The Dissertation Committee for Zhen Xia Certifies that this is the approved version of the following dissertation:

Modeling the Structure, Dynamics, and Interaction of Biological Molecules

Committee:

Pengyu Ren, Supervisor

Robin Gutell

Rick Russell

Aaron Baker

Lydia Contreras

Modeling the Structure, Dynamics, and Interactions of Biological Molecules

by

Zhen Xia, B.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

May 2013

Dedication

To my family

Acknowledgements

First and foremost, I would like to thank my advisor, Dr. Pengyu Ren for his mentorship and for unselfishly sharing his knowledge and experiences with me. The depth of his patience and wisdom in life is unfathomable. He certainly taught me much more than science. I owe my sincere thanks to my willingly served dissertation committee members, Dr. Robin Gutell, Dr. Rick Russell, Dr. Aaron Baker, and Dr. Lydia Contreras for their guidance in all my research endeavors. I would like to thank my intern manager Dr. Ruhong Zhou. It has been an exciting journey to be a part of his research group at IBM T. J. Watson Research Center.

I am grateful to former and current members of the Ren's group, Dr. Michael Schnieders, Dr. Jiajing Zhang, Dr. Johnny Wu, Dr. Chunli Yan, Dr. Qiantao Wang, Yue Shi, Xiaojia Mu, Danny Dykstra, and David Bell for their assistance and advices. Not only did I earn esteemed colleagues but also good friends. I would also like to express my gratitude to my collaborators, Dr. Eugene Shakhnovich, Dr. Isidore Rigoutsos, Dr. David Gardner, Dr. Payel Das, Tien Huynh, and Dr. Seung-gu Kang. I really have been fortunate to have bright minds with expert perspectives offer to work with me and make substantial contributions.

Besides, I would like to acknowledge all the faculties and staffs in the Department of Biomedical Engineering for help. Lastly, I would like to say thanks to my family and friends back in China who always has trust and faith in me.

Modeling the Structure, Dynamics, and Interactions of Biological Molecules

Zhen Xia, Ph.D.

The University of Texas at Austin, 2013

Supervisor: Pengyu Ren

Biological molecules are essential parts of organisms and participate in a variety of biological processes within cells. Understanding the relationship between sequence, structure, and function of biological molecules are of fundamental importance in life science and the health care industry. In this dissertation, a multi-scale approach was utilized to develop coarse-grained molecular models for protein and RNA simulations. By simplifying the atomistic representation of a biomolecular system, the coarse-grained approach enables the molecular dynamics simulations to reveal the biological processes, which occur on the time and length scales that are inaccessible to the all-atom models. For RNA, an "intermediate" coarse-grained model was proposed to provide both accuracy and efficiency for RNA 3D structure modeling and prediction. The overall potential parameters were derived based on structural statistics sampled from experimental structures. For protein, a general, transferable coarse-grain framework based on the Gay-Berne potential and electrostatic point multipole expansion was developed for polypeptide simulations. Next, an advanced atomistic model was developed to model electrostatic interaction with high resolution and incorporates

electronic polarization effect that is ignored in conventional atomistic models. The last part of my thesis work involves applying all-atom molecular simulations to address important questions and problems in biophysics and structural biology. For example, the interaction between protein and miRNA, the recognition mechanism of antigen and antibody, and the structure dynamics of protein in mixed denaturants.

Table of Contents

List of Tables

List of Figures

1 Introduction

Biological molecules (biomolecules), such as proteins and nucleic acids, are essential parts of organisms and participate in most of biological process within cells. One of the factors is that their biological functions highly rely on their structures and dynamics. A traditional way to determine biomolecular structures is using experimental methods, such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and cryo-electron microscopy, which have been successfully applied to determine a broad range of protein and nucleic acid structures. However, determination of three dimensional (3-D) structures by experimental methods is time consuming and expensive. With an increase in the computational power, an alternative way is using computational approaches to model their structures and investigate their biological functions.

Several different computational approaches have been developed to model biomolecular 3-D structures. Molecules can be represented in different scales, from quantum mechanical model to mesoscales model. In Quantum mechanical model, the electrons are considered explicitly, which provide the highest accuracy. However, the system size is always restricted in hundreds atoms and the time scale are limited in pico-second level. The atomistic model use molecular mechanics, the lowest level of information is individual atoms. This model has the good balance between the accuracy and the efficiency, which has been wildly used in studying biological systems. Coarse-grained (CG) models are similar to the atomistic models, with even more crude

representations to increase computational efficiency. Coarse-grained models are usually used in modeling large size molecules or in long-time simulations.

In this dissertation, a coarse-grained approach was utilized to develop simplified molecular models for protein and nucleic acid, respectively. By simplifying the atomistic representation of a biomolecular system, coarse-grained approach enables the molecular dynamics simulations to reveal the biological processes, which occur on the time and length scales that are inaccessible to the all-atom models. Next, those developed models are integrated to current all-atom simulations and applied to three different biological systems: 1) the interaction between protein and miRNA, 2) the recognition mechanism of antigen and antibody, and 3) the structure dynamics of protein in mixed denaturants.

1.1 COARSE-GRAINED (CG) MODELING OF NUCLEIC ACID STRUCTURES

Nucleic acids are linear polymers built with a limited number of nucleotides. Each nucleotide can be divided into two parts: the backbone and the sidechain. The backbone is a series of covalently bonded atoms that together form the continuous chain with the phosphate group and the ribose (RNA) or deoxyribose (DNA) sugar, which are also the common block for each ribonucleotide. The sidechain is built with one type of nucleobase, mainly from adenine (A), cytosine (C), guanine (G), or uracil (U) for RNA molecules.

Coarse-graining approach of the biomolecular system is not unique. It depends on the developers' purposes or the biological phenomena of interest [1]. Some coarse-grained models focus on the backbone structure of nucleic acids, using one or two beads per nucleotide to represent the backbone conformations. Other coarse-grained models

utilize three or more beads to represent each nucleotide, so that the nucleobase (sidechain) information is included in the structural prediction.

There has been a history of modeling DNA at mesoscale [2-18]. Recently studies such as Knotts IV and co-workers successfully predicted several important DNA behaviors, like salt-dependent melting, bubble formation and rehybridization, with a coarse-grained model that uses three interaction sites for phosphate, sugar, and base, respectively.[4] The idea of coarse-grained RNA models can be traced back to 90s, when Harvey and co-workers used a one-bead "molecular mechanics" model to refine large RNA structures with limited low-resolution structural data [19-21]. The model utilized molecular mechanics tools with similar energy function forms as in all-atom potentials, but also incorporated experimental data as restraining factors integrated into the potential energy function. The model was successfully applied to refine the 3-D structure of tRNA, 16S and 23S ribosome RNAs. The one-bead coarse-grained method was then integrated into a molecular simulation program called YUP [22], in which one pseudo-atom is placed at the center of phosphate atom per nucleotide. Later in 2004, McCammon and co-workers combined a similar one-bead RNA coarse-grained model with Monte Carlo (MC) simulations to investigate the distribution of viral RNA inside the capsid of cowpea chlorotic mottle virus [23], where the energy function in the coarse-grained model is described by simple electrostatic potentials of each RNA molecular sphere.

The reduced two-bead virtual bond model was first developed by Cao and Chen in 2006, where two pseudo-atoms are used to represent each nucleotide, one bead standing for the phosphate group and the other for the ribose sugar [24]. Comparing to one-bead per nucleotide model, using two-bead can provide more degrees of freedom to

each nucleotide. Two dihedral angles are available here to describe more complicated backbone conformations, which is somewhat similar to the φ and ψ dihedral angles of the protein backbone. The model treats the helix and loop region separately: the helix parameters were derived from available atomic structures determined by experiments, while the loop conformation was modeled using self-avoiding walks in a diamond lattice. This model is different from some simplified lattice-based models, in which RNA folding energy landscapes and folding thermodynamics properties can be achieved by the statistical mechanical theory. The model is able to give better predictions for simple RNA secondary structures and certain thermodynamics properties such as melting curves. Another advantage of this model is the ability to fold/unfold RNAs at coarse-grained level. The application of the model to the P5abc region of *Tetrahymena* group I ribozyme reveals non-native conformations in the RNA folding process, in which the folding ability can be altered by several important mutations.

Recently, Altman and co-workers developed a nucleic acid simulation tool (NAST) to predict RNA 3-D structure. In this method, each nucleotide is represented as one single pseudo-atom at the center of C3' atom in the ribose sugar [25]. The prediction process is based on the coarse-grained molecular dynamic simulation with a knowledge-based statistical potential function. The parameters for bonds, angles, and torsions were derived from Boltzmann inversion of current available ribosome RNA crystal structures. The non-bonded interaction between each pseudo-atom is represented with a repulsive term from Lennard-Jones potential, to give rise to excluded volume. The last term in the eq. is a distance restraining function for tertiary contacts. In NAST, the RNA secondary-structure is required as the input and some known tertiary contacts are added to improve

the prediction accuracy. The biggest strength of NAST is the ability to model large-sized RNA molecules (>100 nt), which is still a primary limitation for many other coarse-grained models. Obviously, the prediction accuracy largely relies on the correctness of input secondary structure and extra tertiary contacts information. NAST is designed to provide 3-D structure models in conjunction with experiments. The final ranking of the predicted structures can be based on the ideal small-angle X-ray scattering (SAXS) data or experimental solvent accessibility data rather than the NAST energy.

In order to capture the nucleobase conformation, in some coarse-grained model, nucleobase is considered explicitly. Das and Baker developed a fragment assembly of RNA (FARNA) program that allows predicting RNA 3-D structures directly from its primary sequence [26]. The main idea is borrowed from the *Rosetta* low-resolution protein structure prediction method, which was developed by the same group. In FARNA, each nucleobase is represented as a single bead at the geometric center of the base. The backbone conformations were built from known ribosome RNA structure with 3-nt fragments, including the backbone dihedral angles and the conformation of sugar puckering. The 3-nt fragments then are assembled to near-native 3-D structures using Monte Carol simulations. A knowledge-based potential energy, which takes into account backbone conformations and base interaction preferences, is derived from the statistical analysis of experimentally determined RNA structures. Several special terms have been implemented into the energy function, including the radius of gyration, penalty for steric clashes, and terms favoring base stacking and the planarity of both canonical and non-canonical base pairs. Because FARNA is a *de novo* approach, one big advantage of FARNA is that little extra information is needed as the input except the primary

sequence, which made FRARNA very suitable to predict RNA 3-D structures that have very limited secondary structure information, experimental data, or phylogenetic information. After the benchmark test of 20 small-sized RNAs (~30nt), FARNA method reproduces more than 90% of Watson-Crick base pairs and one-third of non-Watson-Crick base pairs ("sheared" base pairs, base triplets, and pseudoknots).

Some RNA coarse-grained models describe both the backbone and nucleobase explicitly, because the base conformation is thought to be at least equally important in determining RNA structures. Structures of most of RNA motifs are actually determined by their base pairing and base stacking conformations. To capture the structural contents of both backbone and nucleobase, three or more beads per nucleotide (pseudo-atoms) are required in the coarse-grained model. The increased number of beads in the backbone-nucleobase hybrid model could substantially enrich the structural details, which will also greatly facilitate the conversion from coarse-grained model to all-atom structures.

Dokholyan's group has developed a Web-based tool, iFoldRNA, to predict RNA 3-D structures [27]. The model uses three pseudo-atoms to represent each nucleotide's phosphate group, sugar ring and the base, respectively. A stepwise potential function is implemented for bonds, angles, and dihedrals, which accounts for base stacking, short-range phosphate–phosphate repulsion, and hydrophobic interactions. The program uses the discrete molecular dynamics (DMD) and the tailored force fields to predict RNA folding dynamics. RNA secondary structure information is not required as input. The replica exchange molecular dynamics (REMD) [28] are implemented to enhance the structure sampling, where multiple simulations or replicas are running simultaneously performed, at different (low to high) temperatures.

Hamelryck and co-workers developed a probabilistic model named BARNACLE to improve the conformational sampling of RNA structures [29]. Seven representative rotatable bonds, six in backbone and one connecting the sugar and base ring, are picked for conformational sampling and calculating the local dependencies between them. The method is based on Bayesian network model using circular distributions and maximum likelihood, where the structural sampling is performed in a continuous space with associated probabilities. This algorithm allows less biased sampling comparing to fragment assembly methods. The model can successfully predict small size RNAs with reasonable native-like structures but lack of handling middle to large size RNAs with long-range contacts.

Another coarse-grained model, called HiRE-RNA, has been developed with similar ideas by Pasquali and Derreumaux in 2010. In this model, an even greater number of pseudo-atoms are used per nucleotide than the five-bead model [30]. Four pseudo-atoms are adopted to give more rotatable bonds for the RNA backbone, which means more conformational flexibility is introduced. Because only one or two bead(s) are used to represent the base, some special non-bonded energy terms such as hydrogen bonding are introduced in the potential function. The model is still under improvement where the parameters and the functional form of the force field are being further optimized.

1.2 Coarse-grained Potentials for Protein Structure and Dynamics

Numerous coarse-grained models have been devised for proteins. The elastic network model (ENM)[31-33] was firstly applied to explore protein dynamics structures based on the seminal work of Flory on polymer networks [34]. It uses a normal mode analysis and simplified harmonic potentials[35], and has been proved to be useful in

analyzing the slow motions of a protein. ENM can be used to calculate vibrational entropy efficiently and accurately [36]. Moreover, ENM can be combined with MARTINI coarse-grained force field to study protein dynamics, and interactions between protein and lipid bilayer [37, 38], which is termed as ELNEDIN[39]. In the ELNEDIN approach, the secondary and tertiary structures of a protein are modeled in the form of elastic network model. A harmonic restraint with a force constant is applied to all the backbone particles within a distance of each other. Nevertheless, both ENM and MARTINI models are only suitable for near native dynamics. They fail to explore the protein folding or predict structures.

Go's models have been quite useful in studying the protein folding pathway, as they employ structure-based approaches [40]. In the native state, a protein structure is composed of backbone atoms with secondary structures (helix, beta strand or coil), and highly well-packed side chains in its core. The non-native protein structure is described with low-level secondary structure packing, which is associated with flexible side chains. Due to the low-level of side-chain packing in the non-native state but high-level packing in the native state, the coarse-grained model can describe a large fluctuation of an unfolded protein. But, if the side-chain information (e.g. shape and specific interaction) is largely lost in coarse-graining, it becomes difficult to capture the native structure of proteins. In Go models, energy terms are constructed in favor of the native contacts, whereas the non-native contacts are termed as less favorable or repellent. In fact, non-native interactions can only contribute to the local structural perturbations or stabilize the protein-folding transition states along the protein-folding pathway. Its transition states are primarily determined by native interactions [41]. Onuchic et al. [42] introduced the off-

lattice Go model, wherein the protein fluctuations near the native state are considered as quasi-harmonic. These approached the funnel-like energy landscape [43] of the protein folding process. Takada et al devised CafeMol, a coarse-grained simulator adopting this off-lattice Go model, with the purpose of developing and simulating proteins [44]. Nevertheless, since the native information of a protein is usually required for structure-based coarse-grained models, these approaches cannot be used to study the protein-folding pathway,, which mean they cannot be used for predicting proteins structure.

Tanaka & Scheraga [28] first proposed the knowledge-based statistical potentials [45] for predicting the structure of proteins. They can be derived from the statistical distribution of native structures in the protein data bank (PDB) through Boltzmann-based methods. Miyazawa & Jernigan [46] followed up the work of Tanaka & Scheraga. They introduced the effect of solvent into the potential. Godzik and Skolnick[47] defined the statistical potentials by adding the residue triplet's term. In addition, the dihedral angle term and other statistical terms, such as solvent accessibility and hydrogen-bonding have been added to the knowledge-based statistical potential [48, 49]. A log-linear model was developed by Bryant and Lawrence [50]. According to this model, it is inappropriate to derive the empirical potential from simple summation over the distribution of residues in the proteins. Rather, protein structures need to be analyzed specifically and separately. These knowledge-based approaches have been employed in simulations of protein folding and protein structure prediction. They had considerable successes [51-53] over the past decade. According to the knowledge-based approaches, interactions can be assumed to be pairwise, and the statistical distributions can be considered Boltzmann-related.

Following the seminal work of Warshel and Levitt [54], a variety of physics-based coarse-grained approaches have been proposed. They have proved to be useful for both the protein folding pathway and protein dynamics near the native state. In fact, they can also be used for predicting protein structure. Among them, one-bead coarse-grained model, developed by Tozzini, V and McCammon, J.A [55], has been employed successfully to study the flap opening of HIV-1 protease. More complex coarse-grained models, such as the UNRES model [56-58], have been developed to predict the protein structures. In UNRES, a polypeptide chain is composed of a sequence of alpha carbons with specific united side-chains and united peptide groups. These coarse-grained particles are connected through virtual bonds. To parameterize the various energy terms, both quantum mechanical ab initio methods and all-atom molecular dynamics simulations were used.

2 Coarse-Grained Model for Simulation of RNA Three-Dimensional Structures

2.1 INTRODUCTION

The importance of RNA has been appreciated since the central dogma was proposed in 1958 [59, 60]. Three RNA molecules, messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA) are associated with the cell's transcription of its DNA into RNA and then translated into proteins[61]. While experiments as early as 1971 suggested that RNA is involved in catalysis during protein synthesis,[62-64] experiments starting in 1982 confirmed that RNA is directly involved in catalysis, in many different RNA systems with difference chemical reactions.[62, 65-73] These RNAs form complex three-dimensional structures.[74-91]

Within the past few years, a large increase in the number of RNAs that form higher-structure, associated with numerous functions in the cell is the foundation for a major paradigm shift in the molecular biology of the cell; RNAs that do not code for proteins are directly associated with the regulation and overall function of the cell [92-102], including different cancers.[103-105] Since more than 90% of the human genome is transcribed into RNAs that do not code for proteins,[106, 107] and the function of an increasing amount of this RNA is now being determined, the prediction of an RNAs higher-order structure and its dynamics will provide great insight into the RNAs contribution to the structure and function of a cell. Towards that end, computational approaches such as molecular modeling have made significant contribution to the understanding of three-dimensional structures and chemical principles of RNA.[19, 20, 108-116] The most successful approaches for protein structure prediction so far have

been based on comparative analysis or reduced models derived from known structures.[117-119]

In recent years, increased effort has been devoted to RNA structure prediction as more and more RNA structures have been determined experimentally. A range of models have been developed for nucleic acids, from fully atomistic models to reduced representations.[3, 120-122] For example, a knowledge-based atomic energy function has been introduced to predict RNA tertiary structures in the FARNA package.[26] Nucleotide cyclic motifs are used in MC-Fold and MC-Sym model to build RNA structure from sequence data.[123] These two models seem successful in predicting the tertiary structure of small RNA molecules. In addition, physics-based atomic force fields such as AMBER[124-127] and CHARMM[128-130] describe the dynamic atomic interaction following traditional molecular mechanics, with parameters derived by fitting to ab initio quantum mechanics calculations and experimental data. It is now feasible with supercomputers to simulate dynamic biological systems as large as an entire virus in atomic detail.[131] However, typical applications of the atomistic force fields are usually limited to small oligomers of nucleic acids or routine simulation times on the order of a few nanoseconds.[4] On the other hand, coarse-grained (CG) methods reduce the number of particles and eliminate high-frequency motions in the system. A CG model enlarges the time step in molecular dynamics simulations while also enhancing intrinsically faster dynamics.[132-134] Several CG approaches, either knowledge or physics based, have been utilized to study the structures of nucleic acids.[2, 4-18, 22, 23, 26, 123, 135-140]

In this chapter, we present an "intermediate" coarse-grained potential for modeling RNA 3-D structure using molecular dynamics. Previous CG RNA models typically used one[19, 23, 141] or two[142] particles for each nucleotide. To optimize the efficiency and accuracy, we developed a model that represents each nucleotide with five pseudo atoms; two of these represent the backbone – one for the sugar and the other for the phosphate, while three pseudo atoms represent the stacking and base pairing for each base. The analytical potential energy functional forms, parameterization with 3-D structural statistics are obtained from experimental structures and initial validate using molecular dynamics simulations of selected RNAs. The model explicitly describes the physical interactions including the electrostatics, hydrogen-bonding and environmental effects. The developed CG model is aimed to predict large-size RNAs with complex tertiary structures. With this CG potential, we are able to predict the 3D structures of small RNAs by *ab initio* folding and capture the tertiary structures of large-size RNA by integrating limited experimental data.

2.2 EXPERIMENTAL METHODS

2.2.1 Data collection and preparation

The CG potential was parameterized using statistics collected from available three dimensional structures of RNA molecules (including both x-ray diffraction structures and nuclear magnetic resonance structures). The RNA structure files were downloaded from The Protein Data Bank (http://www.pdb.org/), Nucleic Acid Database (http://ndbserver.rutgers.edu/index.html), RNA Comparative Analysis Database (rCAD, http://rcat.codeplex.com/, manuscript submitted), and the Comparative RNA Web (CRW) Site[143] (http://www.rna.ccbb.utexas.edu/). Only 668 structure files that

contained more than 5 base pairs and have the resolution records were analyzed for the statistical calculation. All of the coordinates obtained from nuclear magnetic resonance (NMR) structural files were included in the statistical calculations.

2.2.2 Coarse-grained RNA interaction potential

In the CG model, each nucleotide is reduced to five pseudo atoms in RNA (**Figure 2.1**). Two of the five pseudo atoms represent phosphate and sugar respectively, which is the minimum requirement to capture the backbone tertiary structures of RNA.[144] Each base (A, G, C, and T) is represented by three pseudo atoms, connected by three virtual bonds into a triangle. Compared to earlier models with one particle for each base or each residue,[4, 141] the use of three pseudo atoms for each base provides us with better ability to capture the stacking and pairing of bases. As the different bases share some common pseudo atoms, nine unique types of pseudo atoms are needed to represent the four canonical RNA component bases in total (**Figure 2.1**). The improvement in computational efficiency arises from the reduction of number of particles and larger particle mass that enables greater integration time step in molecular dynamics. The topological and physical properties of the pseudo atoms are listed in **Table 2.1**.

The corresponding CG potential energy is calculated by:

$$E_{total} = E_{bonded} + E_{non\text{-}bonded}$$

where E_{bonded} and $E_{non\text{-}bonded}$ are pair-wise bonded and non-bonded energy terms, each representing the sum of contributions of all pairs in the system. The bonded term is further decomposed into:

$$E_{bonded} = E_{bond} + E_{angle} + E_{dihedral}$$

where E_{bond} , E_{angle} and $E_{dihedral}$ are the bond stretching, angle-bending and dihedral energies, respectively.

In classical molecular mechanics, the non-bonded interaction consists of the van der Waals (VDW) and electrostatic contributions. Since our CG model is derived from the 3-D structural statistics of experimental structures, an effective potential is used to represent the potential of mean force of all the non-bonded interactions, including the excluded volume repulsive, the attractive force and the electrostatic force between non-bonded particles, as well as the solvation forces due to the environment. A Buckingham potential is utilized to describe the effective potential.

$$E_{non\text{-}bonded} = E_{effective\text{-}potential}$$

For each term, the parameterization was performed based on the Boltzmann inversion of the corresponding atomistic distribution functions obtained from the experimental structures. The Boltzmann inversion method performs a potential inversion from a set of known distributions of structural parameters to extract effective CG potentials. In our RNA CG system, the potentials calculated from the Boltzmann inversion method[145, 146] need to reproduce the distribution of structural parameters including fourteen different bonds, twenty-five types of angles, twenty-eight dihedral angles, and nineteen intermolecular radial distribution functions extracted from statistical results of all available atomistic RNA structures (please see **Table 2.3** to **Table .25** for more details). All the parameter-fitting works were performed with the software of Matlab Curve Fitting Tool.

The distribution of bond lengths can be represented by the Gaussian function, which is calculated by:

$$P(l) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(b-b_0)^2}{2\sigma^2}} = e^{\frac{-E_{stretch}}{k_B T}}$$

where b, b_0 and σ are the parameters obtained through fitting, k_B is the Boltzmann factor and T is absolute temperature. Taking the logarithm of both sides of equation and dropping the constant term, after performing Boltzmann inversion, we have:

$$E_{bond} = \frac{k_B T}{2\sigma^2}(b-b_0)^2 = K_{bond}(b-b_0)^2$$

where the temperature, T, is set to be 298K.

The distributions of bond angle can be weighted by a factor *sin(θ)* and renormalized by a factor Z_n. The normalized distribution is expressed as:

$$P(\theta) = Z_n p(\theta)/\sin(\theta) = e^{\frac{-E_{bend}}{k_B T}}$$

where θ is the angle between neighboring bonds, while *P(θ)* and *p(θ)* are normalized and un-normalized distribution functions of θ. The distributions of bond angles between CG bonds were also fitted with the Gaussian function, and then the Boltzmann inversion were used to calculate E_{angle}.

$$E_{anlge} = \frac{k_B T}{2\sigma^2}(\theta-\theta_0)^2 = K_a(\theta-\theta_0)^2$$

As in atomic force fields, the tensional energy takes the formula of:

$$E_{dihedral}(\phi) = \sum_{n=1}^{3} K_n[1+\cos(n\phi-\delta_n)]$$

where φ is the dihedral angle, K_n and δ_n *(n = 1, 2, 3)* are force constants and phase angles. The $E_{dihedral}$ were also obtained from performing Boltzmann inversion.

A Buckingham potential,[147] consisting of a 6 term and an exponential term are used to represents the potential of mean forces between a pair of non-bonded atoms, *i* and *j*:

$$E_{non-bonded} = \varepsilon_{ij}[-2.25(\sigma_{ij} / r_{ij})^6 + 1.84 \times 10^5 e^{-12.00\frac{r_{ij}}{\sigma_{ij}}}]$$

where ε_{ij} is the depth of the potential well, σ_{ij} is the radius, and r_{ij} is the distance between a pair of atoms. Note that the above equation is used to describe the potential of mean force even though the symbol "*E*" and the formula are commonly used to represent potential energy. The constants we used here are the same as MM3 force field.[135, 148-150] We use the pair-specific ε and σ parameters instead of the combing rule for unlike atom pair *i* and *j*. The Lennard-Jones (LJ) 6-12 potential and Buckingham potential fitted the non-bonded interactions at the onset were generated. The Buckingham function is "softer" than the LJ 6-12 function in the repulsive region because the exponential term is more suitable to represent the non-bonded potentials in our CG model. However, as shown in **Figure S3**, even the Buckingham potential is not "soft" enough. In addition, some of the interactions (e.g. N2-N2) clearly show a second or more local minima which are ignored by the Buckingham potential. The complicated shape of the non-bonded potentials can be captured much more accurately by using the spline interpolation functions as in the previous statistical potentials for proteins[151]. However, in the current study, we would like to explore the capability of the simple Buckingham potential that is implemented in almost all popular molecular modeling packages. In non-bonded potential fitting, we have chosen to primarily reproduce the global energy minima by using a weighted least-square fit. The data points in minimum energy area (0.5 Å within the potential minimum) were assigned a weight of two while the others one. The final fitting results are shown in **Figure 2.2**. As discussed in the Results and discussion section, a nonlinear optimization was later performed on the non-bonded parameters, after all bonded and non-bonded parameters in the CG model were obtained, by

minimizing the RMSD between the energy minimized and the experimental structures of selected RNAs. The experimental structures were analyzed to generate the radial distribution functions (RDF) or g(r) for selected pairs of coarse-grained particles. The set of 1-2 (directly bonded), 1-3 (separated by 2 bonds), and1-4 (separated by 3 bonds) pairs were not included in RDF calculations. Then the potential of mean force, which corresponds to the Boltzmann inversion of the g(r), is determined from the RDF:

$$E_{non-bonded}(r) = -k_B T \ln g(r)$$

By combining Buckingham potential, we determine the initial values of ε and σ for each pair. We have also combined certain pairs (the same parameters were used) based on the similarity of the RDF obtained.

The Buckingham potential fits well for most of the non-bonded pair interactions; however, it may not be suitable for charged CG atoms. In certain cases the short-range interactions are too soft for the Buckingham equation and the missing details in the analytical representation are also important. We therefore introduced an electrostatic term in our coarse-grained potential by implementing the Debye–Hückel method in the CG model. The Debye–Hückel representation is a good approximation of Poisson-Boltzmann equations, which works well in low salt concentrations such as the physiological environment. In our model the electrostatic term adopts the form of

$$E_{ele} = \sum_{i>j} \frac{q_i q_j}{4\pi D} r_{ij}^{-1} e^{-r_{ij}/\xi}$$

where ξ is the Debye length; D is the dielectric constant for water at room temperature. In comparison to the Buckingham potential, short-range interactions are well-captured for phosphate-phosphate pairs by using the Debye-Huckel potential (**Figure 2.3**). In this

work, we introduced the electrostatic interaction between phosphate-phosphate, sugar-sugar, and phosphate-sugar based on their statistical PMF. The net charge is set to -1.3*e* for phosphate and sugar groups as determined by the fitting (**Figure 2.3**). The Debye length ξ is the response for the salt concentration in the environment. The default value of ξ is set to 10 Å, which mimics the typical *in vivo* environment of 100 mL Na^+. A weak Buckingham potential was added to the phosphate and sugar beads to better reproduce the repulsive force at close distance.

Hydrogen bonding interaction is critical for base pairing and is likely to contribute to certain features that are missed by simply using Buckingham function. To model the hydrogen-bonding, we evaluate the all-atom functional forms and the dipole-dipole interaction approach that treats hydrogen bonding liquids. In the CG model, the hydrogen bond interaction has the form of:

$$E_{H-bond} = E_0 \times \cos(\alpha) \times (\frac{\sigma}{r_{ij}})^3 \qquad (\frac{\pi}{2} < \alpha < \pi)$$

where E_0 is the minimum energy value of the hydrogen bond interaction; α is the relative angle between two base pairs. r_{ij} is the distance between two heave atoms (oxygen or nitrogen) that form hydrogen bonds and σ is the distance at the minimum energy. The defined donors and acceptors in the CG model are listed in **Table 2.2**. For current model, σ is set to 2.9 Å and the strength of hydrogen-bonding E_0 is set to 0.5 kcal/mole according to the previous nucleic acid CG model [4] .

2.2.3 Determination of the RNA coarse-grained model parameters

The probability distribution of all virtual bonds, angles, and torsions (shown in the **Figure 2.4, 2.5, and 2.6**) were used to fit the valence parameters. The fitted parameters for virtual bonds, angles, and torsions are given in **Table 2.3, Table 3.4**, and **Table 2.5**,

respectively. The force constants of bonds and angles for pseudo atoms are smaller than those of atomic bonds and angles, meaning a larger time step could be used during the MD simulations in the CG model. As expected, the larger force constants of bonds and angles within the base make the bases fairly stiff (**Table 2.3** and **Table 2.4**). However, they are still smaller than those of atomic constants by a factor of 3 to 10.

The non-bonded parameters were obtained by mapping the radial distribution functions (RDF) of all the pseudo atoms in existing RNA structures:

$$g_{ij}(r) = \frac{1}{N_i d_j} \frac{n_{ij}(r)}{4\pi r^2 \delta r}$$

where $n_{ij}(r)$ is the number of pairs in the given shell from r to r+dr, Ni is the total number of particle I in the system, and d is the mean bulk density of particle j. The reference state here is the expected number of contacts when two pseudo atoms i and j at long distance, which is approximated as the average density of pseudo atoms j.[152] Therefore, the g(r) could be normalized to 1 at long distance. The results from g(r) are then used get the potential of mean force, which is approximated to be the effective potential function of r. The effective potential functions are shown in **Figure 2.2**.

2.2.4 Optimization of the non-bonded parameters

The RNA structural statistics we utilized to derive the non-bonded and bonded parameters effectively include contributions from all energy terms, although to different extents. For example, the actual conformational distribution is affected by both the torsion and non-bonded energy terms in the CG potential. To remove the "redundancy", we directly compared the structures given by the coarse-grained potential with the experimental structures and adjust the parameters. After we fitted the electrostatic and the

hydrogen bonding parameters using the effective potential (E_{eff}) parameters were first fitted to the potential of mean force, and then optimized by comparing the structures given by the coarse-grained potential with the experimental structures of a set of seven selected RNAs with diverse secondary and tertiary structures. The non-bonded parameters were refined by minimizing the difference between the energy-minimized CG structures and their corresponding experimental structures. First, energy minimization was performed on each of the seven RNA molecules, and the structural root mean square deviation (RMSD) from the experimental structure was calculated based on all pseudo atoms. The average of the RMSD over the seven molecules is used as the target function in the optimization of the non-bonded parameters. An optimally conditioned variable metric nonlinear optimization algorithm in TINKER was utilized.[153-155] The first derivative of the average RMSD with respect to each non-bonded parameter was calculated numerically. In total 38 non-bond parameters were optimized. The average RMSD between the experimental and energy-minimized structures dropped from 3.35 Å to 1.75Å by using the optimized non-bonded parameters.

2.3 RESULTS AND DISCUSSION

2.3.1 Benchmarks and validations of CG potential

After deriving and optimizing all of the parameters, the ability of CG potential to model RNA native structures was tested by fifteen different RNA molecules with various RNA motifs were tested, including frequently found RNA motifs, such as double helices, hairpin loops, interior loops and pseudoknots (see **Table 2**). The molecular dynamics simulations of the CG model were performed with the TINKER software package [156]. For comparison, all-atom simulations were also performed on the same set of RNAs in

the AMBER10 software package with Amber *ff99sb* force field [124, 126, 157]. The TIP3P water model [158, 159] and the Generalized-Born/Surface Area (GB/SA) model [160-165] were used for the explicit and implicit solvent simulations, respectively. The time step was set to 1 fs with a total simulation time of 10 ns for all RNAs. Bussi thermostat was applied to the TINKER simulations [166]. It should be noticed that the time step can be as large as 5 fs in the CG model (see detailed discussion below for the computational efficiency of CG model).

The root mean square deviation (RMSD) was used to examine the structure stability (**Figure 2.7**). The RMSDs were calculated as an average over the MD trajectories using the same atom set of five pseudo atoms per nucleotide for both the coarse-grained and the atomistic model. One base pair at each of the terminals was ignored in all of the RMSD calculations unless specified otherwise. The average RMSD is 2.71Å found from the CG model simulations for all 15 RNAs, which surprisingly, is a little smaller than the all-atom simulations with GB/SA implicit solvent (average RMSD 3.29Å). The all-atom simulations with TIP3P water model give us a slightly better result with an average RMSD of 2.21Å. The overall simulation results have shown that our CG model has comparable accuracy to all-atom models in maintaining the native structure for different RNAs. As we expect, the fluctuation of an RNA molecule is largely related to the number of Watson-Crick base pairs it contains. For example, lower RMSDs (~1.1Å) are found in RNA 1QCU (double helix) [167] and 2JXQ (double helix with a bulge) [168] that form duplex with all canonical Watson-Crick base pairs; while higher RMSDs (about 3Å) are presented in RNA 1KD3 (double helix with an interior loop) and 1LNT (double helix with an interior loop) where 5 or 6 non-canonical pairings (interior loop)

are placed in the center of the duplex [169]. Even though some RNAs had relatively larger RMSDs, the overall native structures were well maintained with reasonable fluctuations at non-canonical pairings regions.

2.3.2 Folding RNA with simulated annealing simulations

In the next step, we tested the ability of the CG model to predict RNA structures. We mimicked RNA folding and unfolding by the simulated annealing method. In our blind test we assumed all the native structures of tested RNAs were unknown. First, we performed 3 to 5 independent simulated annealing simulations on each RNA molecule. During the simulations, the temperature was initially increased to 1,000 K within 10 ns. Then we continued the simulation at 1,000 K for several nanoseconds until the RNA was fully denatured before the actual annealing simulations. Then we gradually cooled down the system to room temperature (298 K) in 100 ns. The time step was set to 2 fs. After that, we minimized the energy of final snapshot that we predicted from each simulation. We then chose the lowest energy conformation among those predicted structures as the final predicted RNA 3D structure. We evaluated our final predicted results by comparing the predicted RNA 3D structure to the native structure (**Figure 2.8**). We found all the tested RNAs were folded to their near-native structures with an average RMSD of 3.31Å from their native conformations. The introduction of the electrostatic term has greatly improved the accuracy of backbone conformations, particularly the short range interactions. Furthermore, the base-pairing conformations are better captured by adding the hydrogen-bond term in current CG potential.

To evaluate the energy landscape for these RNA molecules, we have computed the energy distribution of the random structures sampled from the simulated annealing

simulation. 10,000 structures were taken from each simulated annealing trajectory and minimized, the RMSD were computed with respect to the native structure using all the pseudo particles. Ideally, we expect the structure with lowest RMSD would have the lowest free energy. The examples we show in the **Figure 2.9** indicate that the native structures are mostly near the energy minimum on the CG energy landscape although the energy landscape of RNA can be rather flat instead of funnel-like. In other words, there could be many structures that have energy values very close to that of native structure, which makes the prediction of native structure difficult.

Electrostatic interactions play an important role in RNA structures. As nucleic acids are highly charged, ions in the cellular environment such as K^+ and Mg^{2+} greatly influence the structure and thermodynamics of RNA. Specific interactions with Mg^{2+} ions directly contribute to the tertiary structure stability of certain RNAs [170-173]. In our tested RNA set, magnesium ions greatly stabilized the pseudoknot (PDB: 1L2X)[174] structure. The *ab initio* predicted structure of the pseudoknot was around 8Å RMSD from the native structure with the simulated annealing method. We approached this poor structural prediction by explicitly including the coordination Mg^{2+} ions. The net partial charge is set to +2.0*e*. In order to maintain the Mg^{2+} ion during the simulated annealing simulation at high temperatures, a weak distance restraint with 0.5 kcal/Å was added to its nearest phosphate atoms. Similar Debye–Hückel method was used to treat electrostatic interactions between Mg^{2+} and negatively charged phosphate or sugar. Six pairs of distance restraints were added to reduce the conformational space search during the simulation. The final predicted structure was successfully folded to the near-native conformation, with a 4.8 Å RMSD from the crystal structure using all pseudo atoms

(**Figure 2.10**). The conformation of the pseudoknot stems was thus reasonably captured and the predicted locations of Mg^{2+} ions are near those in experimental structures. However, comprehensive tests on a wide range of pseudoknots and other complex RNAs are needed to fully validate the effectiveness of our approach.

2.3.3 A multi-scale approach to predict large-size RNA structures

It is encouraging that our model successfully folds small RNAs with simulated annealing simulations. We then applied the CG model to predict large-size RNAs (>100 nt). Large-size RNAs usually have complex secondary and more importantly tertiary structures. The free energy landscapes are much more rugged than those of small RNAs with simpler topologies. We introduced limited distance restraints to the canonical Watson-Crick base pairs, based on the nuclear Overhauser effect spectroscopic (NOESY) distance restraints by NMR. Fifteen pairs of distance restraints were applied to predict the 3D structure of 122-nt *H. marismortui* 5S rRNA[75]. The final predicted structure from the 100 ns simulated annealing simulation is shown in **Figure 2.11** in comparison with the crystal structure. Similar global folding was found as the native structure, where the junctions of stems I, II and III, all well as the single quasi-continuous helix form by stems II and III were well predicted by the CG model.

We tested the model with another large-size RNA, the yeast U2/U6 snRNA complex (111-nt) [175]. This time we presented a different prediction strategy (**Figure 7**). First, we performed a number of short simulated annealing simulations of 20 ns starting with a random coil structure of the U2/U6 snRNA complex. In each simulation, 10 pairs of weak distance restraints were added to the canonical Watson-Crick base pairs. After the simulated annealing simulations, each predicted 3D structure was mapped back

to its all-atom structure, in which its' small-angle X-ray scattering (SAXS) amplitudes were predicted using the FOXS web server [176]. Then each predicted SAXS amplitude profile was compared to the experimental SAXS amplitude with χ^2 goodness-of-fit analysis. The predicted structure with the lowest χ^2 was chosen for further all-atom refinement, where all-atom minimization and equilibration was performed to further relax the predicted structure. A good agreement was found by superposing the final predicted structure to the experimental NMR structure (**Figure 2.12**).

Toward this end, we propose a multi-scale approach to predict and model RNA 3D structures (**Figure 2.13**). First, the secondary structure and experimental information are utilized to narrow down the structural sampling; the secondary structure information can be directly obtained from current databases, comparative analyses, or even from secondary structure prediction software (e.g. RNAfold [177]). Stable helices, hairpin loops and other motifs predicted in the secondary structures can be represented as a set of restraints on the canonical base pairs, which significantly reduce the configurational space in 3D. Next, the RNA is subject to coarse-grained molecular dynamics to fold into 3D structure. The restraints from the first step are enforced. After CG modeling, we expect to have a number of candidate 3D structures at a low resolution. In the next step, the candidate structures are chosen to map to their atomic structures. Then the SAXS amplitudes of these predicted structures are filtered by the agreement with the experimental SAXS amplitudes. The structure with best agreement will be further refined using an all-atom force field so that we can expect an accurate RNA 3D structure with atomic details.

2.3.4 Computational efficiency of the coarse-grained model

Increasing computational efficiency is the greatest impetus for the CG model development. With the CG approach, a much smaller number of particles and interactions need to be considered, and the computational efficiency of the CG model can be greatly improved over that of the all-atom model. Our "intermediate" five-bead model can reduce the number of "atoms" by nearly an order of magnitude, and the number of pair interactions can decrease even by two orders of magnitude comparing to the all-atom models. Furthermore, additional speedup was achieved in *ab initio* structural prediction using molecular dynamics simulations, because in CG models, the high-frequency motions in the all-atom model are absent, and the simulation time step can be as large as 5 fs in the MD simulations without a noticeable effect on energy and structural stability. Overall, our model can speed up the simulation by three orders of magnitude compared the atomistic models in TINKER. The computational efficiency of other RNA prediction methods has been discussed in a recent review article by Laing and Schlick [178].

The conformational sampling is another advantage of CG model. In all-atom systems of macromolecules, the conformational sampling is usually prohibitive, because the free energy landscape is extremely rugged, with multiple local minima that may trap a simulation for its entire duration. Even with current parallel supercomputers, successful atomistic 3D folding studies are limited to small size RNA (< 30 residues). In contrast, conformational sampling on the CG energy landscape is more efficient because the energy surface is much "smoother" due to the reduced number of particles and the larger mass of each pseudo-atom.

2.3.5 Modeling RNA structures in *RNA-Puzzles*

The accuracy of our 5-bead model was compared with other 7 RNA structure prediction tools previously applied in *RNA-Puzzles*[179], which is a CASP-like evaluation of RNA 3D structure prediction. We mimicked the prediction process without knowing any extra prior structural information. For each RNA molecule, we performed 5 independent 30-ns simulated annealing simulations, and then we chose the final structure with lowest energy as our best prediction. Limited RNA secondary structures information predicted by *mfold* was adopted as restraints in the 3D structure prediction as explained below[180]. The prediction results are shown in **Figure 2.14** and **Figure 2.17**.

The first RNA molecule is a homodimer that contains two strands of the sequence with blunt ends (C-G closing base pairs). We added 9 weak restraints (k=0.5 kcal/Å) at Watson-Crick base pairs from the *mfold*-predicted secondary structure. The lowest energy structure predicted displayed a RMSD of 5.03 Å from its x-ray crystal structure, which is comparable to the predictions made by the seven other tools (RMSDs ranging from 3.41 Å to 6.94 Å) (**Figure 2.15**). The prediction of second 100-nt square of double-stranded RNA was straightforward, because the secondary structure and the 3D coordinates of the nucleotides in the inner strands were provided and utilized by all prediction methods. Our model predicted a structure with a RMSD of 2.58 Å from the experimental structure, whereas the structures by the other prediction methods displayed RMSDs ranging from 2.3 Å to 3.65 Å (**Figure 2.16**).

The last RNA molecule is a riboswitch domain, which contains several tertiary contacts and was the most complex among the three *RNA-puzzles*. Similar to the first RNA dimer, we again utilized the *mfold* web server to predict the secondary structure of this riboswitch domain from its primary sequence. Then we added 14 weak distance

restraints (0.5 kcal/Å) at the Watson-Crick base pairs from the predicted secondary structure. Interestingly, our predicted structure has a RMSD of 7.66 Å from the experimentally determined structure (**Figure 2.18**), which is very close to the best-predicted result from Chen's group (RMSD=7.24 Å). Considering the mean RMSD is 14.4 Å among all submitted structures, our model performed well in capturing the conformation of this RNA molecule with complex tertiary structures.

2.4 CONCLUSIONS

Understanding RNA three-dimensional structure is crucial to our understanding of the mysterious RNA world. In this work, we proposed an "intermediate" coarse-grained model to provide both accuracy and efficiency for RNA 3D structure modeling and prediction. Each nucleobase was represented with three pseudo-atoms in order to better capture the base stacking and pairing. The overall potential with the bond, angle, torsion, and non-bonded parameters was derived based on structural statistics sampled from experimental structures. The non-bonded interactions now include electrostatics, and hydrogen bonding interactions. The parameters were derived from RNA structural statistics and optimized analytically by comparing the CG minimum-energy structures with the experimental structures. Therefore we have a hybrid potential that is constructed based on structural statistics but also consider the critical physical interactions.

