CONSTANT pH REPLICA EXCHANGE MOLECULAR DYNAMICS STUDY OF PROTEIN STRUCTURE AND DYNAMICS

By

YILIN MENG

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2010

To my family

ACKNOWLEDGMENTS

At the completion of my graduate study at the University of Florida, I would like to take great pleasure in acknowledging the people who have supported me over these years.

I primarily thank my advisor, Professor Adrian E. Roitberg. Throughout the years working in his group, I have learned a tremendous amount from him. His guidance and encouragement supported me to overcome the obstacles not only in research but also in my personal life. There is no way I would have achieved my goal without his support and help.

I am thankful for the support and guidance of my committee members, Professors Kenneth M. Merz Jr., Nicolas C. Polfer, Stephen J. Hagen, and Arthur S. Edison. I also would like to thank Professors So Hirata, Joanna R. Long, Carlos L. Simmerling, and Wei Yang for their guidance in my research. I am very grateful for the assistance and helpful discussions from my colleagues in the Roitberg group, especially Dr. Daniel Sindhikara, Dr. Gustavo Seabra, Dr. Lena Dolghih, Dr. Seonah Kim, Jason Swails, Danial Dashti, Billy Miller, Dwight McGee, and Sung Cho. I appreciate all my friends at the Quantum Theory Project, the Department of Chemistry and Physics.

I thank the source of funding that supported my graduate study. My research was supported by National Institute of Health under Contract 1R01 AI073674. Computer resources and support were provided by the Large Allocations Resource Committee through grant TG-MCA05S010 and the University of Florida High-Performance Computing Center.

I want to acknowledge my wife, Xian who encouraged me and supported me to complete this work. Finally, I am very grateful for my whole family for their love and encouragement.

TABLE OF CONTENTS

page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACE	Analytical Continuum Electrostatic
BAR	Bennett Acceptance Ratio
CD	Circular Dichroism
CE	Continuum Electrostatic
CPHMD	Continuous Constant-pH Molecular Dynamics
CPL	Circularly Polarized Light
DOF	Degree of Freedom
DOS	Density of States
DSSP	Definition of the Secondary Structure of Proteins
EAF	Exchange Attempt Frequency
EFP	Effective Fragment Potential
FEP	Free Enery Perturbation
FDPB	Finite Differece Poisson-Boltzmann
GB	Generalized Born
HEWL	Hen Egg White Lysozyme
HH	Henderson-Hasselbach
H-REMD	Hamiltonian Replica Exchange Molecular Dynamics
LCPL	Left Circularly Polarized Light
MC	Monte Carlo
MCMC	Markov Chain Monte Carlo
MCCE	Multiconformation Continuum Electrostatic
MD	Molecular Dynamics
MDFE	Molecular Dynamics based Free Energy (calculation)

MM	Molecular Mechanics
MUCA	Multicanonical
NMR	Nuclear Magnetic Resonance
NPT	Isothermal-isobaric Ensemble
NVE	Microcanonical Ensemble
NVT	Canonical Ensemble
PB	Poisson-Boltzmann
PBC	Periodic Boundary Condition
PES	Potential Energy Surface
PDF	Probability Distribution Function
PMF	Potential of the Mean Force
QM	Quantum Mechanics
QM/MM	Hybrid Quantum Mechanical Molecular Mechanical
RCPL	Right Circularly Polarized Light
REM	Replica Exchange Method
REMD	Replica Exchange Molecular Dynamics
REX-CPHMD	Replica Exchange Continuous Constant-pH Molecular Dynamics
RF	Radio-Frequency
RMSD	Root-Mean-Square Deviation
TI	Thermodynamic Integration
T-REMD	Temperature Replica Exchange Molecular Dynamics
V-REMD	Viscosity Replica Exchange Molecular Dynamics

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

CONSTANT pH REPLICA EXCHANGE MOLECULAR DYNAMICS STUDY OF PROTEIN STRUCTURE AND DYNAMICS

By

Yilin Meng

August 2010

Chair: Adrian E. Roitberg
Major: Chemistry

Solution pH is a very important thermodynamic variable that affects protein structure, function and dynamics. Enormous effort has been made experimentally and computationally to understand the effect of pH on proteins. One category of computational method to study the effect of pH is the constant-pH molecular dynamics (constant-pH MD) methods. Constant-pH MD employs dynamic protonation in simulations and correlates protein conformations and protonation states. Therefore, constant-pH MD algorithms are able to predict pK_a value of an ionizable residue as well as to study pH-dependence directly.

A replica exchange constant-pH molecular dynamics (constant-pH REMD) method is proposed and implemented to improve coupled protonation and conformational state sampling. By mixing conformational sampling at constant pH (with discrete protonation states) with a temperature ladder, this method avoids conformational trapping. Our method was tested on seven different biological systems. The constant-pH REMD not only predicted pK_a correctly for model peptides but also converged faster than constant pH MD. Furthermore, the constant-pH REMD showed its advantage in the efficiency of conformational samplings. The advantage of utilizing constant-pH REMD is clear.

