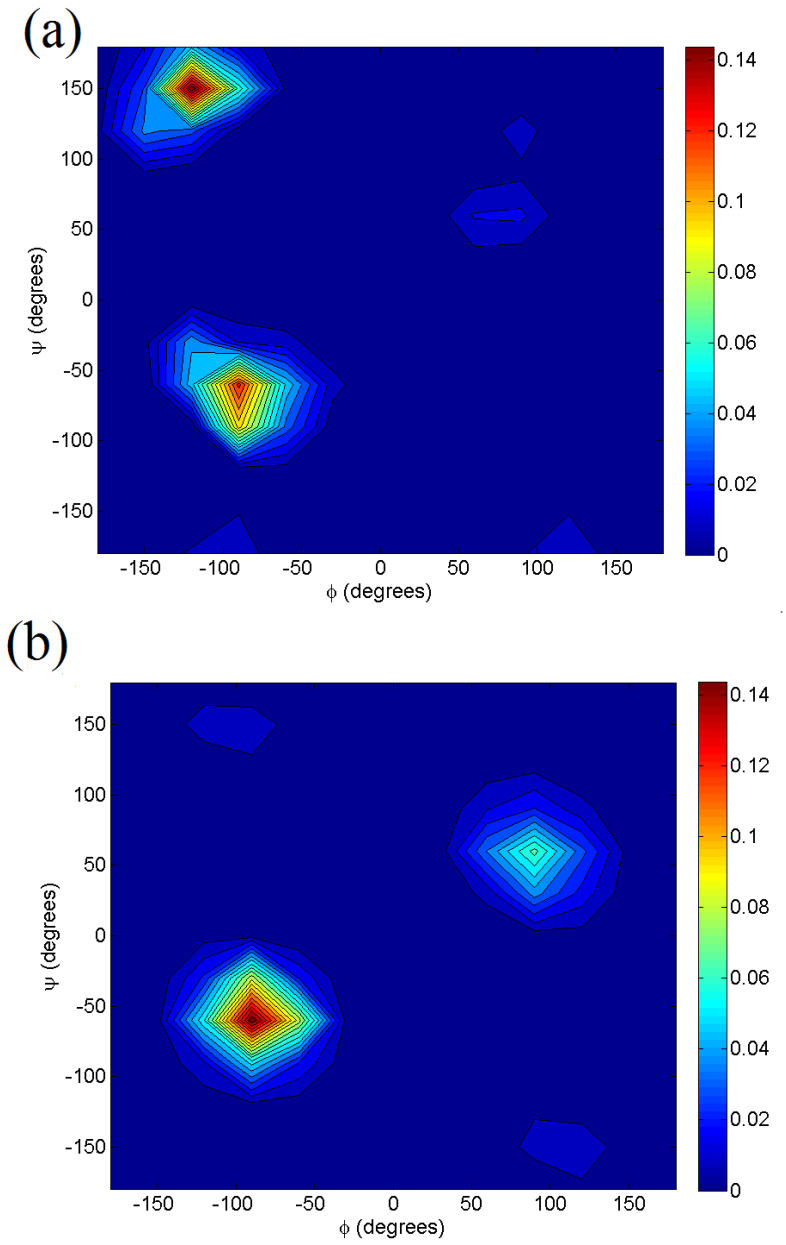


Figure S 1.6: Conformational distributions of 5-mer (a) and 12-mer (b) polyalanine from CG simulations at 298 K. Simulations are started with different initial structures, including the extended conformation, alpha helix, and partial alpha-helical and beta-strand conformations. The total simulation time for 12-mer and 5-mer is 6 μs and 2 μs, respectively. The beginning 1.5 μs of 12-mer trajectory and 0.5 μs of 5-mer trajectory are not included in the calculation. The color bar at right side represents the probability density of ϕ and ψ torsion angles.

2 SUPPLEMENTAL DATA FOR CORRELATION OF RNA SECONDARY STRUCTURE STATISTICS WITH THERMODYNAMIC STABILITY AND APPLICATIONS TO FOLDING

Table S 2.1: Symmetric statistical energy (kcal/mol) derived from tRNA and base-pair stack statistics, in parentheses, taken from comparative sequence analysis. The rows of the tables indicate the base-pair on the 5' end with the first nucleotide on the 5' end and second nucleotide on the 3' end. However, the columns are the base-pair on the 3' end with the first nucleotide on the 3' end and the second nucleotide on the 5' end. This notation allows the energy matrix to be symmetric.

	CG	GC	GU	UG	AU	UA
CG	-2.07	-3.19	-1.52	-0.53	-3.40	-2.86
	(771)	(835)	(106)	(46)	(1141)	(384)
GC	-3.19	-2.69	-1.07	-1.52	-2.37	-3.16
	(1757)	(1426)	(75)	(259)	(286)	(409)
GU	-1.52	-1.07	2.55	1.65	-1.02	-0.77
	(237)	(125)	(5)	(11)	(86)	(44)
UG	-0.53	-1.52	1.65	2.28	-0.50	-0.03
	(67)	(67)	(1)	(4)	(38)	(23)
AU	-3.40	-2.37	-1.02	-0.50	-1.98	-2.41
	(434)	(256)	(17)	(16)	(204)	(111)
UA	-2.86	-3.16	-0.77	-0.03	-2.41	-1.95
	(505)	(848)	(34)	(12)	(306)	(166)

Table S 2.2: Symmetric statistical energy (kcal/mol) derived from Eukaryotic 5S rRNA and base-pair stack statistics, in parentheses, taken from comparative sequence analysis.

	CG	GC	GU	UG	AU	UA
CG	-2.43	-3.34	-2.44	-2.27	-3.09	-3.25
	(459)	(318)	(147)	(186)	(313)	(275)
GC	-3.34	-2.40	-2.37	-1.29	-3.40	-2.73
	(613)	(412)	(92)	(70)	(295)	(269)
GU	-2.44	-2.37	1.91	0.69	-1.45	-1.46
	(165)	(278)	(4)	(10)	(47)	(73)
UG	-2.27	-1.29	0.69	-0.08	-0.93	-1.42
	(158)	(52)	(4)	(31)	(44)	(82)
AU	-3.09	-3.40	-1.45	-0.93	-1.72	-2.36
	(271)	(408)	(56)	(17)	(111)	(75)
UA	-3.25	-2.73	-1.46	-1.42	-2.36	-2.23
	(485)	(114)	(31)	(18)	(141)	(190)

Table S 2.3: Symmetric statistical energy (kcal/mol) derived from Bacterial 5S rRNA and base-pair stack statistics, in parentheses, taken from comparative sequence analysis.

	CG	GC	GU	UG	AU	UA
CG	-1.90	-3.40	-2.14	-1.98	-3.05	-3.38
	(132)	(204)	(67)	(12)	(132)	(94)
GC	-3.40	-2.28	-2.84	-0.71	-3.09	-3.07
	(342)	(107)	(121)	(4)	(78)	(155)
GU	-2.14	-2.84	-1.40	2.27	-3.14	-1.27
	(36)	(102)	(37)	(0)	(43)	(11)
UG	-1.98	-0.71	2.27	-2.22	-1.08	-0.98
	(86)	(23)	(1)	(86)	(18)	(14)
AU	-3.05	-3.09	-3.14	-1.08	-0.98	-2.78
	(68)	(116)	(14)	(2)	(12)	(42)
UA	-3.38	-3.07	-1.27	-0.98	-2.78	-0.36
	(187)	(38)	(13)	(4)	(35)	(7)

Table S 2.4: Symmetric statistical energy (kcal/mol) derived from Bacterial 16S rRNA and base-pair stack statistics, in parentheses, taken from comparative sequence analysis.

	CG	GC	GU	UG	AU	UA
CG	-2.47	-3.40	-2.38	-1.79	-3.15	-3.20
	(6268)	(8038)	(2128)	(1438)	(3488)	(3871)
GC	-3.40	-2.70	-2.42	-2.01	-3.11	-2.85
	(9285)	(7688)	(3323)	(2104)	(3943)	(2756)
GU	-2.38	-2.42	0.07	-1.00	-2.09	-1.68
	(3126)	(2070)	(401)	(476)	(1732)	(1014)
UG	-1.79	-2.01	-1.00	-1.33	-1.41	-1.10
	(1509)	(1677)	(693)	(1664)	(838)	(645)
AU	-3.15	-3.11	-2.09	-1.41	-1.58	-2.49
	(5373)	(4235)	(981)	(525)	(1239)	(1258)
UA	-3.20	-2.85	-1.68	-1.10	-2.49	-2.04
	(4885)	(3702)	(774)	(354)	(1816)	(1955)

Table S 2.5: Symmetric statistical energy (kcal/mol) derived from all-sequences and base-pair stack statistics, in parentheses, taken from comparative sequence analysis.

	CG	GC	GU	UG	AU	UA
CG	-2.42	-3.40	-2.28	-1.66	-3.12	-3.15
	(7630)	(9395)	(2448)	(1682)	(5074)	(4624)
GC	-3.40	-2.69	-2.31	-1.92	-3.08	-2.90
	(11997)	(9633)	(3611)	(2437)	(4602)	(3589)
GU	-2.28	-2.31	0.12	-0.82	-2.03	-1.54
	(3564)	(2575)	(447)	(497)	(1908)	(1142)
UG	-1.66	-1.92	-0.82	-1.23	-1.30	-0.99
	(1820)	(1819)	(699)	(1785)	(938)	(764)
AU	-3.12	-3.08	-2.03	-1.30	-1.65	-2.52
	(6146)	(5015)	(1068)	(560)	(1566)	(1486)
UA	-3.15	-2.90	-1.54	-0.99	-2.52	-2.03
	(6062)	(4702)	(852)	(388)	(2298)	(2318)

Table S 2.6: Asymmetric statistical energy (kcal/mol) derived from tRNA and base-pair stack statistics, in parentheses, taken from comparative sequence analysis. The rows of the tables indicate the base-pair on the 5’ end with the first nucleotide on the 5’ end and second nucleotide on the 3’ end. However, the columns are the base-pair on the 3’ end with the first nucleotide on the 3’ end and the second nucleotide on the 5’ end. This notation allows the energy matrix to be symmetric.

	CG	GC	GU	UG	AU	UA
CG	-2.27	-2.31	-0.25	0.29	-3.4	-2.28
	(771)	(835)	(106)	(46)	(1141)	(384)
GC	-3.13	-2.94	-0.2	-1.44	-1.91	-2.36
	(1757)	(1426)	(75)	(259)	(286)	(409)
GU	-1.35	-0.65	2.3	1.96	-0.9	-0.35
	(237)	(125)	(5)	(11)	(86)	(44)
UG	-0.02	0.03	3.93	2.53	-0.2	0.44
	(67)	(67)	(1)	(4)	(38)	(23)
AU	-2.42	-1.87	0.73	0.67	-2.34	-1.74
	(434)	(256)	(17)	(16)	(204)	(111)
UA	-2.56	-3.05	-0.09	1.05	-2.77	-2.12
	(505)	(848)	(34)	(12)	(306)	(166)

Table S 2.7: Asymmetric statistical energy (kcal/mol) derived from Eukaryotic 5S rRNA and base-pair stack statistics, in parentheses, taken from comparative sequence analysis.

	CG	GC	GU	UG	AU	UA
CG	-2.75	-2.72	-2.14	-1.9	-2.84	-2.58
	(459)	(318)	(147)	(186)	(313)	(275)
GC	-3.26	-2.72	-1.15	-0.84	-2.63	-2.6
	(613)	(412)	(92)	(70)	(295)	(269)
GU	-1.84	-2.38	2.16	1.15	-0.76	-1.25
	(165)	(278)	(4)	(10)	(47)	(73)
UG	-1.72	-0.53	2.16	-0.09	-0.69	-1.39
	(158)	(52)	(4)	(31)	(44)	(82)
AU	-2.61	-3.4	-0.97	0.35	-1.95	-1.51
	(271)	(408)	(56)	(17)	(111)	(75)
UA	-3.18	-1.95	-0.34	0.29	-2.2	-2.52
	(485)	(114)	(31)	(18)	(141)	(190)

Table S 2.8: Asymmetric statistical energy (kcal/mol) derived from Bacterial 5S rRNA and base-pair stack statistics, in parentheses, taken from comparative sequence analysis.

	CG	GC	GU	UG	AU	UA
CG	-2.18	-2.84	-2.01	0.16	-3.01	-2.6
	(132)	(204)	(67)	(12)	(132)	(94)
GC	-3.28	-2.61	-2.55	1.43	-2.46	-3.23
	(342)	(107)	(121)	(4)	(78)	(155)
GU	-1.11	-2.32	-1.6	Inf	-2.18	-0.55
	(36)	(102)	(37)	(0)	(43)	(11)
UG	-2.11	-0.64	2.6	-2.54	-1.12	-0.83
	(86)	(23)	(1)	(86)	(18)	(14)
AU	-2.22	-2.95	-3.18	1.41	-1.12	-2.48
	(68)	(116)	(14)	(2)	(12)	(42)
UA	-3.39	-1.75	-0.74	0.61	-2.27	-0.41
	(187)	(38)	(13)	(4)	(35)	(7)

Table S 2.9: Asymmetric statistical energy (kcal/mol) derived from Bacterial 16S rRNA and base-pair stack statistics, in parentheses, taken from comparative sequence analysis.

	CG	GC	GU	UG	AU	UA
CG	-3.06	-3.3	-1.93	-1.29	-2.75	-2.88
	(6268)	(8038)	(2128)	(1438)	(3488)	(3871)
GC	-3.39	-3.34	-2.32	-1.75	-2.93	-2.48
	(9285)	(7688)	(3323)	(2104)	(3943)	(2756)
GU	-2.24	-1.94	0.08	-0.13	-2.03	-1.39
	(3126)	(2070)	(401)	(476)	(1732)	(1014)
UG	-1.44	-1.51	-0.59	-1.65	-1.15	-0.83
	(1509)	(1677)	(693)	(1664)	(838)	(645)
AU	-3.29	-3.07	-1.33	-0.58	-1.95	-1.98
	(5373)	(4235)	(981)	(525)	(1239)	(1258)
UA	-3.31	-2.86	-1.05	-0.09	-2.44	-2.52
	(4885)	(3702)	(774)	(354)	(1816)	(1955)

Table S 2.10: Asymmetric statistical energy (kcal/mol) derived from all-sequences and base-pair stack statistics, in parentheses, taken from comparative sequence analysis.