With molecular dynamic simulations and simulated annealing simulation protocol, our CG model has shown reasonable successes in folding all 11 tested RNAs to native structures with an average RMSD of 3.3 Å from the native structures. The current model can well predict the structure of 122-nt 5S ribosome with limited restraints on Watson-Crick base pairs. We further proposed a multi-scale approach to predict large-

size RNAs that utilize NOESY atom coupling information and SAXS data and all-atom minimization and molecular dynamic simulations. With the multi-scale approach, we were able to predict the overall fold of the 111-nt U2/U6 snRNA complex from the sequence. In addition, cation-induced RNA folding has been explored in the model. We demonstrated the benefit of explicit consideration of coordinating Mg^{2+} cations in folding a pseudoknot with the CG model.

We have also examined the molecules in the *RNA-Puzzles* retrospectively[179]. With the secondary structure prediction provided by *mfold*, limited, weak distance restraints were applied to WC base pairs in the 3D structure prediction by our CG models. The accuracy of our prediction is about average for the first two RNA molecules; for the RNA molecule with complex tertiary contacts, our model performed as well as the previous top performer.

With CG representations and simplified knowledge-based statistical potential, it remains challenging to accurately capture the tertiary structure of RNAs with long-range interactions. Our model has shown to be effective when the secondary structure and some experimental information are utilized for large-size RNAs with complex structures. Overall incorporation of realistic physical interactions into the statistical potentials has offered notable improvement. RNA structure is quite flexible and sensitive to the change of environments. It is possible that functional structures may not be the same as the "native" states we have determined by X-ray or NMR. Therefore, environment-dependent structure prediction associated with sufficient conformational sampling is required to study the RNA dynamic structures. The CG model reported here lays the ground for such physics based coarse-grain models. Given the simplicity of the proposed

model, a variety of existing sampling techniques (e.g. replica exchange molecular dynamics) and high-performance computing algorithms (e.g. parallelization) can be adopted to further improve the efficiency of the CG model for high-throughput applications.

Table 2.1: The properties of nine coarse-grained (CG) particles

number	CG particle name	mass (amu)	bond connections
1	P	94.970	2
2	S	97.054	3
3	CG	53.022	3
4	N6	42.030	2
5	N2	54.030	2
6	O6	43.014	2
7	O2	42.006	2
8	CU	26.016	3
9	CA	39.015	2

Table 2.2: The donors and acceptors defined in hydrogen bonding in the CG model

	A	G	C	U
Donors	N6	N2	N6	N/A
Acceptors	N/A	O6	O2	O6, O2

Table 2.3: The bond stretching interaction parameters for the CG model of RNA fitted by the gaussian function and obtained from statistical structures

bond	b0	K_{bond}
1 - 2	3.85	11.12
2 - 3	3.74	9.79
2 - 8	3.61	10.89
3 - 4	4.29	57.70
3 - 5	5.66	51.66
3 - 6	4.28	44.60
3 - 9	4.33	109.19
4 - 7	4.55	44.00
4 - 8	3.59	124.29
4 - 9	3.53	93.79
5 - 6	4.57	37.14
6 - 7	4.53	57.10
6 - 8	3.55	89.85
7 - 8	3.52	82.87

Table 2.4: The bond angle interaction parameters for the CG model of RNA fitted by the gaussian function and obtained from statistical structures

angle	θ0	Ka
1-2-1	102.78	1.356
1-2-3	101.75	5.271
1-2-3'[a]	75.89	1.864
1-2-8	100.79	9.115
1-2-8'	74.40	2.386
2-1-2	106.18	2.040
2-3-4	154.72	7.130
2-3-5	104.12	12.734
2-3-6	153.94	8.162
2-3-9	108.78	10.611
2-8-4	163.79	6.794
2-8-6	163.79	6.794
2-8-7	88.99	15.930
3-4-9	66.45	35.882
3-5-6	79.38	16.156
3-6-5	48.06	21.701
4-3-9	48.33	49.428
4-7-8	49.44	24.490
4-8-7	79.78	29.398
4-9-3	65.22	17.290
5-3-6	52.57	50.065
6-7-8	50.54	38.613
6-8-7	79.46	31.109
7-4-8	50.84	29.033
7-6-8	49.98	30.600

[a] The prime in the table indicates the atom comes from its neighbor residue

Table 2.5: The optimized CG parameters for the dihedral interaction term

torsion	V1	δ1	V2	δ2	V3	δ3
1’-2-3-4[a]	3.354	120	-0.606	180	-0.068	120
1’-2-3-9	3.801	120	0.383	180	-0.287	120
1-2-1'-2'	1.358	0	0.944	180	0.574	0
1-2-3-4	2.964	15	-0.099	180	-0.247	15
1'-2-3-5	3.603	120	1.167	180	-0.325	120
1-2-3-5	3.768	0	0.52	180	0.581	0
1'-2-3-6	3.409	120	-0.265	180	-0.226	120
1-2-3-6	3.077	30	0.306	180	0.246	30
1-2-3-9	3.299	15	0.634	180	-0.204	15
1'-2-8-4	3.461	120	-0.617	180	0.294	120
1-2-8-4	3.321	30	1.121	180	-0.156	30
1'-2-8-6	2.737	120	-0.666	180	0.148	120
1-2-8-6	2.51	30	0.518	180	-0.17	30
1'-2-8-7	3.304	120	1.349	180	-0.342	120
1-2-8-7	3.844	0	0.567	180	0.534	0
2-1-2'-1'	-1.626	135	-0.113	180	-0.246	135
2'-1'-2-3	-1.661	60	0.455	180	0.311	60
2'-1-2-3	1.387	120	0.898	180	-0.516	120
2'-1'-2-8	-1.531	45	0.489	180	0.686	45
2'-1-2-8	1.38	135	0.908	180	-0.691	135
2-3-4-9	7.114	150	-2.4	180	0.516	150
2-3-5-6	-3.328	120	0.95	180	0.101	120
2-3-6-5	5.639	150	-2.063	180	-0.009	150
2-3-9-4	2.959	15	-1.022	180	0.666	15
2-8-4-7	5.024	165	-1.509	180	-1.807	165
2-8-6-7	4.756	165	-1.037	180	-1.455	165
2-8-7-4	-4.072	150	0.544	180	-0.144	150
2-8-7-6	3.51	0	0.425	180	0.457	0

[a] The prime in the table indicates the atom comes from its neighbor residue

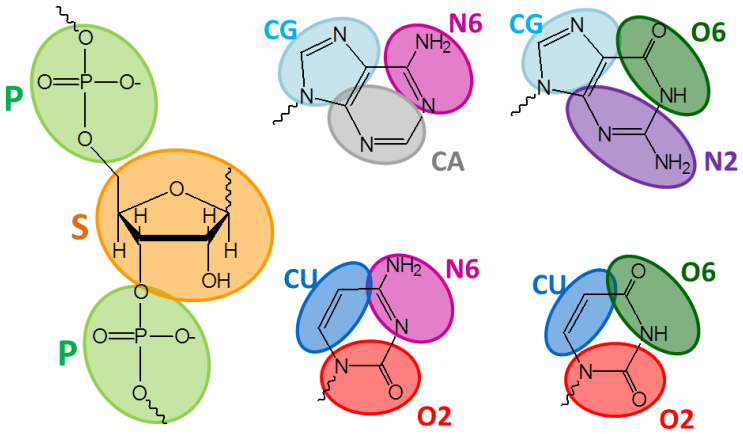


Figure 2.1: Schematic representation of the coarse-grained (CG) model for RNA. Phosphate and sugar are represented as one CG particle. The bases A, G, C, and U are represented as three CG particles for each.

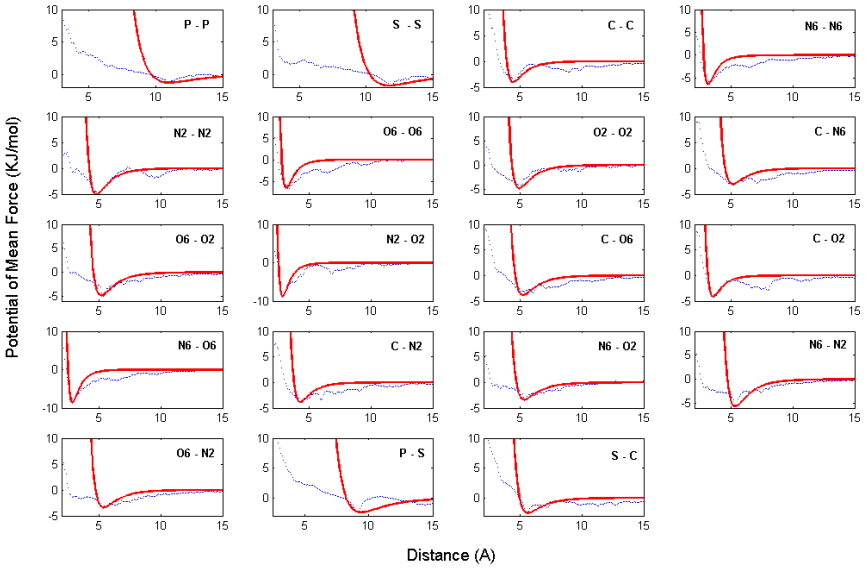


Figure 2.2: Potential of mean force for similar and unlike pairs of CG atoms. The blue dotted lines are the statistical results and the red solid lines are the fitted Buckingham potential curves. The potential of mean force were obtained from the intermolecular RDF for the 9 CG atoms, whose values are used as the initial values of the non-bonded parameters.

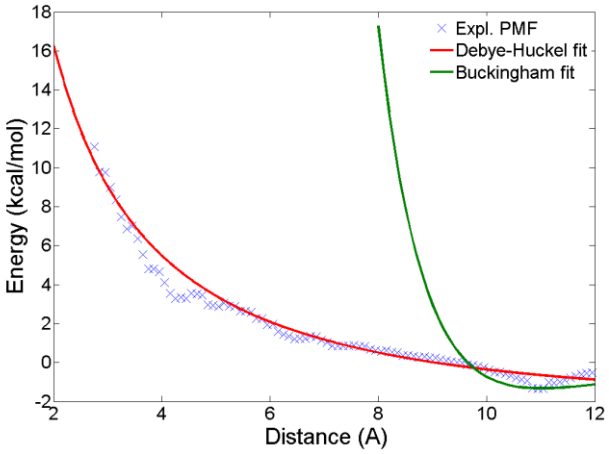


Figure 2.3: Fit the non-bonded interaction between phosphate groups in CG model. The statistical calculated potential of mean force are shown with blue cross, the best fit with Buckingham potential are shown in green line, and the best fit with Debye–Hückel potential are shown in red line. Comparing to the Buckingham potential, short-range interactions are well-captured for phosphate-phosphate pair by using Debye–Hückel potential

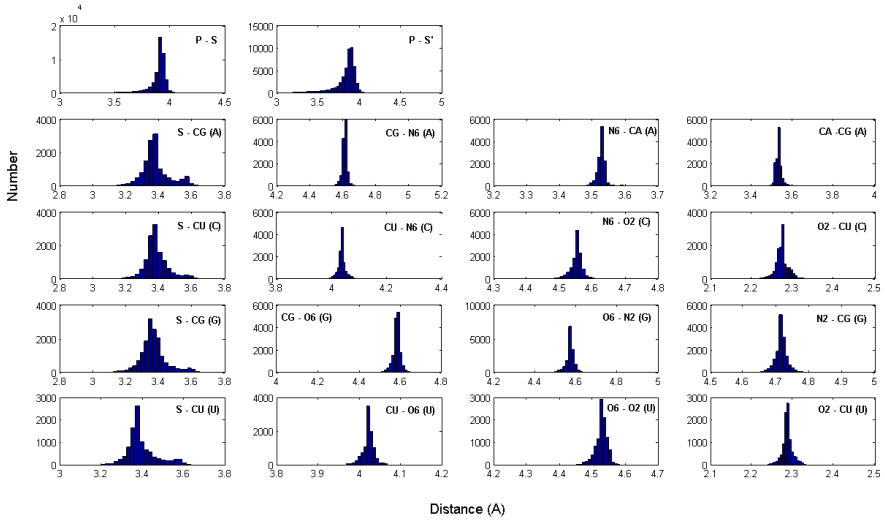


Figure 2.4: Histogram of the bond length distributions between CG atoms obtained from statistical structures.

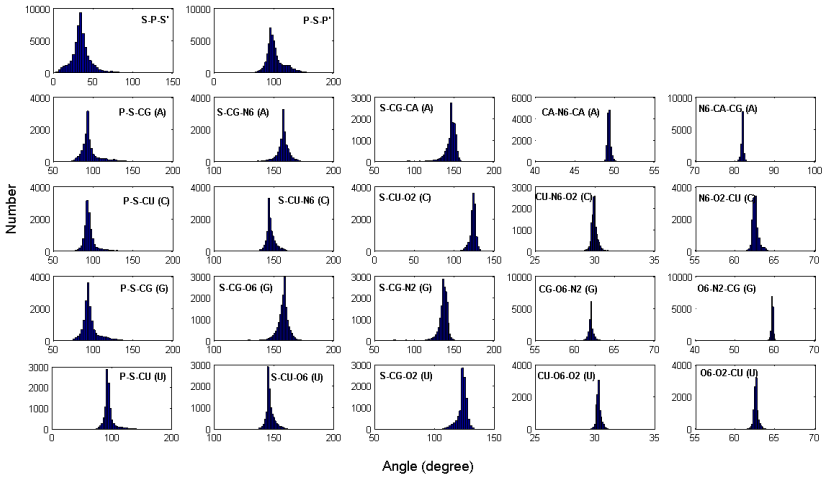


Figure 2.5: Histogram of the bond angle distributions between CG atoms obtained from statistical structures. The primes at atoms S and P indicate the atoms come from their neighbor residues.

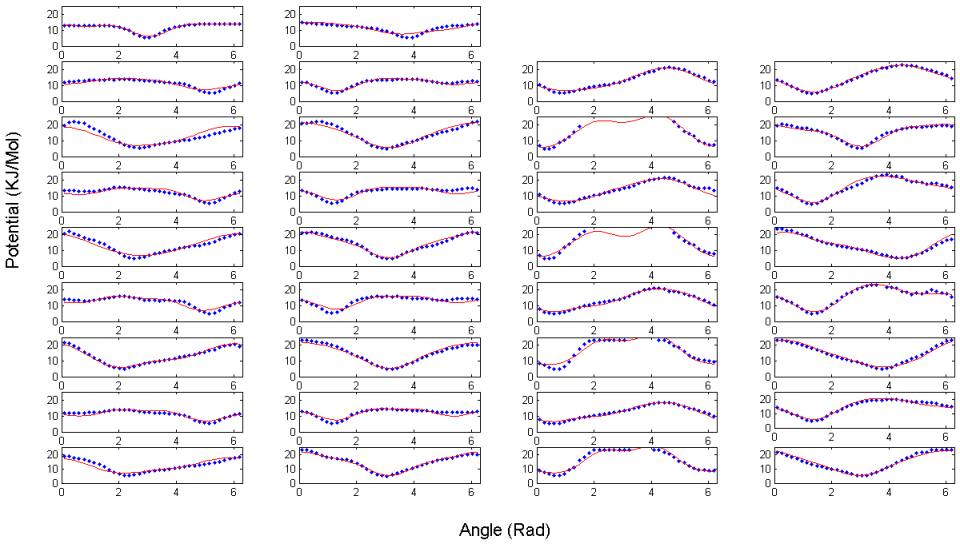


Figure 2.6: Torsion potential between CG atoms obtained from statistical structures. The blue dot is the statistical results and the red line is the fitting curve. The primes at atoms S and P indicate the atoms come from their neighbor residues.

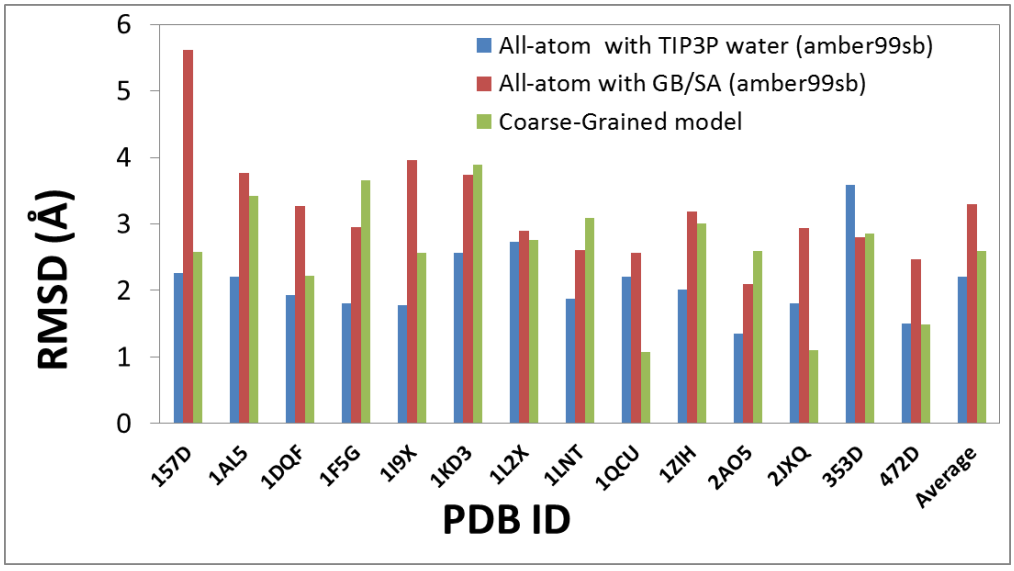
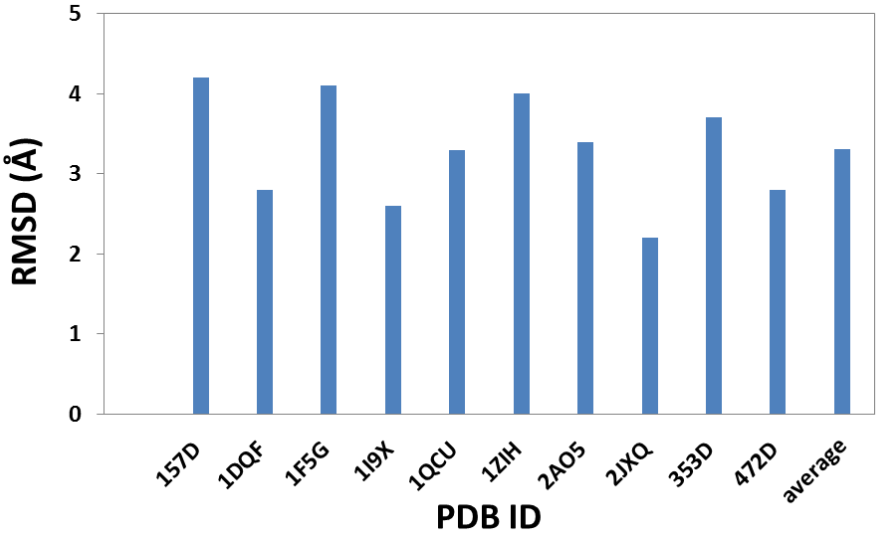


Figure 2.7: Comparison of all-atom average RMSDs from the native crystal structures for both the CG model and the full-atom models. All RMSDs were obtained from all CG atoms (all-atom calculation using the same atom set as CG model).



PDBID	RMSD (Å)
157D	**4.2**
1DQF	**2.8**
1F5G	**4.1**
1I9X	**2.6**
1QCU	**3.3**
1ZIH	**4**
2AO5	**3.4**
2JXQ	**2.2**
353D	**3.7**
472D	**2.8**
average	**3.31**

Figure 2.8: Comparison of RMSDs between the simulated-annealing predicted structures to their native structures. All the CG particles are included in the RMSD calculations.

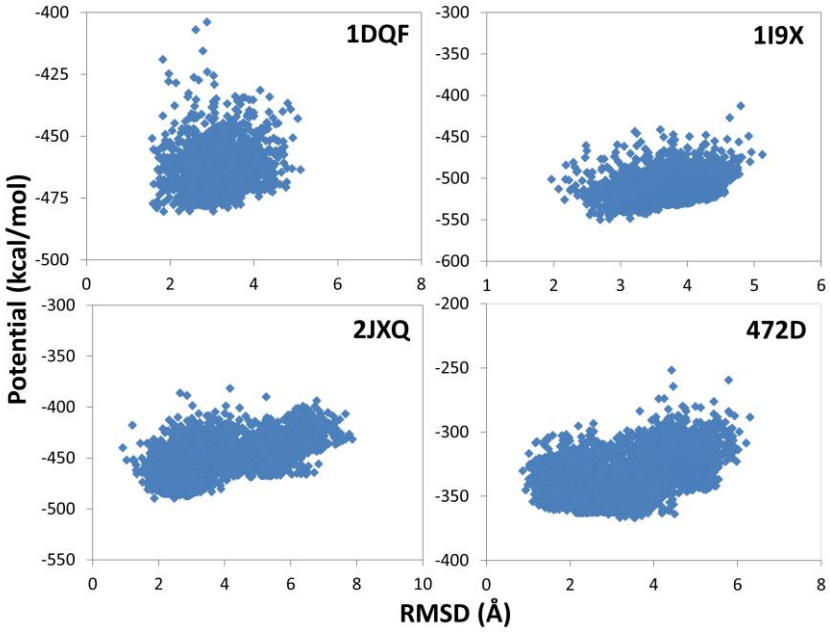


Figure 2.9: Potential energy VS RMSD plot. The near-native structures are picked from the simulated annealing simulations and then minimized. The potential energy and the all-pseudo-atom RMSDs are calculated after the minimizations.

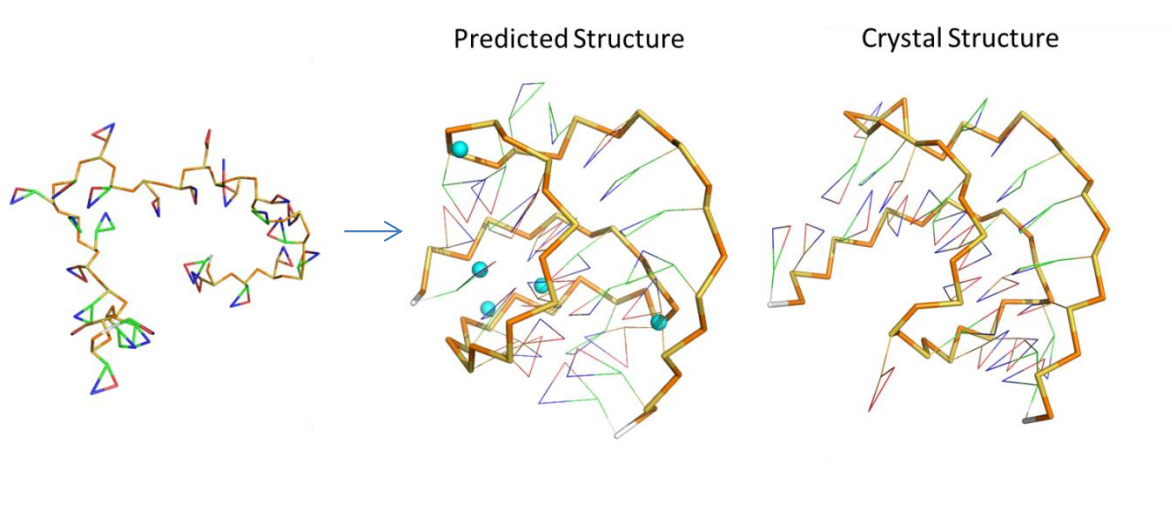


Figure 2.10: Predict the 3D structure of pseudoknot 1L2X with the coordination Mg^{2+} ions explicitly present in the CG model. The predicted structure is obtained from the final snapshot of 100 ns simulated-annealing simulation.

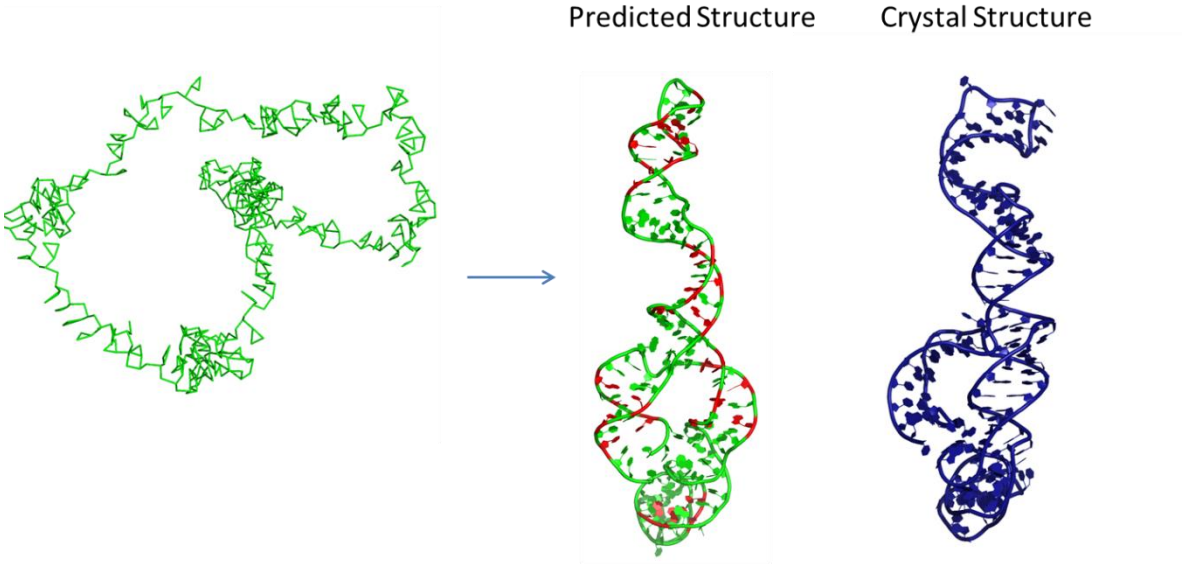


Figure 2.11: Predict the 3D structure of 122-nt *H. marismortui* 5S rRNA with simulated-annealing simulation. The predicted structure is shown in green and the crystal structure is shown in blue. The restrained Watson-Crick base pairs are indicated in red color.

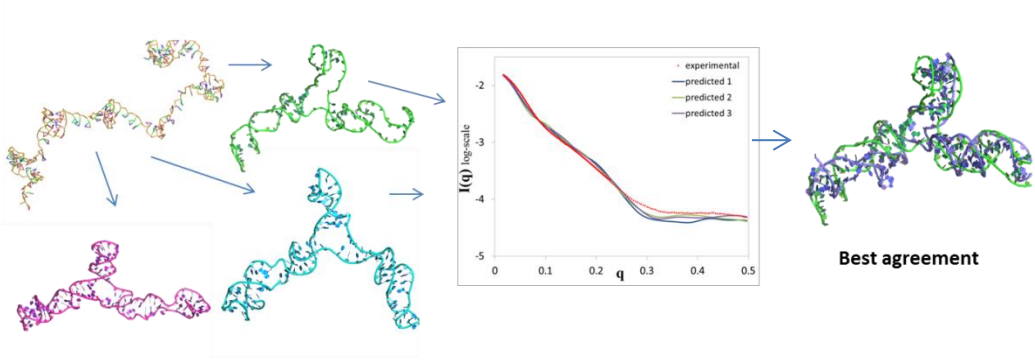


Figure 2.12: Predict the 3D structure of yeast U2/U6 snRNA complex with experimental small-angle X-ray scattering (SAXS) profile and all-atom refinement.

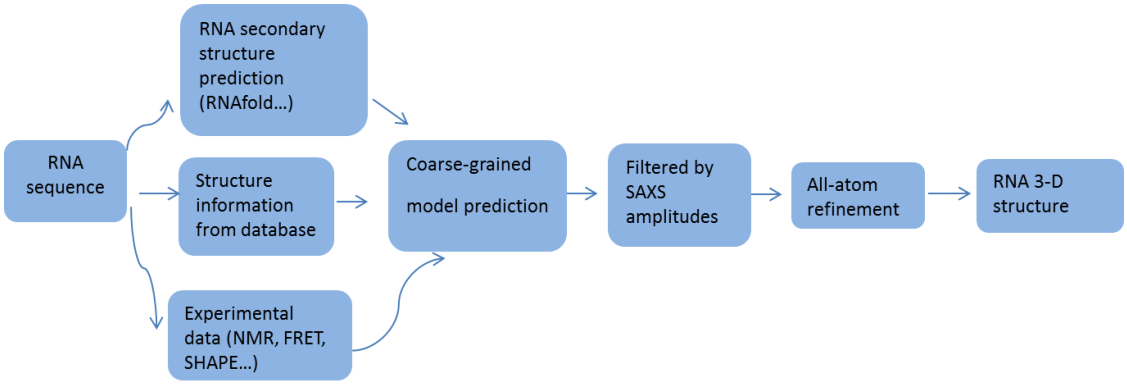


Figure 2.13: Schematic view of the multi-scale approach to predict RNA structures.

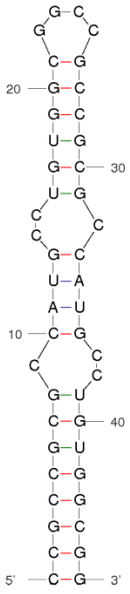


Figure 2.14: Predict the secondary structure of a homodimer in *RNA-Puzzles* that contains two strands of the sequence with blunt ends (C-G closing base pairs). The structure was predicted by *mfold* web server.

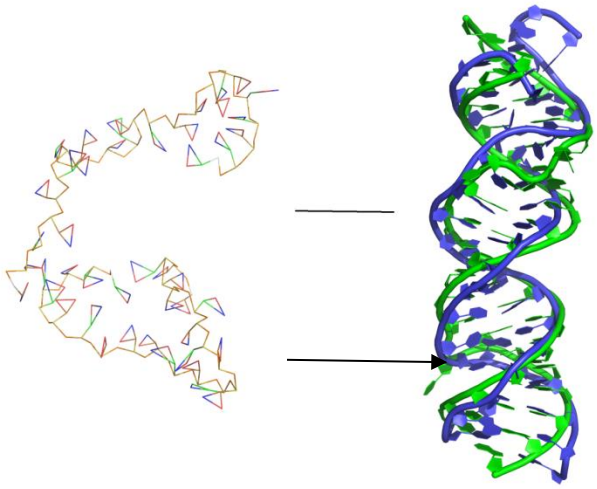


Figure 2.15: Predict the 3D structure of a homodimer in *RNA-Puzzles* that contains two strands of the sequence with blunt ends (C-G closing base pairs). The predicted structure is shown in blue and the crystal structure is shown in green.

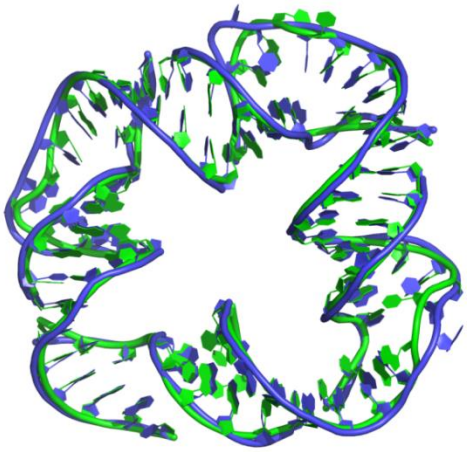


Figure 2.16: Predict the 3D structure of a 100-nt square of double-stranded RNA in *RNA-Puzzles* that self-assembles from four identical inner and four identical outer strands. The predicted structure is shown in blue and the crystal structure is shown in green.

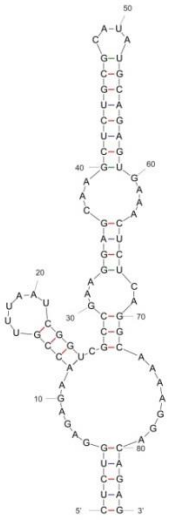


Figure 2.17: Predict the secondary structure of a riboswitch domain in *RNA-Puzzles*. The structure was predicted by *mfold* web server

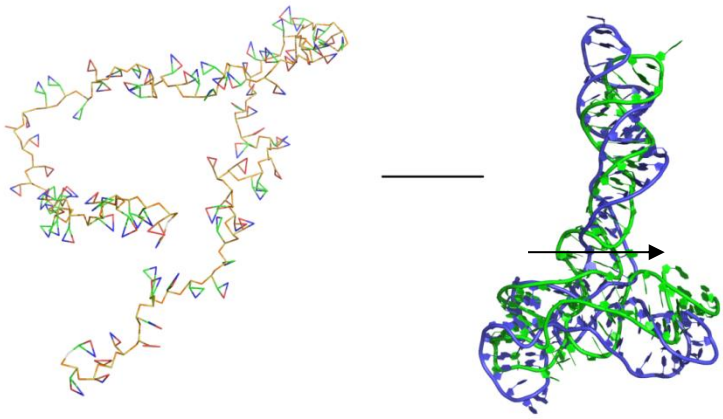


Figure 2.18: Predict the 3D structure of a riboswitch domain in *RNA-Puzzles*. The predicted structure is shown in blue and the crystal structure is shown in green.

3 Gay-Berne and Electrostatic Multipole Based Coarse-grain Potential in Implicit Solvent

3.1 INTRODUCTION

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 [181]. 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 [138, 182]. DeVane and coworkers have recently embarked on a method that employs the Lennard-Jones 9-6 and 12-4 forms to model nonbonded interactions of coarse-grain sites and have thus far validated the model on various amino acid side-chain analogs[183]. 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 [184]. 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 [185-192]. Alternatively, sequence-based statistical potentials have been used as a coarse-grain approach to fold t-RNA, 5S, and 16S ribosomal RNA [122, 193].

In this chapter, a general coarse-grain model, consisting of rigid bodies of anisotropic Gay-Berne particles and point multipoles, has been developed. The Generalized Kirkwood method is applied to account for the solvation effects [194]. While the current coarse-grained (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.

3.2 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 [134, 195] and in the supporting information as well. Based on Gaussian-overlap potential, the potential energy between two particles i and j has the form

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

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

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

and

$$\sigma_0 = \sqrt{d_i^{\,2}+d_j^{\,2}}$$

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

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

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:

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

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

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

The total well-depth parameter is presented as

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

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

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

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

where

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

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

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[196]. 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 [197]. 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 [198].

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

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

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

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

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. A complete description of electrostatic interactions of the GBEMP model are provided in previous work [134, 195].

3.3 Results And Discussion

3.3.1 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 [134, 195], 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 [199, 200].

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 (**Table 3.1**). 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 3.1**). 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%) (**Table 3.2** and **3.3**). More detailed comparison among CG and all-atom simulations, as well as experiment can be found in the supporting information. The GBEMP model is next extended to polyalanine peptides that consist of bonded coarse-grained particles.

3.3.2 Alanine model

In the 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 [148]. To use large time-step in MD simulations, the bond and angle terms can be restrained using rattle algorithm [201].

In a previous work [134], 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 3.2**. 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 3.2**, 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 3.2** 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 [194], a multipolar extension of the Generalized Born approach [160, 202] and is computed for all the multipole sites. The Grycuk effective radius [203] is used in the polar solvation calculations. Nonpolar solvation is evaluated for all Gay-Berne sites with the ACE surface area method [161] and Still method[160, 204] to estimate the effective radius of each particle. All solvation methods as well as effective radii estimation methods are implemented in the TINKER 5[198] 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 3.2** (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 [160, 161]. 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 [205]. 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.

3.3.3 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 3.3**, compared with corresponding energies from all-atom model using the OPLSAA field [206]. 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 3.4b and 3.4e**), the solvation energy contribution (**Figure 3.4c and 3.4f**) 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 3.4a and 3.4d**), including the scale. Likewise, contour maps of the gas-phase electrostatic energy (**Figure 3.4b and 3.4e**), as well as the implicit solvation energy (**Figure 3.4c and 3.4f**), 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 3.4a and 3.4d**). A comparison of the torsional energy contribution (supporting information) between the CG and all-atom models also expresses a consistent behavior. The gas-phase conformational energy captures the C5 local minimum well[207]. 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.

3.3.4 Simulation of polyalanine

The conformation of polyalanine with various lengths has been investigated with both experimental and computational approaches [205, 208-221]. 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) [222] 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 3.5a**). 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 3.4**). Although circular dichroism (CD) spectroscopy and Fourier-transform infrared (FTIR) experiments reported somewhat less alpha-helix conformation [221], 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 [221].

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 fractions 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 3.3a and 3.3c**), a substantial (5-fold) decrease in the beta-strand fraction 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. Five independent simulated annealing trials were performed and an example of the final structure is shown in **Figure 3.6**. The final polyalanine structures after simulated annealing all adopt the alpha-helical conformation at low temperatures (100 K, **Figure 3.7**). A comparison of the RMSD between structures obtained from the simulated annealing trajectory and a canonical alpha-helix (**Figure 3.6**) suggests that the accessible area of phase-space noticeably increases as the temperature rises above 500 K.

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 and is provided in the supplementary material (**Figure 3.8**).

3.3.5 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. 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.

3.4 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. In this CG potential, the fundamental components of intermolecular forces are represented explicitly: 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, including 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 [161, 194]. 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 populations 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. Further speedup is possible if the bonds and angles are restrained. 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 3.1: 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 3.2: 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 [223]

Table 3.3: 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 [224]
[c] From the enthalpy of vaporization in reference [225]
[d] From reference [226]

Table 3.4: Per-residue fractions 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

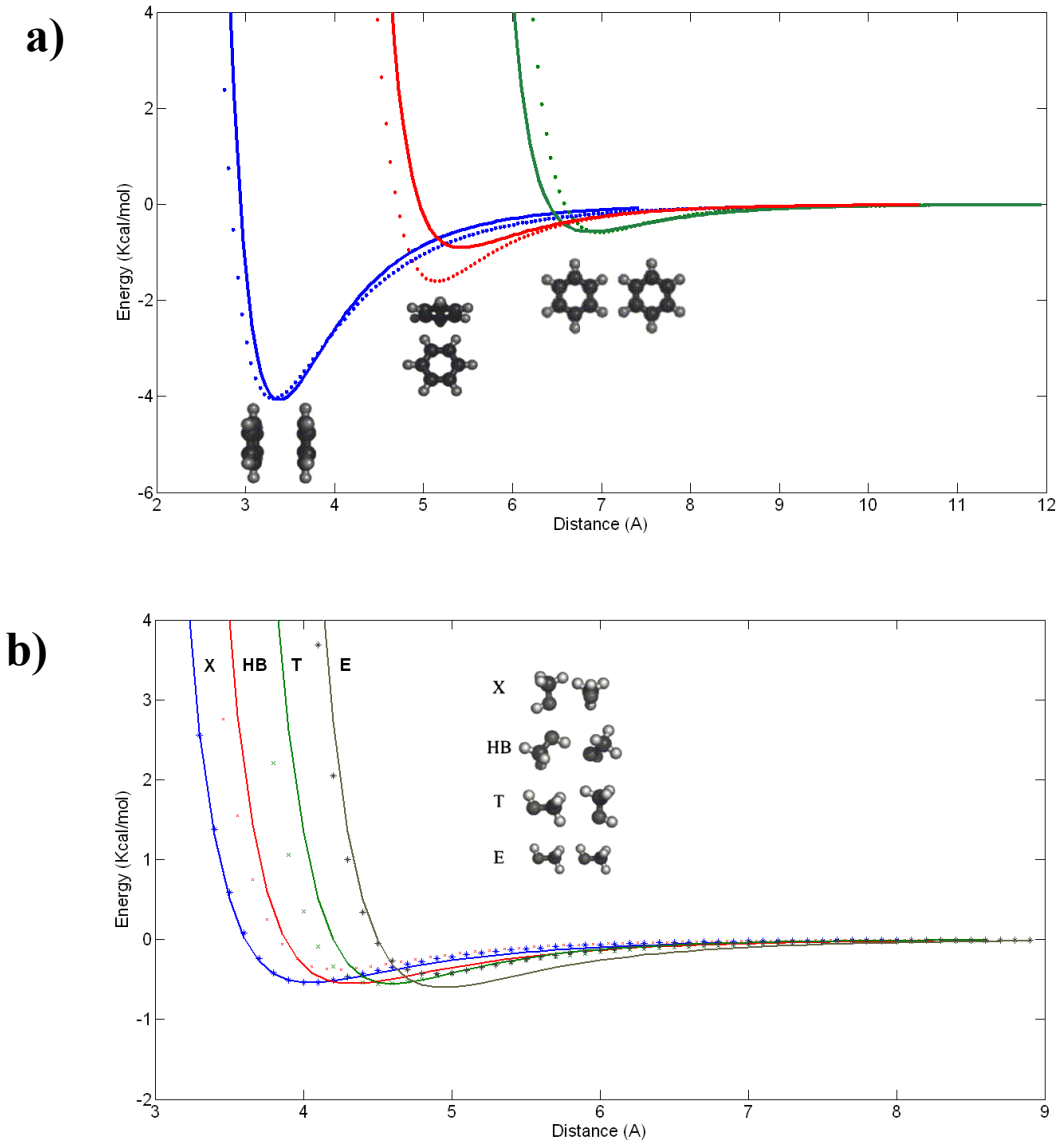


Figure 3.1: Comparison of homodimer interaction energy given by the Gay-Berne model and all-atom model. All atom values are shown as data point, and GB as line in different colors. a. The interaction energy of benzene, the conformations that shown from left to right are: face to face, T shape, side by 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.

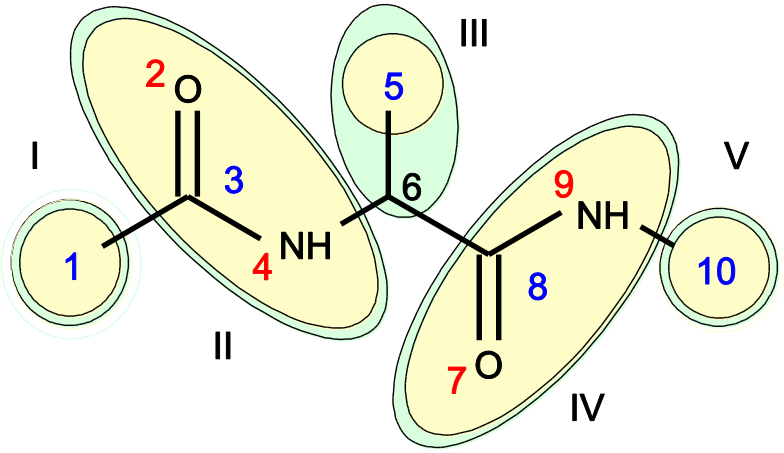


Figure 3.2: Representation of dialanine coarse-grained GBEMP model. Ellipsoids encompass the rigid bodies (green) that contains Gay-Berne (blue) and multipole (red) interaction sites. The Gay-Berne particles are located are at the center of the mass of the corresponding atoms.

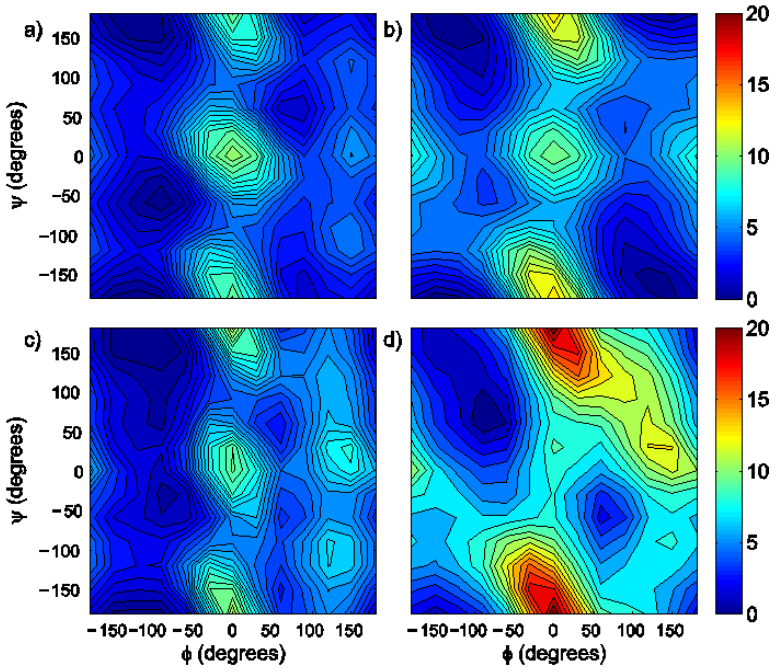


Figure 3.3: Total conformational energy (kcal/mol) of alanine dipeptide: (a) CG model in solution, (b) CG model in gas-phase, (c) all-atom model (OPLSAA) in solution, (d) all-atom model (OPLSAA) in gas-phase.