We have studied the effect of pH on the structure and dynamics of C-peptide from ribonuclease A by constant-pH REMD. The mean residue ellipticity at 222 nm at each pH value is computed, as a direct comparison with experimental measurements. The C-peptide conformational ensembles at pH 2, 5, and 8 are studied. The Glu2-Arg10 and Phe8-His12 interactions and their roles in the helix formation are also investigated.

Constant-pH REMD method is applied to the study of hen egg white lysozyme (HEWL). pK_a values are calculated and compared with experimental values. Factors that could affect pK_a prediction such as hydrogen bond network and interaction between ionizable residues are discussed. Structural feature such as coupling between conformation and protonation states is demonstrated in order to emphasize the importance of accurate sampling of the coupled conformations and protonation states.

CHAPTER 1
INTRODUCTION

1.1 Acid-Base Equilibrium

Acids and bases are common in our daily lives. For example, vinegar is acidic and ammonia is basic. According to the Bronsted-Lowry definition, an acid is a chemical compound that can donate protons and a base is a chemical compound that can accept protons. An acid can be converted to its conjugate base by transferring a proton to a base and a base is converted to its conjugate acid by accepting a proton. For simplicity, the conversion between an acid and its conjugate base can be described by the reaction: $HA \leftrightharpoons H^+ + A^-$ where *HA* is an acid, A^- is its conjugate base, and H^+ represents proton (in aqueous environment, H^+ is hydronium ion H_3O^+).

There exists an equilibrium state between any acid-base conjugate pair. At equilibrium, the concentration of each species is constant. In an acid-base reaction, an acid dissociation constant is used to describe this equilibrium. The acid dissociation constant has the definition of Eq. 1-1.

$$K_a = \frac{(a_{H^+})(a_{A^-})}{a_{HA}} \tag{1-1}$$

Here K_a is the acid dissociation constant and a_{H^+}, a_{A^-} and a_{HA} represent the activity of each species, respectively. In Eq. 1-1, the activity of each individual species (take a_{HA} as an example) can be expressed as:

$$a_{HA} = \gamma_{HA} \frac{[HA]}{c^\circ} \tag{1-2}$$

In Eq. 1-2, γ_{HA} is the activity coefficient of HA, $[HA]$ is the concentration of HA, and c° is the standard concentration which is 1 M. In an ideal solution, the activity coefficients are unity. The concentration of each species is divided by standard

concentration in order to make the acid dissociation constant dimensionless. For simplicity, the acid dissociation constant is expressed using the concentration of each species from now on.

The K_a indicates the strength of an acid: the stronger the acid is, the larger the K_a is. The order of magnitude of K_a can span over a broad range. Therefore, a logarithmic (base 10) measure of the K_a is more frequently adopted:

$$pKa = -\log_{10} K_a \tag{1-3}$$

Combining Eq. 1-1 and Eq. 1-3, we can express the pK_a value as:

$$pK_a = pH - \log_{10}\left(\frac{[A^-]}{[HA]}\right) \tag{1-4}$$

Eq. 1-4 is the Henderson-Hasselbalch (HH) equation. It allows one to solve directly for pH values instead of calculating the concentration of hydronium ions first. When $[A^-] = [HA]$, the HH equation becomes $pK_a = pH$. Therefore, the pK_a value of an acid is numerically equal to the pH value at which the acid and its conjugate base have the same concentrations. The acid dissociation constant represents the thermodynamics of an acid dissociation reaction because the pK_a value is proportional to the Gibbs free energy of the reaction. For simple compounds such as acetic acid, temperature is the most important factor that affects its pK_a value. However, for complex molecules such as proteins and peptides, the effect of environment is also crucial and will be discussed in this dissertation.

1.2 Amino Acids and Proteins

The goal of this dissertation is to study the acid-base equilibrium in peptide and protein systems and its effect on peptide and protein conformations by constant-pH REMD method. Thus, an introduction to peptide and protein, especially their structures

will be helpful. Amino acids have the generic structure as shown in Figure 1-1A. Each amino acid consists of an amino group (-NH2), a carboxylic acid group (-COOH) and a distinctive side chain (-R). All three groups are connected to a carbon atom which is called carbon alpha (C_α). There are twenty naturally occurring side chains and they can be divided into groups based on their physical or chemical properties. For example, one way to categorize the twenty side chains is based on their acid/base properties in aqueous solution. Therefore, an aspartic acid is an acidic amino acid and a lysine is a basic amino acid. For an amino acid, its carboxylic group can react with the amine group of another amino acid. This condensation reaction forms a peptide bond which links the two amino acids and yields a water molecule.

As a consequence of the condensation reaction, proteins are formed. A protein is a string of amino acids connected by peptide bonds and folded into a globular structure. A protein often consists of a minimum of 30 to 50 amino acids.[1] Shorter chains of amino acids are often called peptides. Each amino acid in a protein or peptide is called a residue. The peptide bonds form the backbone of a protein.

