



2018

# Computational Methods in Biomolecules: Study of Hydrophilic Interactions in Protein Folding & Constant-pH Molecular Simulation of pH Sensitive Lipid MORC16

Wei Zhang

University of the Pacific, [w\\_zhang@u.pacific.edu](mailto:w_zhang@u.pacific.edu)

Follow this and additional works at: [https://scholarlycommons.pacific.edu/uop\\_etds](https://scholarlycommons.pacific.edu/uop_etds)

 Part of the [Pharmacy and Pharmaceutical Sciences Commons](#)

## Recommended Citation

Zhang, Wei. (2018). *Computational Methods in Biomolecules: Study of Hydrophilic Interactions in Protein Folding & Constant-pH Molecular Simulation of pH Sensitive Lipid MORC16*. University of the Pacific, Dissertation. [https://scholarlycommons.pacific.edu/uop\\_etds/3145](https://scholarlycommons.pacific.edu/uop_etds/3145)

This Dissertation is brought to you for free and open access by the Graduate School at Scholarly Commons. It has been accepted for inclusion in University of the Pacific Theses and Dissertations by an authorized administrator of Scholarly Commons. For more information, please contact [mgibney@pacific.edu](mailto:mgibney@pacific.edu).

COMPUTATIONAL METHODS IN BIOMOLECULES: STUDY OF HYDROPHILIC  
INTERACTIONS IN PROTEIN FOLDING & CONSTANT-PH MOLECULAR  
SIMULATION OF PH SENSITIVE LIPID MORC16

by

Wei Zhang

A Dissertation Submitted to the

Graduate School

In Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

Chemistry Department  
Bioanalytical and Physical Chemistry

University of the Pacific  
Stockton, CA

2018

COMPUTATIONAL METHODS IN BIOMOLECULES: STUDY OF HYDROPHILIC  
INTERACTIONS IN PROTEIN FOLDING & CONSTANT-PH MOLECULAR  
SIMULATION OF PH SENSITIVE LIPID MORC16

by

Wei Zhang

APPROVED BY:

Dissertation Advisor: Charles M. McCallum, Ph.D.

Committee Member: Jerry Tsai, Ph.D.

Committee Member: Jianhua Ren, Ph.D.

Committee Member: Qinliang Zhao, Ph.D.

Committee Member: Jialing Chan, Ph.D.

Department Chair: Jianhua Ren, Ph.D.

Dean of Graduate School: Thomas Naehr., Ph.D.

COMPUTATIONAL METHODS IN BIOMOLECULES: STUDY OF HYDROPHILIC  
INTERACTIONS IN PROTEIN FOLDING & CONSTANT-PH MOLECULAR  
SIMULATION OF PH SENSITIVE LIPID MORC16

Copyright 2018

by

Wei Zhang

## DEDICATION

This thesis is a dedication to my parents, who have always loved me, and support me since I was born. They gave me many encouragement and suggestions in finishing my study at school. Also to my wife Chau, we met and married in the middle of my study at UOP. She is a kind, hard-working and filial woman, thank her for everything she has done for me.

## ACKNOWLEDGMENTS

My gratitude goes to my adviser Dr. McCallum who have always been a constant source of supporting to my life and research for these years. You will never feel anxious or sad from him, because he always make people think positive and move forward. When he did not agree with something, he convinced me with a mild and tactful way. I also want to thank all my committee members: Dr. Tsai helped me in coordinating my qualify exam and defense, and taught me a lot in my research area. Dr. Ren gave me a lot of support and suggestions for my research, she is more strict and direct, made me dismiss any casual attitude. Dr. Zhao helped me in my life and gave me a lot of suggestions in writing my thesis. Dr. Chan always encouraged me every time we met.

Dr. Joo is the person I talked most about my research during these years, he knows almost everything, he taught me many knowledge about molecular simulation and helped me fixing many problems. Dr. Xiaoning Zhao is my best friend, his help is invaluable and countless.

I also want to thank all PCSP faculties, especially Sue, Kathy and Bret.

I want to thank all my friends, they make my life full of joy and happiness.

COMPUTATIONAL METHODS IN BIOMOLECULES: STUDY OF HYDROPHILIC  
INTERACTION IN PROTEIN FOLDING & CONSTANT-PH MOLECULAR  
SIMULATION OF PH SENSITIVE LIPID MORC16

Abstract

by Wei Zhang

University of the Pacific  
2018

Water molecules play a significant role in biological process and are directly involved with bio-molecules and organic compounds and ions. Recent research has focused on the thermal dynamics and kinetics of water molecules in solution, including experimental (infrared spectroscopy and Raman spectroscopy) and computational (Quantum Mechanics and Molecular Dynamics) approaches. The reason that water molecules are so unique, why they have such a profound influence on bio-activity, why water molecules show some anomalies compared to other small molecules, and where and how water molecules exert their influence on solutes are some of the areas under study. We studied some properties of hydrogen bond networks, and the relationship of these properties with solutes in water. Molecular dynamics simulation, followed by an analysis of “water

bridges”, which represent protein-water interaction have been carried out on folded and unfolded proteins. Results suggest that the formation of transient water bridges within a certain distance helps to consolidate the protein, possibly in transition states, and may help further guide the correct folding of proteins from these transition states. This is supporting evidence that a hydrophilic interaction is the driving force of protein folding. Biological membranes are complex structures formed mostly by lipids and proteins. For this reason the lipid bilayer has received much attention, through computation and experimental studies in recent years. In this dissertation, we report results of a newly designed pH sensitive lipid MORC16, through all-atom and coarse-grained models. The results did not yield a MORC16 amphiphile which flips its conformation in response to protonation. This may be due to imperfect force field parameters for this lipid, an imperfect protonation definition, or formation of hydrogen bond does not responsible for conformation flip in our models. Despite this, some insights for future work were obtained.



## TABLE OF CONTENTS

<u>LIST OF FIGURES.....</u>	<u>11</u>
<u>Chapter 1: Introduction.....</u>	<u>14</u>
<u>1.1 History and Development of MD Simulation.....</u>	<u>14</u>
<u>1.2 General Classical Molecular Dynamics Simulation.....</u>	<u>16</u>
<u>1.3 Accelerated Molecular Dynamics Simulation.....</u>	<u>19</u>
<u>1.3.1 Steered Molecular Dynamics (SMD).....</u>	<u>19</u>
<u>1.3.2 Replica Exchange Molecular Dynamics.....</u>	<u>20</u>
<u>1.3.3 Dissertation Outline.....</u>	<u>22</u>
<u>Chapter 2: The Role of Water Bridges on Protein Folding and Stability.....</u>	<u>23</u>
<u>2.1 Protein Folding Problem.....</u>	<u>23</u>
<u>2.2 Water Bridges and Water Networks.....</u>	<u>26</u>
<u>2.3 Hypothesis.....</u>	<u>33</u>
<u>2.4 Material and Tools.....</u>	<u>34</u>
<u>2.5 Analysis.....</u>	<u>35</u>
<u>2.5.1 Analysis of Hydrogen Bond.....</u>	<u>35</u>
<u>2.5.2 Analysis of Hydrogen Bond Network.....</u>	<u>43</u>

<a href="#">2.5.3 Influence of Ions on Hydrogen Bond Network Property.....</a>	<a href="#">48</a>
<a href="#">2.5.4 Analysis of Water Bridges.....</a>	<a href="#">54</a>
<a href="#">2.6 Discussion and Conclusion.....</a>	<a href="#">69</a>
<a href="#">Chapter 3: Effect of Protein Unfolding on pKa Shifts of BPTI.....</a>	<a href="#">71</a>
<a href="#">3.1 Review of Constant pH Simulations.....</a>	<a href="#">71</a>
<a href="#">3.2 Method - CPMD.....</a>	<a href="#">74</a>
<a href="#">3.3 Analysis.....</a>	<a href="#">76</a>
<a href="#">3.4 Discussion and Conclusion.....</a>	<a href="#">83</a>
<a href="#">Chapter 4: Modeling and Simulation of pH-Sensitive Lipids.....</a>	<a href="#">84</a>
<a href="#">4.1 Review of Lipid Simulations and MORC16.....</a>	<a href="#">84</a>
<a href="#">4.2 Modeling and Simulation of MORC16.....</a>	<a href="#">85</a>
<a href="#">4.2.1 MORC16 Constant pH Simulation using CHARMM.....</a>	<a href="#">86</a>
<a href="#">4.2.2 The Self Assembly of MORC16 with GROMACS.....</a>	<a href="#">94</a>
<a href="#">4.2.3 Stability of MORC16 Micelles and liposomes by AMBER.....</a>	<a href="#">104</a>
<a href="#">APPENDIX A. DATA SET OF 100 PROTEINS.....</a>	<a href="#">124</a>
<a href="#">APPENDIX B. PRINCIPAL COORDINATES ANALYSIS.....</a>	<a href="#">126</a>

## LIST OF TABLES

Table	Page
1. Data fitting of function $P(t)$ with exponential equations.....	43
2. Weighting factors for water bridge following strategy I and II.....	56
3. Calculated pKa values of titratable amino acid fragments.....	77
4. Calculated pKa values of titratable sites in BPTI.....	79
5. Summary of the Coarse-Grain Strategy for MORC16.....	96

## LIST OF FIGURES

Figure	Page
Fig. 1: Schematic of SMD method. From Phys4500 Lecuture, MU(University of Missouri).....	21
Fig. 2: Work flow of replica exchange molecular dynamics. From Wikimedia Commons. .....	22
Fig. 3: Protein folding models.(127).....	25
Fig. 4: Thermodynamics of hydrophobic solvation. From LibreTexts.....	28
Fig. 5: Schematic of one water bridge in protein aggregation. From Arieh Ben-Naim....	29
Fig. 6: Representations of the most stable geometries for (H <sub>2</sub> O) <sub>n</sub> clusters.(59).....	30
Fig. 7: Water bridges (colored in blue) assisted protein binding.(74).....	32
Fig. 8: Water binding in ribonucleotide reductase from C. ammoniagenes.(56).....	33
Fig. 9: Definition of hydrogen bond.....	37
Fig. 10: Degree of hydrogen bonding in pure water, defined by CHARMM.....	40
Fig. 11: Degree of hydrogen bonding in pure water, defined by CHARMM, NAMD and AMBER.....	42
Fig. 12: Survival probability of hydrogen bond lifetime. Red curve is data fitting to $Ae(-t/\tau_1)$ .....	43
Fig. 13: Survival probability of hydrogen bond residence time. Red curve is data fitting to $Ae(-t/\tau_1)$ .....	44
Fig. 14: Survival probability of hydrogen bond residence time. Red curve is data fitting to $Ae(-t/\tau_1)+(1-A)e(-t/\tau_2)$ .....	45

Fig. 15: Topological representation of a water hydrogen bond network of pure water box (9261 water molecules). From igraph in R.....	48
Fig. 16: Community size distribution of a water hydrogen bond network.....	49
Fig. 17: Snap shots of four water communities (green spheres) involving one oxygen atom (red sphere) in four successive frames.....	53
Fig. 18: Influence of ions on degree of hydrogen bonding.....	55
Fig. 19: Influence of ions on hydrogen bond lifetime. Solid curves are density distribution of hydrogen bond life time, dashed vertical lines are calculated average hydrogen bond life time.....	56
Fig. 20: Influence of ions on network community size. x label is community size classified into groups, y label is fraction of water molecules in that group.....	59
Fig. 21: Fraction of direct hydrogen bond and bridge of each anchor atom type. BB: backbone bridge to backbone; SS: side chain bridge to side chain. SB/BS: side chain to backbone.....	63
Fig. 22: Bridge contact map of 20 amino acid residues. The number is hydrophilic interaction strength (from bridge score) with respect to GLU-ARG (defined as one).....	65
Fig. 23: Snapshots of constant velocity pulling of BPTI. Snapshots from step 1 (top), step 250, step 500 and step 1000 (bottom). 1 step=0.002ps.....	67
Fig. 24: Water bridge score along the unfolding of BPTI. Blue line is score from direct hydrogen bond, red line is score from water bridge following strategy I and green line is score from water bridge following strategy II.....	68
Fig. 25: 30 replicas of BPTI after 100 exchanges.....	70
Fig. 26: Heat map of 30 replicas along replica exchange cycle of 100. Legend scale represents RMDS from native structure.....	71
Fig. 27: Energy distribution of 30 replicas of BPTI.....	72
Fig. 28: Potential energy landscape of 30 replicas by Principal Coordinates Analysis. PC1 and PC2 are recalculated coordinates.....	73
Fig. 29: Correlation between RMSD and bridge score for 30 replicas.....	75
Fig. 30: Titration curves for titratable residues in BPTI in the explicit solvent model. From pH-REX simulation with 14 replicas (pH=1,2,...,14) .....	84
Fig. 31: pKa shifts of 8 titratable residues in BPTI during SMD.....	86

Fig. 32: 3D representation of folded BPTI. Hydrogen bonds of ASP3, GLU7 and GLU49 with their neighbors are labeled with a dashed line.....	88
Fig. 33: 3D representation of extended BPTI.....	89
Fig. 34: Formula and 3D representation of MORC16 lipid.....	91
Fig. 35: Protonation of nitrogen on MORC16.....	93
Fig. 36: 3D image of protonated, solvated and ionized (three Cl <sup>-</sup> in cyan , two Na <sup>+</sup> in yellow) MORC16 lipid.....	95
Fig. 37: Data fitting of parameters A, B of PMF function. x is $\theta$ , y is E.....	96
Fig. 38: Titration curve of MORC16 by constant pH replica exchange simulation. pKa=5.05±0.78.....	97
Fig. 39: 3D images of the titratable site in MORC16 at pH=1 (top) and pH=10 (bottom). .....	98
Fig. 40: MARTINI mapping strategy for DPPC, cholesterol and benzene.117.....	100
Fig. 41: Coarse-grained representation of MORC1 superimposed on its all-atom representation.....	101
Fig. 42: Snapshots of 1280 MORC16 lipids in aggregation. Waters are depicted as cyan dots, hydrophilic heads of MORC16 in blue and red, tails of MORC16 in cyan.....	104
Fig. 43: Determination of MORC16 bilayer thickness from simulation.....	105
Fig. 44: Measurement of MORC16 bilayer lateral diffusion from simulation.....	106
Fig. 45: Snapshots from a simulation of aggregation of 12800 MORC16 lipids.....	108
Fig. 46: Snapshots from the self-assembly simulation of MORC16 with Na <sup>+</sup> (blue sphere).....	109
Fig. 47: Snapshots from self-assembly simulation of MORC16 with Cl <sup>-</sup> (cyan sphere). .....	110
Fig. 48: The self assembly simulation of MORC16 lipids(250) in AMBER.....	112
Fig. 49: Cross section of MORC16 micelle and liposome built by PACMOL.....	113
Fig. 50: PMF of pulling one lipid from micelle and liposome surface.....	114
Fig. 51: PMF of pulling one ion through micelle (100 lipids).....	115
Fig. 52: Snapshots of the pulling of Na <sup>+</sup> through a liposome.....	116

## **Chapter 1: Introduction**

### **1.1 History and Development of MD Simulation**

Experiments, along with the guidance of theory and mechanisms, enhance the development of science and technology. The word “experiment” can be explained in a general or in a narrow sense. Here we refer this word to represent all techniques to discover the truth of a specific phenomenon. In case of chemistry research, many instruments are designed and developed in order to investigate the composition or structures of compounds that are interesting, and such experiments may be used to verify hypotheses and and formulate new theory.

With the cooperation of computer science, more and more experimental hypotheses have been modeled and simulated by building a virtual environment on the computer in order to seek the mechanism underneath the phenomenon.

Molecular Dynamics (MD) simulation is a method used to explore chemical phenomena on an atom-level scale. It is a useful method in order to study condensed matter systems, and with the help of this method, we can model the trajectory of atoms, which provides details of microstates. This then produces a supplement to first-principle Quantum-Mechanical (QM) calculations and lab experiments.

The first paper that attempted to apply classical statistics to properties of substances was by Metropolis in 1953<sup>1</sup>. This paper applied the Monte-Carlo technique to physical equations of a system composed of interacting individual molecules with only two-body forces are considered.

The first “proper” molecular dynamics simulation was performed by B.J. Alder and T.E. Wainwright in 1956<sup>2</sup>. The simulation was just able to calculate equilibrium properties and showed significant difference from Monte Carlo results. This was the first time that applying periodic boundary conditions were applied in simulation.

The first realistic molecular simulation of a liquid was performed in 1964 by A. Rahhan<sup>3</sup>, who has become known as father of molecular dynamics. This paper was focused on studying the space and time dependence of two-body correlations within a classical dynamics frame work. The first molecular simulation of water was also achieved in 1971, also by A. Rahhan<sup>4</sup>. Water simulations pose a greater challenge than liquid noble gases, as in addition to van der Waals interactions, there are coulomb and hydrogen bond interactions in liquid water.

The first molecular simulation of a protein was published by Michael Levitt and Arich Warshel<sup>5</sup> in 1975. This paper employed a simple representation of protein structure which averaged over groups of atoms, and successfully renatured BPTI from an open-chain conformation.

The first all-atom molecular simulation of a protein was performed by J. Andrew McCammon<sup>6</sup> in 1977. In this study, the classical equations of motion for all the atoms of



the protein were solved simultaneously for a suitable time around its native conformation, thus being able to observe equilibrium properties.

The 1980s was a booming period of molecular simulation, with a range of biomolecules<sup>7</sup> (protein, DNA) being simulated. A variety implicit solvent model (Poisson-Boltzmann<sup>8</sup>, Generalized-Born<sup>9</sup>) were developed, and continued development of potentials and enhanced sampling techniques (Umbrella Sampling<sup>10</sup>, Replica Exchange<sup>11</sup>).

In the 1990s, more and more molecular simulation packages were developed. Packages such as CHARMM, GROMACS, NAMD and AMBER allow researchers to perform molecular dynamics simulation for their own interest. In the late 1990s, driven by the computer game market, calculation performed on graphic cards became popular, which led to some optimization of MD simulation algorithm on these graphic cards<sup>12-19</sup>.

In 21st century, MD continues to move forward with the fast development of hardware. It is only a matter of time before simulation of whole-cell system<sup>20,21</sup> becomes feasible.

## **1.2 General Classical Molecular Dynamics Simulation**

The key function of molecular dynamics is the creation of a potential function, described by the classical Newton equations. Molecular dynamics assumes atoms move with deterministic trajectories, with the assumption of the ignoring any quantum effects, and utilizing the Born-Oppenheimer adiabatic approximation, which supposes all atoms are in their ground electronic state. The interaction between atoms in MD employs mature

potentials such as Lennard-Jones potential<sup>22</sup>, Morse potential<sup>23</sup> and Embedded Atom Method (EAM)<sup>24</sup>. These potentials provide the relative close results of interaction between atoms while losing some effects of electron coupling.

The general procedure in MD is to start with a set of initial coordinates and velocities of the  $N$  particles contained in system. As the simulation proceeds, the particles evolve through the solvation of Newton's equations of motion in increments of  $\Delta t$  (typically 1-2 fs). Since the force is calculated by taking the gradient of the total potential  $U$ , the specification of  $U$  essentially determines the compromise between physical fidelity and computational efficiency, and this "force field" becomes the most important part of the process.

Here we list the two most commonly-used potential functions : the AMBER<sup>25</sup> and the CHARMM<sup>26,27</sup> force fields respectively.

$$\begin{aligned}
 V_{AMBER} = & \sum_{bonds} k(r - r_{eq})^2 + \sum_{angles} k(\theta - \theta_{eq})^2 + \sum_{dihedrals} \frac{V_2}{2} [1 + \cos(n\phi - \gamma)] \\
 & + \sum_{i < j} \left[ \frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6} \right] + \sum_{i < j} \left[ \frac{q_i q_j}{\epsilon R_{ij}} \right]
 \end{aligned} \tag{1.1}$$

$$\begin{aligned}
 V_{CHARMM} = & \sum_{bonds} k_b(b - b_0)^2 + \sum_{angles} k_\theta(\theta - \theta_0)^2 + \sum_{dihedrals} k_\phi[1 + \cos(n\phi - \delta)] \\
 & + \sum_{Urey-Bradley} k_u(u - u_0)^2 + \sum_{impropers} k_\omega(\omega - \omega_0)^2 + \sum_{\phi, \psi} V_{CMAP} \\
 & + \sum_{nonbonded} \epsilon \left[ \left( \frac{R_{min_{ij}}}{r_{ij}} \right)^{12} - \left( \frac{R_{min_{ij}}}{r_{ij}} \right)^6 \right] + \frac{q_i q_j}{\epsilon r_{ij}}
 \end{aligned} \tag{1.2}$$

The Amber force field is a classical force field employed by many molecular dynamics packages. The CHARMM force field added three additional terms in version CHARMM22<sup>28</sup> (released in 1991). The first term is a two-body Urey-Bradley term<sup>29</sup> which is a cross-term accounting for angle bending using 1,3 nonbonded interactions, where  $k_u$  is the respective force constant and  $u$  is the distance between the 1,3 atoms in the harmonic potential. The second is a four-body quadratic improper term which describes out of plane bending;  $k_\omega$  is the force constant and  $\omega$  is the out of plane angle. The third is a cross term named CMAP<sup>30</sup>, which has been used to improve the treatment of the conformational properties of the  $\varphi/\psi$  dihedral terms in the peptide backbone, explicitly derived in order to correct nonphysical helical structures.

Both CHARMM and AMBER give good results for proteins and DNA, but for some organic compounds like drug molecules, CGenFF<sup>26</sup> (CHARMM General Force Field, an extension of the CHARMM forcefield) or MMFF<sup>31</sup> (Merck Molecular Force Field, developed by Merck Research Laboratories) are better options.

The general limitations in MD methods include limited timescales, which are often far shorter than useful or effective; the force fields are inherently approximations with accuracy errors; and the fact that covalent bonds cannot break or form during conventional MD simulations.

### **1.3 Accelerated Molecular Dynamics Simulation**

Some calculations of properties may not be available using general molecular dynamics simulation. Accelerated molecular dynamics is a class of enhanced MD methods that

attempt to improve the conformational sampling space by either reducing energy barriers or using non-Boltzmann weighting factors to cross any barriers encountered<sup>32-34</sup>.

### **1.3.1 Steered Molecular Dynamics (SMD)**

Steered Molecular dynamics was inspired by atomic force microscopy experiments, which was initially developed in 1990s<sup>35,36</sup>, and has been successfully applied to protein unfolding and other biochemical system including ions or molecule transport<sup>37-40</sup>. The major difference between SMD and regular MD is the constraints on atoms. In regular MD, all atoms are dynamically evolved without constraints or harmonically constrained to a reference point in all three spatial directions. In SMD, tagged atoms have their centers of mass constrained along an outside constraint direction to a reference point which moves with constant velocity. As shown in Fig.1, the atoms aren't pulled directly, but instead move through a dummy atom. The interaction between the dummy atom and the atom of interest is determined through a spring with a corresponding spring constant.

$$U(\vec{r}_i) = \frac{k}{2} \left( v\vec{n}t - [\vec{R}(t) - \vec{R}(0)] \right)^2$$

Fig. 1: Schematic of SMD method. From Phys4500 Lecture,  
MU(University of Missouri)

### 1.3.2 Replica Exchange Molecular Dynamics

Replica Exchange Molecular Dynamics (REMD)<sup>11,41,42</sup> is a parallel simulation technique, which combines the concepts of a multicanonical algorithm simulation, simulated annealing and Monte Carlo methods to attempt to sample a random walk in the energy space, which allows the more efficient sampling of more conformational space. Thus REMD is also referred to as replica Monte Carlo method<sup>43,44</sup> or parallel tempering<sup>45,46</sup>.

The motivation for REMD is that storage has become cheap, but CPU time remains expensive, so simulation time (statistics) may be improved through replication with the fast development of multithreading techniques.

In REMD simulations, a set of parallel independent simulations are conducted over a range of conditions of a single molecular system. The variable here may be any factor that affect the Boltzmann distribution. The most popular condition is temperature, but it

could also be a reaction coordinate such as angle or distance. It may also be physical conditions such as pH or the ionic strength of the system. Here we illustrate the method with a temperature replica exchange simulation (Fig. 2).

A number of systems at different temperatures are allowed to exchange configurations at fixed time intervals with a transition probability following Metropolis criterion:

$$P_{accept} = \min[1, \exp[-\frac{1}{k}(\frac{1}{T_i} - \frac{1}{T_j})(U(X_i) - U(X_j))]] \quad (1.3)$$

where  $U(X_i)$  is the total potential energy of replica  $i$  at temperature  $T_i$ ;  $k$  is Boltzmann's constant.

$P_{accept}=1$  when  $T_i < T_j$ , and  $U(X_i) < U(X_j)$ .

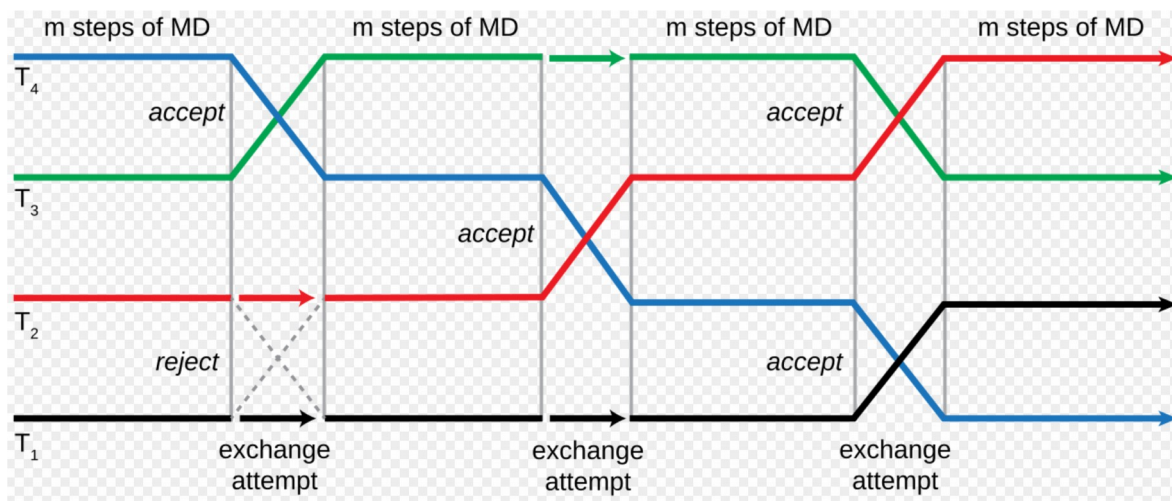


Fig. 2: Work flow of replica exchange molecular dynamics. From Wikimedia Commons.

Each vertical column represents one of the  $M$  replicas of the system. The replicas are distributed in a Boltzmann manner from low to high temperature, and after a certain amount of simulation to time  $t_{ix}$ , exchanges are attempted between two neighboring replicas. If accepted, the temperatures will be exchanged, and the simulation continues with the updated temperatures assigned to different replicas.

### 1.3.3 Dissertation Outline

This dissertation includes two major parts. First is the study of the water-protein interaction through an analysis of water bridges. Second the development of a constant-pH simulation of the MORC16 lipid. We first studied some properties of hydrogen bond network in water including degree of hydrogen bonding and lifetime of hydrogen bonds (Chapter 2.5.1), then studied the influence of solutes in water to these properties (Chapter 2.5.3). We analyzed hydrogen bond network through community detection through R (Chapter 2.5.2). Water bridge information was determined through the development of a tcl script and VMD (Chapter 2.5.4), then analyzed. The role of water bridge in folded, unfolded (through SMD) and misfolded (through REMD) proteins were investigated. The full development and constant-pH simulation of BPTI, from folded to unfolded states, in order to study the pKa shift take up the entirety of Chapter 3.

Constant-pH simulation of MORC16 lipid were performed through CHARMM (Chapter 4.2.1). The development and execution of lipid aggregation simulations through MARTINI coarse-grained model (Chapter 4.2.2) and bilayer stability through AMBER (Chapter 4.2.3) are presented.

## **Chapter 2: The Role of Water Bridges on Protein Folding and Stability**

### **2.1 Protein Folding Problem**

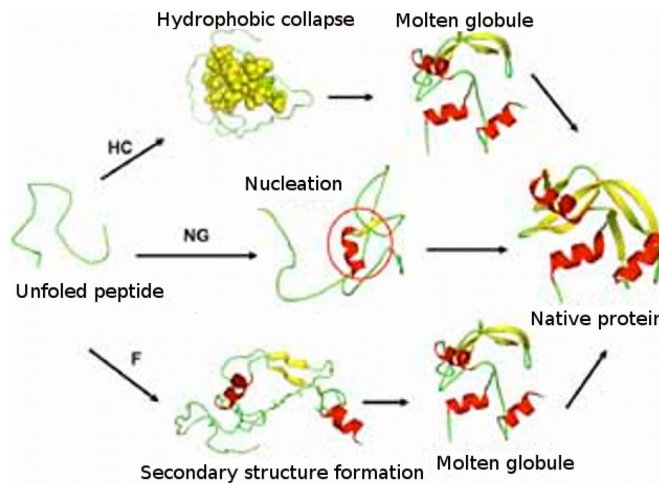
In order to celebrate the journal's 125<sup>th</sup> anniversary in 2005, Science Magazine published a special collection of articles about 125 “big questions”<sup>47</sup> that face scientific inquiry over the next quarter-century. On the list there are two questions regarding protein structures: “Can we predict how proteins will fold?” and “How do proteins find their partners?”.

Most of the presented research focused on globular proteins, which can fold directly in solution. However, fibrous proteins and membrane proteins remain a challenge due to the difficulty in determining their conformation in an isolated formation.

The famous experiment of renaturing of RNase A by Anfinsen<sup>48</sup> in 1961 clearly demonstrated the capability of folding of protein sequence without help from other bio-activity factors using solvent only.