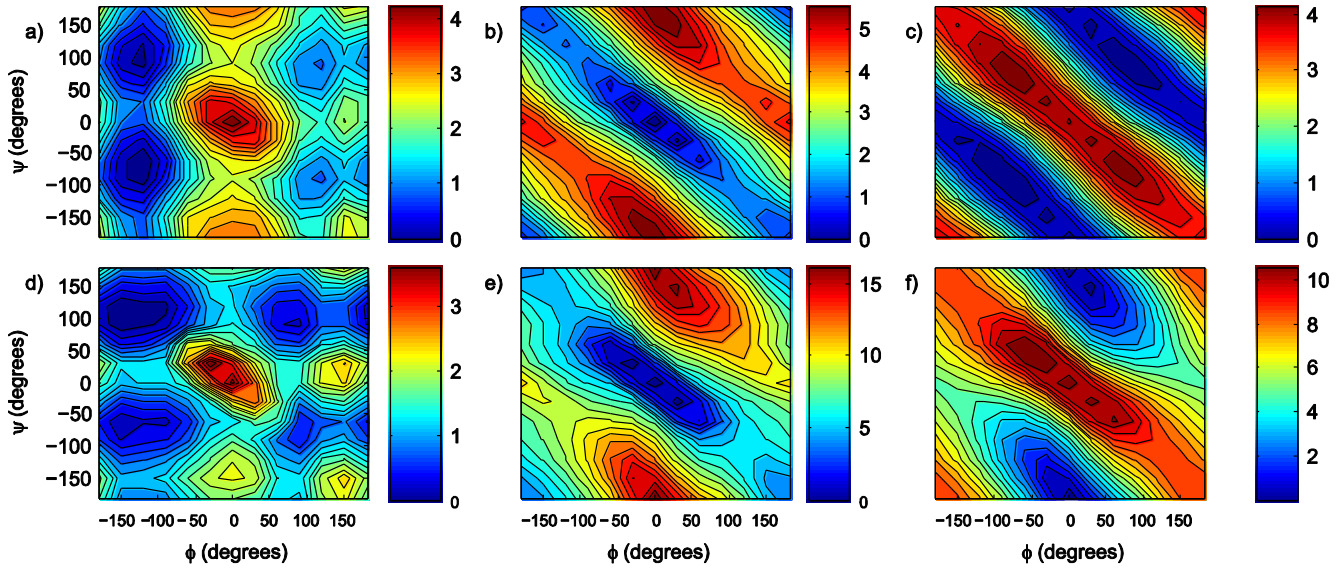


Figure 3.4: Decomposition of alanine dipeptide energy (kcal/mol). Coarse-grain: (a) Gay-Berne energy (b) Gas-phase electrostatic energy (c) implicit solvation energy from GK/SA. All-atom: (d) vdW energy (e) Gas-phase electrostatic energy (f) implicit solvation energy from GB/SA.

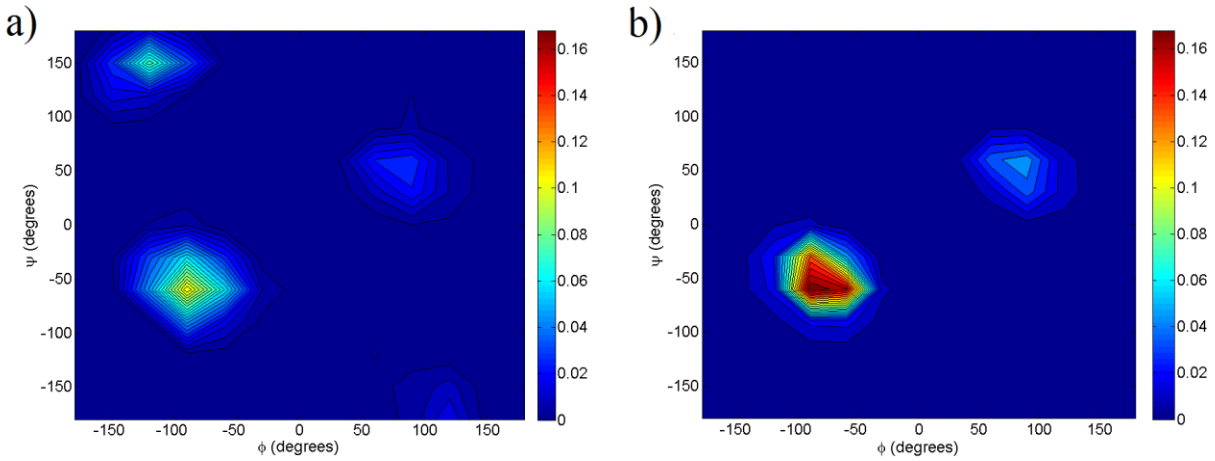


Figure 3.5: Conformational distribution of 5-mer (a) and 12-mer (b) polyalanine from CG REMD simulations.

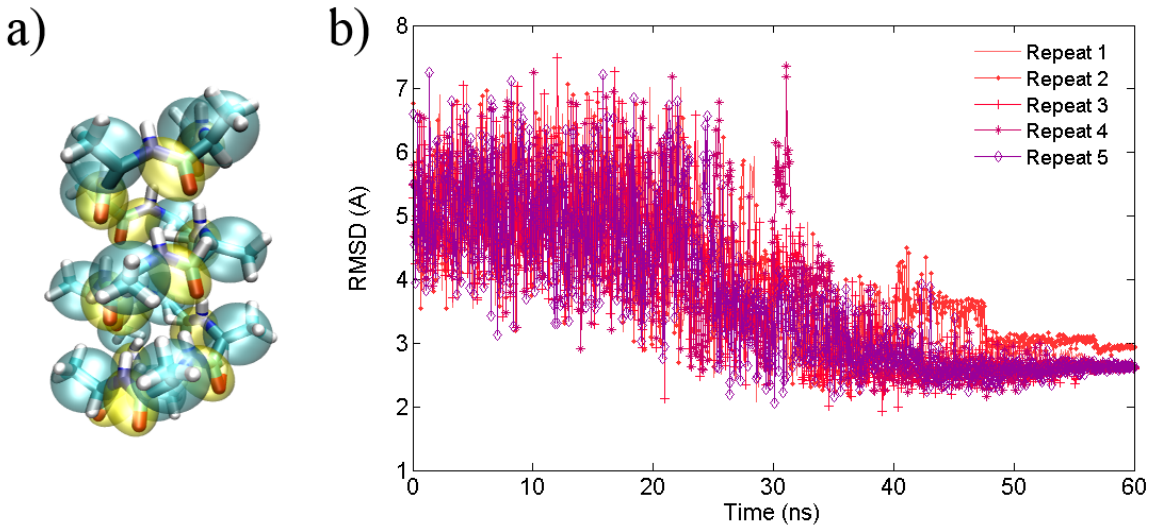


Figure 3.6: Simulated annealing MD simulations were performed to inspect the minimum-energy structure of the peptide after an initial rigid-body energy minimization. (a) A final snapshot of polyalanine from the 60-ns simulated annealing simulations using GBEMP potential. (b) Heavy-atom RMSD of the 12-residue polyalanine from 5 simulated annealing simulations.

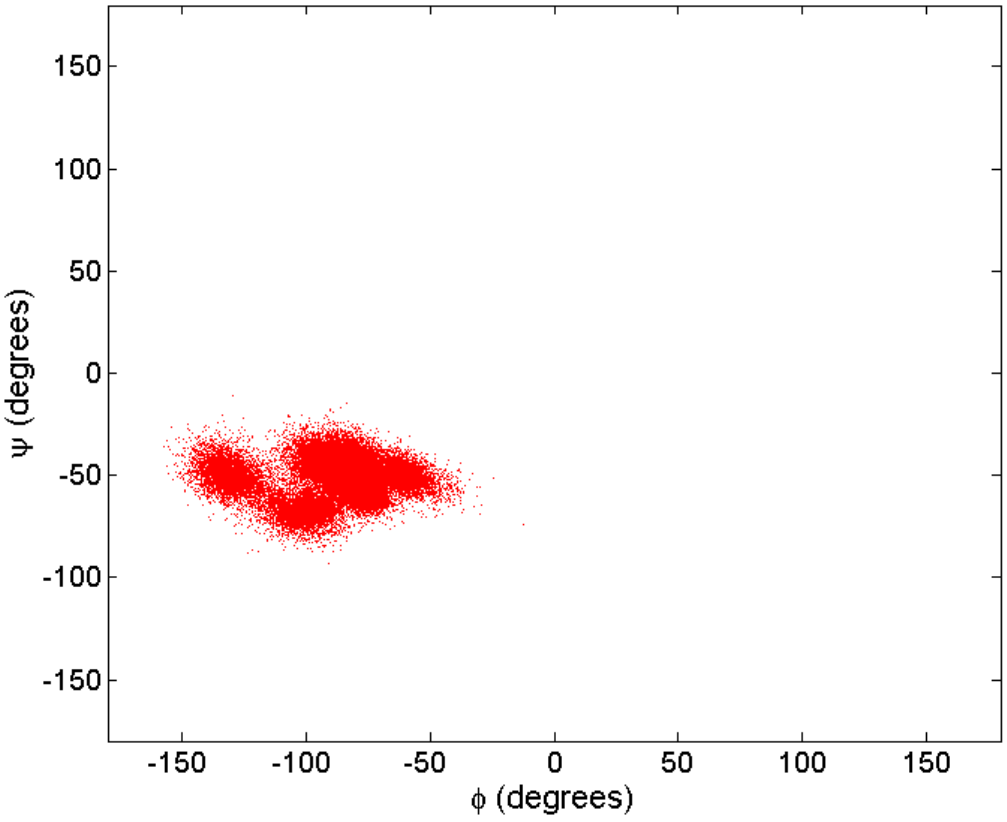


Figure 3.7: Phi and Psi torsion angle distribution of 12-mer polyalanine at temperature of 1 K to 100 K in the simulated annealing simulation. Alpha-helix become the only structure at low temperature for polyalanine.

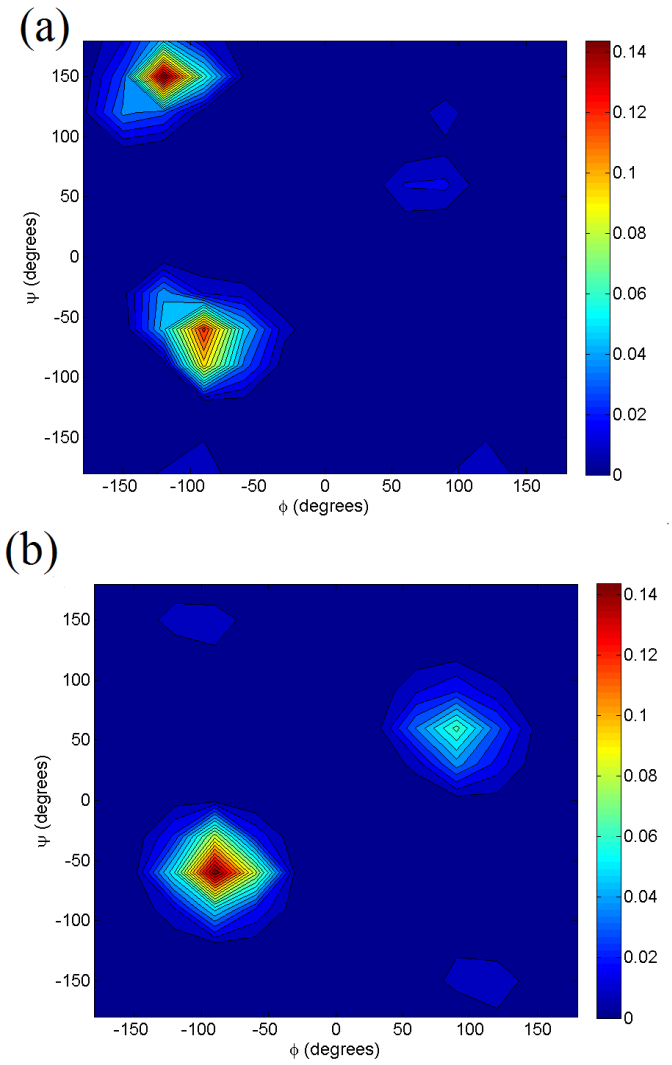


Figure 3.8: Conformational distributions of 5-mer (a) and 12-mer (b) polyalanine from CG simulations at 298 K.

4 Optimize the Torsion Parameters of Amino Acid Backbone in AMOEBA polarizable All-atom Force Field

4.1 INTRODUCTION

Electrostatic forces are crucial intermolecular forces in molecular systems. Accurate representation of the electrostatic interactions remains a grand challenge in molecular modeling and simulations [227]. Fundamentally the electrostatic interaction can be described by Coulumb's law, as employed in most classical force fields, including AMBER [228], CHARMM [229], GROMOS [230], MM3 [149, 150], OPLS [231]. In these force fields, the polarization effect is implicitly included in the parameters, by increasing or decreasing the magnitude of the partial charges in an average fashion. Because many force fields target water environment, the partial charges are typically overestimated compared to the gas-phase values obtained from high-level quantum mechanical calculations. For decades, the classic force fields have gone through extensive refinements, validations and tests [232, 233]. This generation of force fields is now widely used in the studies of molecular structures, dynamics, and interactions. One challenge faced by such force fields, however, is that the electrostatics is unable to respond to environmental changes including dielectric constant, pH value, or nature of solvent.

It is possible to model the non-additive nature of the polarization effect explicitly. Polarization refers to the redistribution of a particle's electron density due to an external electric field. The idea of explicit treatment of electrostatic polarization dates far back [234]. Only in the last decade or so there has been systematic development of polarizable force fields for biomolecular simulations. Different approaches have been introduced to

incorporate the polarization effect, including induced dipole [235-244], Drude oscillator [245-247] and fluctuating charge models [248-250].

Among these different approaches, fluctuating charge and Drude oscillator are easy to implement within existing fixed-charge force field framework. The induced dipole model requires more complicated algorithms however fits nicely into the atomic multipole framework, which offers more accurate description of electrostatic potential than atomic charge models. Gaussian-based approach improves further the electrostatic representation with additional computational cost. An orthogonal issue is how to compute the polarization on the fly in simulations based on molecular dynamics. Either iterative induction or extended Lagrangian treatment is applicable in most of these approaches [251, 252]. The former can be accelerated with advanced linear algebra solver and the later requires smaller time step and a separate low-temperature thermostat to be stable. The parameterization approaches of these force fields also vary in the degrees that relying on QM decomposition and empirical experimental data.

The AMOEBA (Atomic Multipole Optimized Energetics for Biomolecular Applications) force field, developed by Ponder, Ren et al [242, 243, 253], is an induced dipole model accounts for the polarization effect via atomic dipole induction example. In AMOEBA, the atomic multipole consists of charge, dipole and quadrupole moments, which are derived from the ab initio quantum mechanical calculations using procedures such as Stone's Distributed Multipole Analysis (DMA) [254, 255]. With the vdW parameters determined and transferred from liquid simulations of small organic molecules [256], AMOEBA was applied to the simulations of peptides and proteins. In addition to the non-bonded electrostatic and vdW forces, the "torsional" term also

contributes to a fraction of the conformational energy. The torsion term is essentially an error function in the classical force field, and yet it plays a crucial role in determining the detailed conformational properties of peptides and proteins. The recent development of Amber [257, 258] and CHARMM force fields[259, 260] have demonstrated that the conformational populations of small peptides are extremely sensitive to subtle changes (a fraction of kcal/mol) in the torsional parameters.

In this chapter, the torsion energy in protein backbone was parameterized in AMOEBA polarizable force field. The parameters were firstly derived from high-level ab initio peptide energetics and then adjusted with PDB structural statistics. The optimized parameters were then validated by comparing the simulated J coupling constants with NMR experiments of several peptide and protein systems.

4.2 COMPUTATIONAL DETAILS

The TINKER v6 and Amber v10 molecular modeling packages were used for all the molecular mechanics calculations. Particle-Mesh Ewald summation[261-263] was applied for treating the electrostatic interactions, with a real-space cutoff distance of 7.0 Å, grid spacing of 0.8 Å, and a 5^{th} order polynomial. A cutoff with a switching window (from 10.8 Å to 12.0 Å) at 12.0 Å was applied to the vdW interactions. The induced dipoles, which were also computed with PME, were iterated until the root mean square (RMS) changes were below 0.01 Debye per atom. All the molecular dynamics simulations were performed using an integrator based on Velocity Verlet algorithm.[264] The RESPA algorithm was implemented to enable a 2.5 fs time step. The system temperature was controlled via the Nose-Hoover chain thermostat.[265]

For alanine, glycine, proline and some terminals (Ala-COOH, Gly-COOH, Gly-COO^-), the minimum energy map of the dipeptide was calculated on a uniform 15º grid in the ϕ-ψ space. At each of the 576 points, MP2/6-31G* geometry optimization with constrained ϕ and ψ was performed before a single point energy calculation at the RI-TRIM/MP2 CBS level. In proline, fewer grid points were available for QM calculations due to the limited degree of conformational freedom. For the terminals, the single point energy was also calculated for each optimized structure using the Polarizable Continuum Model (PCM) quantum. The torsion parameters for these model compounds were fit to gas-phase *ab initio* conformational energy f irst and then adjusted based on the statistical populations sampled from Protein Data Bank (PDB). For the side chain torsions of all other residues, geometry optimization was performed at MP2/6-31G* level with the specific torsion angle constrained at every 30º from 0º to 360º, followed by the single point RI-TRIM MP2/CBS energy calculations.

The potential of mean force (PMF) of a solvated alanine tripeptide, NH_3^+-Ala-Ala-Ala-COO^-, with respect to ϕ and ψ of the middle Ala was computed using the two-dimensional weighted histogram analysis method (2D-WHAM). [266-268] A total of 576 (on the same grid as in the gas-phase map) independent molecular dynamics simulations of alanine dipeptide plus 206 water molecules in a 26.6 Å octohedron box was carried out at 298K. In each simulation, the ϕ and ψ dihedral angles were restrained to one of the grid point on Ramachandran map using weak harmonic potentials (force constant = 0.01 kcal/mol-deg^2). The resulting alanine conformer population, sampled from the 576 × 70 ps trajectory (after 30 ps equilibration), was utilized to construct the PMF or the relative free energy map via the 2D WHAM.

The PDB PMF was calculated from –ln(P) where P is the torsion distribution sampled from PDB.[269] For the alanine backbone, the PDB PMF was obtained by averaging the data for alanine with either right or left neighbor residue being alanine (Ala-**Ala**-X or X-**Ala**-Ala, X represents any type of residue, same below) – this choice was made to facilitate comparison with the reference peptides simulated, which were oligo-alanines. For proline, the data with either right or left neighbor residue being glycine (Gly-**Pro**-X or X-**Pro**-Gly) was averaged. Similarly, the glycine PDB PMF was calculated by averaging the data for Pro-**Gly**-X and X-**Gly**-Gly. As for alanine, the glycine and proline PMFs were constructed to make them as comparable as possible to simulations of the peptide GPGG (below).

For peptide systems including unblocked and protonated $(Ala)_5$, and NH_3^+-Gly-Pro-Gly-Gly-COO^-. Replica exchange molecular dynamics (REMD)[28, 270] simulations were performed with 36 replicas at temperatures between 278 K and 620 K. The $(Ala)_5$ was unblocked and protonated at both N- and C-termini, corresponding to the experimental conditions of pH 2.[219, 271] For each system, the peptide was soaked in an octahedron water box with ~800 water molecules. The MD simulations were performed under the NVT ensemble for at least 30 ns/per replica, using the PMEMD module in Amber 10.[124] The snapshots were saved every 0.5 ps for analysis purpose.

4.3 TORSIONAL PARAMETERS

Once the electrostatic, vdW, and valence parameters were determined, the last step was to derive the backbone torsional parameters by comparing AMOEBA and *ab initio* conformational energy values. Note that the molecular mechanics conformational energy not only depends on the torsional energy term, but also the treatment of

nonbonded intramolecular interaction, in particular how the 1-4 interactions are handled. In this work the scaling factors for the intramolecular electrostatic and vdW interactions have been chosen so that they 1) minimize (maximize) the contribution of torsional (nonbonded) terms, and 2) also transfer well from dipeptides to tetrapeptides.

The alanine dipeptide is used to parameterize the backbone torsions for all the amino acids except glycine and proline. The *ab initio* (RI-TRIM MP2/CBS) energy of alanine dipeptide was systematically evaluated at different backbone torsion angles over a 24 × 24 grid (15^{o} interval in both ϕ and ψ) as described in the computational details. The AMOEBA energy without the ϕ/ψ torsional contribution was computed for the same conformation using torsion constraints. The difference between AMOEBA and RI-TRIM MP2/CBS energy is taken as the fitting target of the torsional parameters using a three-term Fourier expansion. Subsequently 2D WHAM simulations of an $(Ala)_3$ peptide in explicit water were performed and the Ramachandran potential of mean force (PMF) of the middle residue were obtained. The torsion parameters are further improved by comparing the AMOEBA PMF to the statistical alanine backbone PMF derived from the PDB database.[269] Note that there torsion parameters were not directly fit to the PDB PMF. Instead, the parameter refinement was achieved by assigning relatively higher weight factors to the QM energy of conformers located at the polyproline II (PII), α-helical and β-sheet regions than the other while fitting to the whole QM gas-phase energy map. The torsion parameters are fine-tuned in 3-4 iterations to balance the simulated relative populations in these minimum-energy regions. The gas phase Ramachandran potential energy from RI-TRIM MP2/CBS and AMOEBA with the final torsion parameters are compared in **Figure 4.1**. The simulated solution-phase PMF using

AMOEBA and the PDB statistical PMF maps for alanine backbone are shown in **Figure 4.2**.

The parameterization of proline backbone torsions followed essentially the same procedure as alanine except that fewer grid points are used due to the limited conformational freedom. For glycine, a torsion-torsion spline term is introduced in addition to the Fourier torsional terms for ϕ and ψ. After the Fourier torsions were fit to the gas-phase *ab initio* RI-TRIM MP2/CBS energy, the difference between the *ab initio* and AMOEBA energy were used as the 2-D spline parameters which were fixed in the subsequent optimization of the Fourier torsion parameters. The use of torsion-torsion term improves the fit to both *ab initio* data and solution phase properties. Similar to the parameterization of alanine backbone torsions, the torsion parameters for proline and glycine backbone were refined to match the statistical PMF of proline and glycine from the PDB database, respectively. REMD simulations were performed with a model tetra-peptide GPGG (NH_3^+-Gly-Pro-Gly-Gly-COO^-) to obtain the simulated torsion angles distribution of proline and glycine backbone. All other residues share the same backbone torsion parameters (together with other valence, vdW and electrostatic parameters) as alanine. The parameterization of the –COOH terminal of alanine (and other non-Gly/Pro residues) and –COOH, -COO^- terminals of glycine were also fit to the RI-TRIM MP2/CBS energies on a 12 × 12 torsional grid Since there are no PDB data available, we have optimized these parameters by matching AMOEBA energy in implicit solvent (Generalized Kirkwood surface area) with QM PCM energies at MP2/6-311G(2d,2p) level.

In addition, conformation energies for a benchmark set of 27 alanine tetrapeptides[272] have been assessed. This comparison was made to validate the transferability and adjusting scaling factors for the short-range intramolecular nonbonded interactions. The dipeptide data itself is obviously not useful for this task as its conformational energy surface has been explicitly fit to. The AMOEBA results are compared with those from MP2, LMP2, DFT and RI MP2 calculations in **Table 4.1**. All the *ab initio* calculations are single point energy evaluations of the same HF/6-31G** geometries. AMOEBA calculations were performed with both full geometry optimization and optimization with ϕ and ψ angles constrained at HF geometry. All comparison is made against RI MP2/CBS results. While the AMOEBA-optimized structures deviate only slightly from those of HF/6-31G** (average srms = 0.47 Å), the rms difference between AMOEBA and RI MP2/CBS energies is 1.15 kcal/mol, similar to those of LMP2/cc-pVTZ(-f) and MP2/6-311+G2d2p. Note that the relative conformational energies of the first two conformers (extended) vs. the third (compact) given by RI MP2/CBS lie in between the canonical MP2 and LMP2 results, and so are the AMOEBA predictions.

4.4 Simulation and Validation

4.4.1 Polyalanine conformational free energy in solution

There has been an increasing number of studies of oligopeptide conformational properties in solution to calibrate force field torsional parameters.[205, 219, 221, 257, 259, 260, 273, 274] Simulated results can be directly compared to the experimental nuclear magnetic resonance (NMR) data for the corresponding peptides. Following the

previous work, we have performed simulations on Ala/Gly/Pro based peptides using the current AMOEBA protein force field.

For alanine, we have first examined the solvation of unblocked and protonated $(Ala)_5$ peptide using REMD. The conformational preference is presented as a potential of mean force with respect to ϕ and ψ in Figure 4b, which is calculated from the averaged ϕ and ψ torsion population distributions of Ala-2, Ala-3, and Ala-4 residues. A distinct global minimum is located around the PII conformation. Two other basins with energies about a half kcal/mol higher are in the β-sheet and α-helix region, respectively. The energy barrier between global and the two local minima is about 1-2 kcal/mol. Overall, the upper left region of the Ramachandran map is distinctively flat compared to the rest of conformational space. The shape and location of this highly populated vicinity agree well with the statistical PMF map from the PDB database (Figure 4c), [269] suggesting the transferability from $(Ala)_3$ to $(Ala)_5$ as we expected.

The distributions of ϕ/ψ torsion angles of $(Ala)_5$ have been probed experimentally by NMR.[219] The NMR spin-spin coupling (*J*-coupling) constants, reflecting the ensemble character of the conformational distribution, were compared with those calculated form REMD simulation trajectories of $(Ala)_5$ via Karplus relations.[205, 275] In total, eight NMR *J*-coupling constants were reported: five for the backbone angle ϕ, $^3J(H_N,H_\alpha)$, $^3J(H_N,C')$, $^3J(H_\alpha,C')$, $^3J(C,C')$, $^3J(H_N,C_\beta)$, two for the backbone angle ψ, $^1J(N,C_\alpha)$, $^2J(N,C_\alpha)$ can be measured, and one for both ϕ and ψ $^3J(H_N,C_\alpha)$.[219] The trajectory at 298 K in the $(Ala)_5$ REMD simulation was extracted to calculated the predicted *J*-coupling values. Twenty-seven predicted *J*-coupling values are compared to those measured by NMR experiments in **Table 4.2**. The *J*-coupling constants involved in

the N- and the C-termini (COOH) were also included. The predicted *J*-coupling values are in excellent agreement with those probed by the experiments. The chi-square (χ^2) difference between the simulations and experiments, computed using the experimental uncertainties,[219] is about 0.994, and the overall RMS difference is 0.33 (**Table 4.2**). Note that when using the torsion parameters that were directly fit to the entire gas-phase QM energy map of alanine dipeptide, the χ^2 is 3 or 4 times higher. The torsion refinement has a significant effect in improving the calculated *J*-coupling constants by assigning relatively higher weight factors at the PII, α-helical and β-sheet regions that obtained from the PDB PMF. A notable consequence of the adjustment is the location of α-helix population from simulations shifted lower and to the right toward the (φ, ψ) angles in the PDB distribution. In contrast, with torsion terms fit to QM gas-phase energy alone, the simulated "α-helix" population was much broader than that of PDB and centered at much lower (more negative) φ and higher (less negative) ψ. A similar effect has recently been discussed for the CHARMM 22/CMAP, CHARMM36-MP2 and CHARMM36 force fields and it was suggested that an empirical correction to CMAP approach is important.[259, 260] Improving the agreement with PDB distribution both in terms of shape and location, especially for residues not in actual helices, led to thermodynamic properties and cooperativity in the helix-coil transition that were more consistent with experiments.

4.4.2 Proline and glycine conformational free energy in solution

The ϕ/ψ torsion angles distribution for proline and glycine backbone were validated via REMD simulations of a tetra-peptide GPGG (NH_3^+-Gly-Pro-Gly-Gly-COO^-). For both proline (Pro-2 residue) and glycine (Gly-3 residue), the simulated PMF maps

with respect to the ϕ/ψ torsion angles show good agreement with the PDB statistical PMF maps (**Figure 4.3**). For Pro-2 residue, the relative free energy of the α-helix and PII regions from the PDB data are well reproduced in our simulation and the local minimum in the α-helix region is about 1 kcal/mole higher than the global minimum in the PII region (**Figure 4.3a**). The torsion distributions of the Gly-3 residue from the simulations are also consistent with the PDB data. The global minima are located at the α-helix and the left-handed α-helix regions. Two local minima are located at the PII and the reflection of the PII regions, with about 1 kcal/mole higher than the global (**Figure 4.3c**).

Similar to alanine, the *J*-coupling constants were calculated for Pro-2 and Gly-3 residues. Three *J*-coupling constants, *J*(H_α,C') for Pro-2, *J*(H_α,H_N) and *J*(H_α,C') for Gly-3 were evaluated by using the Karplus coefficients obtained from B972 EPR-III and B3LYP EPR-III calculations.[276] **Table 4.3** compares the *J*-coupling values obtained from the simulations and experiments for the GPGG tetra-peptide. The RMS difference between the calculated and experimental *J*-coupling constants is 0.44 (with B972) and 0.39 (with B3LYP), respectively.

4.5 CONCLUSIONS

AMOEBA was applied to the simulations of peptides and proteins. In addition to the non-bonded electrostatic and vdW forces, the "torsional" term also contributes to a fraction of the conformational energy. The torsion term is essentially an error function in the classical force field, and yet it plays a crucial role in determining the detailed conformational properties of peptides and proteins. The recent development of Amber [257, 258] and CHARMM force fields[259, 260] have demonstrated that the

conformational populations of small peptides are extremely sensitive to subtle changes (a fraction of kcal/mol) in the torsional parameters. In developing the current force field, we have resorted to both high-level *ab initio* (MP2/CBS) peptide energetics and PDB structural statistics in deriving the backbone torsion parameters. The resulting force field overall performed reasonable well compared with NMR *J* coupling constants of several peptide and protein systems. Nonetheless, these are limited validations focusing on conformational properties and torsional parameters. Extensive investigations on more proteins and a broad range of thermodynamic properties will be necessary to understand the various aspect of the potential energy model and to fully determine the successes and failures of the force field. As previously noted,[259, 277] the CMAP style spline torsion allows a force field to reproduce the gas-phase *ab initio* conformational energy exactly. This however may also pick up unphysical errors in the force fields (e.g. in the other valence contributions) that are not transferable to the solution phase. While we have strived to derive a balanced and physical force field, further understanding of the limitations of the molecular mechanics force fields is essential for systematic improvement.

Table 4.1: Comparison of alanine tetrapeptide conformational energy (kcal/mol). The RMSD was computed using the RI MP2/CBS energies as references.

RI MP2/ CBS	**LMP2/ cc-pVTZ(-f)**	**MP2/ 6-311+G2d2p**	**DFT B3LYP/6-31G***	**AMOEBA**	**Struct. RMS (Å)**	**AMOEBA (ϕ/ψ restrained)**
4.13	2.50	4.61	1.62	3.07	0.30	2.54
4.19	2.60	4.21	1.71	3.62	0.42	0.74
0.57	0.00	-0.70	-1.00	0.00	0.21	0.33
5.73	3.87	5.50	3.61	4.07	0.37	3.82
5.26	3.88	5.14	4.25	3.96	0.30	2.27
2.90	2.19	2.10	2.10	2.45	0.53	0.14
6.67	5.73	5.61	6.56	7.64	0.45	0.65
4.64	4.17	3.32	4.99	5.45	0.44	1.06
7.92	6.93	6.98	5.20	10.01	0.25	3.14
7.79	6.99	6.57	7.24	6.34	0.34	0.62
0.00	-0.19	-1.41	0.14	0.75	0.68	0.58
0.29	0.50	-1.07	1.73	0.75	0.91	0.22
3.66	1.77	3.20	1.14	3.56	0.62	0.02
4.68	3.68	4.14	3.89	4.66	0.71	0.00
2.19	2.07	0.65	3.47	2.28	0.59	0.08
3.55	2.83	2.33	3.31	2.93	0.48	0.24
3.42	2.78	2.02	2.00	2.32	0.28	1.09
1.91	0.52	1.15	-0.87	2.19	0.56	0.20
3.82	2.83	2.90	1.13	4.25	0.56	0.19
1.76	0.87	0.88	0.80	3.18	0.47	2.91
2.92	2.11	1.59	1.78	0.00	0.92	8.51
5.82	4.82	4.59	4.84	6.87	0.59	1.60
5.82	4.82	4.57	4.84	6.84	0.33	1.46
3.98	2.98	2.89	3.59	4.11	0.30	0.19
2.50	1.59	1.54	1.92	2.87	0.50	0.35
0.67	0.18	-0.41	1.40	1.60	0.37	1.51
4.02	3.18	3.04	3.53	6.26	0.38	5.57
RMS deviation	1.05	1.06	1.54	*1.15*	*0.47*	*1.22*

Table 4.2. Comparison of *J*-coupling values (Hz) from the AMOEBA simulations and experiments for $(Ala)_5$ peptide. The trajectory at 298 K was extracted for the *J*-coupling calculation.

Rsidue index	*J*-coupling type	*J*-simulation	*J*-expt.[219]
Ala-2	$^{1}J(N,C_\alpha)$	11.07	11.36
Ala-3	$^{1}J(N,C_\alpha)$	10.92	11.26
Ala-4	$^{1}J(N,C_\alpha)$	10.92	11.25
Ala-2	$^{2}J(N,C_\alpha)$	8.45	9.20
Ala-3	$^{2}J(N,C_\alpha)$	8.17	8.55
Ala-4	$^{2}J(N,C_\alpha)$	8.23	8.40
Ala-5	$^{2}J(N,C_\alpha)$	8.25	8.27
Ala-2	$^{3}J(C',C')$	0.87	0.19
Ala-2	$^{3}J(H_\alpha,C')$	1.73	1.85
Ala-3	$^{3}J(H_\alpha,C')$	1.72	1.86
Ala-4	$^{3}J(H_\alpha,C')$	1.71	1.89
Ala-5	$^{3}J(H_\alpha,C')$	1.93	2.19
Ala-2	$^{3}J(H_N,C')$	1.09	1.13
Ala-4	$^{3}J(H_N,C')$	1.32	1.15
Ala-5	$^{3}J(H_N,C')$	1.22	1.16
Ala-2	$^{3}J(H_N,C_\beta)$	1.82	2.30
Ala-3	$^{3}J(H_N,C_\beta)$	1.83	2.24
Ala-4	$^{3}J(H_N,C_\beta)$	1.74	2.14
Ala-5	$^{3}J(H_N,C_\beta)$	1.58	1.96
Ala-2	$^{3}J(H_N,H_\alpha)$	6.27	5.59
Ala-3	$^{3}J(H_N,H_\alpha)$	5.99	5.74
Ala-4	$^{3}J(H_N,H_\alpha)$	6.08	5.98
Ala-5	$^{3}J(H_N,H_\alpha)$	6.61	6.54
Ala-2	$^{3}J(H_N,C_\alpha)$	0.42	0.67
Ala-3	$^{3}J(H_N,C_\alpha)$	0.61	0.68
Ala-4	$^{3}J(H_N,C_\alpha)$	0.65	0.69
Ala-5	$^{3}J(H_N,C_\alpha)$	0.66	0.73
X^2 = 0.994		**RMS=0.33**	

Table 4.3. Comparison of *J*-coupling values (Hz) from AMOEBA simulations and NMR experiments for GPGG tetra-peptide. The trajectory at 298 K was extracted for *J*-coupling calculation.

Residue index	J-coupling type	*J*-simulation (B972 EPR-III)	*J*-simulation (B3LYP EPR-III)	*J*-expt.[276]
Pro-2	$J(H_\alpha,C')$	1.75	1.88	1.30
Gly-3	$J(H_\alpha,H_N)$	4.94	3.67	4.10
Gly-3	$J(H_\alpha,C')$	6.07	6.76	6.30
		RMS=0.44	RMS=0.39	

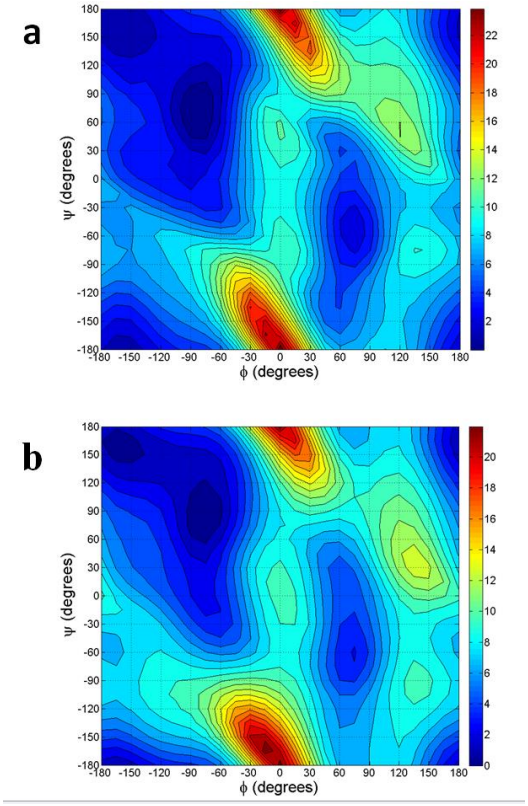


Figure 4.1: Gas-phase energy contours for alanine dipeptide from RI-TRIM MP2/CBS (a) and AMOEBA (b). The energy was computed on a 24 x 24 grid.

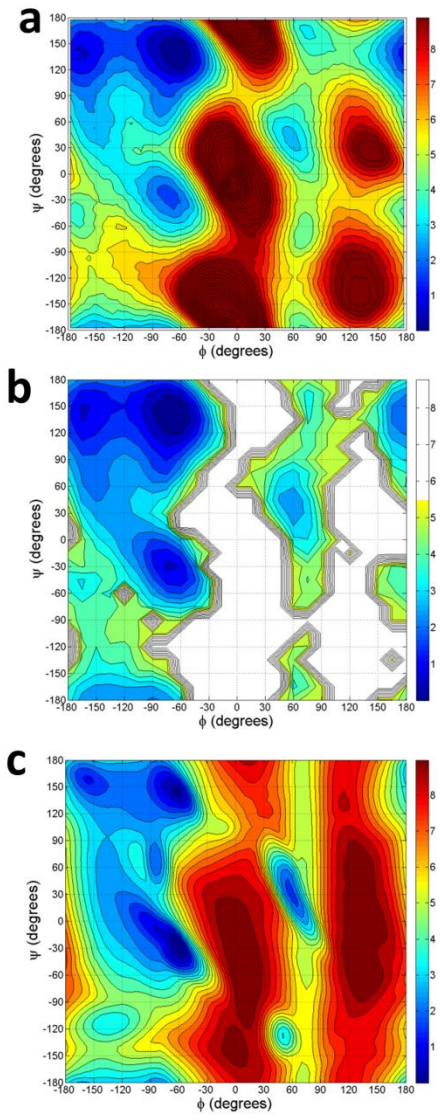


Figure 4.2: Comparison of Ramachandran potential of mean force for alanine. (a) Ala-2 residue of $(Ala)_3$ as predicted by 2D-WHAM simulations. (b) Average of ala-2, ala-3, and ala-4 residues in replica exchange molecular dynamics simulation of the $(Ala)_5$ peptide. The trajectory at 298 K was used. (c) The PDB data are from ref [269].

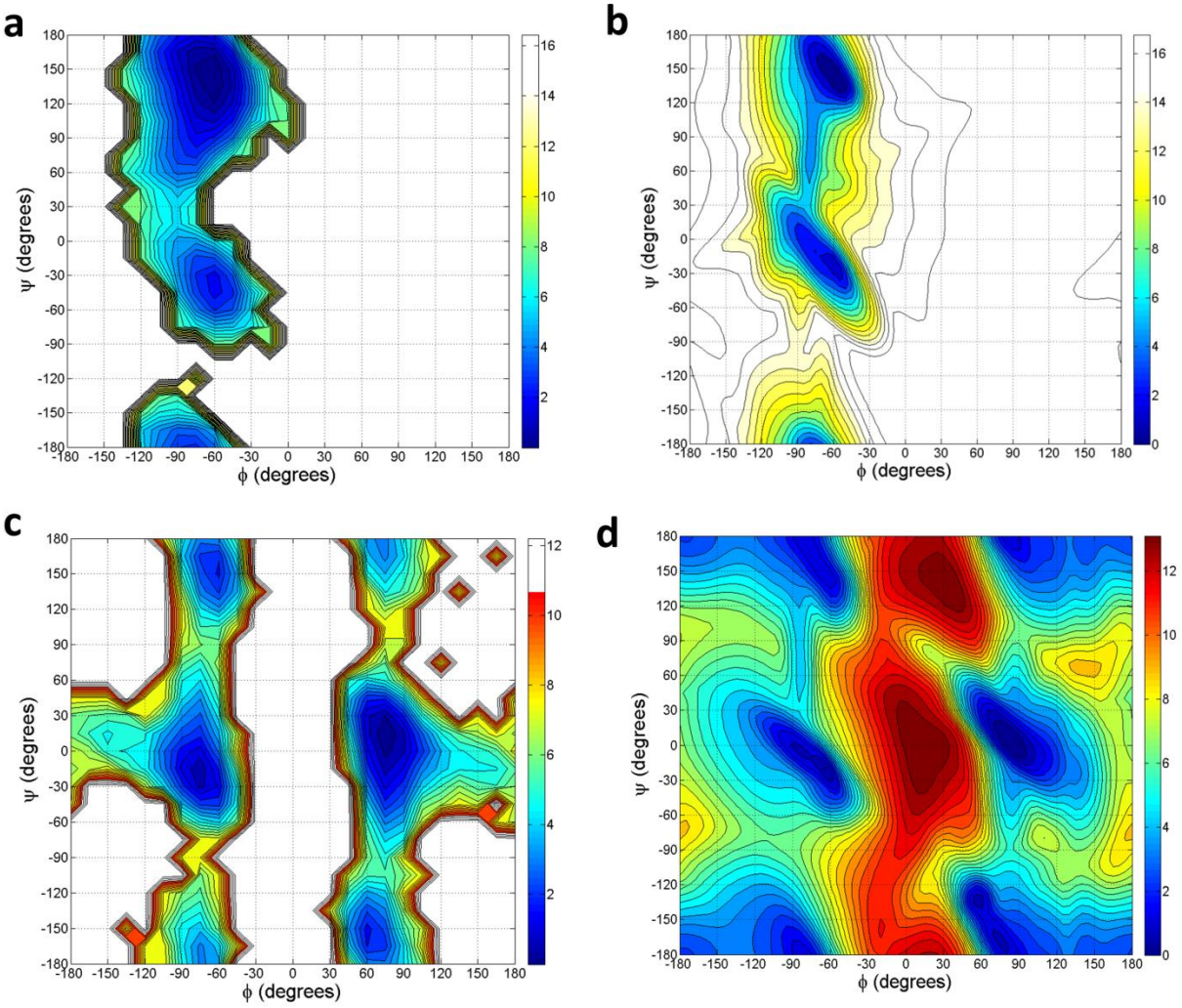


Figure 4.3: Comparison of Ramachandran potential of mean force maps for proline and glycine. (**a**) Pro-2 residue of GPGG from AMOEBA simulations. (b) The PDB data for proline. (c) Gly-3 residue of GPGG from simulations. (d) PDB data for glycine. All the PDB PMF were computed using data from Dunbrack et al. [269] PMFs for glycine have not been symmetrized.

5 Molecular Dynamics Simulations of Ago Silencing Complexes

5.1 INTRODUCTION

In previous three chapters, multi-scale protein and RNA models have been developed for biomolecular simulations. In the following chapters, we applied those developed models to study different biological systems.

One of the important applications is to study protein and microRNA (miRNA) interactions. miRNAs are short RNAs, approximately ~22 nucleotides (nts) in length, which post-transcriptionally regulate their protein-coding targets in a sequence-dependent manner[278]. The typical transcribed precursor molecule (pri-miRNA) assumes a double-stranded RNA (dsRNA) conformation with a characteristic hairpin-like structure; the latter is pre-processed in the nucleus by the "microprocessor" complex into a pre-miRNA before being shuttled into the cytoplasm by XPO5[279]. Mirtrons represent an exception to this rule: their pre-miRNAs are directly generated from introns during the splicing of nascent mRNAs[280]. In the cytoplasm, the pre-miRNA hairpin is cleaved by the DICER endonuclease to form dsRNA, 20-25 nts in length, with 3′ overhangs; one of the two strands, the miRNA, is incorporated into the RISC, also known as the miRNA ribonucleoprotein complex (miRNP), where the interaction with the target mRNA takes place[278]. This interaction typically results in degradation of the target mRNA or inhibition of its translation by the ribosome[281]. Originally believed to form heteroduplexes mainly with the 3′ untranslated region (3′UTR) of the target, miRNAs have since been shown to also target protein-coding regions and 5′UTRs[282-289].