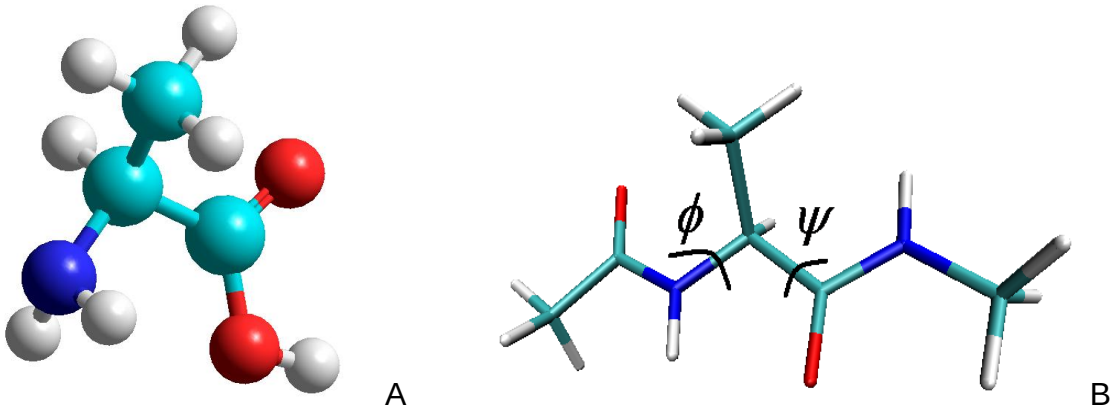


Figure 1-1. A) Structure of an amino acid named alanine. An amino group (-NH2), a carboxylic acid group (-COOH), a side chain (-R, in this case, a methyl group) and a hydrogen atom are bonded to a central carbon atom (C_α). B) Dihedral angles φ and ψ of alanine dipeptide.

A protein usually has four levels of structure which are called primary structure, secondary structure, tertiary structure and quaternary structure. The primary structure is the sequence of amino acids. The folding of a protein is determined by its primary structure.

Next, the secondary structure (e.g. α-helix, β-strand, or loop) is the three-dimensional structure of local segments of a protein. As mentioned earlier, proteins fold themselves into functional structures after they are formed. After folding, protein backbones often possess certain types of fold or alignment. The term of secondary structure is used to describe the three-dimensional structures of such manners. The two most common secondary structures found in proteins are α-helices and β-strands.

The local secondary structure of a particular residue in a protein can be described by a Ramachandran plot which is a two-dimensional histogram (or probability distribution) of backbone dihedral angle pair (ϕ,ψ). As demonstrated in Figure 1-1B, backbones can rotate around the N-C_α and C_α-C bonds, forming dihedral angles ϕ and ψ. Backbone conformations of a residue can be described by specifying (ϕ,ψ). Three main regions are populated in general in a Ramachandran plot, corresponding to the three main stable conformations a residue has: the right-handed α-helix region near (ϕ=-57°, ψ=-47°), the β-strand region near (ϕ=-125°, ψ=150°) and the polyproline II region near (ϕ=-75°, ψ=145°). The most populated region indicates the most stable conformation of a residue. An example of Ramachandran plot is shown in Figure 1-2.

Furthermore, the tertiary structure is the three-dimensional positions of all atoms in a protein. The tertiary structures yield information about protein side chains, for example, salt bridges. Finally, the quaternary structure defines the positions of all atoms

in a protein containing multiple peptide chains, for example, the hemoglobin tetramer. It is the highest level of protein structures.

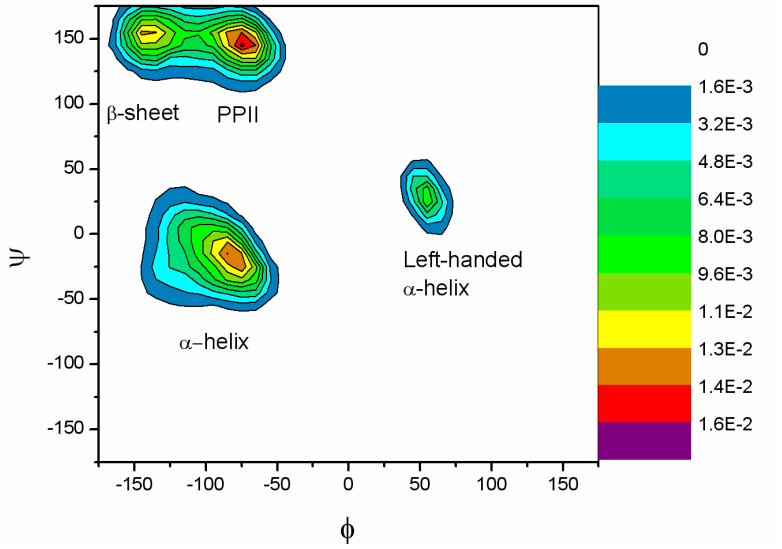


Figure 1-2. A Ramachandran plot (a contour plot showing the probability density of (ϕ,ψ) pairs) of tyrosine generated from the simulation of a heptapeptide which will be described later in chapter 3. In this figure, a left-handed α-helix is also shown.