There are various competing models proposed for protein folding process:





*Fig. 3: Protein folding models.*<sup>(127)</sup>

The hydrophobic-collapse model<sup>49,50</sup> hypothesizes that a protein collapses rapidly due to the sequestration of the hydrophobic side chains from the surrounding water. This collapsed intermediate is also referred to as a molten globule. This hydrophobic collapse is a relatively early event, occurring before the formation of secondary structure.

The nucleation condensation model states that first a local nucleus of folding is formed, then this core acts as scaffold for rapid tertiary structure organization<sup>51</sup>.

The diffusion-collision model proposes that local microdomains (secondary structure or hydrophobic clusters) form independently of tertiary structure<sup>52</sup>. These elements diffuse until they collide, successfully adhering and coalescing to give the tertiary structure<sup>53</sup>.

Proteins are synthesized by ribosomes inside cells and the process takes minutes (single domain) to hours (multidomain). The folding of the N-terminal part of the protein

happens before the completeness of assembly of whole protein sequence. Experiments show the folding unit of a protein is single domain, which is a relatively independent, conserved part of the protein sequence, the assembly process does not move smoothly, and some delay<sup>54</sup> might occur to signal the finishing of one domain to facilitate the maturing of folding for sequence that has already assembled. Since parts of the protein start folding while the assembly still continues elsewhere, the hydrophobic-collapse model might not be the initial process, in which case the protein won't wait until the ending of one assembly to start to fold elsewhere. It is highly likely that some secondary structure forms when the sequence is long enough to form these motifs, and rearrangements through the hydrophobic effect has been happening all the time while the protein is synthesized. As for the nucleation condensation model, it can be seen as a special case of the diffusion-collision model, if there is only one strong microdomain existing, and this domain acts as a nucleus to dominate the rest of folding process.

The chemical synthesis of proteins is one way to acquire small proteins. A similar environment can be cautiously chosen to assure the correct folding of the protein, even though the speed of folding might be hampered without the help from native cell machinery, because all information needed have been already embedded in sequence codes.

Protein folding and stability must be favored by some environmental factors:

temperature, pressure and composition of solvent. Water molecules acting as solvent molecules that are directly in contact with proteins are thought to play a dominant role in determining protein structure and stability. Any other factors have to exert their influence

through water molecules. This triggered our focus on the relationship between water molecules and proteins.

## 2.2 Water Bridges and Water Networks

Water plays an important role in biochemical systems, from DNA and proteins<sup>55</sup> to organic polymers<sup>56</sup>. However, the mechanism of how water molecules affect the folding of proteins or the binding of ligands to proteins is not well understood. Two opinions are actively discussed regarding which of the two dominant protein folding factors matters most: hydrophobic (*HΦO*) or hydrophilic (*HΦI*) effect.

The hydrophobic effect was first introduced by Kauzmann<sup>57</sup> who pointed out that 'hydrophobic bonds' (hydrophobic solvation of non-polar solutes) play a key role in stabilizing protein structure. The source of this interaction comes from a small increase of  $\Delta H$  (through the breaking of old clathrate cages and formation of one big clathrate cage) and a large increase of  $\Delta S$  (closer proximity of two hydrophobes). Fig.4 illustrates the thermodynamics of the hydration process for non-polar hydrophobes. The mixing of hydrophobes and water molecules is not spontaneous, as water hydrogen bonds are broken to make room for the hydrophobes and heat is put into the system makes small increase of enthalpy, while the newly formed ice-like clathrate cage makes the system more structured with an decrease of the total entropy of the system. The coherence of two hydrophobes is a spontaneous process of hydrophobic interaction. When hydrophobes interact with each other, the enthalpy increases as more hydrogen bonds form between

water molecules. Tearing down two structured clathrate cages to form one bigger cage causes the entropy to increase (as it becomes less structured). The clustering of hydrophobic groups results in a negative contribution to the change in entropy of the system, however we must not forget the role of solvent.

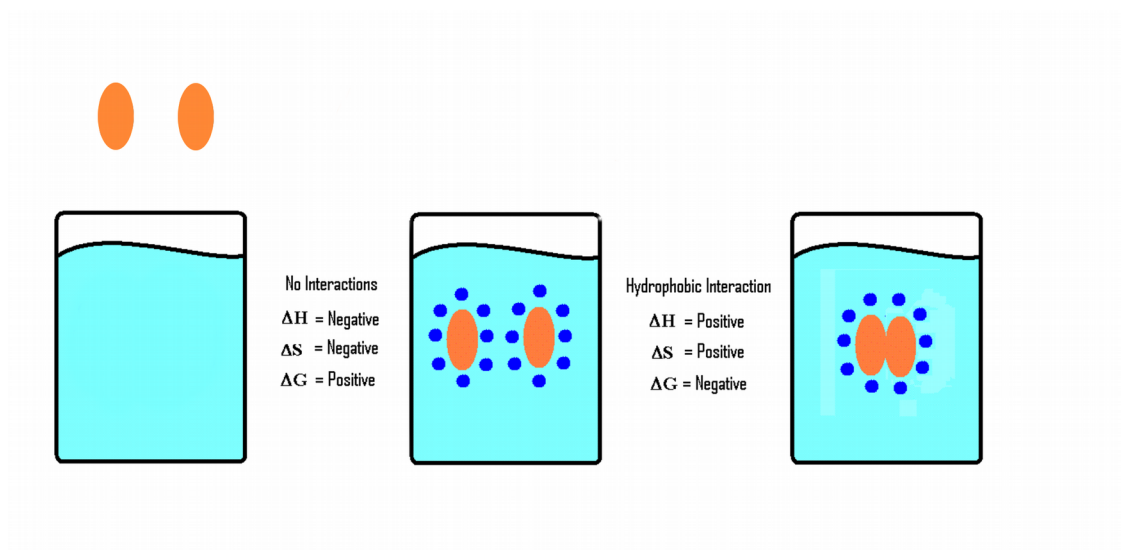
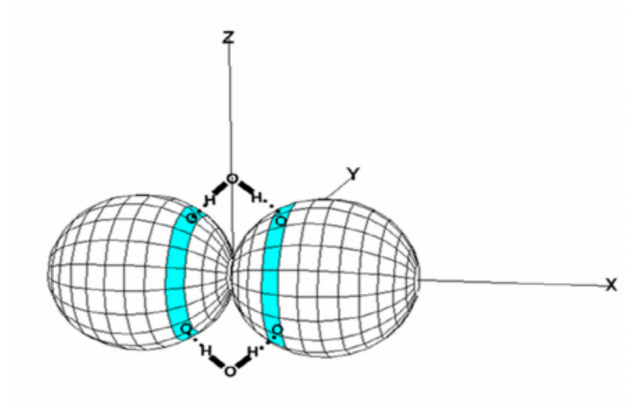


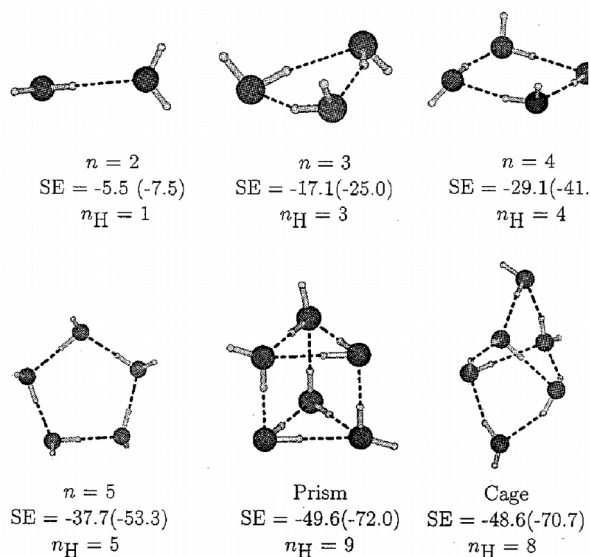
Fig. 4: Thermodynamics of hydrophobic solvation. From LibreTexts.



*Fig. 5: Schematic of one water bridge in protein aggregation. From Arieh Ben-Naim.*

This idea was challenged by Arieh Ben-Naim<sup>58</sup> who is an advocate of the hydrophilic effect regarding protein folding and protein association. He pointed out that the conditional solvation Gibbs energy, rather than Gibbs energy of solvation is the driving force of protein folding, and this varies over one or two orders of magnitude. It is water bridges (-11kJ/mol, per one water bridge) that bring two large globular proteins together (Fig. 5), rather than any hydrophobic effects, since ~70% of the interfacial residues of protein complexes are hydrophilic.

Before investigating the interaction between a water network and its solutes, we need to know the characteristics of a pure water network cluster. Many recent studies have been carried out to investigate the shapes and dynamics of water clusters<sup>59-61</sup> experimentally and computationally.



*Fig. 6: Representations of the most stable geometries for  $(H_2O)_n$  clusters.<sup>(59)</sup>*

Fig.6 is a representation of some small clusters<sup>60</sup>. These small clusters, especially the tetrameric or pentameric rings, are the building blocks of large clusters. The existence of water networks are rather more in-homogeneous than they first appear. Rao<sup>62</sup> found through complex network analysis that the conformation-space of a water network is composed of substructures, These meta-stable states are spatially correlated in two solvation shells, and time correlated in 200~400 ps (the transition time between these meta-stable states are 200~400 ps). This agrees with idea of Ben Ishai<sup>63</sup> who found that water clusters are always in dynamic fluctuation where a few units drift away to break up the existing cluster and leading to formation of new clusters which might contain the core units of old ones.

The formation of hydrogen-bonded networks in a water medium is influenced by solutes such as protons, ions, or proteins in solution. Tuckerman<sup>64</sup> discovered the way  $\text{H}_3\text{O}^+$  serves to allow ion transfer between  $\text{H}_5\text{O}_2^+$  and  $\text{H}_9\text{O}_4^+$ , and this proton transfer is related to the dynamics of the second solvation shell, which is hydrogen bonded to ligand water molecules. The average hydrogen bond life time is about 1ps<sup>65</sup>, the life time of a water network should be smaller than this value, reference<sup>66</sup> reports  $\sim 50\text{fs}$ .

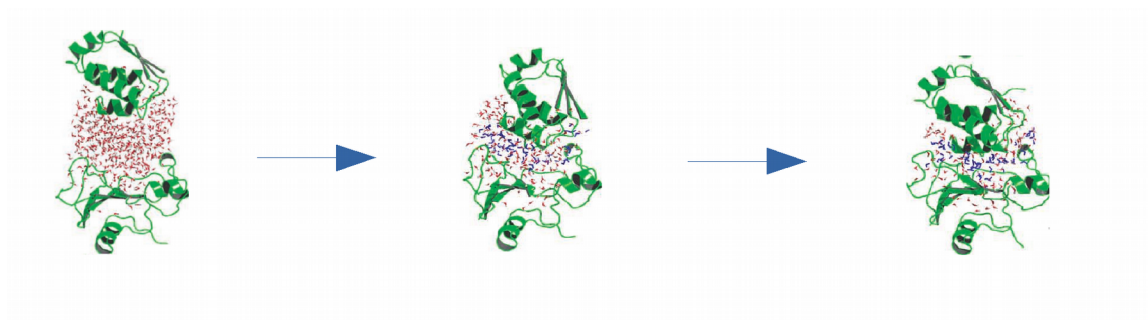
Ions may be classified into two categories: structure-making and structure-breaking species. The definition depends upon the extent of interactions that these ions make with the surrounding water molecules<sup>67</sup>. Some halides could form weak hydrogen bonds with water molecules, thus prolonging the lifetime of hydrogen bonds up to 2.6-25 ps, which is multiples of the lifetime in bulk water. Urbic<sup>68</sup> further found that when monoions are attached to donor water molecules, the strength of hydrogen bond increases, otherwise when monoions are attached to acceptor water molecules, the strength of hydrogen bond decreases. Ions may only influence the strength of hydrogen bonds in the first solvation shell, but they may have influence on water molecules within a few of hydration shells through the influence of the cooperative effect of water molecules and the electric fields exerted by ions. Obrien<sup>69</sup> pointed out that monovalent ions do not affect water structure beyond the first shell, and that ion-water interactions only extend beyond the first shell where both ion and counterion are strongly hydrated (i.e.,  $\text{MgSO}_4$ ).

Proteins being slow, large-scale motion molecules, they not only affect vicinal water molecules which hinder hydrogen bond switching, but also have a long-range effect on

water structures through functional groups<sup>70,71</sup>. In addition to the effects of solutes on water structure, water molecules influence solutes as well.

Sreerama<sup>72</sup> investigated a network of water molecules around an alanine octapeptide in water, and discovered that water bridges help stabilize the P(II) conformation relative to the beta conformation, although this was challenged by Peter<sup>73</sup>. The role of water bridges on the propensity of left-handed polyproline conformation was discussed and contradictory results was found to oppose Sreerama's results.

Ahmad<sup>74</sup> performed extensive molecular dynamics simulations of the association of hydrophilic interfaces of proteins, and discovered the formation of an adhesive hydrogen-bond network between the hydrophilic protein interfaces, which stabilized the early intermediate states before final native states (shown in Fig.7). The water molecules that belong to the hydrogen-bonded hydration shells of both proteins are colored blue.

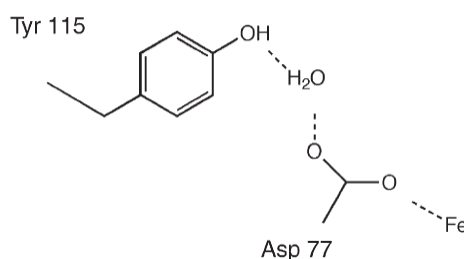


*Fig. 7: Water bridges (colored in blue) assist protein binding.*<sup>(74)</sup>



The water network acts not only as “glue” to bring protein together, but also as a media for electron transfer (ET) between proteins. This is a cornerstone of biological energy transfer. Aurelien<sup>75</sup> discovered that certain protein surface residues organize the solvent structure through water bridges which enhance the ET coupling strength, and a mutation of these residues break this water bridges, resulting in a reduced ET coupling.

Water bridges on the surface of DNA are also reported. Yoshiteru<sup>76</sup> calculate the lifetime of water bridges in the DNA minor groove. He discovered that the water bridge lifetime varies from 1 ~ 300 ps, and depend on two factors: i) hydrogen bond pattern and ii) DNA structural fluctuations. This work also mentioned the electrostatic nature of the surface has no significant influence on the lifetimes of water bridges based on other groups research<sup>54,78</sup>.



*Fig. 8: Water binding in ribonucleotide reductase from C. ammoniagenes.*<sup>(56)</sup>

Josefin<sup>56</sup> found the existence of a water bridge between two functional groups in one re-crystallized organic compound (Fig. 8), and this water bridge has the same effect on electron transfer as direct hydrogen bonding does.

### **2.3 Hypothesis**

Taking all the background information into account, we speculate that as water molecules are the predominant molecules surrounding protein, they must be involved in protein folding, and are related to protein stability and functionality in some fundamental way. As to the role of the hydrophobic or hydrophilic effect in these processes, we are more prone to believe that the hydrophilic interaction is the driving force for protein folding.

However, we must acknowledge the role of hydrophobic effect on protein folding and stability. The hydrophobic effects are over emphasized, and we believe the hydrophobic effect does no more than just arrange the side chain to accommodate the water network. Thus the specificity of 3D structure of protein must be determined by the protein side chains. The overall conjecture is that the hydrophilic interaction folds the protein, but the hydrophobic effect shapes the protein.

Though we investigated the role of water bridges on the protein folding and stability, we are not aiming to predict the exact process of protein folding. We are simply focusing on to what extent water bridges influence and direct this process. To that end we proposed a model which weights hydrogen bonds and water bridges in order to qualitatively account for the thermodynamics of protein folding.

## 2.4 Material and Tools

In this research, four MD simulation packages were used for different purposes: CHARMM C37b2 was used for water network and constant pH simulation; AMBER14 was used for the water bridge and constant pH simulation of proteins and lipids, GROMACS5 was used for the lipid aggregation simulations; and NAMD2.0 was used for steered molecular dynamics (SMD) simulations. Details of the simulation parameters will be given at the beginning of corresponding sections. All scripts to run and analyze the simulations in this thesis will be printed in italic format and double quoted, and can be found at: <http://www.weizhang.us/toolset/>.

We chose TIP3P water model for its simplicity and popularity, generally with the NPT ensemble, using periodic conditions. Particle Mesh Ewald (PME) was turned on for explicit solvent simulations. The temperature of all simulations was 300K and pressure was 1atm unless otherwise specified.

The visual tool to view structure and trajectory of simulations was VMD. The script to calculate water bridges was written in TCL, embedded in VMD. We used R, gnuplot and xmgrace to analyze data and plot the results.

In order to compile water bridge information, we chose a target protein list from the dynamomics server (<http://www.dynamomics.org/>). These molecular simulations of explicitly-solvated proteins were performed with Amber14 using ff14SB forcefield. Ff14SB is a continuing evolution of ff99SB forcefield. The key improvement is minimization of dependence of side chain parameters on particular backbone

conformations. This is crucial to our research as we will look into water bridges among side chains.

## **2.5 Analysis**

### **2.5.1 Analysis of Hydrogen Bond**

The hydrogen bond calculations for pure water simulation was performed by “*water\_bridge.tcl*”. The output from this script were further processed in order to analyze shapes and dynamics of water hydrogen bond network with the complex network analysis tool (igraph) in the program R.

To avoid missing information during the dynamic calculations of water networks, we set the output frequency of coordinates to a trajectory file to be 1, in other words, all coordinates of water molecules were saved every step (0.002 ps). We then extracted the last 10000 steps for detailed analysis. The size of the water box used for the simulation is  $60\text{\AA} \times 60\text{\AA} \times 60\text{\AA}$ , containing 9261 TIP3 water molecules.

### **Degree of Hydrogen Bonding**

Water consists of enormous clusters with different degrees of hydrogen bonding in equilibrium. We consider the degree of hydrogen bonding (for each water molecule, how many other water molecules are connected) as a factor in retaining the characteristics of hydrogen bonding network and should be verified with cautiousness. The purpose of this

analysis is to tune the parameters for hydrogen bond definition. Several different geometric, energetic and combined definition of hydrogen bonding have been tested<sup>79</sup>.

Here we define hydrogen bonds with the following geometric criteria (Fig. 9).  $R_{oo}$  is the bond length between donor and acceptor. The intra-molecular angle  $\Phi_{oo}$  is the obtuse angle of donor-hydrogen-acceptor.

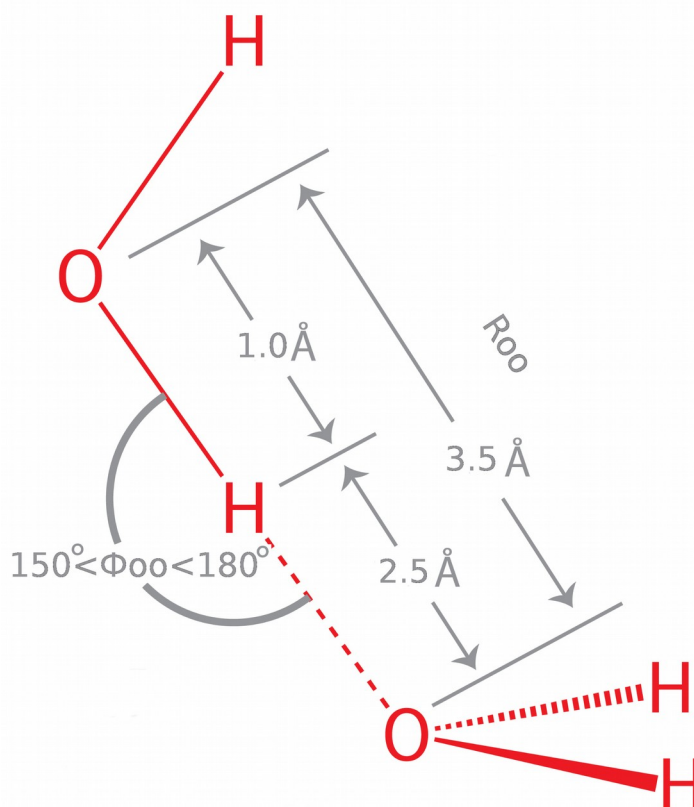
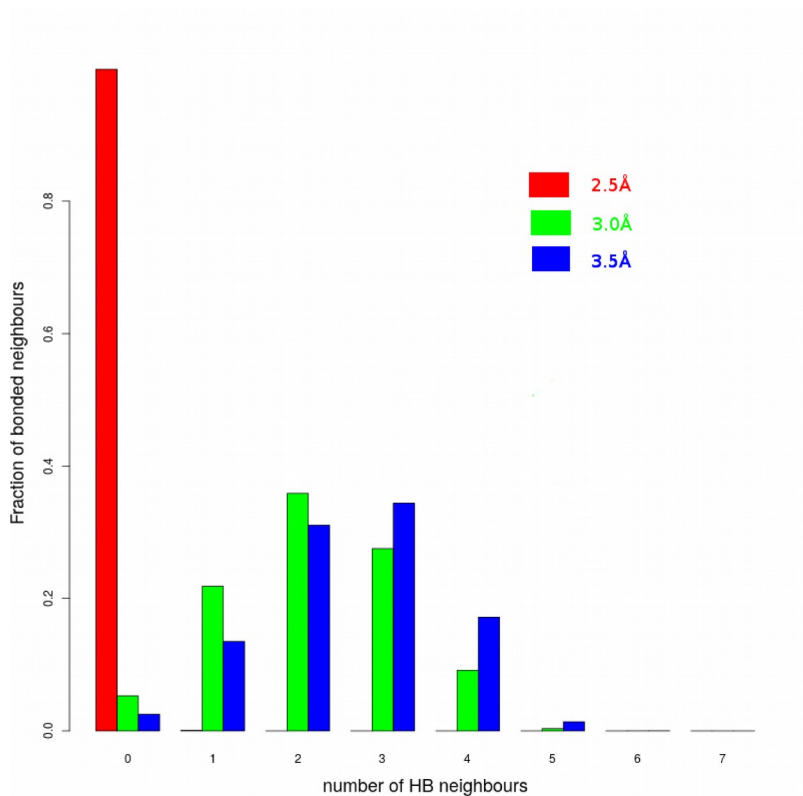


Fig. 9: Definition of hydrogen bond.

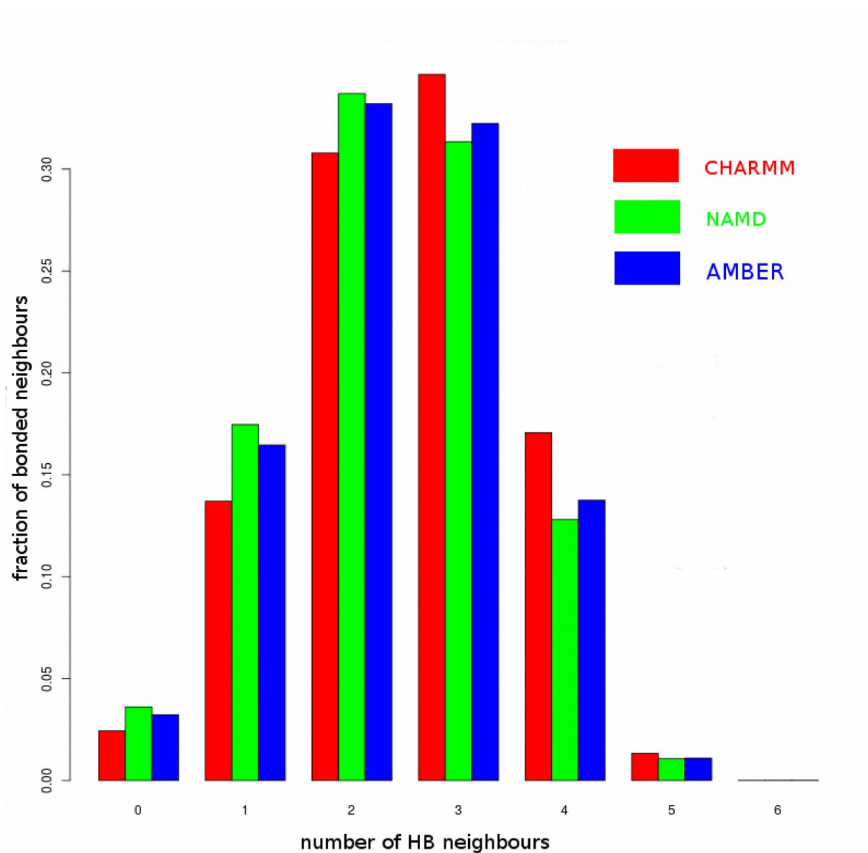


*Fig. 10: Degree of hydrogen bonding in pure water, defined by CHARMM.*

To determine appropriate  $R_{OO}$ , we set  $R_{OO}$  to be 2.5Å, 3.0Å, 3.5Å, and  $150^\circ < \Phi_{OO} < 180^\circ$ , and calculated average degree of hydrogen bonding, and compared these with reference values. The pertinent script is “*deg.R*”. Fig. 10 is the calculated degree of hydrogen bonding follows different value of  $R_{OO}$ . The bond length of 3.5Å has highest population at 3. This is in agreement with the literature value of  $\sim 3$ <sup>80,81</sup> of average degree of hydrogen bonding. Therefore, in this research in all calculation regarding hydrogen bond, we define  $R_{OO} < 3.5\text{Å}$ , and  $150^\circ < \Phi_{OO} < 180^\circ$ . It is worth to mention that the default value for  $R_{OO}$

in vmd is  $3.0\text{\AA}$  which we think is too restrictive in order to study water networks. It may be appropriate for stronger intra-molecular hydrogen bond on proteins, however.

To verify the agreement among the different software packages, we ran a pure water simulation using CHARMM, AMBER and NAMD, and choosing the same condition parameters, Fig. 11 shows that the calculated degree of hydrogen bonding were similar to each other, though minor discrepancy exists. The most probable degree of hydrogen bonding in CHARMM is 3, but both AMBER and NAMD give a most probable of 2. This means if we apply the definition of hydrogen bond to water network calculation in NAMD and AMBER, we might need to adjust the parameters to be more lenient. In this research, we maintain the same hydrogen bond definition amongst all simulation softwares.



*Fig. 11: Degree of hydrogen bonding in pure water, defined by CHARMM, NAMD and AMBER.*

### Hydrogen bond lifetime and residence time

A key factor for hydrogen bond dynamics is the hydrogen bond “lifetime”. One might see some terms related to this time scale such as “hydrogen bond life time” or “hydrogen bond residence time”, with various and sometimes conflicting definitions. Here we refer to the hydrogen bond life time  $t_{HB}$  as the time that the hydrogen bond remains intact at all



time up to time  $t$ , from initial time  $t_0$ . In contrast we define hydrogen bond residence time  $t_R$  as the time that certain hydrogen bond present at time  $t$  from initial time  $t_0$  which is independent of any possible breaking before time  $t$ , thus  $t_R > t_{HB}$ . The relaxation time for the water dipole reorientation at room temperature is about 10 ps, and the average hydrogen bond lifetime should roughly agree with this value as reorientation is impossible without breaking at least one hydrogen bond. We plot the survival probability of hydrogen bond lifetime and residence time which calculated following equation 2.1.

$$P(t) = \frac{\sum h(t)}{\sum h(t_0)} \quad (2.1)$$

$\sum h(t)$  is the frequency of hydrogen bonds from time  $t_0$  at time  $t$ . When  $t=t_0$ ,  $P(t)=1$ .

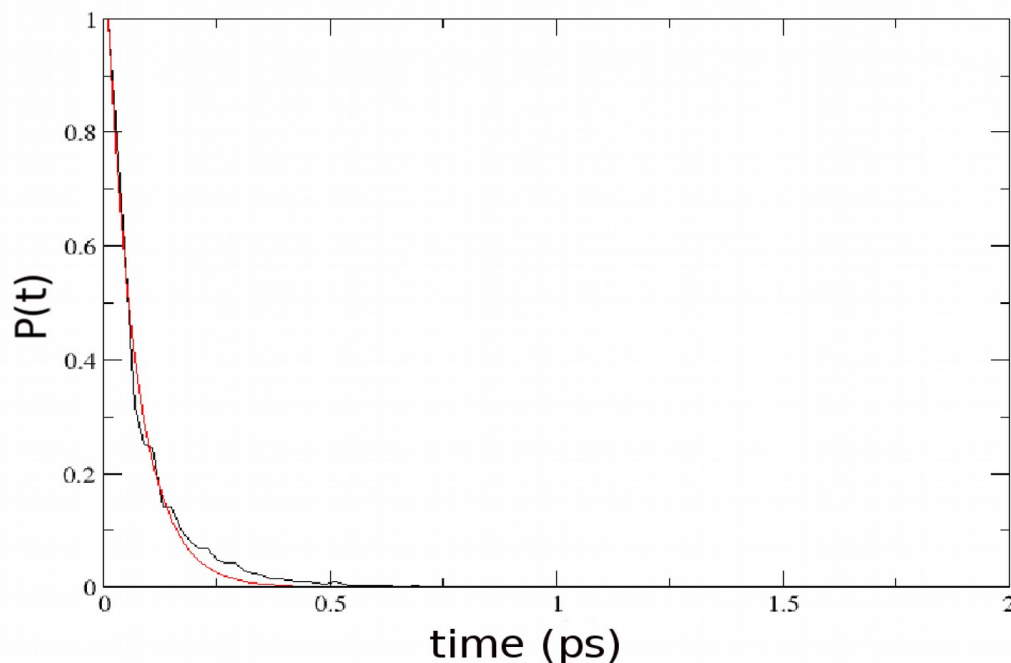


Fig. 12: Survival probability of hydrogen bond lifetime. Red curve is data fitting to  $Ae^{(-t/\tau)}$ .