MiRNAs have been shown to be involved in many fundamental processes that include developmental timing[290-293], the induction of organ asymmetry[294], tumor

suppression and oncogenic activity[295-297], invasion and metastasis[298], modulation of embryonic stem cell differentiation[286], neurodegeneration[299] etc. Moreover, extensive work has revealed tissue- and cellular-state-dependent miRNA profiles in cancers[300-307], cardiovascular disease[308, 309], immunity[310], Alzheimer's[311, 312], Tourette's syndrome[313], schizophrenia[314], and others. Given their involvement in such diverse contexts, and their ever-increasing numbers[315], understanding the identity and cardinality of a given miRNA's targets represents an important endeavor.

Ever since the first report on the lin-4:lin-14 heteroduplex[291, 293], it was clear that the 5′ region of a miRNA played a central role in the recognition of its target. This region, spanning positions 2-7 from the miRNA's 5′ end, was originally referred to as the 'core element' but eventually became known as the 'seed.' Its apparent importance has been a central component of many computer-based miRNA target prediction schemes[316-323] whereby the presence of the reverse complement of a miRNA's seed sequence is used as a filter before generating lists of candidate mRNA targets. Methods that do not rely on such filtering schemes have also been in use[283, 324]. All computational attempts to tackle the problem have met with various degrees of success and for all practical purposes the problem remains open[325, 326].

Some of the very early work[290, 293] indicated that formation of functioning heteroduplexes did not require a strict reverse-complementarity relationship between the miRNA-seed sequence and its target. Since then, evidence has continued to accumulate steadily in support of such an 'expanded' interaction mode[286, 327-340], in turn suggesting a potential repertoire of targets for a given miRNA that is larger than the 'seed'-reverse-complementarity constraint might suggest. We revisit this very question

with the help of molecular dynamics (MD) simulations and the recently released crystal structure of ternary complexes of eubacterial Thermus thermophilus Ago (TtAgo)[341, 342].

5.2 METHODS FOR MOLECULAR DYNAMICS SIMULATIONS

Following similar protocols as in our previous studies[343-346], the X-ray crystal structure of wild-type *Tt*Ago bound to a 21-nt guide DNA and a 20-nt target RNA complexes (PDB entry: 3F73, released in 2008.12) was used as the starting structure for the MD simulation[342]. The DNA and the RNA strands can only be partly traced from position 1 to 11, and the base coordinates at position 10 and 11 are not available in the reported crystal structure[342]. Therefore, the missing coordinates at position 10 and 11 were built from the known backbone structures. We repeated the simulations using the more recently released Ago complexes with longer traceable guide-target duplex (length of the duplex is 15-bp, positions 2-16, PDB entry: 3HK2, released in 2009.10)[347]. The Ago:miRNA:mRNA complexes were generated by replacing the bases of the guide DNA by corresponding RNA (deoxyribose was replaced by ribose in A, C, and G whereas T was replaced by U). All the Ago complexes were solvated in ~110×100×90Å^3 water boxes. A total of 32 Na^+ ions and 29 Cl^- ions were added to neutralize and mimic the biological environment (100 mM NaCl concentration). The solvated systems contain approximately 100,000 atoms. We utilized the NAMD2[348] package for the MD simulations with the NPT ensemble. The CHARMM (parameter set c32b1) force field was used for the protein and nucleic acid[129, 349], and the TIP3P water model was used as the explicit solvent[350]. The Particle Mesh Ewald (PME) method[351] as applied to treat the long-range electrostatic interactions and a 12 Å cutoff was employed for the van

der Waals interactions. All the Ago complexes systems were equilibrated via a 20,000-step energy minimization to remove bad contacts. The minimized configurations were used as the starting point for 1-ns NPT MD equilibrations with 0.5 fs time-step at 1 atm and 310 K. The equilibrated configurations were then subjected to production runs for a minimum of 100 ns. The time step for all production runs was 1.5 fs with SHAKE/RATTLE algorithm[352].

A newer version of the CHARMM force field for RNA parameters, which fixes the too much base-pair opening [353], was brought to our attention after we finished aforementioned simulations. To further validate our findings, we have repeated our simulations for the wild-type as well as the 4-site mismatch mutant with this latest RNA parameter set (CHARMM c36 parameter set). Three independent runs were performed for both the wild-type and 4-site mismatch mutant for at least 100 ns each.

5.3 RESULTS

*Tt*Ago is considered an appropriate model for studying the properties of Ago complexes thanks to the high structural and functional similarities with the eukaryotic Ago families. For this study, we introduced selected modifications to the currently available Ago-*DNA*:mRNA co-crystal structure. The collection of heteroduplexes to simulate was informed by previously reported non-canonical examples[283, 286, 290, 291, 293, 329, 339] and includes heteroduplexes with a) G:U wobbles in the seed in conjunction with adjacent Watson-Crick pairs; b) G:U wobbles in the seed but without adjacent Watson-Crick pairs; c) a single bulge on the *target* side (mRNA) at each of several different seed locations; and, d) a single bulge on the *guide* side (miRNA) at each of several different seed locations. To build our Ago-miRNA:mRNA ternary complexes

we replaced the bases of the guide DNA by the corresponding RNA (miRNA). Each MD simulation spanned a minimum of 100 ns, generating a trajectory that is sufficiently long for the current analysis of the mRNA recognition dynamics and also for observing any potential conformational changes (discussed further below).

5.3.1 The Ago-complex is stable in the presence of multiple seed-region G:U wobbles

In earlier *in vivo* studies, the impact of G:U wobbles in the seed region (positions 2-7) was examined in *D. melanogaster*[354] and in *C. elegans*[329] and led to different conclusions. The fruit-fly study examined the impact of one, two and three G:U wobbles and concluded that "[…] a G:U wobble in the seed region is always detrimental […]"[354]. On the other hand, the worm study examined the impact of one and two wobbles in the seed region and found the mutants to still be functionally regulated by the targeting miRNA lsy-6[329]. In addition to these two studies, luciferase assays were used to demonstrate the validity of functional heteroduplexes involving as many as five G:U wobbles in the seed[283]. For our studies, we created a first mutant with three G:U wobbles at positions 2, 3 and 4 of the seed region (Mutant #1 in **Table 5.1**). In a second experiment, we added a fourth G:U wobble at position 5 of the seed region (Mutant #2). Our analysis shows that in both configurations the resulting structures are stable and their Ago backbones compared to the backbone of the native Ago ternary complex exhibit small RMSD values (~2-3Å) (**see Figure 5.1c, Figure 5.2a, and Figure 5.3**). In the mutants, study of position 6 or 7 of the seed region reveals that the conformations of Watson-Crick pairs remain intact between guide strand and the target strand (**Figure 5.4**).

5.3.2 The Ago-complex is stable in the presence of multiple seed-region G:U wobbles and no compensating Watson-Crick pairs immediately adjacent to the seed

As can be seen from **Figure 5.1b** and **Table 5.1**, the heteroduplexes of Mutants #1 and #2 contain two Watson-Crick pairs immediately past the seed region, at positions 8 and 9. In order to determine the extent to which these two base pairs play a compensatory role that contributes to the stability of the complex we removed both and repeated the previous simulations (Mutant #1 → Mutant #3, Mutant #2 → Mutant #4). The two resulting heteroduplexes, were they stable, would rely primarily on coupling that spans the seed region and is rooted in the present of three- (case of Mutant #3) and four- (case of Mutant #4) G:U wobbles respectively. Our simulations show that these mutants are indeed stable: the RMSD values of the Ago backbones from the wild-type remain low and reach a plateau of ~2.5Å after only ~40 ns (**see Figure 5.1c, Figure 5.2b, and Figure 5.3**). This indicates that the Watson-Crick pairs already present in the seed region (at positions 6 and 7) are sufficient for maintaining the overall stability of the heteroduplex – see also base pair distances in **Figure 5.4.**

5.3.3 The Ago-complex is stable in the presence of only partial seed-region coupling and no compensating Watson-Crick pairs immediately adjacent to it

The observed stability of the ternary complex in the presence of multiple G:U wobbles and without any compensating Watson-Crick pairs adjacent to the seed prompted us to also examine a somewhat extreme situation. In particular, we mutated the miRNA's adenosine at position 7 of the seed to a cytosine, thus "breaking" the base pairing at that location (Mutant #5) – shown in cyan in **Table 5.1**. The resulting heteroduplex, if realized, would be brought about by only five base pairs in the seed region, with four of them being G:U wobbles, and without any compensating Watson-

Crick pairs beyond it. Interestingly, and somewhat surprisingly, we found that this arrangement also leads to a stable structure. In fact, the resulting RMSD is only slightly larger than the wild-type arrangement, remaining well below 3Å for the length of the simulation (green curve of **Figure 5.2b**).

5.3.4 The Ago-complex is stable in the presence of a seed-region bulge on the messenger–RNA–side of the heteroduplex

For let-7, the second miRNA ever reported, it was shown that it regulates the heterochronic gene lin-41 by binding to two locations of lin-41's 3′UTR[292]. These two target locations, referred to as LCS1 and LCS2, were later demonstrated *in vivo* to be simultaneously required for lin-41's regulation[339]. For the purpose of this discussion, the heteroduplex formed between let-7 and LCS1 contains a bulge on the lin-41 (mRNA) side between positions 4 and 5 of the seed region[292, 339]. Subsequently, examples of functioning heteroduplexes comprising messenger-RNA-side bulges in the seed region were reported and validated for mouse *Oct4* (between seed positions 4 and 5) and mouse *Sox2* (between seed positions 5 and 6), and concomitant physiological effects were shown for these heteroduplexes[286]. We thus sought to investigate the impact on the stability of the Ago complex that a bulge located on the target-side (mRNA) might have as a function of the bulge's actual location within the seed. Notably, in these experiments we maintained the three G:U wobbles that were previously introduced in the seed region and removed the two Watson-Crick pairs that were originally adjacent to the seed at positions 8 and 9; arguably this generates a rather demanding context for the complex stability in our simulation study. We investigated three bulge placements in the seed region: between seed positions 6 and 5 (Mutant #6), between seed positions 5 and 4 (Mutant #7), and, finally, between seed positions 4 and 3 (Mutant #8). We found that

all three placements of the bulge generate stable structures, with a slight dependence on the actual position of the bulge within the seed's span. The resulting RMSD from the wild-type arrangement is small and ranges between 2 and 3Å (**Figure 5.2c**). These findings demonstrate that more extreme and challenging scenarios than the one reported very recently[340] are also possible.

5.3.5 The Ago-complex stability is affected minimally by a miRNA–side bulge in the seed region

In one of the very early publications on miRNA-driven RNA interference[290] it was shown that bulged lin-4:lin-14 heteroduplexes, with the bulge being on the side of the targeting miRNA, at position 6 of the seed, were functional and sufficient for lin-14 temporal gradient formation in *C. elegans*. More recently, similarly bulged interactions were shown for mouse miRNA:mRNA heteroduplexes[283, 286]. In this series of experiments we investigated the impact on the stability of the Ago-complex of a single bulge that is increasingly closer to the 5′ end of the miRNA at seed positions 6, 5 and 4 (Mutants #9, #10 and #11, respectively). Just as before, we maintained in all experiments the three G:U wobbles introduced in the seed region thus creating an extreme context for our simulation study. We also preserved a single Watson-Crick base pair immediately adjacent to the seed, at position 8. In all three cases, the resulting RMSD values were similar to what we observed prior to having introduced the miRNA-side bulge (**Figure 5.2d**). Also, there was some distortion of individual base pairs when the bulge was placed at position 4 of the seed (Mutant #11) – see **Figure 5.1c**. Bulge placements at seed positions 6 and 4 (Mutant #9 and Mutant #11, respectively) led to slightly larger RMSD values (~4 Å). Placement of the bulge at position 5 (Mutant #10)

exhibited smaller structural deviations from the native structure for both the Ago protein and the RNA heteroduplex compared to the other two placements (**Figure 5.2d**).

5.3.6 Disruptive mutations lead to a large bending motion of PAZ domain and a subsequent opening of the nucleic-acid-binding channel

We also examined whether our 100 ns simulations are long enough to capture large Ago-complex motions, as would be the case when attempting to simulate unsuitable, disruptive mutations. In order to address this question, we introduced several G→C mutations in the seed region aimed at "disrupting" the structure of the complex. Each G→C mutation broke a triple bond and led to non-bonded bases between the guide (miRNA) strand and the target (mRNA) strand. The first mutation we introduced broke the G-C bond at position 8 immediately adjacent to the seed (Mutant #12). Three more mutations gradually increased the number of non-bonded bases *inside* the seed region from one (Mutant #13) to two (Mutant #14) to three (Mutant #15), while maintaining the mutation at position 8 – see Methods and **Table 5.2** for details. Not surprisingly, as the number of non-bonded bases *increased*, the stability of the miRNA-mRNA heteroduplex decreased. The average RMSD of each nucleotide in the mRNA strand increased significantly and in proportion to the number of mismatches. The comparable structural stability of the wild-type and mutants with a single non-bonded base supports earlier experimental work showing that a single nucleotide mismatch at the seed region only slightly reduces the cleavage activity of Ago complexes[342]. On the contrary, for Mutant #15 (which contains four G-C disruptions) a mere ~10 ns of simulation sufficed to disrupt most of the base pairing and base stacking, even for the canonical Watson-Crick base pairs. The severe distortion of the backbone in the guide-target duplex caused the overall "decoupling" of the miRNA-mRNA heteroduplex, which

in turn indicates that nucleation at the seed region cannot be achieved in the mutant with the four G-C-disruptions. The final snapshot of the wild type and of the four-G-C-disruption mutant is shown in **Figure 5.5a**. We observed significant conformational changes or "unfolding" of the complex both in the three-and the four-non-bonded-position mutants (Mutants #14 and #15). The overall RMSDs increased to more than 6 Å at 100 ns in the case of Mutant #15. To further investigate which region of the Ago protein is responsible, we performed a principal component analysis (PCA) for all the trajectories and then used the 1st and 2nd principal components to characterize the major domain motions. The domain motion analysis was carried out at http://fizz.cmp.uea.ac.uk/dyndom/.

No significant single domain motion was observed for the Ago protein in the wild-type complex. However, large motions of the PAZ domain were observed with Mutant #15 (four-non-bonded seed positions). The 1st principal component showed the PAZ domain bending away from the N domain (rotated by 55.7˚ and translated by 1.7Å, as shown in **Figure 5.5b**). The 2nd principal component also showed that the major motion was in the PAZ domain, which further moved the PAZ domain far away from the PIWI-containing domain (rotated by 42.5˚ and translated by 5.3Å, shown **in Figure 5.5c**). Notably, the native structure of Ago complexes shows no direct contact or interaction between the seed region of the guide-target strands and the PAZ domain.

The domain motion analysis also revealed that the bending was due to the residues 172-174 and 268-279 in both 1st and 2nd principal components (the residue numbering follows the structure of TtAgo, PDB entry: 3F73). These two segments correspond to segments L1 and L2 that connect PAZ to the other domains. We found that

the segments of L1 and L2 closest to PAZ act as a "hinge" that determines the orientation of the PAZ domain (**Figure 5.5a**). In Mutant #15, the strong distortion in the target mRNA and positions 6-10 of the guide miRNA provide enough "room" for the L1 and L2 coils to bend. The final trajectory snapshot shows part of the L1 and L2 segments rotated ~30˚ and ~90˚, respectively (**Figure 5.5a**). These rotations in the "hinge" region finally induced the motions of entire PAZ domain, thus largely opened the nucleic-acid-binding channel between the PAZ- and PIWI-containing lobes.

5.3.7 Further validation with the latest version of RNA force field parameters

A new version of RNA parameters in CHARMM force field is available very recently after we performed all the aforementioned simulations [353]. The original problem of too much Watson-Crick basepair opening was corrected in this latest CHARMM force field (parameter set c36). In order to further validate our current findings, we repeated our MD simulations for the wild-type and the 4-site mismatch mutant using the latest RNA parameters. We found very similar results as those from the original force field (c32b1 parameter set), where again small RMSDs (1.5 Å) and stable A-form heteroduplex were seen in the wild-type, but larger fluctuations (RMSD > 3 Å) and significant seed distortions were observed in the 4-site mismatch mutant (**Figure 5.6**).

5.4 DISCUSSION

We have presented a series of molecular dynamics simulations on Ago ternary complexes that focused on investigating the influence of seed-located wobbles, bulges and combinations thereof on the structural stability of the Ago-miRNA:mRNA complex and the motion of its domains, and, by extension its ability to cleave its target. We found

that introduction of multiple G:U wobbles in the seed region only minimally affects the miRNA-mRNA heteroduplex and does not compromise the stability of the complex. With regard to bulge insertions in the seed region, and for a variety of possible arrangements, we find that they are tolerated on both the miRNA and the mRNA sides. Seed-region bulges that occur on the miRNA side of the heteroduplex give rise to slight distortions in the nucleic acid duplex and induce somewhat larger conformational changes but do not disrupt the complex. Seed-region bulges that occur on the mRNA side appear to be better tolerated by comparison. We also find that arrangements involving simultaneously multiple G:U wobbles and a single bulge lead to stable structures as well. Moreover, we examined the impact of artificially introduced disruptive mutations to the seed region and found a novel recognition mechanism that involves an important bending motion of the PAZ domain along the L1/L2 'hinge' link followed by the opening of the nucleic-acid-binding channel.

Our analyses provide additional evidence in support of and are consistent with earlier work that emphasized the importance of strong base-pairing interactions spanning positions 2 through 7 of a miRNA, or a subset of those positions (e.g. *in vivo* examples involving lin-4 and let-7 comprising seed region bulges). However, it is important to realize that as our molecular dynamics analyses show such strong interactions can be realized in a multitude of ways that obviate the requirement that the exact reverse complement of the miRNA's seed sequence be present in the target. In turn, this suggests that a given miRNA can give rise to non-canonical functioning heteroduplexes with targets that do not contain the miRNA seed. Taken together, these findings indicate that the spectrum of potential targets for a miRNA can admit a wide-spectrum of seed-less

targets and thus substantially differs than what is suggested by the canonical seed model. Consequently, our findings indicate that similar conclusions can be drawn about the potential spectrum of a given siRNA's targets, considering that user-designed siRNAs and miRNAs share the same pathway downstream of the DICER cleavage. In other words, it follows that those mRNAs harboring sequences that are proximal to the seed of the transfected siRNA, either because they would induce G:U wobbles or the introduction of a bulge in the seed region, could also be down-regulated by the siRNA.

Table 5.1: Sequences of the 11-nt guide miRNA and target mRNA heteroduplex used in the simulation.

3 and 4 G:U wobbles in seed	Mutant #1 5′–UCACUACUUUG–3′ 3′–GAUGAUGGGGU–5′ 10987654321	Mutant #2 5′–UCACUAUUUUG–3′ 3′–GAUGAUGGGGU–5′ 10987654321	Wild-Type (mRNA) 5′–UCACUACCUCG–3′ Target 3′–GAUGAUGGAGU–5′ Guide 10987654321 (miRNA)
3, 4 and 5 G:U wobbles in seed with no adjacent Watson-Crick pairs	Mutant #3 5′–UCACUACUUUG–3′ 3′–GAGUAUGGGGU–5′ 10987654321	Mutant #4 5′–UCACUAUUUUG–3′ 3′–GAGUAUGGGGGU–5′ 10987654321	Mutant #5 5′–UCACUAUUUUG–3′ 3′–GAGUCUGGGGU–5′ 10987654321
bulge on mRNA side at different seed positions	Mutant #6 U 5′–UCACUA CUUUG–3′ 3′–GAGUAU-GGGGU–5′ 109876-54321	Mutant #7 G 5′–UCACUAC UUUG–3′ 3′–GAGUAUG-GGGU–5′ 1098765-4321	Mutant #8 G 5′–UCACUACU UUG–3′ 3′–GAGUAUGG-GGU–5′ 10987654-321
bulge on miRNA side at different seed positions	Mutant #9 5′–UCACUA-CUUUG–3′ 3′–GAGUAU GGGGU–5′ A 210987 54321 6	Mutant #10 5′–UCACUAC-UUUG–3′ 3′–GAGUAUG GGGU–5′ C 2109876 4321 5	Mutant #11 5′–UCACUACU-UUG–3′ 3′–GAGUAUGG GGU–5′ C 21098765 321 4

Table 5.2: Sequence of the 11-nt and 15-nt miRNA guides and mRNA targets heteroduplex that are used in the simulations.

Wild-type	Target 5′–UCACUACCUCG–3′ Guide 3′–GAUGAUGGAGU–5′ 10987654321	5′–ACAACCUACUACCUCG–3′ 3′–UGUUGGAUGAUGGAGU–5′ 6543210987654321
G → C mutation at position 8	Mutant #12 5′–UCACUACCUCG–3′ 3′–GAUCAUGGAGU–5′ 10987654321	Mutant #16 5′–ACAACCUACUACCUCG–3′ 3′–UGUUGGAUCAUGGAGU–5′ 6543210987654321
G → C mutations at position 5 and 8	Mutant #13 5′–UCACUACCUCG–3′ 3′–GAUCAUCGAGU–5′ 10987654321	Mutant #17 5′–ACAACCUACUACCUCG–3′ 3′–UGUUGGAUCAUCGAGU–5′ 6543210987654321
G → C mutations at position 4, 5 and 8	Mutant #14 5′–UCACUACCUCG–3′ 3′–GAUCAUCCAGU–5′ 10987654321	Mutant #18 5′–ACAACCUACUACCUCG–3′ 3′–UGUUGGAUCAUCCAGU–5′ 6543210987654321
G → C mutations at position 2, 4, 5 and 8	Mutant #15 5′–UCACUACCUCG–3′ 3′–GAUCAUCCACU–5′ 10987654321	Mutant #19 5′–ACAACCUACUACCUCG–3′ 3′–UGUUGGAUCAUCCACU–5′ 6543210987654321

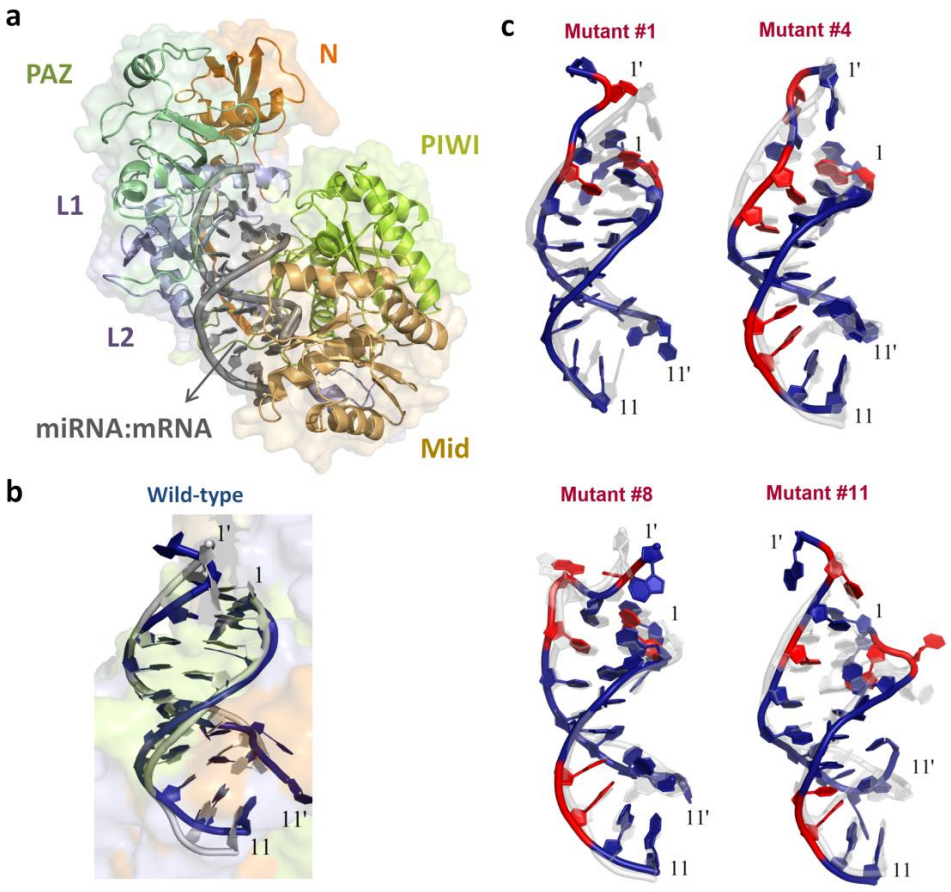


Figure 5.1: Structural views of an 11-nt guide (miRNA) and target (mRNA) heteroduplex for the wild-type and mutants during the simulation. (a) The overall structure of *Tt*Ago-miRNA:mRNA complexes. (b) The structure of the guide-target heteroduplex for the wild-type during the 100-ns molecular dynamics simulation. The conformational change is shown by superimposing the final snapshot (shown in blue) to the starting native structure (shown in gray). (c) The structure of selected mutants in simulations. The conformational changes of the miRNA:mRNA heteroduplex are shown by superimposing the final snapshot (mutated sites are indicated in red) to the starting native structure (colored in light gray) with the ribose and the base shown as plates. Primed (′) numbers indicates bases that belong to the target strand.

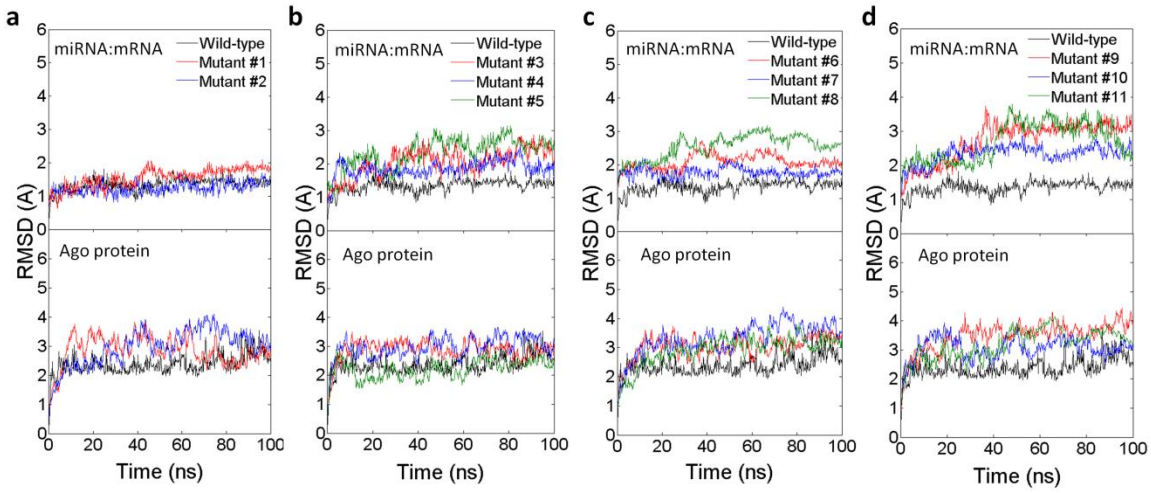


Figure 5.2: Comparison of the structural variations during the simulations for the wild-type and the eleven mutants. (a) Mutants with G:U wobbles in the seed and adjacent Watson-Crick pairs; (b) mutants with G:U wobbles in the seed and with no adjacent Watson-Crick pairs; (c) mutants with one bulge on the target (mRNA) side at different seed positions; (d) mutants with one bulge on the guide (miRNA) side at different seed positions. The plot shows the RMSD values of the miRNA:mRNA heteroduplex (subplot on top) and Ago protein (subplot at bottom) in the ternary complexes.

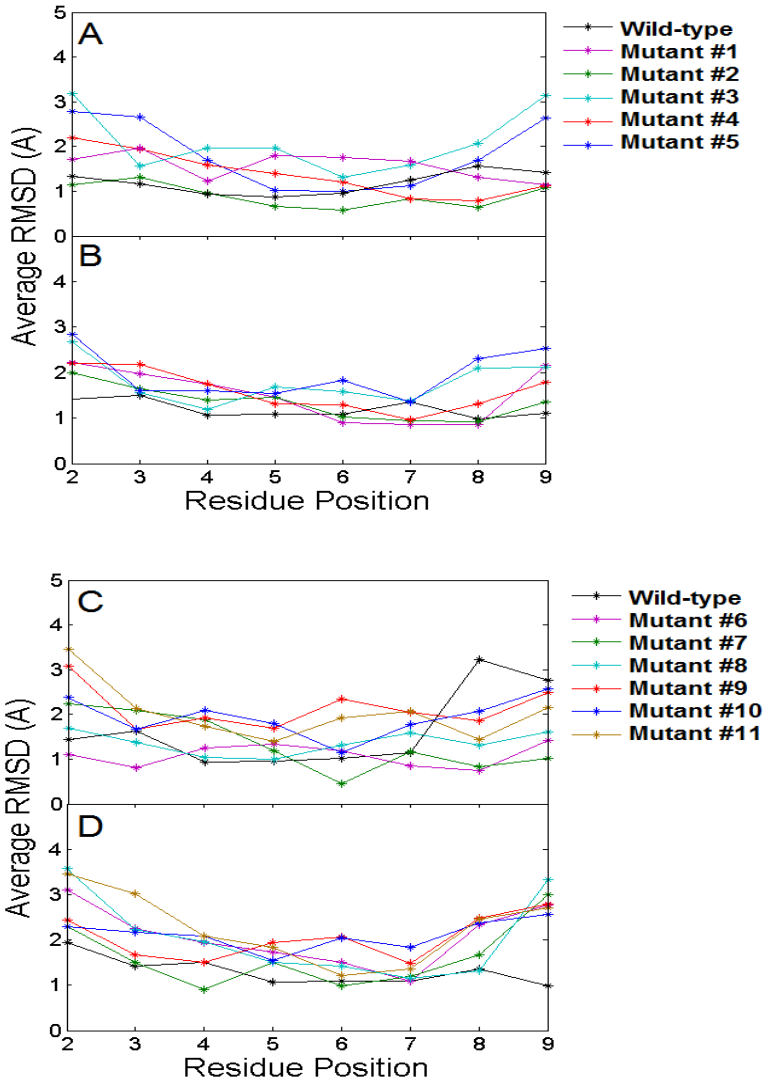


Figure 5.3: Comparison of the average RMSDs. Comparison of the average RMSD values in guide miRNA (a and c) and target mRNA (b and d) from the starting native structures of the wild-type and 11 mutants with 11-nt nucleic acid heteroduplex.

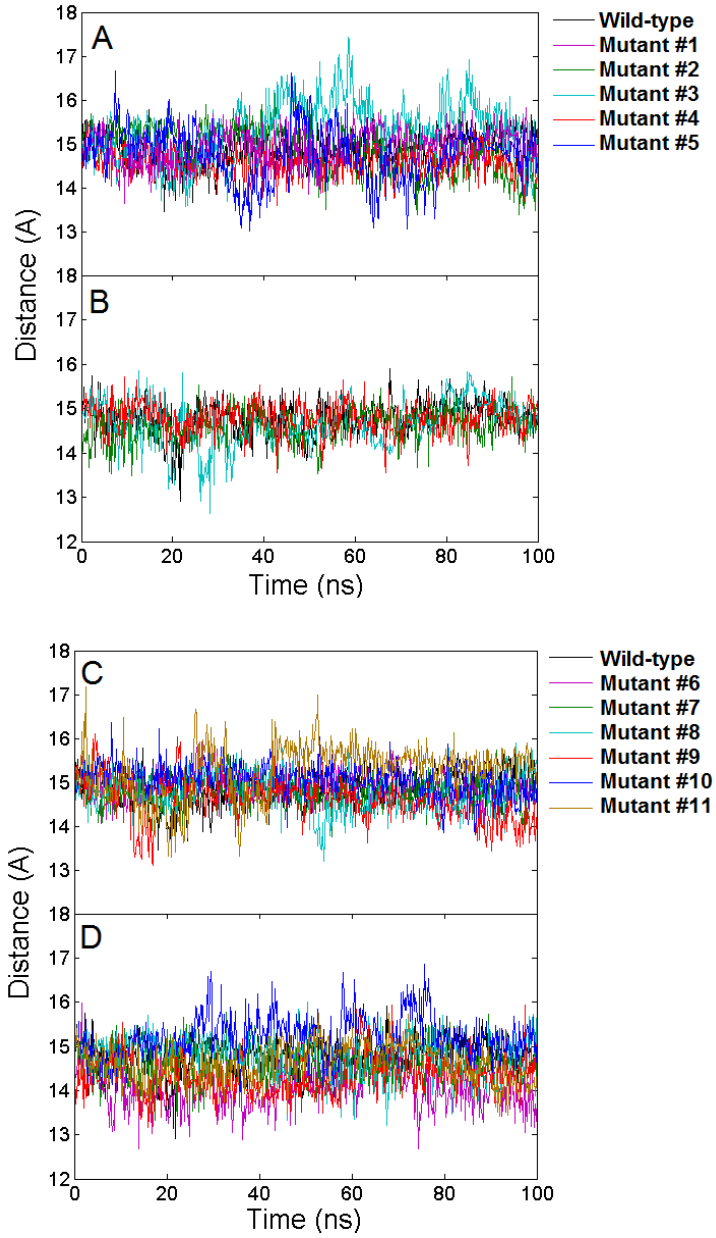


Figure 5.4: The dynamic distance of base pairs for five mutants and the 11-nt heteroduplex. Distances are calculated from the C4′–C4′ atoms between the guide strand and the target strand at the same position. (a) and (c). Distance of A-T base pair at position 6; (b) and (d). Distance of U-A base pair at position 7.

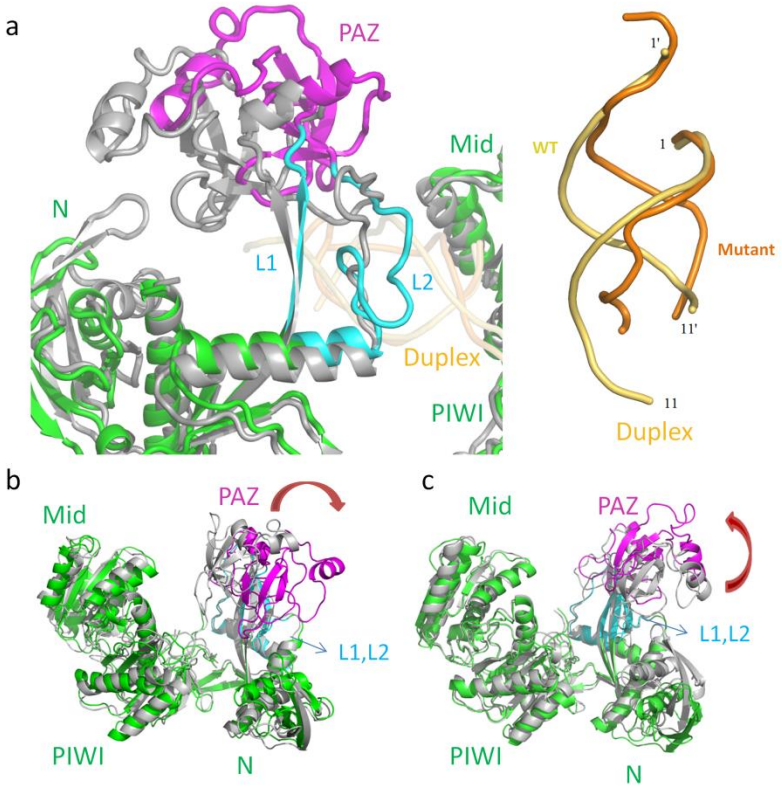


Figure 5.5: Structural views of the guide-target heteroduplex distortion and the domain motions of Ago protein with extreme disruptive mutations. (a) The disassociation of the "hinge-like" L1/L2 segment and the nucleic acid heteroduplex in Mutant #15. (b) and (c) Structural view of the domain motions in the four-G-C-disruptions mutant. Two structures (one colored light gray and the other colored green) are picked from a 100-ns trajectory for each by the principal component analysis (PCA) and the domain motion analysis. The 1st principal component (b) and the 2nd principal component (c) are shown.

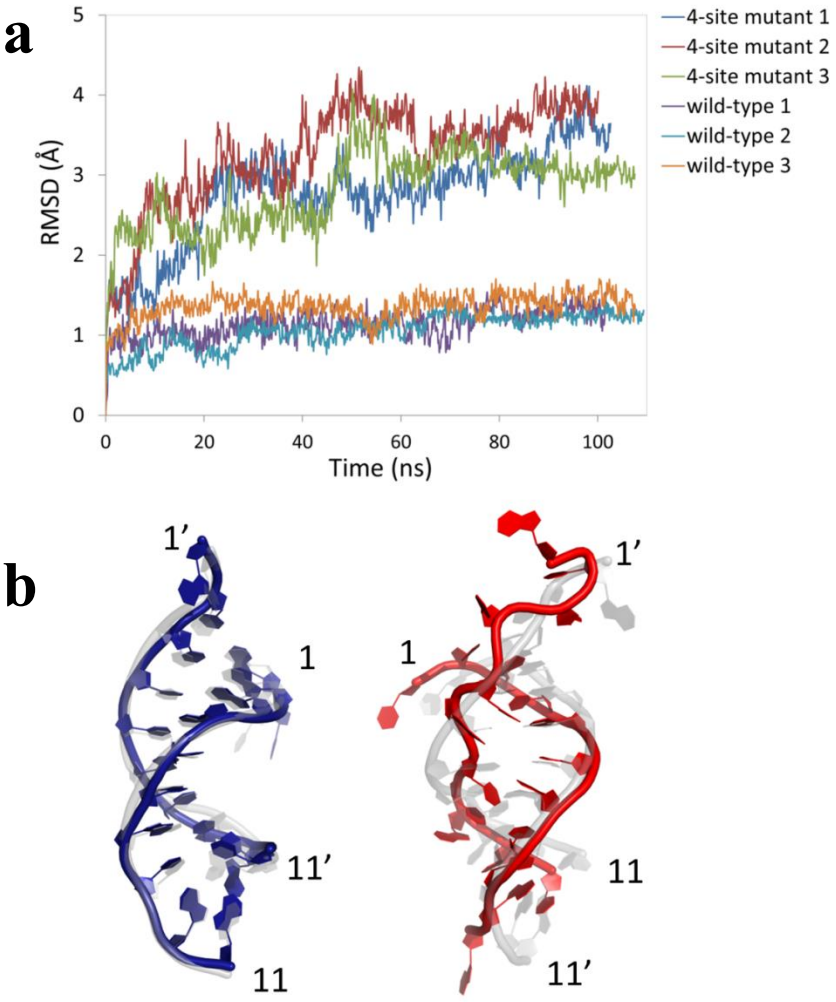


Figure 5.6: Time evolution of the backbone RMSDs of the wild-type and 4-site mismatch mutants from their starting structures. These simulations were performed with new CHARMM force field parameters (set C36) for RNA. (a). RMSDs of the DNA-mRNA heteroduplex in Ago complexes; (b). Superposition of the final snapshot (colored in blue for the wild-type in the left panel and red for the 4-position mismatch mutant in the right) and the starting native structure (colored in light grey) for both the wild-type and the 4-site mismatch mutant.

6 Free Energy Simulations Reveal Important Interactions to Influenza Hemagglutinin Antibody Binding

6.1 INTRODUCTION

Influenza A virus is one of the most fatal infectious diseases in human and poultry [355, 356]. This virus causes deaths of hundreds of thousands in human and tens of millions in avian worldwide each year [357-363]. The recent H1N1 swine flu pandemic and the spread of highly pathogenic H5N1 avian flu have caused a great public health concern [364-372]. Current vaccines usually respond to a limited number of strains and often fail to neutralize emerging new strains because of the rapid genetic evolution [373-378]. Although some neuraminidase (NA) inhibitors, such as oseltamivir (Tamiflu) and zanamivir (Relenza), have shown effective suppression against influenza viruses, their efficacy is often challenged by the rising drug resistances [379-386].

Broad and potent cross-protective host immunity, therefore, is in great demand for influenza prevention and treatment. High affinity neutralizing antibodies (nAbs), such as F10 and CR6261, have been selected by phage display on recombinant H5 HA [367-369]. Both F10 and CR6261 have shown broad neutralization to all Group 1 subtypes, which include the H1N1 'Spanish flu' and the H5N1 'bird flu'. In general, 16 hemagglutinin subtypes of influenza A viruses are categorized into two major phylogenetic groups: Group 1 (H1, H2, H5, H6, H8, H9, H11, H12, H13, and H16) and Group 2 (H3, H4, H7, H10, H14, and H15). The cocrystal structures of H5 HA with antibodies reveal that both antibodies can block the viral infection by inserting their heavy chain into a conserved helical stem region in HA1 and HA2, thereby possibly preventing the membrane fusion [367-369]. We believe that the similar neutralization effect of these nAbs could result from some important common features critical for

developing new vaccines of broad-spectrum. However, several underlying key mechanisms remain unclear. It still needs to be addressed for questions like 'what, at molecular level, is unique for the antibodies to have a broad range of neutralization', and 'why most of Group 2 (i.e. H3 and H7) influenza viruses cannot be neutralized by F10 or CR6261'. To answer these questions, we performed large scale free energy perturbation (FEP) simulations to characterize key residues and their mutation effects on HA-Fab binding at atomic level. The FEP method has been widely used to calculate binding affinities for a variety of chemical and biological systems, such as solvation free energy calculation, ligand-receptor binding, protein–protein interaction, and protein-DNA (RNA) binding [371, 372, 387-398], and is often regarded as the most rigorous and reliable method for free energy calculations. Many previous FEP calculations have achieved high accuracy for various protein-protein and protein-ligand binding affinities when compared to experiments [391, 399-401]. Previous work on H3N2 HA-antibody and H5N1 HA-receptor binding systems also showed excellent agreement between the calculated binding free energies and the experimental values [371, 372]. Among several available computational methods developed in past years, the FEP method based on all-atom explicit solvent model is probably the most accurate approach in estimating the relative antigen-antibody binding affinity[399, 400]. In this chapter, we first validated our FEP protocol by comparing the simulated binding affinity changes with available experimental data. Then we extended our FEP calculation to novel mutations, at the interfacial region, on either Group 1 HA or monoclonical antibody fragment (Fab). We found that the stacking interaction is critical for HA-antibody recognition and the non specific hydrophobic interaction is responsible for the broad neutralization of antibody. In

order to understand why F10-like antibodies successfully neutralize most Group 1 influenza but fail in Group 2 subtypes, we further investigated the role of a highly conserved His residue observed in almost all important Group 1 HA subtypes (such as H1 and H5) by mutating it to Asn in most Group 2 HA subtypes (four out of six Group 2 subtypes, including H3, H7, H10 and H15). Our results show that such a computational approach can serve as a complementary tool to interpret and predict critical mutations for HA-antibody binding.

6.2 METHOD AND SYSTEM

6.2.1 Molecular systems

The H5 HA (Viet04/H5) with Fab F10 complex (pdb entry 3FKU) was used for antigen-antibody binding, in which the HA (HA1 and HA2) monomer is bound to one Fab with both heavy and light chains (**Figure 6.1**) [367]. The HA-Fab complex was solvated in a 71.5 Å x 81.5 Å x 160.0 Å water box, and then 7 sodium ions were added to neutralize the system, with a total number of ~88,000 atoms. The solvated system was first energy minimized by 20,000 steps and then followed by 500,000 step equilibration. The snapshot during the equilibration was randomly picked as the starting point for the FEP calculations. The unbound (free) state was modeled with the HA or F10 only, solvated in water, and equilibrated with a similar process. The particle-mesh Ewald method is used for the long-range electrostatic interactions with a cutoff distance of 12 Å[402]. All underlying molecular dynamics simulations, which are widely used in simulating biological systems[344, 346, 403-410], are performed using NAMD2[411] molecular modeling package with 1.5 fs time step in NPT ensemble at 1 atm and 300 K. The CHARMM22 force field [412] and TIP3P water model [413] are used.

