Developing protein force fields for implicit solvent simulation

Zhang, Haiping

2018

Zhang, H. (2018). Developing protein force fields for implicit solvent simulation. Doctoral thesis, Nanyang Technological University, Singapore.

<http://hdl.handle.net/10356/74803>

<https://doi.org/10.32657/10356/74803>

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

SINGAPORE

**DEVELOPING PROTEIN FORCE FIELDS FOR IMPLICIT SOLVENT
SIMULATION**

ZHANG HAIPING

SCHOOL OF BIOLOGICAL SCIENCES

2018

**DEVELOPING PROTEIN FORCE FIELDS FOR IMPLICIT SOLVENT
SIMULATION**

ZHANG HAIPING

SCHOOL OF BIOLOGICAL SCIENCES

A thesis submitted to the Nanyang Technological University
in fulfillment of the requirement for the degree of
Doctor of Philosophy

2018

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor, A/Prof. Lu Lanyuan, for his invaluable guidance, support and encouragement during every stage of this study.

I am also grateful to A/Prof. Mu Yuguang and Liang Zhaoxun of the School of Biological Sciences (SBS), Nanyang Technological University (NTU) for the constructive comments and insightful suggestions they made as my thesis advisory committee members.

I would also like to express my sincere gratitude to Ass/Prof. Chen Gang and Dr. Zhong ZhenSheng of School of Physical and Mathematical Sciences (SPMS, NTU) for sacrificing their time and resources to provide the data for a collaboration project. I would like to thank Mr Lim Eric Jit-Kai, who has contributed a lot in the dengue's E protein simulation project.

I am also thankful to my parents for their concern and endless encouragement. Every time I feel tired, I will get energized with the thought of their love and support.

The work presented in this report has been carried out at the School of Biological Sciences, Nanyang Technological University (NTU). The financial support from NTU is gratefully acknowledged.

Special thanks also go to the current and past laboratory mates of mine for sharing with me their experiences. I am grateful to Dr. Tong DuDu, Dr. Zhu Guanhua, Dr. Bope Domilongo Christian, Dr. Chen Wenwen, Dr. Nalaparaju Anjaiah. Special thanks to Mr. Ng Tze Yang Justin, Mr. ZHENG LIANGZHEN for the constructive comments and suggestions they provided.

Table of Contents

Acknowledgements	3
Table of Contents	5
List of Tables	10
List of Figures.....	12
Abbreviations	19
Abstract.....	21
1 General Introduction	25
1.1 Molecular dynamics simulation and force fields	25
1.2 Explicit solvent water models	28
1.3 Implicit solvent models	29
1.4 Explicit-implicit solvent comparison	30
1.5 Improving force field in literature	32
1.6 Atomic Contact Energy (ACE) based force field.....	35
1.7 Insufficient and Ineffective Sampling in MD	37
1.8 Metadynamics and well-tempered metadynamics simulation	38

1.9	Replica-exchange molecular dynamics (REMD).....	40
1.10	Conformational change of Dengue E protein.....	41
1.11	Aims and outlines.....	44
2	Improvement of GB simulation performances by adding a cMAP term and strengthening aromatic related interactions	48
2.1	Background	48
2.2	Method details	51
2.2.1	Test Data sets of peptides and small proteins	51
2.2.2	Force field used for the simulation	55
2.2.3	cMAP term and file preparation	56
2.2.4	<i>Ab initio</i> folding with REMD Simulations	59
2.2.5	The data analysis.....	64
2.3	Comparison of results between modified and unmodified force fields ...	66
2.3.1	<i>Ab initio</i> folding of helical peptides.....	71
2.3.2	<i>Ab initio</i> folding of beta hairpins	75
2.3.3	<i>Ab initio</i> folding of other peptides that contain more than one secondary conformation.....	77
2.3.4	<i>Ab initio</i> folding of small proteins with tertiary structures	80

2.3.5	Free energy landscapes (FEL) of test peptides	84
2.4	Conclusion.....	90
3	Solvent-free all-atom force field incorporating experimental native structure information	92
3.1	Background	92
3.2	Method details	98
3.2.1	Data set collection.....	98
3.2.2	Functional forms and parameters.....	102
3.2.3	Parameters of Distance Dependent Dielectric constant and side chain VDW interaction scaling.....	106
3.2.4	Making grid-based dihedral energy correction map (cMAP) for backbone dihedral correction.....	109
3.2.5	Iterative simulations to generate converged cMAPs	110
3.2.6	Further modification of the cMAP by adding a patch	115
3.2.7	Testing the force field by simulating a test set with known experimental secondary structures	116
3.2.8	Convergences test for GB simulation, chapter 2's modified GB simulation and MJ simulation.....	120

3.3	Comparing the simulation results of peptides and small proteins between MJ and GB force field.....	122
3.3.2	<i>Ab initio</i> folding of small proteins with tertiary structure in test set	135
3.3.3	<i>Ab initio</i> folding of other 8 test cases with various sizes and secondary structure components	139
3.3.4	The free energy landscapes (FEL) of test small proteins	143
3.3.5	Future improvement from the clues of failed cases	148
3.4	Conclusions	150
4	Investigating the stability of DENV-2 envelope protein dimer using well-tempered metadynamics simulations	154
4.1	Background	154
4.2	Method details	157
4.2.1	PDB structures collection and topology preparation	157
4.2.2	Molecular simulation procedure	158
4.2.3	Analyzing the convergence, free energy landscape and atomistic details of the interaction.....	160
4.3	FEL for 4 different states of dengue's E protein	161
4.4	FEL for 2 mutations of E protein with antibody	165

4.5	Conclusion.....	168
5	Concluding remarks	170
6	REFERENCES.....	172
7	PUBLICATION LIST	185

List of Tables

Table 2.1 Summary of the performed simulations with the respective number of residues (Nres) and the net charge (Nchg) in water at neutral pH for each model peptide.	52
Table 2.2 Basic information of the small proteins which have basic tertiary structure in the test set.	54
Table 2.3 Temperature values used in the REMD, calculated using the online REMD temperature generator webserver [86].	60
Table 2.4 <i>ab initio</i> folding of peptide test set with the modified force field in comparison with the <i>ab initio</i> folding with the original force field. The RMSD is calculated from the most populated cluster relative to the native conformation.	66
Table 2.5 The <i>ab initio</i> folding of the small protein test set with the modified force field in comparison to the <i>ab initio</i> folding by the original force field. The RMSD is calculated for the predicted structure relative to the native structure.	83
Table 3.1 Summary of the training set peptides with the respective number of residues (Nres) and the net charge (Nchg) in water at neutral pH for each model peptide.	99
Table 3.2 The test set of the 8 fast folding small proteins and the 8 peptides.	101

Table 3.3 The <i>ab initio</i> folding of peptide Set A with the MJ force field in comparison to the <i>ab initio</i> folding by AMBER14_GB force field. The RMSD values are calculated for the most populated cluster related to the native conformation.	127
Table 3.4 The <i>ab initio</i> folding of the 8 small proteins in the test set with the MJ force field in comparison to the <i>ab initio</i> folding by the AMBER14 GB force field. The RMSD values are between predicted structures related to the native ones.....	138
Table 3.5 The <i>ab initio</i> folding of 8 peptides in a test set with the MJ force field in comparison to the <i>ab initio</i> folding by the AMBER14 GB force field. The RMSD values are calculated for the most populated clusters related to the native conformations.	141

List of Figures

Figure 1.1 The time scale needed for the multiscale biomolecule system. It indicates current computational power is challenge for large system such as dengue particle. The dengue viral particle is adopted from PDB 1TG8[12]... 27

Figure 1.2 shows the Ramachandran diagram. The Beta-sheet favor region, right handed alpha-helix, left-handed polyproline II helix (PPII) region and left handed alpha-head region is marked. The figure is generated according to the work in (http://www.cryst.bbk.ac.uk/PPS95/course/3_geometry/rama.html). 33

Figure 1.3 Shows the structure of dengue E protein in different states. Panel a illustrates the structure of monomer E protein; Panel b illustrate the dimer E protein, and the EDE epitope region was marked; Panel c shows the structure of trimeric E protein in acid condition; Panel d show the whole Dengue viral particle, and using the diagram to illustrate different parts of the Dengue viral particle. 43

Figure 2.1 The cMAP energy corrections applied to each residue is shown in the blue color region. The Figure 2.1a panel shows cMAP energy corrections applied in the residues VAL, TRP, THR, PHE, ILE and TYR with strength of -1 kcal/mol at the beta region. The Figure 2.1b panel shows the cMAP energy correction applied in the residues ALA, GLU, ASP, MET, GLN and LEU with energy strength of -1 kcal/mol at the alpha region. The Figure 2.1c panel

shows the cMAP energy correction applied in the residue LYS at the alpha region with energy strength of -0.2 kcal/mol. The Figure 2.1d shows for other residues no cMAP correction was applied. The prePRO is the amino acid residue immediately before PRO residue.	58
Figure 2.2 The superposition of the predicted conformation (Red) by the modified GB onto the native PDB structure (Blue).	69
Figure 2.3 The superposition of the predicted conformation with the original GB (Red) on the native PDB structure (Blue).	71
Figure 2.4 Panel a and b shows the most populated cluster conformations of 5 experimentally known helix peptides and percentages of the most popular clusters over all clusters. Panel c shows the helicity of each residue in those peptides (Red is from the modified force field simulation; Blue is from the original force field simulation).	72
Figure 2.5 The predicted structures of the 8 small protein cases with tertiary structures from the original GB simulations are superposed with the native PDB structures (predicted structures in Red; the native PDB conformations in Blue).	81
Figure 2.6 The free energy landscapes as the function of the all atom radius gyration (RG) and the C α RMSD in the peptides simulation with the modified force field.	85

Figure 2.7 The free energy surfaces as the function of the all atom radius gyration (RG) and the C α RMSD in the 19 small peptide simulations with the original force field.	86
Figure 2.8 a,b shows the free energy surfaces as a function of the all atom radius of gyration (RG) and the C α RMSD in the small protein simulations with the original and modified force fields, respectively.	89
Figure 3.1 The figure shows the non-bonded interaction form. The attractive part is an LJ functional form, and the repulsive part is a stepwise function. For convenience, the interaction measure from the MD simulation of two bead model is shifted up for 0.1 kJ/mol. The energy unit is kJ/mol.	104
Figure 3.2 The dielectric constant versus the distance between the different types of interaction partners. The unit for the x axis is ns.	108
Figure 3.3 The procedure to generate the final cMAP. The ultimate goal of the dipeptide iterative simulations is to obtain the cMAPs that can generate similar phi-psi distributions as in the coil database. The patch adding step is to further correct the cMAP, in order to fold to the native like conformation for a variety of peptide cases.	115
Figure 3.4 The figure shows the final version of cMAP energy correction applied for the 20 residues respectively. The energy unit is kJ/mol, and the applied energy is gradually increased from blue to white.	118

Figure 3.5	The figure shows the final version of cMAP energy correction applied in the prePRO residue (residue immediately before the PRO) and in the sidechain chi1-chi2 of the GLU residue. The energy unit is kJ/mol, and the applied energy is gradually increased from blue to white.....	119
Figure 3.6	The free energy landscapes of each trajectory block of peptide 1fsd in simulation by GB simulation (panel a), chapter 2's modified GB simulation (panel b) and MJ simulation (panel c). The simulation trajectories from these three force field were selected, and divided into 3 blocks . The convergences were roughly checked by comparing the similarity of free energy landscape within blocks.....	121
Figure 3.7	The superposition of the predicted conformation (Red) by the MJ force field and the native PDB structure (Blue).....	123
Figure 3.8	The superposition of the predicted structure (Red) by the original GB and the native PDB structure (Blue).	124
Figure 3.9	Panel a,b show the predicted structures of 5 experimentally known helix peptides and percentages of the most popular clusters over all clusters. Panel c shows the helicity of each residue in those peptides (Red is from the MJ force field simulation, Blue is from the GB force field simulation).	127
Figure 3.10	The predicted structures by the GB simulation for the 8 small protein cases with tertiary structures are superposed with the native PDB	

structures (the most populated cluster in Red, and the native structure in Blue).	137
Figure 3.11 The figure shows the predicted structures by the GB force field for the 8 peptide test cases that are superposed with the native PDB structures (predicted structures in red, and the native structures in blue).	140
Figure 3.12 The free energy surfaces as a function of the all atom radius of gyration (RG) and the C alpha RMSD in the simulation cases with the MJ force field.	144
Figure 3.13 The free energy surfaces as the function of the all atom radius of gyration (RG) and the C alpha RMSD of the 19 small peptide cases from the REMD simulations with the GB force field.	145
Figure 3.14 Panel a,b shows the free energy surfaces as the function of the all atom radius gyration (RG) and the C alpha RMSD in the small protein cases simulated with the GB and the MJ force fields, respectively.....	147
Figure 3.15 Panel a,b shows the free energy surfaces as the function of the all atom radius gyration (RG) and the C alpha RMSD in the 8 peptides of the test set simulated with the GB and the MJ force field respectively.....	148
Figure 4.1 a,b,c,d shows the evolution of Gaussian height over time in the well-tempered metadynamics simulations of wild type dengue E protein dimer with or without antibody at neural or acid conditions respectively. The e, f shows the time evolution of Gaussian height in the well-tempered metadynamics	

simulation of two mutated dengue E protein dimers (W101A and N153N), respectively. In all the simulations, the Gaussian heights were decreased to near zero within the 100 ns simulation, which indicates convergence of the simulation..... 162

Figure 4.2 a,b The FES of the binding between the monomers of the dimer in neutral and acidic conditions respectively. It is computed by the metadynamics simulation through the projection on the two CVs, CA RMSD and interface C α coordination number. The snapshots are the representative conformations by clustering all the conformations that fall in the box region, which is marked with block edge box, with the ranges ($10 < x < 25$, $0.08 < y < 0.12$) and ($5 < x < 20$, $0.1 < y < 0.6$) for Figure 4.2a and Figure 4.2b, respectively. The Figure 4.2c,d show the free energy landscape of the dimer in neutral and acidic conditions with antibody binding. The snapshot is the representative conformation from the region marked by the black edge box, with the ranges ($15 < x < 35$, $0.08 < y < 0.12$) and ($10 < x < 30$, $0.08 < y < 0.12$) for Figure 4.2c and Figure 4.2d respectively. The free energy units are 1 kJ/mol, and the unit for RMSD is nm..... 164

Figure 4.3 a,b shows the FES of the binding between monomers of the dimer in acid condition with antibody for W101A and N153A mutations, respectively. It is computed by the metadynamics simulation through projection on the two CV, C α RMSD and interface C α a coordination number. The Figure 4.3c is

the representative conformation of the most populated cluster. We cluster all the conformations fall in the box region, which is marked by the black edge box, with the ranges ($5 < x < 20$, $0.1 < y < 0.2$) and ($0 < x < 5$, $0.7 < y < 1$) for Figure 4.3a and Figure 4.3b, respectively. The Figure 4.3d is the representative conformation from the region marked by the black edge box, with the ranges ($5 < x < 15$, $0.2 < y < 0.4$) and ($0 < x < 10$, $0.5 < y < 0.7$). The free energy units are 1 kJ/mol, the unit for RMSD is nm. 167

Abbreviations

RMSD	The root-mean-square deviation of atomic positions
REMD	Replica-Exchange Molecular Dynamics
cMAP	Grid-Based Backbone Correction Map term
AMBER	Assisted model building with energy refinement
GROMACS	Groningen Machine for Chemical Simulations
CHARMM	Chemistry at Harvard Macromolecular Mechanics
ACE	Atomic contact energy
GB	Generalized Born model
GBSA	Generalized Born model with solvent accessible surface area
SD	Leap-frog Stochastic Dynamics Integrator
PBC	Periodic boundary conditions
ff14SBonlysc	ff14SB only side chain force field
GB-Neck2	Updated GB-Neck implicit solvent model
LJ	Lennard-Jones potential
VDW	Van der Waals force
EDE	E-Dimer-dependent epitope
E protein	Dengue virus envelope proteins
PDB	Protein Data Bank
FEL	Free energy landscape
LJ-14	Intramolecular interactions occurring between 1-4 pairs

CV	Collective Variable
PPII region	Left-handed polyproline II helix region
prePRO	The amino acid residue immediately before PRO residue

Abstract

Molecular dynamics simulation is widely used in research of biomolecule properties and biomolecular processes, such as protein-protein interactions, protein *ab initio* folding and protein domain-domain interactions with a linker. Previous studies have shown that accuracy and efficiency of such simulations depends heavily on whether the force field is incorporated with explicit solvent. Although computational resources have increased tremendously in recent years, for large systems with explicit water, it is still prohibitive to achieve a simulation time scale between microseconds and milliseconds for many labs, which is the minimal time for most biochemical processes to complete. In contrast, implicit solvent models can significantly increase computational efficiency by reducing the total degrees of freedom in simulations. However, current Generalized Born models (GB) or GB with solvent accessible surface area based (GBSA-based) force fields are usually less accurate than the explicit solvent counterparts and still require improvement. In my study, I developed two methods to improve solvent-free force fields.

One method that I developed is to improve the GB-Neck2 model combined with ff14SBonlysc force field, which is one of the most accurate implicit solvent models in literature. I implemented a cMAP potential energy term to adjust the secondary structure propensity for each type of residue. In addition, non-bonded parameters whose side chains contain aromatic rings were modified to better mimic pi-pi interactions. Results from simulations using my method show significantly

improved performance in predicting secondary structure propensity as well as tertiary structure accuracy. A test set of 19 small peptides with experimentally determined native PDB structures and with diversified secondary structures was constructed to test the accuracy of my force field. In 16 (84.2%) of these cases, the native structures were correctly produced from extended starting structures by replica-exchange molecular dynamics simulations, using a RMSD criterion of 0.45 nm. This result is significantly superior to that using the original force field with the success ratio 47.3 % (9 in 19 cases). Additionally, some small proteins were selected to investigate the force field's performance on larger systems, and better results from my modified parameters were also observed for those cases. Furthermore, the free energy surfaces from the protein simulations illustrate that my force field produces global free energy minima in the vicinities of native structures, while a broad range of conformations can still be sufficiently sampled in the simulations.

Hence, I developed a new GB-based atomistic force field with improved ability to produce secondary and tertiary structure. My force field can efficiently sample the conformational space of peptides and small to medium sized proteins.

The other method is to develop a new force field by incorporating a distance dependent dielectric constant, a pairwise statistical potential and a modified dihedral energy correction (cMAP) term together, to achieve MD simulations of proteins with high practical speed and acceptable accuracy. The pairwise statistical

potential has been widely used as score term in protein-protein docking, while in our study, it shows potential application in molecular dynamics simulation as well. This force field adopts bonded parameters directly from the GROMOS54a7 force field, while adding a newly tuned cMAP to bias the backbone phi-psi distribution to the phi-psi distribution from the Protein Coil Library. The cMAPs of amino acid residues were further manually modified in order to achieve better performance for the training set. The modified force field is able to fold peptides *ab initio* with reasonable alpha helix/beta sheet propensity, maintaining the protein's secondary and tertiary structure.

In Chapter 4, metadynamics simulation approaches were used to explore the conformational dynamics of the dengue virus envelope (E) protein. The E protein undergoes large-scale conformational changes during the viral life cycle, including acidic pH induced changes in the endosome that enable fusion and subsequent infection. Antibodies which block such changes have the potential to inhibit viral infection. The simulations demonstrate the potential for such enhanced sampling approaches to successfully predict differences in the protein complex free energy landscape in response to changes in pH, antibody binding, protein mutations etc., and thus show promise as a tool in future rational therapeutics development.

All the three projects in the thesis are related to enhanced sampling methods in molecular dynamics simulations. The implementation of implicit solvent models is a common strategy to achieve better conformational sampling in simulations, and

the first two projects were designed for the force field developments in solvent-free simulations. In the last project, the Metadynamics approach, which is one of the most popular enhanced sampling methods, was adopted to explore the protein conformational changes in the dengue virus.

1 General Introduction

1.1 *Molecular dynamics simulation and force fields*

Molecular dynamics (MD) simulations calculate the time dependent behavior of molecular systems, providing details of atomic information and processes of biomolecular conformational change[1]. Since after reaching equilibration, the time average of long simulation properties is equivalent to the ensemble average, the trajectories of MD simulations can be used to estimate many thermodynamic properties of the molecular system [2]. This method was first proposed by Fermi E., Pasta J., Ulam S. in the 1950s [3], and further adopted in the 1970s for larger biomolecular systems [4, 5]. The general term of force field can be written as following:

$$E_{mm} = E_{bonded} + E_{non-bonded} \quad (1.1)$$

Where the bonded term usually contains bonded, angle, and dihedral term, and the non-bonded term contains electrical interaction and VDW interaction.

MD simulation has gradually achieved the ability to simulate more complex systems with powerful computers through the decades. Only until recent years, it has advanced from very small peptides with a few hundreds of atoms in vacuo to functionally related biomolecules in a fully explicit solvated environment, including proteins or RNA/DNA complexes in water solvent, membrane proteins in lipid layers [6], whole viral capsids [7], or nucleosomes [8, 9]. In MD

simulations, movements of atoms in a molecular system are calculated as a function of time by solving the Newton's equation of motion [10]. A simulation is converged when the specific properties of the system does not change against the time any further. The atomic interaction as described by the force field can, in most cases, be classified into bonded terms and non-bonded terms [11]. Calculating the non-bonded terms, which include electrostatic terms with Coulombs' law and van der Waals terms with Lennard Jones potential [12], are the most time-consuming and computationally expensive part during an MD simulation.

The force field is the collection of the parameters and functional forms of all bonded and non-bonded terms. The accuracy of the force field directly determines whether the MD simulation can reproduce similar molecular properties as derived from experimental methods. Since most biomolecules and their functions occur in solvent, a common practice in simulation set up is to include explicit solvent in simulation systems, as the solvation free energy and ordering structural properties of water are simply too important to be ignored. However, applying an explicit solvent method on a large biomolecule significantly increase the degrees of freedom of the system during simulation, heavily use of computational resources, and hence making it expensive for large scale industrial applications.

Figure 1.1 The time scale needed for the multiscale biomolecule system. It indicates current computational power is challenge for large system such as dengue particle. The dengue viral particle is adopted from PDB 1TG8[12].

Implicit solvent methods were therefore developed to reduce the expensive requirement of computational resources. The implicit solvent methods reduce the degrees of freedom by reducing solvent particles and approximate the solvation effects through theoretical estimation or through statistical potentials which already take into account solvent effects [13-15].

1.2 *Explicit solvent water models*

Many explicit solvent water models have been developed to approximate properties of water molecules, ranging from two-site to six-site models [16, 17]. These are usually available in the widely used simulation packages [18-20]. For instance, in GROMACS package [20], SPC, SPC/E and TIP3P water models are available for three-point model, TIP4P and TIP5P are four-site and five-site models. Two-site models are less accurate while less computationally demanding and six-site models are the opposite. Nowadays, the three-site models are widely used, which is a tradeoff between complexity and efficiency [17]. Parameters of different force fields are optimized specifically with different solvent models, so a different force field may have better performance with a selected water model. For instance, the AMBER and CHARMM force field often combine with the TIP3P water model [21, 22], while GROMOS96 force field often combines with the simple point charge (SPC) model [23]. A desirable explicit solvent model should correctly approximate all physical properties of liquid water, *e.g.* dielectric constant, density and diffusion constant, the tetrahedral shape (HOH angle) as well as all water

related interactions such as hydrogen bonding [17], while current state-of-the-art explicit solvent models still hard to achieve many of those requirements. Explicit solvent water models are widely used and have achieved many successes in exploring bimolecular processes and revealing structure-to-function relationships, in simulations of free energy calculations, protein folding, drug–receptor interaction, and conformational transitions [24].