	CG	GC	GU	UG	AU	UA
CG	-2.96	-3.18	-1.83	-1.22	-2.92	-2.81
	(7630)	(9395)	(2448)	(1682)	(5074)	(4624)
GC	-3.39	-3.27	-2.14	-1.66	-2.82	-2.52
	(11997)	(9633)	(3611)	(2437)	(4602)	(3589)
GU	-2.13	-1.89	0.17	0.04	-1.9	-1.3
	(3564)	(2575)	(447)	(497)	(1908)	(1142)
UG	-1.39	-1.34	-0.37	-1.48	-1.05	-0.81
	(1820)	(1819)	(699)	(1785)	(938)	(764)
AU	-3.17	-3.02	-1.3	-0.44	-2	-1.95
	(6146)	(5015)	(1068)	(560)	(1566)	(1486)
UA	-3.25	-2.85	-0.94	0.01	-2.47	-2.48
	(6062)	(4702)	(852)	(388)	(2298)	(2318)

Table S 2.11: Potentials (kcal/mol) of consecutive canonical base-pairs obtained from statistics, in parentheses, of RNA crystal structures.[a]

	CG	GC	GU	UG	AU	UA
CG	-2.97	-3.40	-1.79	-1.11	-3.08	-3.04
	(619)	(850)	(149)	(111)	(273)	(296)
GC	-3.40	-3.20	-2.08	-1.31	-3.23	-3.07

	(880)	(740)	(170)	(99)	(313)	(322)
GU	-1.78	-2.08	0.82	0.82	-1.62	-0.27
	(164)	(226)	(13)	(6)	(66)	(10)
UG	-1.11	-1.31	0.82	-0.88	-0.86	-1.19
	(73)	(116)	(20)	(50)	(40)	(60)
AU	-3.08	-3.23	-1.62	-0.86	-2.80	-2.56
	(322)	(357)	(56)	(27)	(106)	(95)
UA	-3.04	-3.07	-0.27	-1.19	-2.56	-2.80
	(282)	(262)	(32)	(27)	(80)	(106)

[a]Mathews, 1999

Table S 2.12: Free energies (kcal/mol) of consecutive canonical base-pairs obtained from experiment.[a]

	CG	GC	GU	UG	AU	UA
CG	-2.40	-3.30	-2.10	-1.40	-2.10	-2.10
GC	-3.30	-3.40	-2.50	-1.50	-2.20	-2.40
GU	-2.10	-2.50	1.30	-0.50	-1.40	-1.30
UG	-1.40	-1.50	-0.50	0.30	-0.60	-1.00
AU	-2.10	-2.20	-1.40	-0.60	-1.10	-0.90
UA	-2.10	-2.40	-1.30	-1.00	-0.90	-1.30

[a]Dima, 2005

Table S 2.13: Statistics and Statistical Energies (SE) derived from Watson-Crick, GU base-pairs

Base-pairs	Number of Base-pairs	BP SE (kcal/mol)
CG	46698	-1.5907
GC	50798	-1.6912
GU	12032	-0.1696
UG	11672	-0.1333
AU	19295	-1.0671
UA	21812	-1.2135

Table S 2.14: Statistics and Statistical Energies (SE) derived from stacks (adjacent nucleotides) in base-pairs

Stacks	Number of Adjacent	ST SE (kcal/mol)

	Nucleotides	
AA	3853	0.4277
AC	12202	0.0677
AG	11621	0.1724
AU	6083	0.2315
CA	10898	0.1027
CC	24129	-0.1466
CG	29656	-0.1208
CU	13691	-0.0228
GA	14036	0.114
GC	22037	-0.0289
GG	36786	-0.0978
GU	22585	-0.088
UA	6602	0.2062
UC	18889	-0.1224
UG	22011	-0.0801
UU	9109	0.0517

Table S 2.15: Minimum and maximum energy values found in experiment**[5]**.

Secondary Structure	Minimum (kcal/mol)	Maximum (kcal/mol)
1x1 Internal Loop	-2.10	1.70
1x2 Internal Loop	0.40	5.50
2x2 Internal Loop	-4.90	3.40
Internal Loop Flank	-1.10	0.70
Hairpin Flank	-2.90	.20

Table S 2.16: Hairpin Flank Statistical Energies (SE) (kcal/mol) derived from tRNA, Eukaryotic 5S rRNA, Bacterial 5S rRNA, Bacterial 16S rRNA, and all-sequences.

Base-pair	HP Flank	tRNA SE	Euk 5S SE	Bact 5S SE	Bact 16S SE	All-sequence SE
AU	AC	0.07	-1.06	-1.30		-1.07
AU	AG				-0.78	-0.41
AU	CA	-2.10	-0.37			-1.92
AU	CC			-1.71		-1.49
AU	CU		-0.72			-0.35
AU	GA		-1.73	-0.38	-1.91	-1.73
AU	GU			-0.89		-0.53

AU	UA				-0.56	-0.21
AU	UC	-0.13	-2.90	-2.90		-2.87
AU	UU	-1.10	-1.46	-1.17	-1.70	-1.45
CG	AA	-2.90		-0.58		-2.81
CG	AC			0.01	-1.95	-1.77
CG	AU				-0.03	0.12
CG	CA	-1.83		0.01		-1.62
CG	CC	-0.86		0.17		-0.50
CG	CU				-0.98	-0.63
CG	GA		0.15	-0.81	-2.90	-2.80
CG	GG			-0.74	-1.42	-1.12
CG	GU				-1.08	-0.74
CG	UC	-0.36		-0.06	-2.44	-2.30
CG	UG				-2.59	-2.51
CG	UU	-0.39			-1.40	-1.09
GC	AA			-1.02	-1.05	-0.89
GC	AC			-0.42	-1.79	-1.54
GC	AG				-2.32	-2.21
GC	CA	-1.07		0.01	-1.77	-1.49
GC	CC	-0.04		0.17		0.11
GC	CG				-0.89	-0.53
GC	CU		0.07	-1.07		-0.72
GC	GA		-2.83	-1.16	-2.03	-2.57
GC	GG				-1.29	-0.99
GC	GU				0.03	0.14
GC	UA			-0.22		0.03
GC	UC	-1.85	-0.80	-1.84	0.02	-1.72
GC	UU	-2.70		-2.34	-2.19	-2.47
GU	AU		-0.08			0.10
GU	GA		0.05		-1.62	-1.37
GU	UC				-0.20	0.04
UA	AA	-0.49				-0.15
UA	AC			-0.43		-0.10
UA	CA	-0.45		-1.94		-1.72
UA	CC	-0.69				-0.32
UA	CU		-0.14			0.07
UA	GA	-1.17			-0.99	-0.95
UA	GC	-1.28				-0.97
UA	GU	-1.07				-0.73
UA	UA			-2.41		-2.31
UA	UC	-1.72	-2.21	-2.28	-0.78	-2.10

UA	UU	-0.23		-0.10	-0.04
UG	AA	-1.32			-1.02
UG	CA	-0.39			-0.07
UG	GA		-0.61	-1.74	-1.47
UG	UC	0.16	-0.29	-0.93	-0.58
UG	UG			-1.18	-0.86
UG	UU			-1.67	-1.45

Table S 2.17: Internal Loop (1x1) Statistical Energies (SE) (kcal/mol) derived from tRNA, Eukaryotic 5S rRNA, Bacterial 5S rRNA, Bacterial 16S rRNA, and all-sequence.

5' End BP	3' End BP	5' Strand IL	3' Strand IL	tRNA SE	Euk 5S SE	Bact 5S SE	Bact 16S SE	All-sequence SE
AU	CG	A	C				-0.81	-0.70
AU	GC	A	A				-1.34	-1.28
AU	GC	C	U				-1.23	-1.16
AU	GC	G	G				-0.60	-0.45
AU	GC	U	C			-1.53		-1.48
CG	AU	A	A			-1.71	-0.80	-1.52
CG	AU	C	C			-1.21		-1.14
CG	AU	G	A			-2.10	-1.26	-1.92
CG	AU	U	C			-1.31		-1.25
CG	CG	A	C		-1.38			-1.32
CG	CG	A	G				-1.50	-1.45
CG	CG	G	A				-1.11	-1.03
CG	CG	U	U				-0.86	-0.75
CG	GC	A	A				-0.90	-0.80
CG	GC	G	A				-0.81	-0.70
CG	GC	G	G				-0.60	-0.45
CG	GC	U	U		-1.50			-1.45
CG	GU	G	A				-1.24	-1.17
CG	UA	A	G				-1.07	-0.99
CG	UA	C	U		-1.55			-1.50
CG	UG	A	G				-1.47	-1.42
CG	UG	U	U				-0.84	-0.73
GC	AU	G	A				-1.82	-1.79
GC	CG	C	A				-0.69	-0.56
GC	CG	C	U				-1.62	-1.58
GC	CG	U	C				-1.55	-1.50

GC	CG	U	U			-2.10	-2.07
GC	GC	A	A			-0.63	-0.49
GC	GC	A	C			-0.96	-0.86
GC	GC	C	A			-0.85	-0.74
GC	GC	G	G			-1.20	-1.13
GC	GU	A	G			-0.99	-0.90
GU	CG	C	A			-1.08	-1.00
GU	GU	A	G			-0.68	-0.55
UA	AU	A	A			-1.04	-0.95
UA	AU	G	A		-1.58		-1.53
UA	CG	A	G			-0.86	-0.75
UA	CG	U	C			-0.97	-0.87
UA	GC	G	G			-0.70	-0.57
UA	GU	A	G			-0.76	-0.64
UG	AU	U	U	-1.63			-1.59
UG	CG	G	A			-1.72	-1.68
UG	CG	U	C		-1.44		-1.39
UG	CG	U	U			-0.90	-0.80
UG	GC	U	U	-2.10			-2.07
UG	GU	U	U	-1.55			-1.50
UG	UA	U	C		-1.30		-1.24

Table S 2.18: Internal Loop (1x2) Statistical Energies (SE) (kcal/mol) derived from Bacterial 16S rRNA and all-sequences. SE for tRNA, Eukaryotic 5S rRNA, and Bacterial 5S rRNA are all at its maximum value of 5.5 kcal/mol since there are no statistics for those structures.

5' End BP	3' End BP	5' Strand IL	3' Strand IL	Bact 16S SE	All-sequence SE
CG	AU	A	AA	0.31	0.31
CG	CG	A	AA	0.38	0.38
CG	CG	G	GA	0.30	0.30
CG	GC	C	AA	0.37	0.37
CG	GC	G	AG	0.27	0.27
CG	UG	G	GA	0.34	0.34
GC	UA	U	UC	0.40	0.40
UA	UG	G	GA	0.25	0.25
UG	UG	G	GA	0.23	0.23

Table S 2.19: Internal Loop (2x2) Statistical Energies (SE) (kcal/mol) derived from Bacterial 16S rRNA and all-sequences. SE for tRNA, Eukaryotic 5S rRNA, and Bacterial 5S rRNA are all at its maximum value of 3.4 kcal/mol since there are no statistics for those structures.

5' End BP	3' End BP	5' Strand IL	3' Strand IL	Bact 16S SE	All-sequence SE
AU	GC	AA	AG	-4.71	-4.70
AU	GC	GA	AG	-3.18	-3.17
AU	GU	AA	AG	-4.09	-4.08
AU	GU	GA	AA	-3.54	-3.53
AU	GU	GA	AG	-4.76	-4.75
CG	AU	AA	AA	-3.54	-3.53
CG	AU	GA	AA	-3.40	-3.39
CG	AU	GA	AG	-3.47	-3.46
CG	CG	AA	GG	-4.09	-4.08
CG	CG	GA	AG	-2.95	-2.94
CG	CG	UU	UU	-3.75	-3.74
CG	GC	AA	AA	-3.43	-3.42
CG	GC	GA	AG	-2.95	-2.94
CG	GU	GA	AA	-3.31	-3.30
CG	GU	GA	AG	-3.03	-3.02
CG	GU	UU	UU	-3.83	-3.82
CG	UA	AA	GG	-3.42	-3.41
CG	UG	AA	AG	-3.66	-3.65
CG	UG	GA	AG	-4.30	-4.29
GC	CG	UU	UU	-3.75	-3.74
GC	GC	GA	AG	-3.24	-3.23
GC	GU	AA	AG	-4.57	-4.56
GC	GU	GA	AA	-3.66	-3.65
GC	GU	GA	AG	-3.54	-3.53
UA	CG	AA	GG	-4.90	-4.89

Table S 2.20: Internal Loop Flank Statistical Energies (SE) (kcal/mol) derived from Eukaryotic 5S rRNA, Bacterial 5S rRNA, and Bacterial 16S rRNA and all-sequences. SE for tRNA are all at its maximum value of 3.4 kcal/mol since there are no statistics for those structures.