Proteins perform vital functions, which are important to our lives. Almost all cell activities depend on proteins. For example, hemoglobin can transport oxygen molecules from lung to cells;[1] many chemical reactions occurring in living organisms are catalyzed by proteins called enzymes; and proteins are also involved in cell signaling. Mutations in the proteins, aggregation and misfolding of proteins can cause many diseases. For example, many cancers result from the mutations in the tumor suppressor p53.[2,3] Thus, understanding protein structures and functions is important.

1.3 Ionizable Residues in Proteins and the Effect of pH on Proteins

An ionizable residue in a protein is a residue with a side chain that can donate or accept proton(s). There are seven ionizable residues: ASP, GLU, HIS, CYS, TYR, LYS and ARG. Ionizable residues define the acid-base properties of that protein.

Consequently, the solution pH value becomes an important thermodynamic variable affecting protein structure, dynamics, folding mechanism, and function[4]. Many biological phenomena such as protein folding/misfolding,[5-8] substrate docking[9] and enzyme catalysis are pH-dependent.[10-12]

A good example of how pH value affects proteins is the pH-dependence of enzyme kinetics. Most enzymes possess an optimal pH value, at which the reaction rate is largest. Enzyme catalysis is pH-dependent because the active sites of enzymes in general contain important acidic or basic residues. Only one form (acidic or basic) of the ionizable residue is catalytically active, thus the concentration of the catalytically active species will affect the kinetics. Consider a simple reaction model (Figure 1-3 and Figure 1-4) to demonstrate how pH value affects enzyme reaction rate. In this model, only the zwitterion form is active; no intermediate exists for the enzyme reaction and the protonation-deprotonation steps are faster than catalysis steps. Furthermore, the rate-determining step does not depend on pH value.

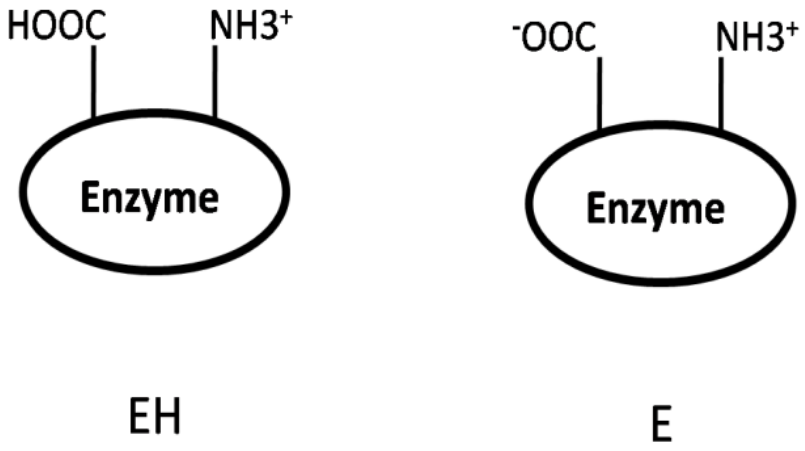


Figure 1-3. A diagram showing the cartoon representation of an enzyme at low pH (acidic) and at around the optimal pH value. *EH* indicates the structure at low pH and *E* stands for the zwitterion form, which is the active species in our model.[13]

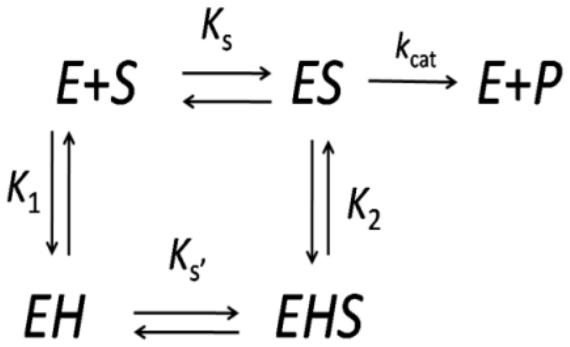


Figure 1-4. The reaction schemes showing the enzyme reactions at which pH values are smaller than the optimal pH value. K_s, $K_{s'}$, K_1 and K_2 are equilibrium constants of corresponding reactions and k_{cat} is the rate constant of the rate-determining step. This model can be used to explain how pH value affects enzyme catalysis in the pH range that is larger than optimal pH.[13,14]

The equilibrium constants shown in Figure 1-4 are not independent of each others. The relationship among them is given by:

$$K_s K_2 = K_{s'} K_1 \tag{1-5}$$

According to the above equation, if $K_1 = K_2$, then the substrate binding will not be affected by pH value of the solution. If it is not the case, then the binding is pH-dependent. After applying steady-state approximation to the $[ES]$, the reaction rate can be written as:

$$v = \frac{k_{cat}\,[E]_0[S]}{K_s + [S](1 + [H^+]/K_2) + K_s[H^+]/K_1} \tag{1-6}$$

where $[E]_0$ is initial concentration of the enzyme and $[H^+]$ is the concentration of hydronium ions. At low pH, increasing the concentration of hydronium ions (pH value decreases) will decrease the reaction rate. The same kind of model can also be applied to derive the effect of pH on reaction rate when the pH is higher than optimal. Likewise, only the zwitterion form is catalytically active. The conclusion is that pH value too high or too low will lower the enzyme catalytic reaction rate.