1.3 Implicit solvent models

Implicit solvent models are alternatives to explicit solvent models, developed to approximate the averaged behavior of water molecules [25]. By reducing the number of explicit water atom, the degrees of freedom in simulation systems have been reduced dramatically. Hence a much higher computational efficiency is expected. The Generalized Born/Surface Area (GBSA) Model is a Generalized Born model combined with the application of hydrophobic solvent accessible surface area (SA) [26-28]. It is one of the most widely used implicit solvent model based methods. The Generalized Born model is an approximation of the Poisson-Boltzmann equation which is relatively rigorous, but computationally expensive [29, 30]. The GBSA model approximates the average behavior of water as a continuous medium, and the SA part is calculated to check the extent of each atom's exposure to solvent [28]. It has the following functional form [28, 31, 32]:

$$G_{pol}^{GB} = -\frac{1}{8\pi\epsilon_0} \left(1 - \frac{1}{\epsilon}\right) \sum_{i,j}^N \frac{q_i q_j}{f_{GB}} \quad (1.1)$$

Where

$$f_{GB} = \sqrt{r_{ij}^2 + a_{ij}^2} e^{-D} \quad (1.2)$$

$$D = \left(\frac{r_{ij}}{2a_{ij}}\right)^2 \quad (1.4)$$

$$a_{ij} = \sqrt{a_i a_j} \quad (1.5)$$

a_i and a_j are the effective Born radii of atom i and j , respectively. ϵ_0 and ϵ are the vacuum permittivity [33] and dielectric constant of the implicit solvent, respectively. While there are various force fields and several Born radii sets of GB or GBSA in AMBER, it is necessary to check which combination of force field and GB or GBSA can yield better fits to the experimental data, *i.e.* the correct *ab initio* folding of proteins and peptides from generally a large and diversified test set [34].

1.4 Explicit-implicit solvent comparison

There are many widely used force fields for explicit solvent or implicit solvent simulation, including OPLS [35], GROMOS [36], AMBER [37], CHARMM [38]. The explicit solvent simulation should be more accurate than the implicit solvent Molecular Dynamics (MD) simulation in most cases [39]. However, the implicit solvent simulation is promising to achieve convergence relatively rapidly, and enhances the sampling efficiency by reducing degrees of freedom and solvent

viscosity. This is especially true in the Replica-Exchange Molecular Dynamics (REMD) simulation, as the replica number is significantly reduced in implicit solvent simulation comparing to that in explicit solvent simulation [40]. This is because replica exchange molecular dynamics simulations require energy overlap between replicas in order for an efficient exchange rate, and the solvent interaction energy accounts for a large portion of the system's energy.

The force fields with explicit solvent have improved continuously during the last decade. Comparing to the original CHARMM force field first introduced, a modified CHARMM force field (CHARMM22*) has improved results in producing correct *ab initio* folding for various proteins, and better transferability among different protein classes [41, 42]. Recently, the residue specific force field RSFF1 or RSFF2 has shown accurate structural prediction for various kinds of peptide and small proteins, some up to sizes reaching 80 residues (*e.g.* λ -repressor) [43-49]. However, to the best of our knowledge, till today, there is no publicized force field model that is able to achieve excellent transferability among different protein classes, when using MD simulation with GB, the implicit solvent model [50-52]. Although GB is widely implemented in MD simulation and available in popular simulation software, its accuracy and efficacy is still questionable [27, 51, 52]. Taking the AMBER force field as an example, even the best combinations of GB models with force fields were only able to successfully predict small peptide conformations in limited cases [51, 52]. Other GB model and force field

combinations give suboptimal results that are either heavily alpha helical biased or beta-sheet biased. Besides secondary structure, the tertiary structure of GB significantly deviates from the native structure in some test cases, making the simulation results questionable[53].

1.5 Improving force field in literature

Residue-specific bonded and non-bonded parameters have been implemented to refine force fields. For instance, the force fields such as RSFF1 and RSFF2, where the residue-specific parameters are iteratively tuned, have correctly preformed *ab initio* folding of peptides and small proteins with diversified secondary and tertiary structures [43, 44, 49]. The Grid-Based Backbone Correction Map term (cMAP), where the backbone dihedral angles of residues are considered, could also be applied to refine force fields [50, 54].

The backbone phi-psi dihedral angle distributions of residues are strongly correlated to the propensities of the secondary structures. For instance, the -100 to -30 degrees of phi and -67 to -7 degrees of psi range correlates to the alpha-helix region; -180 to -100 degrees of phi and 120 to 180 degrees of psi is a beta sheet region; the Polyproline helix (ppII) region (the phi-psi angle region often occurs in proteins that contain repeating proline residues) is defined as -100° to -30° of phi and 100° to 180° of psi [55].

Figure 1.2 shows the Ramachandran diagram. The Beta-sheet favor region, right handed alpha-helix, left-handed polyproline II helix (PPII) region and left handed alpha-head region is marked. The figure is generated according to the work in (http://www.cryst.bbk.ac.uk/PPS95/course/3_geometry/rama.html).

Different papers may have slightly different definitions of which regions of Ramachandran plot correspond to the secondary structures [56]. Also, for different amino acid residues, the secondary structure propensity e.g. helicity or beta sheet propensity is case specific. For instance, the VAL, TRP, PHE, ILE, THR

and TYR residues favor the beta region in the Ramachandran plot, while ALA, MET, LYS, GLU and ASP residues are preferentially found in the helical region.

Next, I will first illustrate the role of phi-psi distribution of protein in force fields, followed by the introduction of cMAP correction. In many MD simulations, the phi-psi distributions of some residues are found to deviate heavily from the experimentally measured ones. This is one reason that those force fields used are biased to alpha helix or beta sheet. The Grid-Based Backbone Correction gives an extra energy term to correct the free energy distribution of the phi-psi angles by a 2d correction data file, which was first applied by CHARMM software [57]. It utilizes the following bi-cubic interpolation function to calculate the value between the known points in the grid; and it smoothens out the first derivatives and makes the second derivatives continue.

$$f(\phi, \psi) = \sum_{i=1}^4 \sum_{j=1}^4 c_{ij} \left(\frac{\phi - \phi_L}{\Delta_\phi} \right)^{i-1} \left(\frac{\psi - \psi_L}{\Delta_\psi} \right)^{j-1} \quad (1.6)$$

c_{ij} is the coefficients, ϕ, ψ are the phi-psi angles of the point, ϕ_L and ψ_L are the lowest values of phi and psi respectively, and Δ_ϕ and Δ_ψ are the grid sizes.

Side chain torsion potentials and side chain hydrophobic interactions would directly influence the tertiary arrangement of protein components. They relate to the force field accuracy as well [58-60]. The hydrophobic interaction refers to the tendency of nonpolar atoms to aggregate together and bury inwards, and is entropically driven by the tendency that water prefers to interact with polar atoms. Residues with long nonpolar side chains, especially the ones with aromatic rings

exhibit strong hydrophobic interaction effects. The hydrophobic interaction is an important driving force in forming both secondary and tertiary structures during biomolecule *ab initio* folding. There are some beta hairpin structures that are primarily maintained by a hydrophobic core, *e.g.* pro-angiogenic β -hairpin [61]. Most tertiary structures are also maintained by aggregation of inward pointing hydrophobic chains and outward facing polar side chains into water solvent. Besides, in some cases it is important for maintaining the protein-protein interaction interface and ligand-protein binding. Hence, an accurate implicit solvent force field should be able to mimic the hydrophobic interaction.

1.6 Atomic Contact Energy (ACE) based force field

The addition of solvent has a critical effect on atomic interactions within protein residues. However, the explicit calculation of solvent interaction would increase the atom number and hence the degrees of freedom in the simulation system significantly and render the simulation excessively calculation intensive. It is desirable to utilize an implicit solvent potential that incorporates similar solvent effects but without intensive calculations. Among the many implicit solvent models, the atomic contact energy is simple while reasonably accurate. The atomic contact energy (ACE) is the atomic free energy difference arising from atoms exposed to water and atoms isolated from water (in the case of inner protein) [62]. It adopts the previous MJ method developed by S. Miyazawa and R. L. Jernigan [63]. MJ

potential is a knowledge based potential in which energies between two amino acid residues are determined through statistical accounts of the contacts of these two residues from known structures in structure databases, such as Protein Data Bank [63]. A major improvement of ACE is that ACE is based on direct atomic interactions, while the original MJ potential is based on indirect residue interaction data. The ACE free energy of solvation is estimated by calculating the energy difference between atom-atom plus solvent-solvent interactions and atom-solvent interactions.

In order to achieve statistical significance, ACE method uses a large dataset of protein structural data from the PDB database to estimate the atom-atom and atom-solvent interactions [64]. The lattice grid model was used for protein atoms. Since there are no explicit solvent molecules, grids not occupied by the protein atoms are taken as occupied by solvent [65]. Those atom pairs of protein atoms or protein and solvent atoms within a cutoff value are accounted as contact pairs. There are different solvent free energy parameters corresponding to different atom type combinations. Hence the atoms are grouped into 18 types in the ACE based on their chemical properties. For instance, the common $C\gamma$ atoms were grouped together since there is no strong partial charge; similarly, most of carbon atoms in the aromatic ring or phenolic ring are grouped together.

Besides the atomic type group, the reference state definition is also critical to remove the composition bias in the PDB database. It further helps to estimate the

solvent-solvent interactions by giving an equilibrium condition, as PDB structures do not contain information necessary to calculate these interactions. Also, ACE model has systematically adopted a scaling factor to better compare model calculated data to the experimental free energy values that are in units of physics [65]. The ACE model is able to quantify the free energy of transferring amino acids from n-octanol to water. The results are highly correlated to that of experimental measurements with a correlation coefficient of 0.89 [65]. In their original report, the ACE model was used to estimate the binding free energy of protein inhibitors, and produced comparable results with the experimental ones. Since free energy calculated by ACE model is rapid and relatively accurate, the ACE model is widely applied.

1.7 Insufficient and Ineffective Sampling in MD

One major concern in MD simulation is the insufficient sampling problem [66-68], in other words, the full convergence of a system is difficult to achieve. With current computing power and a system size for realistic applications, it is still difficult to achieve a simulation time-scale between microseconds and milliseconds, which is the minimal time required for most biochemical processes to complete. The free energy landscape of a protein is rough and bumpy with many local energy minima separated by high energy barriers [69, 70]. Many protein folds through many intermediate states, and misfolding conformations are likely to form during

the folding process [71, 72]. After equilibration, even small proteins may coexist in more than one state. For example, the small proteins with PDB accession code 2A3D 1PRB 1LMB and 1ENH used in our study all have two major folding states in experiments [73-75]. A normal classical MD simulation requires enormous simulation time-scale to overcome these high energy barriers and sampling may be trapped in local minima. In order to solve the sampling problem in MD simulation, many enhance sampling techniques have been developed and employed to study biological systems [76]. Some of the popular sampling techniques such as umbrella sampling [77, 78], replica-exchange molecular dynamics, and simulation annealing are used to address the sampling problem in different cases.

1.8 Metadynamics and well-tempered metadynamics simulation

The metadynamics simulation provides a novel way to solve the sampling problem by filling the free energy wells with Gaussian hills representing a history dependent bias potential[78]. Through this, the already explored spaces become less favorable to be sampled again, and hopefully push the system out of traps at local energy minima. As the low energy wells have been filled, the free energy landscape of the system becomes more flat. Even sampling of different conformations along the collective variables (CV) becomes possible. The free energy landscape can be constructed by addition of the history deposited Gaussian hills. This method assumes that the system can be effectively described by a few CVs which describe

the reaction coordinates. How to define proper CVs that can describe and represent the conformational change of interest is critical for the success of Metadynamics simulations. The following formula shows the total deposited Gaussians potential during the simulation along the CVs' space.

$$V(s, t) = \sum_{k\tau < t} W(k\tau) \exp\left(-\sum_{i=1}^d \frac{(s_i - s_i(q(k\tau)))^2}{2\sigma_i^2}\right) \quad (1.7)$$

The τ is the stride of Gaussian deposition, σ_i is the Gaussian width of the i th CV, and $W(k\tau)$ is the Gaussian height.

Metadynamics simulations were extensively used since the first introduction by Alessandro Laio and Michele Parrinello in 2002 [79]. Previously, early metadynamics simulation is only limited to smaller systems, due to convergence difficulties. As the software and hardware development progresses, metadynamics now can be performed on larger systems such as the G protein complex [80]. Timothy Clark's group has explored the binding free energies of various ligands to G protein complexes embedded in lipid membranes and fully solvated in water for a time-scale of several microseconds, and it shows that the metadynamics simulation results are highly consistent with experimental data [80].

Many modified variants of metadynamics simulation models exist, most of which are trying to solve the long-time convergence difficulty. Well-tempered Metadynamics is one of them [81]. Instead of using a fixed Gaussian hill height, the height is gradually decreased and meanwhile controlled by an adaptive bias. It is expected to achieve smooth convergence by avoiding large fluctuations of

energy while without increasing the simulation time scale required to fill up the energy surface[82]. Another advantage is the simulation tends to focus on regions of interest, instead of continuously pushing the configurations to nonphysical regions [82].

1.9 Replica-exchange molecular dynamics (REMD)

The REMD is a popular enhanced sampling technique [83, 84], through which, enhanced sampling is achieved by periodic exchanges between systems of similar energies, but different conformations or different temperatures according to the detailed balance criterion. By exchanging of conformations at different temperatures, it allows to overcome the energy barrier, hence accelerate the exploration of the entire conformational space. The exchange is controlled according to the following formula [85]:

$$p = \min \left(1, \frac{\exp\left(-\frac{E_j}{kT_i} - \frac{E_i}{kT_j}\right)}{\exp\left(-\frac{E_i}{kT_i} - \frac{E_j}{kT_j}\right)} \right) = \min(1, e^{(E_i - E_j)(\frac{1}{kT_i} - \frac{1}{kT_j})}) \quad (1.8)$$

The exchange is accepted with the exchange probability p , otherwise the replica i and replica j will not exchange. When the energy inverse probability ($\exp\left(-\frac{E_j}{kT_i} - \frac{E_i}{kT_j}\right)$) of state after the exchange is larger than or equal to the energy inverse probability of state before the exchange ($\exp\left(-\frac{E_i}{kT_i} - \frac{E_j}{kT_j}\right)$), the exchange

probability is 1; otherwise the exchange possibility is related to the ratio of the energy inverse probabilities after and before the exchange.

In a Temperature REMD simulation, the temperature range, number of replicas and interval time for attempting the exchanges should first be defined. In most cases, achieving a 20% to 30% exchange rate would be reasonable for REMD simulations. So these parameters should be defined as to guarantee the exchange rate is around the expected optimal rate, *e.g.* 20~30%. A convenient tool to determine these parameters is the REMD temperature generator available as an online server[86]. In addition, the exchange attempt period should be defined to ensure that there is no obvious correlation within the replicas between two neighboring exchange attempts [87].

1.10 Conformational change of Dengue E protein

Infection by the dengue virus can lead a mosquito-borne tropical disease, with possible symptoms, including fever, headache, vomiting, and skin rash *etc* [88]. Millions of people are infected causing tens of thousands of people die each year[88]. Dengue E protein is the unit that composes the dengue's envelope on the surface of dengue virus, which plays a critical role during the viral genome release to the host cell [89]. In body fluid where pH is near neutral, dengue E proteins exist as dimers, assembling a rigid and smooth viral surface structure. They undergo conformational transformation from transmembrane flat dimers to transmembrane

trimers during the acidification process when the virus is being transported to endosome and during endocytosis [89]. The transmembrane E protein trimer contains an outward facing pointing tip and an exposed fusion loop at the tip, and is able to insert the fusion loop into the endosome membrane, assisting the process of viral fusion, releasing the viral genome content into cytoplasm [89]. The dengue E protein structures and dengue particle were shown in the Figure 1.3.

Figure 1.3 The figure shows the structure of the dengue E protein in different states. Panel a illustrates the structure of the monomer E protein; Panel b illustrate the dimer E proteins, and the EDE epitope region is marked; Panel c shows the structure of the trimeric E proteins in the acid condition; Panel d shows the whole dengue viral particle, the diagram illustrate the particle's different parts.

Recent studies in structural biology and immunology show that a dengue virus with a mutated E protein or an antibody bound and neutralized E protein for both of which the E protein dimer is stabilized at low pH without trimerization, disables the virus from infecting normal cells [90, 91]. A potential risk of dengue virus infection, which sometimes is life threatening, is the antibody dependent enhancement, where sub-optimal non-neutralizing antibodies form complexes with the envelope proteins on surface of virus, being actively transported into immune cells and infecting those cells through phagocytosis thus enhancing the infection [92]. Researchers assume that an effective antibody for dengue disease should be able to neutralize all four serotypes of dengue viruses and has to be able to stabilize E protein dimers of all four serotypes. A common E-Dimer-dependent epitope (EDE) was discovered within the relatively conserved region among all four serotypes. A broadly neutralizing antibody that can bind to this common epitope and is able to stabilize the dimer in low pH would be a potential drug candidate [91].

1.11 Aims and outlines

My work aims to improve implicit solvent models enhancing the accuracy and efficiency of current MD simulations to be performed on large and complex biomolecular systems such as proteins, producing more reliable convergence

results than those in previous MD simulations by utilizing a collection of various techniques. The work consists of three parts.

In the first part of my work, I modified the widely used implicit solvent GB force field (ff14SBonlysc with the combination of GB-Neck2 [27, 93, 94]) by adding a cMAP correction term and a scale for aromatic related interactions. The implicit solvent simulation has significantly enhanced the computational efficiency by reducing the total degrees of freedom comparing to explicit solvent MD simulations. However, implicit solvent GB-based force fields are usually less accurate than the explicit solvent counterparts and require further improvement. I improved the GB-Neck2 combined with ff14SBonlysc force field [27, 93, 94], one of the most accurate implicit solvent models according to literature. I implemented a cMAP potential energy term to adjust the secondary structure propensities for each type of residue. I refined non-bonded parameters relating to side chains containing aromatic rings to better mimic the pi electron-related interactions. The MD simulation results showed improvements in predicting secondary structure propensity as well as tertiary structure accuracy. A test set of 27 peptides or proteins with known native PDB structures and with diversified secondary structure types was constructed to test the accuracy and transferability of our force field.

In the second part of my work, I developed a novel alternative implicit solvent force field by incorporating distance dependent dielectric constant, ACE pairwise

statistical potential and a modified grid-based dihedral energy correction map together, enabling MD simulation of protein with high speed and acceptable accuracy, accessing details of protein native structure formation as well as protein-protein association, and hence understanding many biological processes. The new statistics based force field was tested against a comprehensive and diversified sample set of peptides and proteins. My force field showed potential in *ab initio* folding of peptides or small proteins with reasonable alpha-helix or beta-sheet propensity in yielding native like secondary structures and tertiary structures. My force field showed comparable, if not superior, ability comparing to those currently widely use in implicit solvent simulations, e.g. the GB-Neck2 combined with the ff14SBonlysc force field.

In the third part of my work, I explored the dengue virus envelope (E) protein dimer stability, as it is widely assumed that the E protein dimer is stabilized by some drug ligands or antibodies in an acid environment, neutralizing the virus's ability to fuse with human cells, during which a large conformational change of E proteins is required. I performed MD simulations to mimic the conformational change and stability between dengue's E protein dimer in neutral and acid conditions with the well-tempered metadynamics method. Furthermore, as a few neutralizing antibodies discovered from dengue patients are reported [95], I examined the influence of a selected antibody on dimer stability in both neutral and acid conditions with the same simulation method. I also simulated the antibody

influence with a single point mutated E protein that has been reported to interrupt the protein-antibody interaction and resulted in more than 95% loss of the antibody's binding ability in comparison [95]. My simulation results are highly consistent with the experimental conclusion that binding of the antibody to the E protein dimer neutralizes the virus, especially in a low pH condition, while the mutation of W101A or N153A will significantly reduce the antibody's ability in stabilizing the E protein dimer as in the experimental discovery. I demonstrated that well-tempered metadynamics can be used to accurately explore the antibody's interaction on large protein complexes such as E protein, and it is promising in the future to facilitate antibody development.

2 Improvement of GB simulation performances by adding a cMAP term and strengthening aromatic related interactions

2.1 Background

The explicit solvent MD simulation plays an important role in researching details of biological processes, such as protein binding or protein conformational change. Since biological processes usually happen in microseconds to milliseconds and computational resources are still limited, it is difficult to achieve convergence for larger proteins if the explicit solvent is considered. Alternatively, it is practical to use an implicit solvent simulation model that is accurate and efficient enough to mimic the solvent effect.

As the ability of computational power grows, long simulation times may be practically achieved, but certain limitations of implicit solvent models are gradually exposed [51, 52, 96, 97].

For MD simulation with implicit solvent with GB, there is no current utilized force field that can achieve good transferability among different protein classes as far as is known [50-52, 98]. Although the implicit solvent model was implemented in most software packages, the accuracy and efficiency of mostly GB-force field combinations still have much room for improvement [27, 51, 52]. Taking the AMBER force field as an example, some papers show that using the best GB type

and force field combination can only successfully predict small peptide conformations in limited cases [51, 52]. However, some recently developed GB force field combinations, *e.g.* GB model GB-Neck2 with force field ff14SBonlysc [27, 93, 94], has already shown promising results in producing well-defined native secondary and tertiary structures for some peptides or small proteins [98]. With the aid of modern graphics processing (GPU) units, long time simulations with this method also become practical. Yet, there are still great challenges in the research area for further force field improvement. In the *ab initio* folding test of GB model GB-Neck2 with force field ff14SBonlysc, for 7 out of 14 cases, the most populated cluster did not produce native like structures [98]. Others GB type and force field combinations produced structures that are heavily alpha helical biased or beta sheet biased. Besides secondary structure prediction, the tertiary structures that GB models give are often questionable. In some test cases, significant deviation from the native structure was observed [53]. However, the addition of a grid-based energy correction (cMAP) term showed significant improvement for secondary structure formation in the case of AMBER force field combined with GB [50, 54]. The cMAP term was firstly applied in the CHARMM force field, and demonstrates the important role in correcting the phi-psi distribution of amino acid residues, and achieves much better consistent dynamical and structural properties with experiment [99]. Some researchers have attempted to add the cMAP term to AMBER, and it yields better performances in folding nativelylike peptides. Chen's

group has developed a force field (ff99IDPs or ff14IDPs) specifically for intrinsically disordered proteins (IDPs) based on the AMBER force field and the cMAP derived from an IDP database [100-102].

Some residues, e.g. GLY and PRO play a specific role in protein structure formation. The PRO is the only residue whose side chain is connected to the protein backbone twice, which makes it adopt very limited main-chain conformations. The GLY is the only residue that has hydrogen as the sidechain, which provides it more conformational flexibility. The GLY and PRO are often known as helix breakers. The helicity of each standard residue was well studied by statistical or experimental methods. The maintenance of the backbone phi-psi distribution is critical for secondary structure formation in implicit solvent simulations. To develop a more accurate force field based on a current GB model, special attention should be paid to the hydrophobic interactions, since the hydrophobic interactions involve stabilized beta-hairpin structures in some cases, especially the hydrophobic cores that are formed by π related interactions of residues with aromatic rings [103, 104]. It should be noted that to test the accuracy of a force field, a test dataset with diversified secondary structure compositions is necessary. Fortunately, test sets of peptides or small proteins with known structures and suitable for testing the force field performances are available. A Trp-cage protein is used to check the helix-loop balance; the 1fsd protein can indicate the

helix/beta sheet balance; the WW domain protein is a good example to test beta sheet formation.