5' End BP	3' End BP	Euk 5S SE	Bact 5S SE	Bact 16S SE	All-sequence SE
AU	AA	-0.13		-0.47	-0.18
AU	AC	-0.14		0.02	0.09
AU	AG			-0.07	0.27
AU	AU			0.47	0.62
AU	CA	-0.25		-0.49	-0.23
AU	CC	-0.06		0.68	0.28
AU	CG			0.64	0.68
AU	CU			0.13	0.43
AU	GA	-0.19	-0.45	-0.31	-0.27
AU	GG			0.6	0.67
AU	UA	-0.3		0.3	0.04
AU	UC	-0.56	-0.5	0.38	-0.37
AU	UU			0.18	0.47
CG	AA	-0.43	-0.57	-0.91	-0.66
CG	AC	-0.69	-0.03	0.33	-0.35
CG	AG	0.03		-0.73	-0.4
CG	AU	-0.72	-0.22	-0.45	-0.47
CG	CA	-1.08	-0.87	-0.72	-0.89
CG	CC	-0.08	-0.63	0.52	-0.3
CG	CG			-0.15	0.2
CG	CU	-0.21		-0.15	-0.01
CG	GA	-0.56	-1.07	-0.83	-0.85
CG	GC			-0.5	-0.15
CG	GG			-0.57	-0.23
CG	GU	-0.1	-0.28	-0.51	-0.28
CG	UA	-0.43	-0.37	0.04	-0.25
CG	UC	-0.51	-0.14	0.45	-0.21
CG	UG			-0.24	0.12
CG	UU	-0.46		-0.69	-0.46
GC	AA	-0.64		-0.72	-0.55
GC	AC	-0.74		0.07	-0.41
GC	AG	-1.1		0	-0.81
GC	AU			-0.45	-0.09
GC	CA	-0.03	-0.37	-0.67	-0.4
GC	CC	-0.11		-0.51	-0.21
GC	CU		0.04	-0.56	-0.23
GC	GA	-0.84	-0.8	-1.09	-0.9

GC	GC			0	0.33
GC	GG	-0.25	-0.49	0.09	-0.24
GC	GU			-0.72	-0.4
GC	UA			-0.94	-0.67
GC	UC	-0.7	0.04	-0.71	-0.55
GC	UU	-0.23		-1.1	-0.81
GU	AA			0.41	0.59
GU	AC			-0.19	0.17
GU	AG			0.19	0.47
GU	AU			0.17	0.46
GU	CA			-0.57	-0.23
GU	CU			0.62	0.67
GU	GA			0.55	0.65
GU	GU			-0.49	-0.14
GU	UU			0.5	0.63
UA	AA	-0.13		-0.35	-0.09
UA	AC	-0.45		0.38	-0.09
UA	AG		-0.09	-0.57	-0.26
UA	AU			0	0.33
UA	CA		-0.12	0.56	0.22
UA	CC	-0.18			0.17
UA	CG			0.45	0.61
UA	CU	-0.46			-0.1
UA	GA	-0.07	-0.55	-0.73	-0.5
UA	GC			0.14	0.44
UA	GG			-0.25	0.11
UA	GU		-0.13	-0.63	-0.32
UA	UA			0.4	0.59
UA	UC	-0.56		-0.6	-0.44
UA	UG		-0.32	-0.18	-0.09
UA	UU	-0.89		-0.1	-0.58
UG	AA			-0.74	-0.43
UG	AC			-0.22	0.14
UG	AG		-0.2	-0.15	-0.01
UG	AU			-0.65	-0.32
UG	CA			0.56	0.65
UG	CC		0.04		0.36
UG	CU			0.59	0.66
UG	GA		-1.1	-0.72	-0.86
UG	GG			0.36	0.57

UG	GU			0.63	0.68
UG	UA			-0.21	0.15
UG	UC	-0.17	-0.51	0.69	-0.22
UG	UU	-0.89		0.13	-0.57

Table S 2.21: Folding accuracy using non-symmetric Base-pair Stack Statistical Energies (BP-ST SE) derived from tRNA, Eukaryotic 5S rRNA, Bacterial 5S rRNA, Bacterial 16S rRNA, and all-sequence to predict each molecule type.

RNA Molecule	Original mFold	tRNA SE	Euk 5S rRNA SE	Bact 5S rRNA SE	Bact 16S rRNA SE	All-sequence SE
tRNA	.70	.79	.71	.61	.69	0.71
Eukaryotic 5S rRNA	.72	.61	.71	.70	.72	0.74
Bacterial 5S rRNA	.63	.53	.59	.74	.66	0.65
Bacterial 16S rRNA	.49	.38	.38	.37	.50	0.47

Table S 2.22: Folding accuracy using non-symmetric Base-pair Stack, Hairpin, and Internal Loops Statistical Energies (BP-ST-HP-IL SE) derived from tRNA, Eukaryotic 5S rRNA, Bacterial 5S rRNA, Bacterial 16S rRNA, and all-sequences to predict each molecule type.

RNA Molecule	tRNA SE	Euk 5S rRNA SE	Bact 5S rRNA SE	Bact 16S rRNA SE	All-sequences SE
tRNA	.89	.62	.65	.32	0.80
Eukaryotic 5S rRNA	.50	.84	.70	.32	0.79
Bacterial 5S rRNA	.48	.68	.88	.35	0.77
Bacteria 16S rRNA	.32	.32	.35	.56	0.53

Table S 2.23: Folding accuracy of tRNA using original mFold energy values, symmetric BP-ST SE, asymmetric BP-ST SE (rows) each of which including no additional change, with HP SE, IL SE, and HP-IL SE.

	Modification with SE			
tRNA BP-ST SE	No additional changes	HP SE	IL SE	HP-IL SE

Original mFold	.70	.85	.78	.89
Symmetric BP-ST SE	.73	.84	.80	.88
Asymmetric BP-ST SE	.79	.87	.83	.89

Table S 2.24: Folding accuracy of Eurkaryotic 5S rRNA using original mFold energy values, symmetric BP-ST SE, asymmetric BP-ST SE (rows) each of which including no additional change, with HP SE, IL SE, and HP-IL SE.

		Modification with SE		
Eukaryotic 5S rRNA BP Stacking	No additional Changes	HP SE	IL SE	HP-IL SE
Original mFold	.72	.79	.74	.82
Symmetric BP-ST SE	.73	.80	.76	.84
Asymmetric BP-ST SE	.71	.77	.76	.84

Table S 2.25: Folding accuracy of Bacterial 5S rRNA using original mFold energy values, symmetric BP-ST SE, asymmetric BP-ST SE (rows) each of which including no additional change, with HP SE, IL SE, and HP-IL SE.

		Modification with SE		
Bacterial 5S rRNA BP Stacking	No additional changes	HP SE	IL SE	HP-IL SE
Original mFold	.63	.74	.78	.86
Symmetric BP-ST SE	.72	.78	.83	.85
Asymmetric BP-ST SE	.74	.79	.81	.88

Table S 2.26: Folding accuracy of 16S Bacteria rRNA using original mFold energy values, symmetric BP-ST SE, asymmetric BP-ST SE (rows) each of which including no additional change, with HP SE, IL SE, and HP-IL SE.

16S Bacteria BP Stacking	No additional changes	HP SE	IL SE	HP-IL SE
Original mFold	.50	.54	.52	.55
Symmetric BP-ST SE	.45	.49	.47	.51
Asymmetric BP-ST SE	.50	.53	.53	.56

Table S 2.27: Folding accuracy of tRNA using statistical energy derived from all-sequences.

tRNA BP Stacking	No additional changes	Hairpins	Internal Loops	Hairpin + Internal Loops
Original mFold	.70	.83	.67	.79
Symmetric	.59	.70	.57	.68
Non-Symmetric	.71	.79	.69	.80

Table S 2.28: Folding accuracy of Eukaryotic 5S rRNA using statistical energy derived from all-sequences

5S Eukaryote BP Stacking	No additional changes	Hairpins	Internal Loops	Hairpin + Internal Loops
Original mFold	.72	.76	.71	.75
Symmetric	.71	.74	.72	.74
Non-Symmetric	.74	.78	.74	.79

Table S 2.29: Folding accuracy of Bacterial 5S rRNA using SE derived from all-sequences.

5S Bacteria BP Stacking	No additional changes	Hairpins	Internal Loops	Hairpin + Internal Loops
Original mFold	.63	.65	.75	.78
Symmetric	.70	.68	.79	.80
Non-Symmetric	.65	.66	.76	.77

Table S 2.30: Folding accuracy of Bacterial 16S rRNA using SE derived from all-sequences.

16S Bacteria BP Stacking	No additional changes	Hairpins	Internal Loops	Hairpin + Internal Loops
Original mFold	.49	.50	.53	.52
Symmetric	.42	.44	.46	.47
Non-Symmetric	.47	.50	.49	.53

Table S 2.31: Symmetric statistical energy (kcal/mol) derived from All sequence, Bacterial 16S rRNA, Bacterial 5S rRNA and tRNA. BP1 is the 5' base-pair BP2 is the 3' base-pair.

BP1	BP2	All sequence	Bact 16S rRNA	Bact 5S rRNA	Euk 5S rRNA	tRNA
AU	AU	-2.53	-2.49	-2.78	-2.36	-2.81
AU	CG	-3.09	-3.11	-3.09	-3.40	-2.36
AU	GC	-3.19	-3.15	-3.05	-3.09	-3.4
AU	GU	-1.33	-1.41	-1.08	-0.93	-0.51
AU	UA	-1.65	-1.58	-0.98	-1.72	-2.13
AU	UG	-2.03	-2.09	-3.14	-1.45	-0.99
CG	AU	-3.19	-3.20	-3.38	-3.25	-2.85
CG	CG	-3.40	-3.40	-3.4	-3.34	-3.19
CG	GC	-2.43	-2.47	-1.9	-2.43	-2.07
CG	GU	-1.76	-1.79	-1.98	-2.27	-0.53
CG	UA	-3.19	-3.15	-3.05	-3.09	-3.4
CG	UG	-2.32	-2.38	-2.14	-2.44	-1.5
GC	AU	-2.91	-2.85	-3.07	-2.73	-3.16
GC	CG	-2.70	-2.70	-2.28	-2.4	-2.68
GC	GC	-3.40	-3.40	-3.4	-3.34	-3.19
GC	GU	-1.93	-2.01	-0.71	-1.29	-1.51
GC	UA	-3.09	-3.11	-3.09	-3.4	-2.36
GC	UG	-2.35	-2.42	-2.84	-2.37	-1.05
GU	AU	-1.62	-1.68	-1.27	-1.46	-0.85
GU	CG	-2.35	-2.42	-2.84	-2.37	-1.05
GU	GC	-2.32	-2.38	-2.14	-2.44	-1.5
GU	GU	-0.83	-1.00	2.27	0.69	1.66
GU	UA	-2.03	-2.09	-3.14	-1.45	-0.99
GU	UG	0.14	0.07	-1.4	1.91	2.1
UA	AU	-2.04	-2.04	-0.36	-2.23	-1.94
UA	CG	-2.91	-2.85	-3.07	-2.73	-3.16
UA	GC	-3.19	-3.20	-3.38	-3.25	-2.85
UA	GU	-1.07	-1.10	-0.98	-1.42	0
UA	UA	-2.53	-2.49	-2.78	-2.36	-2.81
UA	UG	-1.62	-1.68	-1.27	-1.46	-0.85
UG	AU	-1.07	-1.10	-0.98	-1.42	0
UG	CG	-1.93	-2.01	-0.71	-1.29	-1.51
UG	GC	-1.76	-1.79	-1.98	-2.27	-0.53

UG	GU	-1.22	-1.33	-2.22	-0.08	2.31
UG	UA	-1.33	-1.41	-1.08	-0.93	-0.51
UG	UG	-0.83	-1.00	2.27	0.69	1.66

Table S 2.32: Asymmetric statistical energy (kcal/mol) derived from All sequence, Bacterial 16S rRNA, Bacterial 5S rRNA and tRNA. BP1 is the 5' base-pair BP2 is the 3' base-pair.

BP1	BP2	All sequence	Bact 16S rRNA	Bact 5S rRNA	Euk 5S rRNA	tRNA
AU	AU	-1.95	-1.98	-2.48	-1.51	-1.74
AU	CG	-3.02	-3.07	-2.95	-3.40	-1.87
AU	GC	-3.17	-3.29	-2.22	-2.61	-2.42
AU	GU	-0.44	-0.58	1.41	0.35	0.67
AU	UA	-2.00	-1.95	-1.12	-1.95	-2.34
AU	UG	-1.30	-1.33	-3.18	-0.97	0.73
CG	AU	-2.81	-2.88	-2.60	-2.58	-2.28
CG	CG	-3.18	-3.30	-2.84	-2.72	-2.31
CG	GC	-2.96	-3.06	-2.18	-2.75	-2.27
CG	GU	-1.22	-1.29	0.16	-1.90	0.29
CG	UA	-2.92	-2.75	-3.01	-2.84	-3.40
CG	UG	-1.83	-1.93	-2.01	-2.14	-0.25
GC	AU	-2.52	-2.48	-3.23	-2.60	-2.36
GC	CG	-3.27	-3.34	-2.61	-2.72	-2.94
GC	GC	-3.40	-3.40	-3.28	-3.26	-3.13
GC	GU	-1.66	-1.75	1.43	-0.84	-1.44
GC	UA	-2.82	-2.93	-2.46	-2.63	-1.91
GC	UG	-2.14	-2.32	-2.55	-1.15	-0.20
GU	AU	-1.30	-1.39	-0.55	-1.25	-0.35
GU	CG	-1.89	-1.94	-2.32	-2.38	-0.65
GU	GC	-2.13	-2.24	-1.11	-1.84	-1.35
GU	GU	0.04	-0.13	Inf	1.15	1.96
GU	UA	-1.90	-2.03	-2.18	-0.76	-0.90
GU	UG	0.17	0.08	-1.60	2.16	2.30
UA	AU	-2.48	-2.52	-0.41	-2.52	-2.12
UA	CG	-2.85	-2.86	-1.75	-1.95	-3.05
UA	GC	-3.25	-3.31	-3.39	-3.18	-2.56
UA	GU	0.01	-0.09	0.61	0.29	1.05
UA	UA	-2.47	-2.44	-2.27	-2.20	-2.77
UA	UG	-0.94	-1.05	-0.74	-0.34	-0.09
UG	AU	-0.81	-0.83	-0.83	-1.39	0.44

UG	CG	-1.34	-1.51	-0.64	-0.53	0.03
UG	GC	-1.39	-1.44	-2.11	-1.72	-0.02
UG	GU	-1.48	-1.65	-2.54	-0.09	2.53
UG	UA	-1.05	-1.15	-1.12	-0.69	-0.20
UG	UG	-0.37	-0.59	2.60	2.16	3.93

3 REFERENCES

1. McCool, M.A.; Collings, A.F.; Woolf, L.A., *Pressure and temperature dependence of the self-diffusion of benzene.* Journal of the Chemical Society, Faraday Transactions 1, 1972. **68**: p. 1489 - 1497.
2. Wensink, E.J.W.; Hoffmann, A.C.; van Maaren, P.J.; van der Spoel, D., *Dynamic properties of water/alcohol mixtures studied by computer simulation.* J. Chem. Phys., 2003. **119**(14): p. 7308-7317.
3. Weast, R.C., *CRC handbook of chemistry and physics*. 1st Student ed. 1988, Boca Raton, FL: CRC Press. 1 v. (various pagings).
4. Mikhail, S.Z.; Kimel, W.R., *Densities and Viscosities of Methanol-Water Mixtures.* J. Chem. Eng. Data, 1961. **6**(4): p. 533 - 537.
5. Mathews, D.H.; Sabina, J.; Zuker, M.; Turner, D.H., *Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure.* J. Mol. Biol., 1999. **288**(5): p. 911-940.