Given the importance of the solution pH, knowing the pK_a value of an ionizable residue in a protein is important because it will indicate the average protonation state of that ionizable residue at a certain pH value. However, the pK_a value of an ionizable residue is highly affected by its protein environment.[15,16] Two major factors affect protein pK_a values: one is the desolvation effect and the other is the electrostatic interaction. Other factors such as hydrogen bonding and structural rearrangement are also able to affect protein pK_a values.

An ionizable side chain in the interior of a protein can have a different pK_a value from the isolated amino acid in solution, which is caused by dehydration effect.[17-19] For example, Asp26 of the thioredoxin, which lies in a deep pocket of the protein, has a pK_a value of 7.5[17] while the pK_a value of a water-exposed aspartic acid is 4.0.[20] The Garcia-Moreno group has been employing site-direct mutagenesis method to study the effect of desolvation[18,19,21-23] and will be described later in this chapter. Their research on the buried ionizable residues provides a probe of the dielectric constant inside the protein, which is an important parameter for the pK_a prediction on the basis of the Poisson-Boltzmann equation.

Electrostatic interactions such as salt-bridges are also able to affect pK_a values. For example, His31 and Asp70 form a salt-bridge in the T4 lysozyme.[24] The formation of this salt-bridge shifts the pK_a of Asp70 to 0.5 and changes the pK_a of His31 to 9.1. Interestingly, Asp26 in the thioredoxin has been shown to form a salt-bridge with Lys57 when it is in the deprotonated form.[25] The formation of a salt-bridge should reduce the pK_a value of Asp26. Therefore, the pK_a value of 7.5 is the combined result of desolvation effect and electrostatic interaction.

Each ionizable residue has its own intrinsic pK_a value. The intrinsic pK_a value of an ionizable residue is defined as the pK_a value measured when this residue is fully solvent exposed and is not interacting with any other groups,[20] for example, an aspartate residue with two termini blocked. This kind of dipeptide is often used as reference (or model) compound in the theoretical calculation of protein pK_a values. The intrinsic pK_a values are reported in Table 1-1:

Table 1-1. Intrinsic pK_a values of ionizable residues in proteins.[26]

Residue Name	Intrinsic pK_a value
ASP	4.0
GLU	4.4
HIS	6.7
CYS	8.0
TYR	9.6
LYS	10.4
ARG	12.0

1.4 Measuring pK_a Values of Ionizable Residues

A general way to determining the pK_a value of an acid experimentally is through titration. In experiments, the pH values are measured by a pH meter as a function of the volume of base added to the solution. Therefore, a titration curve will be obtained (Figure 1-5A shows an example of titration curve) and the pK_a value is the pH value at which the deprotonated and protonated species have the same concentrations. Another way of presenting a titration curve is by plotting the fraction of deprotonation (protonation) vs the pH value. A Hill plot (an example is shown in Figure1-4B), which can be obtained by plotting log([A^-]/[HA]) as a function of pH, is used to study titration behavior. After fitting to the modified HH equation: $pH = pK_a + k \cdot \log\left(\frac{A^-}{HA}\right)$, the x-intercept is the pK_a value and the slope (k) is the Hill coefficient which reflects interactions between ionizable residues. The HH equation will be represented as a

straight line in a Hill plot, with a slope of unity. If only one ionizable residue is present in the system of interest, or an ionizable residue does not couple with other ionizable residue(s), the HH equation should be reproduced. A non-zero slope reflects statistical error (random error). Interacting ionizable residues will demonstrate non-HH behavior and possess non-unity slope in a Hill's plot. When $k > 1$, we say the proton binding is positively cooperative which means binding of the first proton will increase the binding affinity of the other one. When $k < 1$, the binding of protons is negatively cooperative which means the binding of one proton will decrease the affinity of the other proton.