If a modified implicit solvent force field is able to fold the small proteins *ab initio*, it can be applied to many applications, such as to explore domain-domain interaction with a linker, and protein interaction with peptides. It would be especially helpful in exploring protein interactions, which begin with unbound partners or even modeled partners [105, 106]. In this study, we first performed a standard GB coupled MD simulation, and examined all the cases that did not form the native-like structure in the most populated cluster. Then, we added cMAP patches to respective residues in order to improve secondary structure propensity. Besides the influence of phi-psi distributions of each residue, the beta strand formation is also influenced by side chain hydrophobic interactions. The beta hairpin of Beta2 peptide is maintained largely by a hydrophobic core formed by four aromatic like rings of TRP [107]. Therefore, the interactions related to atom-type CA (i.e. carbon) in aromatic like rings were scaled to increase the hydrophobic interaction, that better mimics the pi related interactions. The modified force field yields better structural characteristics and dynamic interactions. The balance between different secondary structures is improved in most test cases.

2.2 Method details

2.2.1 Test Data sets of peptides and small proteins

24 peptides (19 cases with native PDB conformation, and 5 without native PDB conformation) and 8 small proteins from various previous MD simulation papers were selected [43, 47, 48, 51, 52, 54, 108-112]. These test samples have known native structure or experimental data about the structure. They contain various secondary structure components, *e.g.* alpha helix, beta sheet, loop, or combinations of them, with a size of 10 to 80 residues. The mechanism of protein folding varies case by case. For some proteins, folding is dominated by electrostatic or hydrophobic contributions while other proteins' folding is highly dependent upon inner residue helicity or beta propensity [113, 114]. It is therefore more favorable to test a force field's transferability and accuracy with a comprehensive test set. Extended PDB structures as well as the topology and coordinate files in AMBER format were generated with tleap in Ambertool 14 [115], and their sequences, some with terminal modifications, are shown in Table 2.1, 2.2.

Table 2.1 Summary of the performed simulations with the respective number of residues (Nres) and the net charge (Nchg) in water at neutral pH for each model peptide.

Peptide	Sequence (Nres/Nchg)	PDB	Structure
CLN025	YYDPETGTWY(10/-2)	2RVD	β -hairpin
TC10b	DAYAQWLADGGPSSGRPPPS(20/-1)	2JOF	α -helix/coil

TC5b(Trp_cage)	NLYIQWLKDGGPSSGRPPPS(20/1)	1L2Y	α -helix/coil
BBA	EQYTAKYKGRTFRNEKELRDFIE KFKGR(28/4)	1fme	Beta/alpha
H1	ACE-KLTWQELYQLKYKGI- NH2(15/2)	*	α -Helix
B1(GB1_hairpin)	GEWTYDDATKTFTVTE(16/-3)	1GB1	β -hairpin
B2(Trpzip2)	SWTWENGKWTWK(12/1)	1le1	β -hairpin
B3	QIFVKTLTGKTITLE(15/1)	1ubq	β -hairpin
WW_domain	GSKLPPGWEKMSRDGRVYYFN HITGTTQFERPSG(35/3)	2F21	β -sheet
Villin_headpiece	MLSDEDFKAVFGMTRSAFANLPL WKQQNLKKEKGLF (36/2)	1vii	α -Helix/loop
EK peptide	YAEAAKAAEAAKAF(14/0)	*	α -Helix
Ribonuclease A C- peptide analog	AETAAAKFLRAHA(13/1)	*	α -Helix
Nrf2 peptide	AQLQLDEETGEFLPIQ(16/-4)	2flu	β -hairpin
1fsd	QQYTAKIKGRTFRNEKELRDFIE KFKGR(28/5)	1fsd	Beta/ alpha
Chignolin	GYDPETGTWG(10/-2)	1uao	β -hairpin
Mbh12	RGKWTYNGITYEGR(14/2)	1k43	β -hairpin
Fs21	Ace-A5[AAARA]3A-NME(21/3)	*	α -Helix
Agd1	Ace-EVLMKVLMEIYLK- NH2(13/0)	*	α -Helix
*The cases with no PDB structure have experimental information related to their secondary structures.			

Table 2.2 Basic information of the small proteins which have basic tertiary structure in the test set.

Peptide	Sequence (<i>Nres/Nchg</i>)	PDB #	Structure
NTL9	MKVIFLKDVKGGMGKKGEIKNVADGYAN NFLFKQGLAIEA (39/3)	2HBA	Helix and Beta sheet
BBL	GSQNNDALSPAIRLLAEWNLDASAIKG TGVGGRLTREDVEKHLAKA(47/1)	2WXC	Helix and Loop
Protein B	LKNAIEDAIAELKKAGITSDFYFNAINKA KTVEEVNALVNEILKAHIEA(49/-2)	1PRB*	Helix and Loop
UVF	MKQWSENVEEKLKEFVKRHQRITQEELH QYAQRLGLNEEAIRQFFEEFEQRK(52/-1)	2P6J	Helix and Loop
EnHD	RPRTAFSSEQLARLKREFNENRYLTERRR QQLSSELGLNEAQIKIWFQNKRAKI(54/7)	1ENH	Helix and Loop
Protein G	MTYKLVIVLNGTTFTYTTEAVDAATAEK VFKQYANDNGVDGEWTYDDATKTFTVT E(56/-5)	1MI0*	Helix and Beta sheet

α 3D	MGSWAEFKQRLAAIKTRLQALGGSEAEL	2A3D	Helix and
	AAFEKEIAAFESELQAYKGKGNPEVEAL		Loop
	RKEAAAIRDELQAYRHN(73/-1)		
λ -	PLTQEQLEAARRLKAIWEKKKNEGLSY	1LMB*	Helix and
repress	ESVADKMGMGQSAVAALFNGINALNAY		Loop
or	NAALLAKILKVSVEEFSPSIAREIY(80/1)		

*have a slightly mutation in sequence from original PBD for faster folding [41].

The dataset was designed to benchmark the force field. Except for Nr2f, all native conformations were retrieved from the PDB database. If a native PDB structure is derived from NMR, the conformation closest to the average is chosen as the representative structure. These data were analyzed for the root mean square fluctuation and root mean square deviation with GROMACS tool `g_rmsf` and `g_rms` respectively. As for Nr2f, the native PDB conformation has several residues missing in the C-terminal, and its native conformation is a beta hairpin according to previous reports [116, 117]. We used the online peptide structure prediction server PEP-FOLD [112, 118] as a tool to obtain the hairpin structure as a reference structure in the RMSD calculation.

2.2.2 Force field used for the simulation

The AMBER package was used for the simulation[115]. The ff14SBonlysc combine GBNeck2 force field [27, 93, 94] with radius type mbondi3 was used to perform the REMD simulation on the comparison group.

The modified ff14SBonlysc combining GBNeck2 force field with radius type mbondi3 was used to perform the REMD simulation on test groups. Two modifications were applied to ff14SBonlysc: adding cMAP for each residue and the double scale of non-bonded force constants for side chain C atom in aromatic residues (atom-type CA in AMBER ff14SBonlysc).

2.2.3 cMAP term and file preparation

The potential energy of AMBER force field combined with cMAP was calculated using the following formula.

$$E_{mm} = E_{bond} + E_{angle} + E_{non-bond} + E_{dihedral} + E_{cmap} \quad (2.1)$$

An additional cMAP correction term was added to the standard AMBER force field. The cMAP is an energy correction matrix derived from the phi-psi angle distribution. The bicubic interpolation method was used to calculate correction energy for each point corresponding to a specific phi-psi value during the simulation [119]. The dihedral angle grids of the correction matrix are in a resolution of 15 degrees.

The patches added to the cMAP for the 20 residue types are shown in Fig 2.1. The patches of ALA, GLU, ASP, MET, GLN and LEU residues were added with the

strength of -1 kcal/mol in the alpha region, and the patch of LYS was added with the strength of -0.2 kcal/mol in the alpha region. For the VAL, TRP, THR, PHE, ILE and TYR residues, the strength of the patches was -1 kcal/mol, added to the beta region. I tested different grid correction strengths, and the above strengths were adopted due to their good performance in forming nativelylike structures. The cMAP correction energy values for the other residues were set to zero (Shown in Figure 2.1). A Perl script inspired by previous Wang et al.'s work was used to add the cMAP into the AMBER topology [101].

Figure 2.1 The cMAP energy corrections applied to each residue is shown in the blue color region. The Figure 2.1a panel shows cMAP energy corrections applied in the residues VAL, TRP, THR, PHE, ILE and TYR with strength of -1 kcal/mol at the beta region. The Figure 2.1b panel shows the cMAP energy correction applied in the residues ALA, GLU, ASP, MET, GLN and LEU with energy

strength of -1 kcal/mol at the alpha region. The Figure 2.1c panel shows the cMAP energy correction applied in the residue LYS at the alpha region with energy strength of -0.2 kcal/mol. The Figure 2.1d shows for other residues no cMAP correction was applied. The prePRO is the amino acid residue immediately before PRO residue.

2.2.4 *Ab initio* folding with REMD Simulations

Two sets of MD simulations were performed to better illustrate the effect of modifications, one with the original force field, and the other with the modified force field.

The original GB simulations were carried out by AMBER package and GPU was used for acceleration. The force field is ff14SBonlysc combined with GB type GBNeck2 (igb=8), and mbondi3 as atomic radius. The cut off was 99.9nm (mimic infinite). The SHAKE algorithm was used to hydrogen is constrained. The Langevin thermostat (ntt=3) was used to control temperature in simulation. The simulation time was 400 ns. First, energy minimization was performed, and followed by a 1 ns equilibration to relax the sidechain, with a time step of 0.002 ps. Then the REMD was performed for the final 400 ns simulation, in order to achieve enhanced sampling. The temperature for each replica was generated by an online server with the range from 300 K to about 700 K [86]

(shown in table 2.3), the time interval for the simulation was 1ns. The replica exchange rate was around 0.2~0.3.

Table 2.3 Temperature values used in the REMD, calculated using the online REMD temperature generator webserver [86].

NAME	Temperature for each replica*						Replica number
1fsd	300.00	327.27	356.30	387.25	421.25	456.62	12
	494.23	534.34	577.13	622.80	671.54	724.54	
1vii	300.00	333.25	369.06	407.52	448.79		12
	493.07	540.65	591.78	646.77	705.83		
	769.42	837.81					
B1	300.00	338.57	380.72	426.81	477.21	532.34	10
	592.70	658.80	731.16	810.54			
B2	300.00	339.81	383.47	431.35	483.90	541.54	10
	604.84	674.09	750.37	834.21			
B3	300.00	341.48	387.09	437.24	492.37	553.20	10
	620.21	694.02	775.43	865.20			
cage	300.00	341.18	386.20	435.49	488.99	548.18	10
	613.01	684.20	762.23	847.88			
H1	300.00	342.47	389.02	440.09	496.21	557.88	10

	625.73	700.34	782.45	872.76			
WW	300.00	325.65	352.88	381.79	412.50	445.11	12
	479.75	516.55	555.67	597.24	641.46	688.45	
Rib	300.00	330.42	363.18	398.52	436.61	477.70	12
	521.98	569.79	621.38	677.04	737.12	801.98	
Nrf2	300.00	337.03	377.36	421.33	469.21	521.41	12
	640.45	708.22	782.16	862.87	950.98		
EK	300.00	338.42	380.28	425.97	475.81	530.20	10
	589.67	654.17	725.22	802.95			
1fme	300.00	327.05	355.84	386.52	419.19	453.97	12
	491.08	530.63	572.80	617.75	665.71	716.85	
1L2Y	300.00	336.27	375.67	418.61	465.28	516.07	10
	571.35	631.59	697.20	768.68			
2JOF	300.00	337.77	378.96	423.92	473.00	526.62	10
	585.21	649.22	719.20	795.95			
2RVD	300.00	345.75	396.50	452.61	514.93	584.20	8
	661.21	746.88					
Agd1	300.00	343.62	391.77	445.03	503.95	569.15	8
	641.37	721.32					
Chignolin	300.00	351.27	408.88	473.75	546.69	628.78	8
	721.25	822.25					

Fs21	300.00 339.56 383.00 430.54 482.60 539.77	10
	602.51 671.34 746.87 829.82	
Mbh12	300.00 339.13 381.91 428.77 480.11 536.34	10
	597.91 665.45 739.48 820.72	
1BY0	300.00 322.18 345.53 370.14 396.05 423.35 452.13	12
	482.46 514.45 548.16 583.70 621.18	
1E0N	300.00 323.74 348.83 375.34 403.38 433.03 464.38	12
	497.55 532.66 569.79 609.09 650.70	
2LD0	300.00 337.50 378.92 424.12 473.56 527.55 586.05	8
	650.50	
2LX5	300.00 325.88 353.49 382.90 414.30 447.55 483.27	10
	521.46 562.17 605.63	
2N0O	300.00 326.96 355.80 386.65 419.61 454.98 492.71	10
	533.02 576.23 622.36	
2N58	300.00 318.12 337.06 356.90 377.77 399.50 422.24	16
	446.02 470.95 497.06 524.38 553.00 582.95 614.34	
	647.15 681.58	
1ENH	300.00 318.34 337.46 357.41 378.23 399.85	20
	422.55 446.25 471.01 496.88 523.89 552.12	
	581.37 612.15 644.37 678.00 713.14 749.89	
	788.29 828.47	

1LMB	300.00	313.45	327.38	341.74	356.58	371.90	24
	387.73	404.10	421.02	438.50	456.56	475.22	
	494.52	514.48	535.10	556.38	578.40	601.18	
	624.71	649.07	674.25	700.27	727.19	755.02	
1MIO	300.00	314.43	329.39	344.90	360.98	377.66	24
	394.97	412.91	431.53	450.84	470.85	491.62	
	513.15	535.52	558.73	582.81	607.80	633.74	
	660.66	688.59	717.57	747.64	778.84	811.24	
1PRB	300.00	316.18	333.02	350.54	368.77	387.74	20
	407.61	428.21	449.62	471.91	495.11	519.29	
	544.47	570.67	597.98	626.45	656.14	687.00	
	719.16	752.65					
2A3D	300.00	318.47	337.84	357.96	379.96	401.94	20
	424.89	448.88	473.95	500.14	527.58	556.20	
	586.12	617.38	650.03	684.20	719.90	757.23	
	796.27	837.09					
2WXC	300.00	323.74	348.83	375.34	403.38	433.03	16
	464.38	497.55	532.66	569.79	609.09	650.70	
	694.72	741.32	790.65	842.88			
2HBA	300.00	315.55	331.65	348.36	365.69	383.67	20
	402.31	421.66	441.75	462.61	484.24	506.70	

	529.98	554.18	579.29	605.36	632.46	660.56	
	689.74	720.03					
2P6J	300.00	320.20	341.39	363.63	386.93	411.40	16
	437.05	463.94	492.16	521.73	552.81	585.50	
	619.87	655.83	693.71	733.34			

*The replica temperature values used here can keep replica exchange rates around 30%.

For the REMD simulation with the modified force field, apart from having added the cMAP in the topology and scaled the VDW interactions of atom-type CA in aromatic rings, all the other simulation procedures and parameters were same as those with the original force field described above.

2.2.5 The data analysis

The g_cluster tool from GROMACS was used to cluster the trajectories with a cutoff of 0.2 nm for backbone atoms by the single linkage algorithm [20]. The RMSD of the native structures with the most populated cluster conformation was calculated using g_rms from GROMACS with the C alpha atoms. The g_confrms tool was used to superpose the simulated cluster structure with the native one. The VMD software was used to generate the snapshot models of the overlapped conformations[120]. For those proteins without a native structure, secondary

structure analysis was carried out by the do_dssp program in GROMACS [121, 122]. The helicity was estimated for each of the residues by measuring the percentage of helix (predicted as A-Helix, 5-Helix, and 3-Helix by dssp) population during the production run. The figure of helicity curve was plotted by Gnuplot [123].

The free energy distribution along the selected two CVs (RMSD and gyration radius) was calculated using inverse Boltzmann formula (free energy = $-RT(\ln(P_i/P_0))$). The counted populations were converted to free energy for all test cases. The radius of gyration value of each frame in the trajectory was calculated with GROMACS tool g_gyrate. The RMSD value of each frame in the trajectory with respect to the native structure was calculated by the GROMACS tool g_rms.

Using the modified force field, 19 peptide cases and 8 small protein cases which have native PDB conformations were simulated from fully extended conformations. *Ab initio* simulations were performed on five more test structures of experimental known helix peptides with no native PDB. Simulations of the entire test set were performed with the original force field for comparison. Here I take the most populated cluster conformation as the predicted native structure.

2.3 Comparison of results between modified and unmodified force fields

Figures 2.2 and 2.3 show the superposition of the most populated cluster conformations with the native ones, and simulation results with the original force field and with the modified force field on the small peptides, respectively. As shown in Fig 2.3 and Table 2.4, with the original force field, the simulated largest clusters are consistent with the native structures ($C\alpha$ RMSD < 0.45 nm) for only 7 out of the 19 peptides. I believe this is because the original force field lacks the ability to account for hydrophobic interactions, which play a critical role in maintaining beta structures. After scaling of the C atom parameters related to aromatic interactions and adding of cMAP, the beta hairpins can be formed by the modified force field (shown in Fig 2.2 and table 2.4).

Table 2.4 *ab initio* folding of peptide test set with the modified force field in comparison with the *ab initio* folding with the original force field. The RMSD is calculated from the most populated cluster relative to the native conformation.

Name	Percentage of most populated	$C\alpha$ RMSD (modified force field)	Percentage of most populated cluster (original	$C\alpha$ RMSD (original force field)	Delta RMSD (nm)*
------	------------------------------------	---	--	---	------------------------

	cluster	(nm)	force field)	(nm)	
	(modified				
	force				
	field)				
1fme	0.578	0.172	0.128	0.477	-0.305
1fsd	0.913	0.294	0.159	0.735	-0.441
1vii	0.46	0.359	0.206	0.763	-0.404
2JOF	0.957	0.403	0.298	0.42	-0.017
2RVD	0.997	0.091	0.999	0.32	-0.229
B1	0.747	0.07	0.66	0.784	-0.714
B2	0.99	0.245	0.978	0.51	-0.265
B3	0.975	0.049	0.902	0.931	-0.882
Cage	0.964	0.074	0.826	0.206	-0.132
Chignolin	0.999	0.087	1	0.355	-0.268
Mbh12	0.986	0.31	0.61	0.196	0.114

Nrf2	0.977	0.498	0.709	0.779	-0.281
WW	0.424	0.417	0.061	0.362	0.055
1BY0	0.339	0.735	0.09	0.484	0.251
1E0N	0.852	0.432	0.206	0.768	-0.336
2LD0	0.904	0.233	0.676	0.422	-0.189
2LX5	0.58	0.394	0.271	0.975	-0.581
2N0O	0.866	0.412	0.549	0.464	-0.052
2N58	0.042	0.596	0.058	0.738	-0.142

*The Delta RMSD is calculated by C α RMSD of the modified force field minus the C α RMSD of the original force field. The more negative value, the greater improvement in terms of RMSD.

Figure 2.2 The superposition of the predicted conformation (Red) by the modified GB onto the native PDB structure (Blue).

The original force field also shows a limitation in forming well defined helical structures as shown in Figure 2.3 and Figure 2.4a. So in the modified force field cMAP patches were added in order to bias the phi-psi distribution of residue ALA LYS MET GLU and LEU to the helix region.

Figure 2.3 The superposition of the predicted conformation with the original GB (Red) on the native PDB structure (Blue).

After the modification, both the helix and beta sheet propensities are found to be more reasonable. Except NRF2, all the test cases can form native-like conformations as the most populated cluster after adding the cMAP and aromatic C alpha atom scaling. Except Mbh12, WW and 1BY0 cases, the other test peptide cases all have improved in terms of the RMSD value (shown in Table 2.4). A total of 78.9 % peptides have improved in terms of RMSD aligning the largest cluster structure to the native one.

2.3.1 *Ab initio* folding of helical peptides

9 helix dominating small peptide samples were used in test simulations, including 4 with native PDB structures available (1BY0, 2LD0, 2LX5 and 2N0O, shown in Figure 2.2, 2.3) and 5 without native PDB structures, but known to be alpha helical in structure (H1, EK, Fs21, Rib and Agd1, shown in Figure 2.4). These 9 peptides of alpha helical structure have diversified amino acid compositions. They have a sequence identity of about or less than 50%.

Figure 2.4 Panel a and b shows the most populated cluster conformations of 5 experimentally known helix peptides and percentages of the most popular clusters over all clusters. Panel c shows the helicity of each residue in those peptides (Red is from the modified force field simulation; Blue is from the original force field simulation).

Among the four peptides with known native PDB structures, 2LD0 and 2LX5 peptides have predicted conformations by the modified force field close to the native ones, while the predicted structures of 2LD0 and 1LX5 peptides have larger RMSD values compared with the native ones when simulated with the original force field (shown in Table 2.4). The predicted structure of 2LX5 peptide is a beta hairpin like conformation, which significantly deviates from the native helix structure (Figure 2.3). Both the original and modified force fields are not performing well in folding 2N00 peptide, with RMSD values 0.464nm and 0.412nm, respectively. The modified force field has poorer performance in terms of RMSD in folding 1BY0 peptide comparing to the original force field, but both force fields successfully predicted the helix feature of the native structure. The poorer prediction by the modified force field is caused by the aromatic like ring of the C terminal residue TRP27 having interaction with the middle region (LEU4, LEU7, LEU11, LYS15) and the turn at the LYS19, which is understandable because I have strengthened aromatic related interactions as well as biased LYS into beta sheet.

All the five helical peptides without known PDB structures are widely taken as test cases in force field validation papers [43, 47, 48, 51, 52, 54, 108-112]. When performing simulations with the original force field, the predicted conformation of H1 peptide is helix like, with a small loop in the C terminal region that is different

from the native structure. The most populated cluster has 82.1% population over the entire cluster (Figure 2.4, Table 2.4). For the EK and RIB peptide cases, the predicted structure is a helix like conformation with flexible loops in two terminal regions and a non-standard helix formed in the medial region. The most populated clusters have 64.7% and 89.3% populations over the entire clusters, respectively. The calculated helicity of EK peptide is relatively low for the residues in the simulation with the original force field (Figure 2.4c). The predicted structure of Fs21 is a beta hairpin like structure, with 1.2% population over the entire cluster. For other non-most populated clusters of Fs21 peptide, no standard helix conformations were found. The predicted structure of Agd1 peptide has formed helix like conformation in two terminal regions, and a turning point in the middle (shown in Figure 2.4). The population of most populated cluster over the entire cluster is 97.9%.

For the REMD simulations of peptides with the modified force field, the predicted structure of H1 and EK is an alpha helix like conformation, shown in Figure 2.4. The most populated clusters have 94.8% and 97.4% populations, respectively. Examining the improvement region of EK peptide, we found the improvement is due to the enhancement of helicity for ALA2 and GLU3 amino acids. The predicted structure of Fs21 is a standard helical conformation, which has 68.3% population over all other clusters. The Fs21 peptides contain 83.4% ALA residues, and its improved performance is mainly due to the helicity enhancement of the

ALA residue. The predicted structure of RIB peptide is helix like conformation in the middle region with a loop in the terminal regions. The predicted structure of Agd1 is helix conformation, which has 99.5% population.

To sum up, the modified force field performs better in forming a native like peptide with helix conformation. For all the five cases, only the folding of Rib has no obvious improvement and both force fields show unsatisfactory performances. The two terminal residue of Rib is ALA. Without explicit solvent the terminal ALA, which has a short sidechain, may have too strong interaction with other polar residues. This may be one possible reason for the poor performance of Rib. From the calculated helicity of each residue shown in Figure 2.4c, it demonstrates higher helicity for most residues in those 5 helix peptides when simulated with the modified force field, comparing to simulating with the original force field. The predicted conformations of H1 and EK peptides from both force fields are helix like (Figure 2.4a,b). The modified force field achieves significant improvement in folding of Fs21 and Agd1 peptides. It indicates the previous force field lacks helicity propensities especially in the ALA residue. By adding the extra patch in the helix region, the helicity of residues can be corrected.