6.2.2 Free energy perturbation protocol

When an antigenic variation occurs, the change in the HA binding affinity to a neutralizing antibody can be calculated by the free energy perturbation (FEP) method [371, 372, 387-389, 391-394]. The Helmholtz free energy of a system can be expressed as,

$$G = -kT \ln Z = kT \ln\{\iint dpdq \exp[-\beta H(p,q)]\}$$

Where Z is the partition function and $H(p,q)$ is the Hamiltonian of the system; p and q represent the momentum and the coordinate, respectively; k is the Boltzmann constant and T is the temperature; β equals to $1/kT$. The binding free energy change ΔG due to a mutation in hemagglutinin or antibody can then be calculated as,

$$\Delta G_\lambda = -kT \ln\left\langle \exp(-\beta[V(\lambda+\Delta\lambda) - V(\lambda)]\right\rangle_\lambda$$

$$\Delta G = \sum_\lambda \Delta G_\lambda$$

where $V(\lambda) = (1-\lambda)\, V_1 + \lambda\, V_2$, and V_1 represents the potential energy of the wild-type, and V_2 represents the potential energy of the mutant. The FEP parameter λ changes from 0 (V_1) to 1 (V_2) when the system changes from the wild-type to the mutant, and $\langle\ldots\rangle_\lambda$ represents the ensemble average at potential $V(\lambda)$. In typical FEP calculations, in order to have a "smooth" transition from state A to B, many perturbation windows have to be used. To avoid singularity at small interaction distances, when λ approaches 0 or 1, a situation often referred to as "endpoint-catastrophe", we have used soft-core potentials for the Lenard-Jones interactions, with the 12-6 LJ function modified as the following [414, 415]:

$$V_{vdW} = \varepsilon_{ij}\left[\left(\frac{R_{ij}^2}{r_{ij}^2 + \delta(1-\lambda)}\right)^6 - \left(\frac{R_{ij}^2}{r_{ij}^2 + \delta(1-\lambda)}\right)^3\right]$$

where ε_{ij} is the depth of the potential well; R_{ij} is the radius; r_{ij} is the distance between a pair of atoms; and δ is the shift parameter which allows a smooth transition from the original Lenard-Jones potential to zero or vice versa. The electrostatic interactions are handled with the normal Coulomb law, but are switched-on for the "appearing atoms" only after $\lambda>0.1$, thus allowing the soft-core Lennard-Jones potentials to repel the possible overlapping before introducing the electrostatic interactions. Similarly, for the "disappearing atoms", the electrostatic interactions are switched off after $\lambda>(1-0.1)=0.9$.

In general, it is difficult to directly calculate the absolute binding affinity change ΔG_A for the binding process between a viral surface protein and an antibody due to the long time scale and complicated binding process. However, we can avoid this problem by designing a thermodynamic cycle to calculate the relative binding free energy change, i.e., $\Delta\Delta G_{AB}$. Instead of calculating the difficult direct binding energies ΔG_A and ΔG_B, we calculate the free energy changes for the same mutation in both the bound state (HA bound to antibody, ΔG_1) and the free state (HA or antibody not bounded, ΔG_2). Within a complete thermodynamic cycle, the total free energy change should be zero, which gives the relative binding affinity due to the mutation from A→ B as:

$$\Delta\Delta G_{bind} = \Delta G_B - \Delta G_A = \Delta G_1 - \Delta G_2$$

In the current setup, a 20-window scheme with soft-core potential has been adopted (λ = 0.00001, 0.0001, 0.001, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99, 0.999, 0.9999, 0.99999, 1). For each mutation, at least five independent runs starting from different initial configurations (taken from the equilibration) are performed for better convergence. The simulation time for each window is 0.3 ns. In each FEP λ-

window, the first 4000 steps are for the further equilibration. Larger window sizes and longer simulation durations have also been tested in our previous studies, and we found that the current setup gives us a reasonable convergence in the final binding affinities. Therefore, at least 60 ns (20windows x 0.3ns x 5runs x 2states) simulation time were generated for each mutation, and the total aggregate simulation time for this study is ~2 μs, which is much longer than most FEP calculations currently reported in the literature.

Despite of the controversy about the meaningfulness of breakdown the total free energy into components in the literature, and the ambiguity associated with a path-dependent decomposition [416-419], a breakup of the total binding free energy into its van der Waals and electrostatic components can offer useful information to the underlying physical interactions involved in HA-Fab binding. In this study, we collected the contributions of free energy change by van der Waals and electrostatic interactions separately. Due to the nonlinearity of the FEP formulation, there might be a small coupling term in this approach.

6.3 Results and Discussion

6.3.1 Validation of the FEP protocol with known experimental data

Mutagenesis studies for H5-F10 complex have shown the importance of the binding region between the αA helix of H5 and Fab, which was proposed to prevent the conformational change of αA helix and the following membrane fusion [367-369]. We noticed that several experimental data are available for mutagenesis studies in the αA helix region. Therefore, we first used these experimental data to validate our FEP protocol by calculating the relative binding free energies for mutations $N50_2A$, $V52_2A$,

$V52_2L$, and $V52_2E$ in the αA helix (subscripts 1 or 2 refer to HA1 or HA2, respectively; and the residue numbering system follows that of H3 (PDB entry: 2HMG)) [367].

The first target site $Asn50_2$ is exposed to the solvent in the αA helix, hence not directly contacting with Fab. Presumably the mutation $N50_2A$ will not affect the binding affinity significantly. Our simulations showed that $N50_2A$ mutation enhanced the binding affinity slightly, with $\Delta\Delta G$ = - 0.28 ± 0.28 kcal/mol (or equivalent to ~1.6 fold enhancement (decrease) in binding dissociation constant K_d, see **Table 6.1**), which agreed well with an experimental result of ~2 fold enhancement in K_d for the same mutation [367]. For the second target site $Val52_2$, which is conserved in Group 1 HA as directly interacting with the complementarity-determining region H1 (CDR-H1) of F10, three different mutations $V52_2A$, $V52_2L$, and $V52_2E$ were examined. Experimental site-directed mutagenesis study on $V52_2A$ revealed a 10~20 fold increase in K_d [367] (equivalent to 1.3~1.6 kcal/mol decrease in binding free energy), which was comparable to our FEP result of $\Delta\Delta G$ = 0.91 ± 0.29 kcal/mol (see **Table 6.1**). A conservative mutation to Leu of similar size as $Val52_2$, however, had little influence to the binding affinity, with only slight reduction in the binding affinity ($\Delta\Delta G$ = 0.57 ± 0.60 kcal/mol), which was also consistent with experimental observation, where the concentration of bound state was lowered by less than 10 % [367]. As one would expect for the $V52_2E$ mutation, the inclusion of charged amino acid in the hydrophobic core at the binding interface severely interfered the antibody association with the αA helix of HA, with a significant decrease in binding affinity, $\Delta\Delta G$ = 9.62 ± 0.93 kcal/mol. Consistently, $V52_2E$ mutation on Fab F10 failed to neutralize the HA in experiment (**Table 6.1**) [367].

We further validated our current FEP approach by performing extra reverse FEP calculations, in which five mutants ($Ala50_2$, $Leu52_2$, $Ala52_2$, $Glu52_2$ and $Ala56_2$) were mutated back to their wild-types ($Asn50_2$, $Val52_2$, $Val52_2$, $Val52_2$, and $Ile56_2$), respectively (see **Table 6.1**). We found that all five backward FEP simulations (mutants to wild-type) and forward FEP calculations (wild-type to mutants) show similar binding affinities but with an opposite sign. They both are in good agreement with experimental results. Considering the large size of the HA-Fab system, our FEP calculation provided a relatively small deviation of less than 0.5 kcal/mol on average with available experimental data. The FEP simulation protocol, therefore, may be employed as a reliable tool for this antigen-antibody complex binding affinity study.

6.3.2 Stacking interaction and hydrophobic environment crucial for HA binding

Inspired by encouraging results from above validation with known mutations, we extended our FEP method to novel mutations found in either HA or Fab side at the binding interface. The co-crystal structures of both H5-F10 and H5-CR6261 complexes suggest that the interactions between the αA helix of HA and CDRs of the Fab might be important for neutralization [367, 368]. As a key interaction site, we first focused on $Trp21_2$ and $Ile56_2$, two conserved residues observed in all subtypes of HAs, which make hydrophobic interactions with Met54 and Phe55 in the CDR-H2 of Fab F10 (equivalent to Ile53 and Phe54, respectively, in antibody CR6261).

In addition to the general hydrophobic interaction, a stacking interaction between the indole ring of $Trp21_2$ and phenyl group of Phe55 (or Phe54) of Fab F10 (or CR6261) is believed to play a critical role in binding recognition between the αA helix of HA and the CDR-H2 of Fab F10 (i.e., providing the binding specificity). Here, computational

mutageneses were attempted for these sites to evaluate how interactions associated with the aromatic side chains (π–π stacking) are helping the recognition of their binding partners. We started with mutating $Trp21_2$ to Ala using our FEP simulation. Knocking out $Trp21_2$ with a smaller amino acid Ala resulted in a binding affinity decrease of $\Delta\Delta G$ = 4.02 ± 0.34 kcal/mol, which corresponds to ~1,000 fold increase in the dissociation constant K_d. Similarly, a counter mutation on antibody's Phe55 with Ala reduced binding affinity by $\Delta\Delta G$ = 4.24 ± 0.76 kcal/mol, implying that the stacking interaction between two aromatic rings are indeed important, contributing ~4 kcal/mol to the binding affinity. The mutation effect on the binding site structure is clearly depicted in **Figure 6.2** for the wild-type and $W21_2A$ at the end of simulation. Due to the lack of an aromatic ring and a shorter side chain in the mutant $Ala21_2$, the inter-molecular hydrophobic cluster made of $Trp21_2$, Phe55 and Met54 is largely disrupted. On the other hand, we noticed enhancement of an intra-molecular interaction between Met54 and Phe55 in CDR-H2 of Fab. A free energy decomposition confirmed the role of vdW interactions (see Methods for detail). Almost all the contribution (>98% of the total) to the binding affinity were originated from the vdW interactions (4.11 kcal/mol), while the electrostatic interactions played only a minor role (-0.05 kcal/mol). The coupling term was 0.04 kcal/mol. Note that our estimated stacking interaction (~4 kcal/mol) and the vdW component of the total free energy (4.11 kcal/mol) are very similar to the previously reported value of stacking interactions between two idealized benzene rings (4.08 kcal/mol) [420]. Overall, our energetic analysis confirmed that the large free energy loss in the $W21_2A$ mutation is attributed to the destruction of stacking interactions between two aromatic residues from both HA and antibody. Meanwhile, the effect of hydrophobic interactions was further

evaluated with two non-conservative mutations, F55E (mutating to an acidic residue) and F55K (mutating to a basic residue). Much larger binding affinity decreases of 13.58 ± 0.55 kcal/mol and 7.66 ± 0.97 kcal/mol were found for F55E and F55K, respectively (**Table 6.2**), which reassures that the recognition of HA by nAb F10 is highly dependent on the hydrophobic interactions at the binding interface.

Given the crucial role of the hydrophobic interactions between these two aromatic residues (HA2's Trp21 and F10's Phe55), we further investigated the nearby non-polar residue Met54 in F10 (Ile53 in CR6261), which displays strong contacts with $Trp21_2$ and $Ile45_2$ in HA2, forming a hydrophobic cluster. As expected, the non-conservative M54E or M54K mutations broke the hydrophobic core and caused about 10 kcal/mol free energy decreases in our FEP simulations (**Table 6.2**). However, substituting Met with Ile, as found in Fab CR6261, or with Ala, were shown to destabilize the protein association by $\Delta\Delta G$ = 1.10 ± 0.31 kcal/mol and 2.13 ± 0.36 kcal/mol, respectively (**Table 6.2**). Note that the free energy decreases in conservative mutations of M54I or M54A are smaller than those of aromatic residue deletions in $W21_2A$ or F55A. This indicates that stacking interaction between aromatic residues is more important for the Fab F10 neutralization, even though a hydrophobic environment around these key aromatic residues is still required.

Overall, our FEP simulations revealed the detailed determining factors in the interaction between CDR-H2 of Fab and HA. The stacking interactions between the aromatic rings from both HA and Fab were essential for the antigen-antibody binding. Preserving a hydrophobic environment around these key aromatic residues was found to

be important as well, i.e., the existence of non-polar residues, such as Met54 in F10 and Ile53 in CR6261, also helped.

6.3.3 Non-specific hydrophobic interactions responsible for the broad antibody neutralization

Besides the hydrophobic interactions between HA and CDR-H2, another hydrophobic core is formed in the αA helix and the CDR-H1 **(Figure 6.3**), where $Pro293_1$, $Val52_2$, and $Ile56_2$ in HA are contacting with Val27 and the methyl group of Thr28 in F10. We first investigated the role of $Val52_2$ by two types of FEP mutations: one with other non-polar residues to keep the hydrophobic environment intact, and the other with charged ones to break the hydrophobic core. The FEP calculation results are listed in **Table 6.3**. In general, for mutations to similar-sized hydrophobic residues, minor binding free energy changes could be expected, but for mutations to charged residues, significant decreases in binding affinity or even disassociations of the HA-Fabs complex could be expected. Indeed, our simulations showed that the polarity of the substituted residue at the site 52 of HA2 is critical in preserving the wild-type binding affinity. More interestingly, the conservative hydrophobic mutations displayed more or less similar binding affinities within about ± 1 kcal/mol variation (**Table 6.3**), indicating the exact hydrophobic residue type is less critical in this binding environment. The mutations to charged amino acids, on the other hand, clearly broke up the antigen-antibody association with high free energy penalty of 7.28~15.47 kcal/mol (see **Table 6.3**). It should be noted that $V52_2I$ substitution can have even stronger binding with Fab than the wild-type, where the binding affinity enhanced by $\Delta\Delta G$ = -1.08 ± 0.59 kcal/mol, indicating that Fab is even more effective for HAs possessing $I52_2$ residue. It is interesting to observe that Ile, Leu, and Val are the most common residue types at site 52 of HA2 for all the

subtypes of Group 1 HA: Ile and Leu are two dominant amino acids in subtype H13/H16 and H12, respectively, while Val is the majority amino acid in other subtypes of Group 1 HAs (**Table 6.4**). Thus, the strong binding affinity and the limited sequence diversity (within aliphatic amino acids Ile, Leu, or Val) at site 52 explains why antibody F10 could have broad neutralization toward various HAs.

To the contrary of the mutational effect at $Val52_2$ which shows a clear binary dependence on residue polarity, the nearby hydrophobic residue $Ile56_2$ displayed a more sophisticated pattern for the individual amino acid substitutions. For example, similar sized $I56_2V$ and $I56_2L$ mutations showed similar binding activities as the wild-type within ± 1 kcal/mol, whereas mutation to a smaller but still hydrophobic Ala residue reduced the binding affinity by about 2.42 ± 0.45 kcal/mol **(Table 6.3**). The free energy decomposition analysis revealed that most of the free energy loss (2.50 kcal/mol) was due to the weaker van der Waals interactions. This is clearly presented in **Figure 6.3**, where the hydrophobic interaction in the wild-type $V56_2$ with $Pro293_1$ in HA1 and Val27 of Fab has been weakened by the short side chain of Ala, and the extra space is partially filled by two water molecules, making contacts with Ala.

The more dramatic difference was demonstrated in the $I56_2K$ mutation, where a non-conservative charged residue was introduced, however, the binding affinity was not affected or rather slightly enhanced by $\Delta\Delta G$ = -0.38 ± 0.65 kcal/mol. **Figure 6.3** shows how the long side chain (also hydrophobic) and positively charged NH3+ group of Lys could be packed in the binding interface. The terminal amine group is pointing out from the hydrophobic binding interface and interacts with hydrophilic residues such as $Asn53_2$ or $Asn60_2$, whereas the long aliphatic chain of Lys still makes favorable contacts with

hydrophobic residues Val27, $Val52_2$ and $Pro293_1$, thereby maintaining a strong interaction between αA helix and CDR-H1.

Compared to the highly conservative aromatic residues in the binding region of HA and Fab-CDR-H2 discussed in the above "stacking interaction" subsection, we found the binding interface between HA and Fab-CDR-H1 exhibits relatively more variability in amino acid selection in HA. Hydrophobic residues were usually required for site 52_2 to keep the Fab binding. However, the binding interface near the site 56_2 needed tightly packed hydrophobic environment. We found either the $V52_2I$ or $I56_2V$ mutation could actually increase the binding affinity by ~1 kcal/mol and ~0.5 kcal/mol, respectively. A strong but nonspecific binding between HA and antibody could be an important principle for designing vaccines with a broad neutralization.

6.3.4 Asn38 in Group 2 HA1 might contribute to the antibody neutralization escape

The molecular mechanism of why most Group 2 HAs (such H3 and H7) cannot be neutralized by F10 or CR6261 is not fully understood. Glycosylation at position 38_1 on Group 2 HA was proposed as a main reason for the neutralization escape in the previous studies [367, 368]. However, we found two of the six subtypes in Group 2 are not glycosylated at position 38_1, yet able to escape the neutralization. That is, the glycosylation may not be the only mechanism to explain the immunity escape of Group 2 subtypes. The possible molecular mechanism beyond the glycosylation was investigated by searching the sequence diversity around the HA-Fab binding interface. The sequences of all HA subtypes were collected from the NCBI Flu database (**Table 6.4**). Our hypothesis is that a residue site in HA would be important in understanding the antibody neutralization escape, if the amino acid types of that specific site are diversified over

different groups, but conserved within each group. That is, antibodies can neutralize HA subtypes in one group with a specific amino acid type conserved at that site, whereas HA subtypes in the other group might escape from the same antibodies due to a different amino acid type conserved at the same site. For the site 38_1, His is a well conserved residue within Group 1 (such as H1, H5, H2, H6, and H9), while Asn is a conserved residue within Group 2 (in 4 out of 6 subtypes, including H3, H7, H10 and H15). Other sites, like 40_1 and 38_2, are not fully conserved either across groups or within individual groups, indicating that they may play a less critical role in the neutralization (**Table 6.4**).

In order to confirm our hypothesis from the above sequence analysis and also to reveal the atomic detail of the binding specificity, we performed the FEP mutation for $His38_1$ in our current Group 1 H5N1 HA to Asn38, mimicking the conservative amino acid in Group 2 HA. The simulated FEP results showed that the binding affinity between HA and Fab decreased by 1.32 ± 0.78 kcal/mol, which is about 10-fold increase in the dissociation constant (K_d) value. Comparing the final structure with the wild-type, we found that the two native hydrogen bonds (nitrogen atoms at the side chain of $His38_1$ to the hydroxyl group of Ser30 or Gln64 in Fab) were broken by the mutation (**Figure 6.4**). The loss of side chain hydrogen bonding was partially compensated by waters entered in the binding interface, but no direct contact was found between mutated $Asn38_1$ and Fab. Therefore, the $H38_1N$ mutation weakened the interaction between HA and Fab by breaking the hydrogen bonds between HA1 and the antibody. It should be noted though that the binding affinity decrease of 1.32 ± 0.78 kcal/mol in $H38_1N$ mutation is not overwhelming, indicating that the antibody neutralization is a complicated process and this His38 residue might be another important contributing factor but probably not the

only factor (other factors, such as glycosylation, might contribute as well[367, 368]). Indeed, several other Group 1 subtypes that are bound/neutralized by F10, such as H8 and H11, do not have His at position 38 (they instead have Gln or Ser, respectively)[367]. Nevertheless, our FEP calculations indicate that the existence of Asn at the position 38_1 might be another important contributing factor for the neutralization escape in the Group 2 HA subtypes, in addition to the glycosylation.

6.4 CONCLUSION

In this paper, we performed rigorous free energy perturbation (FEP) calculations to estimate the influenza antigen-antibody binding affinities (using H5 HA and F10 Fab as a template) and study the characteristics of antibodies with a broad neutralization capability (such as F10 and CR6261). The simulated binding affinities between HA and Fab were in excellent agreement with the currently available experimental data. Several key residues in the HA-Fab binding regions were identified and further examined with *in silico* mutagenesis studies to explore the molecular mechanism of HA-Fab binding.

It was revealed that the stacking interaction of $Trp21_2$ (HA) and Phe55 (Fab) is critical to endow strong binding between the αA helix and the CDR-H2 of antibody. Our FEP simulations suggested that either $W21_2A$ in (HA side) or F55A (antibody side) will cause a significant binding affinity decrease of $\Delta\Delta G > 4.0$ kcal/mol (equivalent to ~1,000 fold increase in binding dissociation constant K_d). Besides, neighboring hydrophobic residues were also required to preserve stable hydrophobic network around the aromatic side chains. Furthermore, more general hydrophobic interactions were observed between HA and the CDR-H1 of Fab.

The HA residue sites 52_2 and 56_2 appeared to be more tolerable with various hydrophobic mutations with similar binding ability as the wild-type, which could elucidate the wide neutralization of Fabs among all Group 1 subtypes. In addition, we found the $V52_2I$ and $I56_2V$ substitutions could increase the binding affinity by ~1 kcal/mol and ~0.5 kcal/mol, respectively, which would be used as a potential way to improve the efficiency of current antibodies.

Besides the hydrophobic interactions, the hydrogen bonding between $His38_1$ and Ser30/Gln64 were also found to be important for the antibody neutralization. When $His38_1$ was mutated to Group 2-like $Asn38_1$, two hydrogen bonds were lost, substituted by hydration around $Asn38_1$ in-between the HA and the Fabs, with a net decrease of ~1.3 kcal/mol in binding affinity. This could be another important contributing factor for the neutralization escape in Group 2 subtypes, in addition to the glycosylation.

Table 6.1: Comparing the FEP simulation results with the experimental data for the HA-nAb binding free energy change due to the mutation in HA[a]

Mutation	Calculated $\Delta\Delta G$ (kcal/mol)	Reverse $\Delta\Delta G$ (kcal/mol)	Exptl $\Delta\Delta G$ (kcal/mol)
$N50_2A$	-0.28 ± 0.28	0.11 ± 0.48 ($A50_2N$)	- 0.4
$V52_2L$	0.57 ± 0.60	-0.20 ± 1.15 ($L52_2V$)	0 ~ 0.5
$V52_2A$	0.91 ± 0.29	-1.08 ± 0.59 ($A52_2V$)	1.3 ~ 1.6
$V52_2E$	9.62 ± 0.93	-8.22 ± 1.61 ($E52_2V$)	no bound states observed
$I56_2A$	2.42 ± 0.45	-1.81 ± 0.46 ($A56_2I$)	N/A

[a] A total of five independent runs has been performed for both the bound and free states for the standard error calculations with each running 6 ns.

Table 6.2: The FEP simulation results for the HA-nAb binding free energy change due to the mutation in HA2/CDR-H2

Mutation	Calculated $\Delta\Delta G$ (kcal/mol)
$W21_2A$	4.02 ± 0.34
M54A	2.13 ± 0.36
M54L	1.10 ± 0.31
M54E	10.02 ± 1.11
M54K	9.29 ± 1.20
F55A	4.24 ± 0.76
F55E	13.58 ± 0.55
F55K	7.66 ± 0.97

Table 6.3: The FEP simulation results for the HA-nAb binding free energy change due to the mutation in HA/CDR-H1

Mutation	Calculated $\Delta\Delta$G (kcal/mol)
H381N	1.32 ± 0.78
V522A	0.91 ± 0.29
V522I	- 1.08 ± 0.59
V522L	0.57 ± 0.60
V522F	1.44 ± 0.87
V522E	10.43 ± 0.83
V522D	15.47 ± 1.09
V522R	13.45 ± 0.98
V522K	7.28 ± 1.51
I562A	2.42 ± 0.45
I562V	-0.48 ± 0.46
I562L	0.84 ± 0.50
I562E	2.24 ± 1.59
I562K	-0.38 ± 0.65

Table 6.4: Comparison of the sequence conservation among 16 hemagglutinin subtypes

Group	Cluster	Subtype	HA1				HA2																
			17	18	38	40	18	19	20	21	38	41	42	44	45	46	48	49	52	53	55	56	111
Group1	H1a	H2	Y	H	H	K	V	D	G	W	K	T	Q	A	I	D	I	T	V	N	V	I	H
		H5	Y	H	H	Q	V	D	G	W	Q	T	Q	A	I	D	V	T	V	N	I	V	H
		H1	Y	H	H	V	V	D	G	W	K	T	Q	A	I	D	I	T	V	N	V	I	H
		H6	Y	H	H	V	I	D	G	W	K	T	Q	A	I	D	I	T	V	N	I	I	H
	H1b	H13	Y	L	S	I	I	N	G	W	K	T	Q	A	I	D	I	T	I	N	I	I	H
		H16	Y	L	S	V	I	N	G	W	K	T	Q	A	I	D	I	T	I	N	I	I	H
		H11	Y	L	S	V	I	N	G	W	K	T	Q	A	I	D	I	T	V	N	I	V	H
	H9	H8	Y	Q	Q	M	I	D	G	W	Q	T	Q	A	I	D	I	T	V	N	I	I	H
		H12	Y	Q	Q	E	V	A	G	W	R	T	Q	A	I	D	M	Q	L	N	V	I	H
		H9	Y	Q	H	K	V	A	G	W	R	T	Q	A	I	D	I	T	V	N	I	V	H
Group 2	H3	H4	H	H	T	Q	I	D	G	W	L	T	Q	A	I	D	I	T	L	N	L	I	T
		H14	H	H	S	K	I	D	G	W	L	T	Q	A	I	D	I	N	L	N	L	I	T
		H3	H	H	N	T	V	D	G	W	L	T	Q	A	I	D	I	N	L	N	V	I	T
	H15	H15	H	H	N	T	I	D	G	W	Y	T	Q	A	I	D	I	T	L	N	L	I	A
		H7	H	H	N	T	I	D	G	W	Y	T	Q	A	I	D	I	T	L	N	L	I	A
		H10	H	H	N	T	V	D	G	W	Y	T	Q	A	I	D	I	T	L	N	L	I	A

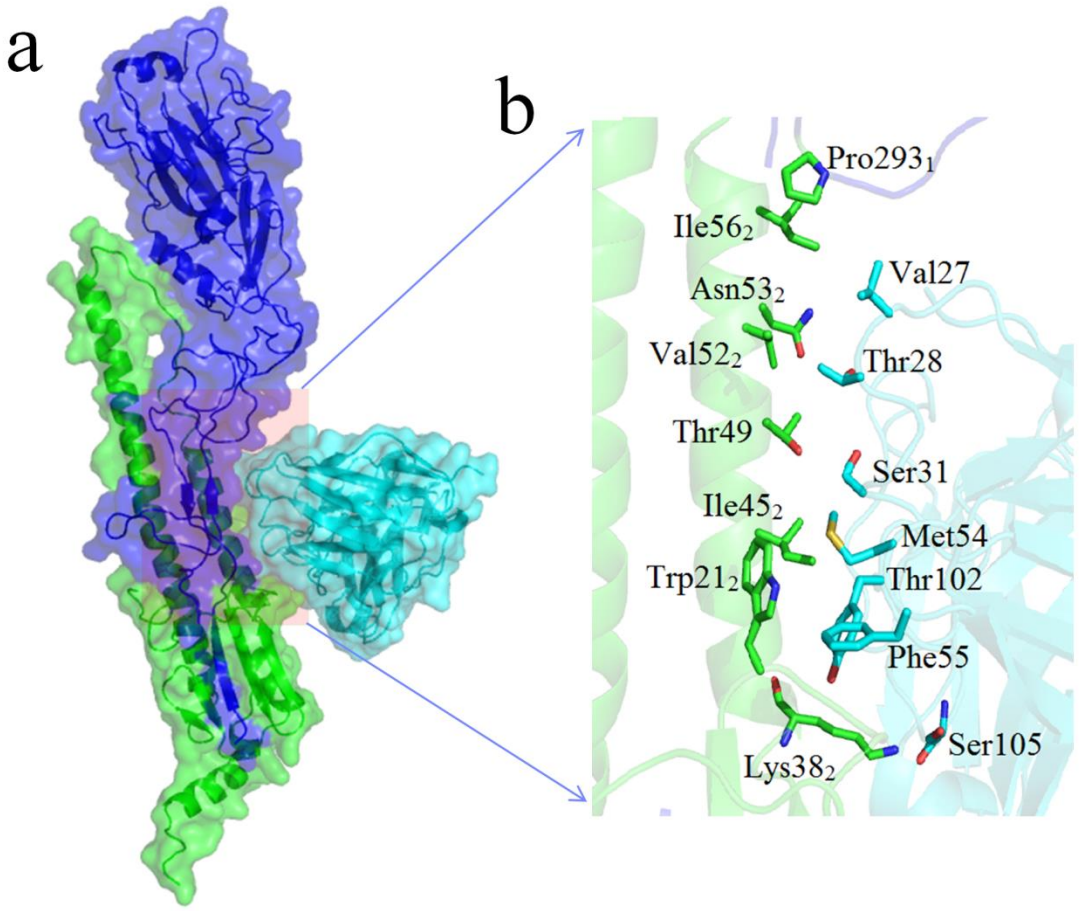


Figure 6.1. Molecular modeling system for the hemagglutinin protein binding with the antibody F10. (a) Overview of the HA-antibody complex structure. The HA and nAb are represented by surface and cartoon; HA1 and HA2 are colored blue and green, respectively, and both the heavy chain and light chain of the antibody are colored cyan. (b) Detailed view of antigen-antibody binding interface; the contact residues at are rendered by sticks.

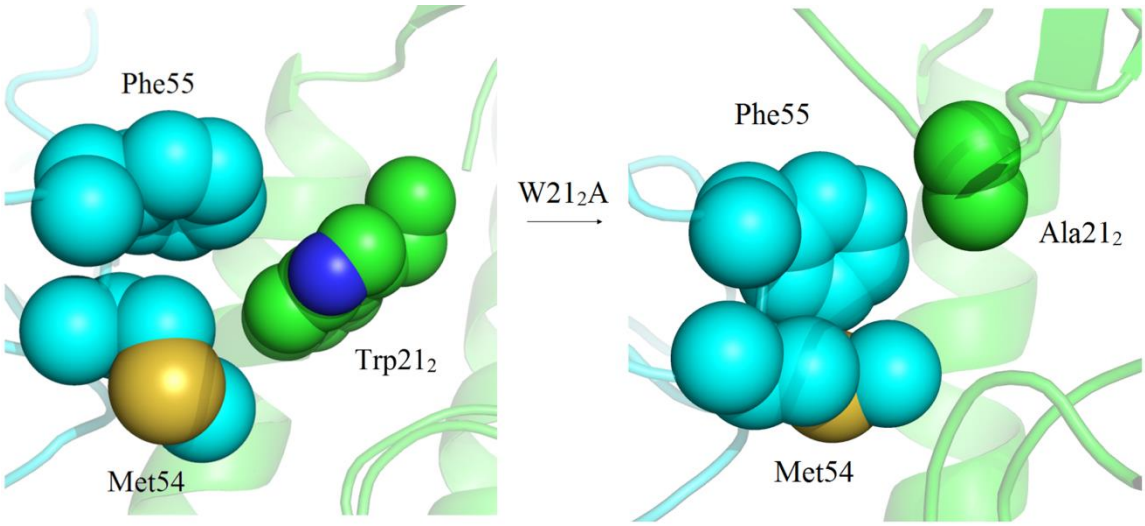


Figure 6.2. Structural comparison of F10 nAb bound to H5 HA around CDR-H2 due to the W21$_2$A mutation (green, the HA part; cyan, the nAb part). The overall complex is represented as cartoon and the residues at interaction face are rendered with spheres. The interactions of Trp21$_2$ (HA) with Met54/Phe55 (nAbs) are largely diminished by W21$_2$A substitution.

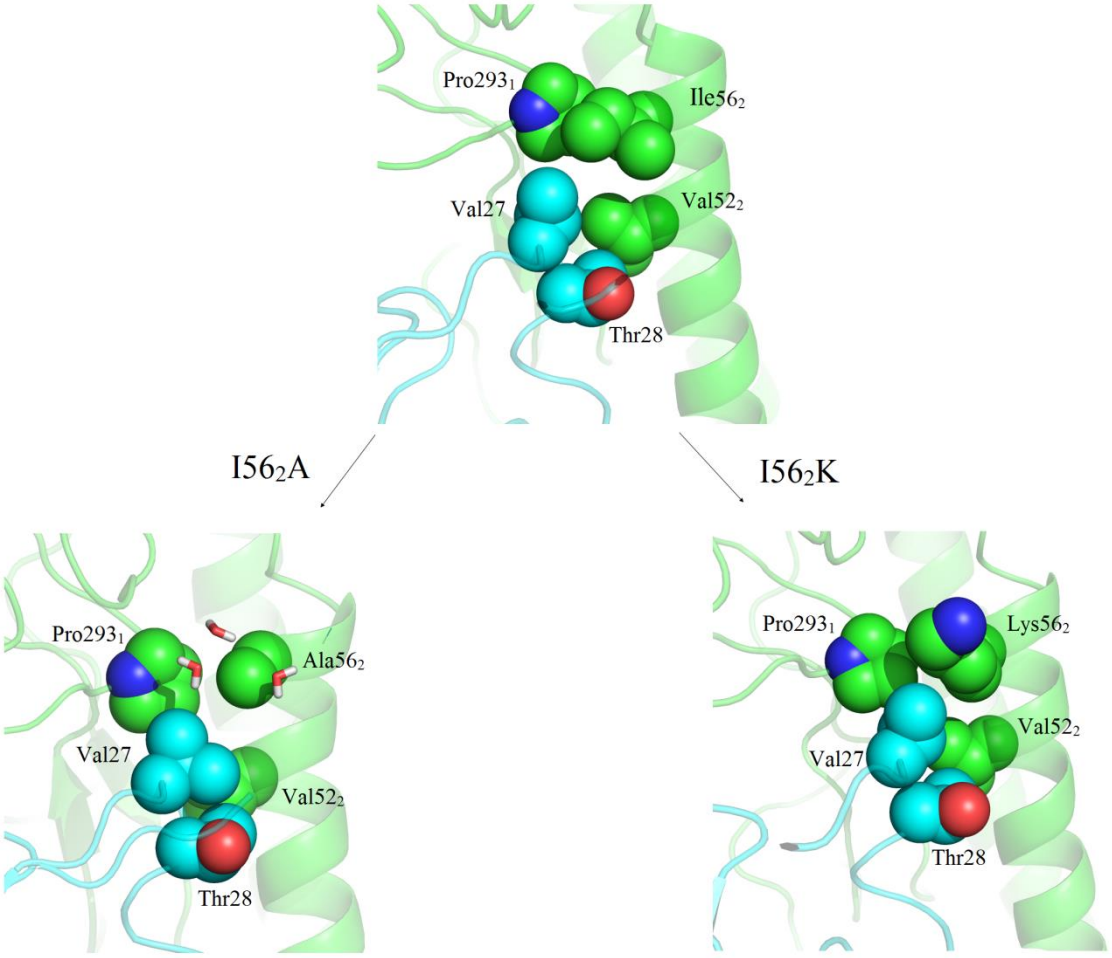


Figure 6.3. Structural comparison of F10 nAb bound to H5 HA around CDR-H2 due to I56$_2$A and I56$_2$K mutations (green, the HA part; cyan, the nAb part). The overall complex is represented as cartoon and the residues at interaction face are rendered with spheres. The hydrophobic core (shown in spheres) is broken by I56$_2$A mutation but is preserved in I56$_2$K mutation.

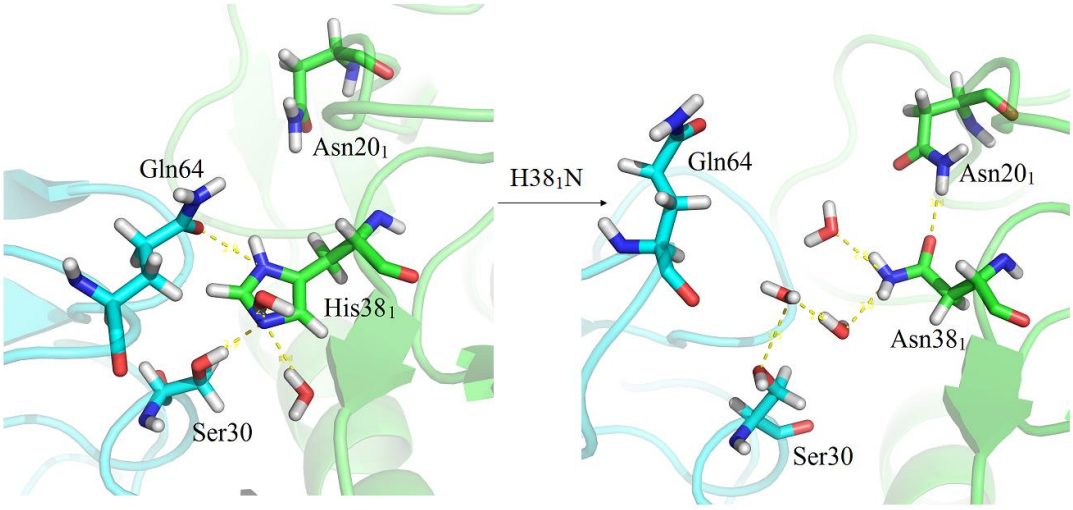


Figure 6.4. Structural comparison of F10 nAb bound to H5 HA around CDR-H2 due to H38$_1$N mutations (green, the HA part; cyan, the nAb part). The overall complex is represented as cartoon and the residues at interaction face are rendered with sticks. The side chain hydrogen bonds between His38$_1$-Gln64 and His38$_1$-Ser30 (shown with yellow dash line) are broken by H38$_1$N mutation.

7 Collapse of Unfolded Proteins in a Mixture of Denaturants

7.1 INTRODUCTION

Urea and guanidinium chloride (GdmCl) are both commonly used as protein denaturants in various studies [421-434]. The molecular denaturation mechanism continues to be controversial in the past decades. In general, two different mechanisms have been proposed: an "indirect mechanism" where denaturants disrupt the structure of water and thus enhance the solubility of hydrophobic groups of proteins [435-440]; and a "direct mechanism" where denaturants directly interact with proteins via electrostatic or van der Waals forces [441-446]. Our recent study on the denaturation of hen egg-white lysozyme in 8M urea solution strongly supports this "direct mechanism," in which urea interacts with the protein backbone and side chains via stronger dispersion interactions than water [446]. A two-stage kinetic model has been proposed, which begins with a "dry-globule" transient state, followed by a global unfolding of the protein [446]. Later experiments have further confirmed the "direct mechanism" by observing urea molecules directly forming hydrogen bonds to the backbone of dialanine (N-acetyl-$_L$-alanine N'-methylamide) [447]. Similar mechanisms have been proposed for GdmCl molecules, which directly interact with proteins [444, 448]. In addition, the planar, charged Gdm^+ is found to interact with aromatic side chains more strongly by "stacking" in both a model helical peptide and in protein L[449, 450] (which is also seen in our simulations, more below). Therefore, Gdm^+ is often considered to be approximately twice as effective as urea in its ability to denature proteins [447].

On the other hand, the behavior of proteins immersed in an aqueous mixture of two denaturants has not been systematically studied. One might be wondering what

happens to the protein conformations solvated in one denaturant (either urea or GdmCl only), when another denaturant (GdmCl or urea) is also added (i.e., in a mixture of both urea and GdmCl denaturants)? One might expect that the unfolded protein chain will remain or be further extended due to the addition of another denaturant. However, Shakhnovich and coworkers previously proposed that a mixture of co-solvents may trigger a collapse of a polymer in a broad range of conditions [451-453]. DeGennes and Brochard have also suggested a possibility of a chain collapse in a mixed solvent near the critical mixing point of the solvent [454]**.** Considering that two denaturants compete with each other in interacting with proteins when they are mixed, the competition might lead to unexpected complex behavior of protein energetics and dynamics. Meanwhile, proteins solvated in mixed solvents are considered to be a more realistic representation of a cellular environment. Thus, it is of fundamental importance to investigate the conformational diversity of a protein in mixed solvents, which may play a crucial role in modulating its function in various environments. In addition, how the interactions among the stronger denaturant (such as GdmCl), the relatively weaker denaturant (such as urea), water, and the protein interplay in the complex solution requires more detailed investigation, as the delicate balance of those interactions will influence the kinetics and thermodynamics of the protein denaturation reaction.

To address these questions, we performed extensive molecular dynamics simulations of proteins in urea/GdmCl mixture with different concentrations. Two independent protein systems, hen egg-white lysozyme and protein L, were simulated as the representative model proteins. We found a collapse of denatured protein conformations in urea/GdmCl mixtures for both lysozyme and protein L, as compared to

their respective conformations in the single denaturant solution. The word collapse here means that the protein form more compact structures. This contraction of the protein chain in mixed denaturant suggests that mixing two denaturants, i.e. good solvents, results in a poor solvent for proteins, which is consistent with the prediction of Brochard and de Gennes. Further analysis reveals that GdmCl is preferentially absorbed onto the charged protein residues due to its stronger electrostatic interactions, whereas urea is preferentially adsorbed to onto the hydrophobic and polar residues. This residue-specific preferential absorption of particular denaturant results in a "urea cloud" near the first solvation shell of the non-charged residues, which constitutes majority of the protein. The largely correlated fluctuations of this "local urea cloud" around particular protein residues induce the protein collapse, mainly by burying the hydrophobic residues.

7.2 METHODS

7.2.1 Preparation of the denatured proteins

The denatured structures of two proteins, hen egg-white lysozyme (PDB entry 193L, with single mutation W62G) [455] and wild-type of protein L (PDB entry 2PTL) [456], were prepared by the following procedure with three steps: 1) Solvating each protein with its native structure in a pre-equilibrated 8M urea water box. The size of the 8M urea solution box was the same as our previous study [345] and same for both proteins, 73.1 Å X 73.1 Å X 73.1 Å, which contained 1,920 urea and 8,192 water molecules, with a density of 1.12g/cm^3. 2) Running molecular dynamics simulations for at least 500 ns with NPT ensemble (1 atm and 310 K) until the proteins were fully unfolded. We defined the structures as the unfolded states when the radius of gyrations (Rgs) of proteins does not increase for at least 50 ns in the simulation. 3) For each

protein, a representative snapshot from the final 50 ns was chosen as the starting denatured structure for the current denaturant mixture study. The Rgs were 24.8 Å and 16.3 Å for denatured lysozyme and protein L, respectively.

For the lysozyme protein, the full length of the protein was used in the simulations (residue index from 1 to 129); for protein L, a fragment of residue 18-78 was used, because the N terminal 17-residue is a long, straight, and Glu-rich loop that was separated from other part of protein L; removal of the N terminal residues was often used by previous studies in order to effectively distinguish denatured states from the native structure, particularly when measuring Rg [457].

7.2.2 Proteins in the mixed-denaturants

In this step, the denatured proteins were then immersed in a mixed-denaturant water box with different combinations of concentrations of urea and guanidinium chloride (GdmCl) for both lysozyme and protein L. A total 4 different urea/GdmCl combinations were studied: 1) 8M urea + 0M GdmCl, 2) 6M urea + 2M GdmCl (replacing one quarter of the initial urea molecules in 8M urea by GdmCl), 3) 4M urea + 4M GdmCl (replacing half of the initial urea molecules in 8M urea by GdmCl), and 4) 0M urea + 6M GdmCl. For each system, we generated 3 independent trajectories, with each about 120 ns to 200 ns long in NPT ensemble (at 310 K and 1 atm). All simulations were performed using the NAMD2 molecular dynamics program [458]. Molecular dynamics simulations have been widely used to complement experiments [345, 446, 459-472], which can provide atomic details that are often inaccessible in experiments due to resolution limits, even with the currently available sophisticated experimental techniques. The CHARMM force field (c32b1 parameter set) [349, 412, 473] is used in the current

study for lysozyme, protein L, and the denaturants urea and GdmCl. A modified TIP3P water model was used for water [350]. The long-range electrostatic interactions were treated with the Particle Mesh Ewald (PME) [351, 474] method and a typical 12 Å cutoff was used for the van der Waals interactions. The time step for all production runs was 1.5 fs. The total aggregated simulation time for this study is more than 4.0 μs.

7.3 RESULTS AND DISCUSSION

7.3.1 Protein conformation collapse in urea and guanidinium chloride mixture

Denatured protein conformations were used as the starting structures for our current simulations. For protein hen egg white lysozyme, a representative snapshot of the unfolded protein was chosen from a previous 1 μs molecular dynamics unfolding trajectory in 8M urea (see **Figure 7.1a**) starting from the crystal structure (PDB entry 193L) [455]. Here, we used the single mutant (W62G) lysozyme for illustration, because it unfolded faster and more globally in 8 M urea than the wild type. The selected denatured structure of lysozyme has a radius of gyration (Rg) of 24.8 Å, which is significantly larger than that of its native state (Rg = 16.0 Å). A similar unfolding simulation in 8M urea was carried out to obtain the starting denatured structure of protein L (with the unfolding simulations starting from the crystal structure, PDB entry 2PTL) [456], where the Rgs for the denatured state and native state were 16.3 Å and 11.0 Å, respectively (**Figure 7.1b**).

For lysozyme, we investigated four different denaturant combinations, with different concentrations of urea and guanidinium chloride (GdmCl). We replaced part of urea with corresponding number of GdmCl molecules in the 8M urea system to generate different combinations of the urea/GdmCl mixture (see Method section for details). In

pure 6M GdmCl solution, the major peak of the Rg distribution is at ~25 Å (25.5 Å on average), whereas the same peak in pure 8M urea is at ~26 Å (Rg = 26.5 Å on average) (**Figure 7.1c**). To our surprise, the distributions of Rgs clearly show that the denatured lysozyme collapsed in all urea/GdmCl mixtures (**Figure 7.1c**). In "4M urea + 4M GdmCl" mixed denaturants (replacing half of the initial urea molecules in 8M urea by GdmCl), the major distribution peak of Rgs was shifted to ~19.4 Å (with average Rg = 21.4 Å), which was significantly smaller than the typical peak of ~24-26 Å in single denaturant solutions, as shown in **Figure 7.1c**. This decrease in Rg can also be seen from the time evolution of Rg in the equimolar mixture of denaturants (**Figure 7.2a**). And another two independent simulations gave us very similar results (see Figure S1). The Rg distribution in the "6M urea + 2M GdmCl" mixture (replacing one quarter of the initial urea molecules in 8M urea by GdmCl) also shows similar behavior, with its major peak shifted to 21~22 Å (with average Rg= 22.8 Å), in between the corresponding peaks of the "4M urea + 4M GdmCl" mixture and the pure denaturant solutions (either 8M urea or 6M GdmCl, see **Figure 7.1c**).