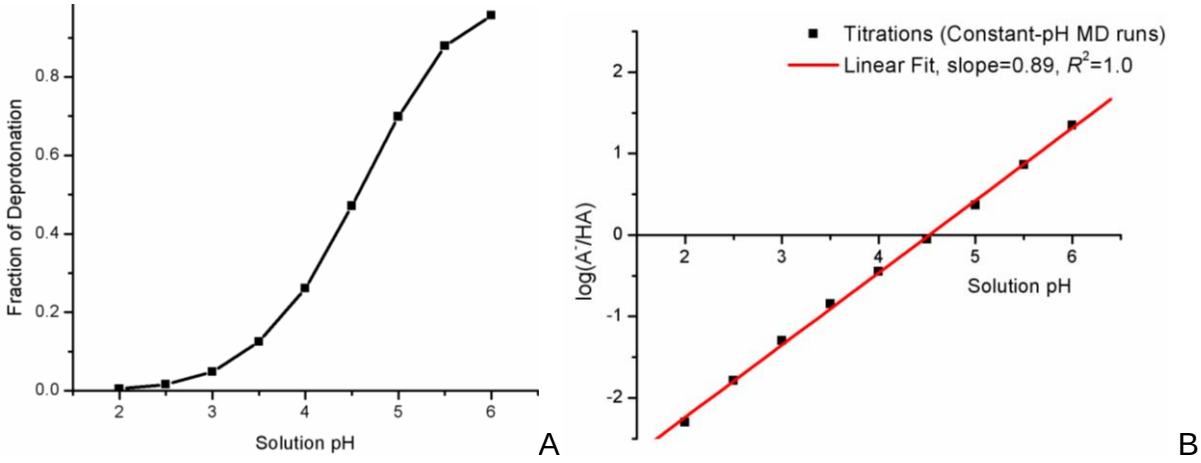


Figure 1-5. A) An example of titration curve. B) An example of Hill's plot on the basis of the titration described in Figure 1-5A. The two plots are generated from constant-pH MD simulations of an aspartic acid in a pentapeptide.

However, determining pK_a value of protein ionizable residues by measuring solution pH as a function of the volume of base is difficult because there are multiple ionizable residues in a protein in general. An experimental technique that is site-specific is preferred.

Nuclear Magnetic Resonance (NMR) is one of the most frequently employed spectroscopic methods in chemistry, physics and biological science. One application of the NMR method is to measure pK_a values of individual ionizable residues. NMR

spectroscopy measures the absorption of radio-frequency (RF) radiation by a nucleus in magnetic field. Only a nucleus with a spin quantum number that equals half of an integer is able to generate NMR signal. Furthermore, the absorption is affected by the chemical environment around that nucleus. Electron density around a nucleus provides a shielding effect to the external magnetic field for the nucleus. Thus, different chemical environment (electron density) around a nucleus will affect its resonance frequency, resulting in chemical shift. Changes in protonation state are able to result in changes in the chemical shift of the nuclei around the ionizable site (for example, C_γ of Asp, C_δ of Glu, and N_δ and N_ε of His). Subsequently, at a given pH value, the equilibrium between the protonated and deprotonated species can yield a weighted average chemical shift,

$$\delta_{obs} = \delta_p + \frac{\Delta\delta}{1+10^{n(pK_a - pH)}} \quad (1\text{-}7)$$

Here δ_{obs}, δ_p and $\Delta\delta$ are the chemical shift observed, chemical shift of the protonated species, the change in chemical shifts caused by titration, respectively, and n is the Hill's coefficient. In Eq. 1-7, the HH equation is implied. Therefore, chemical shifts will be measured at different pH values and a titration curve will be obtained. Figure 1-6 demonstrates a titration curve generated by NMR spectroscopy.

However, in practice, one-dimensional NMR spectra are often too complicated to be interpreted for proteins. Introducing a new spectrum dimension will allow the ability to simplify the spectra and yield more useful information. In two-dimensional NMR spectroscopy, the sample is excited by one or more pulses in the so-called "preparation time". Then the resulting magnetization is allowed to evolve for time t_1, and the signal is not recorded during time t_1. Following the evolution time, one or more pulses will be

applied to the sample and the resulting signal will be measured as a function of a new time variable t_2.

^{1}H, ^{13}C and ^{15}N NMR are frequently employed in experiments to determine protein pK_a values.[14] Proton NMR has shown to be particularly useful in studying histidine pK_a values. It is also employed to study the acid-base equilibrium of tyrosine residues. ^{13}C NMR experiments can be performed to determine the pK_a values of lysine and aspartate.