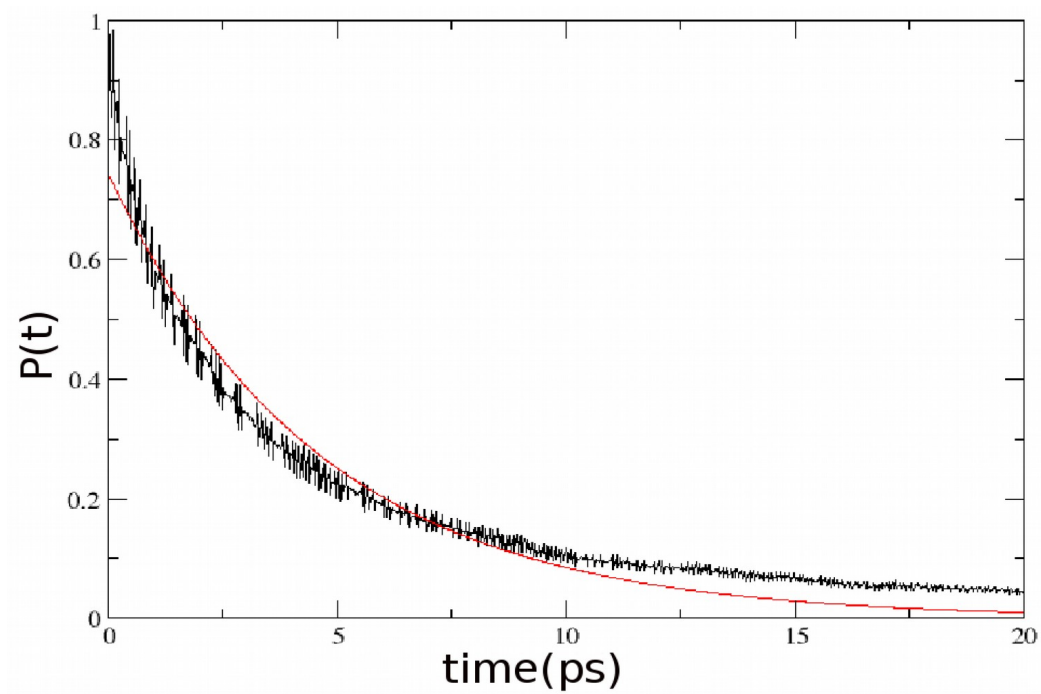


Fig. 13: Survival probability of hydrogen bond residence time. Red curve is data fitting to  $Ae^{-t/\tau}$ .

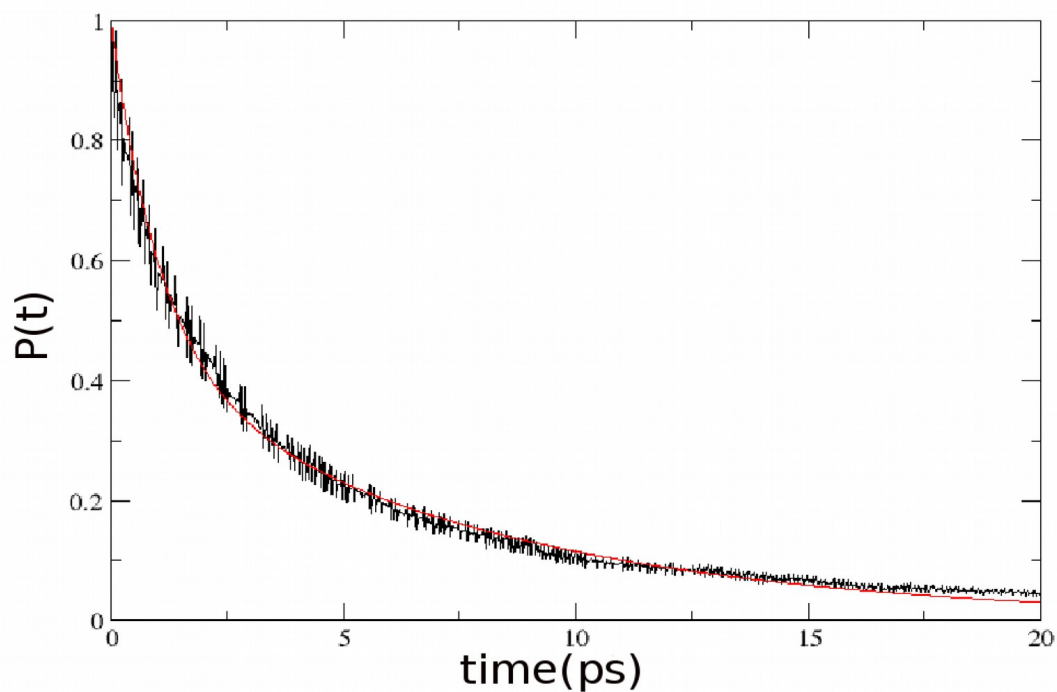


Fig. 14: Survival probability of hydrogen bond residence time. Red curve is data fitting to  $Ae^{(-t/\tau_1)} + (1-A)e^{(-t/\tau_2)}$ .

There are two ways to express lifetime, the first is the half-life method, which is the time required for a given quantity decrease to half of its initial value. The ‘half-life’ life time of a hydrogen bond is about 0.1 ps, and the ‘half-life’ residence time is about 2 ps.

Another way to express lifetime is by fitting decay curves to exponential functions with different characteristic time constants. The parameters of the fitting for various forms is given in Table 1.

Table. 1 Data fitting of function P(t) with exponential equations.

P(t)	A	$\tau_1$	$\tau_2$	Exponential function
Life time	1.07795	0.0578531		$Ae^{(-t/\tau_1)}$
Residence time	0.743134	4.6104		$Ae^{(-t/\tau_1)}$
Residence time	0.441324	7.44046	1.04851	$Ae^{(-t/\tau_1)} + (1-A)e^{(-t/\tau_2)}$

When both hydrogen bond life time and residence time were fitted to two parameter exponential function, hydrogen bond life time decay by a factor of  $e$  every  $\sim 0.06$  ps (or decays to  $e^{-1} \approx 37\%$  of its former value every  $\sim 0.06$  ps), and residence time show a decay by a factor of  $e$  every 4.6 ps (Fig. 12 and Fig. 13). As shown in Fig.13, there is a poor fitting of data points to the two parameter exponential function, but a good fit is found with the three parameter exponential function (Fig. 14). This implies that the compare to hydrogen bond life time which is only dominated by one factor, residence time is indeed dominated by two factors, one fast process and another slower process. The slow process is approximately 7 times slower than the fast one. Speculatively,  $\tau_1$  can be associated with the lifetime of diffusive translational motion and  $\tau_2$  with diffusive rotational motion<sup>82</sup>.

## 2.5.2 Analysis of Hydrogen Bond Network

### Method

As we mentioned in our hypothesis, any solutes in water will interact with water molecules in the form of “solute – water network”, hydrophilic solutes interact with polar

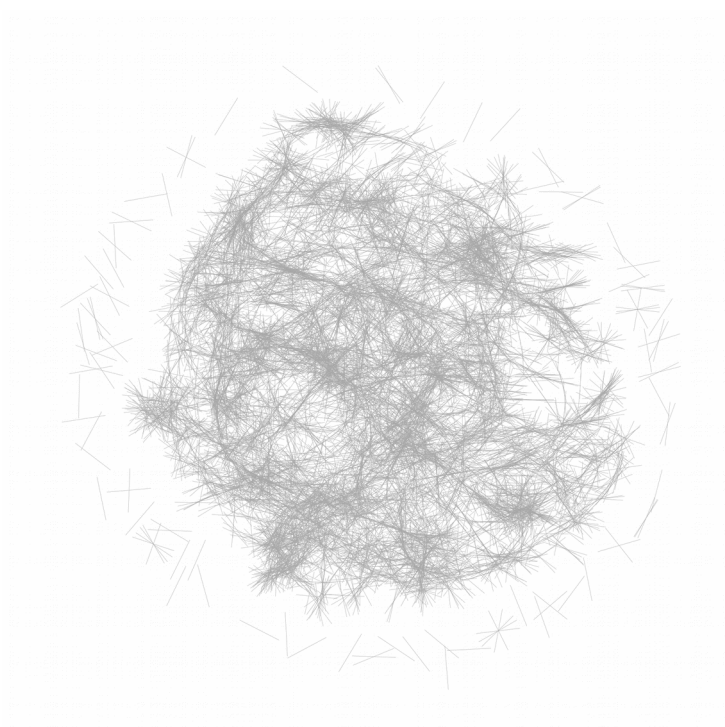
water molecules and be part of water network, while hydrophobic solutes try to avoid water network by wrapping around by a clathrate cage. Therefore study the property of hydrogen bond network in water is a key to understand the behavior of solutes in aqueous environment. There are many network analysis tools available, some examples are: Boost Graph Library, QuickGraph, igraph and NetworkX. We performed a water network clustering analysis using the igraph package in R<sup>81</sup>. To convert our data into an igraph network, we used the `graph_from_data_frame()` function, which takes two data frames: nodes and edges. Nodes are defined as indices of O atoms which form hydrogen bonds, edges are defined as hydrogen bond connection between O atoms.

For example: “18390–22659” represents a node of oxygen atom 18390 which is hydrogen bonded with another node of oxygen atom 22659. The edge between these two nodes is their hydrogen bond connection.

We performed a community detection calculation to study hierarchical structure of water network. There are many ways to cluster the hydrogen bond network. Here we chose the “fast greedy” method<sup>83</sup> to find dense subgraphs. This method assign each nodes as their own community initially, then tries to maximize a quality function called “modularity” in a greedy manner from removing nodes from its community and placing it in the community of its neighbor. Modularity is a measurement of strength of a network into modules (groups, clusters or communities), modules with high modularity have dense connections between nodes within modules but opposite between nodes in different

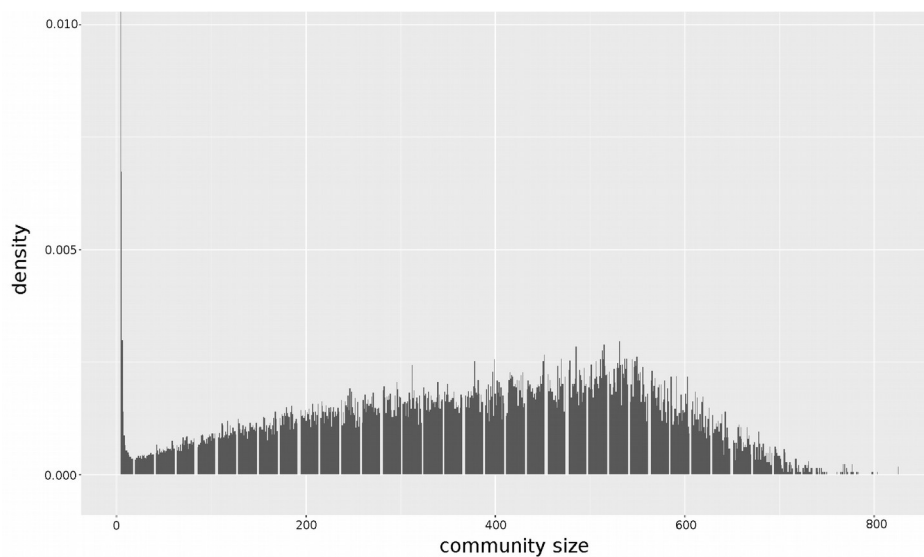
modules. It is fast and easy to start with, since it does not require any parameters to tune. The script for community detection is “*community.R*”.

Fig. 15 is a representation of water network in one frame from simulation. Almost all water molecules are tangled together, but we can tell there are some isolated small water clusters separated from large clusters. The average degree of hydrogen bonding for small water clusters is lower than those in larger clusters, which indicate they have a relatively lower density. We consider these water molecules as “decorating” water molecules, as they are free from the majority of water clusters.

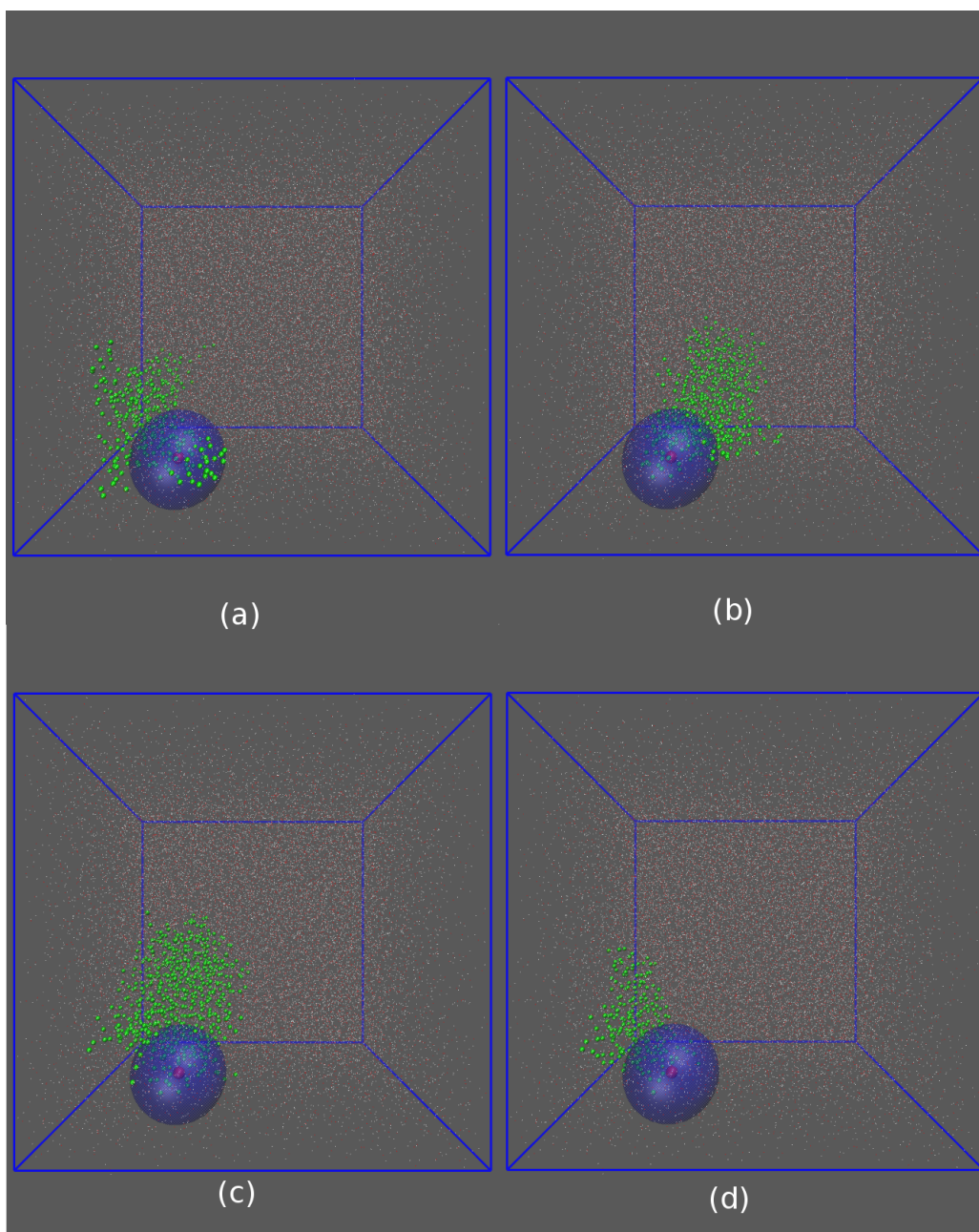


*Fig. 15: Topological representation of a water hydrogen bond network of pure water box (9261 water molecules). From igraph in R.*

Fig.16 shows the size distribution of water communities. The first peak with very high value is the number of communities with least water molecules (2 to 4), these are considered small clusters or “debris” which do not qualify as an actual communities. The density increases as the size of communities increase. The community size of 500 has relatively high density compared to other large size communities. The occurrence of the communities with >800 water molecules are very rare.



*Fig. 16: Community size distribution of a water hydrogen bond network.*

**Community Life Time**

*Fig. 17: Snap shots of four water communities (green spheres) involving one oxygen atom (red sphere) in four successive frames.*



After the detection of communities, they were rendered in VMD. It was found that the evolution time was fast, and the average lifetime of any one community was short (0.002ps). Fig. 17 shows an example of four successive frames from a simulation. Blue spheres represent neighboring groups which includes that specific oxygen atom. This neighboring group (within 5Å) is more conservative (~50%) and updates less frequently compared to the whole community (~20%). From observing the evolution of communities, we found that even though the lifetime of a specific community is short, the evolution of that community might involve the reforming of a new community by keeping some core units of the old community. In this way, we can view the movement of a community as a travelling wave through a medium.

### **2.5.3 Influence of Ions on Hydrogen Bond Network Properties**

Ions in water will obviously interact with any hydrogen bond network. In order to investigate this interaction, we calculated the degree of hydrogen bonding, and the life time of hydrogen bonds in water under different ion atmospheres, including cations of  $\text{Na}^+$ ,  $\text{Mg}^{2+}$  and  $\text{Al}^{3+}$ , and anion  $\text{Cl}^-$ . The results are presented in Fig.18 and Fig.19.

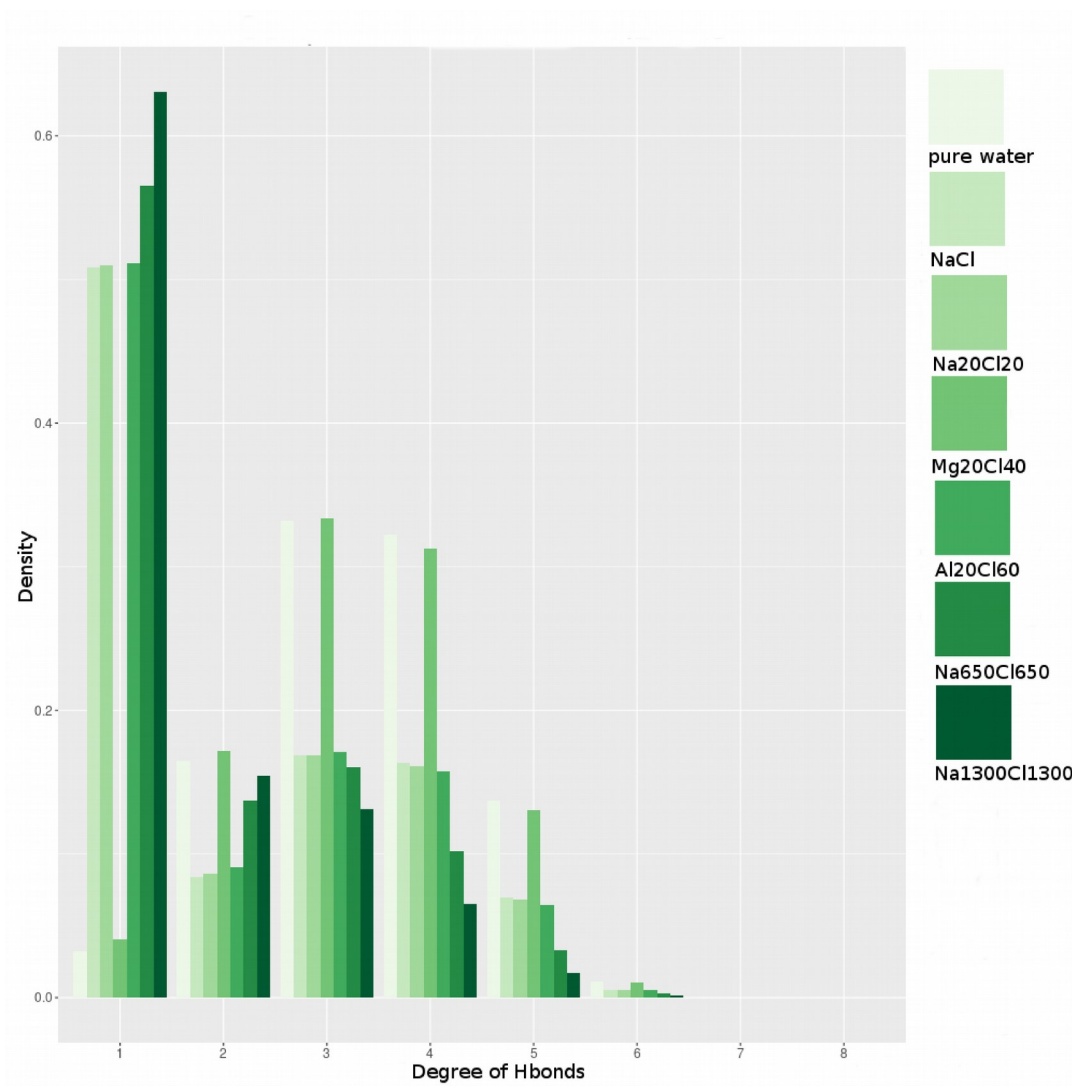
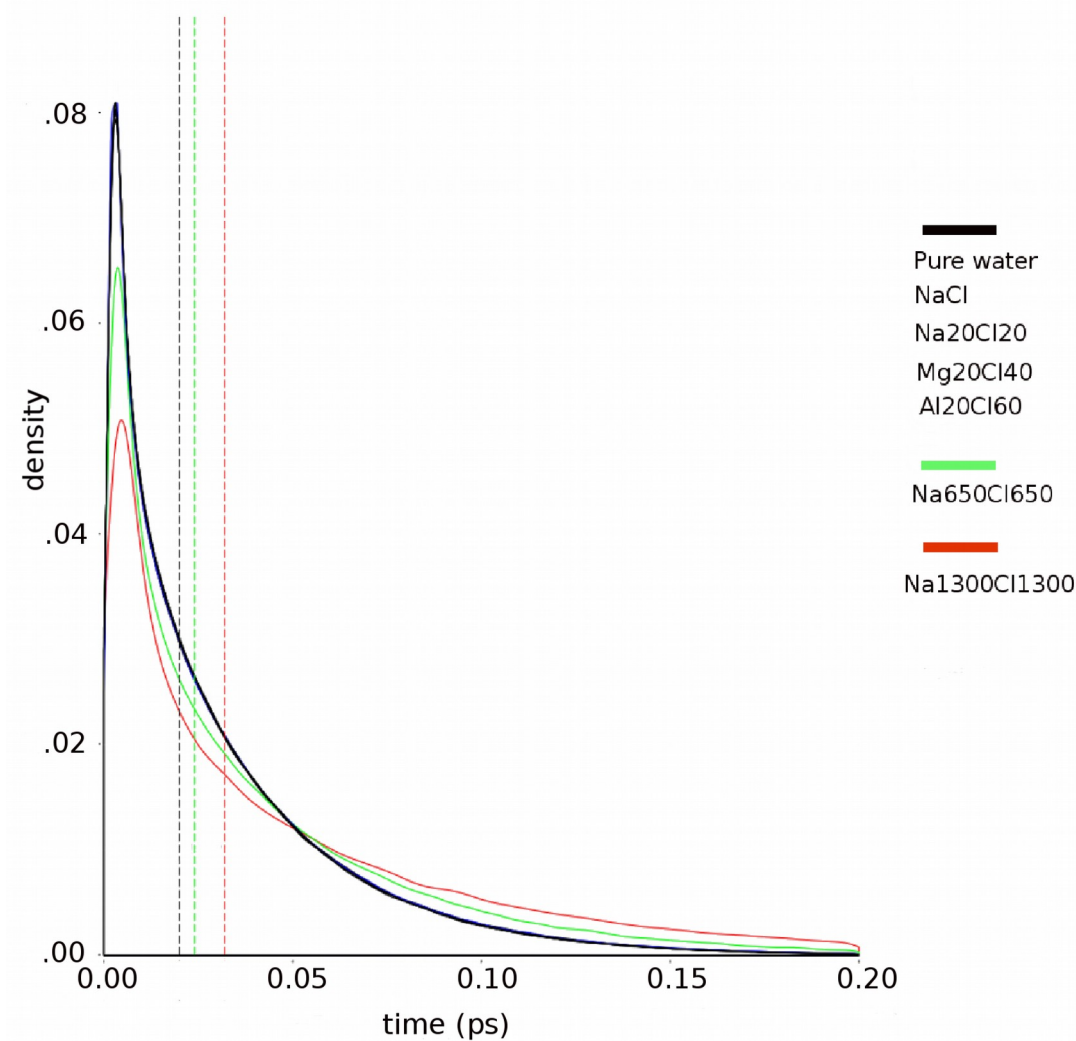


Fig. 18: Influence of ions on degree of hydrogen bonding.



*Fig. 19: Influence of ions on hydrogen bond lifetime. Solid curves are density distribution of hydrogen bond life time, dashed vertical lines are calculated average hydrogen bond life time.*

The systems start with a pure water box, then successively include a single NaCl (~0.007M NaCl), Na<sub>20</sub>Cl<sub>20</sub> (~0.13M NaCl), Mg<sub>20</sub>Cl<sub>40</sub> (~0.13M MgCl<sub>2</sub>), Al<sub>20</sub>Cl<sub>60</sub> (~0.13M

$\text{AlCl}_3$ ),  $\text{Na}_{650}\text{Cl}_{650}$  (~4.1M NaCl), and finally  $\text{Na}_{1300}\text{Cl}_{1300}$  (~8.2M NaCl). The subscript of the ions represents the number of that ions in water box.

Fig.18 shows as the ionic strength increases from left to right in each group, the degree of hydrogen bonding decreases (except for  $\text{MgCl}_2$ ). Ions in water act as a nucleation core with the neighboring water molecules, and this in turn destroys the inherent hydrogen bonding pattern in a bulk water network. Thus, the degree of hydrogen bonding decreases. Water molecules interacting with ions can be organized into different layers. The average hydrogen bonding in first hydration shell will surely deviate from the bulk pattern, but the second shell might exhibit either “positive hydrogen bonding” or “negative hydrogen bonding”. This means there might be more hydrogen bonding in second shell or less hydrogen bonding in second shell compared with the bulk. The results in Fig.18 were different from reference<sup>84</sup>. That paper grouped ions as structure-making, structure-breaking, and borderline ions.  $\text{Na}^+$  was classified to be a borderline ion, but according to our calculation, even the most dilute  $\text{Na}^+$  solution causes a different hydrogen bonding pattern than pure water. According to our results, it seems  $\text{Mg}^{2+}$  should be classified as a borderline ion, since the degree of hydrogen bonding is very close to neat water.

As the ionic strength increases, the distribution of the hydrogen bond lifetime shifts towards longer lifetimes, which means the existence of ions in solution extend the lifetime of hydrogen bonds. The curves in Fig.19 is the calculated hydrogen bond lifetime of all water molecules inside the simulation box, not just hydrogen bonds close to solutes, and that is the reason the differences of lifetime among different solutions are minor.

Fig.18 tells us the existence of ions at least partially destroy the hydrogen bond network. This should lead to freely-moving water molecules, thus less hydrogen bond lifetime. However Fig. 19 seems to indicate opposite results. We believe this is due to the fact that the existence of ions provide additional electric field in space which slows down the reorientation of water molecules within this space.

We also calculated water hydrogen bond network community with existence of these solutes (Fig. 20), and found that the existence of ions influences the community size distribution. In general, as the ionic strength increases, the number of water molecules in larger community decreases (400-600) while the number of water molecules in smaller communities (<400) increase. In one word, ions in solution break a larger water community into smaller communities, As shown in Fig.20, when the ionic strength is very high (Na1300Cl1300), the water network is totally destroyed.

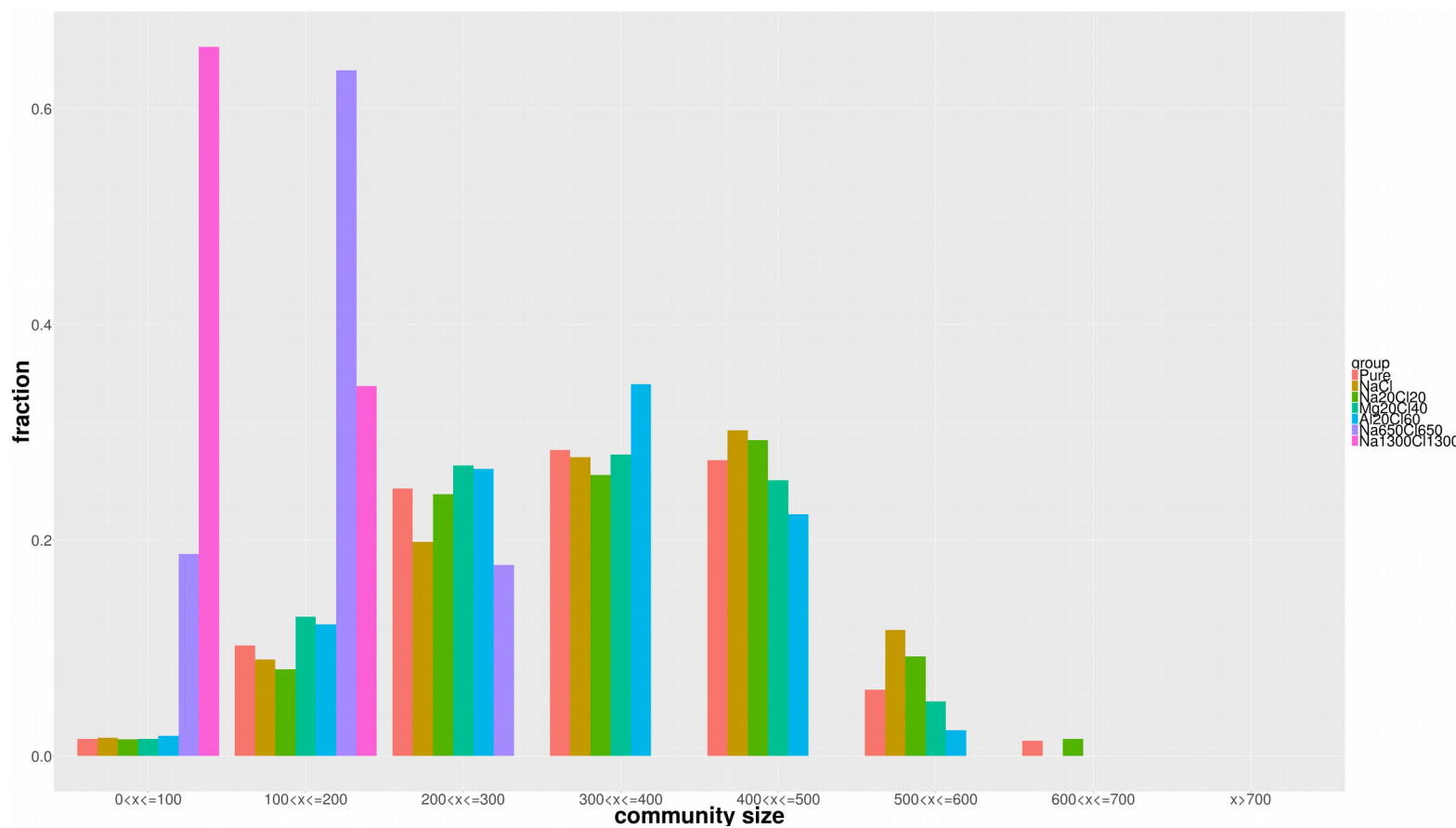


Fig. 20: Influence of ions on network community size.  $x$  label is community size classified into groups,  $y$  label is fraction of water molecules in that group.