2.3.2 *Ab initio* folding of beta hairpins

7 beta hairpins in the test set, including B1, B2, B3, Mbh12, Nrf2, 2RVD, and Chignolin, are shown in Figure 2.2, 2.3. The sequence identity, except 2RVD and

Chignolin (80% identical), between any two of them is equal to or less than 40%. When simulated with the original force field, the top 3 most populated clusters of B1 peptide are all partly helix conformation and no beta hairpin conformations were observed in all the clusters. The predicted conformation of B2 peptide is helix-loop conformation. The representative conformation of the second most populated cluster is a beta hairpin like structure, with one residue shifted from the standard hairpin. The predicted structure of B3 peptide is a beta hairpin like structure, with one or two residues shifted. It contains 90.2% population over the entire cluster. The predicted structure of Mbh12 peptide is a beta hairpin like structure. Similar to the native structure, the terminal region doesn't form a standard hairpin, and the RMSD value between the native conformation and predicted conformation is 0.196 nm. The predicted structures of Nr2f, 2RVD and Chignolin peptides are not beta hairpins, although there is a turning point in the middle.

For simulation of B1 (GB1_hairpin) and B3 peptides with the modified force field, the predicted structure is a standard beta hairpin. The improvement of folding B1 peptide in the modified force field is mainly due to the enhancement of aromatic related interactions in TRP3, TYR5, PHE12 residues. Consistent with previous reports, the sidechains of these three and VAL14 residues have formed a stable hydrophobic core which holds the two strands together [124]. Two modifications in the modified force field may make the improvement on folding of B3 peptide: one

is that the beta sheet propensities of THR, PHE, VAL, ILE residues have been enhanced, and the other is the enhancement of hydrophobic interaction of the aromatic ring in the sidechain of PHE3. The predicted structure of B2 peptide is a beta like conformation, with one residue shift from the native conformation. The predicted structure of Mbh12 is a beta hairpin like conformation, which has RMSD 0.31nm with the native one. The most populated cluster has 98.6% population. The predicted structure of Nr2f is a helix and loop conformation, and there is a standard beta hairpin conformation in a small cluster. The predicted structures of 2RVD and Chignolin peptides are beta hairpin, which have RMSD 0.09 and 0.087 nm with the native one, respectively. The most populated clusters have 99.7% and 99.9% populations for the two peptides. The improvement in folding these two proteins is mainly due to forming a stable hydrophobic core by sidechains of TRP and TYR residues, and the enhanced beta sheet propensity of THR residues.

All in all, the modified force field performs better than the original force field in forming the beta hairpin conformation in most of these peptides, except Mbh12 and Nr2f peptides. The B1, 2RVD and Chignolin which cannot form native like most populated clusters using the original force field, can form well native like hairpin conformations after the cMAP and aromatic CA corrections.

2.3.3 *Ab initio* folding of other peptides that contain more than one secondary conformation

There are 4 peptides 1vii, 2N58, 1L2Y and 2JOF that contain helix-loop in the native structure. 1L2Y and 2JOF are standard trp-cage conformations in the native state, with 80% sequence identity. When using the original force field in the REMD simulation, the predicted structure of 1vii is a helix-loop-helix conformation. The RMSD value between the predicted structure and native one is 0.763nm, which indicates that the two structures are quite different. For the 1L2Y and 2JOF peptides, a standard trp-cage conformation was formed in the most populated cluster, which has a RMSD value 0.206 nm and 0.42 nm with the native structure, respectively. The most populated clusters have populations 82.6% and 29.8% respectively. There are two peptides 1fsd and 1fme that are comprised of helix and beta strand, and they both have 90% sequence identity. When simulated with the original force field, the predicted structure of 1fsd is a helix-turn-helix conformation. Native like helix-turn-beta only occurred in a few small populated clusters. The predicted structure of 1fme peptide is a native like conformation which has the RMSD value 0.477 nm. The most populated cluster has a population 12.8%. The predicted structure of WW domain peptide is a native like beta sheet conformation, which has RMSD 0.362 nm with the native structure. The most populated cluster has a population 6.1%.

When performing REMD simulation with the modified force field, the predicted structures of 1vii and 2N58 peptides are native like conformations composed of several small helices linked by short loops. The RMSD values between the

predicted and native structures are 0.359 nm and 0.596 nm, respectively. The most populated clusters have 46% and 4.2% populations. As for the improvement on the 1vii peptide, accurately forming the first loop region (VAL10, PHE11, GLY12, MET13, THR14, ARG15) and the second loop region (LEU21, PRO22) contributes considerably to the better results in second and tertiary structures. For the 1L2Y and 2JOF peptides, standard trp-cage conformations were formed in their most populated clusters, which have RMSD values 0.074 nm and 0.403 nm compared with the native one. The populations are 96.4% and 95.7% over the entire clusters, respectively. When simulated with the modified force field, the predicted structures of 1fme and 1fsd are conformations of helix-turn-beta strand, close to the native structures. The two conformations have 57.8% and 91.3% populations over the entire clusters respectively. Not only the predicted secondary structures are consistent with native ones, the interaction pattern between the beta strand and helix in the 1fsd and 1fme peptides are similar to the native ones. The RMSD values between the predicted structures and the native ones are 0.172 nm and 0.294 nm, respectively. The predicted structure of WW domain peptide is a beta sheet with a loop at the terminal region, similar to the native structure. Although a small fraction of helix (at the sequence Pro Pro GLY TRP GLU) was formed near the N-terminus that is not the same as the native loop like conformation in the region, the overall conformation is still close to the native one with a RMSD value of 0.417 nm. The predicted structure of 1E0N is native like,

which has a RMSD value of 0.432 nm with the native conformation. The enhancement of beta sheet propensities of TYR, THR, ILE, PHE residues in residues 1-12 range may be an explanation for the improvement on the 1FSD peptide. The correction of beta sheet propensities in TRP3, ILE5 and ILE6 residue may lead to the improvement of 1E0N peptide folding.

In summary, the modified force field has a noticeable improvement in folding of 1vii and 1fsd peptides, and has similar performances on folding of 1L2Y, 2JOF, 1fme and WW peptides.

2.3.4 *Ab initio* folding of small proteins with tertiary structures

Figure 2.5 a/b shows the superposed conformations of predicted structures and the corresponding native ones for the 8 small proteins with the original and modified force fields, respectively.

a Original forcefield

1ENH

1LMB

1MI0

1PRB

2A3D

2HBA

2P6J

2WXC

b Modified forcefield

1ENH

1LMB

1MI0

1PRB

2A3D

2HBA

2P6J

2WXC

Figure 2.5 The predicted structures of the 8 small protein cases with tertiary structures from the original GB simulations are superposed with the native PDB structures (predicted structures in Red; the native PDB conformations in Blue).

When doing REMD simulations with the original force field, among the 8 small proteins, 2 proteins, 1PRB and 1LMB have formed native like predicted conformations. The RMSD values between the predicted conformations and native ones are 0.41 and 0.36 nm, respectively. The predicted structures of 1ENH and 2A3D peptides are also close to the native, while the predicted structures of 1MI0, 2HBA, 2P6J and 2WXC proteins significantly deviate from the native conformations. The superposition of the predicted and the native ones are shown in the Figure 2.5.

Similarly, when doing simulation with the modified force field, the predicted structure 2A3D is native like. The predicted structures of 1LMB, 2WXC, 1ENH, 1MI0 are close to the native. Among them, the 1MI0 protein is a combination of beta sheet and alpha helix. The predicted structures of 2HBA, 2P6J, 1PRB are not native like in terms of the RMSD value with the native structure. The details of RMSD values between predicted structures and native structures are shown in the Table 2.5. Except the 2HBA 2P6J 1PRB proteins, the modified force field has improvement for the other 5 proteins in terms of RMSD difference (Delta RMSD). The enhancement beta sheet propensities of some residues such THR, TYR, ILE and PHE may lead to the improved performance on 1MI0 peptide. However, the misfolding region around 39-45 (VAL, ASP, GLY, GLU, TRP, THR and TYR) indicates the complexity of folding such small-medium size proteins. The poorer performance of the modified force field in 1PRB protein is caused by the failure of

forming the alpha helix in the residue 20-23 region (ASP, PHE, TYR and PHE). The modified beta propensities as well as aromatic ring related interactions of PHE, TYR and PHE residues, may lead to side effects in a few proteins like 1PRB.

Table 2.5 The *ab initio* folding of the small protein test set with the modified force field in comparison to the *ab initio* folding by the original force field. The RMSD is calculated for the predicted structure relative to the native structure.

Name	Percentage		Percentage		
	of most populated cluster (modified force field)	C α RMSD (modified force field) (nm)	of Most populated cluster (original force field)	C α RMSD (original force field) (nm)	Delta RMSD (nm)
1ENH	0.525	0.805	0.088	1.152	-0.347
1LMB	0.266	0.752	0.206	1.159	-0.407
1MIO	0.532	0.746	0.323	1.213	-0.467
1PRB	0.235	0.779	0.480	0.413	0.366
2A3D	0.589	0.32	0.153	0.423	-0.103
2HBA	0.251	1.299	0.303	0.454	0.845

2P6J	0.455	1.042	0.173	0.997	0.045
2WXC	0.127	0.847	0.025	1.031	-0.184

*The Delta RMSD is calculated by C α RMSD of the modified force field minus the C α RMSD of the original force field. The more negative value, the greater improvement in terms of RMSD.

2.3.5 Free energy landscapes (FEL) of test peptides

The Figures 2.6 and 2.7 show the free energy landscapes of the test peptide cases with the original and modified force fields, respectively.

Figure 2.6 The free energy landscapes as the function of the all atom radius gyration (RG) and the $C\alpha$ RMSD in the peptides simulation with the modified force field. The black arrows point at the native RG value.

Figure 2.7 The free energy surfaces as the function of the all atom radius gyration (RG) and the $C\alpha$ RMSD in the 19 small peptide simulations with the original force field. The black arrows point at the native RG value.

The free energy surfaces are obtained by projecting free energy values onto the two reaction coordinates of $C\alpha$ -RMSD and radius gyration. Simulations with the modified force field can produce deep native state basins in the free energy landscape, while still can sample large areas as simulations with the original force field. For instance, the free energy surfaces of 1fme and 1fsd have deep basins around the RMSD 0.2~0.3nm and radius gyration 0.9~1nm using the modified force field, while using the original force field, the low free energy regions mostly are beyond 0.4 nm. The free energy surface analysis is consistent with previous structure analysis for most of other peptides. The folding processes of 1L2Y, 2JOF and WW domain peptides have similar performances in both force field simulations. The free energy minima are observed to focus on a smaller RMSD region with more compact conformations close to native for other peptides, *e.g.* B1, B2, B3, 1vii, 2RVD, Chingnolin, Mbh12, 1E0N, 2LD0 and 2LX5.

The free energy landscapes of B1, B2 and B3 peptides have more than two major basins, while the one close to native is a major basin in the modified force field simulation. The energy barrier between the two basins are 5~20 kJ/mol. It indicates the modified force field still has the ability to efficiently sample various kinds of conformations, with the native conformation corresponding to the lowest free energy region in most of the cases.

Figure 2.8a, b shows the free energy landscape of the small proteins. The simulations with modified force field have explored similar conformational space

for the 8 proteins comparing to simulations with the original force field. However, except for 1PRB and 2HBA proteins, the energy basins of other proteins are deep and close to native. The free energy landscapes indicate that proteins 2A3D, 1PRB, 1LMB, and 1ENH have more than one state during the simulation with both force fields. Those 4 proteins all have two major states according to experimental studies [73-75].

Figure 2.8 a,b shows the free energy surfaces as a function of the all atom radius of gyration (RG) and the $C\alpha$ RMSD in the small protein simulations with the original and modified force fields, respectively. The black arrows point at the native RG value.

2.4 Conclusion

MD Simulation with implicit solvent is attractive in many applications, because of its higher efficiency while still keeping atomistic details. The combination of GB-Neck2 model with ff14SBonlysc force field is a popular choice to perform implicit solvent simulation with reasonable accuracy, and high sampling efficacy due to low viscosity and no extra computational cost for water atoms. Since all the parameters of ff14SBonlysc force field, *e.g.* bonds, angles, dihedrals, are designed for the MD simulation with explicit solvent, it is necessary to add slight modifications to correct helix and beta sheet propensities, which are critical in forming native like structures. Introducing proper cMAPs terms is a known simple while efficient way to improve the force field accuracy. The combination of ff14SBonlysc with GB-Neck2 is one of the best performed force_field-GB_model combination available according to literature reports. So only a few components of the force field should be changed for further improvement. That justifies that I have not done any systematic iteration to improve the cMAP as in other people's work [15]. By observation of the peptides or small proteins that were failed cases in forming native like structures during the simulations with original force field, we examined the possible responsible residues, and proper cMAP patches were added to the residues in order to tune the helicity and beta sheet propensities. Through examining the results of MD simulation with the combination of ff14SBonlysc with GB-Neck2 model, we found non-bonded interactions are mostly reasonable,

while some aromatic related interactions are slightly weaker than necessary. It was solved by scaling the aromatic C α non-bonded force constants twice. Although I only performed a few modifications, the simulation results are significantly improved. The success rate of structure prediction using the modified force field was 84.2% (16 in 19 peptides within RMSD cutoff of 0.4nm) comparing to the original 47.3 % (9 in 19 peptides within RMSD cutoff of 0.4nm) for small peptides. And there is 62.5% (5 in 8 small proteins) improvement in terms of RMSD difference for 8 small proteins with tertiary structures.

The GB implicit solvent model has advantages in computational efficiency of large protein system. The viscosity of the implicit solvent can be much lower than the explicit solvent, which facilitates the sampling efficiency. Furthermore, performing REMD simulation with implicit solvent model requires much fewer replicas. With the increased accuracy, this modified force field is suitable for investigating larger protein systems, especially those proteins which consist of multiple domains with linkers, or proteins involving large conformational changes.

3 Solvent-free all-atom force field incorporating experimental native structure information

3.1 *Background*

The design of force fields for MD simulations has a long history of continuous development, giving its growing importance in biomolecular research. These simulation methods are generally categorized into three categories: all-atom simulations with explicit solvent, all-atom simulations with implicit solvent, and coarse-grained simulations. All-atom simulations with explicit solvent have higher accuracy, but are too computationally expensive to reach a long time scale, if you have too small computers or simulated systems are too large. Coarse-grained simulations can reach time scales in microseconds or milliseconds, which is comparable to the time scale for biological processes to happen. However, Coarse-grained simulations lose detailed atomic interactions, which are important in many biology related events. Atomic simulation with implicit solvation is promising to achieve high computational speed while at the same time keeping the atomic details. An extra advantage of implicit solvent models is that the simulation is easier to converge, because there is no viscous drag from the explicit water that prevents large conformational changes. However, current force fields with atomic details in implicit solvent simulations are usually unsatisfactory for both speed and accuracy [125-129]. Partially because of the reasons mentioned above, researchers

often use docking studies to explore large scale protein-protein interactions rather than molecular dynamic simulation.

A reasonable force field should be able to reproduce the experimental properties of biological systems, e.g. the tertiary structure, the secondary structure propensity, or the native atomic pairs. One widely used way to test the ability of a force field is *ab initio* folding of small proteins and peptides. Only small-sized proteins or peptides can reach convergence and obtain equilibrium properties from MD simulation given current computational resources available. However, *ab initio* folding of various kinds of peptides and small proteins is still challenging for most of the currently available force fields. In current force fields, e.g. AMBER, OPLS, GROMOS, some amino acids' secondary structure propensities are more biased to helix, while residue ALA's propensity is more biased to beta sheet [54]. It was found that performing simulation with AMBER99sb*-ildn or AMBER ff03* results in difficulties to fold the 1FSD peptide, which indicates the alpha/beta propensity is not accurate in these force fields [46]. While MD simulation using AMBER99sb*-ildn with the nmr1 or cMAP shows outstanding performance in the test on 1FSD peptide [46], only one or a few test peptides cannot guarantee the transferability of the force field in general peptide or small protein systems. There is a benchmark simulation of protein Nrf2 using 10 biomolecular force fields with explicit solvent [116], with simulation times of up to 1 μ s. The simulations with OPLS-AA/L and CHARMM27 were difficult in forming the stable beta sheet

conformation of Nrf2 peptide. Even though some force fields successfully formed beta sheet, there is a residue pairing shift problem in some peptides. Lindorff-larsen et al [111] showed that the CHARMM27 and Amber ff03 force fields can properly handle the helix-loop balance *e.g.* villin headpiece, while it is difficult to achieve proper folding of the beta sheet *e.g.* WW domain. Force fields with explicit solvent have shown consistent remarkable improvement [41-49].

The accuracy and transferability of various available GB methods are also unsatisfactory [27, 50-52]. Even *ab initio* folding of small peptides by simulation has a low success rate, and either secondary or tertiary structure is problematic for many cases [51-53]. However, Perez et al. found that adding a cMAP correction term can significantly improve the performance of *ab initio* folding of peptides [50, 54]. The GLY and PRO amino acid residues have an important role in protein structure formation. The GLY has an important role in forming the turning loop in beta sheet conformations, while the PRO and prePRO (the amino acid residue immediately before PRO residue) are often found at the turning point between helix and helix or helix and beta sheet. This indicates maintenance of the backbone phi-psi distribution is critical for secondary structure formation in implicit solvent simulations. To develop an accurate force field with implicit solvent, many aspects should also be considered besides the phi-psi distribution. The hydrophobic interaction should also be given special attention to, since it is sometimes important in stabilizing the beta-hairpin structure, especially the hydrophobic cores formed

by π related interactions of residues with aromatic rings [103, 104]. The dielectric constant also should be carefully determined for interactions between different parts of the protein, since the dielectric constant is not unified in the protein system [108, 109]. A slight change in electrical interaction can result in a different secondary structure conformation [130].

If an implicit solvent force field is able to achieve *ab initio* folding of small proteins, it can be used in many other applications, *e.g.* exploring domain-domain interactions with a linker, refining docking complexes with flexibility, or sampling conformational ensembles of protein. Including conformational flexibility in MD simulations would be an advantage over a rigid protein docking method, *e.g.* Zdock [131]. Zdock is a popular protein docking software using the ACE potential as a score function to select optimum conformations with low energy [131]. Due to imperfect shape complementarity, unbound rigid docking (the two docking units have an unknown experimental complex structure) is often much harder than the bound one [105, 106]. Since a bound conformation is usually unavailable, the unbound docking problem becomes a major challenge in practical applications of docking. This unbound docking problem indicates the importance of introducing flexibility of protein during docking, where MD simulations may help.

Recently, some researchers used the residue based Miyazawa–Jernigan (MJ) potential to explore protein-protein interactions with coarse grained MD simulations. They showed the residue based MJ potential can accurately keep the

native complex structure with optimized alpha and beta parameters for scaling and shifting [132, 133]. Because the MJ potential contains solvation effects, with proper functions, it is hopeful to mimic the solvation effect without explicit solvent. Compared with current implicit solvent GB, much less computational resources would be required. Kim, *et al.* [134] points out, for strong interactions between proteins, explicitly including side chains is necessary. Using similar methods of obtaining the MJ potential, the all atom based effective atomic contact energies (ACE) were developed by Zhang and co-workers [65]. The effective atomic contact energies are desolvation energies that are required to transfer the atoms from water to the protein interior. The ACE parameters were used in scoring functions of the docking software Zdock for the desolvation term.

In this work, I designed a force field derived from atomic based ACE with the functional form adopted from Hummer's work for the hydrophobic interaction [134]. In Hummer's previous work for the residue based MJ potential simulation, the interfacial hydrophobic interactions between two proteins were simulated using an MJ potential. The residue pairs with attraction adopted a normal LJ functional form, while they designed a pairwise function for the residues with repulsive interactions. In Yang's work, they designed a similar function for balancing the repulsive and attractive terms for the residue based MJ. The bonded and LJ-14 interactions were taken from the GROMOS54a7 force field directly. The MJ force field includes all heavy atoms and polar hydrogens; this makes strong interactions,

e.g. side chain interaction and hydrogen bond, possible. It is interesting to note that the electrical terms from both Yang and Hummer use Debye–Hückel like functions, while values of the dielectric constant used in their papers were significantly different, with 10 and 80 for Yang’s and Hummer’s method, respectively[134, 135]. Here in our work, a distance dependent dielectric constant was used, in order to overcome the limitation of such arbitrarily fixed dielectric constants. Hummer points out that only weak interactions can be mimicked by the coarse grained MJ potential, while the strong protein-protein interaction involving side chain effects cannot be accurately determined by this simplified model[134]. Since it is widely known that the backbone dihedral is highly correlated with protein secondary structure, simply combining the ACE non-bonded term with the bonded parameters from other force fields like GROMOS54a7 without fine tuning will cause incorrect distributions of backbone dihedrals. Hence such force fields have difficulties in forming correct secondary conformations during the *ab initio* simulation. The dihedral energy correction (cMAP) was reported to be able to refine the backbone dihedral significantly. With available backbone distributions in a PDB dataset, *e.g.* TOP500 [136], or the Protein Coil Library dataset (a structural database of nonhelix, nonstrand fragments derived from the PDB) [137], it is possible to make a CMAP that biased the backbone phi-psi distribution close to the PHI/PSI distribution from the PDB dataset by an iterative inverse Boltzmann method.

The resultant force field provides a choice to study the assembly of large biomolecular complexes with structural details within reasonable computational time. Comparing to a structure based all atom Gō-type model which is also computationally efficient, our force field can generate much more flexible inter and intra domain interactions, enabling relatively large conformation changes [138]. This is important for some biological processes that involve significant conformational changes. Interface flexibility is also playing a critical role in protein-protein interaction, *e.g.* antibody-antigen recognition. It can also be important in exploring the domain-domain interactions within the linker, in which the linker's contribution can be significant.

The data presented in this report suggest that the statistics based ACE potential can directly be applied to simulations in addition to being limited to the rigid docking scoring function. However, carefully dealing with the dielectric constant, main chain phi-psi distributions and counter-ions is necessary to tune a force field available in most cases. Unlike the previous MJ based force fields, a suitable dielectric constant was carefully chosen for each defined energy group and intra energy group based on the physical meaning and testing of secondary structure maintenance.

3.2 *Method details*

3.2.1 Data set collection

In order to use test cases that can cover most of the protein classes, I choose a large test set from many other reports [43, 47, 48, 51, 52, 54, 108-111]. All the test peptides and small proteins are widely used and have native structures or experimental data related to the native conformation. The native structures of the test cases cover various secondary structure types, *e.g.* alpha helix, beta sheet, loop, and their combinations as well as different sizes range from 10 residues to 80 residues (Table 3.1 and 3.2). Since the *ab initio* folding mechanism of different classes of proteins can be significantly different, it is more reliable to test force field transferability and accuracy with such a comprehensive dataset.

Table 3.1 Summary of the training set peptides with the respective number of residues (Nres) and the net charge (Nchg) in water at neutral pH for each model peptide.

Peptide	Sequence (Nres/Nchg)*	PDB #	Structure
CLN025	YYDPETGTWY(10/-2)	2RVD	β -hairpin
TC10b	DAYAQWLADGGPSSGRPPPS(20/-1)	2JOF	α -helix/coil
TC5b(Trp_cage)	NLYIQWLKDGGPSSGRPPPS(20/1)	1L2Y	α -helix/coil
BBA	EQYTAKYKGRTFRNEKELRDFIEKFKG R(28/4)	1fme	Beta/ alpha

H1	Ace-KLTWQELYQLKYKGI-NH2(15/2)		α -Helix
B1(GB1_hairpin)	GEWTYDDATKTFTVTE(16/-3)	1GB1	β -hairpin
B2(Trpzip2)	SWTWENGKWTWK(12/1)	1le1	β -hairpin
B3	QIFVKTLTGKTTITLE(15/1)	1ubq	β -hairpin
WW_domain	GSKLPPGWEKRMSRDGRVYYFNHITG TTQFERPSG(35/3)	2F21	β -sheet
Villin_head iece	MLSDedFKAVFGMTRSAFANLPLWKQ QNLKKEKGLF (36/2)	1vii	α -Helix/loop
EK peptide	YAEAAKAAEAAKAF(14/0)		α -Helix
Ribonuclease A C-peptide analog	AETAAAKFLRAHA(13/1)		α -Helix
Nrf2 peptide	AQLQLDEETGEFLPIQ(16/-4)	2flu	β -hairpin
1fsd	QQYTAKIKGRTFRNEKELRDFIEKFKG R(28/5)	1fsd	Beta/ alpha
Chignolin	GYDPETGTWG(10/-2)	1uao	β -hairpin
Mbh12	RGKWTYNGITYEGR(14/2)	1k43	β -hairpin
Fs21	Ace-A5[AAARA]3A-NME(21/3)		α -Helix
Agd1	Ace-EVLMKVLMEIYLK-NH2(13/0)		α -Helix

*During the simulation counter ions were used to remove the net-charge.