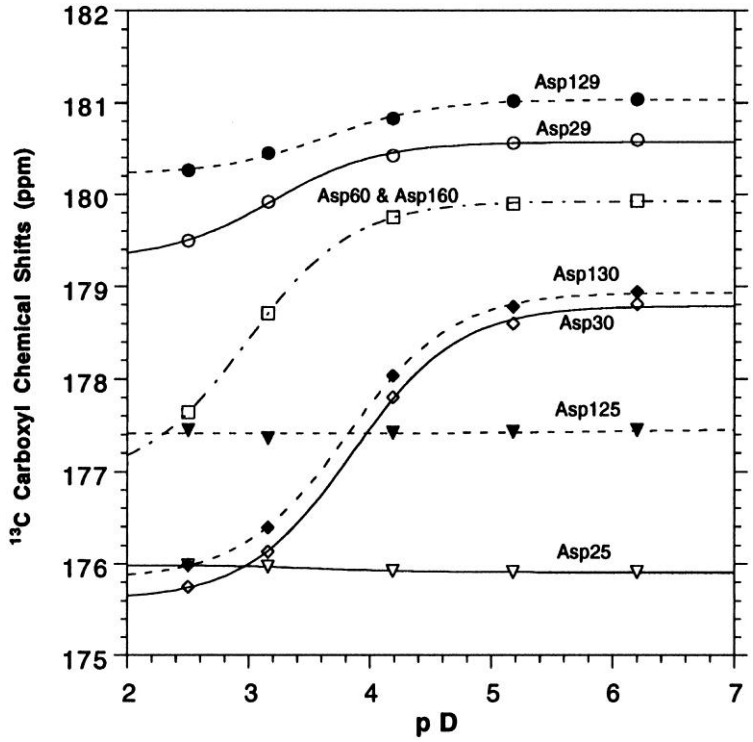


Figure 1-6. ^{13}C NMR titration curves of aspartate residues in HIV-1 protease/KNI-272 complex taken from Wang *et al*.,1996.[27] In this figure, Asp C_γ chemical shifts are plotted as a function of pD. Asp25 and Asp125 do not change protonation states in this pD range. But isotope shift experiments show that Asp25 is protonated and Asp125 is deprotonated in this pD range. "Reprinted with permission from Wang, Y. X.; Freedberg, D. I.; Yamazaki, T.; Wingfield, P. T.; Stahl, S. J.; Kaufman, J. D.; Kiso, Y.; Torchia, D. A. *Biochemistry* **1996**, *35*, 9945-9950. Copyright 1996 American Chemical Society."

One example of measuring the pK_a value of an ionizable residue using NMR technique is the determination of the pK_a value of Asp26 in *Escherichia coli*

thioredoxin.[17,25,28-30] NMR method, especially the 2D-NMR technique, has been intensively employed in the investigations of the pK_a value of Asp26. *Escherichia coli* thioredoxin has two redox forms. The oxidized form has a disulfide bond linking Cys32 and Cys35, while the two cysteine residues are not bonded in the reduced form. Hence, the two cysteine residues are ionizable in the reduced form, which makes the investigations more complicated. Asp26 is located at the bottom of a hydrophobic cavity near the active site disulfide and is completely buried in the protein. In 1991, Dyson *et al.* investigated pH effect on the thioredoxin in the vicinity of active site, using 2D NMR.[28] Both oxidized and reduced thioredoxin have been studied. $C^{\alpha}H$ and $C^{\beta}H$ chemical shifts of Cys32 and Cys35, and NH, $C^{\alpha}H$ and $C^{\beta}H$ chemical shifts of Asp26 as a function of pH value have been measured. Those chemical shifts have been found to titrate with a pK_a value of 7.5. Since the cysteine residues in the oxidized thioredoxin are not ionizable, they proposed that the apparent pK_a is the pK_a value of Asp26. In the same year, experiments performed by Langsetmo *et al.* measured electrophoretic mobility of the wild-type and D26A mutation of the oxidized thioredoxin, as a function of pH. A pK_a of 7.5 has been obtained from their experiments.[17] In 1995, Wilson *et al.* measured the chemical shifts of $C^{\beta}H1$, $C^{\beta}H2$ and C^{β} atoms of Cys32 and Cys35 using the reduced form of thioredoxin.[30] Both the wild-type and D26A mutation have been studied. Comparing the titration curves between the wild-type and the D26A mutation, a titration showing pK_a value > 9 has been found missing in the D26A thioredoxin experiment. Adopting that the cysteine residues in the reduced thioredoxin have pK_a values of 7.1 and 7.9 derived from Raman spectroscopy, they concluded that Asp26 has an apparent pK_a of greater than 9. However, their results were challenged by the

pK_a determinations of Cys32 and Cys35 in the reduced form of thioredoxin. In 1995, Jeng *et al.* studied the titration behaviors of Cys32 and Cys35 in the reduced form of thioredoxin by ^{13}C NMR experiments.[29] Their pK_a values were found to be 7.5 and 9.5. Their pK_a values of Cys32 and Cys35 challenged the results obtained by Wilson *et al*. In order to elucidate the pK_a value of Asp26 in the reduced thioredoxin, Jeng and Dyson measured the pK_a value of Asp26 in 1996 using 2D-NMR.[29] The ^{13}C chemical shift of the carboxylic group, which is bonded to titrating site, as well as the $C^{\beta}H1$ and $C^{\beta}H2$ proton chemical shifts was measured as a function of pH value. The authors believed that the pH effect on ^{13}C chemical shift of the carboxylic group should result from titration due to its close distance to the titrating site. The apparent pK_a value obtained from their experiments has been shown between 7.3 and 7.5, which is the same as the pK_a value of Asp26 in the oxidized form.