References

1. Abraham, M.H.; Whiting, G.S.; Fuchs, R.; Chambers, E.J., *Thermodynamics of Solute Transfer from Water to Hexadecane.* J. Chem. Soc., Perkin Trans. 2, 1990(2): p. 291-300.
2. Ahmed, H.U.; Blakeley, M.P.; Cianci, M.; Cruickshank, D.W.J.; Hubbard, J.A.; Helliwell, J.R., *The determination of protonation states in proteins.* Acta Crystallographica Section D-Biological Crystallography, 2007. **63**: p. 906-922.
3. Albrieux, F.; Calvo, F.; Chirot, F.; Vorobyev, A.; Tsybin, Y.O.; Lepere, V.; Antoine, R.; Lemoine, J.; Dugourd, P., *Conformation of Polyalanine and Polyglycine Dications in the Gas Phase: Insight from Ion Mobility Spectrometry and Replica-Exchange Molecular Dynamics.* J. Phys. Chem. A, 2010. **114**(25): p. 6888-6896.
4. Allinger, N.L.; Yuh, Y.H.; Lii, J.H., *Molecular Mechanics - the MM3 Force-Field for Hydrocarbons .1.* J. Am. Chem. Soc., 1989. **111**(23): p. 8551-8566.
5. Amin, E.A.; Truhlar, D.G., *Zn coordination chemistry: Development of benchmark suites for geometries, dipole moments, and bond dissociation energies and their use to test and validate density functionals and molecular orbital theory.* J. Chem. Theory Comput., 2008. **4**(1): p. 75-85.
6. Andersen, H.C., *Rattle - A Velocity Version of the Shake Algorithm for Molecular-Dynamics Calculations.* J. Comput. Phys., 1983. **52**(1): p. 24-34.
7. Andersson, Y.; Hult, E.; Apell, P.; Langreth, D.C.; Lundqvist, B.I., *Density-functional account of van der Waals forces between parallel surfaces.* Solid State Commun., 1998. **106**(5): p. 235-238.
8. Antao, V.P.; Tinoco, I., *Thermodynamic Parameters For Loop Formation In RNA And DNA Hairpin Tetraloops.* Nucleic Acids Res., 1992. **20**(4): p. 819-824.
9. Antony, J.; Piquemal, J.P.; Gresh, N., *Complexes of thiomandelate and captopril mercaptocarboxylate inhibitors to metallo-beta-lactamase by polarizable molecular mechanics. Validation on model binding sites by quantum chemistry.* J. Comput. Chem., 2005. **26**(11): p. 1131-1147.
10. Asthagiri, D.; Pratt, L.R.; Paulaitis, M.E.; Rempe, S.B., *Hydration structure and free energy of biomolecularly specific aqueous dications, including Zn2+ and first transition row metals.* J. Am. Chem. Soc., 2004. **126**(4): p. 1285-1289.
11. Badyal, Y.S.; Barnes, A.C.; Cuello, G.J.; Simonson, J.M., *Understanding the effects of concentration on the solvation structure of Ca2+ in aqueous solutions. II: Insights into longer range order from neutron diffraction isotope substitution.* J. Phys. Chem. A, 2004. **108**(52): p. 11819-11827.
12. Bagus, P.S.; Illas, F., *Decomposition of the chemisorption bond by constrained variations - Order of the variations and construction of the variational spaces.* J. Chem. Phys., 1992. **96**(12): p. 8962-8970.

13. Battle, D.J.; Doudna, J.A., *Specificity of RNA-RNA helix recognition.* Proc. Natl. Acad. Sci. U. S. A., 2002. **99**(18): p. 11676-11681.
14. Bauer, B.A.; Warren, G.L.; Patel, S., *Incorporating Phase-Dependent Polarizability in Nonadditive Electrostatic Models for Molecular Dynamics Simulations of the Aqueous Liquid-Vapor Interface.* J. Chem. Theory Comput., 2009. **5**(2): p. 359-373.
15. Beck, T., *Hydration Free Energies by Energetic Partitioning of the Potential Distribution Theorem.* Journal of Statistical Physics, 2011: p. 1-20.
16. Becke, A.D., *Density-Functional Exchange-Energy Approximation with Correct Asymptotic-Behavior.* Phys. Rev. A, 1988. **38**(6): p. 3098-3100.
17. Becke, A.D.; Edgecombe, K.E., *A simple measure of electron localization in atomic and molecular-systems.* J. Chem. Phys., 1990. **92**(9): p. 5397-5403.
18. Bennett, C.H., *Efficient Estimation of Free-Energy Differences from Monte-Carlo Data.* J. Comput. Phys., 1976. **22**(2): p. 245-268.
19. Berendsen, H.J.C.; Postma, J.P.M.; Vangunsteren, W.F.; Dinola, A.; Haak, J.R., *Molecular-Dynamics with Coupling to an External Bath.* J. Chem. Phys., 1984. **81**(8): p. 3684-3690.
20. Berka, K.; Laskowski, R.; Riley, K.E.; Hobza, P.; Vondrasek, J., *Representative Amino Acid Side Chain Interactions in Proteins. A Comparison of Highly Accurate Correlated ab Initio Quantum Chemical and Empirical Potential Procedures.* J. Chem. Theory Comput., 2009. **5**(4): p. 982-992.
21. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E., *The Protein Data Bank.* Nucleic Acids Res., 2000. **28**(1): p. 235-242.
22. Best, R.B.; Buchete, N.V.; Hummer, G., *Are current molecular dynamics force fields too helical?* Biophys. J., 2008. **95**(1): p. L7-L9.
23. Beutlera, T.C.; Marka, A.E.; van Schaikb, R.C.; Gerberc, P.R.; van Gunsteren, W.F., *Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations* Chem. Phys. Lett., 1994. **222**(6): p. 529-539.
24. Bloch, F., *Fundamentals of statistical mechanics : manuscript and notes of Felix Bloch*. 2000, London : Singapore :: Imperial College Press ; World Scientific. xii, 302 p. :.
25. Born, M., *Volumen und Hydratationswärme der Ionen.* Zeitschrift für Physik A Hadrons and Nuclei, 1920. **1**(1): p. 45-48.
26. Branlant, C.; Krol, A.; Machatt, M.A.; Pouyet, J.; Ebel, J.P.; Edwards, K.; Kossel, H., *Primary and Secondary Structures of Escherichia-Coli Mre-600-23s Ribosomal-Rna - Comparison with Models of Secondary Structure for Maize Chloroplast 23s Ribosomal-Rna and for Large Portions of Mouse and Human 16s Mitochondrial Ribosomal-Rnas.* Nucleic Acids Res., 1981. **9**(17): p. 4303-4324.
27. Bryant, S.H.; Lawrence, C.E., *The Frequency Of Ion-Pair Substructures In Proteins Is Quantitatively Related To Electrostatic Potential - A Statistical-Model*

For Nonbonded Interactions. Proteins-Structure Function and Genetics, 1991. **9**(2): p. 108-119.
28. Buckingham, A.D., *Permanent and Induced Molecular Moments and Long-Range Intermolecular Forces*. Adv. Chem. Phys. 1967: John Wiley & Sons, Inc. 107-142.
29. Burnham, C.J.; Li, J.C.; Xantheas, S.S.; Leslie, M., *The parametrization of a Thole-type all-atom polarizable water model from first principles and its application to the study of water clusters (n=2-21) and the phonon spectrum of ice Ih.* J. Chem. Phys., 1999. **110**(9): p. 4566-4581.
30. Bush, B.L.; Bayly, C.I.; Halgren, T.A., *Consensus bond-charge increments fitted to electrostatic potential or field of many compounds: Application to MMFF94 training set.* J. Comput. Chem., 1999. **20**(14): p. 1495-1516.
31. C. Czaplewski; A. Liwo; M. Makowski; S. Oldziej; H.A. Scheraga, *Coarse-Grained Models of Proteins: Theory and Applications* in *Multiscale Approaches to Protein Modeling*, A. Kolinski, Editor. 2010, Springer.
32. Caminiti, R.; Licheri, G.; Piccaluga, G.; Pinna, G., *X-ray-diffraction study of a 3-ion aqueous-solution.* Chem. Phys. Lett., 1977. **47**(2): p. 275-278.
33. Cech, T.R., *Conserved Sequences and Structures of Group-I Introns - Building an Active-Site for Rna Catalysis - a Review.* Gene, 1988. **73**(2): p. 259-271.
34. Chambers, C.C.; Hawkins, G.D.; Cramer, C.J.; Truhlar, D.G., *Model for aqueous solvation based on class IV atomic charges and first solvation shell effects.* J. Phys. Chem., 1996. **100**(40): p. 16385-16398.
35. Chou, P.Y.; Fasman, G.D., *Conformational Parameters for Amino-Acids in Helical, Beta-Sheet, and Random Coil Regions Calculated from Proteins.* Biochemistry, 1974. **13**(2): p. 211-222.
36. Chu, J.W.; Izvekov, S.; Voth, G.A., *The multiscale challenge for biomolecular systems: coarse-grained modeling.* Mol. Simul., 2006. **32**(3-4): p. 211-218.
37. Cieplak, P.; Dupradeau, F.Y.; Duan, Y.; Wang, J.M., *Polarization effects in molecular mechanical force fields.* Journal of Physics-Condensed Matter, 2009. **21**(33): p. 333102.
38. Cisneros, G.A.; Darden, T.A.; Gresh, N.; Pilmé, J.; Reinhardt, P.; Parisel, O.; Piquemal, J.P., *Design Of Next Generation Force Fields From AB Initio Computations: Beyond Point Charges Electrostatics*, in *Multi-scale Quantum Models for Biocatalysis*. 2009, Springer Netherlands. p. 137-172.
39. Cleaver, D.J.; Care, C.M.; Allen, M.P.; Neal, M.P., *Extension and generalization of the Gay-Berne potential.* Physical Review E, 1996. **54**(1): p. 559-567.
40. Constanciel, R.; Contreras, R., *Self-Consistent Field-Theory of Solvent Effects Representation by Continuum Models - Introduction of Desolvation Contribution.* Theor. Chim. Acta, 1984. **65**(1): p. 1-11.
41. Costa, M.; Michel, F., *Frequent Use Of The Same Tertiary Motif By Self-Folding RNAs.* EMBO J., 1995. **14**(6): p. 1276-1285.

42. Counterman, A.E.; Clemmer, D.E., *Large anhydrous polyalanine ions: Evidence for extended helices and onset of a more compact state.* J. Am. Chem. Soc., 2001. **123**(7): p. 1490-1498.
43. D'Angelo, P.; Barone, V.; Chillemi, G.; Sanna, N.; Meyer-Klaucke, W.; Pavel, N.V., *Hydrogen and Higher Shell Contributions in Zn2+, Ni2+, and Co2+ Aqueous Solutions: An X-ray Absorption Fine Structure and Molecular Dynamics Study.* J. Am. Chem. Soc., 2002. **124**(9): p. 1958-1967.
44. Darden, T.; York, D.; Pedersen, L., *Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems.* J. Chem. Phys., 1993. **98**(12): p. 10089-10092.
45. Das, R.; Baker, D., *Automated de novo prediction of native-like RNA tertiary structures.* Proc. Natl. Acad. Sci. U. S. A., 2007. **104**(37): p. 14664-14669.
46. De Courcy, B.; Gresh, N.; Piquemal, J.P., *Importance of lone pair interactions/redistribution in hard and soft ligands within the active site of alcohol dehydrogenase Zn-metalloenzyme: Insights from electron localization function.* Interdisciplinary Sciences: Computational Life Sciences, 2009. **1**(1): p. 55-60.
47. de Courcy, B.; Pedersen, L.G.; Parisel, O.; Gresh, N.; Silvi, B.; Pilme, J.; Piquemal, J.P., *Understanding Selectivity of Hard and Soft Metal Cations within Biological Systems Using the Subvalence Concept. 1. Application to Blood Coagulation: Direct Cation-Protein Electronic Effects versus Indirect Interactions through Water Networks.* J. Chem. Theory Comput., 2010. **6**(4): p. 1048-1063.
48. de Courcy, B.; Piquemal, J.P.; Gresh, N., *Energy Analysis of Zn Polycoordination in a Metalloprotein Environment and of the Role of a Neighboring Aromatic Residue. What Is the Impact of Polarization?* J. Chem. Theory Comput., 2008. **4**(10): p. 1659-1668.
49. DeVane, R.; Klein, M.L.; Chiu, C.C.; Nielsen, S.O.; Shinoda, W.; Moore, P.B., *Coarse-Grained Potential Models for Phenyl-Based Molecules: I. Parametrization Using Experimental Data.* J. Phys. Chem. B, 2010. **114**(19): p. 6386-6393.
50. Dima, R.I.; Hyeon, C.; Thirumalai, D., *Extracting stacking interaction parameters for RNA from the data set of native structures.* J. Mol. Biol., 2005. **347**(1): p. 53-69.
51. Do, C.B.; Woods, D.A.; Batzoglou, S., *CONTRAfold: RNA secondary structure prediction without physics-based models.* Bioinformatics, 2006. **22**(14): p. E90-E98.
52. Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M.C.; Xiong, G.M.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.M.; Kollman, P., *A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations.* J. Comput. Chem., 2003. **24**(16): p. 1999-2012.