Table 3.2 The test set of the 8 fast folding small proteins and the 8 peptides.

Peptide	Sequence (Nres/Nchg)*	PDB #	Structure
NTL9	MKVIFLKDVKGMGKKGEIKNVADGYA NNFLFKQGLAIEA (39/3)	2HBA	Helix and Beta sheet
BBL	GSQNNDALSPAIRRLAEWNLDAIAIK GTGVGGRLTREDVEKHLAKA(47/1)	2WXC	Helix and Loop
Protein B	LKNAIEDAIAELKKAGITSDFYFNAINK AKTVEEVNALVNEILKAHIEA(49/-2)	1PRB*	Helix and Loop
UVF	MKQWSENVEEKLKEFVKRHQRITQE ELHQYAQRLGLNEEAIRQFFEEFEQR K(52/-1)	2P6J	Helix and Loop
EnHD	RPRTAFSSEQLARLKREFNENRYLTE RRRQQLSSELGLNEAQIKIWFQNKRA KI(54/7)	1ENH	Helix and Loop
Protein G	MTYKLVIVLNGTTFTYTTEAVDAAT AEKVFKQYANDNGVDGEWTYDDAT KTFTVTE(56/-5)	1MI0*	Helix and Beta sheet
α 3D	MGSWAEFKQRLAAIKTRLQALGGSE AELAAFEKEIAAFESELQAYKGKGNP EVEALRKEAAAIRDELQAYRHN(73/- 1)	2A3D	Helix and Loop
λ -repressor	PLTQEQLEAARRLKAIWEKKKNELGL SYESVADKMGMGQSAVAALFNGINA LNAYNAALLAKILKVSVEEFSPSIARE IY(80/1)	1LMB*	Helix and Loop
1F8P	YPSKPDNPGEDAPAEDLARYYSALRH	1F8P	Helix-loop

	YINLITRQRY-NH2(36/1)		
1J5B	DVASDAKAAAELVAANAKAAAELV	1J5B	helix
	AANAKAAAEAVAR-NH2(37/0)		
1LE3	GEWTWDDATKTWTWTE-NH2(16/-2)	1LE3	Beta hairpin
1LOI	MPLVDFFCETCSKPWLVGWWDQFK	1LOI	Helix-turn-helix
	R-NH2(25/1)		
1MBJ	VKKTSWTEEDRIIYQAHKRLGNRW	1MBJ	Helices and Loops
	AEIAKLLPGRTDNAIKNHWNSTMRR		
	KV-NH2(52/7)		
2N31	MATTVSTQRGPVYIGELPQDF(21/-1)	2N31	Beta hairpin-loop
2N5L	GPHMGDLAKERAGVYTKLCGVFPPH	2N5L	Helices and Loops
	LVEAVMRRFPQLLDPQQLAAEILSYK		
	SQHLS(56/1)		
2N88	GLEYYAVAESVIGKRVGDDGKTIEYLV	2N88	Beta sheet -helix-loop
	KWTDMSDATWEPQDNVDSTLVLLY		
	QQQQPMN(57/-6)		

*The number of residues (Nres) and the net-charge (Nchg) at neutral pH for each model peptide were shown in the bracket.

During the simulation counter ions were used to neutralize the net charge in the system.

3.2.2 Functional forms and parameters

The function for hydrophobic interaction was adopted from Hummer's method [134]. The hydrophobic interaction is divided into attractive and repulsive classes depending on the sign of $\epsilon_1(i, j)$. The attractive interaction is a normal LJ function, while the repulsive interaction is a stepwise function as follow (Figure 3.1):

When $\epsilon_1(i, j) > 0$:

$$E_{ace} = \sum_{i < j - 3}^{native} \varepsilon_1(i, j) \left[\left(\frac{\sigma_{ij}^{vdw}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}^{vdw}}{r_{ij}} \right)^6 \right] \quad (3.1)$$

When $\varepsilon_1(i, j) < 0$ and $r_{ij} < \sigma_{ij}^{vdw}$:

$$E_{ace} = \sum_{i < j - 3}^{native} \varepsilon_1(i, j) \left[\left(\frac{\sigma_{ij}^{vdw}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}^{vdw}}{r_{ij}} \right)^6 \right] + 2\varepsilon_1(i, j) \quad (3.2)$$

When $\varepsilon_1(i, j) < 0$ and $r_{ij} > \sigma_{ij}^{vdw}$:

$$E_{ace} = \sum_{i < j - 3}^{native} -\varepsilon_1(i, j) \left[\left(\frac{\sigma_{ij}^{vdw}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}^{vdw}}{r_{ij}} \right)^6 \right] \quad (3.3)$$

Where $\varepsilon_1(i, j)$ is the original ACE energy value multiplied by -1, unit was converted to kJ/mol by multiplying 1/21.

Figure 3.1 The figure shows the non-bonded interaction form. The attractive part is an LJ functional form, and the repulsive part is a stepwise function. For convenience, the interaction measure from the MD simulation of two bead model is shifted up for 0.1 kJ/mol. The energy unit is kJ/mol.

The atom types in the ACE were adopted from the work of Zhang and co-workers[65]. The Van der Waals radii were adopted from the united atom force field GROMOS54a7.

The bond, angle, dihedral parameters and charges were adopted from the united atom force field GROMOS54a7. The LJ14 is also adopted from the GROMOS54a7 force field.

There are 12 types of atom in the origin Zhang's paper [62]. When combined with different VDW radii, we generated more than 67 types of atoms, with about 1942 non-bonded interaction pairs. The non-bonded interactions were written in the [nonbond parameter] section of topology of GROMACS.

The ACE parameters were from the Zhang's paper [65], and the values are converted to values with units of kJ/mol. The 1-4 interaction was directly from the GROMOS54a7 force field, and the electrical interaction was scaled by 0.5. The GROMOS54a7 is a united atom force field, as no non-polar hydrogen is included. The polar hydrogens were kept for their importance in hydrogen bond formation, and hence for the stability of secondary structure. There are 3 polar hydrogen types, the backbone Nitrogen hydrogen (H), the aromatic connected hydrogen (HC) and other Nitrogen/Oxygen connected hydrogen (HH). As the original atomic contact energy (ACE) has no parameters of hydrogen, the non-bonded VDW parameter on HC, is directly chosen from the GROMOS54a7 force field. HH and H in GROMOS54a7 are defined as zero for C6 and C12 of LJ function. So HH and H

are given C12 values 10^{-6} and 10^{-12} for corresponding non-bonded VDW parameters. HH and H interactions with ION are set to 10^{-6} . All these small modifications are intended to avoid unwanted steric clashes. The repulsive value of polar hydrogen related interactions will directly influence the strength of hydrogen bonds, through which secondary structure is maintained. It serves as an alternative adjustment to helix and beta sheet propensities.

3.2.3 Parameters of Distance Dependent Dielectric constant and side chain

VDW interaction scaling

Since dielectric constants in different pairs of atoms are highly dependent on extent of atoms' exposure and distances, here a widely used sigmoidal distance dependent dielectric function was applied to approximate the dielectric constant [139-141]. For different parts of the protein we use different parameter sets for the function inspired by a previous work [140], shown in Figure 3.2.

The electrostatic function for the 1-4 interaction is the standard electrostatic interaction function defined in GROMACS with a simple 0.5 scale. We divided the simulation system into four subsets of atoms (CHG, MainChain, Prores and ION) based on the extent of exposure to water and the charge group of GROMOS54a7 force field. The charge group information can be found in the topology of GROMOS54a7 force field. The CHG group was defined as all the atoms in a charge group without zero total charge. All atoms within the same charge group as

MainChain+H were defined as MainChain group. The Ion group was defined as all the ions used in the system. The Prores is defined as all the rest atoms besides CHG MainChain and ION groups.

The parameters of electrostatic interaction within the MainChain group atoms were $M=1$, $N=77$, $d=1.7$, $f=0.2$, $c_1=7.6$, $c_2=5.8$, $c_3=4$, $c_4=2.47$, $c_5=89$.

The parameters of electrical interaction within the charge group atoms were $M=15$, $N=63$, $d=2$, $f=0.2$, $c_1=7.6$, $c_2=5.8$, $c_3=4$, $c_4=2.47$, $c_5=89$.

The parameters of Ion related interaction were $M=15$, $N=63$, $d=1.6$, $f=0.05$, $c_1=7.6$, $c_2=5.8$, $c_3=4$, $c_4=2.47$, $c_5=89$.

The residues not in CHG MainChain or Ion were defined as Prores. The parameters of interaction between CHG and Prores or MainChain were $M=10$, $N=68$, $d=2$, $f=0.2$, $c_1=7.6$, $c_2=5.8$, $c_3=4$, $c_4=2.47$, $c_5=89$.

The parameters of other interaction besides above were $M=3$, $N=75$, $d=2$, $f=0.2$, $c_1=7.6$, $c_2=5.8$, $c_3=4$, $c_4=2.47$, $c_5=89$.

$$D_{\text{eff}} = M + N \frac{(r/A)^n}{1+(r/A)^n} \quad (3.4)$$

Where

$$A = c_1 \ln[1.5 + c_2(d - 1.59)] + c_3 f \quad (3.5)$$

$$n = c_4 + \frac{r^2}{c_5} \quad (3.6)$$

This schedule is an aggressive approximation, but it still can keep acceptable accuracy.

Figure 3.2 The dielectric constant versus the distance between the different types of interaction partners. The unit for the x axis is ns.

For the interaction within side chain beginning from atom-type CB (called side chain here), I scale the MJ parameter by 7 except CB in the Proline, and the side chain and main chain LJ interaction was scaled by 2. This helps to form the hydrophobic core within a protein, hence maintaining the tertiary structure. The VDW interactions related to CB atom in the Proline were scaled by 12, in order to better mimic the interaction of aromatic like rings in Proline which is similar to the Pi interaction in many cases [142].

3.2.4 Making grid-based dihedral energy correction map (cMAP) for backbone dihedral correction

The experimental and computational procedures I used to compute grid-based phi/psi energy distributions followed a previous study [143].

Firstly, I derived a phi/psi experimental Ramachandran Plot from the coil lib dataset [136, 144] as an initial cMAP target distribution. The coil lib dataset is from the online website (<http://folding.chemistry.msstate.edu/coil/>) [144], and it contains coil fragments extracted from high-resolution crystal structures. This data allows us to represent the tendencies for each type of amino acids, whether they are involved in secondary structure or not, and separates special cases like prePRO (the amino acids immediately before proline).

From the population densities in the Ramachandran plot, an approximation to the shape of the free energy landscape can be built qualitatively by binning the space and using the following equation.

$$\Delta G_{\text{pdb}} = RT(\ln(n/N)) \quad (3.7)$$

The temperature T was 300K, and R is the Boltzmann constant multiplied Avogadro constant with the unit of KJ/mol. The parameter n is the population of phi-psi falling in a specific bin box; N is the total population. Both phi and psi range from -180 degrees to 165 degrees according to the cMAP format in GROMACS. The function provides an experimental distribution in a qualitative way. 21 unique target cMAPs (taken as initial cMAPs for later iterations) were derived from the coil lib using this method, including 20 standard amino acids and the prePRO case.

3.2.5 Iterative simulations to generate converged cMAPs

The 20 dipeptides (ACE-X-NHE) and one Tripeptide (ACE-X-PRO-NHE) were created by Tleap for each amino acid with ACE and NHE as terminal blocks. All the peptides were used to perform REMD simulations with the initial cMAP, and the cMAPs were iteratively refined using the following method, until convergence. The convergence is defined by no further improvement for 5 iterations. The REMD simulation is with 6 replicas and temperature from 300K to 800K.

The inverse Boltzmann method was widely used for this type of iteration [145]. But here an extra patch (with the small value 0.1 kJ/mol) was added to the alpha region [146] (phi from -160 to -20 and psi from -120 to 50) and beta region (phi from -180 to -90 and psi from 50 to 240) based on alpha/beta ratio at each iteration. The Ramachandran plot phi/psi distribution was converted into the cMAP format in GROMACS, and was taken as an initial cMAP to run the simulation using the following procedure:

1. Energy minimization.
2. Equilibrate for 100 ns for each replica in the REMD with corresponding temperature.
3. Simulation with 300 ns with REMD.

After the simulation, the phi/psi distribution in the trajectory is derived by the following formula:

$$\Delta G_{\text{sim}} = RT(\ln(n_s/N_s)) \quad (3.8)$$

The temperature T was 300K, R is the Boltzmann constant multiplied Avogadro constant with unit of kJ/mol. The parameter n_s is the number of frames that have phi-psi value falling in a specific bin box over the simulation trajectory; N_s is the total number of simulation frames.

$$\Delta G_{\text{diff}} = KT \left(\ln \left(\frac{B}{A} \right) \right) - KT \left(\ln \left(\frac{B_t}{A_t} \right) \right) \quad (3.9)$$

Update cMAP by

If $\Delta G_{\text{diff}} > 0$,

Alpha helix region:

$$E_{\text{cmap}} = E_{\text{cmap}} - (\Delta G_{\text{sim}} - \Delta G_{\text{pdb}}) \times \$\text{scale} - 0.1 \quad (3.10)$$

Beta sheet region:

$$E_{\text{cmap}} = E_{\text{cmap}} - (\Delta G_{\text{sim}} - \Delta G_{\text{pdb}}) \times \$\text{scale} + 0.1 \quad (3.11)$$

Other region:

$$E_{\text{cmap}} = E_{\text{cmap}} - (\Delta G_{\text{sim}} - \Delta G_{\text{pdb}}) \times \$\text{scale} \quad (3.12)$$

If $\Delta G_{\text{diff}} < 0$

Alpha helix region:

$$E_{\text{cmap}} = E_{\text{cmap}} - (\Delta G_{\text{sim}} - \Delta G_{\text{pdb}}) \times \$\text{scale} + 0.1 \quad (3.13)$$

Beta sheet region:

$$E_{\text{cmap}} = E_{\text{cmap}} - (\Delta G_{\text{sim}} - \Delta G_{\text{pdb}}) \times \$\text{scale} - 0.1 \quad (3.14)$$

Other region:

$$E_{emap} = E_{emap} - (\Delta G_{sim} - \Delta G_{pdb}) \times \$scale \quad (3.15)$$

B in the formula is the beta region population in simulation; A is the alpha region population in simulation; B_t is the beta region population of the target; A_t is the alpha region population of the target; The \$scale is a scale factor (with value 1 in this work).

The whole iterative process is demonstrated in Figure 3.3.

Figure 3.3 The procedure to generate the final cMAP. The ultimate goal of the dipeptide iterative simulations is to obtain the cMAPs that can generate similar phi-psi distributions as in the coil database. The patch adding step is to further correct the cMAP, in order to fold to the native like conformation for a variety of peptide cases.

With the ΔG_{diff} related adjustment, I aimed to generate a cMAP that can produce a phi-psi distribution with similar ratio of alpha helix and beta sheet propensity as in the target distribution.

3.2.6 Further modification of the cMAP by adding a patch

In order to add proper patches to maintain the secondary structure balance in simulation, a test set of 16 peptides with known experimental secondary structures were chosen as the training set. The training set cases were chosen from various literature reports related to peptide or small protein MD simulations [34, 46, 147]. The native structures were downloaded from the PDB database bank [148]. If the NMR structure contains several frames, only the structure has the lowest RMSD with reference to the average structure was chosen as a representative structure (done by GROMACS tools `g_rms` and `g_rmsf`).

Since the sidechain environment of backbone in the loop lib is more complex than the sidechain environment of simple dipeptide, a training data set was used to test the accuracy of the force field with cMAP. The cMAP was generated by iteratively

biasing to the phi-psi distribution from the loop lib containing deviation from real peptide or protein simulations. Here I manually added extra values to the helix region, beta region or PII region in the cMAP. I used a narrow definition for these regions, which follows previous literature work [55]. The helix region for phi ranges from -100 to -30, and for psi it ranges from -67 to -70. The beta region for phi ranges from -180 to -100, and for psi it ranges from 120 to 180. The PII region ranges from -100 to -30, and for psi it ranges from 120 to 180. The strength of the patch for each residue is case specific, and is partly guided by already known amino acid helicity and beta sheet propensity. Each time after modifying the cMAP, *ab initio* simulations with the new cMAP are taken for the training set. Only if the success rate is improved (using RMSD to the native state and percentage of secondary structure consistent to native structure as main criteria), the cMAP was kept. The final cMAPs we used are shown in Figures 3.4, 3.5.

3.2.7 Testing the force field by simulating a test set with known experimental secondary structures

The final MJ force field parameters were determined by iterative training over the 19 peptides, which include various secondary structures. In order to validate the ability of the force field in predicting or maintaining peptide and small protein native structure, the training set and test set were taken, and simulations were performed with the final parameters. The replica exchange molecular dynamics

(REMD) simulation method was used in the simulation to enhance the sampling efficacy [85, 149]. After the simulation, a clustering analysis was performed by the single linkage method which is implemented in GROMACS tools. For proteins with length < 20 , between 20~40, and > 40 amino acids, the corresponding cutoffs of 2.0 Å, 3.0 Å, 4.0 Å were used respectively. If the most populated cluster conformations are close to the native one, it indicates the force field can be used for *ab initio* folding of this peptide.

For each REMD simulation, the NVT ensemble simulation was carried out at 300 K to equilibrate the densities of initial periodic box. Subsequently, an NPT ensemble simulation was conducted for 1 ns with the temperature of each replica respectively. The temperature for each replica was adopted according to a recent study [86] to give nearly uniform exchange rates about 20~30%. The exchange interval is 10 ps. The temperature ranges from about 300K to 600K or up to 800K. The electrostatic and van der Waals interaction was treated using a cutoff of 1.6 nm in all simulations. The velocity rescaling thermostat was adopted for maintaining constant temperature. The sd (leap-frog stochastic dynamics integrator) was used as the integrator, and the time constant for coupling (τ) was set to 0.05. All bonds involving hydrogen were constrained using LINCS (52), and a time step of 2 fs was used. The ionic side chains of ARG, LYS, ASP, and GLU amino acid residues were neutralized with counter-ions (Cl^- or Na^+) according to the pH 7 condition.

Figure 3.4 The figure shows the final version of cMAP energy correction applied for the 20 residues respectively. The energy unit is kJ/mol, and the applied energy is gradually increased from blue to white.

prePRO

GLU chi1 chi2

Figure 3.5 The figure shows the final version of cMAP energy correction applied in the prePRO residue (residue immediately before the PRO) and in the sidechain chi1-chi2 of the GLU residue. The energy unit is kJ/mol, and the applied energy is gradually increased from blue to white.

Using the final MJ force field parameter set, all the cases which have native PDB conformations in the training set were simulated from fully extended conformations. Another 5 experimentally known helical peptides which have no native PDB conformations were simulated for *ab initio* folding as well. For the test data set, 8 small fast folding protein cases and 8 peptide cases were simulated from fully extended conformations to further validate the transferability of the MJ force field. All the peptides or small proteins in the training set and the test set were also simulated with a GB force field (the ff14SBonlysc with the combination of GB-Neck2) for comparison. Here we take the most populated cluster conformation as the predicted structure.

3.2.8 Convergences test for GB simulation, chapter 2's modified GB simulation and MJ simulation

In order to check the convergences of the simulation, the last 300 ns simulation trajectory was divided into 3 blocks (101 ns to 200 ns, 201 ns to 300 ns, 301 ns to 400ns). The free energy landscape as function of the all atom radius of gyration (RG) and the C α RMSD was calculated for each block. By examining the free energy landscape similarity within those 3 blocks, it can roughly check the convergences. The results indicate most of the simulation is converged in terms of radius of gyration (RG) and C α RMSD property. For instance, the free energy landscapes of each trajectory block of peptide 1fsd in simulation by GB simulation, chapter 2's modified GB simulation and MJ simulation were shown in Figure 3.6.

a. GB model

b. Modified GB model

c. MJ model

Figure 3.6 The free energy landscapes of each trajectory block of peptide 1fsd in simulation by GB simulation (panel a), chapter 2's modified GB simulation (panel b) and MJ simulation (panel c). The simulation trajectories from these three force field were selected, and divided into 3 blocks . The convergences were roughly checked by comparing the similarity of free energy landscape within blocks.

3.3 Comparing the simulation results of peptides and small proteins between MJ and GB force field

Figures 3.7, 3.8 shows the superposition of the most populated cluster conformation and the native one, which are from the MJ force field and the GB force field simulations for the small peptide cases, respectively.

Figure 3.7 The superposition of the predicted conformation (Red) by the MJ force field and the native PDB structure (Blue).

The simulation result with the GB force field is performed well for about 7 out of the 19 cases as shown in Figure 3.7, 3.8 and Table 3.3.

Figure 3.8 The superposition of the predicted structure (Red) by the original GB and the native PDB structure (Blue).

One serious limitation of the GB force field is in maintaining beta sheet, where hydrophobic interactions play a critical role. The GB force field also shows limitations in forming well defined alpha helical structure as shown in Figure 3.7 and Fig 3.8a. While for the MJ force field, both the helix and beta sheet propensities are more reasonable in many cases. Except the 1BY0 and 2N58, most of training cases can form nativelylike conformations as the most populated clusters. The Mbh12 cage WW 1BY0 1fme and 2N58 cases have shown no improvement compared to GB, while all other test peptide cases have improvements in terms of the RMSD value (shown in Table 3.3).

There are totally about 73.7 % (14 out of 19) improved cases according to the RMSD of the most populated cluster conformation as compared to the native structure, for the small peptides in the training set. The performances of the MJ force field have also improved in terms of the secondary structure measurement, for the 5 helix cases without native PDB conformation. About 75% (6 out of 8) cases have improved over the GB force field in terms of RMSD for the small proteins in the test set, and about 75% (6 in 8) cases show better performance in terms of RMSD for the other 8 peptides in the test set.

3.3.1.1 *Ab initio* folding of helical peptides in training set

There are nine small alpha helix dominant peptides used in our training set simulation, including 4 cases with native PDB structures available (1BY0, 2LD0,

2LX5, 2N0O, shown in Figure 3.7,3.8) and 5 cases without native PDB structure available but with known alpha helix conformations in experiments (H1 EK Fs21 Rib and Agd1, shown in Figure 3.9). The nine helix peptides have various amino acid compositions, with sequence identity equal to or smaller than 50%.

a By GBSA force field

b By MJ force field

c

Figure 3.9 Panel a,b show the predicted structures of 5 experimentally known helix peptides and percentages of the most popular clusters over all clusters. Panel c shows the helicity of each residue in those peptides (Red is from the MJ force field simulation, Blue is from the GB force field simulation).

Among the four peptides with known native PDB structures, 2LD0, 2LX5, and 2N0O peptides have predicted conformations by our force field close to the native structures. The predicted structures of 2LD0 and 1LX5 have larger RMSD with the native ones when simulated with the GB force field (shown in Table 3.3), especially 2LX5 peptide formed a beta hairpin like most populated conformation that is far away from the native one (Figure 3.8). The 2N0O peptide did not perform well with both force fields, with RMSD values 0.738 nm and 0.83 nm respectively. The folding of 1BY0 peptide is an exception, which has poorer performance in term of RMSD by the MJ force field compared to the GB one, but the predicted conformation with the MJ force field still contains partial helix conformation.