This collapse of the denatured protein conformations in the mixed denaturants was then further confirmed by simulating another protein, protein L. There, we found a similar denaturant mixture-induced collapse. In the "4M urea + 4M GdmCl" mixture, the average Rg value was shifted from 16.2 Å in pure 8M urea and 16.9 Å in pure 6M GdmCl to a lower value of 14.5 Å (**Figure 7.1d**). Similarly, the time evolution of Rg shows the same decreasing trend when the mixture of denaturants was used **(Figure 7.2b).**

7.3.2 Denaturant mixture triggers a decrease in solvent exposure of protein hydrophobic residues

The collapse of denatured proteins in mixed denaturants can also be seen from the lowering of protein solvent-accessible surface areas (SASA) during the simulation. Upon immersing lysozyme in the "4M urea + 4M GdmCl" mixture, the overall protein SASA decreased from initial value of ~13,300 $Å^2$ (i.e., in pure 8M urea) to ~11,500 $Å^2$. Similar trend was also found for protein L, where the corresponding SASA dropped from ~6,600 $Å^2$ in pure 8M urea to 6,100 $Å^2$ in the mixture. We further decomposed the total SASA into different amino acid types: hydrophobic, polar, and charged residues. We found that the reduction in SASA was mainly contributed by the hydrophobic residues (**Figure 7.2c** and **Figure 7.3**). For example, the SASA dropped by ~1,100 $Å^2$ in lysozyme for hydrophobic residues, which contributed to ~65% total loss in SASA, with the remaining 35% loss from the polar and charged residues. In other words, the solubility of hydrophobic residues decreased significantly in the mixture. Similar SASA decrease could be seen for protein L, in which hydrophobic residues contributed ~60% of total loss (the SASA drop from 1622 $Å^2$ at the beginning to 1347±32 $Å^2$ for the last 10 ns). Meanwhile, some non-native hydrophobic core was formed during the protein collapse. For further analysis, we calculated the radial distribution function (rdf) of urea and Gdm+ around the side chain of one representative hydrophobic residue Phe38 in lysozyme. A significant decrease in the solvent exposure can be seen for residue Phe38 after the protein collapse (**Figure 7.2d**). In addition, the number of backbone-backbone hydrogen bonds increased 70% for the hydrophobic residues, (from 3.07±0.83 at the beginning 10 ns to 5.26±1.10 for the last 10 ns, see **Figure 7.4**), which further confirms enhanced self-interactions among them. On the contrary, the SASA remained mostly

unchanged for charged residues (due to their strong electrostatic interactions with GdmCl) and only decreased slightly for polar residues during the same collapsing process for both protein systems. To further confirm the solubility decrease of hydrophobic residues in mixtures, but not in GdmCl alone, we also calculated the SASA of proteins in a pure 6M GdmCl system ("0M urea + 6M GdmCl"). No obvious decrease of SASA was observed for all three amino acid types (See **Figure 7.5** and **Figure 7.6**). In summary, our simulation results revealed a reduced solubility and increased self-interactions of hydrophobic residues in mixed denaturants, indicating the hydrophobic collapse as the underlying main factor for the consequent collapse of the entire protein.

7.3.3 The increased contacts during the collapse are mostly non-native

The numbers of native contacts and total local contacts (including both native and non-native) were calculated for both lysozyme and protein L in the "4M urea + 4M GdmCl" mixture during the collapse process (**Figure 7.2a** and **7.2b**). Two residues are considered to be in contact, if their Cα-Cα distance is less than 6.5 Å. For the lysozyme system, we noticed a significant increase in the number of total local contacts after ~40 ns of the simulation time. The number of total local contacts increased 21.5%, from the initial 107 to the final 130 with an average number of 125 ± 10, in the "4M urea + 4M GdmCl" mixture (**Figure 7.2a**). Coincidently, we found that the radius of gyration Rg dropped from ~25 Å to ~19 Å after ~40 ns (**Figure 7.2a**), which was correlated to the increase in the number of local contacts. However, the native contact number remained roughly constant during this process, fluctuating around 34 with a standard deviation of 3.0. For the protein L mixture system, very similar trends were found (**Figure 7.2b**). The number of local contacts increased by 16%, from the initial 56 to the final 65, with an

average of 62 ± 2.4; while the final number of native contacts even decreased slightly, from the initial 26 to the final 23 with an average number of 23 ± 2.3. Therefore, it seems that both lysozyme and protein L collapsed towards non-native structures, indicating the proteins were still in their denatured states, albeit more compact structures. In other words, the induced collapse of the denaturant mixture is not a refolding process, but a mere collapse with more non-native contacts formed.

7.3.4 Rearrangement of denaturants near protein surface and enhanced local crowding induce the protein collapse

The driving force toward the protein collapse was then investigated from an energetic perspective. The total interaction energy distributions between the individual solvent molecules (guanidinium, urea, or water) and protein were calculated for lysozyme in the "4M urea + 4M GdmCl" mixture system. **Figure 7.7a** shows the comparison of the interaction energy distributions in the first and last 10ns of the trajectory, which represent the initial unfolded state and the final collapsed state of protein, respectively. The energy values are normalized in per molecule scale that using totally interaction energy divided by the number of each solvent. It is apparent from this figure that GdmCl overall has a significantly more favorable interaction with the protein than the urea, confirming that GdmCl is a relatively stronger denaturant. For the interaction between water and protein, we found the distribution of the total interaction energies almost the same before and after the collapse (**Figure 7.7a)**. For the interaction between urea and protein, we observed a slightly higher energy distribution (i.e., less favorable) after the protein collapse, with the interaction energy weakened by ~0.1 kcal/mol per urea molecule on average, indicating that some of the initial nearby urea molecules were replaced by the stronger denaturant GdmCl. The largest changes were observed in the interaction

between guanidinium and protein. Interaction energy between guanidinium and protein was much enhanced by the collapse, with the average interaction energy enhanced by 1.07 kcal/mol per guanidinium (-2.72 kcal/mol in the beginning and -3.79 kcal/mol after the collapse). In order to better understand whether the primary driving force for this process was electrostatic or van der Waals (vdW), the total interaction energy was further decomposed into electrostatic and vdW contributions. We found most of the interaction energy enhancements (1.00 kcal/mol out of 1.07 kcal/mol total) stemmed from the electrostatic interaction between guanidinium and protein (**Figure 7.7b)**, with the average electrostatic energy changed from -2.53 kcal/mol in the beginning to -3.63 kcal/mol after the protein collapse. Meanwhile, the interaction energy between chloride and protein was not changed significantly, with only a 0.18 kcal/mol decease. Therefore, the protein collapse is favored by an enhancement in the electrostatic interaction energy between guanidinium and protein. In the mean time, overall density of urea consistently increases near the first solvation shell (FSS) of protein lysozyme (while the density of guanidinium stays nearly the same) from 10 to 30 ns until the protein collapses, as evident from the density of particular denaturant molecules shown in **Figure 7.8.** Similar trend of denaturant density redistribution before the collapse can also be seen for other repeated runs and protein L (see FigureS6 and S7). This rearrangement of denaturants near the protein stems from two different factors: (1) protein residue-specific affinity for a particular denaturant, which originates from the chemical composition of an amino acid. For example, the acidic residues interact with guanidinium with much more higher affinity than with urea. On the other hand, hydrophobic residues have a preference for urea over guanidinium. (2) Self-aggregation tendency of a particular denaturant. It is

known that urea can form large clusters in solution, whereas guanidinium prefers to stay as a homo-dimer (ref). Thus, the preferential adsorption of urea onto the non-charged amino acids, that are major constituent of the protein chain, creates a "urea cloud" around those residues, as evident from Fig 4a. As suggested by Brochard and de Gennes, near the critical temperature of the denaturant mixture, the solvent (i.e. urea) density fluctuations become more and more correlated, resulting in a solvent correlation length ξ that is comparable to the size of the protein. The large-scale solvent density fluctuations near criticality result in indirect long-range attractions between protein residues, which shield the excluded-volume interactions. As a result, the protein collapses by burying the hydrophobic residues due to the solvent-fluctuation mediated attractive interactions between protein segments. Meanwhile, this collapsed protein conformation is also accompanied by the most energetically favorable environment for both guanidinium and urea surrounding the protein, as shown by the interaction energy analysis.

The protein-solvent interaction was further investigated by calculating the ratio of GdmCl to urea molecules ($\rho_{gdm/urea}$) at the first solvation shell (FSS) of each protein residue. Any water, urea, or GdmCl molecule is considered to be in the FSS, if that is within 5Å of any protein atom. For both lysozyme and protein L solvated in the "4M urea + 4M GdmCl" mixtures, we found that the $\rho_{gdm/urea}$ increased for most of the protein residues when half of the original urea molecules in 8M urea were replaced by GdmCl. **Figure 7.9** shows the comparison of $\rho_{gdm/urea}$ in the first and last 10 ns, which indicates that Gdm+ replaced urea in the FSS of proteins due to its stronger electrostatic interactions with the protein, as discussed above. Interestingly, we also noticed that $\rho_{gdm/urea}$ dropped at the locations of some residues in protein L during the simulation,

which are found to be mainly lysine residues (marked with * in **Figure 7.9b**). Considering the unfavorable electrostatic interactions between two positively charged groups, Gdm^+ and $–NH3^+$ of lysine, this seems reasonable (see more data and discussions below with **Figure 7.10,** as well as **Figure 7.11**for Arg).

This can also be seen from the time evolution of detailed atomic radial distribution functions (rdf). **Figure 7.9c** shows the rdf between the oxygen of urea (OU) and the backbone amide hydrogen (HB) [$g_{OU\text{-}HB}(r)$], which experiences a noticeable reduction after the protein collapse, indicating the loss of overall urea-backbone interactions during this process due to the burying of hydrophobic residues. On the other hand, a slight increase in the first peak was observed for the corresponding $g_{HG\text{-}OB}(r)$ between the amide hydrogen (HG) of GdmCl and the carbonyl oxygen (OB) of the protein backbone (**Figure 7.9d**). Therefore, the stronger denaturant GdmCl replaces part of the weaker denaturant urea, and enhances its overall interaction with the protein during this collapsing process. Our previous denaturation studies on urea-induced lysozyme unfolding have suggested a ‘‘direct interaction mechanism,’’ in which urea has stronger interactions with protein than water [446, 475-480]. In the current urea/GdmCl mixtures, GdmCl has even stronger interactions with protein than urea, indicating that GdmCl acts as the leading player among water, urea, and GdmCl in terms of their direct interactions with proteins. Therefore, it appears that guandiniums are “sucked in” by the protein to maximize the total number of strongly interacting guanidiniums on the protein surface. Interestingly, the protein also collapses somewhat by burying its own hydrophobic residues to accommodate more denaturant molecules near its surface, which is the most energetically favored state for the protein and denaturant mixture.

Finally, since guanidinium is positively charged, we further investigated the role of charged protein residues during this collapse. We analyzed the solvation of both glutamic acid (E) and lysine (K) side chains as an example (see **Figure 7.10)**. The negatively charged side-chain oxygens (OE) of glutamic acid were highly solvated by Gdm+ for both protein systems, which is 20~30 fold higher than that of urea. Following the protein collapse, guanidinium replaced urea molecules originally in contact with glutamic acid side chains. On the contrary, the positively charged side-chain (-NH3+) of lysine was mainly solvated by urea rather than Gdm^+, with more urea and less Gdm+ near –NH3+ at the end of the simulations for both lysozyme and protein L (**Figure 7.10**). We noticed that the solvation of another positively charged amino acid, arginine, mostly remained the same during the protein collapse (**Figure 7.11**). A possible explanation is that the unfavorable electrostatic interaction was compensated by the favorable stacking interaction between Gdm^+ and the guanidinium group from the arginine side chain. These detailed results with charged residues further support that the electrostatic interactions play the dominant role in guanidinium's interaction with proteins.

7.4 CONCLUSIONS

In general, guanidinium chloride (GdmCl) is considered to be a stronger denaturant than urea for proteins; therefore, adding more GdmCl to the solution or replacing part of the urea with GdmCl would presumably cause the protein to unfold further to a more stretched state (or at least keep at the current stretched state) if a simple additive denaturation effect is in action. However, a counter-intuitive phenomenon was observed in our molecular dynamics simulations with both hen egg-white lysozyme and protein L, where the unfolded proteins collapsed in urea/GdmCl mixture as compared to

that in the single denaturant (either GdmCl or urea). We then found the collapse was accompanied by a burying of hydrophobic residues at the protein surface, and an increase of local non-native contacts, indicating that it was not a refolding process but rather a simple collapse of the denatured state.

Detailed energetic and structural analyses then showed that GdmCl molecules replaced some urea molecules in the FSS of proteins through their stronger electrostatic interactions with protein backbones. Meanwhile, the urea molecules, though some replaced by the stronger denaturant GdmCl, still accumulate near the protein surface, creating a more crowded local environment for the protein. This rearrangement of denaturants near the protein surface and the crowded local environment induce the protein collapse, mainly by burying the hydrophobic residues, resulting in an enhanced self-interaction among the protein residues as seen in its increased non-native local contacts. These findings from detailed molecular simulations not only confirm the predictions from analytical statistical mechanics models, but also provide us with a deeper molecular picture of this denaturation process: when the two denaturants compete with each other to interact with the protein, the stronger denaturant (GdmCl) has a greater tendency to move closer to the protein surface than the weaker one (urea); in order to accommodate more GdmCl near the protein surface, the protein itself will collapse somewhat (i.e., yielding in space as well) in order to "suck in" more GdmCl molecules. The redistribution of denaturants near the protein surface and crowded local environment induce the protein hydrophobic collapse and result in an overall minimized free energy state.

Mixed solvents are considered to be a more realistic environment for protein in nature. It is thus of fundamental importance to investigate the conformational diversity of a protein, which plays a crucial role in modulating its function in various environments. Our study indicates that the dynamic structure of a protein in mixed solvents is more complicated than we have expected, and such a detailed study of the protein solvated in different denaturants may have provided new insights into the mechanisms of protein folding and unfolding.

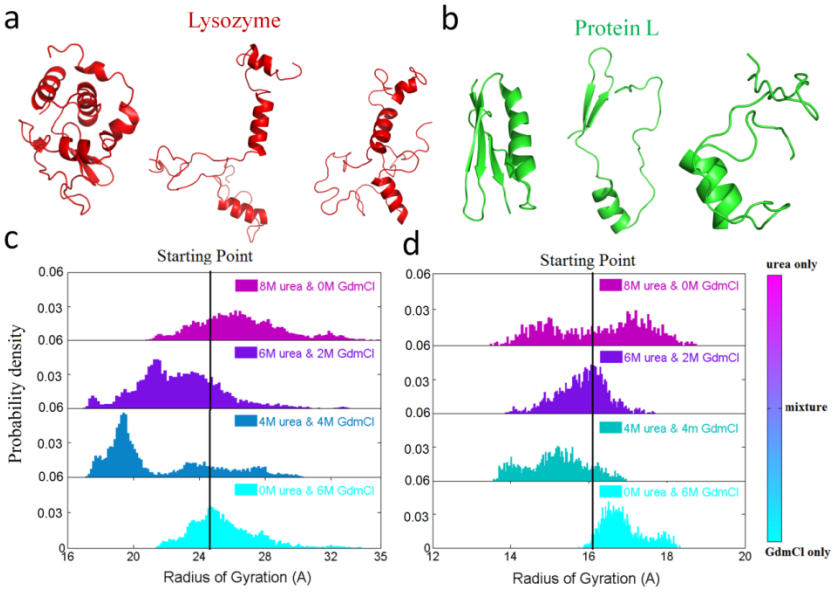


Figure 7.1: Protein collapse in urea/GdmCl mixture. (a) and (b) The structure of lysozyme and protein L. The native structure is shown on the left, and the denatured structure is in the middle which is used as the starting point for the simulations with different denaturant combinations. The collapsed structure is shown on the right. (c) and (d) The distribution of the radius of gyration (Rg) of proteins under different concentrations of urea and GdmCl for lysozyme and protein L systems, respectively. The black line is the starting point to all the simulations.

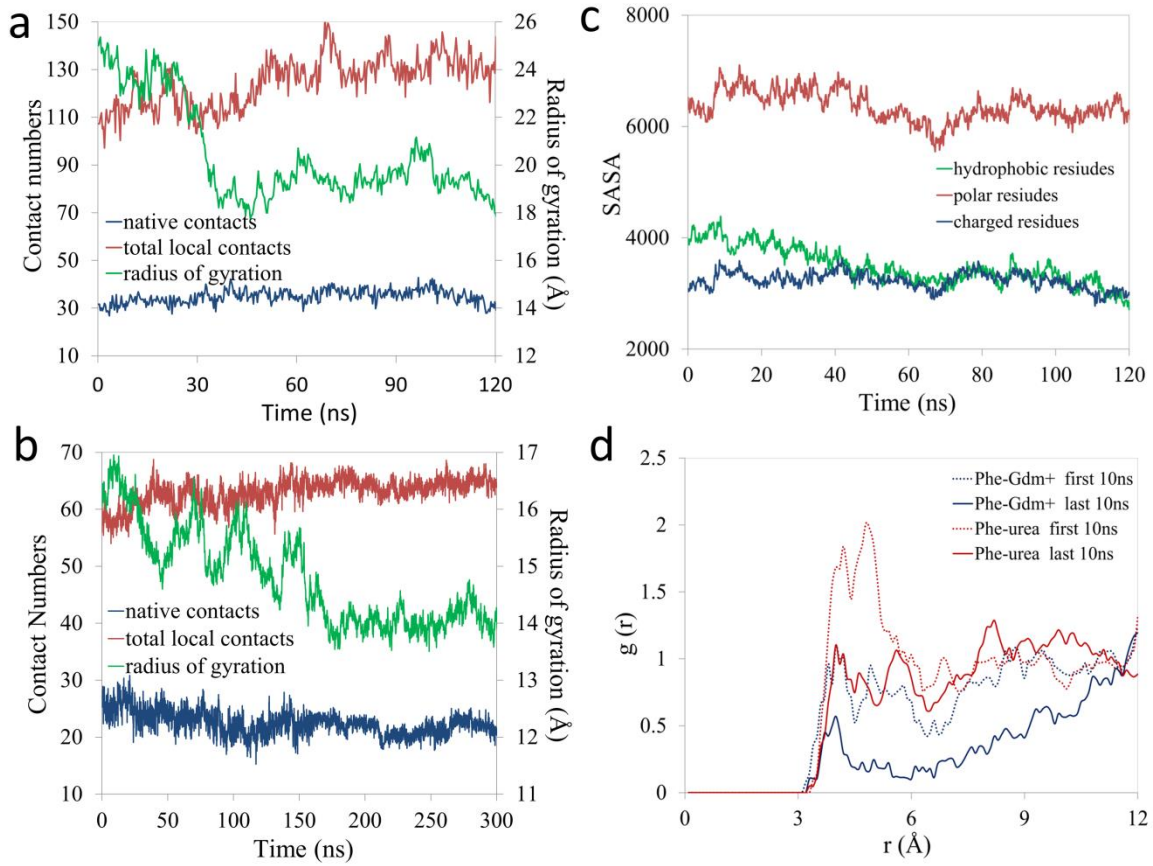


Figure 7.2: The time dependent of total local contacts (red), native contacts (blue), and radius of gyrations (green) in "4M urea + 4M GdmCl" mixture for lysozyme (a) and protein L (b), respectively. (c) Protein solvent-accessible surface area (SASA) of different type of residues in "4M urea + 4M GdmCl" mixture for lysozyme. (d) Time dependent pair radial distribution function g(r) between side chain of Phe38 and the carbon atom of urea /Gdm+.

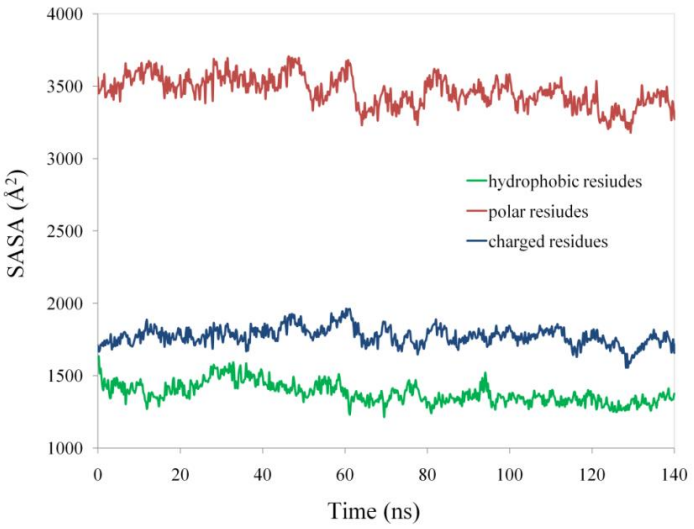


Figure 7.3: Protein solvent-accessible surface area of different type of residues in "4M urea + 4M GdmCl" mixture for protein L during the simulation time.

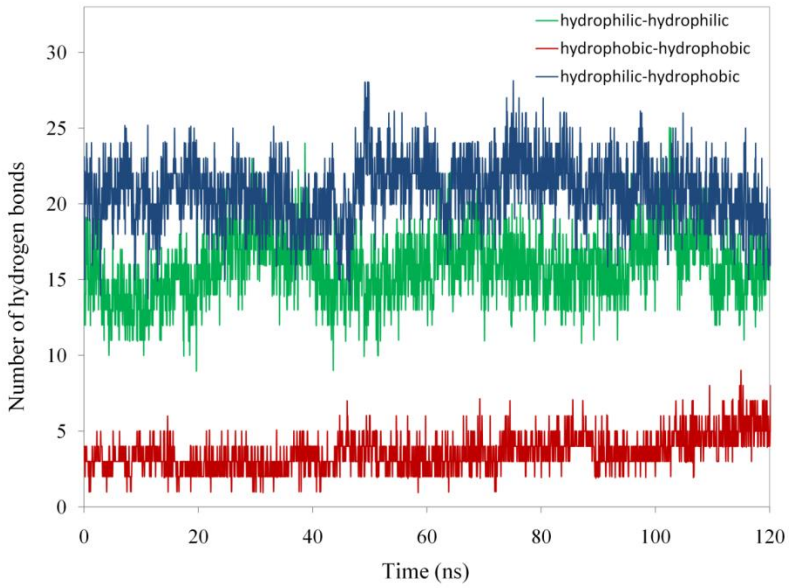


Figure 7.4: The time dependence of the backbone–backbone hydrogen bonds formed by the residue pairs of hydrophobic–hydrophobic (red), hydrophilic–hydrophilic (green), and hydrophobic–hydrophilic (blue), respectively in "4M urea + 4M GdmCl" mixture for lysozyme.

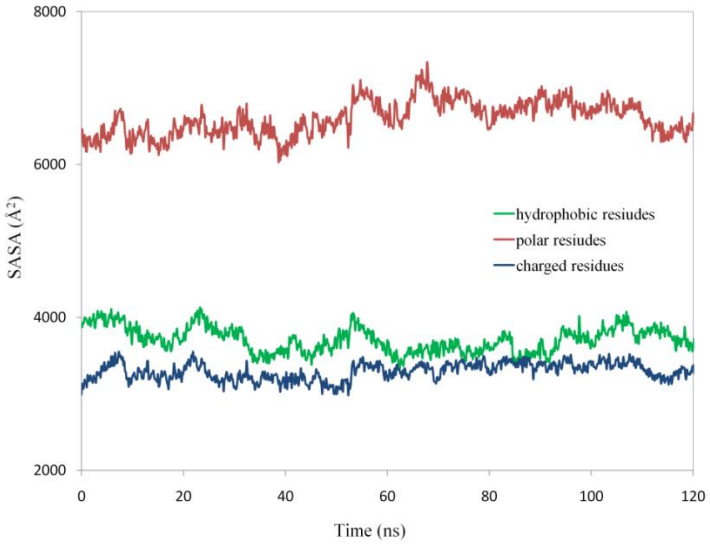


Figure 7.5: Protein solvent-accessible surface area of different type of residues in "0M urea + 6M GdmCl" single denaturant system for lysozyme during the simulation time.

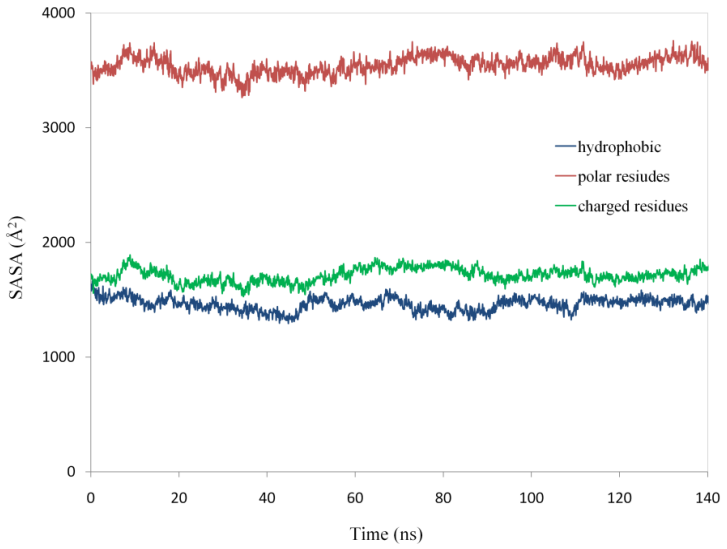


Figure 7.6: Protein solvent-accessible surface area of different type of residues in "0M urea + 6M GdmCl single denaturant system for protein L during the simulation time.

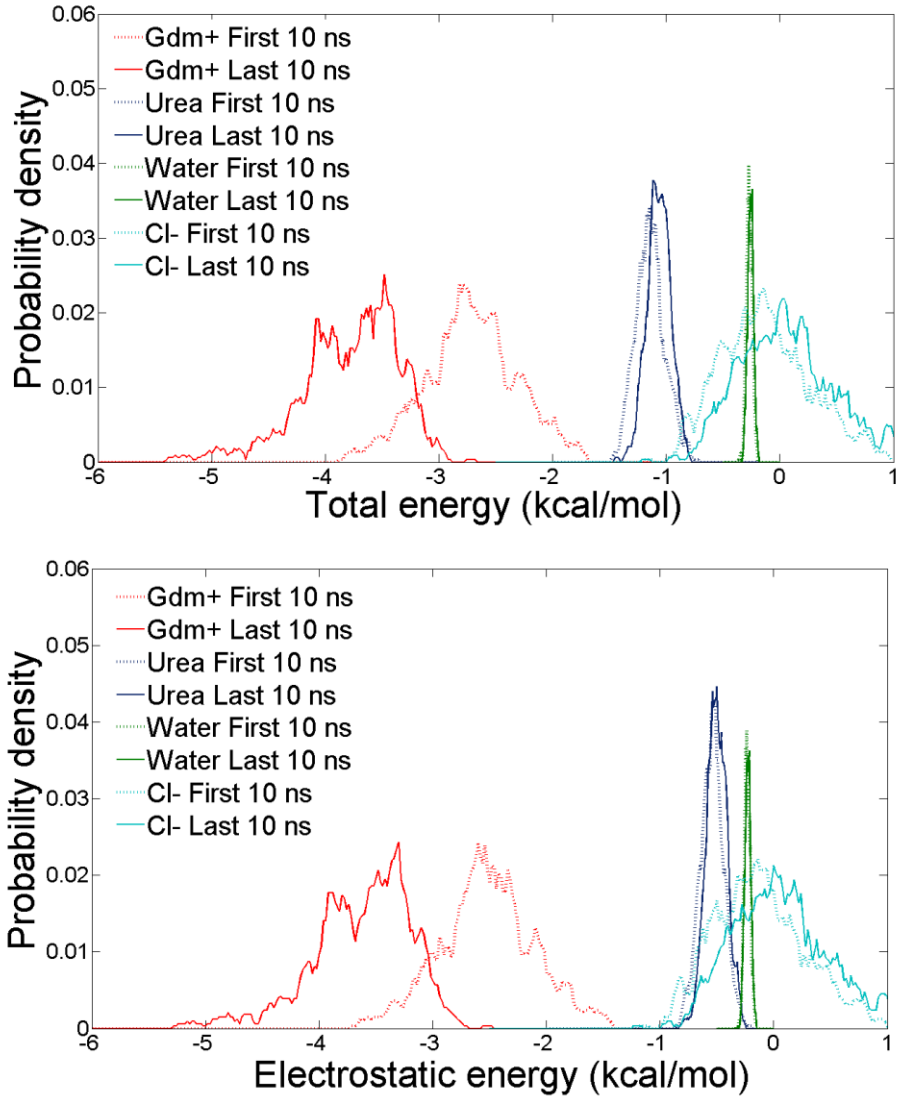


Figure 7.7: The distributions of total interaction energy (a) and electrostatic component energy (b) between solvents (guanidinium, urea, water, and chloride ion) and protein for "4M urea + 4M GdmCl" mixture lysozyme system. The interaction energies are normalized in per molecule level, with individual total interaction energies divided by the number of each solvent molecule.

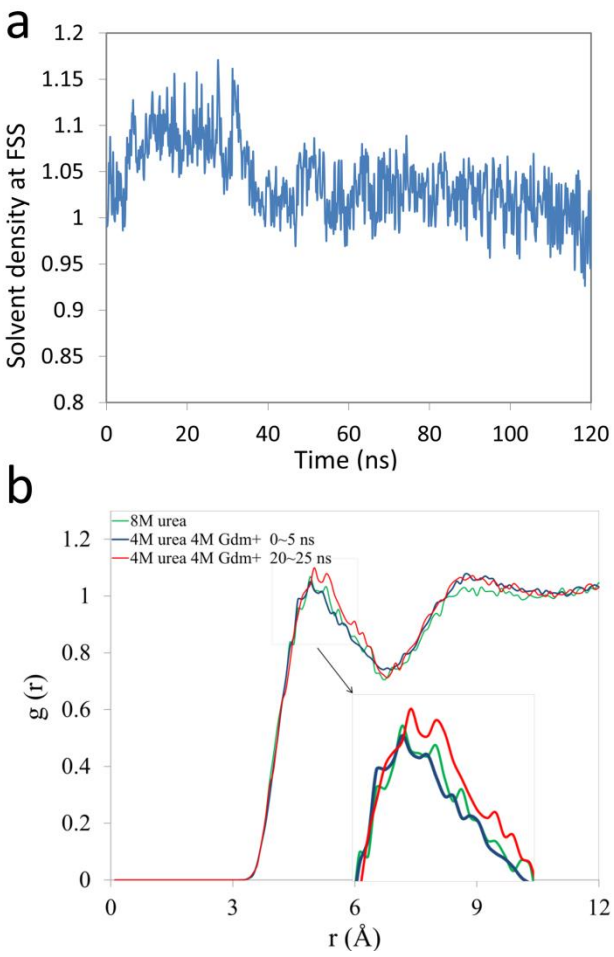


Figure 7.8: The local crowding effect at the surface of protein lysozyme. (a) The time dependent density of urea and guanidinium molecules at the first solvation shell of protein. The density is calculated from the total number of urea and guanidinium molecules and then normalized by the solvent-accessible surface area of the protein. (b) Time dependent pair radial distribution function g(r) between the α carbon atoms of the protein backbone and the carbon atoms of urea (and guanidinium if any) in "8M urea + 0M GdmCl" mixture (green) and "4M urea + 4M GdmCl" mixture (at t=0-5ns in blue as reference, and 20-25ns in red), respectively.

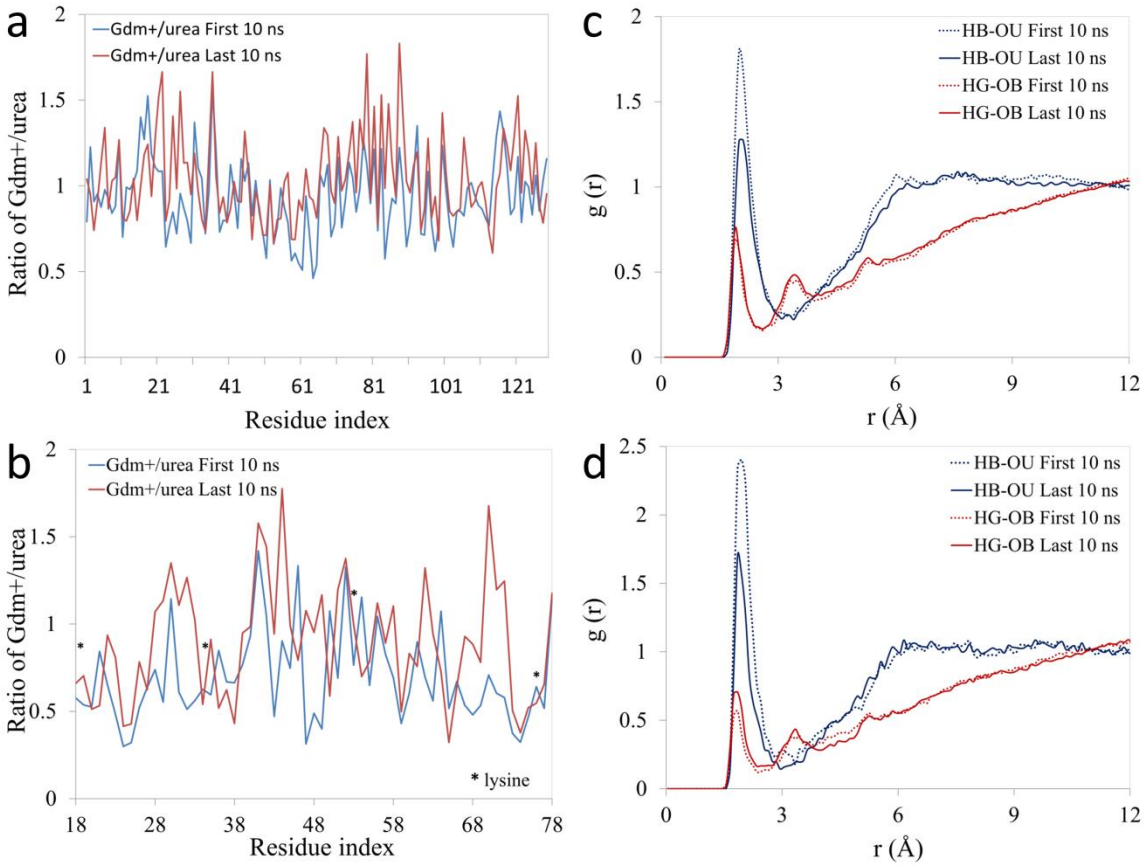


Figure 7.9: (a) and (b) The ratio of GdmCl to urea molecules ($\rho_{gdm/urea}$) at the first solvation shell for each protein residue in "4M urea + 4M GdmCl" mixture for lysozyme and protein L, respectively. Amino acid lysine is labeled as * for protein L. (c) and (d) Time dependent pair radial distribution function g(r) between backbone amide hydrogen HB and urea oxygen OU (blue), as well as between backbone carbonyl oxygen OB and Gdm^+ hydrogen HG (red) residue in "4M urea + 4M GdmCl" mixture for lysozyme and protein L, respectively.

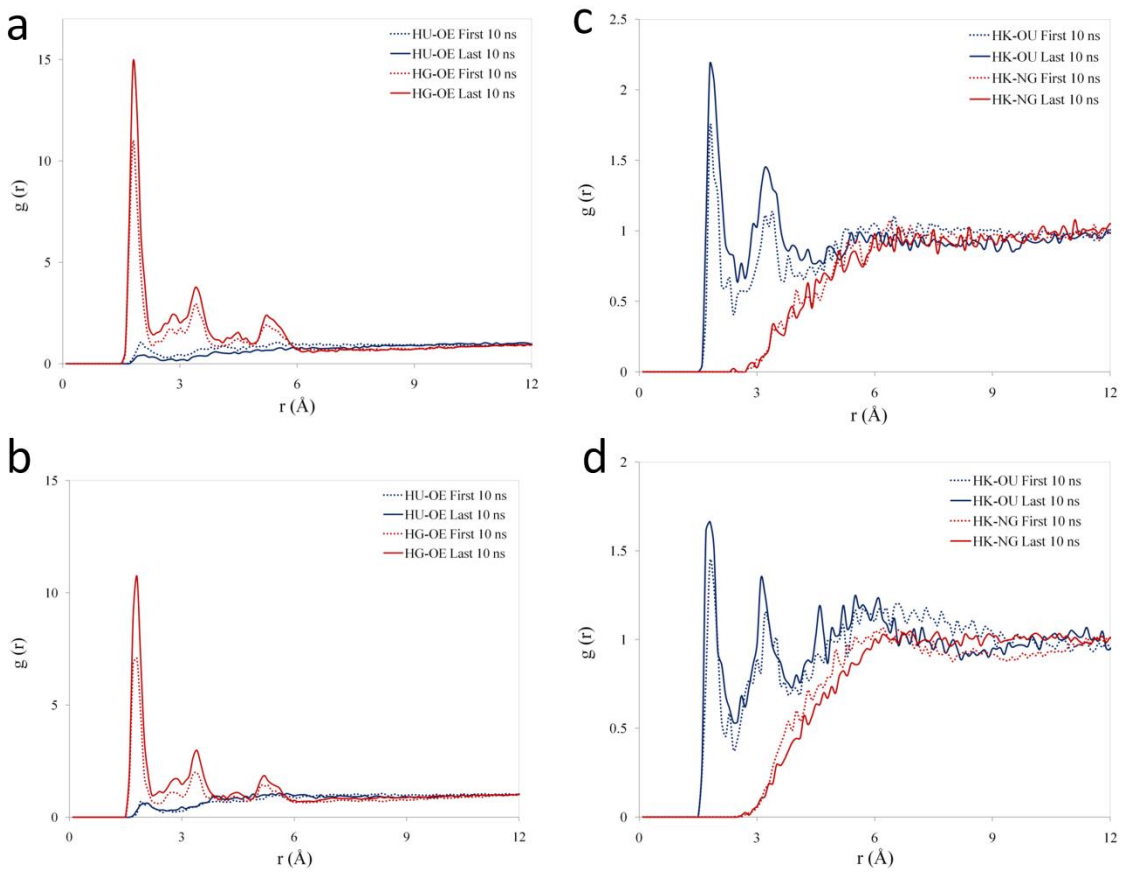


Figure 7.10: Time dependent radial distribution functions of urea and GdmCl to charged side chains in "4M urea + 4M GdmCl" mixture. (a) and (b) The pair radial distribution function g(r) between negatively charged glutamic acid side-chain oxygen OE and urea hydrogen HU (blue), as well as Gdm^+ hydrogen HG (red) for lysozyme and protein L, respectively. (c) and (d) The pair radial distribution function g(r) between positively charged lysine side-chain hydrogen HK and urea oxygen OU(blue), as well as Gdm^+ oxygen OG (red) for lysozyme and protein L, respectively.

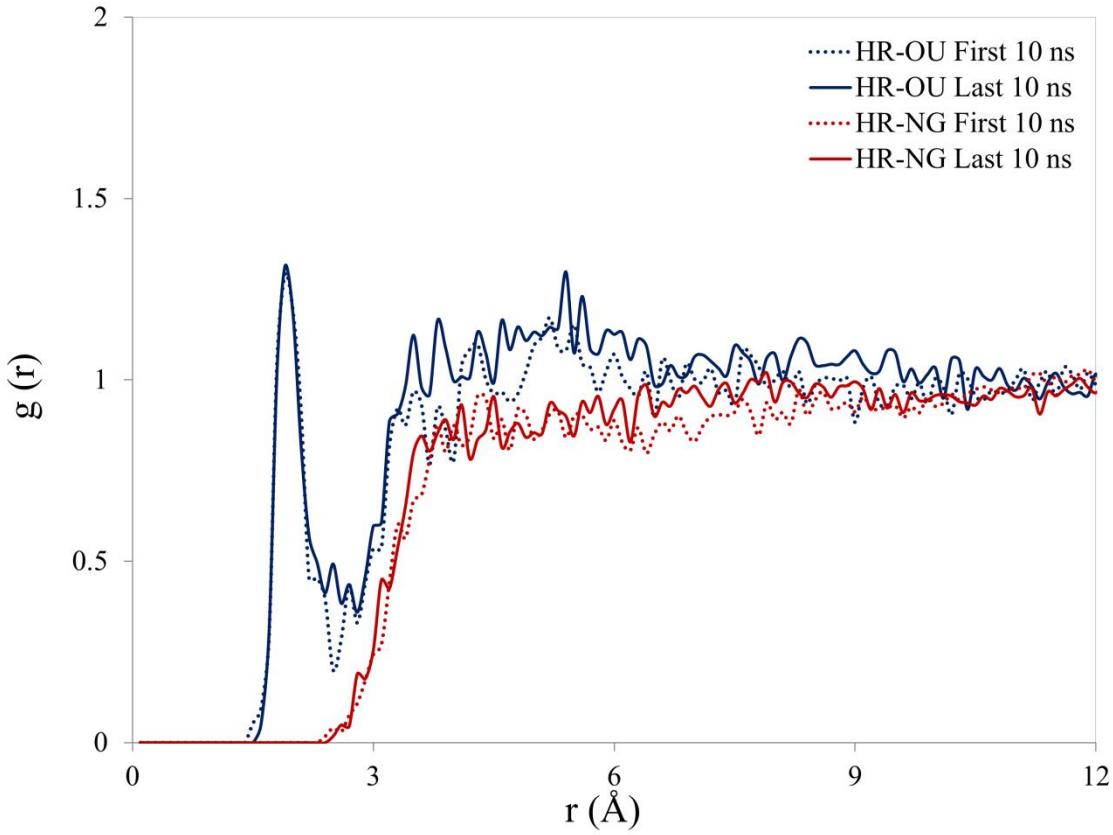


Figure 7.11: Time dependent pair radial distribution function g(r) between the hydrogen atom (HR) at the side chain of arginine and the oxygen atom of urea (OU), or the nitrogen atom (NG) of Gdm+.

8 Conclusion

The accurate prediction of an RNAs three dimensional structure from its "primary structure" will have a tremendous influence on the experimental design and its interpretation, and ultimately our understanding of the many functions of RNA. RNAs form complex secondary and three-dimensional structures, and their biological functions highly rely on their structures and dynamics. Therefore, a general coarse-grained potential is developed for modeling RNA 3-D structures. Each nucleotide is represented by five pseudo atoms, two for the backbone (one for the phosphate and another for the sugar), and three for the base to represent base-stacking interactions. We introduce a hybrid coarse-grained model that explicitly describes the physics electrostatics, and hydrogen bond interactions and is constructed based on experimental structural statistics. The overall potential is derived and optimized to structural statistics calculated from known structures. The developed model can stably capture the RNA native structures with various motifs, which has shown comparable abilities to the atomistic model. With a simulated annealing simulation protocol, the model successfully folds most of tested RNAs of less than 30-nt to within 4.0 Å of the native structures. In addition, with limited restraints on Watson-Crick basepairing based on the data from NMR spectroscopy, the current model can successfully predict complex tertiary structures of large size RNAs. The model was then further refined in atomistic scale using an approach that integrates small-angle X-ray scattering, all-atom minimization and molecular dynamic simulations. The model also demonstrates that some specific motifs such as pseudoknots can be captured when the coordinated Mg^{2+} cations and limited basepairing restraints are explicitly included in our coarse-grained model. The accuracy of our model has been

compared with other RNA structure prediction methods presented in the previous study of *RNA-Puzzles*; the results are comparable to the best predictions given by the other methods. Therefore the coarse-grained model presented here has the potential to offer both improved accuracy and efficiency for RNA structure modeling and prediction.

For proteins, a general, transferable coarse-grain framework based on the Gay-Berne potential and electrostatic point multipole expansion is developed for polypeptide simulations. The solvent effect is described by the Generalized Kirkwood theory. The coarse-grain 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 coarse-grain level. We demonstrate that the coarse-grain 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 and microsecond dynamic simulations of polyalanine of 5 and 12 residues reveal that the coarse-grain 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 coarse-graining approach presented has the potential to offer both accuracy and efficiency for biomolecular modeling.