53. Dunning, T.H., *Gaussian-Basis Sets for Use in Correlated Molecular Calculations .1. The Atoms Boron through Neon and Hydrogen.* J. Chem. Phys., 1989. **90**(2): p. 1007-1023.
54. Durbin, R.; Eddy, S.R.; Krogh, A.; Mitchison, G., *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* 1998: Cambridge University Press.
55. Elgavish, T.; Cannone, J.J.; Lee, J.C.; Harvey, S.C.; Gutell, R.R., *AA.AG@Helix.Ends: A : A and A : G base-pairs at the ends of 16 S and 23 S rRNA helices.* J. Mol. Biol., 2001. **310**(4): p. 735-753.
56. Essmann, U.; Perera, L.; Berkowitz, M.L.; Darden, T.; Lee, H.; Pedersen, L.G., *A Smooth Particle Mesh Ewald Method.* J. Chem. Phys., 1995. **103**(19): p. 8577-8593.
57. Estiu, G.; Suarez, D.; Merz, K.M., *Quantum mechanical and molecular dynamics simulations of ureases and Zn beta-lactamases.* J. Comput. Chem., 2006. **27**(12): p. 1240-1262.
58. Fatmi, M.Q.; Hofer, T.S.; Randolf, B.R.; Rode, B.M., *An extended ab initio QM/MM MD approach to structure and dynamics of Zn(II) in aqueous solution.* J. Chem. Phys., 2005. **123**(5): p. 054514-8.
59. Fatmi, M.Q.; Hofer, T.S.; Randolf, B.R.; Rode, B.M., *Temperature Effects on the Structural and Dynamical Properties of the Zn(II)−Water Complex in Aqueous Solution: A QM/MM Molecular Dynamics Study.* J. Phys. Chem. B, 2006. **110**(1): p. 616-621.
60. Faver, J.C.; Benson, M.L.; He, X.A.; Roberts, B.P.; Wang, B.; Marshall, M.S.; Kennedy, M.R.; Sherrill, C.D.; Merz, K.M., *Formal Estimation of Errors in Computed Absolute Interaction Energies of Protein-Ligand Complexes.* J. Chem. Theory Comput., 2011. **7**(3): p. 790-797.
61. Fenn, T.D.; Schnieders, M.J.; Brunger, A.T.; Pande, V.S., *Polarizable Atomic Multipole X-Ray Refinement: Hydration Geometry and Application to Macromolecules.* Biophys. J., 2010. **98**(12): p. 2984-2992.
62. Finkelstein, A.V.; Badretdinov, A.Y.; Gutin, A.M., *Why Do Protein Architectures Have Boltzmann-Like Statistics.* Proteins-Structure Function and Genetics, 1995. **23**(2): p. 142-150.
63. Floudas, C.A.; Fung, H.K.; McAllister, S.R.; Monnigmann, M.; Rajgaria, R., *Advances in protein structure prediction and de novo protein design: A review.* Chem. Eng. Sci., 2006. **61**(3): p. 966-988.
64. Fountain, M.A.; Serra, M.J.; Krugh, T.R.; Turner, D.H., *Structural features of a six-nucleotide RNA hairpin loop found in ribosomal RNA.* Biochemistry, 1996. **35**(21): p. 6539-6548.
65. Fox, G.E.; Woese, C.R., *5S-RNA Secondary Structure.* Nature, 1975. **256**(5517): p. 505-507.
66. Friedman, H., *Hydration complexes - some firm results and some pressing questions.* Chemica Scripta, 1985. **25**(1): p. 42-48.

67. Friesner, R.A.; Guallar, V., *Ab initio quantum chemical and mixed quantum mechanics/molecular mechanics (QM/MM) methods for studying enzymatic catalysis.* Annu. Rev. Phys. Chem., 2005. **56**: p. 389-427.
68. Frisch, M.J.; Trucks, G.W.; Schlegel, H.B.; Scuseria, G.E.; Robb, M.A.; Cheeseman, J.R.; J. A. Montgomery, J.; Vreven, T.; Kudin, K.N.; Burant, J.C.; Millam, J.M.; Iyengar, S.S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G.A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J.E.; Hratchian, H.P.; Cross, J.B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R.E.; Yazyev, O.; Austin, A.J.; Cammi, R.; Pomelli, C.; Ochterski, J.W.; Ayala, P.Y.; Morokuma, K.; Voth, G.A.; Salvador, P.; Dannenberg, J.J.; Zakrzewski, V.G.; Dapprich, S.; Daniels, A.D.; Strain, M.C.; Farkas, O.; Malick, D.K.; Rabuck, A.D.; Raghavachari, K.; Foresman, J.B.; Ortiz, J.V.; Cui, Q.; Baboul, A.G.; Clifford, S.; Cioslowski, J.; Stefanov, B.B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R.L.; Fox, D.J.; Keith, T.; Al-Laham, M.A.; Peng, C.Y.; Nanayakkara, A.; Challacombe, M.; Gill, P.M.W.; Johnson, B.; Chen, W.; Wong, M.W.; Gonzalez, C.; Pople, J.A., *Gaussian 03*. 2004, Gaussian, Inc.: Wallingford, CT.
69. Frisch, M.J.; Trucks, G.W.; Schlegel, H.B.; Scuseria, G.E.; Robb, M.A.; Cheeseman, J.R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G.A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H.P.; Izmaylov, A.F.; Bloino, J.; G. Zheng; Sonnenberg, J.L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; M. Ishida, T.N.; Y. Honda; O. Kitao, H.N.; Vreven, T.; J. A. Montgomery, J.; Peralta, J.E.; Ogliaro, F.; Bearpark, M.; J. J. Heyd, E.B.; Kudin, K.N.; Staroverov, V.N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J.C.; Iyengar, S.S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J.M.; Klene, M.; Knox, J.E.; Cross, J.B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R.E.; Yazyev, O.; Austin, A.J.; Cammi, R.; Pomelli, C.; Ochterski, J.W.; Martin, R.L.; Morokuma, K.; Zakrzewski, V.G.; Voth, G.A.; Salvador, P.; Dannenberg, J.J.; Dapprich, S.; A. D. Daniels; Farkas, Ö.; Foresman, J.B.; Ortiz, J.V.; Cioslowski, J.; Fox, D.J., *Gaussian 09*. 2009, Gaussian, Inc.: Wallingford CT.
70. Gallicchio, E.; Zhang, L.Y.; Levy, R.M., *The SGB/NP hydration free energy model based on the surface generalized born solvent reaction field and novel nonpolar hydration free energy estimators.* J. Comput. Chem., 2002. **23**(5): p. 517-29.
71. Garde, S.; Hummer, G.; Paulaitis, M.E., *Free energy of hydration of a molecular ionic solute: Tetramethylammonium ion.* J. Chem. Phys., 1998. **108**(4): p. 1552-1561.
72. Gate, J.H.; Gooding, A.R.; Podell, E.; Zhou, K.H.; Golden, B.L.; Szewczak, A.A.; Kundrot, C.E.; Cech, T.R.; Doudna, J.A., *RNA tertiary structure mediation by adenosine platforms.* Science, 1996. **273**(5282): p. 1696-1699.

73. Gautheret, D.; Konings, D.; Gutell, R.R., *A Major Family Of Motifs Involving G-Center-Dot-A Mismatches In Ribosomal-RNA.* J. Mol. Biol., 1994. **242**(1): p. 1-8.
74. Glotz, C.; Zwieb, C.; Brimacombe, R.; Edwards, K.; Kossel, H., *Secondary Structure of the Large Subunit Ribosomal-Rna from Escherichia-Coli, Zea-Mays Chloroplast, and Human and Mouse Mitochondrial Ribosomes.* Nucleic Acids Res., 1981. **9**(14): p. 3287-3306.
75. Golubkov, P.A.; Ren, P.Y., *Generalized coarse-grained model based on point multipole and Gay-Berne potentials.* J. Chem. Phys., 2006. **125**(6): p. 64103.
76. Golubkov, P.A.; Wu, J.C.; Ren, P.Y., *A transferable coarse-grained model for hydrogen-bonding liquids.* Phys. Chem. Chem. Phys., 2008. **10**(15): p. 2050-2057.
77. Gordon, M.S.; Schmidt, M.W., *Advances in electronic structure theory: GAMESS a decade later*, in *Theory and Applications of Computational Chemistry, the first forty years*, C.E. Dykstra, Frenking, G., Kim, K. S., Scuseria, G. E., Editor. 2005, Elsevier: Amsterdam.
78. Graf, J.; Nguyen, P.H.; Stock, G.; Schwalbe, H., *Structure and dynamics of the homologous series of alanine peptides: A joint molecular dynamics/NMR study.* J. Am. Chem. Soc., 2007. **129**(5): p. 1179-1189.
79. Grater, F.; Schwarzl, S.M.; Dejaegere, A.; Fischer, S.; Smith, J.C., *Protein/ligand binding free energies calculated with quantum mechanics/molecular mechanics.* J. Phys. Chem. B, 2005. **109**(20): p. 10474-83.
80. Gresh, N., *Energetics of Zn2+ Binding to a Series of Biologically Relevant Ligands - a Molecular Mechanics Investigation Grounded on Ab-Initio Self-Consistent-Field Supermolecular Computations.* J. Comput. Chem., 1995. **16**(7): p. 856-882.
81. Gresh, N.; Cisneros, G.A.; Darden, T.A.; Piquemal, J.P., *Anisotropic, polarizable molecular mechanics studies of inter- and intramoecular interactions and ligand-macromolecule complexes. A bottom-up strategy.* J. Chem. Theory Comput., 2007. **3**(6): p. 1960-1986.
82. Gresh, N.; Garmer, D.R., *Comparative binding energetics of Mg2+, Ca2+, Zn2+, and Cd2+ to biologically relevant ligands: Combined ab initio SCF supermolecule and molecular mechanics investigation.* J. Comput. Chem., 1996. **17**(12): p. 1481-1495.
83. Gresh, N.; Piquemal, J.; Krauss, M., *Representation of Zn(II) complexes in polarizable molecular mechanics. Further refinements of the electrostatic and short-range contributions. Comparisons with parallel ab initio computations.* J. Comput. Chem., 2005. **26**(11): p. 1113-1130.
84. Grossfield, A., *Dependence of ion hydration on the sign of the ion's charge.* J. Chem. Phys., 2005. **122**(2): p. 024506.
85. Grossfield, A.; Ren, P.Y.; Ponder, J.W., *Ion solvation thermodynamics from simulation with a polarizable force field.* J. Am. Chem. Soc., 2003. **125**(50): p. 15671-15682.

86. Grycuk, T., *Deficiency of the Coulomb-field approximation in the generalized Born model: An improved formula for Born radii evaluation.* J. Chem. Phys., 2003. **119**(9): p. 4817-4826.
87. Guckian, K.M.; Schweitzer, B.A.; Ren, R.X.F.; Sheils, C.J.; Tahmassebi, D.C.; Kool, E.T., *Factors contributing to aromatic stacking in water: Evaluation in the context of DNA.* J. Am. Chem. Soc., 2000. **122**(10): p. 2213-2222.
88. Gutell, R.R., *COLLECTION OF SMALL-SUBUNIT (16S- AND 16S-LIKE) RIBOSOMAL-RNA STRUCTURES - 1994.* Nucleic Acids Res., 1994. **22**(17): p. 3502-3507.
89. Gutell, R.R.; Cannone, J.J.; Konings, D.; Gautheret, D., *Predicting U-turns in ribosomal RNA with comparative sequence analysis.* J. Mol. Biol., 2000. **300**(4): p. 791-803.
90. Gutell, R.R.; Cannone, J.J.; Shang, Z.; Du, Y.; Serra, M.J., *A story: Unpaired adenosine bases in ribosomal RNAs.* J. Mol. Biol., 2000. **304**(3): p. 335-354.
91. Gutell, R.R.; Lee, J.C.; Cannone, J.J., *The accuracy of ribosomal RNA comparative structure models.* Curr. Opin. Struct. Biol., 2002. **12**(3): p. 301-310.
92. Gutell, R.R.; Noller, H.F.; Woese, C.R., *Higher order structure in ribosomal RNA.* EMBO J., 1986. **5**(5): p. 1111-3.
93. Gutell, R.R.; Weiser, B.; Woese, C.R.; Noller, H.F., *Comparative Anatomy of 16-S-Like Ribosomal-RNA.* Prog. Nucleic Acid Res. Mol. Biol., 1985. **32**: p. 155-216.
94. Gutell, R.R.; Woese, C.R., *Higher-Order Structural Elements in Ribosomal-RNAs - Pseudo-Knots and the Use of Noncanonical Pairs.* Proc. Natl. Acad. Sci. U. S. A., 1990. **87**(2): p. 663-667.
95. Halgren, T.A., *Representation of Vanderwaals (Vdw) Interactions in Molecular Mechanics Force-Fields - Potential Form, Combination Rules, and Vdw Parameters.* J. Am. Chem. Soc., 1992. **114**(20): p. 7827-7843.
96. Headgordon, T.; Headgordon, M.; Frisch, M.J.; Brooks, C.L.; Pople, J.A., *Theoretical-Study of Blocked Glycine and Alanine Peptide Analogs.* J. Am. Chem. Soc., 1991. **113**(16): p. 5989-5997.
97. Hegefeld, W.A.; Chen, S.E.; DeLeon, K.Y.; Kuczera, K.; Jas, G.S., *Helix Formation in a Pentapeptide Experiment and Force-field Dependent Dynamics.* J. Phys. Chem. A, 2010. **114**(47): p. 12391-12402.
98. Helm, L.; Merbach, A.E., *Water exchange on metal ions: experiments and simulations.* Coord. Chem. Rev., 1999. **187**: p. 151-181.
99. Henzler, K.A.; Lee, D.K.; Ramamoorthy, A., *Conformational stability of solid-state poly(l-alanine).* Biophys. J., 2001. **80**(1): p. 187a-187a.
100. Hills, R.D.; Lu, L.Y.; Voth, G.A., *Multiscale Coarse-Graining of the Protein Energy Landscape.* PLoS Comput. Biol., 2010. **6**(6).
101. Hobza, P.; Sponer, J., *Structure, energetics, and dynamics of the nucleic Acid base pairs: nonempirical ab initio calculations.* Chem Rev, 1999. **99**(11): p. 3247-76.