#### **2.5.4 Analysis of Water Bridges**

As we mentioned in our hypothesis, we want to investigate the hydrophilic role in protein folding and stability. There are many interactions can be attributed to hydrophilic effects: charge-charge, charge-polar interaction and the hydrogen bond interaction. In this research we focus on the hydrogen bond interaction, since this interaction is fundamental in water solution, even though it is not as strong as charge-charge interaction. We started with all the hydrogen bond information that exists in a water box, and then looked for water bridges from one heavy atom to another heavy atom, which belongs to a different amino acid on one protein. We consider these water bridges as a long-distance hydrophilic interaction.

#### **Method**

We chose a particular target list from the Dynameomics website which contains 100 proteins. See Appendix A for details.

In order to prepare the structures, they were processed first by pdb4amber. Options included taking out any crystal waters, removing TER cards, choosing MODEL 1 if multiple models exist in PDB file. The terminal ends of protein were capped with an acetyl group (ACE) and a N-methylamide group (NME) at N-terminal and C-terminal ends respectively.

A steepest-descent energy minimization of the system was performed for 10,000 steps before switching to conjugated gradient minimization for 50,000 steps. Next, heating of the system was performed from 0K to 300K for 5,000 steps, and then an equilibration

over 5,000 to 30,000 steps. 1,000,000 steps (which equal to 2ps) of MD simulation was then performed at a temperature of 300K and pressure of 1atm. Intra-molecular bonds involving hydrogen were constrained using the SHAKE algorithm.

The minimization and heating procedures were performed by pmemd.MPI, and the MD simulation was performed by pmemd.cuda (The CUDA version of pmemd can not adjust the cell dimension automatically). The total simulation time on a server with 12-core Xeon E5 V3 and 3 GTX 980 graphic cards was ~1 week, and the size of data generated was ~700G.

The script to run simulation is “*waterbridge\_pipeline.sh*”.

### **Theory and Analysis**

In order to investigate the role of water bridges, a variable called the “bridge score” was defined in equation 2.2.

$$S_{bridge} = \sum w \cdot T_{bridge} \quad (2.2)$$

$S_{bridge}$  is the total bridge score.

$T_{bridge}$  is the bridge type.

0 water bridge which means a direct hydrogen bond; 1 water bridge; 2 water bridge and 3 water bridge.

$w$  is weighting factor for different bridge type.



Table 2 . Weighting factors for water bridge following strategy I and II.

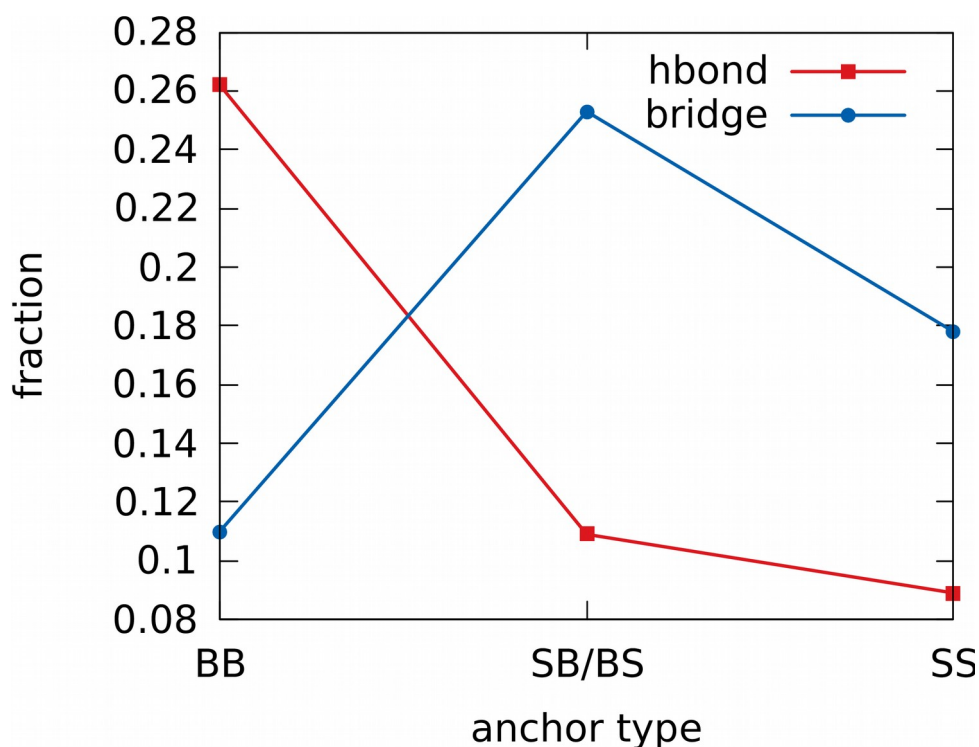
	$w_0$	$w_1$	$w_2$	$w_3$
Strategy I	1.0	1.0	1.0	10
Strategy II	1.0	0.5	0.25	0.125

The weighting factor for different bridge type is based on general Gibbs energy estimation. The contribution of direct hydrogen bonds to Gibbs free energy to be around -10.9kJ/mol, the contribution of indirect hydrogen bond through one water molecule is about -11.7kJ/mol<sup>85,86</sup>. We set 1.0 to direct hydrogen bond, but different values to water bridge. In strategy I, we set 1.0 to all types of water bridge which will maximize the influence of water bridge. In strategy II, we set 0.5 to one water bridge and 0.25 to two water bridge and 0.125 to three water bridge to minimized the influence of water bridge. The reason we cut in half as water increase is due to the fact that direct hydrogen bond lasts longer than water bridge. In this research, we used strategy I for most calculation unless specified. The eventual goal is to connect this bridge score to the thermodynamics of protein folding process. We consider bridge as the accommodation of protein to order of water network.

### **Property of Bridge Anchor Atoms**

In order to study the accommodation of protein to water network, we need to first investigate where these bridges located in protein. The characteristics about bridge sites in protein help unveil protein folding process in sense of point of action.

We define anchor atoms as heavy atoms on the protein backbone or side-chain, which were bridged (hydrogen bonded) through water molecules. Statistics about these anchor atoms help us to understand where and how often these bridges appear.



*Fig. 21: Fraction of direct hydrogen bond and bridge of each anchor atom type. BB: backbone bridge to backbone; SS: side chain bridge to side chain. SB/BS: side chain to backbone.*

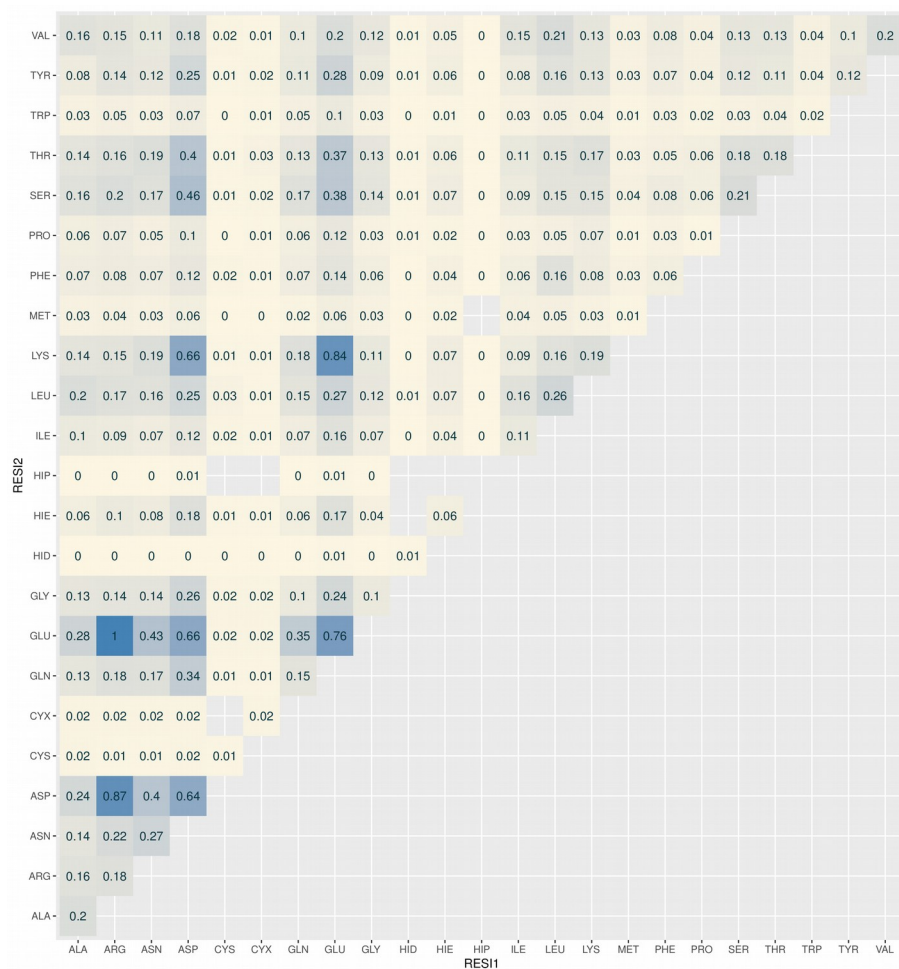
Fig.21 shows the frequency of direct hydrogen bonds appear on backbone-backbone is higher than mixed type of backbone side chain and side chain to side chain. However, water bridges show a significantly different pattern, water bridges happen in backbone-

side chain more often than side chain-side chain, and least likely happen in backbone-backbone. The ratio of frequency of hydrogen bonds to bridges is 0.46:0.54. It is more likely that a backbone try to “touch” another backbone with a direct hydrogen bond, while a side chain tries to “contact” another side chain or backbone through a water bridge. In determining a hydrophilic interaction, the backbone carries more weight than a side chain, since every residue has at least three atoms: carbonyl oxygen (strong), amide nitrogen (strong) and alpha carbon (weak) to be able to form hydrogen bond.

### **Protein Residue Bridge Contact Map**

In the field of structural proteomics, structure prediction have shown to be promising in solving ab initio protein structure prediction. Protein residue contacts is one of the methods used to study this prediction problem<sup>87-91</sup>. Inspired by Tsai<sup>92,93</sup>, we defined a “bridge contact”, which is different from geometric distance based contact, in that it is a hydrophilic interaction contact.

We calculated bridge scores for all amino acid pairs, divided by the highest score which is from glutamic acid and arginine, then plot the matrix as a map (Fig.22). The number in the map is actually the strength of hydrophilic interaction compared to the strongest GLU-ARG.



*Fig. 22: Bridge contact map of 20 amino acid residues. The number is hydrophilic interaction strength (from bridge score) with respect to GLU-ARG (defined as one).*

Arginine is positively charged, and Glutamic acid is negatively charged, and their interaction is the strongest among all residue pairs. The next strongest one is Arginine

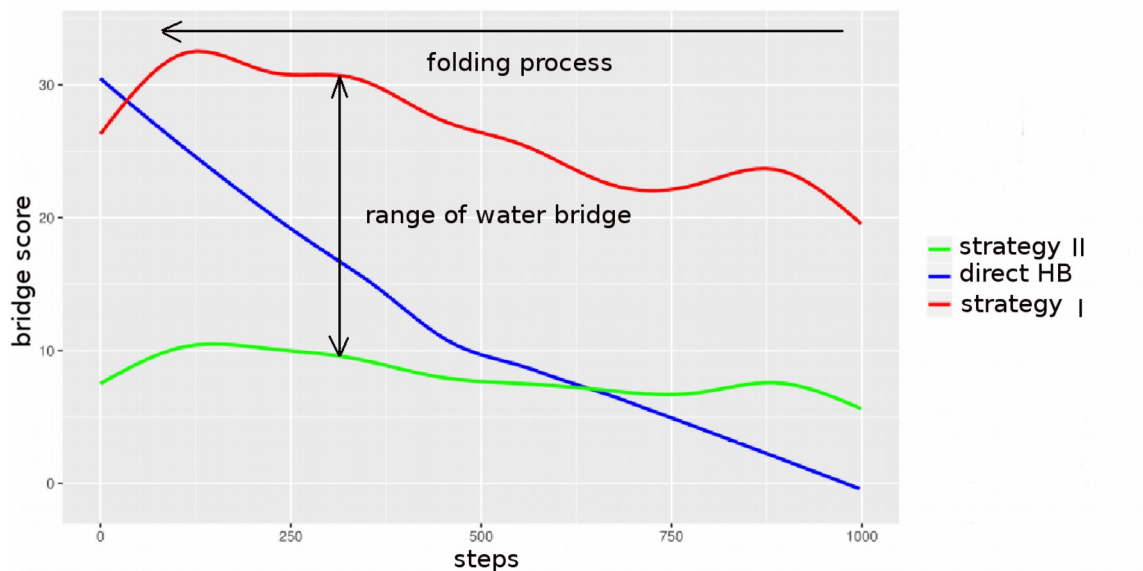
with Aspartic Acid. Essentially, the bridge score is high for oppositely charged amino acid residues, and lower for hydrophobic amino acid residues.

### **Bridge Score Along Protein Extension**

We performed steered molecular dynamics simulations using NAMD on BPTI with a constant velocity of  $10\text{\AA}/\text{ns}$  by fixing the CA at N-terminus, and pulling CA at C-terminus, and analyzed the bridge score along the trajectory of pulling in order to see the characteristics of the role of the water bridges on folded and unfolded protein. Fig.23 is four snapshots from the pulling of BPTI.



*Fig. 23: Snapshots of constant velocity pulling of BPTI. Snapshots from step 1 (top), step 250, step 500 and step 1000 (bottom). 1 step=0.002ps.*



*Fig. 24: Water bridge score along the unfolding of BPTI. Blue line is score from direct hydrogen bond, red line is score from water bridge following strategy I and green line is score from water bridge following strategy II.*

Fig.24 shows the contribution from direct hydrogen bond and water bridge as BPTI is unfolded. Here we show the two scores from the different scoring strategies. As BPTI was pulled, the score of direct hydrogen bonding decreases from about 30 to 0. However the score of water bridges following strategy I (red) only slowly drops from about 30 to 20, and the score of water bridges following strategy II (green) slowly drops from about 10 to 5. We believe the best bridge score should be in between these two curves. If we look at the curve from the right side to the left side as the folding process of BPTI, the contribution of water bridge to folding is higher than direct hydrogen bond at early stage (600 - 1000 steps). At the early stage of protein folding, residues are further away from

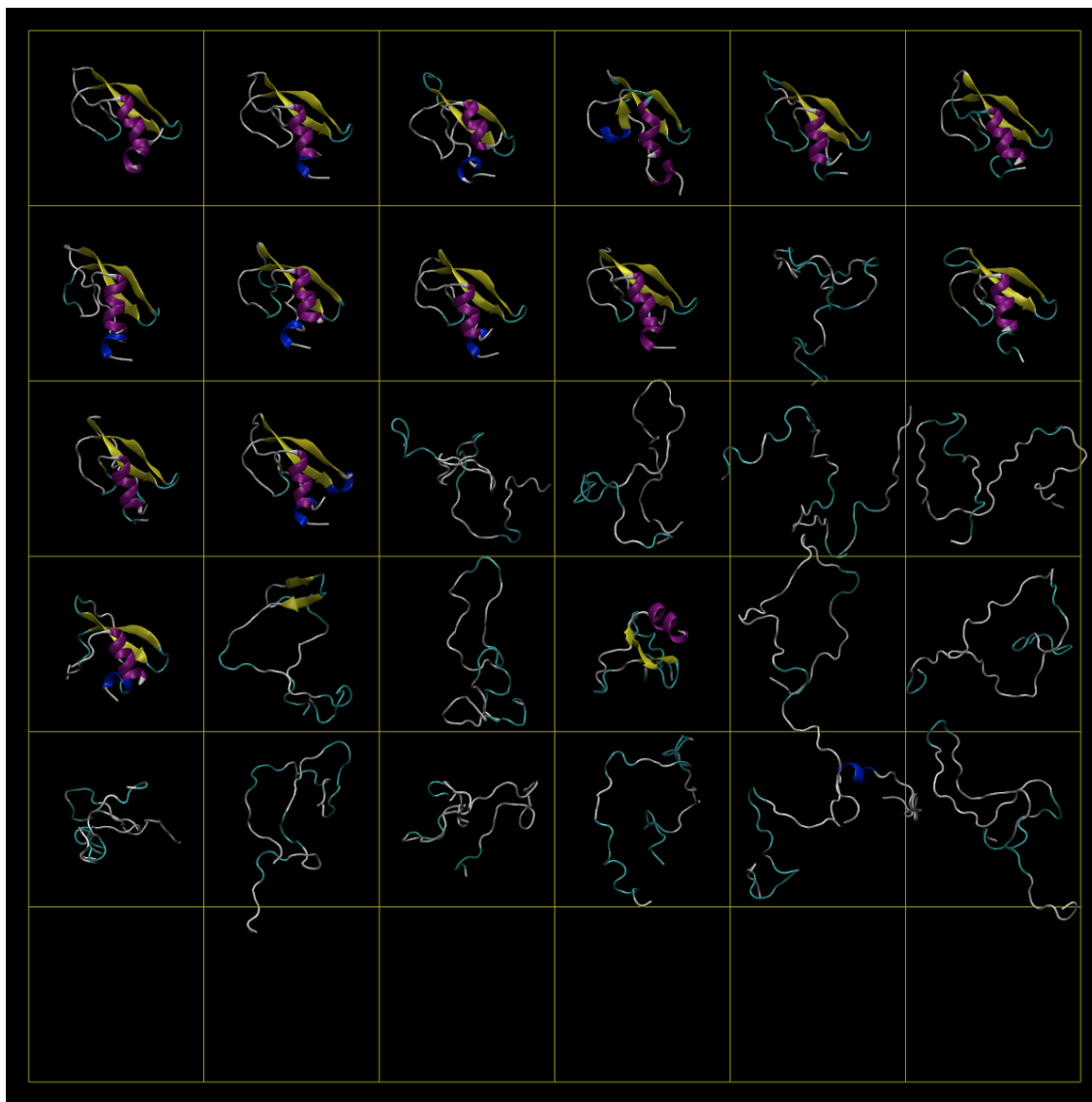
each other, and long distance water bridges help to pull the residues together. As the protein continues to fold, more and more direct hydrogen bonds help consolidate the protein, and this becomes the dominant factor for protein stability. Some amount of water bridges still exist in folded protein. When the conformation of the protein is close to a completely folded state or an extended state, the bridge score show a significant decrease. When the protein folding is accomplished, the forces from water bridges acting on the protein reaches a minimum level, so that protein can avoid the influence from water solvent. When the protein is completely extended, the forces from the water bridge are also at a minimum level, because the distance between hydrophilic sites are furthestmost, but after the initial bending of peptide chain from random movement, water bridges begin their role, and from then on are continuously acting on the folding of proteins.

### **Bridge Score of Misfolded Protein**

Apart from an extended protein, misfolded proteins are more important as they provide useful information when in comparison to native folded protein.

In order to simulate a misfolded protein, we performed a temperature replica exchange simulation of BPTI, the starting structures of all 30 replicas were native structures. The temperatures were distributed from 300K to 850K. The total number of exchange trials was 100 steps, and there were 1000 dynamic steps between exchanges.



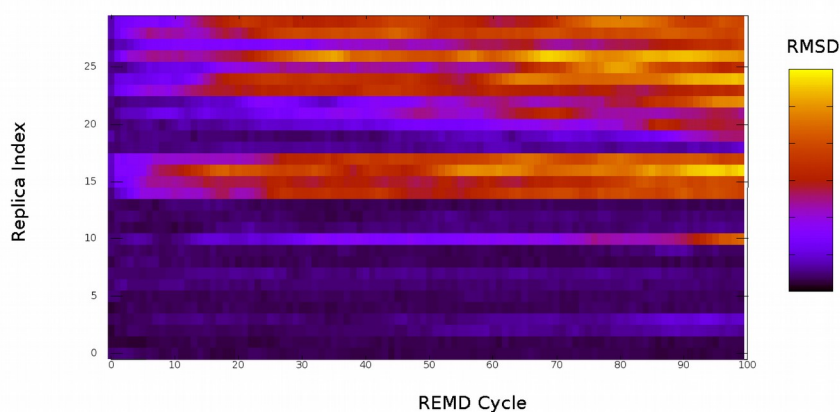


*Fig. 25: 30 replicas of BPTI after 100 exchanges.*

Fig. 25 shows 30 replicas (named aa1 to aa30) of final configurations after 100 exchanges. From top-left to bottom-right is replica aa1 to aa30. The reason initial replicas are all native structures rather than random coils is to achieve meta stable

structures that are close to native structure. About half of the replicas remained their native states, and the rest replicas unfolded and reached misfolded states.

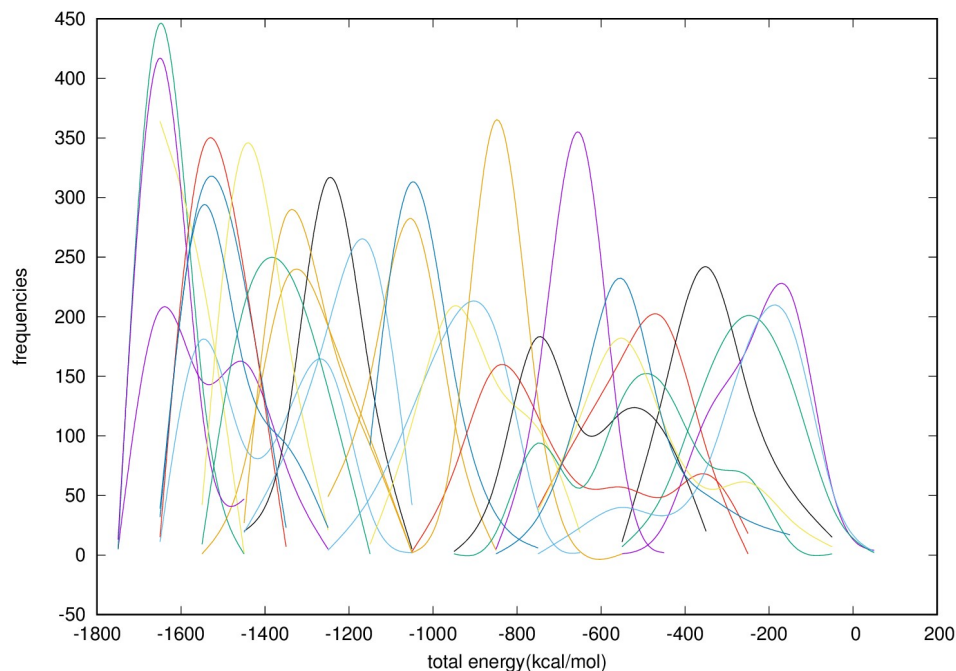
### RMSD of 30 Replicas of BPTI



*Fig. 26: Heat map of 30 replicas along replica exchange cycle of 100. Legend scale represents RMSD from native structure.*

Fig. 26 is a way to quantify the degree of misfoldness. The more exchange attempts, the more replicas jump out their native states and sample misfolded states. The rmsd of most misfolded BPTI is as high as 30Å.

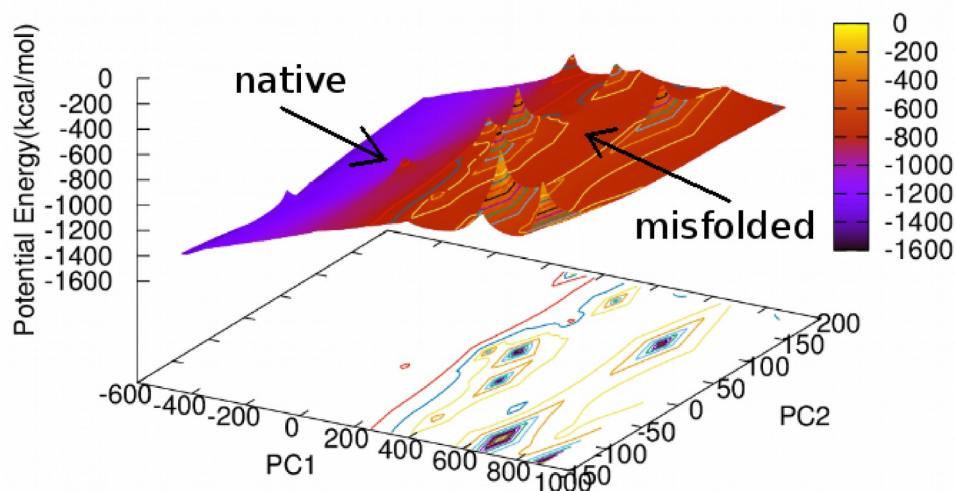
### Energy Distribution of 30 Replicas of BPTI



*Fig. 27: Energy distribution of 30 replicas of BPTI.*

Fig.27 shows the energy distribution of all replicas (The line with the same color does not necessarily mean the same replica). In order to efficiently sample in replica exchanging, the energy distributions must overlap. Due to limited data points, it is difficult to fit the data to a gaussian function. For this reason, the curves in graph are smooth expression of potential energetics. All replicas cover the energy range from  $\sim -1800$  kcal/mol to  $\sim 0$  kcal/mol, and the acceptance ratio is  $\sim 20-50\%$ . Since we start from native structures, the overlap of energetics becomes quite large for some replicas.

### 3D Potential Energy Landscape of 30 Replicas



*Fig. 28: Potential energy landscape of 30 replicas by Principal Coordinates Analysis. PC1 and PC2 are recalculated coordinates.*

To make sure misfolded proteins are close to native folded protein, we plot the potential energy surface in 3D space, x and y are two new calculated coordinates labeled “PC1” and “PC2”, the overall goal is to make PC1 has the largest variance and PC2 has the second largest variance. Z axis is potential energies of 30 replicas from last final step of each replica. See Appendix B for details of PCA analysis. From the potential surface we should be able to see valleys of native proteins and misfolded proteins, separated by

energy barriers. Our results show a population of native stated proteins at left side, and misfolded protein on right side.

It is difficult to depict the actual potential energy landscape of a protein folding due to the inefficient sampling of conformation space, so this is only a landscape of these 30 replicas. To get more accurate potential, more conformations from all trajectories should be included, but this will slow down the calculation of matrix eigenvalues significantly in PCA analysis.

To observe the role of water bridge in misfolded proteins, we calculated water bridge score for all 30 replicas, and compared the score from direct hydrogen bonds and from water bridges. Fig. 29 shows that direct hydrogen bond are more directly related to RMSD, in other words, misfolded protein involves significantly fewer number of direct hydrogen bonds than the native protein. Water bridges are less related to RMSD compared to direct hydrogen bond, and the contribution of water bridges in native structure are lower than in misfolded structures. Misfolded proteins also have a compact arrangement as native structure, but their arrangement lacks of direct hydrogen bonds to remain its stability. In addition to this, the forces applied to a protein from water bridges are stronger than in native structure, which makes them even more unstable.

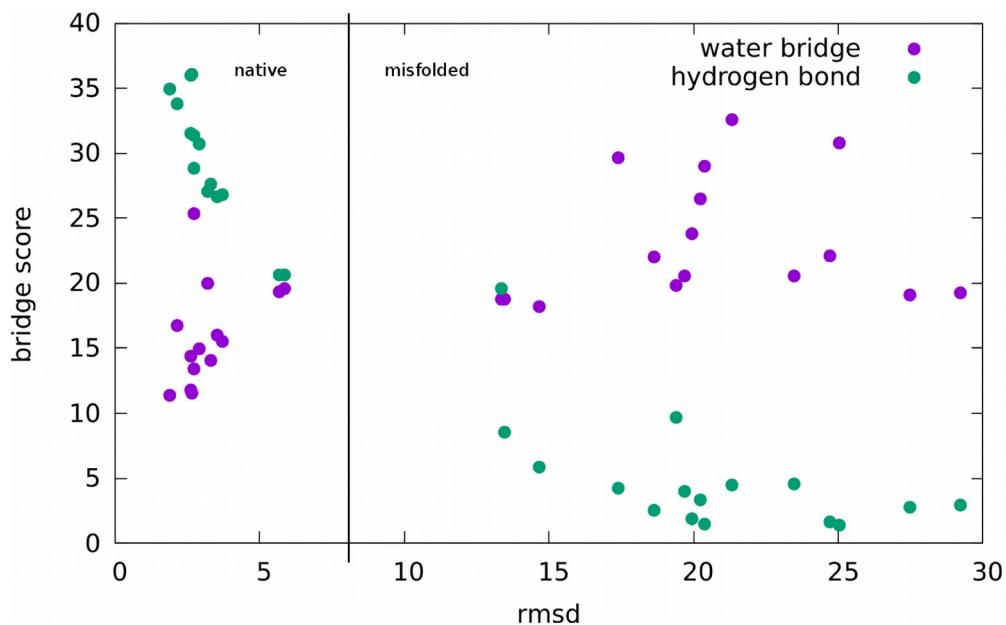


Fig. 29: Correlation between RMSD and bridge score for 30 replicas.