Table 3.3 The *ab initio* folding of peptide Set A with the MJ force field in comparison to the *ab initio* folding by AMBER14_GB force field. The RMSD

values are calculated for the most populated cluster related to the native conformation.

Name	Percentage of most populated cluster (MJ force field)	C α RMSD (MJ force field)	Percentage of most populated cluster (original force field)	C α RMSD (original force field) (nm)	Delta RMSD* (nm)
1fme	0.596	0.606	0.128	0.477	0.129
1fsd	0.532	0.454	0.159	0.735	-0.281
1vii	0.783	0.553	0.206	0.763	-0.210
2JOF	0.99	0.355	0.298	0.420	-0.065
2RVD	1	0.11	0.999	0.320	-0.21

B1	1	0.192	0.66	0.784	-0.592
B2	1	0.072	0.978	0.510	-0.438
B3	0.467	0.375	0.902	0.931	-0.556
1L2Y	0.975	0.356	0.826	0.206	0.150
Nrf2	0.53	0.461	0.709	0.779	-0.318
Chignolin	0.997	0.23	1	0.355	-0.125
Mbh12	0.999	0.168	0.61	0.196	-0.028
WW	0.384	0.745	0.061	0.362	0.383
1BY0	0.532	0.932	0.09	0.484	0.448

1E0N	0.64	0.438	0.206	0.768	-0.330
2LD0	1	0.405	0.676	0.422	-0.017
2LX5	1	0.333	0.271	0.975	-0.642
2N0O	0.515	0.374	0.549	0.464	-0.090
2N58	0.7	0.830	0.058	0.738	0.092

*The Delta RMSD is calculated as C α RMSD of the MJ force field minus the C α RMSD of the GB force field.

All the five cases without known PDB structures are widely taken as test cases in peptide simulation related papers [43, 47, 48, 51, 52, 54, 108-112]. In the GB force field, the H1 has formed a helix like conformation in the most populated cluster, while the C terminal region forms a small loop. The most populated cluster has 82.1% population over the entire cluster (Figure 3.9). For the EK and RIB peptide cases, they have formed helix like conformations in the most populated clusters,

while the residues in the two terminal regions are flexible loops; even the helices formed in the medial region are not standard. The most populated clusters have 64.7% and 89.3% populations over the entire cluster, respectively. The helicity calculated by our method also shows the helix does not form quite well for the EK peptide (Figure 3.9c). The predicted structure of the Fs21 peptide is a beta hairpin like conformation, but it only has 1.2% over the entire cluster. No standard helix conformations were found in other non-most populated clusters of Fs21. The predicted structure contains only a small fraction of helix and the region near two terminals are loop conformations (shown in Figure 3.9). The predicted structure of Agd1 is a helix like conformation in two terminal regions and a turning point in the middle (shown in Figure 3.9). The most populated cluster has 97.9% population over the entire cluster.

For the REMD simulation with the MJ force field, H1 and EK peptides have formed a helix like conformation in the most populated conformation, shown in Figure 3.9. The most populated clusters have 98.7% and 94.5%. The Fs21 has formed a standard helix in the largest conformation which has 100% population over all other clusters. The most populated structures of RIB and Agd1 peptides are helix like conformations as well. From the calculated helicity of each residue in Figure 3.9c by DSSP software, the result demonstrates higher helicity for most residues in those 5 helix cases when simulated with the MJ force field compared to those simulated with the GB force field. The predicted conformations from both

force fields are helix like for the H1 and EK peptides (Figure 3.9 a,b). Significant improvement was achieved for the Fs21 and Agd1 peptides. This indicates that the GB force field lacks helicity propensities in some peptides, while the helicity of each residue can be carefully adjusted by the cMAP correction in our MJ force field.

To summarize, the MJ force field performs better in folding native like helix conformations for the tested peptides.

3.3.1.2 *Ab initio* folding of beta hairpin in training set

There are 7 beta hairpins including B1 B2 B3 Mbh12 Nrf2 2RVD and Chignolin peptides in the training set, shown in Figures 3.7, 3.8. Except 2RVD and Chignolin peptides (sequence identity 80%), the sequence identity between each pair of them is equal to or smaller than 40%. When simulated with the GB force field, the top 3 most populated clusters of B1 peptide all form partial helix conformation and no beta hairpin conformation was observed in all the clusters. The predicted structure of B2 is a helix-loop conformation. The second most populated cluster has formed a beta hairpin like conformation, but has one residue shifted from the standard hairpin. The predicted structure of B3 peptide is a beta hairpin like conformation, which has one or two residue shifted. It has 90.2% population over the entire cluster. The most populated cluster of Mbh12 peptide is a beta hairpin like conformation. Similar to the native one, the terminal region did not form standard

hairpin conformation, and the RMSD value between the native structure and the predicted structure is 0.196 nm. The predicted structures of Nr2f, 2RVD and Chignolin peptides are not beta hairpins. The predicted structure of Nr2f peptide is a helix-turn-helix conformation.

When simulated with the MJ force field, the predicted structure is a standard beta hairpin for B1 and B2 peptides. The predicted structure of B3 is partly beta like conformation, with loops in the middle part, and it has 46.7% population. The predicted structure of Mbh12 peptide is a beta hairpin like conformation, which has an RMSD value of 0.168 nm with the native structure. The most populated cluster has 99.9% population. The predicted structure of Nr2f peptide is a partly beta sheet conformation. A standard beta hairpin conformation was formed in the second populated cluster of Nr2f peptide. The predicted structures of 2RVD and Chignolin peptides are beta hairpins, which have RMSD values of 0.11 nm and 0.23 nm with the native structures, respectively. The most populated clusters have 100% and 99.7% populations, respectively.

All in all, the modified force field performs better than the original force field in forming the beta hairpin conformation in most of the cases except Mbh12 and B3 peptides with similar performances with both force fields. Especially for the B1, B2, 2RVD and Chignolin which cannot form native structure like most populated clusters with the GB force field, they can form well native like hairpin conformations with the MJ force field.

3.3.1.3 *Ab initio* folding of other cases that contains more than one secondary conformation

The performance of the GB force field for the 4 peptides (1vii 2N58 1L2Y and 2JOF) which contain both loops and helix in the native structures is described in Chapter 2.3.3.

When simulated with the MJ force field, the 1vii peptide can form a native like conformation composed of several small helices linked by a short loop, which has an RMSD value of 0.553 nm with the native structure. The most populated cluster has 46% population. For the 2N58 peptide, the predicted structure is a helix conformation with population 70%, which is one of the few cases not consistent with the native structure. The third populated cluster is close to the native conformation with the 4% cluster population, and the RMSD value 0.515 nm. The 1L2Y and 2JOF peptides both form standard trp-cage like conformations in their most populated clusters, which have RMSD values 0.356 nm and 0.355 nm with native structures and have populations 97.5% and 99% over the entire clusters, respectively. When simulated with the modified force field, the predicted structures of 1fme and 1fsd do not successfully formed the beta hairpin parts in the N terminus, and they consisted of 59.6% and 53.2% over the entire clusters, respectively. They both formed native like conformations in the second most populated clusters, which are composed of structures of the helix and beta strand linked by a short turn. They consist of 25.3% and 36.8% over the entire clusters

respectively. Of their second most populated conformations, not only the secondary structure is consistent with the native conformation, the interaction patterns between the beta strand and helix in the 1fsd and 1fme cases are similar as the native one. They have RMSD values 0.172 nm and 0.294 nm with the native structure, respectively. For the WW domain, the predicted structure is a beta sheet with a loop similar to the native structure. Although there is a small fraction of helices (at the sequence PRO, PRO, GLY, TRP, GLU) formed, which is not the same as the native conformation in this region, the overall conformation is still close to the native one, which has 0.745 nm RMSD with the native conformation. The predicted structure of 1E0N peptide is native like, which has 0.432 nm RMSD with the native conformation.

In summary, for the peptides containing more than one secondary structure in the training set, the 1E0N has noticeable improvement using the MJ force field, while 1vii, 1L2Y, 2JOF, 1fme, 1fsd, and WW cases have similar performances with the MJ force field comparing to the GB force field.

3.3.2 *Ab initio* folding of small proteins with tertiary structure in test set

The Figure 3.9a, b shows the superposed conformations of predicted structures and the native ones for the 8 small protein cases with the GB and MJ force fields respectively. When simulated with the GB force field, among the 8 small protein cases, 2 proteins (1PRB and 1LMB) form native like conformations in the most

populated clusters. The two proteins have RMSD values of 0.41 and 0.36 nm with their corresponding native structures, respectively (Table 3.4). The predicted structures of 1ENH and 1MI0 proteins are also close to the native, while the predicted structure of 2A3D 2HBA 2P6J and 2WXC deviate from the native conformations significantly.

Figure 3.10 The predicted structures by the GB simulation for the 8 small protein cases with tertiary structures are superposed with the native PDB structures (the most populated cluster in Red, and the native structure in Blue).

Table 3.4 The *ab initio* folding of the 8 small proteins in the test set with the MJ force field in comparison to the *ab initio* folding by the AMBER14 GB force field. The RMSD values are between predicted structures related to the native ones.

Name	Percentage of most populated cluster (modified force field)	C α RMSD (modified force field) (nm)	Percentage of most populated cluster (original force field)	C α RMSD (original force field) (nm)	Delta RMSD (nm)*
1ENH	0.46	0.646	0.088	1.152	-0.506
1LMB	0.279	1.137	0.206	1.159	-0.022
1MI0	0.248	0.565	0.323	1.213	-0.648
1PRB	0.263	1.137	0.48	0.413	0.724
2A3D	0.808	0.4	0.153	0.423	-0.023
2HBA	0.328	0.819	0.303	0.454	0.365
2P6J	0.553	0.285	0.173	0.997	-0.712
2WXC	0.706	0.604	0.025	1.031	-0.427

*The Delta RMSD is calculated by C α RMSD of the MJ force field minus the C α RMSD of the GB force field.

Similarly, when simulated with the modified force field, the predicted structures of the 2P6J and 2A3D proteins are native like. The predicted structures of 1LMB, 2WXC, 1ENH, and 1MI0 proteins are close to the native. Among these proteins, the 1MI0 is a combination of beta sheet and helix. The predicted structures of 2HBA and 1PRB proteins are not native like in terms of RMSD values with the native structures. However, the second most populated cluster conformations are close to the native structure for the 1PRB.

3.3.3 *Ab initio* folding of other 8 test cases with various sizes and secondary structure components

Similar to the previous simulations, I choose another 8 test peptides which have NMR native conformations for the further validation of the robustness of this force field. The superimposition of the simulated conformations with the native ones is shown in the Figure 3.11.

Figure 3.11 The figure shows the predicted structures by the GB force field for the 8 peptide test cases that are superposed with the native PDB structures (predicted structures in red, and the native structures in blue).

Table 3.5 The *ab initio* folding of 8 peptides in a test set with the MJ force field in comparison to the *ab initio* folding by the AMBER14 GB force field. The RMSD values are calculated for the most populated clusters related to the native conformations.

Name	Percentage of most populated cluster (modified force field)	C α RMSD (modified force field) (nm)	Percentage of most populated cluster (original force field) (nm)	C α RMSD (original force field) (nm)	Delta RMSD (nm)*
1F8P	0.696	0.651	0.067	1.161	-0.51
1J5B	1	0.241	0.014	1.283	-1.042
1LE3	0.997	0.228	0.656	0.8	-0.572
1LOI	0.825	0.517	0.155	0.471	0.046
1MBJ	0.288	1.037	0.348	1.089	-0.052
2N31	0.998	0.348	0.131	0.404	-0.056
2N5L	0.372	1.08	0.536	0.605	0.475
2N88	0.225	1.061	0.136	1.254	-0.193

*The Delta RMSD is calculated by C α RMSD of the MJ force field minus the C α RMSD of the GB force field.

It shows all the small cases which have residue numbers around or below 30, can form very well native like conformations.

The 1J5B and 2P6J peptides form standard helix or beta hairpin conformations, respectively, which have RMSD values 0.241 nm and 0.228 nm with the native structures, respectively (Table 3.5). The 1F8P, 1LOI, and 2N31 peptides are more challenging, since there are loops or turning regions. However, these three peptides have achieved accurate predictions. The predicted structure of 1F8P has a helix region overlapping well with the native, and the N-terminal loop region has been correctly predicted with the RMSD value 0.651 nm with the native conformation. The deviation is mostly caused by the flexible loop region, which can be expected. The predicted structure 1LOI has a correct turning loop in the middle as the native structure, which has the RMSD value 0.517 nm with the native one. The predicted structure of 2N31 has correct predicted the beta hairpin region as well the loop region. It has the RMSD value 0.348 with the native conformation.

For the complex cases as 2N5L, 2N88, and 1MBJ, which are relatively large and have complicated combinations of beta sheet, helix, as well as loop, making it more challenging for the current MJ force field. The failure of folding native like conformations may be caused by the difficulty of convergence in *ab initio* folding simulations of larger cases, or by that our MJ force field still needs further improvement. For instance, most failed cases contained similar residue types, *e.g.* GLY, LYS, ARG, or TYR.

3.3.4 The free energy landscapes (FEL) of test small proteins

Figures 3.11, 3.12 shows the free energy landscape of the test peptides with the GB and MJ force fields, respectively. It is obtained by projecting onto the two reaction coordinates of C α -RMSD and radius gyration.

Figure 3.12 The free energy surfaces as a function of the all atom radius of gyration (RG) and the C alpha RMSD in the simulation cases with the MJ force field. The black arrows point at the native RG value.

The modified force field can produce deep native state basin in the free energy landscape, while still being able to sample large areas as the GB force field. For instance, the free energy surfaces of 1fme and 1fsd have low energy basin regions close to native (RMSD 0.2~0.5nm and radius gyration 0.9~1nm) in the simulations with the MJ force field, while most of the low free energy regions have a RMSD value beyond 0.5 nm in the simulation with the GB force field. The free energy surface analysis is quite consistent with previous cluster analysis. For 1L2Y, 2JOF, and WW domain proteins the two force fields have similar performance in terms of sampling. Nevertheless, for some other cases such as B1, B2, B3, 1vii, 2RVD, Chingnolin, Mbh12, 1E0N, 2LD0, and 2LX5 peptides, the MJ force field samples smaller RMSD regions with more compact conformations, corresponding to the native structures.

The B1, B2, and B3 peptides have more than two major basins, and the one close to native is the major basin in the modified force field simulation. The energy barrier between the two basins is 5~20 kJ/mol. It indicates the modified force field still has the ability to efficiently sample various kinds of conformations, with the native conformation corresponding to the lowest free energy region in most of the cases.

Figure 3.13 The free energy surfaces as the function of the all atom radius of gyration (RG) and the C alpha RMSD of the 19 small peptide cases from the REMD simulations with the GB force field. The black arrows point at the native RG value.

Figure 3.14a, b shows the free energy landscapes of the small proteins. The MJ force field has explored similar conformational space with the GB one for the 8 proteins. However, except for 1PRB and 2HBA proteins, the energy basins are deep and close to native for most other proteins.

Figure 3.14 Panel a,b shows the free energy surfaces as the function of the all atom radius gyration (RG) and the C alpha RMSD in the small protein cases simulated with the GB and the MJ force fields, respectively. The black arrows point at the native RG value.

The Figure 3.15a, b shows the free energy landscapes of the 8 peptides of the test set. The simulations with the MJ force field have explored similar conformational space with the GB ones for the 8 peptides. However, except for 2N5L, 2N88 and 1MBJ peptides, the energy basins of most other peptides are deep and close to native.

Figure 3.15 Panel a,b shows the free energy surfaces as the function of the all atom radius gyration (RG) and the C alpha RMSD in the 8 peptides of the test set simulated with the GB and the MJ force field respectively.

3.3.5 Future improvement from the clues of failed cases

The simulation results of cases such as 1BYO, 1MBJ, 2N58, 1PRB, and 2HBA show that the force field is still limited in some situations. Some of the failed cases may provide clues for future improvement. One most challenging part was often related to the region composed of several charged residues. For instance, the region LYS-LYS-LYS in the middle part of 1BYO is a loop instead of the native like alpha helix. And the GLU-GLU-GLU-ASP region also failed to form the helix conformation for the 1MBJ. The 1PRB protein also has difficulty to form the helix in the LYS-ALA-LYS region. These 3 cases show clear shortage of charged residue related parameters, which is an expected limitation of ACE. Since the ACE is derived from monomer proteins, which statistically underestimates charge-charge residue interactions as the original paper has discussed. This can be improved in future by incorporating some interface ACE parameters. Several reasons may be responsible for the failure of folding native like conformations for 2N58 and 2HBA proteins with the MJ force field. For instance, the loop propensities and the helicities of some specific residues are much harder to solve. However, our method is a useful approximation and improvement compared with other explicit solvent approaches. It is a tradeoff of the transferability and accuracy of various kinds of peptides and small proteins.

3.4 Conclusions

In contrast to the traditional continuum solvent models such as the PB and GB models, the ACE combined with distance dependent dielectric constant and the cMAP based bonded interaction can keep the secondary structure of the native protein with a simple force field functional form. As the ACE already implicitly contains the solvent effect, no extra computational time will be spent to mimic the solvent effect. GB was claimed to have higher speed due to reduced degrees of freedom from explicit water. However, in order to accurately calculate the GB energy values, relative complex force field terms are implemented, and GB also requires a large non-bonded interaction cutoff. The consequence of the above issues is usually slow simulation speed, especially for large proteins. With much higher simulation speed and comparable accuracy, our ACE based force field would have potential in a wide range of applications, including protein stability tests, protein-protein interactions, and sampling conformational changes of large protein structures. Similar to the traditional continuum solvent models, in our method, there is no need for equilibrating the water around the solute as the explicit solvent was eliminated, which saves a large amount of time. Furthermore, the protein will explore the conformations more efficiently for the reason of the absence of viscosity of water. The training set and testing set both show high accuracy of the force field to *ab initio* fold peptides or small proteins. Good balance between different secondary structures was achieved. At the same time, the

electrostatic interaction, hydrogen bond and hydrophobic interaction are well represented in most test cases.

Compared to GB simulations in terms of RMSD difference for the 19 peptides in the training set, 73.7% (14 in 19) cases have improved performance. For another 5 peptides with no native PDB structures, all have improved performance in terms of helicity. For the 8 peptides and 8 small fast folding small proteins in the test set, 75% (6 in 8) cases are improved for both test sets in term of RMSD comparing to the GB simulations. All these show the new force field has an advantage over the GB approach. Comparing to the docking method with the ACE as desolvation term, using MD simulations to explore the protein-protein interaction can fully consider the flexibility of the protein. As we known, the protein-protein interaction has an inducing effect, so the interface between the interacting proteins may have large conformational changes. The current rigid docking methods can predict protein-protein interaction accurately when the proteins are from the experimental crystallographic or NMR complexes, while once the two proteins are from individual experiments or by homology modelling, much less accuracy can be achieved. Compared with the coarse-grained models, the all atom models should be more accurate due to the consideration of atomistic details and hydrogen bonds. Certain strong and specific binding complexes, *e.g.* the one sensitive to single residue mutation are hard to predict by the coarse grained models. The reason was attributed to the loss of details of atomic interaction as described by Hummer's

paper [150]. Unlike in the traditional GB model, our force field has introduced ions to neutralize the system. This can prevent unphysical conformations around highly charged residues by reducing the net charge.

However, our method still needs improvement in many aspects. The protein database for generating ACE parameters was all monomeric proteins. This indicates the ACE may overlook some interface interactions between proteins. This may influence the polar and charge residue interactions among the interfaces. It is also a challenge to select the most suitable parameters for non-bonded terms, dielectric constant, as well as cMAP.

In future, I will focus on applying this force field or its modification to systems with domain-domain interactions and flexible linkers, as well as protein-protein interactions, *e.g.* antibody-antigen. A 100 ns MD simulation for the protein 2KC0 (with domain-domain interactions and a linker, containing 242 residues) by the MJ force field was carried with same simulation conditions as in previous MJ simulation cases. The clustering method and parameters used were also same as in previous MJ simulation analysis. A preliminary result shows that the representative structure of the most populated cluster is close to the native conformation (with a RMSD value 0.462 nm). The most populated cluster has a population of 23.2%.

All in all, our method provides an alternative method for exploring protein dynamics and protein-protein interaction with faster simulation speed and relatively accurate structures. Since it is incorporated in the GOMACS package, the

usage is straightforward and user-friendly. Furthermore, more parameters may be induced, such as interface ACE and ligand protein ACE, enabling wider applications of this method in molecular dynamics research or drug design.

4 Investigating the stability of DENV-2 envelope protein dimer using well-tempered metadynamics simulations

4.1 Background

The dengue virus which is spread by the mosquito is a growing health threat to the people near equatorial regions [151]. There are mainly four serotypes of dengue virus, and sequential infection by several serotypes can cause a more serious syndrome, which is described as antibody-dependent enhancement (ADE) [152]. Currently, a very limited number of effective drugs or vaccines can combat the dengue viral infection yet, and only a vaccine of dengue called Dengvaxia was licensed in some countries [153, 154]. The virus contains 3 structural proteins and 7 nonstructural proteins. The E protein is a constituent protein on the virus outer surface, which is an ideal target for antibody to bind. The E protein undergoes large conformational changes prior to the fusion process, which is initiated by low pH in the endosome and transfer from dimer to trimer[89]. Hence the effective antibody is designed to block the large conformational change of E protein complex by stabilizing the dimer complex. Because of the ADE effect, the ideal antibody must neutralize the four serotypes simultaneously. According to recent research, the most effective epitopes may be on the interface of two monomer E proteins, called E-Dimer-dependent epitope (EDE). EDE is divided into two classes, EDE1 which is regardless of the glycosylation state of position N153, and EDE2 which depends

on glycosylation on position N153 for tight binding. Since the EDE epitope is a conserved region among the four serotypes, the antibody that targets this region is promising to bind the four serotypes. Interestingly, the epitope exists and is conserved in the E protein of Zika virus as well [155], which indicates the research on this epitope also help in developing antibodies that targets Zika.

Currently, most research on effects of dengue E protein dimer with or without antibodies is based on experimental methods. It is relatively expensive to explore the binding with atomistic details in large scale molecular simulations. In recent years, the accuracy as well as efficacy of MD simulation has gradually increased. Hence, molecular dynamics simulation is promising in discovering the atomic details of the protein-protein interactions. As the computational power increasing, some researchers have applied the classical molecular dynamics simulation in such large systems [156, 157], though the simulations would be computationally expensive and the systems may remain trapped in local minima. Convergence usually requires timescales of microseconds for such large complexes; otherwise, the sampling is often insufficient. However, the metadynamics method developed in recent years helps to accelerate sampling speed by adding Gaussian heights to the low energy regions. The free energy landscape can be recovered by counting the negative of overall Gaussian sands added [78]. Later, developed well-tempered Metadynamics methods further improved the sampling efficiently by adjusting the Gaussian heights automatically during the simulation [82]. The metadynamics

simulation is suitable to research the energy landscape of large conformational changes, *e.g.* it was used to explore the monomeric dengue E protein conformational transition with a pH change [158].

Previous researches have implemented short MD simulations combined with an MMGB model to explore the pH influence on the dimer stability [159]. In our research, we first use well-tempered metadynamics to explore the conformational change and free energy landscape of four situations: dengue E protein dimer with or without antibody at neutral or acid pH conditions. The simulation time we used is much longer than previous research, and the sampling technique is much more efficient. Mutational experimental data are available for the antibody E protein binding, *e.g.* W101A or N153A mutation will cause the loss of more than 95% binding [95]. We have carried out two extra simulations of mutated E proteins with antibody to validate the accuracy of these metadynamics settings. The result indicates the potential of this method in researching the prevention of conformational changes by antibody binding. The method would be promising in pharmaceutical development. Furthermore, free energy of protein binding can be influenced by factors such as pH or antibody binding. In such cases, it cannot be easily explored by other computational techniques such as docking, which is designed to discover binding affinity between two partners only, and is limited by scoring function approximations.