102. Holley, R.W.; Apgar, J.; Everett, G.A.; Madison, J.T.; Marquise.M; Merrill, S.H.; Penswick, J.R.; Zamir, A., *Structure Of A Ribonucleic Acid.* Science, 1965. **147**(3664): p. 1462.
103. Huang, S.G.; Wang, Y.X.; Draper, D.E., *Structure of a hexanucleotide RNA hairpin loop conserved in ribosomal RNAs.* J. Mol. Biol., 1996. **258**(2): p. 308-321.
104. Hudgins, R.R.; Ratner, M.A.; Jarrold, M.F., *Design of helices that are stable in vacuo.* J. Am. Chem. Soc., 1998. **120**(49): p. 12974-12975.
105. Jaeger, L.; Michel, F.; Westhof, E., *Involvement Of A GNRA Tetraloop In Long-Range Tertiary Interactions.* J. Mol. Biol., 1994. **236**(5): p. 1271-1276.
106. Jakalian, A.; Bush, B.L.; Jack, D.B.; Bayly, C.I., *Fast, efficient generation of high-quality atomic Charges. AM1-BCC model: I. Method.* J. Comput. Chem., 2000. **21**(2): p. 132-146.
107. Jakalian, A.; Jack, D.B.; Bayly, C.I., *Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation.* J. Comput. Chem., 2002. **23**(16): p. 1623-1641.
108. Jalilehvand, F.; Spangberg, D.; Lindqvist-Reis, P.; Hermansson, K.; Persson, I.; Sandstrom, M., *Hydration of the calcium ion. An EXAFS, large-angle X-ray scattering, and molecular dynamics simulation study.* J. Am. Chem. Soc., 2001. **123**(3): p. 431-441.
109. James, B.D.; Olsen, G.J.; Liu, J.S.; Pace, N.R., *The Secondary Structure of Ribonuclease-P Rna, the Catalytic Element of a Ribonucleoprotein Enzyme.* Cell, 1988. **52**(1): p. 19-26.
110. Jarrold, M.F., *Helices and sheets in vacuo.* Phys. Chem. Chem. Phys., 2007. **9**(14): p. 1659-1671.
111. Jenkins, L.; Hara, T.; Durell, S.; Hayashi, R.; Inman, J.; Piquemal, J.; Gresh, N.; Appella, E., *Specificity of acyl transfer from 2-mercaptobenzamide thioesters to the HIV-1 nucleocapsid protein.* J. Am. Chem. Soc., 2007. **129**(36): p. 11067-11078.
112. Jensen, L.; Astrand, P.O.; Osted, A.; Kongsted, J.; Mikkelsen, K.V., *Polarizability of molecular clusters as calculated by a dipole interaction model.* J. Chem. Phys., 2002. **116**(10): p. 4001-4010.
113. Jiao, D.; Golubkov, P.A.; Darden, T.A.; Ren, P., *Calculation of protein-ligand binding free energy by using a polarizable potential.* Proc. Natl. Acad. Sci. U. S. A., 2008. **105**(17): p. 6290-6295.
114. Jiao, D.; King, C.; Grossfield, A.; Darden, T.A.; Ren, P.Y., *Simulation of Ca2+ and Mg2+ solvation using polarizable atomic multipole potential.* J. Phys. Chem. B, 2006. **110**(37): p. 18553-18559.
115. Jiao, D.; Zhang, J.; Duke, R.E.; Li, G.; Schnieders, M.J.; Ren, P., *Trypsin-ligand binding free energies from explicit and implicit solvent simulations with polarizable potential.* J. Comput. Chem., 2009. **30**(11): p. 1701-11.
116. Jonikas, M.A.; Radmer, R.J.; Laederach, A.; Das, R.; Pearlman, S.; Herschlag, D.; Altman, R.B., *Coarse-grained modeling of large RNA molecules with knowledge-*

based potentials and structural filters. RNA-Publ. RNA Soc., 2009. **15**(2): p. 189-199.
117. Jonikas, M.A.; Radmer, R.J.; Laederach, A.; Das, R.; Pearlman, S.; Herschlag, D.; Altman, R.B., *Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters.* RNA, 2009. **15**(2): p. 189-99.
118. Jorgensen, W.L.; Maxwell, D.S.; 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): p. 11225-11236.
119. Kaminski, G.A.; Friesner, R.A.; Tirado-Rives, J.; Jorgensen, W.L., *Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides.* J. Phys. Chem. B, 2001. **105**(28): p. 6474-6487.
120. Katz, B.A.; Elrod, K.; Luong, C.; Rice, M.J.; Mackman, R.L.; Sprengeler, P.A.; Spencer, J.; Hataye, J.; Janc, J.; Link, J.; Litvak, J.; Rai, R.; Rice, K.; Sideris, S.; Verner, E.; Young, W., *A novel serine protease inhibition motif involving a multi-centered short hydrogen bonding network at the active site.* J. Mol. Biol., 2001. **307**(5): p. 1451-86.
121. Keilin, D.; Mann, T., *Carbonic anhydrase. Purification and nature of the enzyme.* Biochem. J., 1940. **34**(8-9): p. 1163-1176.
122. Kelly, C.P.; Cramer, C.J.; Truhlar, D.G., *Aqueous solvation free energies of ions and ion-water clusters based on an accurate value for the absolute aqueous solvation free energy of the proton.* J. Phys. Chem. B, 2006. **110**(32): p. 16066-16081.
123. Kern, D.; Zuiderweg, E.R.P., *The role of dynamics in allosteric regulation.* Curr. Opin. Struct. Biol., 2003. **13**(6): p. 748-757.
124. Kiparisov, S.; Petrov, A.; Meskauskas, A.; Sergiev, P.V.; Dontsova, O.A.; Dinman, J.D., *Structural and functional analysis of 5S rRNA in Saccharomyces cerevisiae.* Mol. Genet. Genomics, 2005. **274**(3): p. 235-247.
125. Klimovich, P.V.; Mobley, D.L., *Predicting hydration free energies using all-atom molecular dynamics simulations and multiple starting conformations.* J. Comput-Aided. Mol. Des., 2010. **24**(4): p. 307-316.
126. Knudsen, B.; Hein, J., *RNA secondary structure prediction using stochastic context-free grammars and evolutionary history.* Bioinformatics, 1999. **15**(6): p. 446-454.
127. Knudsen, B.; Hein, J., *Pfold: RNA secondary structure prediction using stochastic context-free grammars.* Nucleic Acids Res., 2003. **31**(13): p. 3423-3428.
128. Kollman, P.A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D.A.; Cheatham, T.E., 3rd, *Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models.* Acc. Chem. Res., 2000. **33**(12): p. 889-97.

129. Krishnamoorthy, B.; Tropsha, A., *Development of a four-body statistical pseudo-potential to discriminate native from non-native protein conformations.* Bioinformatics, 2003. **19**(12): p. 1540-8.
130. Krishnan, C.V.; Friedman, H.L., *Solvation Enthalpies of Various Ions in Water and Heavy Water.* J. Phys. Chem., 1970. **74**(11): p. 2356.
131. Kuzmin, A.; Obst, S.; Purans, J., *X-ray absorption spectroscopy and molecular dynamics studies of Zn2+ hydration in aqueous solutions.* J. Phys.: Condens. Matter, 1997. **9**(46): p. 10065-10078.
132. Lee, C.T.; Yang, W.T.; Parr, R.G., *Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron-Density.* Physical Review B, 1988. **37**(2): p. 785-789.
133. Lee, J.C.; Cannone, J.J.; Gutell, R.R., *The lonepair triloop: A new motif in RNA structure.* J. Mol. Biol., 2003. **325**(1): p. 65-83.
134. Lee, J.C.; Gutell, R.R., *Diversity of base-pair conformations and their occurrence in rRNA structure and RNA structural motifs.* J. Mol. Biol., 2004. **344**(5): p. 1225-1249.
135. Lee, J.C.; Gutell, R.R.; Russell, R., *The UAA/GAN internal loop motif: A new RNA structural element that forms a cross-strand AAA stack and long-range tertiary interactions.* J. Mol. Biol., 2006. **360**(5): p. 978-988.
136. Leiros, H.K.; Brandsdal, B.O.; Andersen, O.A.; Os, V.; Leiros, I.; Helland, R.; Otlewski, J.; Willassen, N.P.; Smalas, A.O., *Trypsin specificity as elucidated by LIE calculations, X-ray structures, and association constant measurements.* Protein Sci., 2004. **13**(4): p. 1056-70.
137. Leontis, N.B.; Stombaugh, J.; Westhof, E., *The non-Watson-Crick base pairs and their associated isostericity matrices.* Nucleic Acids Res., 2002. **30**(16): p. 3497-3531.
138. Levitt, M., *Detailed Molecular Model For Transfer Ribonucleic Acid.* Nature, 1969. **224**(5221): p. 759.
139. Levy, Y.; Jortner, J.; Becker, O.M., *Solvent effects on the energy landscapes and folding kinetics of polyalanine.* Proc. Natl. Acad. Sci. U. S. A., 2001. **98**(5): p. 2188-2193.
140. Liang, X.G.; Kuhn, H.; Frank-Kamenetskii, M.D., *Monitoring single-stranded DNA secondary structure formation by determining the topological state of DNA catenanes.* Biophys. J., 2006. **90**(8): p. 2877-2889.
141. Lightstone, F.C.; Schwegler, E.; Allesch, M.; Gygi, F.; Galli, G., *A first-principles molecular dynamics study of calcium in water.* ChemPhysChem, 2005. **6**(9): p. 1745-1749.
142. Lipscomb, W.N.; Strater, N., *Recent advances in zinc enzymology.* Chem. Rev. (Washington, DC, U. S.), 1996. **96**(7): p. 2375-2433.
143. Liwo, A.; Czaplewski, C.; Oldziej, S.; Rojas, A.V.; Kazmierkiewicz, R.; Makowski, M.; Murarka, R.K.; Scheraga, H.A., *Simulation of protein structure and dynamics with the coarse-grained UNRES force field* in *Coarse-Graining of*

Condensed Phase and Biomolecular Systems, G. Voth, Editor. 2008, CRC Press, Taylor & Francis Group: Farmington, CT. p. 107-122.
144. Liwo, A.; Czaplewski, C.; Pillardy, J.; Scheraga, H.A., *Cumulant-based expressions for the multibody terms for the correlation between local and electrostatic interactions in the united-residue force field.* J. Chem. Phys., 2001. **115**(5): p. 2323-2347.
145. Liwo, A.; Khalili, M.; Czaplewski, C.; Kalinowski, S.; Oldziej, S.; Wachucik, K.; Scheraga, H.A., *Modification and optimization of the united-residue (UNRES) potential energy function for canonical simulations. I. Temperature dependence of the effective energy function and tests of the optimization method with single training proteins.* J. Phys. Chem. B, 2007. **111**(1): p. 260-285.
146. Lopes, P.E.M.; Roux, B.; MacKerell, A.D., *Molecular modeling and dynamics studies with explicit inclusion of electronic polarizability: theory and applications.* Theor. Chem. Acc., 2009. **124**(1-2): p. 11-28.
147. MacKerell, A.D.; Brooks, B.; Brooks, C.L.; Nilsson, L.; Roux, B.; Won, Y.; Karplus, M., *CHARMM: The Energy Function and Its Parameterization*, in *Encyclopedia of Computational Chemistry*. 2002, John Wiley & Sons, Ltd.
148. Maisuradze, G.G.; Liwo, A.; Oldziej, S.; Scheraga, H.A., *Evidence, from simulations, of a single state with residual native structure at the thermal denaturation midpoint of a small globular protein.* J. Am. Chem. Soc., 2010. **132**(27): p. 9444-52.
149. Makowski, M.; Sobolewski, E.; Czaplewski, C.; Oldziej, S.; Liwo, A.; Scheraga, H.A., *Simple physics-based analytical formulas for the potentials of mean force for the interaction of amino acid side chains in water. IV. Pairs of different hydrophobic side chains.* J. Phys. Chem. B, 2008. **112**(36): p. 11385-11395.
150. Marcus, Y., *A Simple Empirical-Model Describing the Thermodynamics of Hydration of Ions of Widely Varying Charges, Sizes, and Shapes.* Biophys. Chem., 1994. **51**(2-3): p. 111-127.
151. Marini, G.W.; Texler, N.R.; Rode, B.M., *Monte Carlo simulations of Zn(II) in water including three-body effects.* J. Phys. Chem., 1996. **100**(16): p. 6808-6813.
152. Mathews, D.H.; Sabina, J.; Zuker, M.; Turner, D.H., *Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure.* J. Mol. Biol., 1999. **288**(5): p. 911-940.
153. Maynard, A.T.; Covell, D.G., *Reactivity of zinc finger cores: Analysis of protein packing and electrostatic screening.* J. Am. Chem. Soc., 2001. **123**(6): p. 1047-1058.
154. McCool, M.A.; Collings, A.F.; Woolf, L.A., *Pressure and temperature dependence of the self-diffusion of benzene.* Journal of the Chemical Society, Faraday Transactions 1, 1972. **68**: p. 1489 - 1497.
155. McQuarrie, D.A., *Statistical Mechanics*. 2000: University Science Books.
156. Michel, F.; Jacquier, A.; Dujon, B., *Comparison of Fungal Mitochondrial Introns Reveals Extensive Homologies in Rna Secondary Structure.* Biochimie, 1982. **64**(10): p. 867-881.