Next, those developed models are integrated to current all-atom models and applied to different biological systems. One of the important applications is to study

protein and microRNA interactions. The recognition mechanism and cleavage activity of argonaute (Ago), miRNA, and mRNA complexes are the core processes to the small non-coding RNA world. The 5' nucleation at the 'seed' region (position 2-8) of miRNA was believed to play a significant role in guiding the recognition of target mRNAs to the given miRNA family. To better understand the recognition mechanism of RISC and the repertoire of guide-target interactions we introduced G:U wobbles and mismatches at various positions of the microRNA (miRNA) 'seed' region and performed all-atom molecular dynamics simulations of the resulting Ago-miRNA:mRNA ternary complexes. Our simulations reveal that a wide variety of modifications, including combinations of multiple G:U wobbles and mismatches in the seed region, are admissible and result in only minor structural fluctuations that do not affect overall complex stability. Lastly, introduction of disruptive mutations revealed a bending motion of the PAZ domain along the L1/L2 'hinge' and a subsequent opening of the nucleic-acid-binding channel. These findings suggest that the spectrum of a miRNA's admissible targets is different from what is currently anticipated by the canonical seed-model. Moreover, they provide a likely explanation for the previously reported sequence-dependent regulation of unintended targeting by siRNAs.

Antigen-antibody binding is a good example to study protein-protein interactions. Antibodies binding to conserved epitopes can provide a broad range of neutralization to existing influenza subtypes and may also prevent propagations of potential pandemic viruses by fighting against emerging strands. Here we propose a computational framework to study structural binding patterns and detailed molecular mechanisms of viral surface glycoprotein hemagglutinin (HA) binding with a broad-spectrum of

neutralizing monoclonical antibody fragments (Fab). Rigorous free energy perturbation (FEP) methods have been used to calculate the antigen-antibody binding affinities, with an aggregate underlying molecular dynamics simulation time of several microseconds (~2μs) using all-atom, explicit solvent, models. A high accuracy has been achieved in the validation of our free energy perturbation protocol against a series of known binding affinities for this complex system, with less than 0.5 kcal/mol errors on average. Then novel mutations onto the interfacial region are introduced to study further the binding mechanism. It is found that the stacking interaction between Trp21 in HA2 and Phe55 in the CDR-H2 of Fab is crucial to the antibody-antigen association. A single mutation of either W21A or F55A can cause a binding affinity decrease of ~4.0 kcal/mol (equivalent to ~1,000-fold increase in dissociation constant Kd). Moreover, for Group 1 HA subtypes (which include both the H1N1 'Swine flu' and the H5N1 'bird flu'), the relative binding affinities only change slightly (< ± 1 kcal/mol) when non-polar residues at the αA helix of HA mutate to conservative amino acids with similar size, which explains the broad neutralization capability of antibodies like F10 and CR6261. Finally, the hydrogen bonding network between His38 (in HA1) and Ser30/Gln64 (in Fab) is found to be important in preserving the strong binding of Fab against Group 1 HAs, whereas the lack of such hydrogen bonds with Asn38 in most Group 2 HAs might be responsible for the escape of antibody neutralization. These large scale simulations might have provided new insight into the antigen-antibody binding mechanism at atomic level, which could be essential in designing more effective vaccines for influenza.

Finally, dynamics structures of protein in mixed denaturants are investigated by molecular mechanics. As we know, both urea and guanidinium chloride (GdmCl) are

frequently used as protein denaturants. Given that, proteins generally adopt extended and/or unfolded conformations in either aqueous urea or GdmCl, one might expect that the unfolded protein chains will remain or become further extended due to the addition of another denaturant. However, a collapse of denatured proteins has been revealed using atomistic molecular dynamics simulations, when a mixture of denaturants is used. Both hen egg-white lysozyme and Protein L are found to undergo collapse in the denaturant mixture. The collapse of the protein conformational ensembles is accompanied by a decreased solubility and increased non-native self-interactions of hydrophobic residues in the urea/GdmCl mixture. The increase of non-native interactions rather than the native contacts indicates that the proteins experience a simple collapse transition from the fully denatured states. During the protein collapse, the relatively stronger denaturant GdmCl displays a higher tendency to be absorbed onto the protein surface due to their stronger electrostatic interactions with proteins. At the same time, urea molecules also accumulate near the protein surface, resulting in an enhanced "local crowding" for the protein near its first solvation shell. This rearrangement of denaturants near the protein surface and crowded local environment induce the protein collapse, mainly by burying their hydrophobic residues. These findings from molecular simulations are then further explained by a simple analytical model based on statistical mechanics.

References

1. Merchant, B.A. and J.D. Madura, A Review of Coarse-Grained Molecular Dynamics Techniques to Access Extended Spatial and Temporal Scales in Biomolecular Simulations. Annual Reports in Computational Chemistry, 2011. **7**: p. 67-85.
2. Olson, W.K. and V.B. Zhurkin, Modeling DNA deformations. Current Opinion in Structural Biology, 2000. **10**(3): p. 286-297.
3. Orozco, M., et al., Theoretical methods for the simulation of nucleic acids. Chemical Society Reviews, 2003. **32**(6): p. 350-64.
4. Knotts, T.A., et al., A coarse grain model for DNA. Journal of Chemical Physics, 2007. **126**(8): p. 084901.
5. Maroun, R.C. and W.K. Olson, Base sequence effects in double-helical DNA. III. Average properties of curved DNA. Biopolymers, 1988. **27**(4): p. 585-603.
6. Maroun, R.C. and W.K. Olson, Base sequence effects in double-helical DNA. II. Configurational statistics of rodlike chains. Biopolymers, 1988. **27**(4): p. 561-84.
7. Hao, M.H. and W.K. Olson, Modeling DNA supercoils and knots with B-spline functions. Biopolymers, 1989. **28**(4): p. 873-900.
8. Tan, R.K. and S.C. Harvey, Molecular mechanics model of supercoiled DNA. Journal of Molecular Biology, 1989. **205**(3): p. 573-91.
9. Sprous, D. and S.C. Harvey, Action at a distance in supercoiled DNA: effects of sequence on slither, branching, and intramolecular concentration. Biophys. J., 1996. **70**(4): p. 1893-908.
10. Sprous, D., R.K. Tan, and S.C. Harvey, Molecular modeling of closed circular DNA thermodynamic ensembles. Biopolymers, 1996. **39**(2): p. 243-58.
11. Tan, R.K., D. Sprous, and S.C. Harvey, Molecular dynamics simulations of small DNA plasmids: effects of sequence and supercoiling on intramolecular motions. Biopolymers, 1996. **39**(2): p. 259-78.
12. Matsumoto, A. and W.K. Olson, Sequence-dependent motions of DNA: a normal mode analysis at the base-pair level. Biophys. J., 2002. **83**(1): p. 22-41.
13. Coleman, B.D., W.K. Olson, and D. Swigon, Theory of sequence-dependent DNA elasticity. Journal of Chemical Physics, 2003. **118**(15): p. 7127-7140.

14. Mergell, B., M.R. Ejtehadi, and R. Everaers, Modeling DNA structure, elasticity, and deformations at the base-pair level. Physical Review E, 2003. **68**(2): p. 021911.

15. Flammini, A., A. Maritan, and A. Stasiak, Simulations of action of DNA topoisomerases to investigate boundaries and shapes of spaces of knots. Biophys. J., 2004. **87**(5): p. 2968-75.

16. LaMarque, J.C., T.V. Le, and S.C. Harvey, Packaging double-helical DNA into viral capsids. Biopolymers, 2004. **73**(3): p. 348-55.

17. Peyrard, M., Nonlinear dynamics and statistical physics of DNA. Nonlinearity, 2004. **17**(2): p. R1-R40.

18. Vologodskii, A., Brownian dynamics simulation of knot diffusion along a stretched DNA molecule. Biophys. J., 2006. **90**(5): p. 1594-7.

19. Malhotra, A., R.K. Tan, and S.C. Harvey, Modeling large RNAs and ribonucleoprotein particles using molecular mechanics techniques. Biophys. J., 1994. **66**(6): p. 1777-95.

20. Malhotra, A. and S.C. Harvey, A quantitative model of the Escherichia coli 16 S RNA in the 30 S ribosomal subunit. J. Mol. Biol., 1994. **240**(4): p. 308-40.

21. Harvey, S.C., R.K.Z. Tan, and A. Malhotra, Succinct Models for Very Large Nucleic-Acids, with Application to Supercoiled DNA and to the Structure of the Ribosome. Biophys. J., 1990. **57**(2): p. A10-A10.

22. Tan, R.K.Z., A.S. Petrov, and S.C. Harvey, YUP: A molecular simulation program for coarse-grained and multiscaled models. Journal of Chemical Theory and Computation, 2006. **2**(3): p. 529-540.

23. Zhang, D.Q., et al., Electrostatic interaction between RNA and protein capsid in cowpea chlorotic mottle virus simulated by a coarse-grain RNA model and a Monte Carlo approach. Biopolymers, 2004. **75**(4): p. 325-337.

24. Cao, S. and S.J. Chen, Predicting RNA folding thermodynamics with a reduced chain representation model. Rna-a Publication of the Rna Society, 2005. **11**(12): p. 1884-1897.

25. Jonikas, M.A., et al., Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. Rna-a Publication of the Rna Society, 2009. **15**(2): p. 189-199.

26. Das, R. and D. Baker, Automated de novo prediction of native-like RNA tertiary structures. Proc. Natl. Acad. Sci. USA, 2007. **104**(37): p. 14664-14669.

27. Ding, F., et al., Ab initio RNA folding by discrete molecular dynamics: From structure prediction to folding mechanisms. Rna-a Publication of the Rna Society, 2008. **14**(6): p. 1164-1173.

28. Sugita, Y. and Y. Okamoto, Replica-exchange molecular dynamics method for protein folding. Chemical Physics Letters, 1999. **314**(1-2): p. 141-151.

29. Frellsen, J., et al., A Probabilistic Model of RNA Conformational Space. Plos Computational Biology, 2009. **5**(6).

30. Pasquali, S. and P. Derreumaux, HiRE-RNA: A High Resolution Coarse-Grained Energy Model for RNA. Journal of Physical Chemistry B, 2010. **114**(37): p. 11957-11966.

31. Tirion, M.M., Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. Phys. Rev. Lett., 1996. **77**: p. 1905-1908.

32. Bahar, I., R. Atilgan, and B. Erman, Direct evaluation of thermal fluctuations in protein using a single parameter harmonic potential. Folding & Design, 1997. **2**: p. 173-181.

33. Haliloglu, T., I. Bahar, and B. Erman, Gaussian dynamics of folded proteins. Phys. Rev. Lett., 1997. **79**: p. 3090-3093.

34. Flory, P.J., M. Gordon, and N.G. McCrum, Statistical thermodynamics of random networks. Proc. Roy. Soc. Lond. A., 1976. **351**: p. 351-380.

35. Go, N., T. Noguti, and T. Nishikawa, Dynamics of a small globular protein in terms of low-frequency vibrational modes. Proc. Natl. Acad. Sci. USA., 1983. **80**: p. 3696-3700.

36. Xu, B., et al., Fast and accurate computation schemes for evaluating vibrational entropy of proteins. J. Comp. Chem., 2011. **32**: p. 3188-3193.

37. Balali-Mood, K., P.J. Bond, and M.S.P. Sansom, Interaction of Monotopic Membrane Enzymes with a Lipid Bilayer: A Coarse-Grained MD Simulation Study. Biochemistry., 2009. **48**: p. 2135-2145.

38. Bond, P.J., et al., Assembly of lipoprotein particles revealed by coarse-grained molecular dynamics simulations. J. Struct. Biol., 2007. **157**: p. 579-592.

39. Periole, X., et al., Combining an elastic network with a coarse-grained molecular force field: structure, dynamics and intermolecular recognition. J. Chem. Theory. Comput., 2009. **5**: p. 2531-2543.

40. Go, N. and H. Taketomi, Respective roles of short-range and longrange interactions in protein folding. Proc. Natl. Acad. Sci. USA., 1978. **75**(2): p. 559-563.

41. Karanicolas, J. and C.L. Brooks, Improved Go -like Models Demonstrate the Robustness of Protein Folding Mechanisms Towards Nonnative Interactions. J. Mol. Biol., 2003. **334**: p. 309-325.

42. Clementi, C., H. Nymeyer, and J. Onuchic, Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. J. Mol. Biol., 2000. **298**: p. 937-953.

43. Leopold, P., M. Montal, and J. Onuchic, Protein folding funnels: a kinetic approach to the sequence-structure relationship. Proc. Natl. Acad. Sci. USA., 1992. **89**: p. 8721-8725.

44. Kenzaki, H., et al., CafeMol: A Coarse-Grained Biomolecular Simulator for Simulating Proteins at Work. J. Chem. Theory. Comput., 2011. **7**: p. 1979-1989.

45. Sippl, M.J., Knowledge-based potentials for proteins. Curr. Opin. Struct. Biol., 1995. **5**: p. 229-235.

46. Miyazawa, S. and R.L. Jernigan, Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules, 1985. **18**: p. 534-552.

47. Godzik, A. and J. Skolnick, Sequence-structure matching in globular proteins: application to supersecondary and tertiary structure determination. Proc. Natl. Acad. Sci. USA., 1992. **89**: p. 12098-12102.

48. Kocher, J.P., M.J. Rooman, and S.J. Wodak, Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. J. Mol. Biol., 1994. **16**: p. 1598-1613.

49. Nishikawa, K. and Y. Matsuo, Development of pseudoenergy potentials for assessing protein 3-D-1-D compatibility and detecting weak homologies. Protein Eng., 1993. **6**: p. 811-820.

50. Bryant, S.H. and C.E. Lawrence, An empirical energy function for threading protein sequence through the folding motif. Proteins: Struct. Funct. Genet., 1993. **16**: p. 92-112.

51. Buchete, N.V., S.J. E., and D. Thirumalai, Development of novel statistical potentials for protein fold recognition. Curr. Opin. Struct. Biol., 2004. **14**: p. 225-232.

52. Lazaridis, T. and M. Karplus, Effective energy functions for protein structure prediction. Curr. Opin. Struct. Biol., 2000. **10**: p. 139-145.

53. Skolnick, J., In quest of an empirical potential for protein structure prediction. Curr. Opin. Struct. Biol., 2006. **16**: p. 166-171.

54. Levitt, M. and A. Warshel, Computer simulation of protein folding. Nature, 1975. **235**: p. 694-8.

55. Tozzini, V. and J.A. McCammon, A coarse-grained model for the dynamics of flap opening in HIV-1 protease. Chem. Phys. Lett. , 2005. **413**: p. 123–128.

56. Liwo, A., et al., Calculation of protein backbone geometry from alpha-carbon coordinates based on peptide-group dipole alignment. Protein Sci., 1993. **2**: p. 1697-1714.

57. Liwo, A., et al., Prediction of protein conformation on the basis of a search for compact structures: test on avian pancreatic polypeptide. Protein Sci., 1993. **2**: p. 1715-1731.

58. Ołdziej, S., et al., Physics-based protein-structure prediction using a hierarchical protocol based on the UNRES force field: assessment in two blind tests. Proc. Natl. Acad. Sci. USA., 2005. **102**: p. 7547-7552.

59. Crick, F., Ideas on Protein Synthesis. Symp. Soc. Exp. Biol., 1958. **XII**: p. 138-163.

60. Crick, F., Central dogma of molecular biology. Nature, 1970. **227**(5258): p. 561-3.

61. James D. W, et al., Molecular Biology of the Gene (6th Edition) 2007: Benjamin Cummings.

62. Stark, B.C., et al., Ribonuclease P: an enzyme with an essential RNA component. Proc. Natl. Acad. Sci. USA, 1978. **75**(8): p. 3717-21.

63. Noller, H.F. and J.B. Chaires, Functional modification of 16S ribosomal RNA by kethoxal. Proc. Natl. Acad. Sci. USA, 1972. **69**(11): p. 3115-8.

64. Noller, H.F., et al., Chemical modification of the transfer RNA and polyuridylic acid binding sites of Escherichia coli 30 s ribosomal subunits. J. Mol. Biol., 1971. **61**(3): p. 669-79.

65. Kruger, K., et al., Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. Cell, 1982. **31**(1): p. 147-57.

66. Guerrier-Takada, C., et al., The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. Cell, 1983. **35**(3 Pt 2): p. 849-57.

67. Nahvi, A., et al., Genetic control by a metabolite binding mRNA. Chem. Biol., 2002. **9**(9): p. 1043.

68. Winkler, W., A. Nahvi, and R.R. Breaker, Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. Nature, 2002. **419**(6910): p. 952-6.

69. Winkler, W.C., S. Cohen-Chalamish, and R.R. Breaker, An mRNA structure that controls gene expression by binding FMN. Proc. Natl. Acad. Sci. USA, 2002. **99**(25): p. 15908-13.

70. Rodionov, D.A., et al., Regulation of lysine biosynthesis and transport genes in bacteria: yet another RNA riboswitch? Nucleic. Acids. Res., 2003. **31**(23): p. 6748-57.

71. Vitreschak, A.G., et al., Riboswitches: the oldest mechanism for the regulation of gene expression? Trends Genet, 2004. **20**(1): p. 44-50.

72. Doudna, J.A. and J.W. Szostak, RNA-catalysed synthesis of complementary-strand RNA. Nature, 1989. **339**(6225): p. 519-22.

73. Noller, H.F., V. Hoffarth, and L. Zimniak, Unusual resistance of peptidyl transferase to protein extraction procedures. Science, 1992. **256**(5062): p. 1416-9.

74. Hingerty, B., R.S. Brown, and A. Jack, Further Refinement of Structure of Yeast Transfer-Rna Phe. Journal of Molecular Biology, 1978. **124**(3): p. 523-534.

75. Ban, N., et al., The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. Science, 2000. **289**(5481): p. 905-20.

76. Wimberly, B.T., et al., Structure of the 30S ribosomal subunit. Nature, 2000. **407**(6802): p. 327-39.

77. Brodersen, D.E., et al., Crystal structure of the 30 S ribosomal subunit from Thermus thermophilus: structure of the proteins and their interactions with 16 S RNA. Journal of Molecular Biology, 2002. **316**(3): p. 725-68.

78. Kazantsev, A.V., et al., Crystal structure of a bacterial ribonuclease P RNA. Proc. Natl. Acad. Sci. USA, 2005. **102**(38): p. 13392-7.

79. Torres-Larios, A., et al., Crystal structure of the RNA component of bacterial ribonuclease P. Nature, 2005. **437**(7058): p. 584-7.

80. Serganov, A., L. Huang, and D.J. Patel, Coenzyme recognition and gene regulation by a flavin mononucleotide riboswitch. Nature, 2009. **458**(7235): p. 233-7.

81. Cate, J.H., et al., Crystal structure of a group I ribozyme domain: principles of RNA packing. Science, 1996. **273**(5282): p. 1678-85.

82. Vidovic, I., et al., Crystal structure of the spliceosomal 15.5kD protein bound to a U4 snRNA fragment. Mol Cell, 2000. **6**(6): p. 1331-42.

83. Serganov, A., et al., Structural basis for discriminative regulation of gene expression by adenine- and guanine-sensing mRNAs. Chem Biol, 2004. **11**(12): p. 1729-41.

84. Bessho, Y., et al., Structural basis for functional mimicry of long-variable-arm tRNA by transfer-messenger RNA. Proc Natl Acad Sci U S A, 2007. **104**(20): p. 8293-8.

85. Klein, D.J. and A.R. Ferre-D'Amare, Structural basis of glmS ribozyme activation by glucosamine-6-phosphate. Science, 2006. **313**(5794): p. 1752-6.

86. Serganov, A., L. Huang, and D.J. Patel, Structural insights into amino acid binding and gene control by a lysine riboswitch. Nature, 2008. **455**(7217): p. 1263-7.

87. Dann, C.E., 3rd, et al., Structure and mechanism of a metal-sensing regulatory RNA. Cell, 2007. **130**(5): p. 878-92.

88. Thore, S., M. Leibundgut, and N. Ban, Structure of the eukaryotic thiamine pyrophosphate riboswitch with its regulatory ligand. Science, 2006. **312**(5777): p. 1208-11.

89. Montange, R.K. and R.T. Batey, Structure of the S-adenosylmethionine riboswitch regulatory mRNA element. Nature, 2006. **441**(7097): p. 1172-5.

90. Hainzl, T., S. Huang, and A.E. Sauer-Eriksson, Structure of the SRP19 RNA complex and implications for signal recognition particle assembly. Nature, 2002. **417**(6890): p. 767-71.

91. Kim, S.H., et al., Three-dimensional tertiary structure of yeast phenylalanine transfer RNA. Science, 1974. **185**(149): p. 435-40.

92. Costa, F.F., Non-coding RNAs: meet thy masters. Bioessays, 2007. **32**(7): p. 599-608.

93. Spizzo, R., et al., SnapShot: MicroRNAs in Cancer. Cell, 2009. **137**(3): p. 586-586 e1.

94. Frohlich, K.S. and J. Vogel, Activation of gene expression by small RNA. Current Opinion in Microbiology, 2009. **12**(6): p. 674-82.

95. Georg, J., et al., Evidence for a major role of antisense RNAs in cyanobacterial gene regulation. Mol. Syst. Biol., 2009. **5**: p. 305.

96. Khraiwesh, B., et al., Transcriptional control of gene expression by microRNAs. Cell, 2010. **140**(1): p. 111-122.

97. Hale, C.R., et al., RNA-guided RNA cleavage by a CRISPR RNA-cas protein complex. Cell, 2009. **139**(5): p. 945-956.

98. Marraffini, L.A. and E.J. Sontheimer, CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. Nature Reviews Genetics, 2010. **11**(3): p. 181-190.

99. Hamilton, A.J. and D.C. Baulcombe, A species of small antisense RNA in posttranscriptional gene silencing in plants. Science, 1999. **286**(5441): p. 950-2.

100. Lecellier, C.H., et al., A cellular microRNA mediates antiviral defense in human cells. Science, 2005. **308**(5721): p. 557-60.

101. Buchon, N. and C. Vaury, RNAi: a defensive RNA-silencing against viruses and transposable elements. Heredity, 2006. **96**(2): p. 195-202.

102. Mattick, J.S., R.J. Taft, and G.J. Faulkner, A global view of genomic information - moving beyond the gene and the master regulator. Trends Genet., 2010. **26**(1): p. 21-8.

103. Gupta, R.A., et al., Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature, 2010. **464**(7291): p. 1071-6.

104. Croce, C.M., Causes and consequences of microRNA dysregulation in cancer. Nature Reviews Genetics, 2009. **10**(10): p. 704-14.

105. Sassen, S., E.A. Miska, and C. Caldas, MicroRNA: implications for cancer. Virchows Arch., 2008. **452**(1): p. 1-10.

106. Hahn, M.W. and G.A. Wray, The g-value paradox. Evol Dev, 2002. **4**(2): p. 73-5.

107. Taft, R.J., M. Pheasant, and J.S. Mattick, The relationship between non-protein-coding DNA and eukaryotic complexity. Bioessays, 2007. **29**(3): p. 288-99.

108. Michel, F. and E. Westhof, Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. Journal of Molecular Biology, 1990. **216**(3): p. 585-610.

109. Levitt, M., Detailed molecular model for transfer ribonucleic acid. Nature, 1969. **224**(5221): p. 759-63.

110. Fink, D.L., et al., Computational methods for defining the allowed conformational space of 16S rRNA based on chemical footprinting data. RNA, 1996. **2**(9): p. 851-66.

111. Lehnert, V., et al., New loop-loop tertiary interactions in self-splicing introns of subgroup IC and ID: a complete 3D model of the Tetrahymena thermophila ribozyme. Chem. Biol., 1996. **3**(12): p. 993-1009.

112. Wang, R., et al., Three-dimensional placement of the conserved 530 loop of 16 S rRNA and of its neighboring components in the 30 S subunit. Journal of Molecular Biology, 1999. **286**(2): p. 521-40.

113. Sommer, I. and R. Brimacombe, Methods for refining interactively established models of ribosomal RNA towards a physico-chemically plausible structure. Journal of Computational Chemistry, 2001. **22**(4): p. 407-417.

114. Stagg, S.M., J.A. Mears, and S.C. Harvey, A structural model for the assembly of the 30S subunit of the ribosome. Journal of Molecular Biology, 2003. **328**(1): p. 49-61.

115. Shapiro, B.A., et al., Bridging the gap in RNA structure prediction. Current Opinion in Structural Biology, 2007. **17**(2): p. 157-65.

116. Devkota, B., et al., Structural and electrostatic characterization of pariacoto virus: implications for viral assembly. Biopolymers, 2009. **91**(7): p. 530-8.

117. Meller, J. and R. Elber, Protein recognition by sequence-to-structure fitness: Bridging efficiency and capacity of threading models. Computational Methods for Protein Folding, 2002. **120**: p. 77-130.

118. Kryshtafovych, A., et al., Progress over the first decade of CASP experiments. Proteins-Structure Function and Bioinformatics, 2005. **61**: p. 225-236.

119. Moult, J., A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Current Opinion in Structural Biology, 2005. **15**(3): p. 285-289.

120. Zwieb, C. and F. Muller, Three-dimensional comparative modeling of RNA. Nucleic. Acids. Symp. Ser., 1997. **36**(36): p. 69-71.

121. Jossinet, F. and E. Westhof, Sequence to Structure (S2S): display, manipulate and interconnect RNA data from sequence to structure. Bioinformatics, 2005. **21**(15): p. 3320-1.

122. Wu, J.C., et al., Correlation of RNA Secondary Structure Statistics with Thermodynamic Stability and Applications to Folding. Journal of Molecular Biology, 2009. **391**(4): p. 769-783.

123. Parisien, M. and F. Major, The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. Nature, 2008. **452**(7183): p. 51-55.

124. Pearlman, D.A., et al., Amber, a Package of Computer-Programs for Applying Molecular Mechanics, Normal-Mode Analysis, Molecular-Dynamics and Free-Energy Calculations to Simulate the Structural and Energetic Properties of Molecules. Computer Physics Communications, 1995. **91**(1-3): p. 1-41.

125. Cheatham, T.E., 3rd, P. Cieplak, and P.A. Kollman, A modified version of the Cornell et al. force field with improved sugar pucker phases and helical repeat. Journal of Biomolecular Structure and Dynamics, 1999. **16**(4): p. 845-62.

126. Case, D.A., et al., The Amber biomolecular simulation programs. Journal of Computational Chemistry, 2005. **26**(16): p. 1668-88.

127. Perez, A., et al., Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. Biophys. J., 2007. **92**(11): p. 3817-29.

128. Brooks, B.R., et al., CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. Journal of Computational Chemistry, 1983. **4**(2): p. 187-217.

129. MacKerell, A.D., Jr., N. Banavali, and N. Foloppe, Development and current status of the CHARMM force field for nucleic acids. Biopolymers, 2000. **56**(4): p. 257-65.

130. Brooks, B.R., et al., CHARMM: the biomolecular simulation program. Journal of Computational Chemistry, 2009. **30**(10): p. 1545-614.

131. Freddolino, P.L., et al., Molecular dynamics simulations of the complete satellite tobacco mosaic virus. Structure, 2006. **14**(3): p. 437-49.

132. Golubkov, P.A. and P.Y. Ren, Generalized coarse-grained model based on point multipole and Gay-Berne potentials. Journal of Chemical Physics, 2006. **125**(6): p. -.

133. Scheraga, H.A., M. Khalili, and A. Liwo, Protein-folding dynamics: Overview of molecular simulation techniques. Annual Review of Physical Chemistry, 2007. **58**: p. 57-83.

134. Golubkov, P.A., J.C. Wu, and P.Y. Ren, A transferable coarse-grained model for hydrogen-bonding liquids. Physical Chemistry Chemical Physics, 2008. **10**(15): p. 2050-2057.

135. Norman L. Allinger, Young H. Yuh, and J.-H. Lii, Molecular Mxhanics. The MM3 Force Field for Hydrocarbon 1. Journal of the American Chemical Society, 1989. **111**(23): p. 8851-8566.

136. Massire, C. and E. Westhof, MANIP: An interactive tool for modelling RNA. Journal of Molecular Graphics & Modelling, 1998. **16**(4-6): p. 197-205, 255-257.

137. Tanaka, I., et al., Matching the crystallographic structure of ribosomal protein S7 to a three-dimensional model of the 16S ribosomal RNA. RNA, 1998. **4**(5): p. 542-550.

138. Nielsen, S.O., et al., Coarse grain models and the computer simulation of soft materials. Journal of Physics-Condensed Matter, 2004. **16**(15): p. R481-R512.

139. Tepper, H.L. and G.A. Voth, A coarse-grained model for double-helix molecules in solution: Spontaneous helix formation and equilibrium properties. Journal of Chemical Physics, 2005. **122**(12): p. 124906.

140. Li, X.J., et al., Developing a coarse-grained force field for the diblock copolymer poly(styrene-b-butadiene) from atomistic simulation. Journal of Chemical Physics, 2006. **124**(20): p. -.

141. Jonikas, M.A., et al., Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. RNA, 2009. **15**(2): p. 189-199.

142. Cao, S. and S.J. Chen, Predicting RNA folding thermodynamics with a reduced chain representation model. RNA, 2005. **11**(12): p. 1884-1897.

143. Cannone, J.J., et al., The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinformatics, 2002. **3**: p. 2.

144. Duarte, C.M. and A.M. Pyle, Stepping through an RNA structure: A novel approach to conformational analysis. Journal of Molecular Biology, 1998. **284**(5): p. 1465-1478.

145. Tschop, W., et al., Simulation of polymer melts. I. Coarse-graining procedure for polycarbonates. Acta Polymerica, 1998. **49**(2-3): p. 61-74.

146. Muller-Plathe, F., Coarse-graining in polymer simulation: From the atomistic to the mesoscopic scale and back. Chemphyschem, 2002. **3**(9): p. 754-769.

147. Buckingham, R.A., The Classical Equation of State of Gaseous Helium, Neon and Argon. Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences, 1938. **168**: p. 264-283.

148. Allinger, N.L., Y.H. Yuh, and J.H. Lii, Molecular Mechanics - the Mm3 Force-Field for Hydrocarbons .1. Journal of the American Chemical Society, 1989. **111**(23): p. 8551-8566.

149. Lii, J.H. and N.L. Allinger, Molecular Mechanics - the Mm3 Force-Field for Hydrocarbons .2. Vibrational Frequencies and Thermodynamics. Journal of the American Chemical Society, 1989. **111**(23): p. 8566-8575.

150. Lii, J.H. and N.L. Allinger, Molecular Mechanics - the Mm3 Force-Field for Hydrocarbons .3. The Vanderwaals Potentials and Crystal Data for Aliphatic and Aromatic-Hydrocarbons. Journal of the American Chemical Society, 1989. **111**(23): p. 8576-8582.

151. Summa, C.M. and M. Levitt, Near-native structure refinement using in vacuo energy minimization. Proc. Natl. Acad. Sci. USA, 2007. **104**(9): p. 3177-3182.

152. Skolnick, J., et al., Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct? Protein. Sci., 1997. **6**(3): p. 676-88.

153. Shanno, D.F. and K.-H. Phua, Matrix Conditioning and Nonlinear Optimization. Mathematical Programming, 1977. **25**: p. 507-518.

154. Davidon, W.C., Optimally Conditioned Optimization Algorithms without Line Searches. Mathematical Programming, 1975. **9**(1): p. 1-30.

155. Shanno, D.F. and K.H. Phua, Numerical Comparison of Several Variable-Metric Algorithms. Journal of Optimization Theory and Applications, 1978. **25**(4): p. 507-518.

156. J.W.Ponder, TINKER molecular modeling package. Washington University Medical School.

157. D.A. Case, et al., AMBER 10. University of California, San Francisco, 2008.

158. Jorgensen, W.L., et al., Comparison of simple potential functions for simulating liquid water. Journal of Chemical Physics, 1983. **79**(2): p. 926-935.

159. Neria, E. and M. Karplus, A position dependent friction model for solution reactions in the high friction regime: Proton transfer in triosephosphate isomerase (TIM). Journal of Chemical Physics, 1996. **105**(24): p. 10812-10818.

160. Still, W.C., et al., Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics. Journal of the American Chemical Society, 1990. **112**(16): p. 6127-6129.

161. Schaefer, M. and M. Karplus, A comprehensive analytical treatment of continuum electrostatics. Journal of Physical Chemistry, 1996. **100**(5): p. 1578-1599.

162. Scarsi, M., J. Apostolakis, and A. Caflisch, Continuum electrostatic energies of macromolecules in aqueous solutions. Journal of Physical Chemistry A, 1997. **101**(43): p. 8098-8106.

163. Dominy, B.N. and C.L. Brooks, Development of a generalized born model parametrization for proteins and nucleic acids. Journal of Physical Chemistry B, 1999. **103**(18): p. 3765-3773.

164. Bashford, D. and D.A. Case, Generalized born models of macromolecular solvation effects. Annual Review of Physical Chemistry, 2000. **51**: p. 129-152.

165. Feig, M., W. Im, and C.L. Brooks, Implicit solvation based on generalized Born theory in different dielectric environments. Journal of Chemical Physics, 2004. **120**(2): p. 903-911.

166. Bussi, G., T. Zykova-Timan, and M. Parrinello, Isothermal-isobaric molecular dynamics using stochastic velocity rescaling. Journal of Chemical Physics, 2009. **130**(7).

167. Klosterman, P.S., S.A. Shah, and T.A. Steitz, Crystal structures of two plasmid copy control related RNA duplexes: An 18 base pair duplex at 1.20 A resolution and a 19 base pair duplex at 1.55 A resolution. Biochemistry, 1999. **38**(45): p. 14784-92.

168. Popenda, L., R.W. Adamiak, and Z. Gdaniec, Bulged adenosine influence on the RNA duplex conformation in solution. Biochemistry, 2008. **47**(18): p. 5059-67.

169. Deng, J., et al., Structure of an RNA dodecamer containing a fragment from SRP domain IV of Escherichia coli. Acta. Crystallogr. D. Biol. Crystallogr., 2003. **59**(Pt 6): p. 1004-11.

170. Conn, G.L., et al., A compact RNA tertiary structure contains a buried Backbone-K+ complex. Journal of Molecular Biology, 2002. **318**(4): p. 963-973.

171. Draper, D.E., RNA Folding: Thermodynamic and Molecular Descriptions of the Roles of Ions. Biophys. J., 2008. **95**(12): p. 5489-5495.

172. Draper, D.E., D. Grilley, and A.M. Soto, Ions and RNA folding. Annual Review of Biophysics and Biomolecular Structure, 2005. **34**: p. 221-243.

173. Misra, V.K. and D.E. Draper, A thermodynamic framework for Mg2+ binding to RNA. Proc. Natl. Acad. Sci. U.S.A., 2001. **98**(22): p. 12456-12461.

174. Egli, M., et al., Metal ions and flexibility in a viral RNA pseudoknot at atomic resolution. Proc. Natl. Acad. Sci. U.S.A., 2002. **99**(7): p. 4302-4307.

175. Burke, J.E., et al., Structure of the yeast U2/U6 snRNA complex. RNA, 2012. **18**(4): p. 673-83.

176. Schneidman-Duhovny, D., M. Hammel, and A. Sali, FoXS: a web server for rapid computation and fitting of SAXS profiles. Nucleic. Acids. Res., 2010. **38**(Web Server issue): p. W540-4.

177. *http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi*. RNAfold WebServer.

178. Laing, C. and T. Schlick, Computational approaches to 3D modeling of RNA. J. Phys. Condens. Matter, 2010. **22**(28): p. 283101.

179. Cruz, J.A., et al., RNA-Puzzles: A CASP-like evaluation of RNA three-dimensional structure prediction. Rna-a Publication of the Rna Society, 2012. **18**(4): p. 610-625.

180. Zuker, M., Mfold web server for nucleic acid folding and hybridization prediction. Nucleic. Acids. Res., 2003. **31**(13): p. 3406-15.

181. Kern, D. and E.R.P. Zuiderweg, The role of dynamics in allosteric regulation. Current Opinion In Structural Biology, 2003. **13**(6): p. 748-757.

182. Shelley, J.C., et al., A coarse grain model for phospholipid simulations. Journal of Physical Chemistry B, 2001. **105**(19): p. 4464-4470.

183. DeVane, R., et al., Coarse-Grained Potential Models for Phenyl-Based Molecules: I. Parametrization Using Experimental Data. Journal of Physical Chemistry B, 2010. **114**(19): p. 6386-6393.

184. Hills, R.D., L.Y. Lu, and G.A. Voth, Multiscale Coarse-Graining of the Protein Energy Landscape. PLoS Computational Biology, 2010. **6**(6).

185. Makowski, M., et al., 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. Journal of Physical Chemistry B, 2008. **112**(36): p. 11385-11395.

186. Maisuradze, G.G., et al., Evidence, from simulations, of a single state with residual native structure at the thermal denaturation midpoint of a small globular protein. Journal of the American Chemical Society, 2010. **132**(27): p. 9444-52.

187. Liwo, A., et al., Cumulant-based expressions for the multibody terms for the correlation between local and electrostatic interactions in the united-residue force field. Journal of Chemical Physics, 2001. **115**(5): p. 2323-2347.

188. Wu, C. and J.E. Shea, Coarse-grained models for protein aggregation. Current Opinion In Structural Biology, 2005. **21**(2): p. 209-220.

189. Liwo, A., et al., 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. Journal of Physical Chemistry B, 2007. **111**(1): p. 260-285.

190. Liwo, A., et al., 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.

191. C. Czaplewski, et al., Coarse-Grained Models of Proteins: Theory and Applications in Multiscale Approaches to Protein Modeling, A. Kolinski, Editor 2010, Springer.

192. Voth, G., ed. Coarse-Graining of Condensed Phase and Biomolecular Systems. 2008, CRC Press, Taylor & Francis Group: Farmington, CT.

193. Xia, Z., et al., Coarse-grained model for simulation of RNA three-dimensional structures. Journal of Physical Chemistry B, 2010. **114**(42): p. 13497-506.

194. Schnieders, M.J. and J.W. Ponder, Polarizable atomic multipole solutes in a generalized Kirkwood continuum. Journal of Chemical Theory and Computation, 2007. **3**(6): p. 2083-2097.

195. Golubkov, P.A. and P.Y. Ren, Generalized coarse-grained model based on point multipole and Gay-Berne potentials. Journal of Chemical Physics, 2006. **125**(6): p. 64103.

196. Cleaver, D.J., et al., Extension and generalization of the Gay-Berne potential. Physical Review E, 1996. **54**(1): p. 559-567.

197. Halgren, T.A., Representation of Vanderwaals (Vdw) Interactions in Molecular Mechanics Force-Fields - Potential Form, Combination Rules, and Vdw Parameters. Journal of the American Chemical Society, 1992. **114**(20): p. 7827-7843.

198. Ponder, J.W., TINKER molecular modeling package. Washington University Medical School, 2010.

199. Ren, P. and J.W. Ponder, Polarizable Atomic Multipole Water Model for Molecular Mechanics Simulation. Journal of Physical Chemistry B, 2003. **107**: p. 5933-5947.

200. Ponder, J.W., et al., Current Status of the AMOEBA Polarizable Force Field. Journal of Physical Chemistry B, 2010. **114**(8): p. 2549-2564.

201. Andersen, H.C., Rattle - a Velocity Version of the Shake Algorithm for Molecular-Dynamics Calculations. Journal Of Computational Physics, 1983. **52**(1): p. 24-34.

202. Constanciel, R. and R. Contreras, Self-Consistent Field-Theory of Solvent Effects Representation by Continuum Models - Introduction of Desolvation Contribution. Theoretica Chimica Acta, 1984. **65**(1): p. 1-11.

203. Grycuk, T., Deficiency of the Coulomb-field approximation in the generalized Born model: An improved formula for Born radii evaluation. Journal of Chemical Physics, 2003. **119**(9): p. 4817-4826.

204. Qiu, D., et al., The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii. Journal of Physical Chemistry A, 1997. **101**(16): p. 3005-3014.

205. Best, R.B., N.V. Buchete, and G. Hummer, Are current molecular dynamics force fields too helical? Biophysical Journal, 2008. **95**(1): p. L7-L9.

206. Jorgensen, W.L., D.S. Maxwell, and J. TiradoRives, Development and testing of the OPLS all-atom force field on conformational energetics and properties of

organic liquids. Journal of the American Chemical Society, 1996. **118**(45): p. 11225-11236.

207. Headgordon, T., et al., Theoretical-Study of Blocked Glycine and Alanine Peptide Analogs. Journal of the American Chemical Society, 1991. **113**(16): p. 5989-5997.

208. Chou, P.Y. and G.D. Fasman, Conformational Parameters for Amino-Acids in Helical, Beta-Sheet, and Random Coil Regions Calculated from Proteins. Biochemistry, 1974. **13**(2): p. 211-222.

209. Richardson, J.S. and D.C. Richardson, Amino-Acid Preferences for Specific Locations at the Ends of Alpha-Helices. Science, 1988. **240**(4859): p. 1648-1652.

210. Hudgins, R.R., M.A. Ratner, and M.F. Jarrold, Design of helices that are stable in vacuo. Journal of the American Chemical Society, 1998. **120**(49): p. 12974-12975.

211. Levy, Y., J. Jortner, and O.M. Becker, Solvent effects on the energy landscapes and folding kinetics of polyalanine. Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(5): p. 2188-2193.

212. Counterman, A.E. and D.E. Clemmer, Large anhydrous polyalanine ions: Evidence for extended helices and onset of a more compact state. Journal of the American Chemical Society, 2001. **123**(7): p. 1490-1498.

213. Henzler, K.A., D.K. Lee, and A. Ramamoorthy, Conformational stability of solid-state poly(l-alanine). Biophysical Journal, 2001. **80**(1): p. 187a-187a.

214. Nguyen, H.D., A.J. Marchut, and C.K. Hall, Solvent effects on the conformational transition of a model polyalanine peptide. Protein Science, 2004. **13**(11): p. 2909-2924.

215. Soto, P., A. Baumketner, and J.E. Shea, Aggregation of polyalanine in a hydrophobic environment. Journal of Chemical Physics, 2006. **124**(13): p. 134904.

216. Zhou, J., et al., Coarse-grained peptide modeling using a systematic multiscale approach. Biophysical Journal, 2007. **92**(12): p. 4289-4303.

217. Chu, J.W., S. Izvekov, and G.A. Voth, The multiscale challenge for biomolecular systems: coarse-grained modeling. Molecular Simulation, 2006. **32**(3-4): p. 211-218.

218. Jarrold, M.F., Helices and sheets in vacuo. Physical Chemistry Chemical Physics, 2007. **9**(14): p. 1659-1671.

219. Graf, J., et al., Structure and dynamics of the homologous series of alanine peptides: A joint molecular dynamics/NMR study. Journal of the American Chemical Society, 2007. **129**(5): p. 1179-1189.

220. Albrieux, F., et al., Conformation of Polyalanine and Polyglycine Dications in the Gas Phase: Insight from Ion Mobility Spectrometry and Replica-Exchange Molecular Dynamics. Journal of Physical Chemistry A, 2010. **114**(25): p. 6888-6896.

221. Hegefeld, W.A., et al., Helix Formation in a Pentapeptide Experiment and Force-field Dependent Dynamics. Journal of Physical Chemistry A, 2010. **114**(47): p. 12391-12402.

222. Penev, E.S., S. Lampoudi, and J.E. Shea, TiREX: Replica-exchange molecular dynamics using TINKER. Computer Physics Communications, 2009. **180**(10): p. 2013-2019.

223. McCool, M.A., A.F. Collings, and L.A. Woolf, Pressure and temperature dependence of the self-diffusion of benzene. Journal of the Chemical Society, Faraday Transactions 1, 1972. **68**: p. 1489 - 1497.

224. Wensink, E.J.W., et al., Dynamic properties of water/alcohol mixtures studied by computer simulation. Journal of Chemical Physics, 2003. **119**(14): p. 7308-7317.

225. Weast, R.C., CRC handbook of chemistry and physics. 1st Student ed1988, Boca Raton, FL: CRC Press. 1 v. (various pagings).

226. Mikhail, S.Z. and W.R. Kimel, Densities and Viscosities of Methanol-Water Mixtures. Journal of Chemical and Engineering Data, 1961. **6**(4): p. 533 - 537.

227. Ren, P., et al., Biomolecular electrostatics and solvation: a computational perspective. Quarterly Reviews of Biophysics, 2011. **45**(4): p. 427-491.

228. Cornell, W.D., et al., A 2nd Generation Force-Field for the Simulation of Proteins, Nucleic-Acids, and Organic-Molecules. Journal of the American Chemical Society, 1995. **117**(19): p. 5179-5197.

229. MacKerell, A.D., et al., All-atom empirical potential for molecular modeling and dynamics studies of proteins. Journal of Physical Chemistry B, 1998. **102**(18): p. 3586-3616.

230. Valdes, H., et al., Benchmark database on isolated small peptides containing an aromatic side chain: comparison between wave function and density functional theory methods and empirical force field. Physical Chemistry Chemical Physics, 2008. **10**(19): p. 2747-2757.