## 2.6 Discussion and Conclusion

In this chapter, we first investigated some properties of water network, and the influence of solutes on these properties. The results tells us that any solute in water solution reacts with the water network, and different ionic atmosphere provide different environment for proteins to function. Ions in solution disrupt the native water network through decreasing degree of hydrogen bonding, though this is mitigated by an enhanced hydrogen bond lifetime. The tearing effect of ions is not surprising. However, the exact picture of how ions interact with water network still needs to be investigated, especially the different roles of cations and anions. We also investigated the role that water bridges play in folded

and unfolded protein through molecular dynamics simulation. The fact that there are more hydrophilic groups on proteins leads us to believe that the hydrophilic interaction must have a profound effect on protein folding or stability. We developed a scoring system to quantify hydrogen bond interaction which is the dominant factor of hydrophilic interaction. We classified hydrogen bond interaction into “direct hydrogen bond” and “water bridge” with one to three water molecules on that bridge. In this research we tried two different scoring strategies, so that we did not overestimate the bridge contribution. From the calculation we conclude that the formation of water bridges at early stage of protein folding is part of driving force of folding, when intra-molecular hydrogen bonds have not formed yet. After the formation of a transient structure which includes some direct hydrogen bonds, water bridge starts to act as a subsidiary factor in protein folding. From the fact that misfolded protein show more water bridge than native protein, we can tell that the folding of a protein must be directed by the water bridges, until a minimum influence from water bridges are reached.

In this research we did not consider any factor of hydrophobic interaction, which we believe is crucial to side chain packing, This will be studied in future research.

## Chapter 3: Effect of Protein Unfolding on pKa Shifts of BPTI

### 3.1 Review of Constant pH Simulations

The folding, function and stability of proteins depends upon the solution pH. This pH-dependent process is largely due to the relationship between the charged nature of both species (protein and solution) and the conformation response. The protonated or deprotonated states are determined by intrinsic pKa ( $pK_{1/2}$ ) values and the electrostatic field exerted by the surrounding environment. Any factor influencing the electrostatic field will have an effect on the protonated/deprotonated states of the side chain. Among these factors, the ionic strength is the most crucial one. The electrostatic field is not only modified by ions in solution, but also by protons binding to charged groups. The overall effect of mobile ions and polar water molecules on the protein is that they shield the fixed protein charges from each other by distributing and orienting themselves.<sup>94</sup>

The favored method for pKa determination through experiment is by NMR<sup>95-99</sup>. The overall idea is the alteration of chemical shifts of NMR-active nuclei ( $^1\text{H}$ ,  $^{13}\text{C}$ ) upon protonation and deprotonation varies with pH. In contrast to the experimental approach, some in-silico pKa prediction methods<sup>100</sup> are available to offer fast and preliminary pKa values for titratable sites. These methods take experimental pKa values of model compounds as input, and then further improved by empirical methods or free energy calculation.



No matter experimental method or predictive methods, the major drawback is structural correlations are less well developed. This makes MD method for pKa calculation stand out.

Due to the complexity of estimating the energetics of protonation/deprotonation events, as well as limitation on forming or breaking bonds in a simulation, it has only recently been possible to model pH effects through molecular dynamics simulation. In principle, the most rigorous way to estimate an individual pKa value for a protein side chain would involve a free energy simulation connecting the protonated and deprotonated forms of the molecule:

$$pKa = -\log_{10} Ka = \frac{\Delta G}{k_B T \ln 10} \quad (3.1)$$

Any thermodynamic quantity that involves entropy in free energies cannot be derived from one simulation, as an evaluation over all multidimensional phase space is impossible. It is however possible to find the difference in free energy between two systems with slightly different potentials  $V_1$  and  $V_2$ :

$$G_1 - G_2 = -k_B T \ln \frac{Q_1}{Q_2} = -k_B T \ln \frac{\int \exp(-\beta V_1) dr}{\int \exp(-\beta V_2) dr} \quad (3.2)$$

The fundamental theory of constant pH simulation defines  $V_1$  and  $V_2$  as the potential of protonated and deprotonated states respectively. Within this definition, there are two common models to describe the protonation states, the continuous and discrete models.

### **Continuous-Protonation State Models**

Baptista<sup>101</sup> calculates a charge distribution with a continuum electrostatics method (pH-CE), which is dependent on the configurations of the protein. The probability of finding the system with this charge distribution is only a parameter of Potential of Mean Force (PMF), which can simultaneously treat titration and conformational freedom.

The ‘Acidostat’ method developed by Borjesson<sup>102</sup> defines a linear combination of the fraction  $\lambda$  of the deprotonated-state, and a fraction  $(1-\lambda)$  of the protonated-state. This fractional protonation states represents the mean influence of a specific ionizable group on its environment. The evolution of these  $\lambda$  values is determined by relaxation to  $\lambda^0$ , the corresponding equilibrium value, and this  $\lambda^0$  is determined by pH, and the chemical nature of groups and the environment (the configuration and ionization states of other groups). Compared to pH-CE model, this approach includes structural conformations into the free energy evaluation.

Brooks<sup>103</sup> developed a new method based on early work by Mertz and Pettitt<sup>104</sup>, which also defines a coordinate  $\lambda$ , lying between 0 to 1 corresponds to a completely protonated ( $\lambda=0$ ) and a completely unprotonated ( $\lambda=1$ ) state. To avoid a convergence to an intermediate charge state, they introduced an energetic barrier to force  $\lambda$  towards values representing a fully protonated or fully deprotonated state, and this continuous titration coordinate represents an instantaneous microstate rather than a fractional protonation population as stated in the “Acidostat” method.

### **Discrete-Protonation State Models**

The Stochastic model developed by Baptista<sup>105</sup> proposed Monte Carlo (MC) sampling after a short period (0.2-5.0 ps) of explicit solvent MD simulation. After a protonation state has changed, the solvent is equilibrated (~ps) and then the MD simulation is continued. The protonation states are clearly defined as either protonated or deprotonated, and equilibration is necessary to avoid energy “shock”.

Another model similar to the stochastic model was developed by Dlugosz<sup>106,107</sup>. It also employs discrete protonation states. The solvent is represented as continuum with a dielectric constant, ionic strength and pH (an implicit solvent algorithm). Assignment of protonation states happens after parallel simulations (constant-pH simulations of 100 initial structures each with different pH) are all done, the molecule makes a random jump to the defined protonation states and then moves on to the next sub-trajectory. Compared to the stochastic model, which only changes the partial charges of the titratable groups, this model explicitly adds or removes protons from molecules. Thus the number of atoms of molecules at each beginning stage may not remain constant. Resetting and reassigning all the velocities of the solute atoms is needed to avoid a sudden change in energetics due to adding or removing the hydrogens.

### **3.2 Method - CPMD**

In this research, we performed constant-pH simulation using continuous Constant PH Molecular Dynamics (CPMD) from Brooks. This function is implemented in C37b and later version of CHARMM. The basis of this method is a set of continuous titration

coordinates that describe transitions between protonated and deprotonated states of a group labeled  $i$ . In order to vary the value of  $\lambda$  in between 0 and 1, it is defined as a function of  $\theta$ :

$$\lambda_i = \sin^2(\theta_i) \quad (3.3)$$

To restrict  $\lambda$  to the endpoints at 0 or 1, the  $\lambda_p$  is defined to occupy one of three states:

Protonated States= $(0 \leq \lambda_i \leq \lambda_p)$

Mixed States= $(\lambda_p \leq \lambda_i \leq 1-\lambda_p)$

Unprotonated States= $(1-\lambda_p \leq \lambda_i \leq 1)$

So this  $\lambda_p$  is an empirical value that must be determined through trial and error. In our research we set it to 0.1 which is its default value. The fractional population of unprotonated states  $S_i^{unprot}$ , which provides a link between calculation of pKa and simulation statistics about  $\lambda$  is defined as

$$\frac{1}{1 + \frac{N(\lambda_i < \lambda_p)}{N(\lambda_i > 1 - \lambda_p)}} = S_i^{unprot} = \frac{1}{1 + 10^{(pKa^i - pH)}} \quad (3.4)$$

The Potential of Mean Force (PMF) of a model compound is essentially quadratic in nature with respect to  $\lambda$  (the derivative of PMF is linear to  $\lambda$ ), and hence can be fit to a two-parameter parabolic function<sup>103</sup>.

$$U^{model}(\lambda_i) = -A_i(\lambda_i - B_i)^2 \quad (3.5)$$

The change of Gibbs free energy of protonating the model compound can be defined as:

$$\Delta G_{class}^i(model) = U_{model}^i(\lambda = 0) - U_{model}^i(\lambda = 1) = -A_i(2B_i - 1) \quad (3.6)$$

All parameters for titratable amino residues are stored in a charmm control file *phmd.in*.

In this chapter, we will calculate pKa of BPTI with residues that have already been parameterized. In Chapter 4 we will describe parameterization of a new compound.

### 3.3 Analysis

The newly prepared input scripts to run the simulation include:

“*solvate.sh*”: which prepare a clean starting structure.

“*add\_h.inp*”: this adds protons to all titratable sites.

“*solvate.inp*” & “*add\_ions\_hphmd.inp*” & “*minimize.inp*”: these three scripts solvate the protein with TIP3P water molecules and ions (Na<sup>+</sup>,Cl<sup>-</sup>) in order to neutralize the system. This step is for the hybrid model and is not necessary for any implicit solvent model.

“*hphmd.inp*” & “*hphmd\_phrex.inp*”: these two scripts run the constant pH simulation for one pH value or a series of pH values through the replica exchange technique.

#### **Constant pH simulation of monomer and dimer blocked amino acids**

We calculated the pKa values of titratable amino acids monomers and dimers (Table 2).

The C-terminal and N-terminal ends were capped with ACE and CT3.

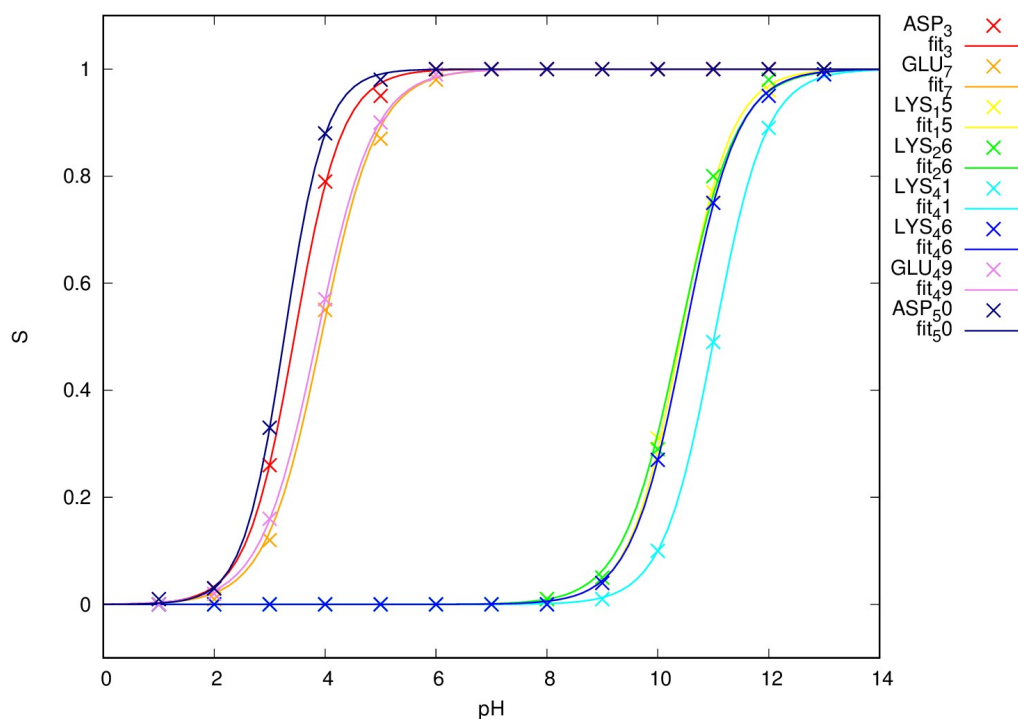
Table. 3 Calculated pKa values of titratable amino acid fragments.

Amino Acid	Monomer		Dimer		Reference <sup>108</sup>
	Implicit	Explicit	Implicit	Explicit	CHARMM
ARG	13.55±0.02	13.36±0.14	13.61±0.10	13.36±0.14	12.0
ASP	2.87±0.23	3.63±0.12	3.27±0.16	3.74±0.14	4.0
GLU	3.65±0.05	3.86±0.04	3.66±0.04	3.88±0.07	4.4
HSD	6.88±0.06	6.88±0.08	6.54±0.18	6.56±0.26	6.3
LYS	10.69±0.07	10.52±0.05	12.21±0.40	10.48±0.03	10.4
TYR	10.03±0.05	10.05±0.06	9.97±0.06	9.98±0.05	9.6

The calculated pKa's agree with experimental measured pKa's very well. Errors in the measurements are typically  $\pm 0.1-0.4$  pH units. The accuracy of implicit and explicit solvent model are comparable. The implicit solvent model is faster than explicit solvent model, but it can not deal with hydrogen bonds, as it is a factor that influence pKa value.

### Constant pH simulation of native BPTI

We calculated pKa values of titratable sites in native BPTI, the titration curve is shown in Fig. 30, and compared the pKa values with monomer-state of those sites (Table 3).



*Fig. 30: Titration curves for titratable residues in BPTI in the explicit solvent model. From pH-REX simulation with 14 replicas (pH=1,2,...,14).*

The average unprotonated fractions  $S$  (shown as crosses) at 14 pH values were fit to the Henderson-Hasselbach equation and shown as solid lines.

Table 4. Calculated pKa values of titratable sites in BPTI

Residue	pKa(explicit)	monomer	reference <sup>100</sup> (CHARMM)
ASP_3	3.45±0.96	3.63±0.12	4.0
GLU_7	3.93±0.34	3.86±0.04	4.4
LYS_15	10.41±0.64	10.52±0.05	10.4
LYS_26	10.39±1.61		
LYS_41	11.02±2.24		
LYS_46	10.48±0.21		
GLU_49	3.86±0.86		
ASP_50	3.26±1.58		

The ASP and GLU residues show smaller pKa values in BPTI than the reference value.

The average pKa values of ASP, GLU, and LYS decrease, increase, and decrease respectively from calculated pKa of monomer, but these changes were within the range of standard error.

### **pKa shifts Upon Unfolding Of BPTI**

The protonation behavior of titratable groups is sensitive to the environment, which includes the properties of solute and solvent. It is useful to know what the pKa values of titratable groups would be in a protein if their environment is changed. In order to



understand this, we performed SMD of BPTI, and calculated pKa values of 100 snapshots from unfolding trajectories.

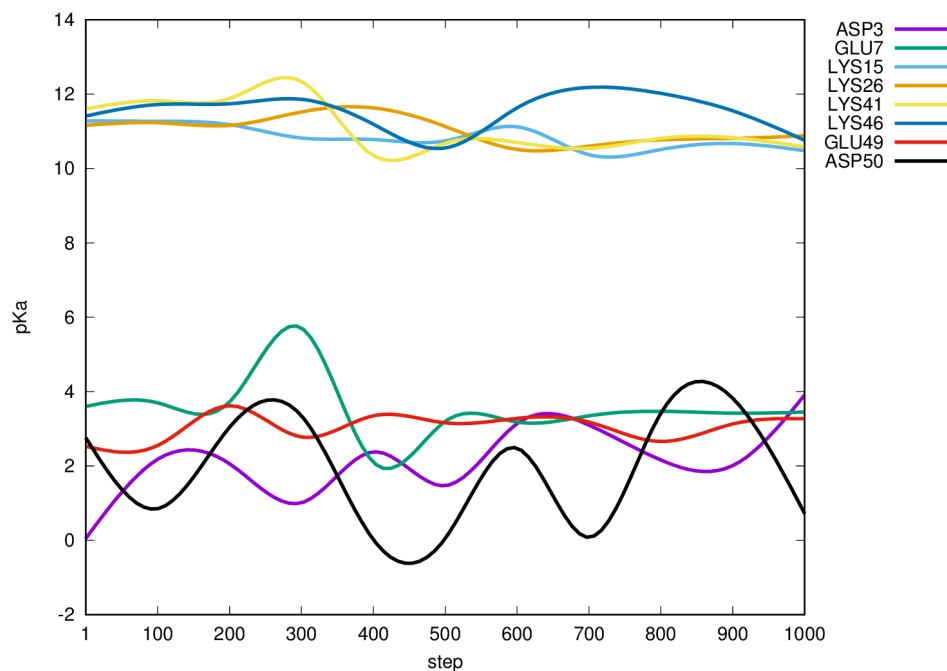


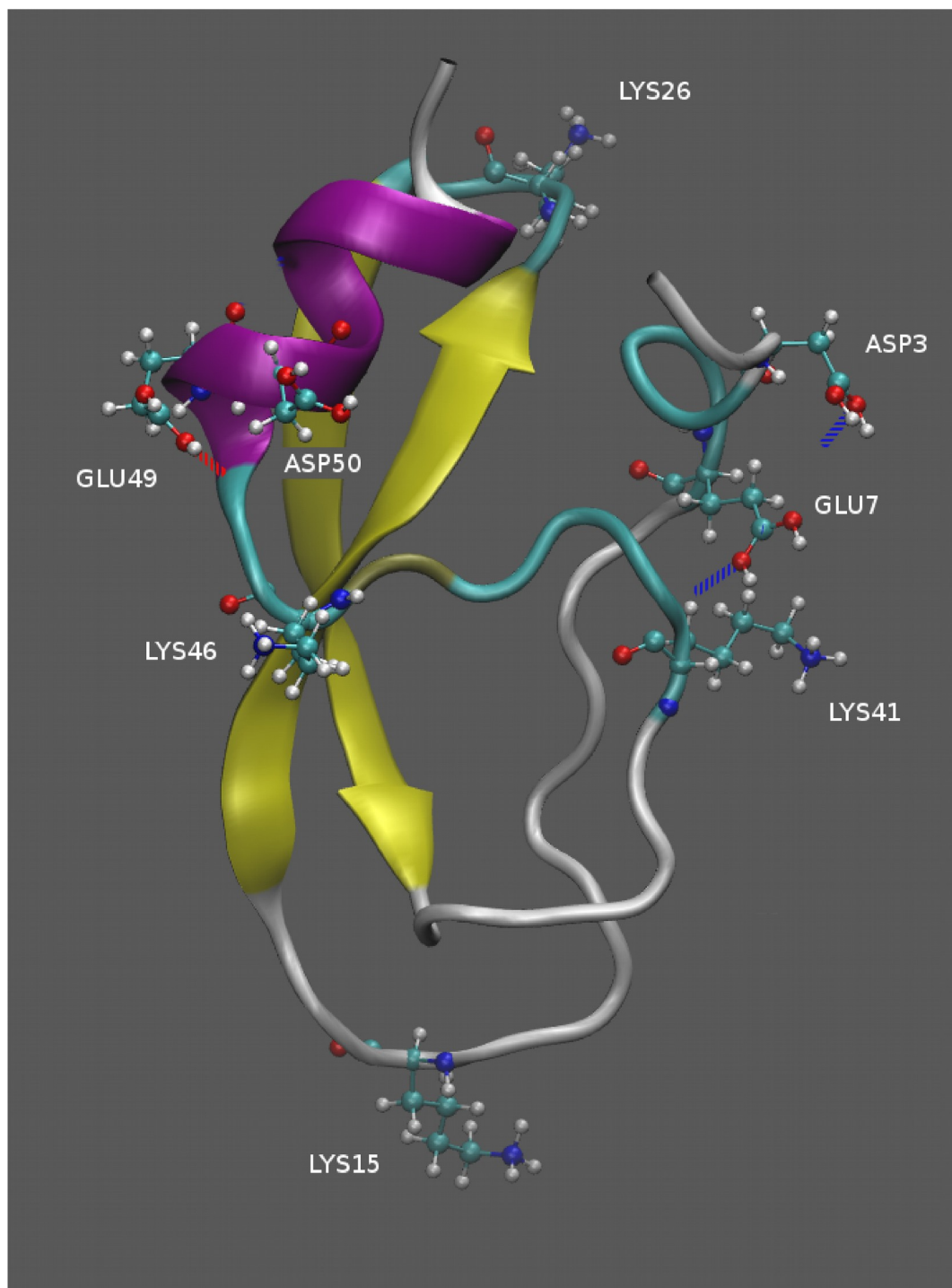
Fig. 31: pKa shifts of 8 titratable residues in BPTI during SMD.

There are some factors that influence pKa values<sup>109</sup>: 1) Dehydration (through the Born Effect): pKa shifts upon the environment change between hydrophilic and hydrophobic atmosphere; 2) Charge-Charge interaction (Coulombic): pKa shifts as the free energy of Coulombic interaction of charges on titratable sites goes lower; 3) Charge-Dipole Interaction (Hydrogen Bonding): pKa shifts so that the interactions are more favorable with protonated states.

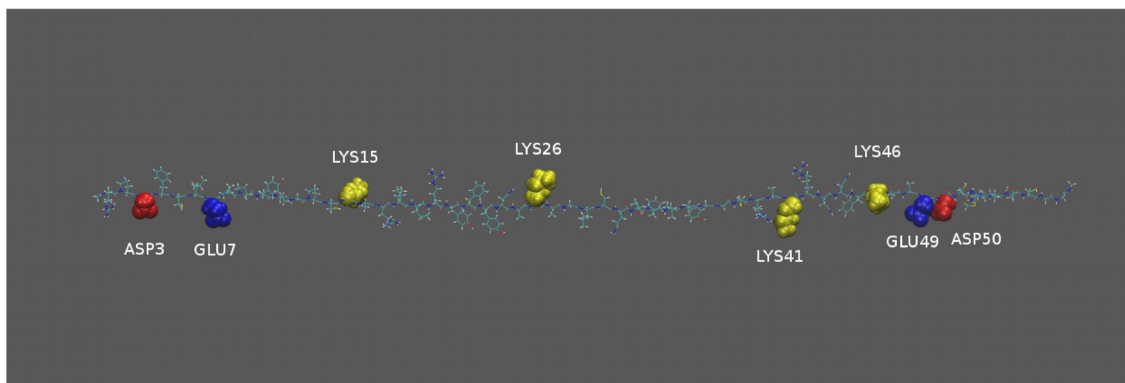
Fig. 31 shows that the pKa values of all titratable residues does not show a universal trend over the trajectory of unfolding. ASP and GLU residues show more fluctuation than

LYS. Fig. 32 and Fig.33 show that BPTI is a small protein, and the side chain of the pertinent residues (ASP3, GLU7, LYS15, LYS26, LYS42, LYS46, GLU49 and ASP50) are not buried inside completely. This means that the hydrophobic environment of all the titratable residues remains almost the same before and after pulling, the factor of dehydration does not affect pKa significantly.

The energy of interaction of the ionizable group can be calculated with Coulomb's law. When two opposite charges are  $4.2\text{\AA}$  apart in water,  $\Delta G = -1\text{kcal/mol}$ , and this reduced to  $-0.5\text{kcal/mol}$  inside a cell due to electric constant change. As BPTI is pulled, the ASP and GLU residues (especially ASP3 and GLU7) that are close to the LYS residues (LYS26, LYS41) in folded BPTI become further apart. This leads to an pKa increase of ASP and GLU (they are more easily protonated), and a corresponding decrease in LYS (it is more difficult to protonate). The loss of hydrogen bonds which do not need protons for the ASP and GLU residues during the extension process leads to decrease of their pKa values. With opposite effects from factor two and three, the apparent value of the pKa shift is difficult to predict. From the curve in Fig. 31, we can tell that the overall effect of unfolding on the pKa is negligible.



*Fig. 32: 3D representation of folded BPTI. Hydrogen bonds of ASP3, GLU7 and GLU49 with their neighbors are labeled with a dashed line.*



*Fig. 33: 3D representation of extended BPTI.*

### 3.4 Discussion and Conclusion

We measured the pKa values of titratable groups of amino acids in monomer and dimer states, and compared these with the pKa values of titratable sites on BPTI. The calculated pKa values of the monomers agree with experimental values well. The calculated pKa values on native and unfolded BPTI deviate slightly from monomer, but this deviation is completely subjected to high standard error, thus it is difficult to tell the pKa shift in folded or unfolded protein. Some large pKa shifts were reported<sup>110</sup>. To observe pKa shift, we need to choose a larger protein that has significant hydrophobic environment change, or charge-interaction change for some residues, and continue to improve the calculation algorithm to get more accurate value with high precision.

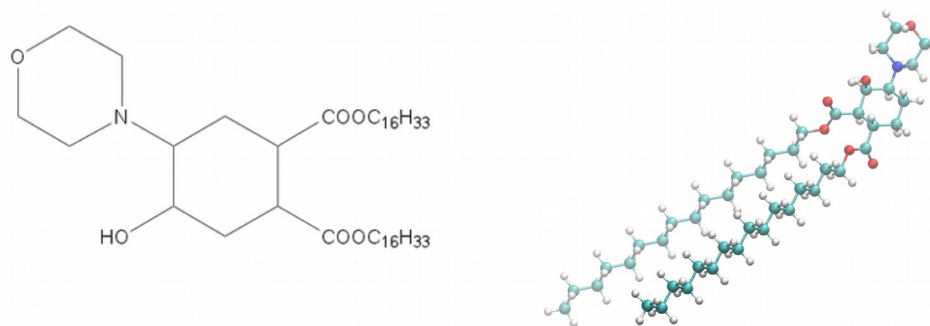
## Chapter 4: Modeling and Simulation of pH-Sensitive Lipids

### 4.1 Review of Lipid Simulations and MORC16

In contrast to proteins, lipids are molecules that are able to dissolve in both organic and polar solvents. They are one of the most important biomolecules in cellular signaling, energy storage, and are building blocks of cellular membranes, and are able to form more complex structures, such as micelles and liposomes. The first molecular dynamics simulation of a realistic lipid monolayer was by Kox<sup>111</sup>. The first molecular dynamics simulation of bilayers was by Heller<sup>112</sup>. Even though all-atom molecular simulation can be successfully applied to small lipid systems, approximation methods are needed to simulate complex lipid system with millions of atoms. Coarse-grained models are currently a very popular method in simulating lipid systems, providing a good accuracy within an affordable time<sup>113-119</sup>. Among these methods, the Martini model is the most successful in lipid simulation<sup>117,120,121</sup>. Here, we report the developed of simulation framework for a novel designed lipid, MORC16 synthesized by Samoshin<sup>122</sup>, shown in Fig.34. MORC16 is a derivative of trans-2-aminocyclohexanol. The liposome formed by MORC16 is also referred to as “fliposome”, due to MORC16’s unique property of pH responsive conformation flipping. This fliposome has the potential to serve as a vessel in a drug delivery system.

## 4.2 Modeling and Simulation of MORC16

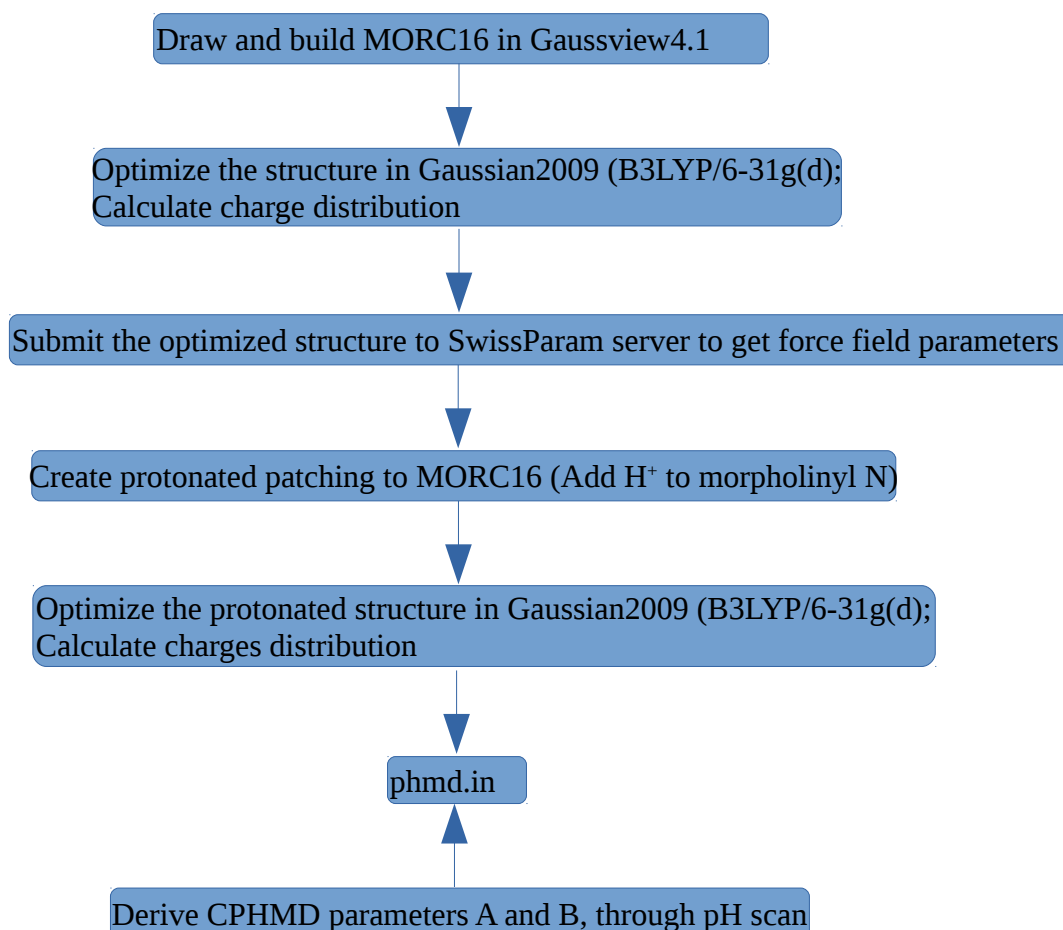
We performed a series of all-atom, constant pH simulations using CHARMM. Because CHARMM has better support for constant-pH simulations, it provides the best framework for these simulations. However, as the size of system increases, the efficiency of any simulation slows down rapidly, and other considerations take precedence. Thus our self-assembly simulations were performed using GROMACS along with a coarse-grained MARTINI forcefield. MARTINI provides good results for lipid bilayer simulation with satisfying speed. However, GROMACS does not support pH simulation in its published version. We performed steered molecular simulations by AMBER to study stability of micelles and liposomes. Amber has good support for CUDA acceleration, and it also supports constant pH simulations. We used PACKMOL to assist in the modeling of micelles and liposomes.