4.2 Method details

4.2.1 PDB structures collection and topology preparation

The PDB structure was retrieved from PDB database [160] with PDBID 4UT6 [161]. The structure of 4UT6 is an E protein dimer of dengue serotype 2 with antibody EDE2 B7 binding at the EDE2 epitope. To get the initial structure of E protein dimer simulation without the antibody, I just removed the antibody part in the 4UT6 PDB structure. The acid and neutral states were mimicked by changing the protonated state of the histidine residue. Since the system was at about PH 5, only the HIS protonated state should change [162]. Tleap from AMBERtool package was used to generate the topology. All the n-linked glycan structures were included in the PDB file. The Glycan residues include NAG (N-acetyl-D-glucosamine), BMA (beta-D-mannose), MAN (alpha-D-mannose) and FUC (fucose) The force field parameters for the glycans are from GLYCAM 06j [163], which are already available in Ambertools [115]. The linked bond between ASN and glycan or between glycan and glycan can be accomplished by writing and executing a tleap script. The generated topology and coordinate files are in AMBER format with AMBER 14 force field (the glycan parameters and topologies adopt GLYCAM 06j force field as mentioned above). It is later converted into GROMACS format by the glycam2gmx.pl script written by Wehle et al. [164], which properly treated the improper dihedrals as well as ensuring compatibility

with GLYCAM. The 2 mutation systems (W101A and N153A) are taken as negative control, that will decrease the binding affinity after mutation according to a literature report [1]. The mutations are done in each monomer in the E protein dimer with the swapaa command in chimera software [2].

4.2.2 Molecular simulation procedure

The simulation was performed with Plumed [3] incorporated within the GROMACS 4.6 software package [4] using AMBER14 force field. The system was solvated in a dodecahedron box of TIP3P water, and the box size is defined by keeping the closest protein atom to the box boundary as 1 nm. The PME method was implemented to calculate electrostatic interactions. The periodic boundary conditions were applied in xyz dimensions. A 1.4 nm cutoff was used to evaluate the non-bonded interactions. The covalent bonds connecting hydrogen atoms were restrained by the LINCS algorithm. Energy minimization was used to refine the initial structure. A canonical velocity-rescaling thermostat[165] was used to control temperature. The Parrinello-Rahman Barostat[166] was used in NPT simulation to control pressure. Then a 1 ns NVT ensemble simulation was carried out to equilibrate the water box. A 10 ns NPT ensemble simulation was used to further equilibrate the system. Another 100 ns NPT ensemble of normal MD was used for the production simulation. Then the metadynamics simulation was applied by Plumed incorporated into the GROMACS package with AMBER14 force field

to continue exploration of the free energy landscape. Two collective variables (CVs) relevant to binding were used: one is the C α distance RMSD of the E protein dimer, and the other is the interface coordination number of C α of the E protein dimer. The distance RMSD estimates the similarity of a structure with respect to a reference structure by calculating the distances between all the pairs of atoms. It is computationally efficient compared to the normal RMSD which requires first superimposing two structures by translations and rotations. And the method also overcomes the problem caused by a bad alignment. The function for the distance RMSD is as follows from the Plumed software manual:

$$d(X^A, X^B) = \sqrt{\frac{1}{N(N-1)} \sum_{i \neq j} [d(x_i^a, x_j^a)] - d(x_i^b, x_j^b)]^2}$$

where N is the number of protein atoms, $d(x_i, x_j)$ is the distance between atoms i and j. The coordination number is defined as follows by Plumed:

$$Coordination\ number = \sum_{i \in A} \sum_{j \in B} S_{ij} ,$$

where $S_{ij} = \frac{1 - (\frac{r_{ij} - d_0}{r_0})^n}{1 - (\frac{r_{ij} - d_0}{r_0})^m}$, and n is 6, m is 12 and r_0 is 0.5 in the simulation.

The extent of contacts between two groups of atoms can be estimated by the function, and it is differentiable [167]. The interface coordination number calculation is only limited to interface residues within a cutoff 1 nm, which is calculated with Pymol's script InterfaceResidues.py [168]. 80~150 ns well-

tempered metadynamics simulation was carried out following the NPT ensemble equilibration. The simulation time used is dependent on whether the system reaches convergence. During the well-tempered metadynamic simulation, Gaussian values were deposited every 1 ps with a starting height of 1.5 kJ/mol and gradually decreased by a bias factor 15. The width of the Gaussians was 8 for the coordination number and 0.02 nm for the distance RMSD. We determined the convergence and termination of simulation when the height of the deposited Gaussian in the space of the employed CVs decreased close to zero [5].

4.2.3 Analyzing the convergence, free energy landscape and atomistic details of the interaction

We analyzed the free energy landscape for all 7 simulation sets. The free energy along the 2 CVs was generated by the Plumed program [169]. The free energy landscape was later plotted by Gnuplot program [6]. The conformation falling in the region of interest in the CVs space was extracted, using GROMACS tool `g_cluster`, to perform the clustering with a cutoff 0.2nm by the linkage method. The snapshot picture of the representative conformation of the most populated cluster was generated by the VMD program [7].

4.3 FEL for 4 different states of dengue's E protein

The convergence of well-tempered metadynamics simulation is checked by observing whether the Gaussian height is close to zero at the end of simulation according to previous researchers' practice [170, 171]. It shows all the simulations reach convergence within 100 ns (Figure 4.1).

Figure 4.1 a,b,c,d shows the evolution of Gaussian height over time in the well-tempered metadynamics simulations of wild type dengue E protein dimer with or without antibody at neutral or acid conditions respectively. The e, f shows the time evolution of Gaussian height in the well-tempered metadynamics simulation of two mutated dengue E protein dimers (W101A and N153N), respectively. In all the simulations, the Gaussian heights were decreased to near zero within the 100 ns simulation, which indicates convergence of the simulation.

The free energy landscape of the E protein dimer system in neutral condition against two CVs, CA RMSD and interface C α coordination number, is shown in Figure 4.2a. Free energy minima correspond to the native like conformation in the simulation trajectory. The distance RMSD is relatively low and the coordination number is close to the initial structure value. The snapshot for the low energy basin shows the complexes remain in a good contact in the interface of dimer.

The free energy landscape of the E protein dimer system in acidic conditions against two CVs, CA RMSD and interface C α coordination number, is shown in Figure 4.2b. The free energy minima correspond to a region that has large distance RMSD values and small values of Coordination Number. It indicates the dimer is destabilized by the low pH. The conformations with free energy minima correspond to the separated E proteins in the simulation trajectory (the snapshot was shown in Figure 4.1b).

The free energy landscape of the system of E protein dimer in neutral condition binding with antibody EDE2 B7 against two CVs, CA RMSD and interface C α coordination number, is shown in Figure 4.2c. The lowest energy basin is highly focused on the region which corresponds to the native like conformation. This is consistent with our expectation. Since the E protein is relatively stable in neutral conditions, with an extra stabilization of antibody binding, the energy basin becomes deep.

The free energy landscape of E protein dimer in acidic condition bound with antibody EDE2 B7 against two CVs, CA RMSD and interface C α coordination number, is shown in Figure 4.2d. The energy basin is around the native conformation region in the free energy surface. It indicates the antibody has effectively stabilized the dimer complex in the low pH condition. The representative conformation from the low energy region also shows the contact interface in the two parts of dimer is well preserved. It is consistent with the experimental conclusion that the tight binding of EDE2 B7 antibody on the E protein EDE region of E protein can reverse the destabilization of the E protein dimer in low pH.

Figure 4.2 a,b The FES of the binding between the monomers of the dimer in **neutral and acidic conditions respectively**. It is computed by the metadynamics simulation through the projection on the two CVs, CA RMSD and interface C α coordination number. The snapshots are the representative conformations by

clustering all the conformations that fall in the box region, which is marked with black edge box, with the ranges ($10 < x < 25$, $0.08 < y < 0.12$) and ($5 < x < 20$, $0.1 < y < 0.6$) for Figure 4.2a and Figure 4.2b, respectively. The Figure 4.2c,d show the free energy landscape of the dimer in neutral and acidic conditions with antibody binding. The snapshot is the representative conformation from the region marked by the black edge box, with the ranges ($15 < x < 35$, $0.08 < y < 0.12$) and ($10 < x < 30$, $0.08 < y < 0.12$) for Figure 4.2c and Figure 4.2d respectively. The free energy units are 1 kJ/mol, and the unit for RMSD is nm.

4.4 FEL for 2 mutations of E protein with antibody

Figure 4.3a and Figure 4.2d revealed that the energy landscape is much dispersed in the W101A mutation compared to the wild-type.

The lowest energy region is close to an unstable conformation region in the mutant E protein. These results suggest that the antibody cannot hold the lowest free energy basin around the native conformation region in acidic conditions. This is consistent with the experimental result, that the W101A mutation will reduce the EDE2 B7 binding $\leq 95\%$. A previous report stated that the W101 residue was critical in stabilizing the E-dimer contact by the interaction with the K310 in the domain III [161] .

We found that mutation of N153A will lead to abort N-linked glycosylation branch in ASN153 of E protein. And the EDE2 epitope depends on the N-linked glycosylation at position N153 for the binding [161]. Figure4.3b shows the free energy basin is significantly shifted to the region corresponding to the non-native like conformation. This suggests that the N153A mutation has significantly reduced the antibody's stabilizing effect on the dimer. It is consistent with the experimental result, that the N153A mutation will decrease more than 95% EDE2 B7 antibody binding.

Figure 4.3 a,b shows the FES of the binding between monomers of the dimer in acid condition with antibody for W101A and N153A mutations, respectively.

It is computed by the metadynamics simulation through projection on the two CV, C α RMSD and interface C α a coordination number. The Figure 4.3c is the representative conformation of the most populated cluster. We cluster all the

conformations fall in the box region, which is marked by the black edge box, with the ranges ($5 < x < 20$, $0.1 < y < 0.2$) and ($0 < x < 5$, $0.7 < y < 1$) for Figure 4.3a and Figure 4.3b, respectively. The Figure 4.3d is the representative conformation from the region marked by the black edge box, with the ranges ($5 < x < 15$, $0.2 < y < 0.4$) and ($0 < x < 10$, $0.5 < y < 0.7$). The free energy units are 1 kJ/mol, the unit for RMSD is nm.

4.5 Conclusion

Our simulations reveal that the E protein dimer is more stable in neutral conditions compared with the one in acidic condition. This is consistent with experimental research and the hypothesis about the process of virus fusion. This indicates that the molecular dynamics simulations can detect pH influence on dimer stability at an atomic level. However, it requires a long time simulation to reach convergence with an accurate force field and an appropriate sampling technique. The stabilization of the dimer by the EDE2 B7 antibody in acidic conditions is also confirmed by simulation. This is consistent with the experimental observation that antibody can prevent the virus fusion process.

The W101A or N153A mutation in E protein leads to significant reduction (95%) in dimer stability with EDE2 B7 antibody, which was observed by simulation. This indicates the well-tempered metadynamics simulation can be used to screen antibodies with specific binding to antigens. However, with currently available

computational facilities, any large scale screen of a large and complex antibody is still challenging.

In conclusion, we can identify the stability of the E protein dimer with different conditions such as pH change, antibody binding, or mutation by well-tempered metadynamics simulation. As computational power increases continuously and atomic force fields improve gradually in accuracy and efficiency, MD simulation is promising to study large systems associated with complicated biological processes in future.

5 Concluding remarks

MD simulation is appealing for investigating dynamic properties and processes of biomolecules. However, classical MD simulations with explicit solvent require large computational resources, in order to sample various states that are separated by high energy barriers. Implicit solvent model is one straightforward method to alleviate such a computational burden. Employing an advanced sampling technique in MD simulation is another widely used way to overcome the high energy barriers between energy minima, and hence reduce total simulation time.

In my first project, I improved the accuracy of an implicit solvent GB model (ff14SBonlysc with the combination of GB-Neck2). It demonstrated that the accuracy of the current GB models still has room for improvement. In the second project, I focused on developing a new solvent free force field with high efficiency and reasonable accuracy. It indicates the possibility of incorporating experimental information, *e.g.* statistics based non-bonded interaction, or phi-psi energy correction map derived from experimental structures, to a solvent free force field. According to recent reports, the metadynamics-based approach can significantly accelerate sampling efficiency in MD simulation, and it is suitable for investigating drug target associations, *e.g.* drug-GPCR Binding [172, 173]. It shows growing importance in drug discovery. So, in the third project, I used well-tempered metadynamics to investigate the E protein dimer stability in various conditions. The simulation results are consistent with the experimental findings.

In future works, we will apply our implicit solvent methods in some practically important applications, such as antibody-antigen associations and domain-domain interactions with linkers. We also plan to employ the well-tempered metadynamics method to find possible drugs (*e.g.* antibodies, peptides, or small molecules) that target the EDE epitope region and prevent E protein dimer conformational changes in low pH conditions.

6 REFERENCES

1. Karplus, M. and J.A. McCammon, *Molecular dynamics simulations of biomolecules*. Nat Struct Biol, 2002. **9**(9): p. 646-52.
2. Molnar, S.P., *The art of molecular dynamics simulation* (vol 76, pg 171, 1999). Journal of Chemical Education, 1999. **76**(4): p. 471-471.
3. Fermi E., P.J., Ulam S., in *Los Alamos report LA-1940* 1955.
4. McCammon, J.A., B.R. Gelin, and M. Karplus, *Dynamics of folded proteins*. Nature, 1977. **267**(5612): p. 585-90.
5. Warshel, A. and M. Levitt, *Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme*. J Mol Biol, 1976. **103**(2): p. 227-49.
6. Lindahl, E. and M.S.P. Sansom, *Membrane proteins: molecular dynamics simulations*. Current Opinion in Structural Biology, 2008. **18**(4): p. 425-431.
7. Freddolino, P.L., et al., *Molecular dynamics simulations of the complete satellite tobacco mosaic virus*. Structure, 2006. **14**(3): p. 437-449.
8. Roccatano, D., A. Barthel, and M. Zacharias, *Structural flexibility of the nucleosome core particle at atomic resolution studied by molecular dynamics simulation*. Biopolymers, 2007. **85**(5-6): p. 407-421.
9. Sun, R., Z.L. Li, and T.C. Bishop, *Molecular Dynamics Simulation Workflow for Nucleosome Studies*. Biophysical Journal, 2017. **112**(3): p. 289a-289a.
10. Haile, J.M., *Molecular-Dynamics Simulations of Simple Fluids with 3-Body Interactions Included*. Abstracts of Papers of the American Chemical Society, 1978. **175**(Mar): p. 11-11.
11. MacKerell, A.D., et al., *All-atom empirical potential for molecular modeling and dynamics studies of proteins*. J Phys Chem B, 1998. **102**(18): p. 3586-616.
12. Zhang, Y., et al., *Conformational changes of the flavivirus E glycoprotein*. Structure, 2004. **12**(9): p. 1607-1618.
13. Decherchi, S., et al., *Implicit solvent methods for free energy estimation*. European Journal of Medicinal Chemistry, 2015. **91**: p. 27-42.
14. Lee, M.S. and M. Olson, *New hybrid explicit/implicit solvent method for biomolecular simulations*. Abstracts of Papers of the American Chemical Society, 2004. **228**: p. U507-U507.
15. Lee, M.S. and M. Feig, *New trends in generalized born implicit solvent methods*. Abstracts of Papers of the American Chemical Society, 2003. **225**: p. U791-U791.
16. Ouyang, J.F. and R.P. Bettens, *Modelling Water: A Lifetime Enigma*. Chimia (Aarau), 2015. **69**(3): p. 104-11.

17. Guillot, B., *A reappraisal of what we have learnt during three decades of computer simulations on water*. Journal of Molecular Liquids, 2002. **101**(1-3): p. 219-260.
18. Brooks, B.R., et al., *CHARMM: the biomolecular simulation program*. J Comput Chem, 2009. **30**(10): p. 1545-614.
19. Salomon-Ferrer, R., D.A. Case, and R.C. Walker, *An overview of the Amber biomolecular simulation package*. Wiley Interdisciplinary Reviews-Computational Molecular Science, 2013. **3**(2): p. 198-210.
20. Van der Spoel, D., et al., *GROMACS: Fast, flexible, and free*. Journal of Computational Chemistry, 2005. **26**(16): p. 1701-1718.
21. Foloppe, N. and A.D. MacKerell, *All-atom empirical force field for nucleic acids: I. Parameter optimization based on small molecule and condensed phase macromolecular target data*. Journal of Computational Chemistry, 2000. **21**(2): p. 86-104.
22. Gong, Z., Y.J. Zhao, and Y. Xiao, *RNA Stability Under Different Combinations of Amber Force Fields and Solvation Models*. Journal of Biomolecular Structure & Dynamics, 2010. **28**(3): p. 431-441.
23. Schuler, L.D., X. Daura, and W.F. Van Gunsteren, *An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase*. Journal of Computational Chemistry, 2001. **22**(11): p. 1205-1218.
24. Cao, L.R., et al., *Recent Developments in Using Molecular Dynamics Simulation Techniques to Study Biomolecules*. Acta Physico-Chimica Sinica, 2017. **33**(7): p. 1354-1365.
25. Cramer, C.J. and D.G. Truhlar, *Implicit Solvation Models: Equilibria, Structure, Spectra, and Dynamics*. Chem Rev, 1999. **99**(8): p. 2161-2200.
26. Weiser, J., P.S. Shenkin, and W.C. Still, *Approximate atomic surfaces from linear combinations of pairwise overlaps (LCPO)*. Journal of Computational Chemistry, 1999. **20**(2): p. 217-230.
27. Nguyen, H., D.R. Roe, and C. Simmerling, *Improved Generalized Born Solvent Model Parameters for Protein Simulations*. Journal of Chemical Theory and Computation, 2013. **9**(4): p. 2020-2034.
28. 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.
29. Feig, M. and C.L. Brooks, 3rd, *Recent advances in the development and application of implicit solvent models in biomolecule simulations*. Curr Opin Struct Biol, 2004. **14**(2): p. 217-24.
30. Hou, T., et al., *Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations*. J Chem Inf Model, 2011. **51**(1): p. 69-82.

31. Onufriev, A., D.A. Case, and D. Bashford, *Effective Born radii in the generalized Born approximation: The importance of being perfect*. Journal of Computational Chemistry, 2002. **23**(14): p. 1297-1304.
32. Tsui, V. and D.A. Case, *Theory and applications of the generalized Born solvation model in macromolecular Simulations*. Biopolymers, 2001. **56**(4): p. 275-291.
33. Sumner, W.Q., *On the Variation of Vacuum Permittivity in Friedmann Universes*. Astrophysical Journal, 1994. **429**(2): p. 491-498.
34. Maffucci, I. and A. Contini, *An Updated Test of AMBER Force Fields and Implicit Solvent Models in Predicting the Secondary Structure of Helical, beta-Hairpin, and Intrinsically Disordered Peptides*. J Chem Theory Comput, 2016. **12**(2): p. 714-27.
35. Jorgensen, W.L. and J. Tirado-Rives, *The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin*. J Am Chem Soc, 1988. **110**(6): p. 1657-66.
36. Scott, W.R.P., et al., *The GROMOS biomolecular simulation program package*. Journal of Physical Chemistry A, 1999. **103**(19): p. 3596-3607.
37. D.A. Case, D.S.C., T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, D. Greene, N. Homeyer, S. Izadi, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D. Mermelstein, K.M. Merz, G. Monard, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, C.L. Simmerling, W.M. Botello-Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, L. Xiao, D.M. York and P.A. Kollman, *AMBER 2018*. University of California, San Francisco., 2018.
38. Brooks, B.R., et al., *CHARMM: The Biomolecular Simulation Program*. Journal of Computational Chemistry, 2009. **30**(10): p. 1545-1614.
39. Zhou, R.H., *Free energy landscape of protein folding in water: Explicit vs. implicit solvent*. Abstracts of Papers of the American Chemical Society, 2004. **227**: p. U336-U336.
40. Lei, H. and Y. Duan, *Improved sampling methods for molecular simulation*. Curr Opin Struct Biol, 2007. **17**(2): p. 187-91.
41. Lindorff-Larsen, K., et al., *How fast-folding proteins fold*. Science, 2011. **334**(6055): p. 517-20.
42. Piana, S., K. Lindorff-Larsen, and D.E. Shaw, *How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization?* Biophysical Journal, 2011. **100**(9): p. L47-L49.
43. Jiang, F. and Y.D. Wu, *Folding of Fourteen Small Proteins with a Residue-Specific Force Field and Replica-Exchange Molecular Dynamics*. Journal of the American Chemical Society, 2014. **136**(27): p. 9536-9539.

44. Jiang, F., C.Y. Zhou, and Y.D. Wu, *Residue-Specific Force Field Based on the Protein Coil Library. RSFF1: Modification of OPLS-AA/L*. Journal of Physical Chemistry B, 2014. **118**(25): p. 6983-6998.
45. Li, S.X. and A.H. Elcock, *Residue-Specific Force Field (RSFF2) Improves the Modeling of Conformational Behavior of Peptides and Proteins*. Journal of Physical Chemistry Letters, 2015. **6**(11): p. 2127-2133.
46. Yang, C., E. Kim, and Y. Pak, *Probing alpha/beta Balances in Modified Amber Force Fields from a Molecular Dynamics Study on a beta beta alpha Model Protein (1FSD)*. Bulletin of the Korean Chemical Society, 2014. **35**(6): p. 1713-1719.
47. Zeng, J., F. Jiang, and Y.D. Wu, *Folding Simulations of an alpha-Helical Hairpin Motif alpha t alpha with Residue-Specific Force Fields*. Journal of Physical Chemistry B, 2016. **120**(1): p. 33-41.
48. Zhou, C.Y., F. Jiang, and Y.D. Wu, *Folding Thermodynamics and Mechanism of Five Trp-Cage Variants from Replica-Exchange MD Simulations with RSFF2 Force Field*. Journal of Chemical Theory and Computation, 2015. **11**(11): p. 5473-5480.
49. Zhou, C.Y., F. Jiang, and Y.D. Wu, *Residue-Specific Force Field Based on Protein Coil Library. RSFF2: Modification of AMBER ff99SB*. Journal of Physical Chemistry B, 2015. **119**(3): p. 1035-1047.
50. Buck, M., et al., *Importance of the CMAP correction to the CHARMM22 protein force field: Dynamics of hen lysozyme*. Biophysical Journal, 2006. **90**(4): p. L36-L38.
51. Maffucci, I. and A. Contini, *An Updated Test of AMBER Force Fields and Implicit Solvent Models in Predicting the Secondary Structure of Helical, beta-Hairpin, and Intrinsically Disordered Peptides*. Journal of Chemical Theory and Computation, 2016. **12**(2): p. 714-727.
52. Shell, M.S., R. Ritterson, and K.A. Dill, *A test on peptide stability of AMBER force fields with implicit solvation*. Journal of Physical Chemistry B, 2008. **112**(22): p. 6878-6886.
53. Lwin, T.Z., R.H. Zhou, and R. Luo, *Is Poisson-Boltzmann theory insufficient for protein folding simulations?* Journal of Chemical Physics, 2006. **124**(3).
54. Perez, A., et al., *Grid-Based Backbone Correction to the ff12SB Protein Force Field for Implicit-Solvent Simulations*. Journal of Chemical Theory and Computation, 2015. **11**(10): p. 4770-4779.
55. 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.
56. Best, R.B., N.V. Buchete, and G. Hummer, *Are current molecular dynamics force fields too helical?* Biophys J, 2008. **95**(1): p. L07-9.