157. Michel, F.; Umesono, K.; Ozeki, H., *Comparative and Functional-Anatomy of Group-Ii Catalytic Introns - a Review.* Gene, 1989. **82**(1): p. 5-30.
158. Michel, F.; Westhof, E., *Modeling of the 3-Dimensional Architecture of Group-I Catalytic Introns Based on Comparative Sequence-Analysis.* J. Mol. Biol., 1990. **216**(3): p. 585-610.
159. Mikhail, S.Z.; Kimel, W.R., *Densities and Viscosities of Methanol-Water Mixtures.* J. Chem. Eng. Data, 1961. **6**(4): p. 533 - 537.
160. Miller, T.F.; Eleftheriou, M.; Pattnaik, P.; Ndirango, A.; Newns, D.; Martyna, G.J., *Symplectic quaternion scheme for biophysical molecular dynamics.* J. Chem. Phys., 2002. **116**(20): p. 8649-8659.
161. Mobley, D.L.; Bayly, C.I.; Cooper, M.D.; Shirts, M.R.; Dill, K.A., *Small Molecule Hydration Free Energies in Explicit Solvent: An Extensive Test of Fixed-Charge Atomistic Simulations.* J. Chem. Theory Comput., 2009. **5**(2): p. 350-358.
162. Mohammed, A.M.; Loeffler, H.H.; Inada, Y.; Tanada, K.-i.; Funahashi, S., *Quantum mechanical/molecular mechanical molecular dynamic simulation of zinc(II) ion in water.* J. Mol. Liq., 2005. **119**(1-3): p. 55-62.
163. Morgan, H.L., *The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service.* Journal of Chemical Documentation, 1965. **5**(2): p. 107-113.
164. Moriarty, N.W.; Grosse-Kunstleve, R.W.; Adams, P.D., *electronic Ligand Builder and Optimization Workbench (eLBOW): a tool for ligand coordinate and restraint generation.* Acta Crystallographica Section D-Biological Crystallography, 2009. **65**: p. 1074-1080.
165. Naor, M.M.; Van Nostrand, K.; Dellago, C., *Car-Parrinello molecular dynamics simulation of the calcium ion in liquid water.* Chem. Phys. Lett., 2003. **369**(1-2): p. 159-164.
166. Neely, J.; Connick, R., *Rate of water exchange from hydrated magnesium ion.* J. Am. Chem. Soc., 1970. **92**(11): p. 3476-3478.
167. Nguyen, H.D.; Marchut, A.J.; Hall, C.K., *Solvent effects on the conformational transition of a model polyalanine peptide.* Protein Sci., 2004. **13**(11): p. 2909-2924.
168. Nielsen, S.O.; Lopez, C.F.; Srinivas, G.; Klein, M.L., *Coarse grain models and the computer simulation of soft materials.* Journal Of Physics-condensed Matter, 2004. **16**(15): p. R481-R512.
169. Nissen, P.; Ippolito, J.A.; Ban, N.; Moore, P.B.; Steitz, T.A., *RNA tertiary interactions in the large ribosomal subunit: The A-minor motif.* Proc. Natl. Acad. Sci. U. S. A., 2001. **98**(9): p. 4899-4903.
170. Noller, H.F.; Kop, J.; Wheaton, V.; Brosius, J.; Gutell, R.R.; Kopylov, A.M.; Dohme, F.; Herr, W.; Stahl, D.A.; Gupta, R.; Woese, C.R., *Secondary Structure Model for 23s Ribosomal-Rna.* Nucleic Acids Res., 1981. **9**(22): p. 6167-6189.

171. Noury, S.; Krokidis, X.; Fuster, F.; Silvi, B., *Computational tools for the electron localization function topological analysis.* Comput. Chem., 1999. **23**(6): p. 597-604.
172. Obst, S.; Bradaczek, H., *Molecular dynamics simulations of zinc ions in water using CHARMM.* Journal of Molecular Modeling, 1997. **3**(6): p. 224-232.
173. Ohtaki, H.; Radnai, T., *Structure and dynamics of hydrated ions.* Chem. Rev. (Washington, DC, U. S.), 1993. **93**(3): p. 1157-1204.
174. Oostenbrink, C.; Villa, A.; Mark, A.E.; van Gunsteren, W.F., *A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6.* J. Comput. Chem., 2004. **25**(13): p. 1656-76.
175. Ota, N.; Stroupe, C.; Ferreira-da-Silva, J.M.; Shah, S.A.; Mares-Guia, M.; Brunger, A.T., *Non-Boltzmann thermodynamic integration (NBTI) for macromolecular systems: relative free energy of binding of trypsin to benzamidine and benzylamine.* Proteins, 1999. **37**(4): p. 641-53.
176. Parisien, M.; Major, F., *The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data.* Nature, 2008. **452**(7183): p. 51-55.
177. Penev, E.S.; Lampoudi, S.; Shea, J.E., *TiREX: Replica-exchange molecular dynamics using TINKER.* Comput. Phys. Commun., 2009. **180**(10): p. 2013-2019.
178. Pilme, J.; Piquemal, J.P., *Advancing beyond charge analysis using the electronic localization function: Chemically intuitive distribution of electrostatic moments.* J. Comput. Chem., 2008. **29**(9): p. 1440-1449.
179. Piquemal, J.; Marquez, A.; Parisel, O.; Giessner-Prettre, C., *A CSOV study of the difference between HF and DFT intermolecular interaction energy values: The importance of the charge transfer contribution.* J. Comput. Chem., 2005. **26**(10): p. 1052-1062.
180. Piquemal, J.P.; Perera, L.; Cisneros, G.A.; Ren, P.Y.; Pedersen, L.G.; Darden, T.A., *Towards accurate solvation dynamics of divalent cations in water using the polarizable amoeba force field: From energetics to structure.* Journal of Chemical Physics, 2006. **125**(5): p. 054511.
181. Piquemal, J.P.; Pilme, J.; Parisel, O.; Gerard, H.; Fourre, I.; Berges, J.; Gourlaouen, C.; De La Lande, A.; Van Severen, M.C.; Silvi, B., *What can be learnt on biologically relevant systems from the topological analysis of the electron localization function?* Int. J. Quantum Chem., 2008. **108**(11): p. 1951-1969.
182. Piquemal, J.P.; Williams-Hubbard, B.; Fey, N.; Deeth, R.J.; Gresh, N.; Giessner-Prettre, C., *Inclusion of the ligand field contribution in a polarizable molecular mechanics: SIBFA-LF.* J. Comput. Chem., 2003. **24**(16): p. 1963-1970.
183. Ponder, J., *TINKER: Software Tools for Molecular Design*. 2009: Saint Louis, MO.
184. Ponder, J.W., *TINKER molecular modeling package.* Washington University Medical School, 2010.

185. Ponder, J.W.; Wu, C.J.; Ren, P.Y.; Pande, V.S.; Chodera, J.D.; Schnieders, M.J.; Haque, I.; Mobley, D.L.; Lambrecht, D.S.; DiStasio, R.A.; Head-Gordon, M.; Clark, G.N.I.; Johnson, M.E.; Head-Gordon, T., *Current Status of the AMOEBA Polarizable Force Field.* Journal of Physical Chemistry B, 2010. **114**(8): p. 2549-2564.
186. Qiu, D.; Shenkin, P.S.; Hollinger, F.P.; Still, W.C., *The GB/SA Continuum Model for Solvation. A Fast Analytical Method for the Calculation of Approximate Born Radii.* J. Phys. Chem. A, 1997. **101**(16): p. 3005-3014.
187. Qiu, D.; Shenkin, P.S.; Hollinger, F.P.; Still, W.C., *The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii.* J. Phys. Chem. A, 1997. **101**(16): p. 3005-3014.
188. Rajamani, S.; Ghosh, T.; Garde, S., *Size dependent ion hydration, its asymmetry, and convergence to macroscopic behavior.* J. Chem. Phys., 2004. **120**(9): p. 4457-4466.
189. Rapcewicz, K.; Ashcroft, N.W., *Fluctuation Attraction in Condensed Matter - a Nonlocal Functional-Approach.* Physical Review B, 1991. **44**(8): p. 4032-4035.
190. Rayon, V.M.; Valdes, H.; Diaz, N.; Suarez, D., *Monoligand Zn(II) complexes: Ab initio benchmark calculations and comparison with density functional theory methodologies.* J. Chem. Theory Comput., 2008. **4**(2): p. 243-256.
191. Reinhardt, P.; Piquemal, J.; Savin, A., *Fragment-Localized Kohn-Sham Orbitals via a Singles Configuration-interaction Procedure and Application to Local Properties and Intermolecular Energy Decomposition Analysis.* J. Chem. Theory Comput., 2008. **4**(12): p. 2020-2029.
192. Ren, P.; Ponder, J.W., *Consistent treatment of inter- and intramolecular polarization in molecular mechanics calculations.* J. Comput. Chem., 2002. **23**(16): p. 1497-506.
193. Ren, P.; Ponder, J.W., *Polarizable Atomic Multipole Water Model for Molecular Mechanics Simulation.* Journal of Physical Chemistry B, 2003. **107**: p. 5933-5947.
194. Ren, P.Y.; Ponder, J.W., *Consistent treatment of inter- and intramolecular polarization in molecular mechanics calculations.* Journal of Computational Chemistry, 2002. **23**(16): p. 1497-1506.
195. Ren, P.Y.; Ponder, J.W., *Polarizable atomic multipole water model for molecular mechanics simulation.* J. Phys. Chem. B, 2003. **107**(24): p. 5933-5947.
196. Ren, P.Y.; Ponder, J.W., *Temperature and pressure dependence of the AMOEBA water model.* Journal of Physical Chemistry B, 2004. **108**(35): p. 13427-13437.
197. Richardson, J.S.; Richardson, D.C., *Amino-Acid Preferences for Specific Locations at the Ends of Alpha-Helices.* Science, 1988. **240**(4859): p. 1648-1652.
198. Rick, S.W.; Stuart, S.J.; Berne, B.J., *Dynamical Fluctuating Charge Force-Fields - Application to Liquid Water.* J. Chem. Phys., 1994. **101**(7): p. 6141-6156.
199. Rogers, D.M.; Beck, T.L., *Quasichemical and structural analysis of polarizable anion hydration.* J. Chem. Phys., 2010. **132**(1): p. 014505.

200. Roux, B.; Simonson, T., *Implicit solvent models.* Biophys. Chem., 1999. **78**(1-2): p. 1-20.
201. Roux, C.; Gresh, N.; Perera, L.; Piquemal, J.; Salmon, L., *Binding of 5-phospho-D-arabinonohydroxamate and 5-phospho-D-arabinonate inhibitors to zinc phosphomannose isomerase from Candida albicans studied by polarizable molecular mechanics and quantum mechanics.* J. Comput. Chem., 2007. **28**(5): p. 938-957.
202. Ryde, U., *Combined quantum and molecular mechanics calculations on metalloproteins.* Curr. Opin. Chem. Biol., 2003. **7**(1): p. 136-142.
203. Sagui, C.; Darden, T.A., *Molecular dynamics simulations of biomolecules: Long-range electrostatic effects.* Annu. Rev. Biophys. Biomol. Struct., 1999. **28**: p. 155-179.
204. Sagui, C.; Pedersen, L.G.; Darden, T.A., *Towards an accurate representation of electrostatics in classical force fields: Efficient implementation of multipolar interactions in biomolecular simulations.* J. Chem. Phys., 2004. **120**(1): p. 73-87.
205. Sakakibara, Y.; Brown, M.; Hughey, R.; Mian, I.S.; Sjolander, K.; Underwood, R.C.; Haussler, D., *Stochastic Context-Free Grammars for Transfer-RNA Modeling.* Nucleic Acids Res., 1994. **22**(23): p. 5112-5120.
206. Salmon, P.S.; Bellissentfunel, M.C.; Herdman, G.J., *The Dynamics of Aqueous Zn-2+ Solutions - a Study Using Incoherent Quasi-Elastic Neutron-Scattering.* Journal of Physics-Condensed Matter, 1990. **2**(18): p. 4297-4309.
207. Santalucia, J.; Kierzek, R.; Turner, D.H., *Stabilities Of Consecutive A.C, C.C, G.G, U.C, And U.U Mismatches In RNA Internal Loops - Evidence For Stable Hydrogen-Bonded U.U And C.C+ Pairs.* Biochemistry, 1991. **30**(33): p. 8242-8251.
208. Savin, A.; Nesper, R.; Wengert, S.; Fassler, T.F., *ELF: The electron localization function.* Angewandte Chemie-International Edition in English, 1997. **36**(17): p. 1809-1832.
209. Schaefer, M.; Karplus, M., *A comprehensive analytical treatment of continuum electrostatics.* J. Phys. Chem., 1996. **100**(5): p. 1578-1599.
210. Schmid, R.; Miah, A.M.; Sapunov, V.N., *A new table of the thermodynamic quantities of ionic hydration: values and some applications (enthalpy-entropy compensation and Born radii).* Phys. Chem. Chem. Phys., 2000. **2**(1): p. 97-102.
211. Schnieders, M.J.; Fenn, T.D.; Pande, V.S., *Polarizable Atomic Multipole X-Ray Refinement: Particle Mesh Ewald Electrostatics for Macromolecular Crystals.* J. Chem. Theory Comput., 2011. **7**(4): p. 1141-1156.
212. Schnieders, M.J.; Ponder, J.W., *Polarizable Atomic Multipole Solutes in a Generalized Kirkwood Continuum.* J. Chem. Theory Comput., 2007. **3**(6): p. 2083-2097.
213. Schnieders, M.J.; Ponder, J.W., *Polarizable atomic multipole solutes in a generalized Kirkwood continuum.* Journal of Chemical Theory and Computation, 2007. **3**(6): p. 2083-2097.

214. Schwarzl, S.M.; Tschopp, T.B.; Smith, J.C.; Fischer, S., *Can the calculation of ligand binding free energies be improved with continuum solvent electrostatics and an ideal-gas entropy correction?* J. Comput. Chem., 2002. **23**(12): p. 1143-9.
215. Senn, H.M.; Thiel, W., *QM/MM Methods for Biomolecular Systems.* Angewandte Chemie-International Edition, 2009. **48**(7): p. 1198-1229.
216. Serra, M.J.; Lyttle, M.H.; Axenson, T.J.; Schadt, C.A.; Turner, D.H., *RNA Hairpin Loop Stability Depends On Closing Base-Pair.* Nucleic Acids Res., 1993. **21**(16): p. 3845-3849.
217. Shelley, J.C.; Shelley, M.Y.; Reeder, R.C.; Bandyopadhyay, S.; Klein, M.L., *A coarse grain model for phospholipid simulations.* J. Phys. Chem. B, 2001. **105**(19): p. 4464-4470.
218. Shen, M.Y.; Sali, A., *Statistical potential for assessment and prediction of protein structures.* Protein Sci., 2006. **15**(11): p. 2507-2524.
219. Shi, Y.; Jiao, D.A.; Schnieders, M.J.; Ren, P.Y., *Trypsin-Ligand Binding Free Energy Calculation with AMOEBA.* Embc: 2009 Annual International Conference of the Ieee Engineering in Medicine and Biology Society, Vols 1-20, 2009: p. 2328-2331.
220. Shi, Y.; Wu, C.J.; Ponder, J.W.; Ren, P.Y., *Multipole Electrostatics in Hydration Free Energy Calculations.* J. Comput. Chem., 2011. **32**(5): p. 967-977.
221. Shirts, M.R.; Bair, E.; Hooker, G.; Pande, V.S., *Equilibrium free energies from nonequilibrium measurements using maximum-likelihood methods.* Phys. Rev. Lett., 2003. **91**(14): p. 140601-1-4.
222. Shirts, M.R.; Bair, E.; Hooker, G.; Pande, V.S., *Equilibrium Free Energies from Nonequilibrium Measurements Using Maximum-Likelihood Methods.* Phys. Rev. Lett., 2003. **91**: p. 140601.
223. Silvi, B.; Savin, A., *Classification of chemical-bonds based on topological analysis of electron localization functions.* Nature, 1994. **371**(6499): p. 683-686.
224. Singh, J.; Thornton, J.M., *Atlas of Protein Side-Chain Interactions*. Vol. I & II. 1992, Oxford: IRL press.
225. Singh, R.K.; Tropsha, A.; Vaisman, II, *Delaunay tessellation of proteins: four body nearest-neighbor propensities of amino acid residues.* J. Comput. Biol., 1996. **3**(2): p. 213-21.
226. Sinnokrot, M.O.; Sherrill, C.D., *Highly accurate coupled cluster potential energy curves for the benzene dimer: Sandwich, T-shaped, and parallel-displaced configurations.* J. Phys. Chem. A, 2004. **108**(46): p. 10200-10207.
227. Smith, M.W.; Meskauskas, A.; Wang, P.; Sergiev, P.V.; Dinman, J.D., *Saturation Mutagenesis of 5S rRNA in Saccharomyces cerevisiae.* Mol. Cell. Biol., 2001. **21**(24): p. 8264-8275.
228. Soto, P.; Baumketner, A.; Shea, J.E., *Aggregation of polyalanine in a hydrophobic environment.* J. Chem. Phys., 2006. **124**(13): p. 134904.
229. Stevens, W.J.; Basch, H.; Krauss, M., *Compact Effective Potentials and Efficient Shared-Exponent Basis-Sets for the 1st-Row and 2nd-Row Atoms.* J. Chem. Phys., 1984. **81**(12): p. 6026-6033.