*Fig. 34: Formula and 3D representation of MORC16 lipid.*

#### 4.2.1 MORC16 Constant pH Simulation using CHARMM

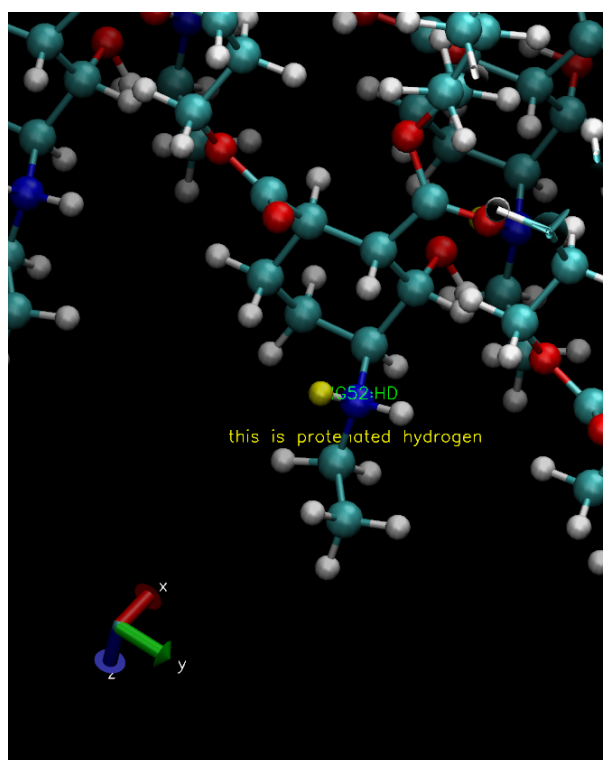
MORC16 is a pH-sensitive molecule with one titratable site (Samoshin<sup>122</sup>). Because it is a newly-synthesized molecule, the parameters must be constructed for this molecule. This includes both the parameters for regular dynamics and for constant pH simulation. The flow chart of modeling of MORC16 is as follows:



The parameters for atoms SwissParam<sup>123</sup> server uses are derived from Merck Molecular ForceField (MMFF)<sup>124</sup>. It is useful to know that MMFF is not optimized for one use, but tries to perform well over a wide range of calculation. This means it may require some

improvement for specific purposes. In this research we did not require any modification to the parameter file, however. The compressed archive retrieved from the server includes a parameter file, which include bond, angle, dihedral, improper and nonbonded terms and some coordinate files in different formats.

According to Samoshin<sup>122</sup>, the driving force of this acid-triggered flip is strong intramolecular hydrogen bonding of hydroxyl oxygen and protonated morpholinyl nitrogen showed in Fig. 35. The major difference between protonated and deprotonated states is a change of charge distribution upon protonation. We calculated the charge distribution of protonated and deprotonated MORC16 by the NBO method.



*Fig. 35: Protonation of nitrogen on MORC16.*



The CPHMD method relies on the calculation of the deprotonation Gibbs free energy of a tritratable site in reference to model compound. The parameters needed are the experimental pKa value, parameter A, B and  $\lambda_p$  (see section 3.2).

### Step 1. Input file preparation

create an entry in *phmd.in* file, the first line is defined as follows:

```
'title pKa A B  $\lambda_p$ '
```

pKa is 5, A, B and  $\lambda_p$  are initially set to 0, so the first line looks like this:

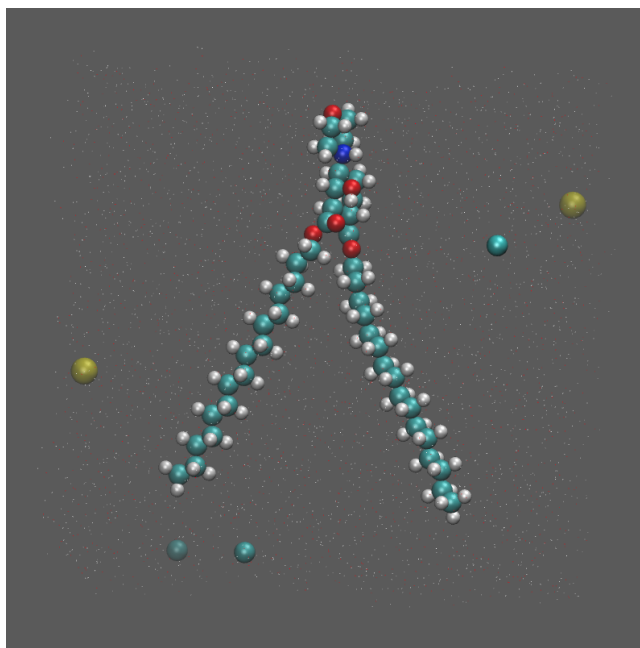
```
'MOR 5.0 0.0 0.0 0.0'
```

The atom names and the partial charges of protonated and deprotonated state follow.

One line that is very crucial for this entry is definition of protonated proton:

```
'HD 0.46058 0.0 1.0 0.0'
```

HD is the name of the protonated hydrogen, (+)0.46058 is the charge of the protonated state, 0.0 is the charge for the deprotonated state, 1.0 is mass of the protonated state, 0.0 is mass of the deprotonated state. In this way we can see that the deprotonated state effectively ionizes the proton by setting its mass and charge to zero. After running the scripts *add\_h.inp* (define protonated state), *solvate.inp* (solvate in explicit water box), *add\_ions\_hphmd.inp* (neutralize the system to defined ionic strength, 0.1M) and *minimize.inp* (combine water box and ions together and equilibrate the system), the system looks like Fig. 36.



*Fig. 36: 3D image of protonated, solvated and ionized (three Cl<sup>-</sup> in cyan , two Na<sup>+</sup> in yellow) MORC16 lipid.*

## **Step 2. Derivation of the Parameters for MORC16**

The pipeline of the derivation of the parameters for MORC16 is as following:

Set  $\theta$  value to 0.4(0.6,0.8,1.0,1.2,1.4) in hphmd\_derive.inp

Run the simulation with hphmd\_derive.inp

Calculate average  $dE/d\theta$

Analytically fit the data to  $dU_{mod}/d\theta = 2 * A * \sin(2 * \theta) * (\sin\theta^2 - B)$

*hphmd\_derive.inp* is a script to calculate average  $dE/d\theta$  at each fixed value of  $\theta$ . To obtain parameters A and B, we fit the values of  $dE/d\theta$  to function

$2 * A * \sin(2 * \theta) * (\sin\theta^2 - B)$  through Grace, results shown in Fig.37.  $A = -188.747$ ,  $B = 0.990891$ .

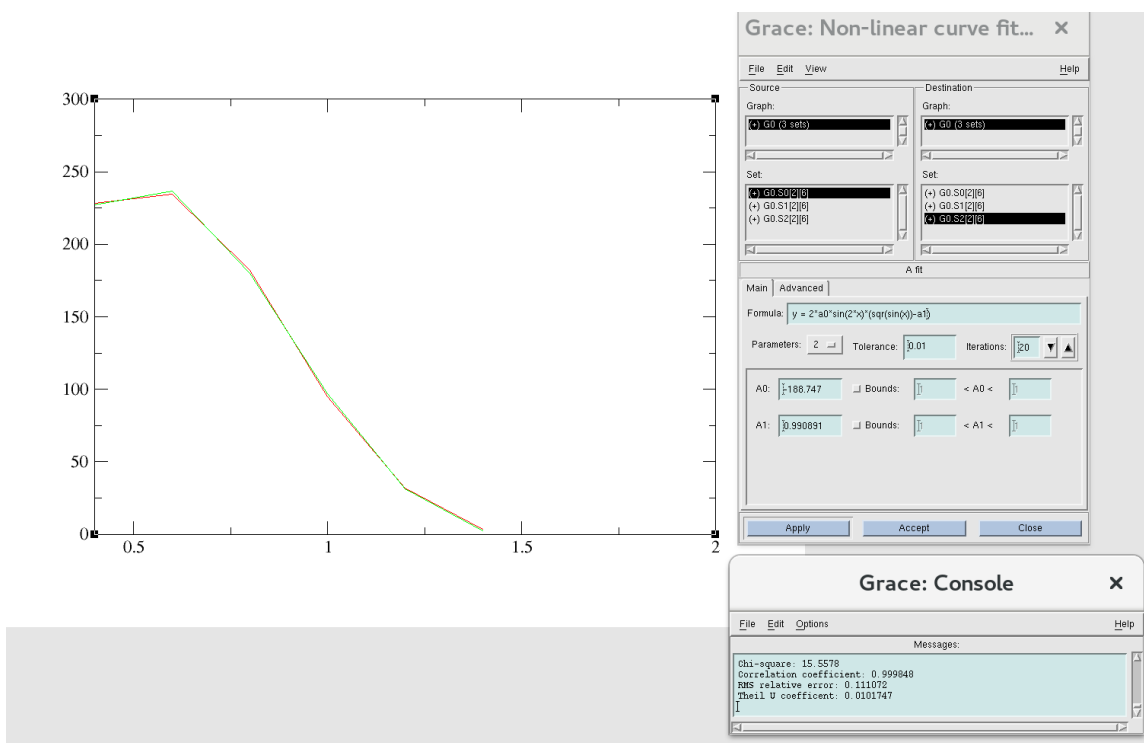


Fig. 37: Data fitting of parameters A, B of PMF function. x is  $\theta$ , y is E.

The newly derived parameters were then updated in the file `phmd.in`. A constant pH replica exchange simulation by `hphmd_phrex.inp` was then run, where the pH was increased from 1 to 8 increase by 1 for each replica. The calculated pKa value is 5.05 (shown in Fig.38). We also checked the data to ensure the ionization occurs 50% protonated at its experimental  $pK_{1/2}$ .

### Titration Curve for MORC16

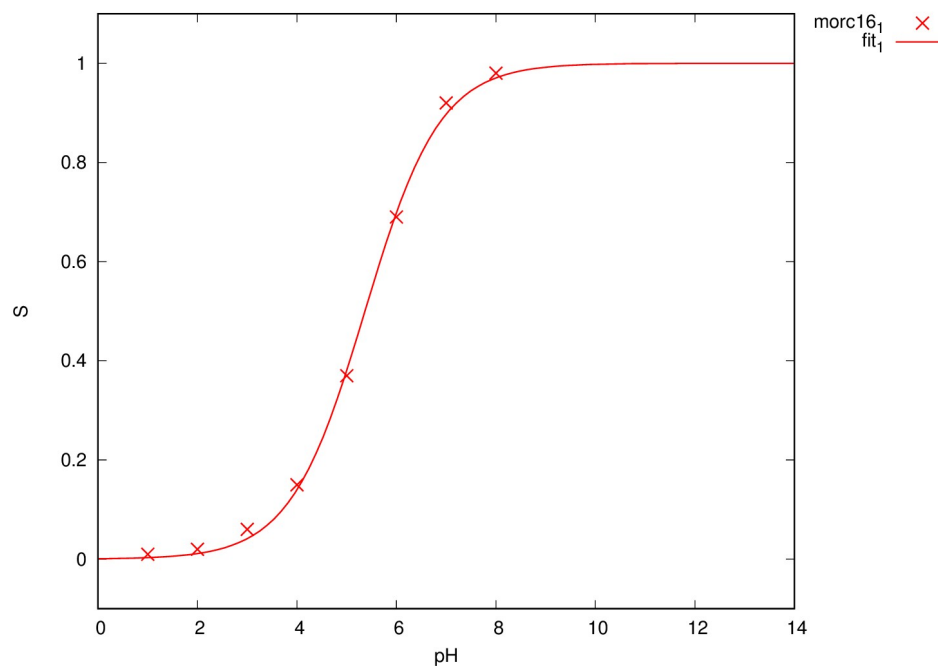
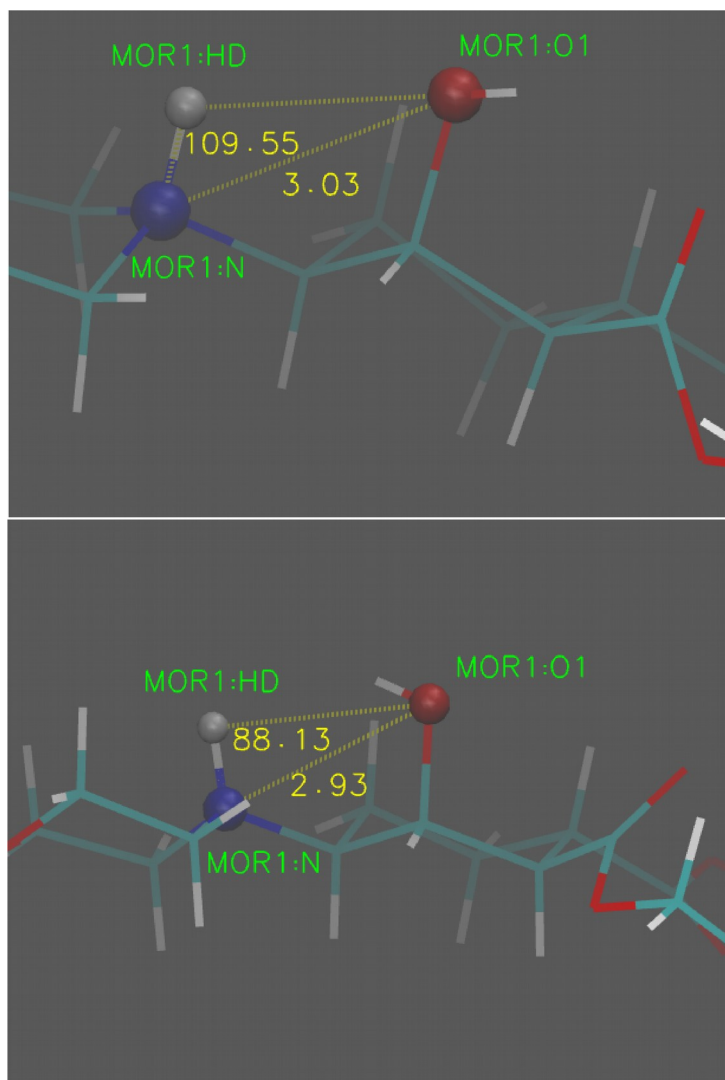


Fig. 38: Titration curve of MORC16 by constant pH replica exchange simulation.  $pK_a=5.05\pm 0.78$ .

### Geometric Characteristics of MORC16 for Different pH Values



*Fig. 39: 3D images of the titratable site in MORC16 at pH=1 (top) and pH=10 (bottom).*

To investigate the structural change of MORC16 at different pH, we show two representative configurations at pH=1 and pH=10, which represents the protonated and deprotonated states respectively in Fig. 39. When MORC16 is protonated, the distance between N and O1 is around 3Å, which is a typical distance to form a direct hydrogen bond. However, the angle of N-HD-O1 is too small for a direct hydrogen bond between the N and O1 atoms. The only difference between protonated and deprotonated states is the position of hydrogen connected to O1. When MORC16 is protonated, the proton connected to O1 is facing away from proton HD, while it is facing towards the N when MORC16 is deprotonated. The chair conformation of the cyclohexane does not flip to the the other chair conformation, which would cause the two lipid tails to be in axial position. The distance is also smaller than 4.5 Å, which is the minimum distance needed to accommodate a water bridge, so there is no way to correlate these two atoms by direct hydrogen bond or a water bridge.

#### 4.2.2 The Self Assembly of MORC16 with GROMACS

The Martini Model<sup>117,121,125</sup> is a popular coarse-grained model for molecular dynamics simulations of biomolecular system. This model uses a four-to-one mapping strategy to simplify the all-atom model. This means that on average, four heavy atoms (and any associated hydrogens) are represented by a single interaction center (see Fig.40). Only four main types of interaction centers are considered: polar(P), non-polar(N), apolar(C) (apolar compounds are strongly hydrophobic, while non-polar are miscible with both water and organic solvent) and charged(Q), thus simplifying considerably, the number of interactions, while maintaining important characteristics of the system.

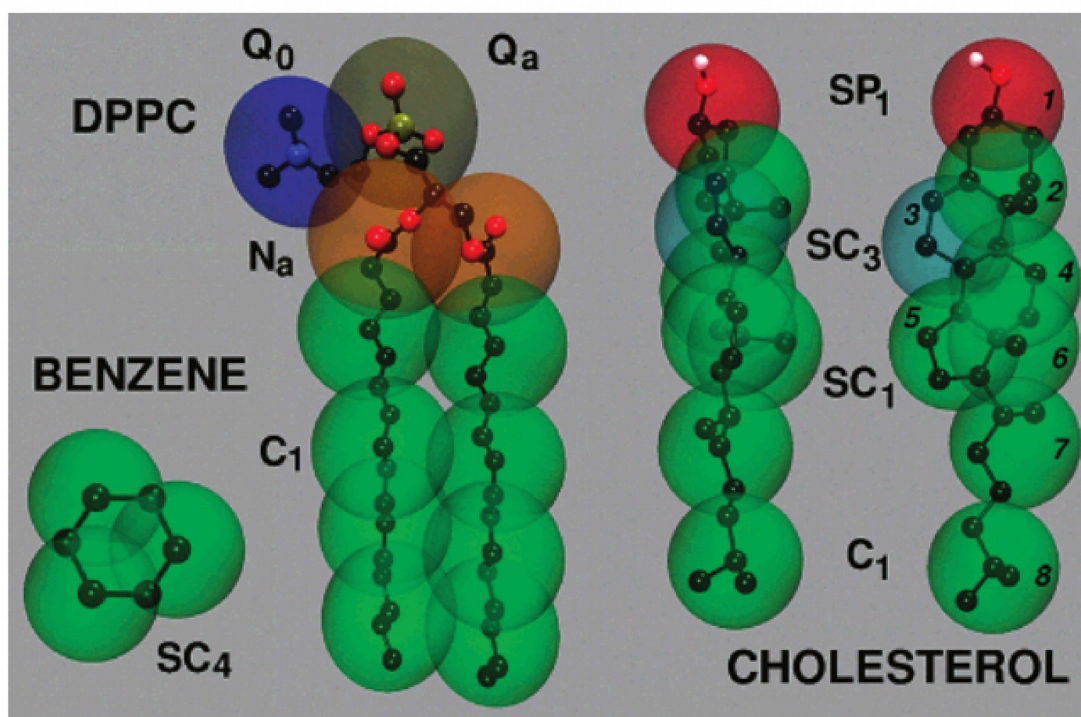
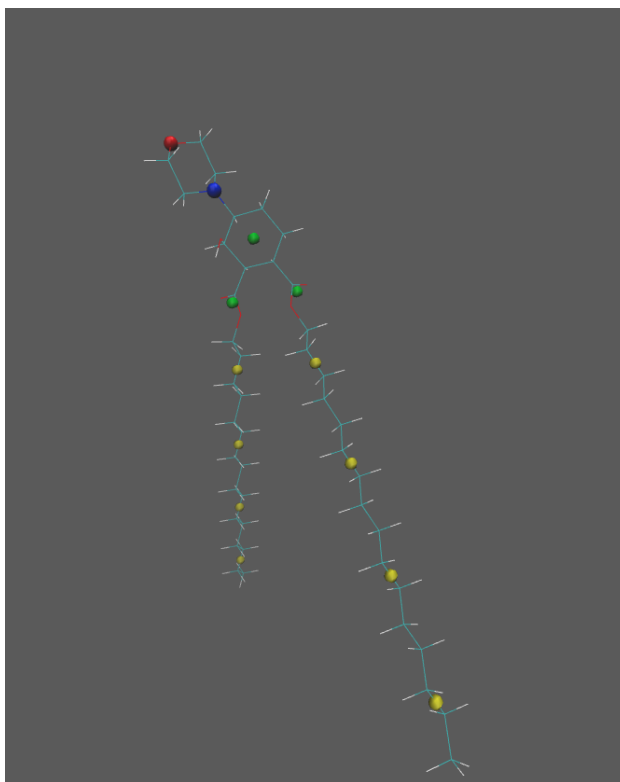


Fig. 40: MARTINI mapping strategy for DPPC, cholesterol and benzene.<sup>117</sup>

## Coarse-grained modeling of MORC16

The script used to create a coarse-grained model simulation for MORC16 is *martini\_morc16.tcl*, and the forcefield for MORC16 is contained in *MOC.itp*. We did not follow exactly the four-to-one strategy, as we were trying keep the molecule “pH triggerable” for future study. For this reason, the N atom was represented explicitly. Table 4. shows our mapping strategy used to represent MORC16 in coarse grained model, and Fig. 41 shows the representation of all beads in coarse grained model.



*Fig. 41: Coarse-grained representation of MORC1 superimposed on its all-atom representation.*

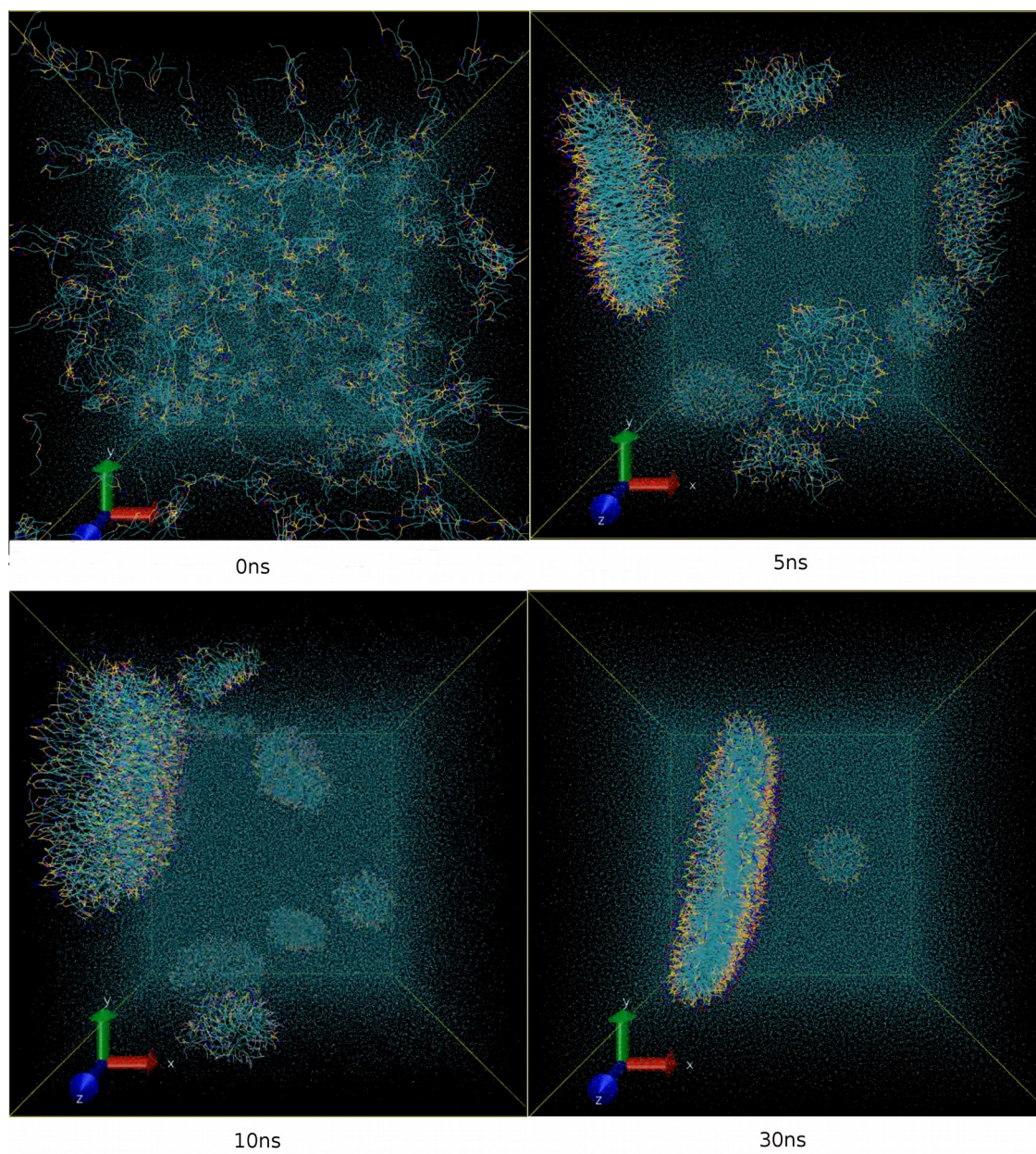


Table 5. Summary of the Coarse-Grain Strategy for MORC16

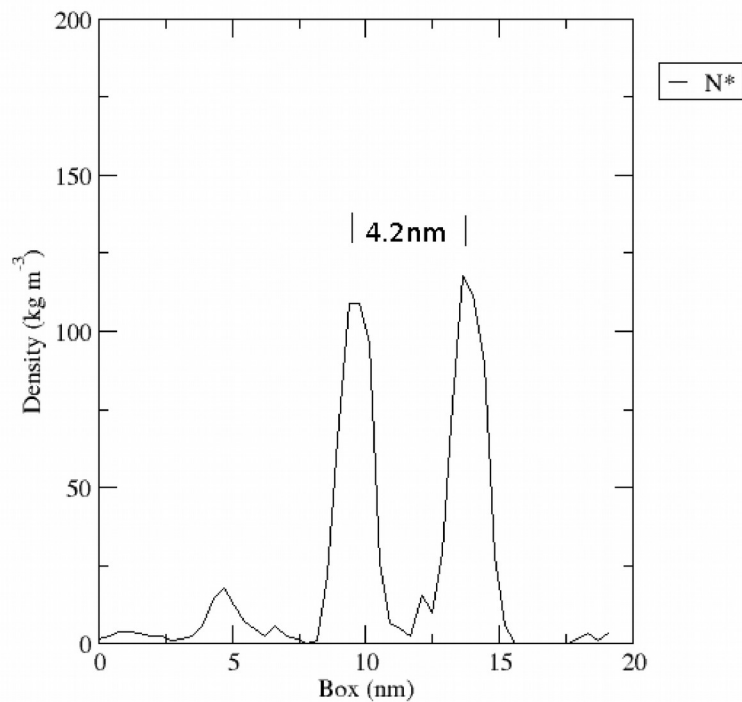
name of CG model	type	Coordinates from all-atom MORC16
OB1	P1	the same as O
NB1	Qa	the same as N
SB1	Qa	geometric center of (C,C1,C2,C3,C4,C5)
GL1	Na	geometric center of (C10,N2,O4)
GL2	Na	geometric center of (C11,O3,O5)
C1A	C1	geometric center of (C12,C13,C14,C15)
C2A	C1	geometric center of (C16,C17,C18,C19)
C3A	C1	geometric center of (C20,C21,C22,C23)
C4A	C1	geometric center of (C24,C25,C26,C27)
C1B	C1	geometric center of (C28,C29,C30,C31)
C2B	C1	geometric center of (C32,C33,C34,C35)
C3B	C1	geometric center of (C36,C37,C38,C39)
C4B	C1	geometric center of (C40,C41,C42,C43)

### **Bilayer Formation of MORC16 in Water**

The first step taken was to verify that single MORC16 lipids do form bilayer assemblies. We started with a water box filled with 1280 randomly distributed MORC16 lipids at a water:lipid ratio of 100:1. There was a rapid ( $\sim 5$  ns) initial formation of a liquid lipid phase separated into water solution driven by strong thermodynamic forces that isolate the hydrophobic tails from the aqueous environment. This was followed by a slower rearrangement of the lipid phase into bilayer (5 to 30 ns). The aggregation process is illustrated in Fig. 42. We also calculated the thickness and lateral diffusion through this bilayer, the results are shown in Fig. 43 and Fig. 44.

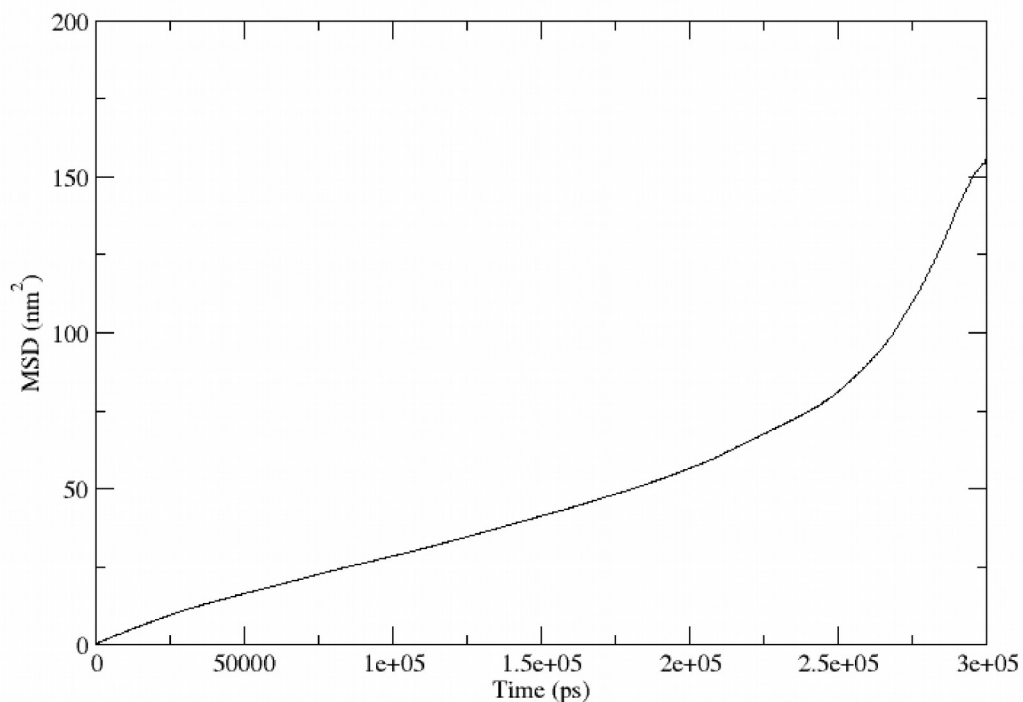


*Fig. 42: Snapshots of 1280 MORC16 lipids in aggregation. Waters are depicted as cyan dots, hydrophilic heads of MORC16 in blue and red, tails of MORC16 in cyan.*



*Fig. 43: Determination of MORC16 bilayer thickness from simulation.*

In order to determine the bilayer thickness, we calculated the density over a number of different functional groups. The bilayer thickness can thus be obtained from the distance between the atom N peaks in the density profile, averaged over the course of a simulation run.