231. Jorgensen, W.L., D.S. Maxwell, and J. Tirado-Rives, Development and testing of the OPLS all-atom force field on conformational energetics and properties of

organic liquids. Journal of the American Chemical Society, 1996. **118**(45): p. 11225-11236.

232. Rezac, J., et al., Quantum Chemical Benchmark Energy and Geometry Database for Molecular Clusters and Complex Molecular Systems (*Www.Begdb.Com):* A Users Manual and Examples. Collection of Czechoslovak Chemical Communications, 2008. **73**(10): p. 1261-1270.

233. Ponder, J.W. and D.A. Case, Force Fields for Protein Simulations, in Advances in Protein Chemistry, D. Valerie, Editor 2003, Academic Press. p. 27-85.

234. Barker, J.A., Statistical Mechanics of Interacting Dipoles. Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences, 1953. **219**(1138): p. 367-372.

235. Caldwell, J.W. and P.A. Kollman, Structure and Properties of Neat Liquids Using Nonadditive Molecular Dynamics: Water, Methanol, and N-Methylacetamide. Journal of Physical Chemistry, 1995. **99**: p. 6208-6219.

236. Cieplak, P., et al., Polarization effects in molecular mechanical force fields. Journal of Physics-Condensed Matter, 2009. **21**(33): p. 333102-333123.

237. Friesner, R.A., Modeling polarization in proteins and protein-ligand complexes: Methods and preliminary results. Advances in Protein Chemistry and Structural Biology, 2006. **72**: p. 79-104.

238. Holt, A. and G. Karlström, Inclusion of the quadrupole moment when describing polarization. The effect of the dipole-quadrupole polarizability. Journal of Computational Chemistry, 2008. **29**(12): p. 2033-2038.

239. Kaminski, G.A., et al., Development of a polarizable force field for proteins via ab initio quantum chemistry: First generation model and gas phase tests. Journal of Computational Chemistry, 2002. **23**(16): p. 1515-1531.

240. Moghaddam, S., et al., New ultrahigh affinity host-guest complexes of cucurbit[7]uril with bicyclo[2.2.2]octane and adamantane guests: thermodynamic analysis and evaluation of M2 affinity calculations. Journal of the American Chemical Society, 2011. **133**(10): p. 3570-81.

241. Molnar, L.F., et al., Further analysis and comparative study of intermolecular interactions using dimers from the S22 database. The Journal of Chemical Physics, 2009. **131**(6): p. 065102.

242. Ren, P. and J.W. Ponder, Consistent treatment of inter- and intramolecular polarization in molecular mechanics calculations. Journal of Computational Chemistry, 2002. **23**(16): p. 1497-1506.

243. Ren, P. and J.W. Ponder, Polarizable atomic multipole water model for molecular mechanics simulation. Journal of Physical Chemistry B, 2003. **107**(24): p. 5933-5947.

244. Wang, Z.X., et al., Strike a balance: Optimization of backbone torsion parameters of AMBER polarizable force field for simulations of proteins and peptides (vol 27, pg 781, 2006). Journal of Computational Chemistry, 2006. **27**(8): p. 994-994.

245. Geerke, D.P., W.F. van Gunsteren, and Sk, Calculation of the free energy of polarization: Quantifying the effect of explicitly treating electronic polarization on the transferability of force-field parameters. Journal of Physical Chemistry B, 2007. **111**(23): p. 6425-6436.

246. Lamoureux, G., et al., A polarizable model of water for molecular dynamics simulations of biomolecules. Chemical Physics Letters, 2006. **418**(1-3): p. 245-249.

247. Lamoureux, G., A.D. MacKerell, and B. Roux, A simple polarizable model of water based on classical Drude oscillators. Journal of Chemical Physics, 2003. **119**(10): p. 5185-5197.

248. Banks, J.L., et al., Parametrizing a Polarizable Force Field from ab Initio Data. I. The Fluctuating Point Charge Model. Journal of Chemical Physics, 1999. **110**(2): p. 741-754.

249. Patel, S. and C.L. Brooks III, CHARMM fluctuating charge force field for proteins: I parameterization and application to bulk organic liquid simulations. Journal of Computational Chemistry, 2004. **25**(1): p. 1-15.

250. Rappe, A.K. and W.A. Goddard, Charge Equilibration for Molecular-Dynamics Simulations. Journal of Physical Chemistry, 1991. **95**(8): p. 3358-3363.

251. van Belle, D. and S.J. Wodak, Extended Lagrangian Formalism Applied to Temperature Control and Electronic Polarization Effects in Molecular Dynamics Simulations. Computer Physics Communications, 1995. **91**: p. 253-262.

252. Vanbelle, D., et al., Molecular-Dynamics Simulation of Polarizable Water by an Extended Lagrangian Method. Molecular Physics, 1992. **77**(2): p. 239-255.

253. Ponder, J.W., et al., Current Status of the AMOEBA Polarizable Force Field. The Journal of Physical Chemistry B, 2010. **114**(8): p. 2549-2564.

254. Stone, A.J., Distributed Multipole Analysis: Methods and Applications. Molecular Physics, 1985. **56**(5): p. 1047-1064.

255. Stone, A.J., Distributed multipole analysis: Stability for large basis sets. Journal of Chemical Theory and Computation, 2005. **1**(6): p. 1128-1132.

256. Ren, P., C. Wu, and J.W. Ponder, Polarizable Atomic Multipole-based Molecular Mechanics for Organic Molecules. Journal of Chemical Theory and Computation, 2011. **7**(10): p. 3143-3161 PMCID: PMC3196664.

257. Hornak, V., et al., Comparison of multiple amber force fields and development of improved protein backbone parameters. Proteins, 2006. **65**(3): p. 712-725.

258. Lindorff-Larsen, K., et al., Improved side-chain torsion potentials for the Amber ff99SB protein force field. Proteins., 2010. **78**(8): p. 1950-8.

259. Best, R.B., et al., Inclusion of Many-Body Effects in the Additive CHARMM Protein CMAP Potential Results in Enhanced Cooperativity of alpha-Helix and beta-Hairpin Formation. Biophysical Journal, 2012. **103**(5): p. 1045-1051.

260. Best, R.B., et al., Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone phi, psi and Side-Chain chi(1) and chi(2) Dihedral Angles. Journal of Chemical Theory and Computation, 2012. **8**(9): p. 3257-3273.

261. Essmann, U., et al., A Smooth Particle Mesh Ewald Method. Journal of Chemical Physics, 1995. **103**(19): p. 8577-8593.

262. Darden, T., D. York, and L. Pedersen, Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems. Journal of Chemical Physics, 1993. **98**(12): p. 10089-10092.

263. Sagui, C., L.G. Pedersen, and T.A. Darden, Towards an accurate representation of electrostatics in classical force fields: Efficient implementation of multipolar interactions in biomolecular simulations. Journal of Chemical Physics, 2004. **120**(1): p. 73-87.

264. Verlet, L., Computer Experiments on Classical Fluids .I. Thermodynamical Properties of Lennard-Jones Molecules. Physical Review, 1967. **159**(1): p. 98-&.

265. Martyna, G.J., et al., Explicit Reversible Integrators for Extended Systems Dynamics. Molecular Physics, 1996. **87**: p. 1117-1157.

266. Roux, B., The Calculation of the Potential of Mean Force Using Computer-Simulations. Computer Physics Communications, 1995. **91**(1-3): p. 275-282.

267. Kumar, S., et al., The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules .1. The Method. Journal of Computational Chemistry, 1992. **13**(8): p. 1011-1021.

268. Kumar, S., et al., Multidimensional Free-Energy Calculations Using the Weighted Histogram Analysis Method. Journal of Computational Chemistry, 1995. **16**(11): p. 1339-1350.

269. Ting, D., et al., Neighbor-Dependent Ramachandran Probability Distributions of Amino Acids Developed from a Hierarchical Dirichlet Process Model. PLoS Comput. Biol., 2010. **6**(4).

270. Patriksson, A. and D. van der Spoel, A temperature predictor for parallel tempering simulations. Physical Chemistry Chemical Physics, 2008. **10**(15): p. 2073-2077.

271. Best, R.B. and G. Hummer, Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. Journal of Physical Chemistry B, 2009. **113**(26): p. 9004-15.

272. Beachy, M.D., et al., Accurate ab initio quantum chemical determination of the relative energetics of peptide conformations and assessment of empirical force fields. Journal of the American Chemical Society, 1997. **119**(25): p. 5908-5920.

273. Best, R.B. and J. Mittal, Balance between a and beta Structures in Ab Initio Protein Folding. Journal of Physical Chemistry B, 2010. **114**(26): p. 8790-8798.

274. Lindorff-Larsen, K., et al., Systematic Validation of Protein Force Fields against Experimental Data. Plos One, 2012. **7**(2).

275. Karplus, M., Contact Electron-Spin Coupling of Nuclear Magnetic Moments. Journal of Chemical Physics, 1959. **30**(1): p. 11-15.

276. Aliev, A.E. and D. Courtier-Murias, Experimental Verification of Force Fields for Molecular Dynamics Simulations Using Gly-Pro-Gly-Gly. Journal of Physical Chemistry B, 2010. **114**(38): p. 12358-12375.

277. Mackerell, A.D., M. Feig, and C.L. Brooks, Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. Journal of Computational Chemistry, 2004. **25**(11): p. 1400-1415.

278. Bartel, D.P., MicroRNAs: target recognition and regulatory functions. Cell, 2009. **136**(2): p. 215-33.

279. Bartel, D.P., MicroRNAs: genomics, biogenesis, mechanism, and function. Cell, 2004. **116**(2): p. 281-97.

280. Berezikov, E., et al., Mammalian mirtron genes. Molecular Cell, 2007. **28**(2): p. 328-336.

281. Filipowicz, W., S.N. Bhattacharyya, and N. Sonenberg, Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nature Reviews: Genetics, 2008. **9**(2): p. 102-14.

282. Orom, U.A., F.C. Nielsen, and A.H. Lund, MicroRNA-10a binds the 5'UTR of ribosomal protein mRNAs and enhances their translation. Molecular Cell, 2008. **30**(4): p. 460-471.

283. Miranda, K.C., et al., A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. Cell, 2006. **126**(6): p. 1203-17.

284. Duursma, A.M., et al., miR-148 targets human DNMT3b protein coding region. RNA, 2008. **14**(5): p. 872-7.

285. Lal, A., et al., p16(INK4a) translation suppressed by miR-24. PLoS ONE, 2008. **3**(3): p. e1864.

286. Tay, Y., et al., MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. Nature, 2008. **455**(7216): p. 1124-8.

287. Forman, J.J., A. Legesse-Miller, and H.A. Coller, A search for conserved sequences in coding regions reveals that the let-7 microRNA targets Dicer within its coding sequence. Proc. Natl. Acad. Sci. USA, 2008. **105**(39): p. 14879-84.

288. Shen, W.F., et al., MicroRNA-126 regulates HOXA9 by binding to the homeobox. Molecular and Cellular Biology, 2008. **28**(14): p. 4609-19.

289. Rigoutsos, I., New tricks for animal microRNAS: targeting of amino acid coding regions at conserved and nonconserved sites. Cancer Res, 2009. **69**(8): p. 3245-8.

290. Ha, I., B. Wightman, and G. Ruvkun, A bulged lin-4/lin-14 RNA duplex is sufficient for Caenorhabditis elegans lin-14 temporal gradient formation. Genes Dev, 1996. **10**(23): p. 3041-50.

291. Lee, R.C., R.L. Feinbaum, and V. Ambros, The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell, 1993. **75**(5): p. 843-54.

292. Reinhart, B.J., et al., The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. Nature, 2000. **403**(6772): p. 901-6.

293. Wightman, B., I. Ha, and G. Ruvkun, Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. Cell, 1993. **75**(5): p. 855-62.

294. Didiano, D. and O. Hobert, Molecular architecture of a miRNA-regulated 3' UTR. RNA, 2008. **14**(7): p. 1297-317.

295. Esquela-Kerscher, A. and F.J. Slack, Oncomirs - microRNAs with a role in cancer. Nat Rev Cancer, 2006. **6**(4): p. 259-69.

296. Hammond, S.M., MicroRNAs as tumor suppressors. Nat Genet, 2007. **39**(5): p. 582-3.

297. Poliseno, L., et al., A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature, 2010. **465**(7301): p. 1033-8.

298. Ma, L., J. Teruya-Feldstein, and R.A. Weinberg, Tumour invasion and metastasis initiated by microRNA-10b in breast cancer. Nature, 2007. **449**(7163): p. 682-8.

299. Nelson, P.T., W.X. Wang, and B.W. Rajeev, MicroRNAs (miRNAs) in neurodegenerative diseases. Brain Pathol, 2008. **18**(1): p. 130-8.

300. Calin, G.A. and C.M. Croce, MicroRNA signatures in human cancers. Nat Rev Cancer, 2006. **6**(11): p. 857-66.

301. Deng, S., et al., Mechanisms of microRNA deregulation in human cancer. Cell Cycle, 2008. **7**(17): p. 2643-6.

302. Godshalk, S.E., et al., MicroRNAs and cancer: a meeting summary of the eponymous Keystone Conference. Epigenetics, 2010. **5**(2): p. 164-8.

303. Ryan, B.M., A.I. Robles, and C.C. Harris, Genetic variation in microRNA networks: the implications for cancer research. Nat Rev Cancer, 2010. **10**(6): p. 389-402.

304. Saey, T., Cancer's little helpers: Tiny pieces of RNA may turn cells to the dark side. Science News, 2010.

305. Spizzo, R., et al., SnapShot: MicroRNAs in Cancer. Cell, 2009. **137**(3): p. 586-586.e1.

306. Ventura, A. and T. Jacks, MicroRNAs and cancer: short RNAs go a long way. Cell, 2009. **136**(4): p. 586-91.

307. Voorhoeve, P. and R. Agami, Classifying microRNAs in cancer: The good, the bad and the ugly. Biochimica et Biophysica Acta (BBA)-Reviews on Cancer, 2007. **1775**(2): p. 274-282.

308. Small, E.M., R.J.A. Frost, and E.N. Olson, MicroRNAs add a new dimension to cardiovascular disease. Circulation, 2010. **121**(8): p. 1022-32.

309. Small, E. and E. Olson, Pervasive roles of microRNAs in cardiovascular biology. Nature, 2011. **469**(21248840): p. 336-42.

310. Taganov, K.D., M.P. Boldin, and D. Baltimore, MicroRNAs and immunity: tiny players in a big field. Immunity, 2007. **26**(2): p. 133-7.

311. Wang, W.-X., et al., miR-107 regulates granulin/progranulin with implications for traumatic brain injury and neurodegenerative disease. Am J Pathol, 2010. **177**(1): p. 334-45.

312. Wang, W.X., et al., The expression of microRNA miR-107 decreases early in Alzheimer's disease and may accelerate disease progression through regulation of beta-site amyloid precursor protein-cleaving enzyme 1. J Neurosci, 2008. **28**(5): p. 1213-23.

313. Abelson, J.F., et al., Sequence variants in SLITRK1 are associated with Tourette's syndrome. Science, 2005. **310**(5746): p. 317-20.

314. Perkins, D.O., et al., microRNA expression in the prefrontal cortex of individuals with schizophrenia and schizoaffective disorder. Genome Biol, 2007. **8**(2): p. R27.

315. Griffiths-Jones, S., miRBase: the microRNA sequence database. Methods in molecular biology, 2006. **342**: p. 129-38.

316. Burgler, C. and P.M. Macdonald, Prediction and verification of microRNA targets by MovingTargets, a highly adaptable prediction method. BMC Genomics, 2005. **6**(1): p. 88.

317. Enright, A.J., et al., MicroRNA targets in Drosophila. Genome Biol, 2003. **5**(1): p. R1.

318. John, B., et al., Human MicroRNA targets. PLoS Biol, 2004. **2**(11): p. e363.

319. Kertesz, M., et al., The role of site accessibility in microRNA target recognition. Nat Genet, 2007. **39**(10): p. 1278-84.

320. Lewis, B.P., et al., Prediction of mammalian microRNA targets. Cell, 2003. **115**(7): p. 787-98.

321. Moxon, S., V. Moulton, and J. Kim, A scoring matrix approach to detecting miRNA target sites. Algorithms for Molecular Biology, 2008. **3**(1): p. 3.

322. Rajewsky, N. and N.D. Socci, Computational identification of microRNA targets. Dev Biol, 2004. **267**(2): p. 529-35.

323. Stark, A., et al., Identification of Drosophila MicroRNA targets. PLoS Biol, 2003. **1**(3): p. E60.

324. Rehmsmeier, M., et al., Fast and effective prediction of microRNA/target duplexes. RNA, 2004. **10**(10): p. 1507-17.

325. Ritchie, W., S. Flamant, and J.E.J. Rasko, Predicting microRNA targets and functions: traps for the unwary. Nat Methods, 2009. **6**(6): p. 397-398.

326. Rigoutsos, I. and A. Tsirigos, MicroRNA Target Prediction in MicroRNAs in Development and Cancer, F.J. Slack, Editor 2010, Imperial College Press.

327. Selbach, M., et al., Widespread changes in protein synthesis induced by microRNAs. Nature, 2008. **455**(7209): p. 58-63.

328. Baek, D., et al., The impact of microRNAs on protein output. Nature, 2008. **455**(7209): p. 64-71.

329. Didiano, D. and O. Hobert, Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions. Nat Struct Mol Biol, 2006. **13**(9): p. 849-51.

330. Easow, G., A.A. Teleman, and S.M. Cohen, Isolation of microRNA targets by miRNP immunopurification. RNA, 2007. **13**(8): p. 1198-204.

331. Lal, A., et al., miR-24 Inhibits cell proliferation by targeting E2F2, MYC, and other cell-cycle genes via binding to "seedless" 3'UTR microRNA recognition elements. Molecular Cell, 2009. **35**(5): p. 610-625.

332. Chi, S.W., et al., Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. Nature, 2009. **460**(7254): p. 479-86.

333. Zisoulis, D.G., et al., Comprehensive discovery of endogenous Argonaute binding sites in Caenorhabditis elegans. Nature Structural & Molecular Biology, 2010. **17**(2): p. 173-9.

334. Thomas, M., J. Lieberman, and A. Lal, Desperately seeking microRNA targets. Nature Structural & Molecular Biology, 2010. **17**(10): p. 1169-74.

335. Hafner, M., et al., Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell, 2010. **141**(1): p. 129-41.

336. Stark, A., et al., Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. Cell, 2005. **123**(6): p. 1133-46.

337. Farh, K.K., et al., The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. Science, 2005. **310**(5755): p. 1817-21.

338. Fabian, M.R., N. Sonenberg, and W. Filipowicz, Regulation of mRNA translation and stability by microRNAs. Annual Review of Biochemistry, 2010. **79**: p. 351-79.

339. Vella, M.C., et al., The C. elegans microRNA let-7 binds to imperfect let-7 complementary sites from the lin-41 3 ' UTR. Genes & Development, 2004. **18**(2): p. 132-137.

340. Chi, S.W., G.J. Hannon, and R.B. Darnell, An alternative mode of microRNA target recognition. Nature Structural & Molecular Biology, 2012. **19**(3): p. 321-327.

341. Wang, Y., et al., Structure of the guide-strand-containing argonaute silencing complex. Nature, 2008. **456**(7219): p. 209-13.

342. Wang, Y., et al., Structure of an argonaute silencing complex with a seed-containing guide DNA and target RNA duplex. Nature, 2008. **456**(7224): p. 921-6.

343. Das, P., J.A. King, and R. Zhou, Aggregation of gamma-crystallins associated with human cataracts via domain swapping at the C-terminal beta-strands. Proc Natl Acad Sci U S A, 2011. **108**(26): p. 10514-9.

344. Liu, P., et al., Observation of a dewetting transition in the collapse of the melittin tetramer. Nature, 2005. **437**(7055): p. 159-62.

345. Zhou, R., et al., Destruction of long-range interactions by a single mutation in lysozyme. Proc Natl Acad Sci U S A, 2007. **104**(14): p. 5824-9.

346. Zhou, R., et al., Hydrophobic collapse in multidomain protein folding. Science, 2004. **305**(5690): p. 1605-9.

347. Wang, Y., et al., Nucleation, propagation and cleavage of target RNAs in Ago silencing complexes. Nature, 2009. **461**(7265): p. 754-61.

348. Kumar, S., et al., Scalable Molecular Dynamics with NAMD on Blue Gene/L IBM J. Res. Dev., 2007. **52**(No. 1/2).

349. Brooks, B.R., et al., CHARMM: the biomolecular simulation program. J. Comput. Chem., 2009. **30**(10): p. 1545-614.

350. Jorgensen, W.L., et al., Comparison of simple potential functions for simulating liquid water. J Chem Phys, 1983. **79**: p. 926-935.

351. Darden, T.A., D.M. York, and L.G. Pedersen, Particle mesh Ewald: An NlogN method for Ewald sums in large systems. J Chem Phys, 1993. **98**: p. 10089-10092.

352. Ryckaert, J.P., G. Ciccotti, and H.J.C. Berendsen, Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J. of Comp. Phys., 1977. **23**(3): p. 327-341.

353. Denning, E.J., et al., Impact of 2 '-Hydroxyl Sampling on the Conformational Properties of RNA: Update of the CHARMM All-Atom Additive Force Field for RNA. Journal of Computational Chemistry, 2011. **32**(9): p. 1929-1943.

354. Brennecke, J., et al., Principles of microRNA-target recognition. PLoS Biol, 2005. **3**(3): p. e85.

355. Cox, N.J. and K. Subbarao, Global epidemiology of influenza: past and present. Annu. Rev. Med., 2000. **51**: p. 407-21.

356. Hilleman, M.R., Realities and enigmas of human viral influenza: pathogenesis, epidemiology and control. Vaccine, 2002. **20**(25-26): p. 3068-87.

357. Stevens, J., et al., Glycan microarray analysis of the hemagglutinins from modern and pandemic influenza viruses reveals different receptor specificities. Journal of Molecular Biology, 2006. **355**(5): p. 1143-1155.

358. Fleury, D., et al., Antigen distortion allows influenza virus to escape neutralization. Nature Structural Biology, 1998. **5**(2): p. 119-23.

359. Yang, Z.Y., et al., Immunization by avian H5 influenza hemagglutinin mutants with altered receptor binding specificity. Science, 2007. **317**(5839): p. 825-828.

360. Tumpey, T.M., et al., A two-amino acid change in the hemagglutinin of the 1918 influenza virus abolishes transmission. Science, 2007. **315**(5812): p. 655-659.

361. Stevens, J., et al., Structure of the uncleaved human H1 hemagglutinin from the extinct 1918 influenza virus. Science, 2004. **303**(5665): p. 1866-70.

362. Stevens, J., et al., Structure and receptor specificity of the hemagglutinin from an H5N1 influenza virus. Science, 2006. **312**(5772): p. 404-10.

363. Gamblin, S.J., et al., The structure and receptor binding properties of the 1918 influenza hemagglutinin. Science, 2004. **303**(5665): p. 1838-42.

364. Lin, T., et al., The hemagglutinin structure of an avian H1N1 influenza A virus. Virology, 2009. **392**(1): p. 73-81.

365. Igarashi, M., et al., Predicting the antigenic structure of the pandemic (H1N1) 2009 influenza virus hemagglutinin. PLoS One, 2010. **5**(1): p. e8553.

366. Xu, R., et al., Structural basis of preexisting immunity to the 2009 H1N1 pandemic influenza virus. Science, 2010. **328**(5976): p. 357-60.

367. Sui, J., et al., Structural and functional bases for broad-spectrum neutralization of avian and human influenza A viruses. Nature Structural & Molecular Biology, 2009. **16**(3): p. 265-73.

368. Ekiert, D.C., et al., Antibody recognition of a highly conserved influenza virus epitope. Science, 2009. **324**(5924): p. 246-51.

369. Throsby, M., et al., Heterosubtypic neutralizing monoclonal antibodies cross-protective against H5N1 and H1N1 recovered from human IgM+ memory B cells. PLoS One, 2008. **3**(12): p. e3942.

370. Russell, R.J., et al., Structure of influenza hemagglutinin in complex with an inhibitor of membrane fusion. Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(46): p. 17736-41.

371. Das, P., et al., Free energy simulations reveal a double mutant avian H5N1 virus hemagglutinin with altered receptor binding specificity. J. Comput. Chem., 2009. **30**: p. 1654-1663.

372. Zhou, R., P. Das, and A.K. Royyuru, Single Mutation Induced H3N2 Hemagglutinin Antibody Neutralization: A Free Energy Perturbation Study. J. Phys. Chem. B, 2008. **112**: p. 15813–15820.

373. Chen, R. and E.C. Holmes, Avian influenza virus exhibits rapid evolutionary dynamics. Molecular Biology and Evolution, 2006. **23**(12): p. 2336-41.

374. Xia, Z., et al., Using a mutual information-based site transition network to map the genetic evolution of influenza A/H3N2 virus. Bioinformatics, 2009. **25**(18): p. 2309-17.

375. Smith, D.J., et al., Mapping the antigenic and genetic evolution of influenza virus. Science, 2004. **305**(5682): p. 371-6.

376. Du, X., et al., Networks of genomic co-occurrence capture characteristics of human influenza A (H3N2) evolution. Genome Research, 2008. **18**(1): p. 178-87.

377. Yamada, S., et al., Haemagglutinin mutations responsible for the binding of H5N1 influenza A viruses to human-type receptors. Nature, 2006. **444**(7117): p. 378-82.

378. Apisarnthanarak, A., et al., Seroprevalence of anti-H5 antibody among Thai health care workers after exposure to avian influenza (H5N1) in a tertiary care center. Clinical Infectious Diseases, 2005. **40**(2): p. e16-8.

379. Mehta, T., et al., Detection of oseltamivir resistance during treatment of 2009 H1N1 influenza virus infection in immunocompromised patients: utility of cycle threshold values of qualitative real-time reverse transcriptase PCR. J. Clin. Microbiol., 2010. **48**(11): p. 4326-8.

380. Stephenson, I., et al., Neuraminidase inhibitor resistance after oseltamivir treatment of acute influenza A and B in children. Clin. Infect Dis., 2009. **48**(4): p. 389-96.

381. Beigel, J. and M. Bray, Current and future antiviral therapy of severe seasonal and avian influenza. Antiviral Res., 2008. **78**(1): p. 91-102.

382. Lowen, A.C. and P. Palese, Influenza virus transmission: basic science and implications for the use of antiviral drugs during a pandemic. Infect Disord Drug Targets, 2007. **7**(4): p. 318-28.

383. Fick, J., et al., Antiviral oseltamivir is not removed or degraded in normal sewage water treatment: implications for development of resistance by influenza A virus. PLoS One, 2007. **2**(10): p. e986.

384. Bright, R.A., et al., Adamantane resistance among influenza A viruses isolated early during the 2005-2006 influenza season in the United States. Jama, 2006. **295**(8): p. 891-4.

385. de Jong, M.D., et al., Oseltamivir resistance during treatment of influenza A (H5N1) infection. New England Journal of Medicine, 2005. **353**(25): p. 2667-72.

386. Kiso, M., et al., Resistant influenza A viruses in children treated with oseltamivir: descriptive study. Lancet, 2004. **364**(9436): p. 759-65.

387. Jorgensen, W.L., Free-energy calculations - a breakthrough for modeling organic-chemistry in solution. Acc. Chem. Res., 1989. **22**(5): p. 184-189.

388. Deng, Y. and B. Roux, Calculation of standard binding free energies: aromatic molecules in the T4 lysozyme L99A mutant. J. Chem. Theo. Comp. , 2006. **2**: p. 1255-1273.

389. Kollman, P., Free-energy calculations - applications to chemical and biochemical phenomena. Chemical Reviews, 1993. **93**(7): p. 2395-2417.

390. Pathiaseril, A. and R.J. Woods, Relative energies of binding for antibody-carbohydrate-antigen complexes computed from free-energy simulations. Journal of the American Chemical Society, 2000. **122**(2): p. 331-338.

391. Simonson, T., G. Archontis, and M. Karplus, Free energy simulations come of age: Protein-ligand recognition. Acc. Chem. Res., 2002. **35**(6): p. 430-437.

392. Tembe, B.L. and J.A. McCammon, Ligand receptor interactions. Computers & Chemistry, 1984. **8**(4): p. 281-283.

393. Warshel, A., Simulating the Energetics and Dynamics of Enzymatic Reactions in Specificity in Biological Interactions, Pontificiae Academiae Scientiarum Scripta Varia, 1984. **55**: p. 60-81.

394. Warshel, A., et al., Modeling Electrostatic Effects in Proteins. Biochimica et Biophysica Acta, 2006. **1764**(11): p. 1647-1676.

395. Xia, Z., et al., Recognition mechanism of siRNA by viral p19 suppressor of RNA silencing: a molecular dynamics study. Biophysical Journal, 2009. **96**(5): p. 1761-9.

396. Zheng, L., M. Chen, and W. Yang, 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. 20227-20232.

397. Jiao, D., et al., Calculation of protein-ligand binding free energy by using a polarizable potential. Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(17): p. 6290-5.

398. Chodera, J.D., et al., Alchemical free energy methods for drug discovery: progress and challenges. Current Opinion in Structural Biology, 2011. **21**(2): p. 150-60.

399. Almlof, M., et al., Probing the Effect of Point Mutations at Protein-Protein Interfaces with Free Energy Calculations. Biophysical Journal, 2006. **90**(2): p. 433-442.

400. Brandsdal, B.O. and A.O. Smalas, Evaluation of protein-protein association energies by free energy perturbation calculations. Protein Engineering, 2000. **13**(4): p. 239-45.

401. Wang, J., Y. Deng, and B. Roux, Absolute binding free energy calculations using molecular dynamics simulations with restraining potentials. Biophysical Journal, 2006. **91**(8): p. 2798-814.

402. Darden, T.A., D.M. York, and L.G. Pedersen, Particle mesh Ewald: An NlogN method for Ewald sums in large systems. J. Chem. Phys., 1993. **98**: p. 10089-10092.

403. Eleftheriou, M., et al., Thermal denaturing of mutant lysozyme with both the OPLSAA and the CHARMM force fields. Journal of the American Chemical Society, 2006. **128**(41): p. 13388-95.

404. Zhou, R., et al., Destruction of long-range interactions by a single mutation in lysozyme. Proc. Natl. Acad. Sci. U.S.A., 2007. **104**(14): p. 5824-9.

405. Gao, Y.Q., W. Yang, and M. Karplus, A structure-based model for the synthesis and hydrolysis of ATP by F_1-ATPase. Cell, 2005. **123**(2): p. 195-205.

406. Hummer, G., J.C. Rasaiah, and J.P. Noworyta, Water conduction through the hydrophobic channel of a carbon nanotube. Nature, 2001. **414**: p. 188-190.

407. Kamberaj, H. and A. van der Vaart, An optimized replica exchange molecular dynamics method. J. Chem. Phys., 2009. **130**(7): p. 074906.

408. Karplus, M., et al., Protein structural transitions and their functional role. Philos. Transact. A Math. Phys. Eng. Sci., 2005. **363**: p. 331-355; discussion 355-356.

409. Hua, L., et al., Nanoscale dewetting transition in protein complex folding. J Phys Chem B, 2007. **111**(30): p. 9069-77.

410. Li, X., et al., Hydration and dewetting near fluorinated superhydrophobic plates. J Am Chem Soc, 2006. **128**(38): p. 12439-47.

411. Kumar, S., et al., Scalable Molecular Dynamics with NAMD on Blue Gene/L. IBM Journal of Research and Development: Applications of Massively Parallel Systems, 2008. **52**: p. 177-188.

412. MacKerell, A.D., et al., All-atom empirical potential for molecular modeling and dynamics studies of proteins. J. Phys. Chem. B, 1998. **102**: p. 3586-3616.

413. Jorgensen, W.L., et al., Comparison of simple potential functions for simulating liquid water. J. Chem. Phys., 1983. **79**: p. 926-935.

414. T. C. Beutler, et al., Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. Chem. Phys. Lett., 1994. **222**(222).

415. Zacharias, M., T.P. Straatsma, and J.A. McCammon, Separation-Shifted Scaling, a New Scaling Method for Lennard-Jones Interactions in Thermodynamic Integration. J. Chem. Phys., 1994. **100**(12): p. 9025-9031.

416. Boresch, S. and M. Karplus, The meaning of component analysis - decomposition of the free-energy in terms of specific interactions. Journal of Molecular Biology, 1995. **254**(5): p. 801-807.

417. Brady, G.P. and K.A. Sharp, Decomposition of interaction free-energies in proteins and other complex-systems. Journal of Molecular Biology, 1995. **254**(1): p. 77-85.

418. Bren, M., et al., Do all pieces make a whole? Thiele cumulants and the free energy decomposition. Theor. Chem. Acc., 2007. **117**(4): p. 535-540.

419. Mark, A.E. and W.F. Vangunsteren, Decomposition of the free-energy of a system in terms of specific interactions - implications for theoretical and experimental studies. Journal of Molecular Biology, 1994. **240**(2): p. 167-176.

420. Golubkov, P.A. and P.Y. Ren, Generalized coarse-grained model based on point multipole and Gay-Berne potentials. J. Chem. Phys., 2006. **125**(6): p. -.

421. Tanford, C., Protein denaturation. C. Theoretical models for the mechanism of denaturation. Adv Protein Chem, 1970. **24**: p. 1-95.

422. Pace, C.N., Determination and analysis of urea and guanidine hydrochloride denaturation curves. Methods Enzymol, 1986. **131**: p. 266-80.

423. Alonso, D.O. and K.A. Dill, Solvent denaturation and stabilization of globular proteins. Biochemistry, 1991. **30**(24): p. 5974-85.

424. Auton, M., L.M. Holthauzen, and D.W. Bolen, Anatomy of energetic changes accompanying urea-induced protein denaturation. Proc Natl Acad Sci U S A, 2007. **104**(39): p. 15317-22.

425. Courtenay, E.S., M.W. Capp, and M.T. Record, Jr., Thermodynamics of interactions of urea and guanidinium salts with protein surface: relationship between solute effects on protein processes and changes in water-accessible surface area. Protein Sci, 2001. **10**(12): p. 2485-97.

426. Schellman, J.A., The stability of hydrogen-bonded peptide structures in aqueous solution. Compt. rend. trav. Lab. Carlsberg, 1955. **29**(14-15): p. 230-59.

427. Scholtz, J.M., et al., Urea unfolding of peptide helices as a model for interpreting protein unfolding. Proc Natl Acad Sci U S A, 1995. **92**(1): p. 185-9.

428. Kuharski, R.A. and P.J. Rossky, Molecular-Dynamics Study of Solvation in Urea Water Solution. Journal of the American Chemical Society, 1984. **106**(20): p. 5786-5793.

429. Kuharski, R.A. and P.J. Rossky, Solvation of Hydrophobic Species in Aqueous Urea Solution - a Molecular-Dynamics Study. Journal of the American Chemical Society, 1984. **106**(20): p. 5794-5800.

430. TiradoRives, J., M. Orozco, and W.L. Jorgensen, Molecular dynamics simulations of the unfolding of barnase in water and 8 M aqueous urea. Biochemistry, 1997. **36**(24): p. 7313-7329.

431. Caflisch, A. and M. Karplus, Structural details of urea binding to barnase: a molecular dynamics analysis. Structure with Folding & Design, 1999. **7**(5): p. 477-488.

432. Tobi, D., R. Elber, and D. Thirumalai, The dominant interaction between peptide and urea is electrostatic in nature: A molecular dynamics simulation study. Biopolymers, 2003. **68**(3): p. 359-369.

433. Kokubo, H. and B.M. Pettitt, Preferential solvation in urea solutions at different concentrations: Properties from simulation studies. Journal of Physical Chemistry B, 2007. **111**(19): p. 5233-5242.

434. Tran, H.T., A. Mao, and R.V. Pappu, Role of backbone - Solvent interactions in determining conformational equilibria of intrinsically disordered proteins. Journal of the American Chemical Society, 2008. **130**(23): p. 7380-7392.

435. Wetlaufer D, et al., Nonpolar group participation in the denaturation of proteins by urea and guanidinium salts. Model compound studies. J Am Chem Soc, 1964. **86**(86): p. 508-514.

436. Frank H and F. F, Structural approach to the solvent power of water for hydrocarbons; urea as a structure breaker. J Chem Phys, 1968. **48**(48): p. 4746–4757.

437. Barone, G., E. Rizzo, and Vitaglia.V, Opposite Effect of Urea and Some of Its Derivatives on Water Structure. Journal of Physical Chemistry, 1970. **74**(10): p. 2230-&.

438. Finer, E.G., F. Franks, and M.J. Tait, Nuclear Magnetic-Resonance Studies of Aqueous Urea Solutions. Journal of the American Chemical Society, 1972. **94**(13): p. 4424-&.

439. Hammes, G.G. and P.R. Schimmel, An Investigation of Water-Urea and Water-Urea-Polyethylene Glycol Interactions. Journal of the American Chemical Society, 1967. **89**(2): p. 442-&.

440. Bennion, B.J. and V. Daggett, The molecular basis for the chemical denaturation of proteins by urea. Proc Natl Acad Sci U S A, 2003. **100**(9): p. 5142-5147.

441. Robinson, D.R. and W.P. Jencks, The Effect of Compounds of the Urea-Guanidinium Class on the Activity Coefficient of Acetyltetraglycine Ethyl Ester and Related Compounds. J Am Chem Soc, 1965. **87**: p. 2462-70.

442. Wallqvist, A., D.G. Covell, and D. Thirumalai, Hydrophobic interactions in aqueous urea solutions with implications for the mechanism of protein denaturation. Journal of the American Chemical Society, 1998. **120**(2): p. 427-428.

443. Klimov, D.K., J.E. Straub, and D. Thirumalai, Aqueous urea solution destabilizes A beta(16-22) oligomers. Proc Natl Acad Sci U S A, 2004. **101**(41): p. 14760-14765.

444. O'Brien, E.P., et al., Interactions between hydrophobic and ionic solutes in aqueous guanidinium chloride and urea solutions: Lessons for protein denaturation mechanism. Journal of the American Chemical Society, 2007. **129**(23): p. 7346-7353.

445. Stumpe, M.C. and H. Grubmuller, Interaction of urea with amino acids: implications for urea-induced protein denaturation. J Am Chem Soc, 2007. **129**(51): p. 16126-31.

446. Hua, L., et al., Urea denaturation by stronger dispersion interactions with proteins than water implies a 2-stage unfolding. Proc Natl Acad Sci U S A, 2008. **105**(44): p. 16928-33.

447. Lim, W.K., J. Rosgen, and S.W. Englander, Urea, but not guanidinium, destabilizes proteins by forming hydrogen bonds to the peptide group. Proc Natl Acad Sci U S A, 2009. **106**(8): p. 2595-600.

448. Godawat, R., S.N. Jamadagni, and S. Garde, Unfolding of hydrophobic polymers in guanidinium chloride solutions. J Phys Chem B, 2010. **114**(6): p. 2246-54.

449. Camilloni, C., et al., Urea and guanidinium chloride denature protein L in different ways in molecular dynamics simulations. Biophys J, 2008. **94**(12): p. 4654-61.

450. Mason, P.E., et al., The interaction of guanidinium ions with a model peptide. Biophys J, 2007. **93**(1): p. L4-L6.

451. Grosberg, A.Y., I.Y. Erukhimovitch, and E.I. Shakhnovitch, On the Theory of Psi-Condensation. Biopolymers, 1982. **21**(12): p. 2413-2432.

452. Finkelstein, A.V. and E.I. Shakhnovich, Theory of Cooperative Transitions in Protein Molecules .2. Phase-Diagram for a Protein Molecule in Solution. Biopolymers, 1989. **28**(10): p. 1681-1694.

453. Grosberg, A.Y., I.Y. Erukhimovich, and E.I. Skahnovich, On DNA Compactization in Diluted Polymeric Solutions. Biofizika, 1981. **26**(3): p. 415-420.

454. Brochard, F. and P.G. Degennes, Collapse of One Polymer Coil in a Mixture of Solvents. Ferroelectrics, 1980. **30**(1-4): p. 33-47.

455. Vaney, M.C., et al., High-resolution structure (1.33 A) of a HEW lysozyme tetragonal crystal grown in the APCF apparatus. Data and structural comparison with a crystal grown under microgravity from SpaceHab-01 mission. Acta Crystallogr D Biol Crystallogr, 1996. **52**(Pt 3): p. 505-17.

456. Wikstrom, M., et al., Three-dimensional solution structure of an immunoglobulin light chain-binding domain of protein L. Comparison with the IgG-binding domains of protein G. Biochemistry, 1994. **33**(47): p. 14011-7.

457. Voelz, V.A., et al., Unfolded-state dynamics and structure of protein L characterized by simulation and experiment. J Am Chem Soc, 2010. **132**(13): p. 4702-9.

458. Kumar, S., et al., Scalable Molecular Dynamics with NAMD on Blue Gene/L. IBM J. Res.Dev., 2008. **52**: p. 177-188.

459. Brooks, C.L., et al., Chemical physics of protein folding. Proc Natl Acad Sci U S A, 1998. **95**(19): p. 11037-11038.

460. Brooks, C.L., J.N. Onuchic, and D.J. Wales, Statistical thermodynamics - Taking a walk on a landscape. Science, 2001. **293**(5530): p. 612-613.

461. Fersht, A.R. and V. Daggett, Protein folding and unfolding at atomic resolution. Cell, 2002. **108**(4): p. 573-582.

462. Snow, C.D., et al., Absolute comparison of simulated and experimental protein-folding dynamics. Nature, 2002. **420**(6911): p. 102-106.

463. Garcia, A.E. and J.N. Onuchic, Folding a protein in a computer: An atomic description of the folding/unfolding of protein A. Proc Natl Acad Sci U S A, 2003. **100**(24): p. 13898-13903.

464. Kokubo, H. and Y. Okamoto, Prediction of transmembrane helix configurations by replica-exchange simulations. Chemical Physics Letters, 2004. **383**(3-4): p. 397-402.

465. Zhou, R.H., et al., Hydrophobic collapse in multidomain protein folding. Science, 2004. **305**(5690): p. 1605-1609.

466. Liu, P., et al., Observation of a dewetting transition in the collapse of the melittin tetramer. Nature, 2005. **437**(7055): p. 159-162.

467. Zhou, R. and B.J. Berne, Can a continuum solvent model reproduce the free energy landscape of a beta -hairpin folding in water? Proc Natl Acad Sci U S A, 2002. **99**(20): p. 12777-82.

468. Zhou, R., B.J. Berne, and R. Germain, The free energy landscape for beta hairpin folding in explicit water. Proc Natl Acad Sci U S A, 2001. **98**(26): p. 14931-6.

469. Zhou, R., Trp-cage: folding free energy landscape in explicit water. Proc Natl Acad Sci U S A, 2003. **100**(23): p. 13280-5.

470. Li, J., et al., Electrostatic gating of a nanometer water channel. Proc Natl Acad Sci U S A, 2007. **104**(10): p. 3687-92.

471. Tu, Y., et al., Water-mediated signal multiplication with Y-shaped carbon nanotubes. Proc Natl Acad Sci U S A, 2009. **106**(43): p. 18120-4.

472. Ge, C., et al., Binding of blood proteins to carbon nanotubes reduces cytotoxicity. Proc Natl Acad Sci U S A, 2011. **108**(41): p. 16968-73.

473. Brooks, B.R., et al., CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J. Comput. Chem., 1983. **4**(2): p. 187-217.

474. Deserno, M. and C. Holm, How to mesh up Ewald sums. I. A theoretical and numerical comparison of various particle mesh routines. Journal of Chemical Physics, 1998. **109**(18): p. 7678-7693.

475. Zhou, R.H., et al., Comment on "Urea-Mediated Protein Denaturation: A Consensus View". Journal of Physical Chemistry B, 2011. **115**(5): p. 1323-1326.

476. Xiu, P., et al., Urea-Induced Drying of Hydrophobic Nanotubes: Comparison of Different Urea Models. Journal of Physical Chemistry B, 2011. **115**(12): p. 2988-2994.

477. Das, A. and C. Mukhopadhyay, Reply to the "Comment on 'Urea-Mediated Protein Denaturation: A Consensus View'". Journal of Physical Chemistry B, 2011. **115**(5): p. 1327-1328.

478. Gao, M., Z.S. She, and R.H. Zhou, Key Residues that Play a Critical Role in Urea-Induced Lysozyme Unfolding. Journal of Physical Chemistry B, 2010. **114**(47): p. 15687-15693.

479. Das, P. and R.H. Zhou, Urea-Induced Drying of Carbon Nanotubes Suggests Existence of a Dry Globule-like Transient State During Chemical Denaturation of Proteins. Journal of Physical Chemistry B, 2010. **114**(16): p. 5427-5430.

480. Zangi, R., R.H. Zhou, and B.J. Berne, Urea's Action on Hydrophobic Interactions. Journal of the American Chemical Society, 2009. **131**(4): p. 1535-1541.