230. Stevens, W.J.; Fink, W.H., *Frozen fragment reduced variational space analysis of hydrogen-bonding interactions - Application to the water dimer.* Chem. Phys. Lett., 1987. **139**(1): p. 15-22.
231. Stiegler, P.; Carbon, P.; Zuker, M.; Ebel, J.P.; Ehresmann, C., *Secondary Structure And Topography Of 16S-Ribosomal RNA From Escherichia-Coli.* Comptes Rendus Hebdomadaires Des Seances De L Academie Des Sciences Serie D, 1980. **291**(12): p. 937-940.
232. Still, W.C.; Tempczyk, A.; Hawley, R.C.; Hendrickson, T., *Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics.* J. Am. Chem. Soc., 1990. **112**(16): p. 6127-6129.
233. Stone, A.J., *The Theory of Intermolecular Forces*. 1997, USA: Oxford University Press.
234. Stone, A.J., *Distributed multipole analysis: Stability for large basis sets.* J. Chem. Theory Comput., 2005. **1**(6): p. 1128-1132.
235. Stone, A.J.; Alderton, M., *Distributed Multipole Analysis - Methods and Applications.* Mol. Phys., 1985. **56**(5): p. 1047-1064.
236. Stone, J.E.; Gohara, D.; Shi, G., *OpenCL: A Parallel Programming Standard for Heterogeneous Computing Systems.* Comput Sci Eng, 2010. **12**(3): p. 66-72.
237. Sykes, M.T.; Levitt, M., *Describing RNA structure by libraries of clustered nucleotide doublets.* J. Mol. Biol., 2005. **351**(1): p. 26-38.
238. Talhout, R.; Engberts, J.B., *Thermodynamic analysis of binding of p-substituted benzamidines to trypsin.* Eur. J. Biochem., 2001. **268**(6): p. 1554-60.
239. Tanaka, S.; Scheraga, H.A., *Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins.* Macromolecules, 1976. **9**(6): p. 945-50.
240. Thole, B.T., *Molecular Polarizabilities Calculated with a Modified Dipole Interaction.* Chem. Phys., 1981. **59**(3): p. 341-350.
241. Thornton, J.M. 2009; Available from: http://www.ebi.ac.uk/thornton-srv/databases/sidechains/.
242. Tidor, B.; Karplus, M., *The contribution of vibrational entropy to molecular association. The dimerization of insulin.* J. Mol. Biol., 1994. **238**(3): p. 405-14.
243. Tiraboschi, G.; Gresh, N.; Giessner-Prettre, C.; Pedersen, L.G.; Deerfield, D.W., *Parallel ab initio and molecular mechanics investigation of polycoordinated Zn(II) complexes with model hard and soft ligands: Variations of binding energy and of its components with number and charges of ligands.* J. Comput. Chem., 2000. **21**(12): p. 1011-1039.
244. Tissandier, M.D.; Cowen, K.A.; Feng, W.Y.; Gundlach, E.; Cohen, M.H.; Earhart, A.D.; Tuttle, T.R.; Coe, J.V., *The proton's absolute aqueous enthalpy and Gibbs free energy of solvation from cluster ion solvation data (vol 102A, pg 7791, 1998).* J. Phys. Chem. A, 1998. **102**(46): p. 9308-9308.
245. Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M., *Ab initio calculations of structures and interaction energies of toluene dimers including CCSD(T) level electron correlation correction.* J. Chem. Phys., 2005. **122**(14): p. 144323.

246. Tuccinardi, T.; Martinelli, A.; Nuti, E.; Carelli, P.; Balzano, F.; Uccello-Barretta, G.; Murphy, G.; Rossello, A., *Amber force field implementation, molecular modelling study, synthesis and MMP-1/MMP-2 inhibition profile of (R)and (S)-N-hydroxy-2-(N-isopropoxybiphenyl-4-ylsulfonamido)-3-methylbutanamides.* Bioorg. Med. Chem., 2006. **14**(12): p. 4260-4276.
247. Tuerk, C.; Gauss, P.; Thermes, C.; Groebe, D.R.; Gayle, M.; Guild, N.; Stormo, G.; Daubentoncarafa, Y.; Uhlenbeck, O.C.; Tinoco, I.; Brody, E.N.; Gold, L., *CUUCGG Hairpins - Extraordinarily Stable RNA Secondary Structures Associated With Various Biochemical Processes.* Proc. Natl. Acad. Sci. U. S. A., 1988. **85**(5): p. 1364-1368.
248. Vallurupalli, P.; Moore, P.B., *The solution structure of the loop E region of the 5 S rRNA from spinach chloroplasts.* J. Mol. Biol., 2003. **325**(5): p. 843-856.
249. Voth, G., ed. *Coarse-Graining of Condensed Phase and Biomolecular Systems*. 2008, CRC Press, Taylor & Francis Group: Farmington, CT.
250. Wagoner, J.A.; Baker, N.A., *Assessing implicit models for nonpolar mean solvation forces: the importance of dispersion and volume terms.* Proc. Natl. Acad. Sci. U. S. A., 2006. **103**(22): p. 8331-6.
251. Walter, A.E.; Wu, M.; Turner, D.H., *The Stability and Structure of Tandem Ga Mismatches in Rna Depend on Closing Base-Pairs.* Biochemistry, 1994. **33**(37): p. 11349-11354.
252. Wang, J.M.; Wang, W.; Kollman, P.A.; Case, D.A., *Automatic atom type and bond type perception in molecular mechanical calculations.* J. Mol. Graph. Model., 2006. **25**(2): p. 247-260.
253. Wang, J.M.; Wolf, R.M.; Caldwell, J.W.; Kollman, P.A.; Case, D.A., *Development and testing of a general amber force field.* J. Comput. Chem., 2004. **25**(9): p. 1157-1174.
254. Warshel, A.; Levitt, M., *Theoretical studies of enzymic reactions - dielectric, electrostatic and steric stabilization of carbonium-ion in reaction of lysozyme.* J. Mol. Biol., 1976. **103**(2): p. 227-249.
255. Weast, R.C., *CRC handbook of chemistry and physics*. 1st Student ed. 1988, Boca Raton, FL: CRC Press. 1 v. (various pagings).
256. Weininger, D., *Smiles, a Chemical Language and Information-System .1. Introduction to Methodology and Encoding Rules.* J. Chem. Inf. Comput. Sci., 1988. **28**(1): p. 31-36.
257. Wensink, E.J.W.; Hoffmann, A.C.; van Maaren, P.J.; van der Spoel, D., *Dynamic properties of water/alcohol mixtures studied by computer simulation.* J. Chem. Phys., 2003. **119**(14): p. 7308-7317.
258. Wesenberg, J.H.; Ardavan, A.; Briggs, G.A.; Morton, J.J.; Schoelkopf, R.J.; Schuster, D.I.; Molmer, K., *Quantum computing with an electron spin ensemble.* Phys. Rev. Lett., 2009. **103**(7): p. 070502.
259. Wilson, E.B.; Decius, J.C.; Cross, P.C., *Molecular Vibrations: TheTheory of Infrared and Raman Vibrational Spectra*. 1955, New York: McGraw-Hill.

260. Wimberly, B.; Varani, G.; Tinoco, I., *The Conformation Of Loop-E Of Eukaryotic 5S-Ribosomal RNA.* Biochemistry, 1993. **32**(4): p. 1078-1087.
261. Woese, C.R.; Gutell, R.; Gupta, R.; Noller, H.F., *Detailed Analysis of the Higher-Order Structure of 16s-Like Ribosomal Ribonucleic-Acids.* Microbiological Reviews, 1983. **47**(4): p. 621.
262. Woese, C.R.; Magrum, L.J.; Gupta, R.; Siegel, R.B.; Stahl, D.A.; Kop, J.; Crawford, N.; Brosius, J.; Gutell, R.; Hogan, J.J.; Noller, H.F., *Secondary Structure Model For Bacterial 16S Ribosomal-RNA - Phylogenetic, Enzymatic And Chemical Evidence.* Nucleic Acids Res., 1980. **8**(10): p. 2275-2293.
263. Woese, C.R.; Winker, S.; Gutell, R.R., *Architecture of Ribosomal-Rna - Constraints on the Sequence of Tetra-Loops.* Proc. Natl. Acad. Sci. U. S. A., 1990. **87**(21): p. 8467-8471.
264. Woodcock, H.L.; Miller, B.T.; Hodoscek, M.; Okur, A.; Larkin, J.D.; Ponder, J.W.; Brooks, B.R., *MSCALE: A General Utility for Multiscale Modeling.* J. Chem. Theory Comput., 2011. **7**(4): p. 1208-1219.
265. Wu, C.; Shea, J.E., *Coarse-grained models for protein aggregation.* Curr. Opin. Struct. Biol., 2005. **21**(2): p. 209-220.
266. Wu, J.C.; Chattree, G.; Ren, P., *Automation of AMOEBA polarizable force field parameterization for small molecules.* Theoretical Chemistry Accounts, In press.
267. Wu, J.C.; Gardner, D.P.; Ozer, S.; Gutell, R.R.; Ren, P.Y., *Correlation of RNA Secondary Structure Statistics with Thermodynamic Stability and Applications to Folding.* J. Mol. Biol., 2009. **391**(4): p. 769-783.
268. Wu, J.C.; Piquemal, J.P.; Chaudret, R.; Reinhardt, P.; Ren, P., *Polarizable molecular dynamics simulation of Zn(II) in water using the AMOEBA force field.* J. Chem. Theory Comput., 2010. **6**(7): p. 2059-2070.
269. Xia, T.B.; SantaLucia, J.; Burkard, M.E.; Kierzek, R.; Schroeder, S.J.; Jiao, X.Q.; Cox, C.; Turner, D.H., *Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs.* Biochemistry, 1998. **37**(42): p. 14719-14735.
270. Xia, Z.; Gardner, D.P.; Gutell, R.R.; Ren, P., *Coarse-grained model for simulation of RNA three-dimensional structures.* J. Phys. Chem. B, 2010. **114**(42): p. 13497-506.
271. Xie, W.; Orozco, M.; Truhlar, D.G.; Gao, J., *X-Pol Potential: An Electronic Structure-Based Force Field for Molecular Dynamics Simulation of a Solvated Protein in Water.* J. Chem. Theory Comput., 2009. **5**(3): p. 459-467.
272. Xie, W.S.; Pu, J.Z.; MacKerell, A.D.; Gao, J.L., *Development of a polarizable intermolecular potential function (PIPF) for liquid amides and alkanes.* J. Chem. Theory Comput., 2007. **3**(6): p. 1878-1889.
273. Xin, Y.R.; Olson, W.K., *BPS: a database of RNA base-pair structures.* Nucleic Acids Res., 2009. **37**: p. D83-D88.
274. Yakovchuk, P.; Protozanova, E.; Frank-Kamenetskii, M.D., *Base-stacking and base-pairing contributions into thermal stability of the DNA double helix.* Nucleic Acids Res., 2006. **34**(2): p. 564-574.

275. Yongyai, Y.P.; Kokpol, S.; Rode, B.M., *Zinc Ion in Water - Intermolecular Potential with Approximate 3-Body Correction and Monte-Carlo Simulation.* Chem. Phys., 1991. **156**(3): p. 403-412.
276. Yuan-Ping, P., *Successful molecular dynamics simulation of two zinc complexes bridged by a hydroxide in phosphotriesterase using the cationic dummy atom method.* Proteins: Struct., Funct., Genet., 2001. **45**(3): p. 183-189.
277. Zhao, Z.; Rogers, D.M.; Beck, T.L., *Polarization and charge transfer in the hydration of chloride ions.* J. Chem. Phys., 2010. **132**(1): p. 014502.
278. Zheng, L.; Chen, M.; Yang, W., *Random walk in orthogonal space to achieve efficient free-energy simulation of complex systems.* Proc. Natl. Acad. Sci. U. S. A., 2008. **105**(51): p. 20227-32.
279. Zhou, J.; Thorpe, I.F.; Izvekov, S.; Voth, G.A., *Coarse-grained peptide modeling using a systematic multiscale approach.* Biophys. J., 2007. **92**(12): p. 4289-4303.
280. Zuker, M.; Jaeger, J.A.; Turner, D.H., *A comparison of optimal and suboptimal RNA secondary structures predicted by free energy minimization with structures determined by phylogenetic comparison.* Nucleic Acids Res., 1991. **19**(10): p. 2707-14.
281. Zwieb, C.; Glotz, C.; Brimacombe, R., *Secondary Structure Comparisons Between Small Subunit Ribosomal-RNA Molecules From 6 Different Species.* Nucleic Acids Res., 1981. **9**(15): p. 3621-3640.