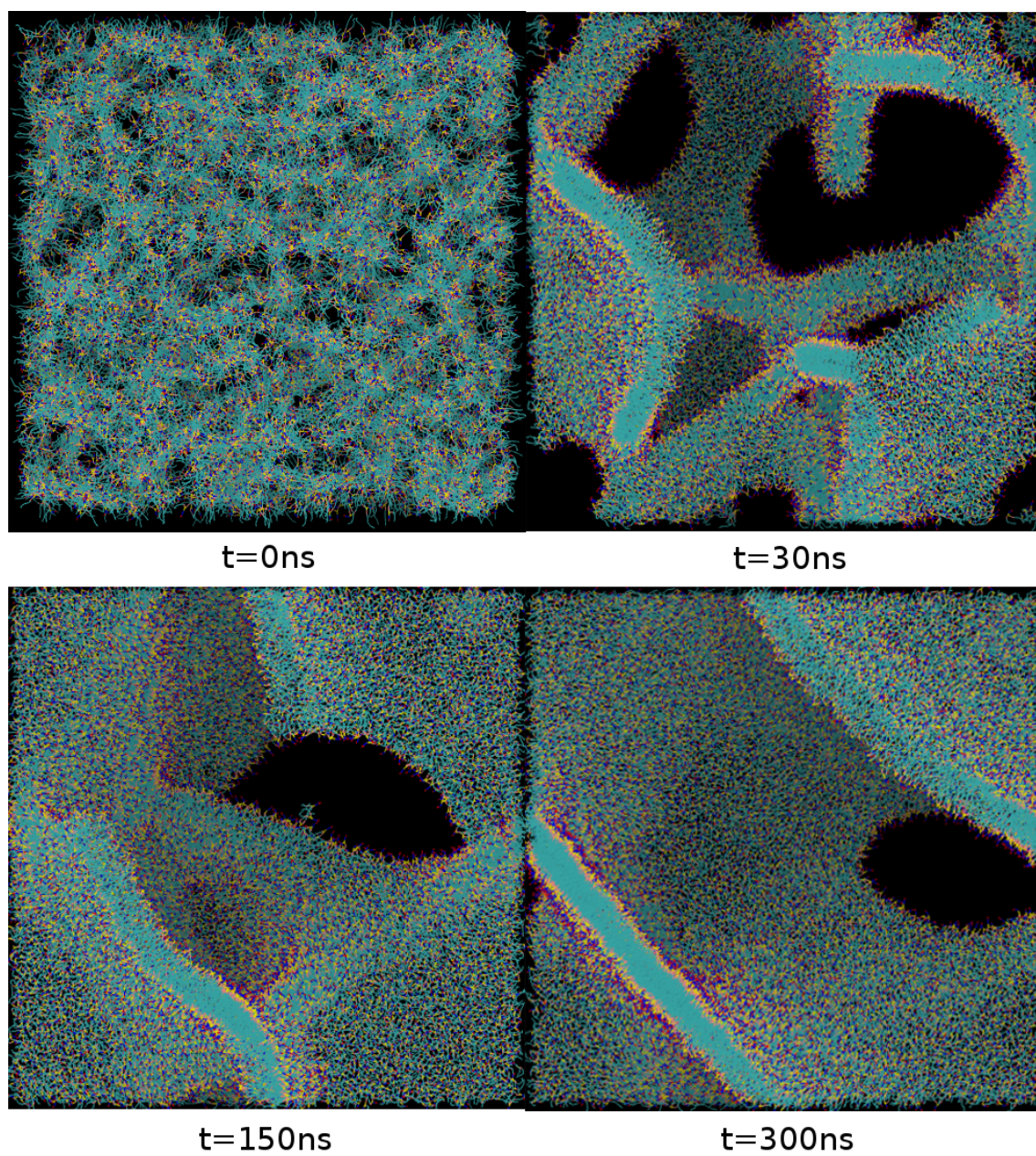
*Fig. 44: Measurement of MORC16 bilayer lateral diffusion from simulation.*

Lateral diffusion refers to the lateral movement of lipids and proteins within the membrane. In order to measure this lateral diffusion, we ran a 300 ns simulation of the MORC16 lipid bilayer, removing any jumps caused by the box boundaries. The lateral diffusion was calculated using *g\_msd* in GROMACS. Fig. 44 shows that before 250 ns, the lipids move at a steady speed ( $\sim 2500 \mu\text{m}^2/\text{s}$ ), then the movement significantly accelerates after 250 ns. We were unable to run any longer to see the whole picture of the

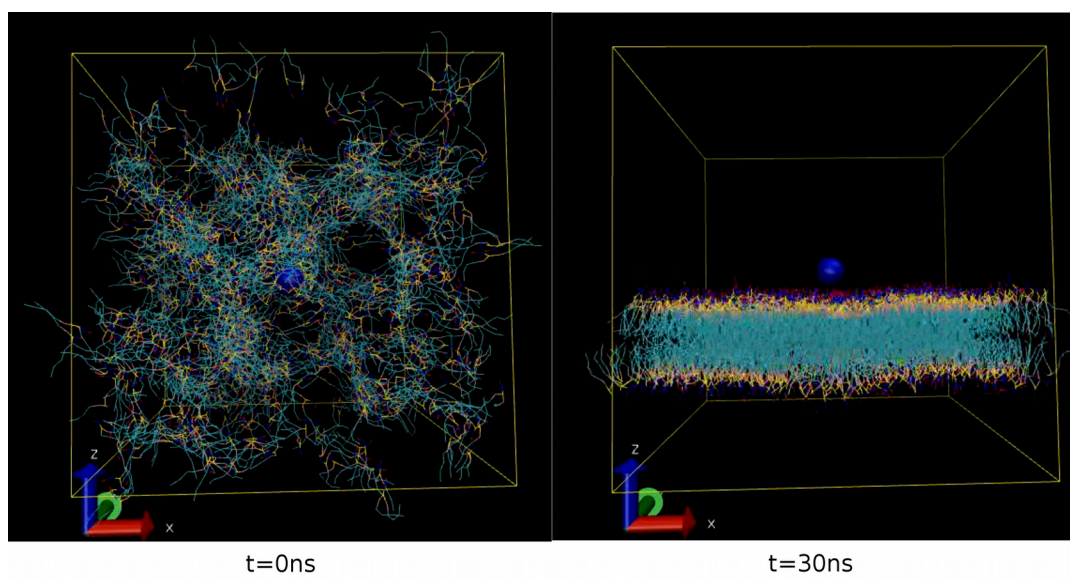
movement, but we anticipate there may be some fluctuation of lateral diffusion along timeline.

In an attempt to model a small vesicle, we increased the number of lipids to 12800 and the size of water box correspondingly. The total simulation time was 300 ns, and the snapshots are shown in Fig. 45.

Periodic boundary conditions (PBC) are useful for approximating the behavior of a macro-scale system, however, it can also introduce correlational artifacts. Using PBC, one side of lipid bilayer may interact with the other side of bilayer, which creates a crystal-like structure, freezing the system into a local minimum. However, one solution to this problem is to increase the size of simulation box, this will increase the calculation cost significantly. The second option is to apply some constrains or introduce a bias to avoid this artifacts. In an effort to bias the system, we added ions to the simulation box, and fixed the ions in the center of the box. The purpose was to see if the addition of ions would nucleate the formation of bilayer around the ions, thus inducing the formation of a micelle or liposome.

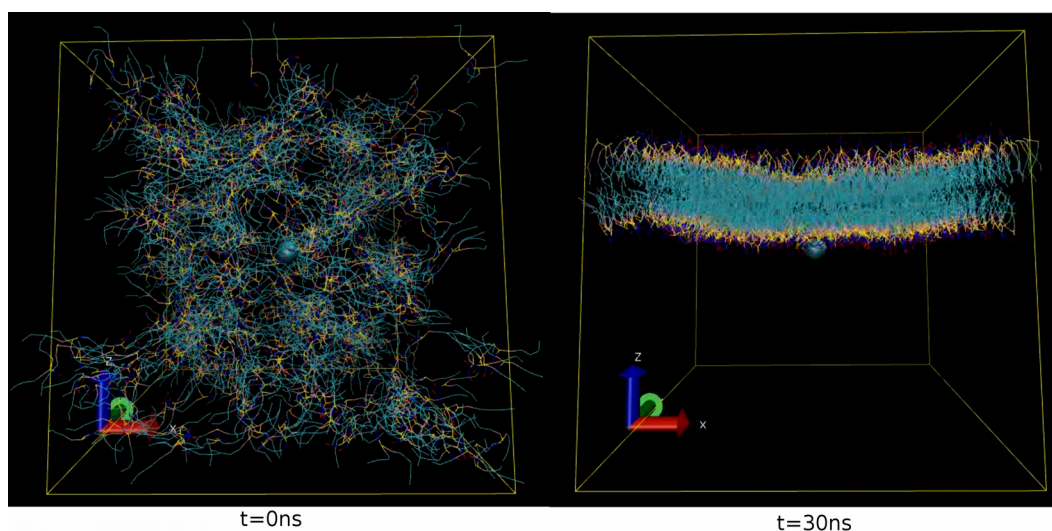


*Fig. 45: Snapshots from a simulation of aggregation of 12800 MORC16 lipids.*



*Fig. 46: Snapshots from the self-assembly simulation of MORC16 with  $\text{Na}^+$  (blue sphere).*





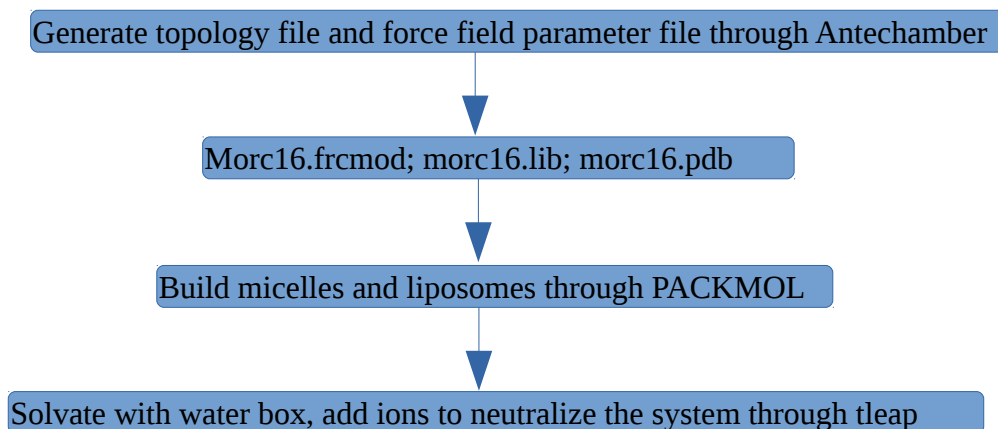
*Fig. 47: Snapshots from self-assembly simulation of MORC16 with Cl<sup>-</sup>(cyan sphere).*

We expected these ions would act as a nucleation center, gathering lipids around them through charge-charge interaction. However, the results show that no matter how many charges on the ions (1-10), it is very difficult to influence the shape of lipid bilayer.

#### **4.2.3 Stability of MORC16 Micelles and Liposomes by AMBER**

The characterization of stable liposome is important for drug delivery. Many factors such as temperature, cholesterol content and pH etc. could influence liposome stability. Here we tested the physical stability ('compactness' to be more specific) of micelles and liposomes by pulling one lipid out from the surface and pulling ions go through the surface from outside to inside, The MARTINI model is fast compared to most all-atom models, but still not fast enough for us to study larger system like liposomes. In order to

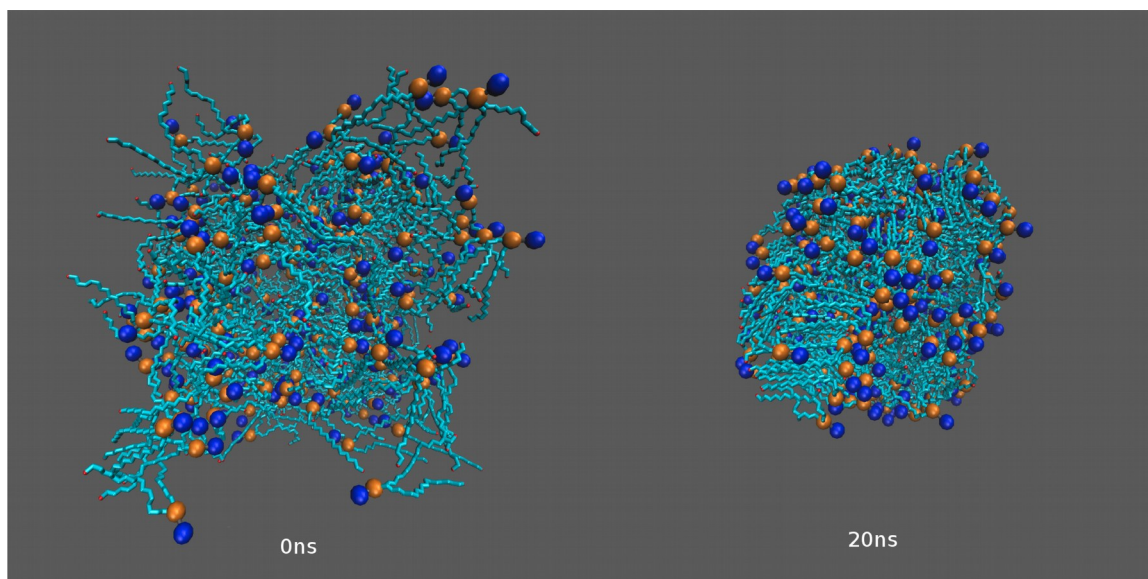
break this computertime barrier, we turned our focus to AMBER, which has very good support of CUDA (on graphic cards) acceleration. The speed can be 10-100 times faster than CPU based methods. In addition to the speed up, AMBER supports pH-based simulation natively. Our plan was to first run non-pH simulation of liposome with all-atom model in AMBER, then run constant-pH simulation. The flow chart to prepare MORC16 lipid simulation in AMBER is as following:



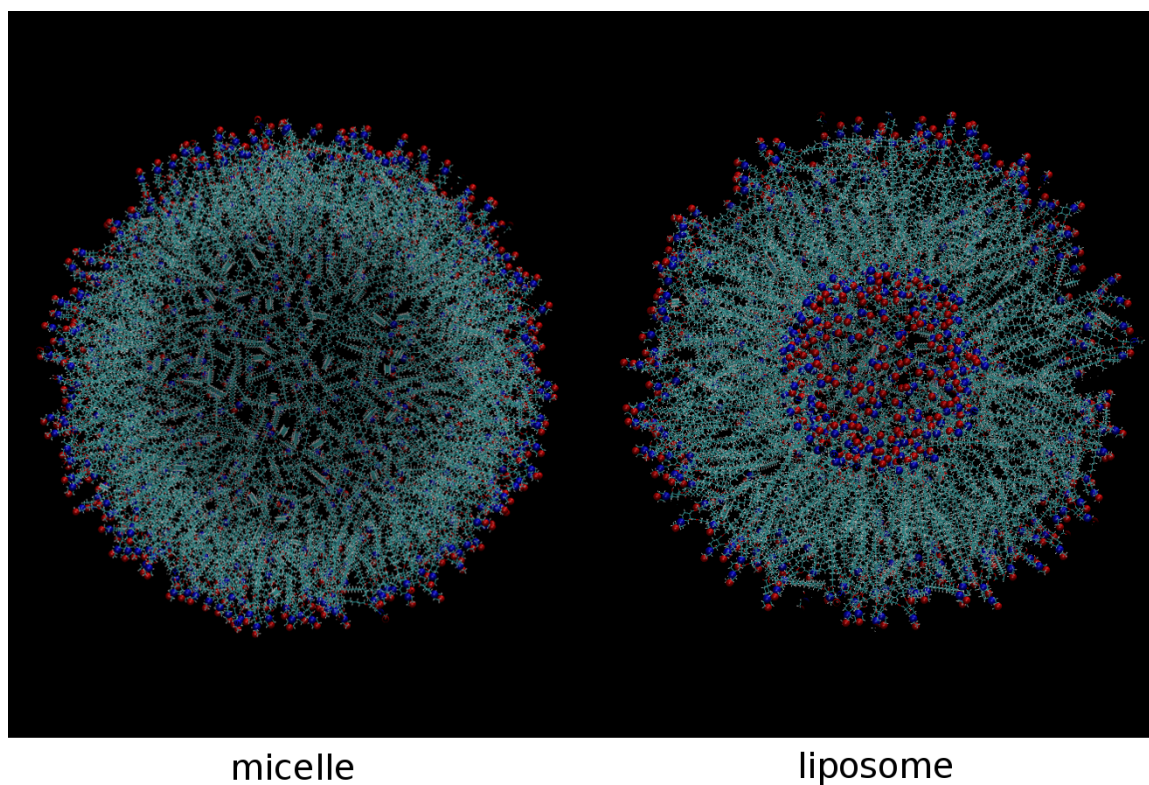
As an auxiliary program for MD simulation, Antechamber can generate input files automatically for most organic molecules. But one must always check the generated prmtop file for safety. In this research there was something wrong with charge assignment (total charge  $\neq 0$ ). We changed partial charge of atom "O" from -0.417600 to -0.421600 to make the total charge to be 0 to fix this problem.

PACKMOL is a tool to create an initial structure of variety type of ordered system, and helps guarantee that short-range repulsive interactions will not destroy the created system. The input file for PACKMOL is *morc16\_micelle\_tleap.in*. We were hoping to get

micelles and liposomes from aggregation simulation, but later found it is not feasible from random lipids, the speed is slow and the system stuck into a tangled conformation (Fig. 48), so we build the micelles and liposomes from PACKMOL (Fig. 49).



*Fig. 48: The self assembly simulation of MORC16 lipids(250) in AMBER.*

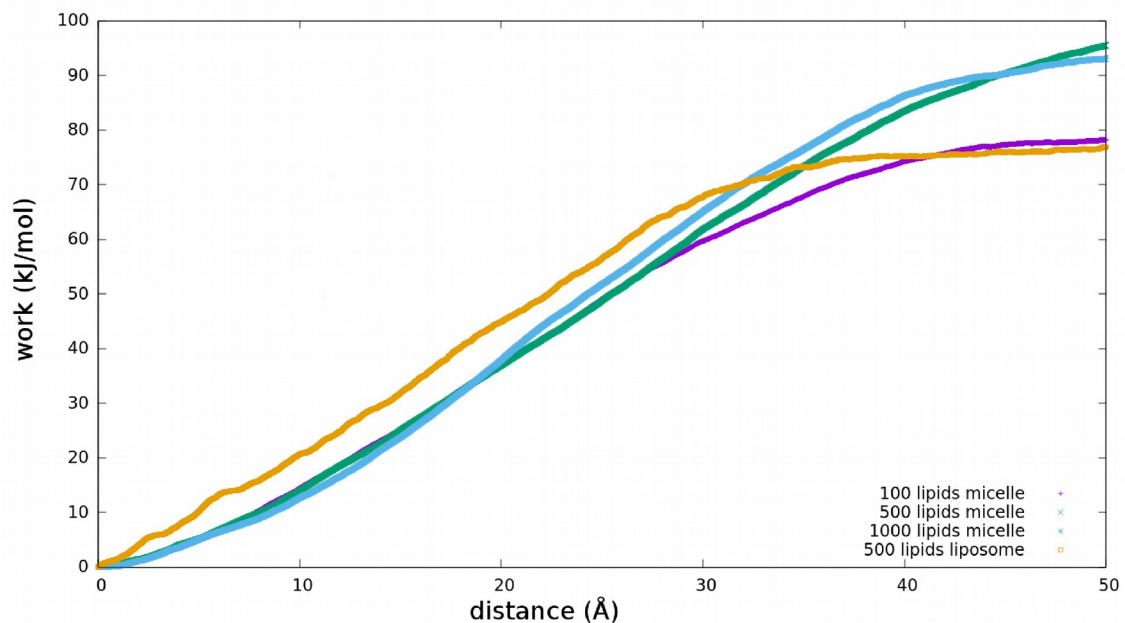


*Fig. 49: Cross section of MORC16 micelle and liposome built by PACMOL.*

### **Steered Molecular Dynamics Analysis of MORC16 Vesicles**

Two types of SMD simulations were performed: first pulls one lipid out of the micelle or liposome surface; the second is pulls one ion from the outside to the inside of the micelle or liposome. We ran an average SMD of 10 lipids, the pulling speed is  $50\text{\AA} / 1000000$  steps =  $25\text{\AA} / \text{ns}$  (1step = 0.002ps).

### Pull One Lipid Out From Micelle and Liposome Surface



*Fig. 50: PMF of pulling one lipid from micelle and liposome surface.*

Fig. 50 shows that the work required to pull one lipid from liposome is slightly higher than micelle. The liposome has a bilayer structure, which makes each layer more stable than a single layer micelle. The bigger the micelle the more work is required since bigger micelles are more dense and tight, but this difference is not significant.

### Pull Ions Through Micelle and Liposome Surface

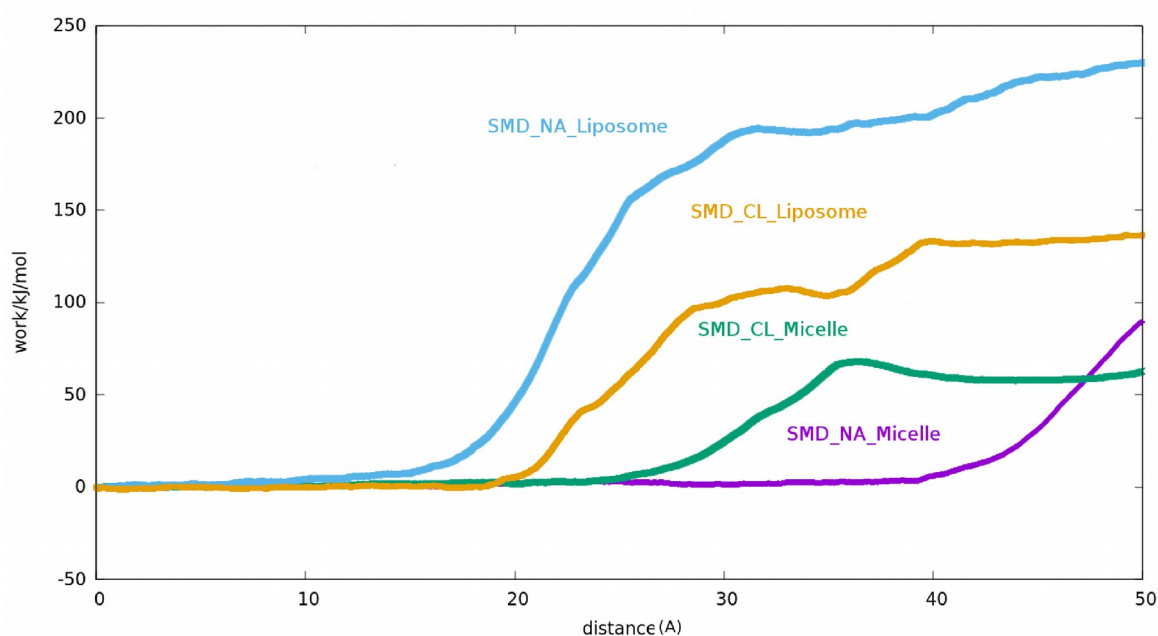
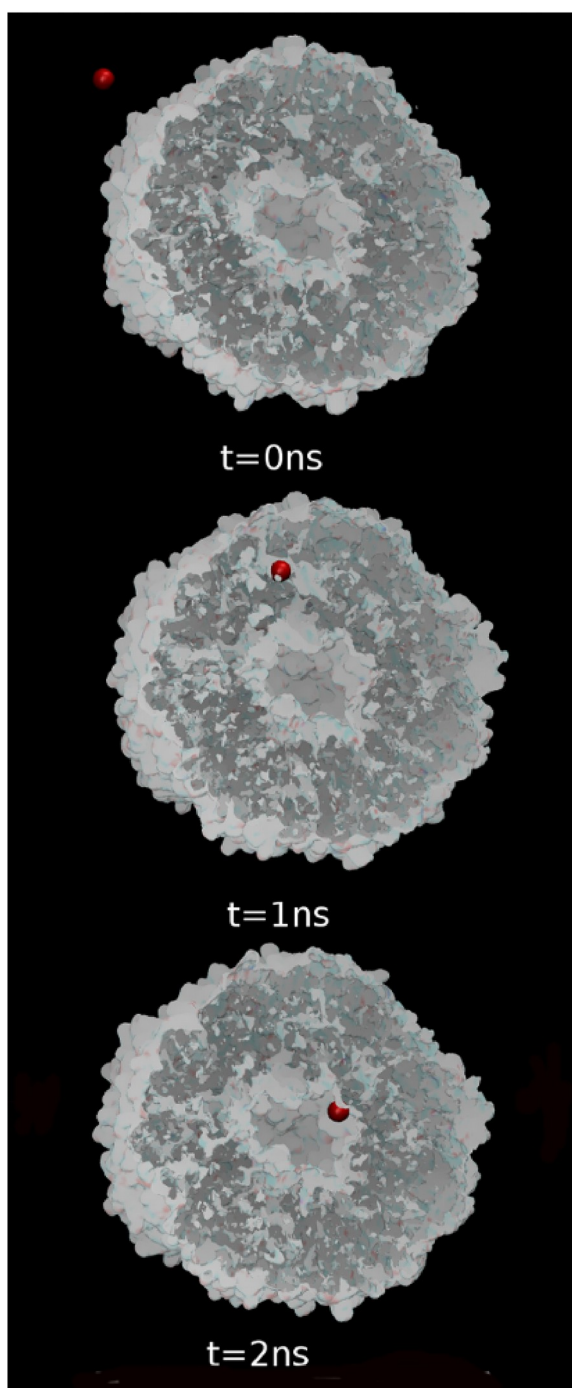


Fig. 51: PMF of pulling one ion through micelle (100 lipids).

Fig. 51 shows that even though  $\text{Na}^+$  is smaller than  $\text{Cl}^-$ , the work required for  $\text{Na}^+$  to transport through the micelle and liposome is higher than  $\text{Cl}^-$ . This may be because the heads of lipid carry more negative charge (through the oxygen atoms), and therefore has a higher charge-charge interaction with  $\text{Na}^+$ , making it more difficult for  $\text{Na}^+$  to cross the surface. The snapshots of  $\text{Na}^+$  travelling through a liposome are shown in Fig.52. The SMD of  $\text{Na}^+$  through micelle failed, because the  $\text{Na}^+$  circled around the micelle rather than go into the core of micelle.



*Fig. 52: Snapshots of the pulling of  $\text{Na}^+$  through a liposome.*

#### 4.4 Discussion and Conclusion

In this chapter, We parameterized and modeled a newly synthesized lipid, MORC16, and performed constant-pH simulations on systems of MORC16 using CHARMM. We were able to determine the pKa from titration curve, and even though it matched experimental value very well, the “correct” behavior of molecule in different pH was not observed.

One conclusion from the simulation is the protonation of N atom will not lead to formation of hydrogen bond between N and O1, in order for the chair conformation of the ring to flip at least in our model. Even if the hydrogen bond does form, the energy will not be enough to induce the flip of the chair conformation.

The simulation of large vesicles of micelles and liposomes are still challenging. We have not considered a pH based simulation of a liposome yet, as the computing cost has already exceeded our capability. The promising way to simulate function of pH sensitive liposome is to develop a feasible method for constant pH simulation using a coarse grained model which is still under development<sup>126</sup>.



## REFERENCES

1. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of State Calculations by Fast Computing Machines. *J Chem Phys*. 1953;21(6):1087-1092. doi:10.1063/1.1699114.
2. Alder BJ, Wainwright TE. Phase Transition for a Hard Sphere System. *J Chem Phys*. 1957;27(5):1208-1209. doi:10.1063/1.1743957.
3. Rahman A. Correlation in the Motion of Atoms in Liquid Argon. *Phys Rev*. 1964;136(2A):405-411. doi:10.1103/PhysRev.136.A405.
4. Rahman A, Stillinger FHJ. Molecular dynamics study of liquid water. *J Chem Phys*. 1971;55(7):3336-3359. doi:10.1063/1.1676585.
5. Levitt M, Warshel A. Computer simulation of protein folding. *Nature*. 1975;253(5494):694-698. doi:10.1038/253694a0.
6. McCammon JA, Gelin BR, Karplus M. Dynamics of folded proteins. *Nature*. 1977;267(5612):585-590. doi:10.1038/267585a0.
7. Karplus M, McCammon JA. Molecular dynamics simulations of biomolecules. *Nat Struct Biol*. 2002;9(9):646-652. doi:10.1038/nsb0902-646.
8. Fogolari F, Brigo A, Molinari H. The Poisson-Boltzmann equation for biomolecular electrostatics: A tool for structural biology. *J Mol Recognit*. 2002;15(6):377-392. doi:10.1002/jmr.577.
9. Tsui V, Case DA. Theory and applications of the Generalized Born solvation model in macromolecular simulations. *Biopolymers*. 2000;56(4):275-291. doi:10.1002/1097-0282(2000)56:4<275::AID-BIP10024>3.0.CO;2-E.
10. Kästner J. Umbrella sampling. *Wiley Interdiscip Rev Comput Mol Sci*. 2011;1(6):932-942. doi:10.1002/wcms.66.
11. Sugita Y, Okamoto Y. Replica-exchange molecular dynamics method for protein folding. *Chem Phys Lett*. 1999;314(1-2):141-151. doi:10.1016/S0009-2614(99)01123-9.
12. Shkurti A, Orsi M, MacLi E, Ficarra E, Acquaviva A. Acceleration of coarse grain molecular dynamics on GPU architectures. *J Comput Chem*. 2013;34(10):803-818. doi:10.1002/jcc.23183.