57. 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.
58. Lindorff-Larsen, K., et al., *Improved side-chain torsion potentials for the Amber ff99SB protein force field*. Proteins-Structure Function and Bioinformatics, 2010. **78**(8): p. 1950-1958.
59. Berhanu, W.M. and U.H.E. Hansmann, *Side-chain hydrophobicity and the stability of A beta(16-22) aggregates*. Protein Science, 2012. **21**(12): p. 1837-1848.
60. Shao, Q.A., H.Y. Wei, and Y.Q. Gao, *Effects of Turn Stability and Side-Chain Hydrophobicity on the Folding of beta-Structures*. Journal of Molecular Biology, 2010. **402**(3): p. 595-609.
61. Diana, D., et al., *Long range Trp-Trp interaction initiates the folding pathway of a pro-angiogenic beta-hairpin peptide*. Sci Rep, 2015. **5**: p. 16651.
62. Zhang, C., et al., *Determination of atomic desolvation energies from the structures of crystallized proteins*. J Mol Biol, 1997. **267**(3): p. 707-26.
63. Miyazawa, S. and R.L. Jernigan, *Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading*. J Mol Biol, 1996. **256**(3): p. 623-44.
64. Berman, H.M., et al., *The Protein Data Bank*. Nucleic Acids Research, 2000. **28**(1): p. 235-242.
65. Zhang, C., et al., *Determination of atomic desolvation energies from the structures of crystallized proteins*. Journal of Molecular Biology, 1997. **267**(3): p. 707-726.
66. Cheatham, T.E. and M.A. Young, *Molecular dynamics simulation of nucleic acids: Successes, limitations, and promise*. Biopolymers, 2001. **56**(4): p. 232-256.
67. Durrant, J.D. and J.A. McCammon, *Molecular dynamics simulations and drug discovery*. BMC Biology, 2011. **9**.
68. Jorgensen, W.L. and J. TiradoRives, *Monte Carlo vs molecular dynamics for conformational sampling*. Journal of Physical Chemistry, 1996. **100**(34): p. 14508-14513.
69. Leopold, P.E., M. Montal, and J.N. Onuchic, *Protein folding funnels: a kinetic approach to the sequence-structure relationship*. Proc Natl Acad Sci U S A, 1992. **89**(18): p. 8721-5.
70. Dinner, A.R., et al., *Understanding protein folding via free-energy surfaces from theory and experiment*. Trends Biochem Sci, 2000. **25**(7): p. 331-9.
71. St-Pierre, J.F., N. Mousseau, and P. Derreumaux, *The complex folding pathways of protein A suggest a multiple-funnelled energy landscape*. Journal of Chemical Physics, 2008. **128**(4).

72. Bryngelson, J.D., et al., *Funnels, Pathways, and the Energy Landscape of Protein-Folding - a Synthesis*. Proteins-Structure Function and Genetics, 1995. **21**(3): p. 167-195.
73. Xi, L.L., et al., *Recognition of protein folding kinetics pathways based on amino acid properties information derived from primary sequence*. Chemometrics and Intelligent Laboratory Systems, 2013. **126**: p. 76-82.
74. Huang, L.T. and M.M. Gromiha, *Analysis and prediction of protein folding rates using quadratic response surface models*. Journal of Computational Chemistry, 2008. **29**(10): p. 1675-1683.
75. Zheng, O.Y. and L. Jie, *Predicting protein folding rates from geometric contact and amino acid sequence*. Protein Science, 2008. **17**(7): p. 1256-1263.
76. Bernardi, R.C., M.C.R. Melo, and K. Schulten, *Enhanced sampling techniques in molecular dynamics simulations of biological systems*. Biochimica Et Biophysica Acta-General Subjects, 2015. **1850**(5): p. 872-877.
77. Torrie, G.M. and J.P. Valleau, *Non-Physical Sampling Distributions in Monte-Carlo Free-Energy Estimation - Umbrella Sampling*. Journal of Computational Physics, 1977. **23**(2): p. 187-199.
78. Laio, A. and F.L. Gervasio, *Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science*. Reports on Progress in Physics, 2008. **71**(12).
79. Laio, A. and M. Parrinello, *Escaping free-energy minima*. Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(20): p. 12562-12566.
80. Clark, T., *G-Protein coupled receptors: answers from simulations*. Beilstein J Org Chem, 2017. **13**: p. 1071-1078.
81. Barducci, A., G. Bussi, and M. Parrinello, *Well-tempered metadynamics: a smoothly converging and tunable free-energy method*. Phys Rev Lett, 2008. **100**(2): p. 020603.
82. Barducci, A., G. Bussi, and M. Parrinello, *Well-tempered metadynamics: A smoothly converging and tunable free-energy method*. Physical Review Letters, 2008. **100**(2).
83. Zhou, R., *Replica exchange molecular dynamics method for protein folding simulation*. Methods Mol Biol, 2007. **350**: p. 205-23.
84. Rhee, Y.M. and V.S. Pande, *Multiplexed-replica exchange molecular dynamics method for protein folding simulation*. Biophys J, 2003. **84**(2 Pt 1): p. 775-86.
85. Sugita, Y. and Y. Okamoto, *Replica-exchange molecular dynamics method for protein folding*. Chemical Physics Letters, 1999. **314**(1-2): p. 141-151.
86. Patriksson, A. and D. van der Spoel, *A temperature predictor for parallel tempering simulations*. Phys Chem Chem Phys, 2008. **10**(15): p. 2073-7.

87. Abraham, M.J. and J.E. Gready, *Ensuring mixing efficiency of replica-exchange molecular dynamics simulations*. Journal of Chemical Theory and Computation, 2008. **4**(7): p. 1119-1128.
88. Lum, L.C.S.L., et al., *Risk factors for hemorrhage in severe dengue infections*. Journal of Pediatrics, 2002. **140**(5): p. 629-631.
89. Modis, Y., et al., *Structure of the dengue virus envelope protein after membrane fusion*. Nature, 2004. **427**(6972): p. 313-319.
90. Dejnirattisai, W., et al., *A new class of highly potent, broadly neutralizing antibodies isolated from viremic patients infected with dengue virus (vol 16, pg 171, 2015)*. Nature Immunology, 2015. **16**(7): p. 785-785.
91. Rouvinski, A., et al., *Recognition determinants of broadly neutralizing human antibodies against dengue viruses*. Nature, 2015. **520**(7545): p. 109-U260.
92. Flipse, J., et al., *Antibody-Dependent Enhancement of Dengue Virus Infection in Primary Human Macrophages; Balancing Higher Fusion against Antiviral Responses*. Scientific Reports, 2016. **6**.
93. Nguyen, H., et al., *Folding Simulations for Proteins with Diverse Topologies Are Accessible in Days with a Physics-Based Force Field and Implicit Solvent*. Journal of the American Chemical Society, 2014. **136**(40): p. 13959-13962.
94. Maier, J.A., et al., *ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB*. Journal of Chemical Theory and Computation, 2015. **11**(8): p. 3696-3713.
95. Dejnirattisai, W., et al., *A new class of highly potent, broadly neutralizing antibodies isolated from viremic patients infected with dengue virus*. Nature Immunology, 2015. **16**(2): p. 170-+.
96. Shell, M.S., R. Ritterson, and K.A. Dill, *A test on peptide stability of AMBER force fields with implicit solvation*. J Phys Chem B, 2008. **112**(22): p. 6878-86.
97. Nguyen, H., D.R. Roe, and C. Simmerling, *Improved Generalized Born Solvent Model Parameters for Protein Simulations*. J Chem Theory Comput, 2013. **9**(4): p. 2020-2034.
98. Nguyen, H., et al., *Folding simulations for proteins with diverse topologies are accessible in days with a physics-based force field and implicit solvent*. J Am Chem Soc, 2014. **136**(40): p. 13959-62.
99. Buck, M., et al., *Importance of the CMAP correction to the CHARMM22 protein force field: dynamics of hen lysozyme*. Biophys J, 2006. **90**(4): p. L36-8.
100. Song, D., et al., *ff14IDPs force field improving the conformation sampling of intrinsically disordered proteins*. Chem Biol Drug Des, 2017. **89**(1): p. 5-15.

101. Wang, W., et al., *New force field on modeling intrinsically disordered proteins*. Chem Biol Drug Des, 2014. **84**(3): p. 253-69.
102. Ye, W., et al., *Test and Evaluation of ff99IDPs Force Field for Intrinsically Disordered Proteins*. J Chem Inf Model, 2015. **55**(5): p. 1021-9.
103. Macias, A.T. and A.D. MacKerell, *CH/pi interactions involving aromatic amino acids: Refinement of the CHARMM tryptophan force field*. Journal of Computational Chemistry, 2005. **26**(14): p. 1452-1463.
104. Popp, A., et al., *Effect of Hydrophobic Interactions on the Folding Mechanism of beta-Hairpins*. Journal of Physical Chemistry B, 2014. **118**(49): p. 14234-14242.
105. Chen, R. and Z.P. Weng, *Docking unbound proteins using shape complementarity, desolvation, and electrostatics*. Proteins-Structure Function and Genetics, 2002. **47**(3): p. 281-294.
106. Vajda, S., *Classification of protein complexes based on docking difficulty*. Proteins-Structure Function and Bioinformatics, 2005. **60**(2): p. 176-180.
107. Xiao, Y., C. Chen, and Y. He, *Folding mechanism of beta-hairpin trpzip2: heterogeneity, transition state and folding pathways*. Int J Mol Sci, 2009. **10**(6): p. 2838-48.
108. Wang, H.W. and S.S. Sung, *Molecular dynamics simulations of three-strand beta-sheet folding*. Journal of the American Chemical Society, 2000. **122**(9): p. 1999-2009.
109. Wang, H.W., et al., *Molecular dynamics simulations of beta-hairpin folding*. Proteins-Structure Function and Genetics, 1999. **37**(3): p. 325-333.
110. Geng, H., F. Pang, and Y.D. Wu, *Accurate Structure Prediction and Conformational Analysis of Cyclic Peptides with Residue-Specific Force Fields*. Journal of Physical Chemistry Letters, 2016. **7**(10): p. 1805-1810.
111. Lindorff-Larsen, K., et al., *Systematic Validation of Protein Force Fields against Experimental Data*. Plos One, 2012. **7**(2).
112. Shen, Y., et al., *Improved PEP-FOLD Approach for Peptide and Miniprotein Structure Prediction*. J Chem Theory Comput, 2014. **10**(10): p. 4745-58.
113. Grantcharova, V., et al., *Mechanisms of protein folding*. Current Opinion in Structural Biology, 2001. **11**(1): p. 70-82.
114. Broglia, R.A. and G. Tiana, *Mechanism of folding and aggregation of proteins*. Protein Folding, Evolution and Design, 2001. **333**: p. 69-101.
115. D.A. Case, D.S.C., T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, D. Greene, N. Homeyer, S. Izadi, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D. Mermelstein, K.M. Merz, G. Monard, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, C.L. Simmerling, W.M. Botello-Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, L.

- Xiao, D.M. York and P.A. Kollman *Amber 2017; University of California: San Francisco, CA*. 2017.
116. Cino, E.A., W.Y. Choy, and M. Karttunen, *Comparison of Secondary Structure Formation Using 10 Different Force Fields in Microsecond Molecular Dynamics Simulations*. Journal of Chemical Theory and Computation, 2012. **8**(8): p. 2725-2740.
 117. Lo, S.C., et al., *Structure of the Keap1 : Nrf2 interface provides mechanistic insight into Nrf2 signaling*. Embo Journal, 2006. **25**(15): p. 3605-3617.
 118. Maupetit, J., P. Derreumaux, and P. Tuffery, *PEP-FOLD: an online resource for de novo peptide structure prediction*. Nucleic Acids Res, 2009. **37**(Web Server issue): p. W498-503.
 119. Mackerell, A.D., Jr., M. Feig, and C.L. Brooks, 3rd, *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*. J Comput Chem, 2004. **25**(11): p. 1400-15.
 120. Humphrey, W., A. Dalke, and K. Schulten, *VMD: visual molecular dynamics*. J Mol Graph, 1996. **14**(1): p. 33-8, 27-8.
 121. Joosten, R.P., et al., *A series of PDB related databases for everyday needs*. Nucleic Acids Research, 2011. **39**: p. D411-D419.
 122. Kabsch, W. and C. Sander, *Dictionary of Protein Secondary Structure - Pattern-Recognition of Hydrogen-Bonded and Geometrical Features*. Biopolymers, 1983. **22**(12): p. 2577-2637.
 123. Racine, J., *gnuplot 4.0: A portable interactive plotting utility*. Journal of Applied Econometrics, 2006. **21**(1): p. 133-141.
 124. Best, R.B. and J. Mittal, *Free-energy landscape of the GB1 hairpin in all-atom explicit solvent simulations with different force fields: Similarities and differences*. Proteins-Structure Function and Bioinformatics, 2011. **79**(4): p. 1318-1328.
 125. Chen, J.H., C.L. Brooks, and J. Khandogin, *Recent advances in implicit solvent-based methods for biomolecular simulations*. Current Opinion in Structural Biology, 2008. **18**(2): p. 140-148.
 126. Swanson, J.M.J., J. Mongan, and J.A. McCammon, *Limitations of atom-centered dielectric functions in implicit solvent models*. Journal of Physical Chemistry B, 2005. **109**(31): p. 14769-14772.
 127. Tanner, D.E., et al., *Parallel Generalized Born Implicit Solvent Calculations with NAMD*. Journal of Chemical Theory and Computation, 2011. **7**(11): p. 3635-3642.
 128. Zhou, R.H., G. Krilov, and B.J. Berne, *Comment on "Can a continuum solvent model reproduce the free energy landscape of a beta-hairpin*

- folding in water?" The Poisson-Boltzmann equation. Journal of Physical Chemistry B, 2004. 108(22): p. 7528-7530.*
129. Zhou, R.H. and B.J. Berne, *Can a continuum solvent model reproduce the free energy landscape of a beta-hairpin folding in water?* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(20): p. 12777-12782.
 130. Monticelli, L., et al., *Assessing the influence of electrostatic schemes on molecular dynamics simulations of secondary structure forming peptides.* Journal of Physics-Condensed Matter, 2006. **18**(14): p. S329-S345.
 131. Chen, R., L. Li, and Z.P. Weng, *ZDOCK: An initial-stage protein-docking algorithm.* Proteins-Structure Function and Genetics, 2003. **52**(1): p. 80-87.
 132. Kim, Y.C. and G. Hummer, *Coarse-grained models for simulations of multiprotein complexes: application to ubiquitin binding.* Journal of Molecular Biology, 2008. **375**(5): p. 1416-1433.
 133. Ravikumar, K.M., W. Huang, and S.C. Yang, *Coarse-Grained Simulations of Protein-Protein Association: An Energy Landscape Perspective.* Biophysical Journal, 2012. **103**(4): p. 837-845.
 134. Abd El-Aty, A.M., et al., *Determination of volatile organic compounds generated from fresh, white and red Panax ginseng (C. A. Meyer) using a direct sample injection technique.* Biomed Chromatogr, 2008. **22**(5): p. 556-62.
 135. Ravikumar, K.M., W. Huang, and S. Yang, *Coarse-grained simulations of protein-protein association: An energy landscape perspective.* Biophys J, 2012. **103**(4): p. 837-45.
 136. Lovell, S.C., et al., *Structure validation by Calpha geometry: phi,psi and Cbeta deviation.* Proteins, 2003. **50**(3): p. 437-50.
 137. Swindells, M.B., M.W. MacArthur, and J.M. Thornton, *Intrinsic phi, psi propensities of amino acids, derived from the coil regions of known structures.* Nat Struct Biol, 1995. **2**(7): p. 596-603.
 138. Wu, L., et al., *Folding of proteins with an all-atom Go-model.* Journal of Chemical Physics, 2008. **128**(23).
 139. Mallik, B., A. Masunov, and T. Lazaridis, *Distance and exposure dependent effective dielectric function.* J Comput Chem, 2002. **23**(11): p. 1090-9.
 140. Ha-Duong, T., *Protein Backbone Dynamics Simulations Using Coarse-Grained Bonded Potentials and Simplified Hydrogen Bonds.* Journal of Chemical Theory and Computation, 2010. **6**(3): p. 761-773.
 141. Sandberg, L. and O. Edholm, *A fast and simple method to calculate protonation states in proteins.* Proteins-Structure Function and Genetics, 1999. **36**(4): p. 474-483.
 142. Zondlo, N.J., *Aromatic-proline interactions: electronically tunable CH/pi interactions.* Acc Chem Res, 2013. **46**(4): p. 1039-49.

143. Perez, A., et al., *Grid-based backbone correction to the ff12SB protein force field for implicit-solvent simulations*. J Chem Theory Comput, 2015. **11**(10): p. 4770-9.
144. Fitzkee, N.C., P.J. Fleming, and G.D. Rose, *The protein coil library: A structural database of nonhelix, nonstrand fragments derived from the PDB*. Proteins-Structure Function and Bioinformatics, 2005. **58**(4): p. 852-854.
145. Moore, T.C., C.R. Iacovella, and C. McCabe, *Derivation of coarse-grained potentials via multistate iterative Boltzmann inversion*. Journal of Chemical Physics, 2014. **140**(22).
146. 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.
147. Zhou, C.Y., F. Jiang, and Y.D. Wu, *Residue-specific force field based on protein coil library. RSFF2: modification of AMBER ff99SB*. J Phys Chem B, 2015. **119**(3): p. 1035-47.
148. Berman, H.M., et al., *The Protein Data Bank*. Nucleic Acids Res, 2000. **28**(1): p. 235-42.
149. Earl, D.J. and M.W. Deem, *Parallel tempering: Theory, applications, and new perspectives*. Physical Chemistry Chemical Physics, 2005. **7**(23): p. 3910-3916.
150. Kim, Y.C. and G. Hummer, *Coarse-grained models for simulations of multiprotein complexes: application to ubiquitin binding*. J Mol Biol, 2008. **375**(5): p. 1416-33.
151. Hemungkorn, M., U. Thisyakorn, and C. Thisyakorn, *Dengue infection: a growing global health threat*. Biosci Trends, 2007. **1**(2): p. 90-6.
152. Kliks, S.C., et al., *Antibody-Dependent Enhancement of Dengue Virus Growth in Human-Monocytes as a Risk Factor for Dengue Hemorrhagic-Fever*. American Journal of Tropical Medicine and Hygiene, 1989. **40**(4): p. 444-451.
153. Flasche, S., et al., *The Long-Term Safety, Public Health Impact, and Cost-Effectiveness of Routine Vaccination with a Recombinant, Live-Attenuated Dengue Vaccine (Dengvaxia): A Model Comparison Study*. Plos Medicine, 2016. **13**(11).
154. Aguiar, M. and N. Stollenwerk, *Dengvaxia: age as surrogate for serostatus*. Lancet Infectious Diseases, 2018. **18**(3): p. 245-245.
155. Swanstrom, J.A., et al., *Dengue Virus Envelope Dimer Epitope Monoclonal Antibodies Isolated from Dengue Patients Are Protective against Zika Virus*. MBio, 2016. **7**(4).
156. Marzinek, J.K., et al., *Pushing the Envelope: Dengue Viral Membrane Coaxed into Shape by Molecular Simulations*. Structure, 2016. **24**(8): p. 1410-1420.

157. Huber, R.G., et al., *Multiscale molecular dynamics simulation approaches to the structure and dynamics of viruses*. Progress in Biophysics & Molecular Biology, 2017. **128**: p. 121-132.
158. Prakash, M.K., A. Barducci, and M. Parrinello, *Probing the Mechanism of pH-Induced Large-Scale Conformational Changes in Dengue Virus Envelope Protein Using Atomistic Simulations*. Biophysical Journal, 2010. **99**(2): p. 588-594.
159. Dubey, K.D., A.K. Chaubey, and R.P. Ojha, *Role of pH on dimeric interactions for DENV envelope protein: An insight from molecular dynamics study*. Biochimica Et Biophysica Acta-Proteins and Proteomics, 2011. **1814**(12): p. 1796-1801.
160. Sussman, J.L., et al., *Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules*. Acta Crystallogr D Biol Crystallogr, 1998. **54**(Pt 6 Pt 1): p. 1078-84.
161. Rouvinski, A., et al., *Recognition determinants of broadly neutralizing human antibodies against dengue viruses*. Nature, 2015. **520**(7545): p. 109-13.
162. Moss, T., et al., *Ph-Dependent Interaction between Histones H2a and H2b Involving Secondary and Tertiary Folding*. European Journal of Biochemistry, 1976. **71**(2): p. 337-350.
163. Kirschner, K.N., et al., *A Glycam-Based Force Field for Simulations of Lipopolysaccharide Membranes: Parametrization and Validation*. J Chem Theory Comput, 2012. **8**(11): p. 4719-31.
164. Wehle, M., et al., *Mechanical Compressibility of the Glycosylphosphatidylinositol (GPI) Anchor Backbone Governed by Independent Glycosidic Linkages*. Journal of the American Chemical Society, 2012. **134**(46): p. 18964-18972.
165. Bussi, G., D. Donadio, and M. Parrinello, *Canonical sampling through velocity rescaling*. Journal of Chemical Physics, 2007. **126**(1).
166. Parrinello, M. and A. Rahman, *Polymorphic Transitions in Single-Crystals - a New Molecular-Dynamics Method*. Journal of Applied Physics, 1981. **52**(12): p. 7182-7190.
167. Tribello, G.A., et al., *PLUMED 2: New feathers for an old bird*. Computer Physics Communications, 2014. **185**(2): p. 604-613.
168. Vertrees, J., *InterfaceResidues.py*. 2009.
169. Bonomi, M., et al., *PLUMED: A portable plugin for free-energy calculations with molecular dynamics*. Computer Physics Communications, 2009. **180**(10): p. 1961-1972.
170. Incerti, M., et al., *Metadynamics for Perspective Drug Design: Computationally Driven Synthesis of New Protein-Protein Interaction Inhibitors Targeting the EphA2 Receptor*. Journal of Medicinal Chemistry, 2017. **60**(2): p. 787-796.

171. Fidelak, J., et al., *Free-Energy-Based Methods for Binding Profile Determination in a Congeneric Series of CDK2 Inhibitors*. Journal of Physical Chemistry B, 2010. **114**(29): p. 9516-9524.
172. Incerti, M., et al., *Metadynamics for Perspective Drug Design: Computationally Driven Synthesis of New Protein-Protein Interaction Inhibitors Targeting the EphA2 Receptor*. J Med Chem, 2017. **60**(2): p. 787-796.
173. Cavalli, A., et al., *Investigating Drug-Target Association and Dissociation Mechanisms Using Metadynamics-Based Algorithms*. Accounts of Chemical Research, 2015. **48**(2): p. 277-285.

7 PUBLICATION LIST

Zhensheng Zhong, Lixia Yang, Haiping Zhang, Jiahao Shi, Jeya Vandana, Do Thuy Uyen Ha Lam, René C. L. Olsthoorn, Lanyuan Lu, and Gang Chen* (2016) Mechanical unfolding kinetics of the SRV-1 gag-pro mRNA pseudoknot: possible implications for -1 ribosomal frameshifting stimulation, *Sci Rep*, 6, 39549.

Haiping Zhang, Lanyuan Lu*, Improvement of GBSA simulation performances by adding a cMAP term and strengthening aromatic related interactions, (in preparation).

Haiping Zhang, Lanyuan Lu*, Solvent-free all-atom force field incorporating experimental native structure information, (in preparation).

Haiping Zhang, Lanyuan Lu*, Investigating the stability of DENV-2 envelope protein dimer using well-tempered metadynamics simulations, (in preparation).

Kiran M. Patil,⁼ Desiree-Faye Kaixin Toh,⁼ Zhen Yuan,⁼ Zhenyu Meng,⁼ Zhiyu Shu, Haiping Zhang, Lanyuan Lu, Yunpeng Lu,* and Gang Chen,* Incorporating uracil and 5-halouracils into short peptide nucleic acids for enhanced recognition of A-U pairs in dsRNAs, (in preparation).