13. Levine BG, Lebard DN, Devane R, Shinoda W, Kohlmeyer A, Klein ML. Micellization studied by GPU-accelerated coarse-grained molecular dynamics. *J Chem Theory Comput.* 2011;7(12):4135-4145. doi:10.1021/ct2005193.
14. Glaser J, Nguyen TD, Anderson JA, et al. Strong scaling of general-purpose molecular dynamics simulations on GPUs. *Comput Phys Commun.* 2015;192:97-107. doi:10.1016/j.cpc.2015.02.028.
15. Zheng M, Li X, Guo L. Algorithms of GPU-enabled reactive force field (ReaxFF) molecular dynamics. *J Mol Graph Model.* 2013;41:1-11. doi:10.1016/j.jm gm.2013.02.001.
16. Harger M, Li D, Wang Z, et al. Tinker-OpenMM: Absolute and relative alchemical free energies using AMOEBA on GPUs. *J Comput Chem.* 2017;38(23):2047-2055. doi:10.1002/jcc.24853.
17. Rodrigues C, Hardy D, Stone J, Schulten K, Hwu W-MW. GPU acceleration of cutoff pair potentials for molecular modeling applications. *5th Conf Comput Front.* 2008:273-282. doi:10.1145/1366230.1366277.
18. Anderson JA, Lorenz CD, Travesset A. General purpose molecular dynamics simulations fully implemented on graphics processing units. *J Comput Phys.* 2008;227(10):5342-5359. doi:10.1016/j.jcp.2008.01.047.
19. Stone JE, Phillips JC, Freddolino PL, Hardy DJ, Trabuco LG, Schulten K. Accelerating molecular modeling applications with graphics processors. *J Comput Chem.* 2007;28(16):2618-2640. doi:10.1002/jcc.20829.
20. Macklin DN, Ruggero NA, Covert MW. The future of whole-cell modeling. *Curr Opin Biotechnol.* 2014;28:111-115. doi:10.1016/j.copbio.2014.01.012.
21. Karr JR, Sanghvi JC, MacKlin DN, et al. A whole-cell computational model predicts phenotype from genotype. *Cell.* 2012;150(2):389-401. doi:10.1016/j.cell.2012.05.044.
22. Jones JE. On the Determination of Molecular Fields. I. From the Variation of the Viscosity of a Gas with Temperature. *Proc R Soc A Math Phys Eng Sci.* 1924;106(738):441-462. doi:10.1098/rspa.1924.0081.
23. Morse PM. Diatomic molecules according to the wave mechanics. II. Vibrational levels. *Phys Rev.* 1929;34(1):57-64. doi:10.1103/PhysRev.34.57.

24. Daw MS, Baskes MI. Embedded-atom method: Derivation and application to impurities, surfaces, and other defects in metals. *Phys Rev B*. 1984;29(12):6443-6453. doi:10.1103/PhysRevB.29.6443.
25. Cornell WD, Cieplak P, Bayly CI, et al. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J Am Chem Soc*. 1995;117(19):5179-5197. doi:10.1021/ja00124a002.
26. Vanommeslaeghe K, Hatcher E, Acharya C, et al. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J Comput Chem*. 2010;31(4):671-690. doi:10.1002/jcc.21367.
27. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J Comput Chem*. 1983;4(2):187-217. doi:10.1002/jcc.540040211.
28. MacKerell AD, Bashford D, Bellott M, et al. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins †. *J Phys Chem B*. 1998;102(18):3586-3616. doi:10.1021/jp973084f.
29. K. N. The Urey — Bradley Force Field: Its Significance and Application. *Dev Appl Spectrosc Dev Appl Spectrosc*. 1964;3. doi:https://doi.org/10.1007/978-1-4684-8688-9\_12.
30. Buck M, Bouguet-Bonnet S, Pastor RW, MacKerell AD. Importance of the CMAP correction to the CHARMM22 protein force field: Dynamics of hen lysozyme. *Biophys J*. 2006;90(4). doi:10.1529/biophysj.105.078154.
31. Halgren T a. Merck Molecular Force Field. *J Comput Chem*. 1996;17(5-6):490-519. doi:10.1002/(SICI)1096-987X(199604)17:5/6<520::AID-JCC2>3.0.CO;2-W.
32. Voter AF. Parallel replica method for dynamics of infrequent events. *Phys Rev B - Condens Matter Mater Phys*. 1998;57(22):R13985-R13988. doi:10.1103/PhysRevB.57.R13985.
33. Sørensen MR, Voter AF. Temperature-accelerated dynamics for simulation of infrequent events. *J Chem Phys*. 2000;112(21):9599-9606. doi:10.1063/1.481576.
34. Voter AF. Hyperdynamics: Accelerated molecular dynamics of infrequent events. *Phys Rev Lett*. 1997;78(20):3908-3911. doi:10.1103/PhysRevLett.78.3908.

35. Isralewitz B, Gao M, Schulten K. Steered molecular dynamics and mechanical functions of proteins. *Curr Opin Struct Biol.* 2001;11(2):224-230. doi:10.1016/S0959-440X(00)00194-9.
36. Grubm ller H, Heymann B, Tavan P. Ligand Binding: Molecular Mechanics Calculation of the Streptavidin-Biotin Rupture Force. *Science (80- ).* 1996;271(5251):997-999. doi:10.1126/science.271.5251.997.
37. Jensen MØ, Yin Y, Tajkhorshid E, Schulten K. Sugar transport across lactose permease probed by steered molecular dynamics. *Biophys J.* 2007;93(1):92-102. doi:10.1529/biophysj.107.103994.
38. De Fabritiis G, Coveney P V., Villà-Freixa J. Energetics of K<sup>+</sup> permeability through Gramicidin A by forward-reverse steered molecular dynamics. *Proteins Struct Funct Genet.* 2008;73(1):185-194. doi:10.1002/prot.22036.
39. Gao M, Craig D, Vogel V, Schulten K. Identifying unfolding intermediates of FN-III10 by steered molecular dynamics. *J Mol Biol.* 2002;323(5):939-950. doi:10.1016/S0022-2836(02)01001-X.
40. Lu H, Isralewitz B, Krammer A, Vogel V, Schulten K. Unfolding of Titin Immunoglobulin Domains by Steered Molecular Dynamics Simulation. *Biophys J.* 1998;75(2):662-671. doi:10.1016/S0006-3495(98)77556-3.
41. Nadler W, Hansmann UHE. Optimizing replica exchange moves for molecular dynamics. *Phys Rev E - Stat Nonlinear, Soft Matter Phys.* 2007;76(5). doi:10.1103/PhysRevE.76.057102.
42. Kamberaj H, van der Vaart A. An optimized replica exchange molecular dynamics method. *J Chem Phys.* 2009;130(7):74906. doi:10.1063/1.3077857.
43. Swendsen RH, Wang JS. Replica Monte Carlo simulation of spin-glasses. *Phys Rev Lett.* 1986;57(21):2607-2609. doi:10.1103/PhysRevLett.57.2607.
44. Wang J-S, Swendsen RH. Replica Monte Carlo Simulation (Revisited). *Exch Organ Behav Teach J.* 2004;157(1):7. doi:10.1143/PTPS.157.317.
45. Marinari E, Parisi G, Ruiz-Lorenzo J, Ritort F. Numerical evidence for spontaneously broken replica symmetry in 3D spin glasses. *Phys Rev Lett.* 1996;76(5):843-846. doi:10.1103/PhysRevLett.76.843.
46. Papakonstantinou T, Malakis a. Parallel tempering and 3D spin glass models. *J Phys Conf Ser.* 2014;487(1):12010. doi:10.1088/1742-6596/487/1/012010.

47. Kennedy D, Norman C. What don't we know? *Science (80- )*. 2005;309(5731):75. doi:10.1126/science.309.5731.75.
48. Anfinsen CB, Haber E, Sela M, White FH. THE KINETICS OF FORMATION OF NATIVE RIBONUCLEASE DURING OXIDATION OF THE REDUCED POLYPEPTIDE CHAIN. *Proc Natl Acad Sci*. 1961;47(9):1309-1314. doi:10.1073/pnas.47.9.1309.
49. Zhou R. Hydrophobic Collapse in Multidomain Protein Folding. *Science (80- )*. 2004;305(5690):1605-1609. doi:10.1126/science.1101176.
50. Ferguson N, Fersht AR. Early events in protein folding. *Curr Opin Struct Biol*. 2003;13(1):75-81. doi:10.1016/S0959-440X(02)00009-X.
51. Fersht a R. Nucleation mechanisms in protein folding. *Curr Opin Struct Biol*. 1997;7:3-9. doi:10.1016/S0959-440X(97)80002-4.
52. Kim P, Baldwin R. Intermediates In The Folding Reactions Of Small Proteins. *Annu Rev Biochem*. 1990;59(1):631-660. doi:10.1146/annurev.biochem.59.1.631.
53. Karplus M, Weaver DL. Protein folding dynamics: the diffusion-collision model and experimental data. *Protein Sci*. 1994;3(4):650-668. doi:10.1002/pro.5560030413.
54. Ciryam P, Morimoto RI, Vendruscolo M, Dobson CM, O'Brien EP. In vivo translation rates can substantially delay the cotranslational folding of the Escherichia coli cytosolic proteome. *Proc Natl Acad Sci*. 2013;110(2):E132-E140. doi:10.1073/pnas.1213624110.
55. Petukhov M, Cregut D, Soares CM, Serrano L. Local water bridges and protein conformational stability. *Protein Sci*. 1999;8(10):1982-1989. doi:10.1110/ps.8.10.1982.
56. Utas JE, Kritikos M, Sandström D, Kermack B. Water as a hydrogen bonding bridge between a phenol and imidazole. A simple model for water binding in enzymes. *Biochim Biophys Acta - Bioenerg*. 2006;1757(12):1592-1596. doi:10.1016/j.bbabi.2006.08.005.
57. Kauzmann W. Some factors in the interpretation of protein denaturation. *Adv Protein Chem*. 1959;14:1-63.

58. Ben-Naim A. Inversion of the hydrophobic/hydrophilic paradigm demystifies the protein folding and self-assembly of problems. *Int J Phys*. 2013;1(3):66-71. doi:10.12691/ijp-1-3-2.
59. Xantheas SS. Cooperativity and hydrogen bonding network in water clusters. *Chem Phys*. 2000;258(2-3):225-231. doi:10.1016/S0301-0104(00)00189-0.
60. Maheshwary S, Patel N, Sathyamurthy N, Kulkarni AD, Gadre SR. Structure and stability of water clusters (H<sub>2</sub>O)(n), n=8-20: An ab initio investigation. *J Phys Chem A*. 2001;105(46):10525-10537. doi:10.1021/jp013141b.
61. Keutsch FN, Saykally RJ. Water clusters: Untangling the mysteries of the liquid, one molecule at a time. *Proc Natl Acad Sci*. 2001;98(19):10533-10540. doi:10.1073/pnas.191266498.
62. Rao F, Garrett-Roe S, Hamm P. Structural inhomogeneity of water by complex network analysis. *J Phys Chem B*. 2010;114(47):15598-15604. doi:10.1021/jp1060792.
63. Ben Ishai P, Tripathi SR, Kawase K, Puzenko A, Feldman Y. What is the primary mover of water dynamics? *Phys Chem Chem Phys*. 2015;17(23):15428-15434. doi:10.1039/C5CP01871D.
64. Tuckerman M, Laasonen K, Sprik M, Parrinello M. Ab Initio Molecular Dynamics Simulation of the Solvation and Transport of H<sub>3</sub>O<sup>+</sup> and OH<sup>-</sup> Ions in Water. *J Phys Chem*. 1995;99(16):5749-5752. doi:10.1021/j100016a003.
65. Starr FW, Nielsen JK, Stanley HE. Fast and Slow Dynamics of Hydrogen Bonds in Liquid Water. *Phys Rev Lett*. 1999;82(11):2294-2297. doi:10.1103/PhysRevLett.82.2294.
66. Zhao L, Ma K, Yang Z. Changes of Water Hydrogen Bond Network with Different Externalities. *Int J Mol Sci*. 2015;16(4):8454-8489. doi:10.3390/ijms16048454.
67. Marcus Y. Effect of ions on the structure of water. *Pure Appl Chem*. 2010;82(10). doi:10.1351/PAC-CON-09-07-02.
68. Urbic T. Ions increase strength of hydrogen bond in water. *Chem Phys Lett*. 2014;610-611:159-162. doi:10.1016/j.cplett.2014.06.054.
69. O'Brien JT, Williams ER. Effects of Ions on Hydrogen-Bonding Water Networks in Large Aqueous Nanodrops. *J Am Chem Soc*. 2012;134(24):10228-10236. doi:10.1021/ja303191r.

70. Bondar A-N, White SH. Hydrogen bond dynamics in membrane protein function. *Biochim Biophys Acta - Biomembr.* 2012;1818(4):942-950. doi:10.1016/j.bbamem.2011.11.035.
71. Zhao L, Ma K, Yang Z. Changes of water hydrogen bond network with different externalities. *Int J Mol Sci.* 2015;16(4):8454-8489. doi:10.3390/ijms16048454.
72. Sreerama N, Woody RW. Molecular dynamics simulations of polypeptide conformations in water: A comparison of alpha, beta, and poly(pro)II conformations. *Proteins.* 1999;36(4):400-406.
73. Law PB, Daggett V. The relationship between water bridges and the polyproline II conformation: a large-scale analysis of molecular dynamics simulations and crystal structures. *Protein Eng Des Sel.* 2010;23(1):27-33. doi:10.1093/protein/gzp069.
74. Ahmad M, Gu W, Geyer T, Helms V. Adhesive water networks facilitate binding of protein interfaces. *Nat Commun.* 2011;2:261. doi:10.1038/ncomms1258.
75. de la Lande A, Babcock NS, Rezac J, Sanders BC, Salahub DR. Surface residues dynamically organize water bridges to enhance electron transfer between proteins. *Proc Natl Acad Sci.* 2010;107(26):11799-11804. doi:10.1073/pnas.0914457107.
76. Yonetani Y, Kono H. What determines water-bridge lifetimes at the surface of DNA? Insight from systematic molecular dynamics analysis of water kinetics for various DNA sequences. *Biophys Chem.* 2012;160(1):54-61. doi:10.1016/j.bpc.2011.09.006.
77. Schröder C, Rudas T, Boresch S, Steinhauser O. Simulation studies of the protein-water interface. I. Properties at the molecular resolution. *J Chem Phys.* 2006;124(23):234907. doi:10.1063/1.2198802.
78. Kovacs H, Mark AE, van Gunsteren WF. Solvent structure at a hydrophobic protein surface. *Proteins.* 1997;27(3):395-404.
79. Kumar R, Schmidt JR, Skinner JL. Hydrogen bonding definitions and dynamics in liquid water. *J Chem Phys.* 2007;126(20):204107. doi:10.1063/1.2742385.
80. Bakó I, Megyes T, Bálint S, et al. Hydrogen bonded network properties in liquid formamide. *J Chem Phys.* 2010;132(1):14506. doi:10.1063/1.3268626.
81. Bakó I, Bencsura A, Hermansson K, et al. Hydrogen bond network topology in liquid water and methanol: a graph theory approach. *Phys Chem Chem Phys.* 2013;15(36):15163-15171. doi:10.1039/c3cp52271g.

82. Kaiser A, Ismailova O, Koskela A, et al. Ethylene glycol revisited: Molecular dynamics simulations and visualization of the liquid and its hydrogen-bond network. *J Mol Liq.* 2014;189:20-29. doi:10.1016/j.molliq.2013.05.033.
83. Bentley JL. An Introduction to Algorithm Design. *Computer (Long Beach Calif).* 1979;12(2):66-78. doi:10.1109/MC.1979.1658620.
84. Marcus Y. Effect of ions on the structure of water: Structure making and breaking. *Chem Rev.* 2009;109(3):1346-1370. doi:10.1021/cr8003828.
85. Ben-Naim A. *Molecular Theory of Water and Aqueous Solutions Part II.*; 2011.
86. Arieh Ben-Naim. *Molecular Theory of Water and Aqueous Solutions Part I.*; 2009.
87. Adhikari B, Cheng J. Protein residue contacts and prediction methods. In: *Methods in Molecular Biology.* Vol 1415. ; 2016:463-476. doi:10.1007/978-1-4939-3572-7\_24.
88. NARAYANA SVL, ARGOS P. Residue contacts in protein structures and implications for protein folding. *Int J Pept Protein Res.* 1984;24(1):25-39. doi:10.1111/j.1399-3011.1984.tb00924.x.
89. Miyazawa S, Jernigan RL. 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):623-644. doi:10.1006/jmbi.1996.0114.
90. Warne PK, Morgan RS. A survey of atomic interactions in 21 proteins. *J Mol Biol.* 1978;118(3):273-287. doi:10.1016/0022-2836(78)90228-0.
91. Tanaka S, Scheraga H a. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules.* 1976;9(6):945-950. doi:10.1021/ma60054a013.
92. Joo H, Chavan AG, Phan J, Day R, Tsai J. An amino acid packing code for  $\alpha$ -helical structure and protein design. *J Mol Biol.* 2012;419(3-4):234-254. doi:10.1016/j.jmb.2012.03.004.
93. Joo H, Tsai J. An amino acid code for  $\beta$ -sheet packing structure. *Proteins Struct Funct Bioinforma.* 2014;82(9):2128-2140. doi:10.1002/prot.24569.
94. Matthew JB, Gurd FR, Garcia-Moreno B, Flanagan M a, March KL, Shire SJ. pH-dependent processes in proteins. *CRC Crit Rev Biochem.* 1985;18(2):91-197. doi:10.3109/10409238509085133.



95. Zhang M, Vogel HJ. Determination of the side chain pK(a) values of the lysine residues in calmodulin. *J Biol Chem*. 1993;268(30):22420-22428.
96. Pascal JW, Dorman DE. Determination of pKa Values Using <sup>15</sup>N and <sup>13</sup>C Nuclear Magnetic Resonance Spectroscopy. The case of Apramycin. *Org Magn Reson*. 1978;11(12):632-634. doi:10.1002/mrc.1270111210.
97. Tossell JA. Boric acid, "carbonic" acid, and N-containing oxyacids in aqueous solution: Ab initio studies of structure, pKa, NMR shifts, and isotopic fractionations. *Geochim Cosmochim Acta*. 2005;69(24):5647-5658. doi:10.1016/J.GCA.2005.08.004.
98. Sheff JT, Lucius AL, Owens SB, Gray GM. Generally applicable NMR titration methods for the determination of equilibrium constants for coordination complexes: Syntheses and characterizations of metallacrown ethers with  $\alpha,\omega$ -bis(phosphite)- Polyether ligands and determination of equilibrium binding. *Organometallics*. 2011;30(21):5695-5709. doi:10.1021/om200580t.
99. Tomlinson JH, Green VL, Baker PJ, Williamson MP. Structural origins of pH-dependent chemical shifts in the B1 domain of protein G. *Proteins Struct Funct Bioinforma*. 2010;78(14):3000-3016. doi:10.1002/prot.22825.
100. Lee AC, Crippen GM. Predicting pKa. *J Chem Inf Model*. 2009;49(9):2013-2033. doi:10.1021/ci900209w.
101. Baptista AM, Martel PJ, Petersen SB. Simulation of protein conformational freedom as a function of pH: Constant-pH molecular dynamics using implicit titration. *Proteins Struct Funct Genet*. 1997;27(4):523-544. doi:10.1002/(SICI)1097-0134(199704)27:4<523::AID-PROT6>3.0.CO;2-B.
102. Börjesson U, Hünenberger PH. pH-dependent stability of a decalysine  $\alpha$ -helix studied by explicit-solvent molecular dynamics simulations at constant pH. *J Phys Chem B*. 2004;108(35):13551-13559. doi:10.1021/jp037841n.
103. Lee MS, Salsbury FR, Brooks CL. Constant-pH molecular dynamics using continuous titration coordinates. *Proteins Struct Funct Genet*. 2004;56(4):738-752. doi:10.1002/prot.20128.
104. Mertz JE, Pettitt BM. Molecular dynamics at a constant pH. *Int J High Perform Comput Appl*. 1994;8(1):47-53. doi:10.1177/109434209400800106.















































105. Baptista AM, Teixeira VH, Soares CM. Constant-pH molecular dynamics using stochastic titration. *J Chem Phys.* 2002;117(9):4184. doi:10.1063/1.1497164.
106. Długosz M, Antosiewicz JM. Constant-pH molecular dynamics simulations: a test case of succinic acid. *Chem Phys.* 2004;302(1–3):161-170. doi:http://dx.doi.org/10.1016/j.chemphys.2004.03.031.
107. Długosz M, Antosiewicz JM, Robertson AD. Constant-pH molecular dynamics study of protonation-structure relationship in a heptapeptide derived from ovomucoid third domain. *Phys Rev E - Stat Nonlinear, Soft Matter Phys.* 2004;69(2 1). doi:10.1103/PhysRevE.69.021915.
108. Thurlkill RL, Grimsley GR, Scholtz JM, Pace CN. pK values of the ionizable groups of proteins. *Protein Sci.* 2006;15(5):1214-1218. doi:10.1110/ps.051840806.
109. Pace CN, Grimsley GR, Scholtz JM. Protein ionizable groups: pK values and their contribution to protein stability and solubility. *J Biol Chem.* 2009;284(20):13285-13289. doi:10.1074/jbc.R800080200.
110. Isom DG, Castaneda CA, Cannon BR, Garcia-Moreno E. B. Large shifts in pKa values of lysine residues buried inside a protein. *Proc Natl Acad Sci.* 2011;108(13):5260-5265. doi:10.1073/pnas.1010750108.
111. Kox a. J, Michels JPJ, Wiegel FW. Simulation of a lipid monolayer using molecular dynamics. *Nature.* 1980;287(5780):317-319. doi:10.1038/287317a0.
112. Heller H, Schaefer M, Schulten K. Molecular-Dynamics Simulation of A Bilayer of 200 Lipids in the Gel and in the Liquid-Crystal Phases. *J Phys Chem.* 1993;97(31):8343-8360. doi:10.1021/j100133a034.
113. Marrink SJ, de Vries AH, Mark AE. Coarse Grained Model for Semiquantitative Lipid Simulations. *J Phys Chem B.* 2004;108(2):750-760. doi:10.1021/jp036508g.
114. Stevens MJ. Coarse-grained simulations of lipid bilayers. *J Chem Phys.* 2004;121(23):11942-11948. doi:10.1063/1.1814058.
115. Lu L, Voth GA. Systematic coarse-graining of a multicomponent lipid bilayer. *J Phys Chem B.* 2009;113(5):1501-1510. doi:10.1021/jp809604k.
116. Wang S, Larson RG. Coarse-grained molecular dynamics simulation of tethered lipid assemblies. *Soft Matter.* 2013;9(2):480-486. doi:10.1039/C2SM26850G.

117. Marrink SJ, Risselada HJ, Yefimov S, Tieleman DP, De Vries AH. The MARTINI force field: Coarse grained model for biomolecular simulations. *J Phys Chem B*. 2007;111(27):7812-7824. doi:10.1021/jp071097f.
118. Shih AY, Arkhipov A, Freddolino PL, Schulten K. Coarse grained protein-lipid model with application to lipoprotein particles. *J Phys Chem B*. 2006;110(8):3674-3684. doi:10.1021/jp0550816.
119. Orsi M, Haubertin DY, Sanderson WE, Essex JW. A quantitative coarse-grain model for lipid bilayers. *J Phys Chem B*. 2008;112(3):802-815. doi:10.1021/jp076139e.
120. Monticelli L, Kandasamy SK, Periole X, Larson RG, Tieleman DP, Marrink SJ. The MARTINI coarse-grained force field: Extension to proteins. *J Chem Theory Comput*. 2008;4(5):819-834. doi:10.1021/ct700324x.
121. Marrink SJ, Risselada HJ, Yefimov S, Tieleman DP, Vries AH De. The MARTINI Force Field : Coarse Grained Model for Biomolecular Simulations The MARTINI Force Field : Coarse Grained Model for Biomolecular Simulations. *J Phys Chem B*. 2007;111(June):7812-7824. doi:10.1021/jp071097f.
122. Samoshin V V. Flipsomes: Stimuli-triggered conformational flip of novel amphiphiles causes an instant cargo release from liposomes. *Biomol Concepts*. 2014;5(2):131-141. doi:10.1515/bmc-2014-0002.
123. Zoete V, Cuendet MA, Grosdidier A, Michielin O. SwissParam: A fast force field generation tool for small organic molecules. *J Comput Chem*. 2011;32(11):2359-2368. doi:10.1002/jcc.21816.
124. Halgren TA. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J Comput Chem*. 1996;17(5-6):490-519. doi:10.1002/(SICI)1096-987X(199604)17:5/6<490::AID-JCC1>3.0.CO;2-P.
125. Periole X, Marrink SJ. The martini coarse-grained force field. *Methods Mol Biol*. 2013;924:533-565. doi:10.1007/978-1-62703-17-5\_20.
126. Bennett WFD, Chen AW, Donnini S, Groenhof G, Tieleman DP. Constant pH simulations with the coarse-grained MARTINI model — Application to oleic acid aggregates. *Can J Chem*. 2013;91(April):839-846. doi:10.1139/cjc-2013-0010.

127. Cattaneo M, Dominici R, Cardano M, Diaferia G, Rovida E, Biunno I. Molecular chaperones as therapeutic targets to counteract proteostasis defects. *J Cell Physiol.* 2012;227(3):1226-1234. doi:10.1002/jcp.22856.

## APPENDIX A. DATA SET OF 100 PROTEINS

name	structure	name	structure	name	structure
1a6n		1r4v		1dyn	
1enh		2adr		1vid	
4icb		1uxc		1ebd	
1cun		1ehe		1qau	
1was		1bhd		2pth	
2giw		1ihb		3chy	
1hh8		1bp5		1ypi	
1cok		1snb		1ris	
1vap		1pma		1sac	
1ier		1ifc		1ubq	
1dj1		1tmc		1okt	
1lbd		3grs		1arb	
1hgu		1axj		3tgl	
1qaz		1agi		1uu2	
1bo9		2go0		1ril	
1ev4		1jam		1gff	
1bs2		1cvz		1f8d	
1jd1		1ceq		2trc	

1ypr		1d0n		1hcc	
1bfd		11as		1ezg	
1eqk		1j1y		1elp	
1gad		1p99		1ntn	
1ab2		1cuk		1fkb	
1iad		1wq4		1ixa	
1b6n		1u9a		1d1n	
1b5u		2lao		4wbc	
1nr2		1bf0		1ep0	
1l8l		1php		1mjc	
1fzw		1ddg		1shf	
2hnp		1g5b		1wit	
1bsg		2tgi			
1p9g		1d8v			
1d6t		1bqg			
1p88		1fzt			
1esj		1byl			

## APPENDIX B. PRINCIPAL COORDINATES ANALYSIS

The goal of PCA is to permit the positioning of objects in a space of reduced dimensionality while preserving their distance relationships as well as possible.

The value of PCA is that it permits the use of all types of variables, provided that a coefficient of appropriate type has been used to compute the resemblance hemi-matrix.

(1) The initial matrix needs to be a distance matrix  $\mathbf{A} = [\mathbf{d}_{ij}]$ .

(2) Matrix D is transformed into a new matrix A by defining:

$$\mathbf{A}_{ij} = -(1/2)\mathbf{d}_{ij}^2$$

(3) Matrix A is centered so that the sum of every row and of every column of A is 0, using equation:

$$\mathbf{A}_{ij}^* = \mathbf{A}_{ij} - \langle \mathbf{A}_{ij} \rangle_i - \langle \mathbf{A}_{ij} \rangle_j + \langle \mathbf{A}_{ij} \rangle_{ij}$$

$\langle \mathbf{A}_{ij} \rangle_i$  is average of *i*th row.  $\langle \mathbf{A}_{ij} \rangle_j$  is average of *j*th column.

(4) The eigenvalues and eigenvectors are computed and the latter are scaled to lengths equal to the square roots of the respective eigenvalues

(5) After scaling, if the eigenvectors are written as columns, the rows of the resulting table are the coordinates of the objects in PCO space.

Step 1. is accomplished by *rmsd\_pairwise.pl* and *RMSD\_ener.pl*

example: *rmsd\_pairwise.pl -dir PDB\_DIR -o mix\_rmsd\_pairwise.txt*

the directory PDB\_DIR includes all 'final.pdb' from each replica, the file

'mix\_rmsd\_pairwise.txt' contains data of rmsd value of all comparison of all replicas in matrix format. *RMSD\_ener.pl* is basically retrieve energy value from 'energy.log' of all replicas.

Step 2-5 is accomplished by *eig.pl*.

This script will do all calculations and generate file 'eigplot\_mix\_rmsd.txt' for gnuplot.

The final plot of landscape is finished by following command in gnuplot.

```
gnuplot>set dgrid3d 100,100
```

```
gnuplot>set pm3d
```

```
gnuplot>set hidden3d
```

```
gnuplot>set contour both
```

```
gnuplot>set xlabel "PC1"
```

```
gnuplot>set ylabel "PC2"
```

```
gnuplot>set zlabel "Potential Energy(kcal/mol)" rotate by 90
```

```
gnuplot>set cntrparam levels incr -800,50,0
```

```
gnuplot>splot "eigplot_mix_rmsd.txt" u 2:3:4 w lines notitle
```