Fluorescence spectroscopy can be utilized to determine pK_a values as well. Fluorescence is the emission of light by a substance when it is relaxing from electronic excited state (S_1) to electronic ground state (S_0). In fluorescence spectroscopy, the substance is first excited from S_0 to one of many vibrational states of S_1 by absorbing a photon. Following the excitation, relaxation to the vibrational ground state S_1 occurs through collisions with other molecules. Once in the ground vibrational state of S_1, the substance will return to one of many vibrational states of S_0 by emitting a photon. Since the substance can return to various vibrational states in the electronic ground state, a band of emission wavelengths will be observed. The absorption and emission wavelengths are different (emission photons have a larger wavelength) and the

difference in wavelength is called Stokes shift. The average time the substance stays in its electronic excited state is called the fluorescence lifetime.

In biophysical chemistry, the tryptophan fluorescence is frequently employed to study the conformational changes in proteins. In general, tryptophan has a maximal absorption wavelength of 280 nm[31] and maximal emission wavelength of 300~350 nm.[32,33] Changes in the environment of a tryptophan residue will affect the emission wavelength and/or intensity. Furthermore, it has been noticed that tryptophan fluorescence is sensitive to the polarity of the local environment. One advantage of tryptophan fluorescence spectroscopy is that the chromophore is intrinsic; no change is made to the protein.

If the change in protonation state of an ionizable residue affects the spectrum of a neighboring tryptophan residue, which is the main fluorescent species in a protein, then fluorescence spectroscopy can be employed to generate a titration curve. Therefore, the pK_a value will be obtained. One example of determining pK_a value by fluorescence spectroscopy is measuring the pK_a of Glu35 in HEWL performed by the Imoto group.[34] The Trp108 is in van der Waals contact with Glu35. Changes in protonation state of Glu35 can induce a large shift in intensity of Trp108 fluorescence signal.

Another way of obtaining a titration curve is the potentiometric method. The potentiometric titration measures pH value as a function of the volume of titrant added. The volume of titrant added at each dosing can be used to calculate moles of hydrogen ion released from (or bound by) a peptide or protein, and hence number of hydrogen ions released (or bound) per molecule. Plotting number of hydrogen ions released (or bound) per molecule as a function of pH will generate a titration curve. By utilizing

potentiometric titration, a titration curve of the entire peptide or protein can be obtained. The Garcia-Moreno group has been utilizing the potentiometric method, combined with other experimental techniques and protein pK_a calculations, to investigate pK_a values of ionizable residues buried deep in a protein.[18,19,21-23] As mentioned earlier in the last section, protein environment can shift the pK_a value of an ionizable residue. In nature, a small portion of the ionizable residues are buried in the deep pockets of the protein, inaccessible to water.[22,35] Those buried ionizable residues are crucial to the protein functions such as catalysis,[12,36] and ion or electron transport.[37,38] Determining and understanding the pK_a values of buried ionizable residues is important for biological research. The Garcia-Moreno group performed site-directed mutagenesis experiments, mutating a nonpolar residue which is inaccessible to water to an ionizable residue. The pK_a value of the mutated ionizable residue is determined experimentally and predicted theoretically. By combining experimental and theoretical determination, the dielectric effect and electrostatic interactions will be elucidated. One example of the mutagenesis experiment is mutating Val66 in a “hyperstable variant” of the staphylococcal nuclease (SNase) to glutamate.[19,21] The original and mutated forms of the “hyperstable variant” of SNase are called PHS and PHS/V66E. The PHS nuclease can be made by mutating three residues of the wild-type SNase: P117G, H124L, and S128A. Val66 has been found in the core region of the SNase and inaccessible to aqueous environment. The potentiometric titrations have been performed on both PHS and PHS/V66E. The difference between the two titration curves represents the Glu66 titration plus other titrations affected by the mutation, although it is assumed that the latter effect is not

significant. The difference in hydrogen ions ($\Delta\upsilon_i$) bound to PHS and PHS/V66E was fitted to the following equation,

$$\Delta\upsilon_i = \frac{10^{n(pH-pK_a)}}{1+10^{n(pH-pK_a)}} \quad (1\text{-}8)$$

where n is the Hill's coefficient, pH is the solution pH value, and pK_a in this case is the pK_a value of Glu66.

The pH-dependence of PHS and PHS/V66E stability was also demonstrated by the guanidine hydrochloride denaturation free energy profiles. The Trp140 fluorescence was recorded as a probe of the denaturation. The difference in denaturation free energy profiles was also fitted nonlinearly to obtain the pK_a value of Glu66. The pK_a value of Glu66 has been determined to be 8.8 from potentiometric titration and 8.5 from the protein stability study. The pK_a shift of 4.4 (on the basis of the potentiometric measurements, and glutamate has an intrinsic pK_a value of 4.4) is among the largest ones for acidic ionizable residues. Once the experimental pK_a value is accurate obtained, a "reverse pK_a prediction" can be performed to investigate the dielectric constant inside the protein, which is an important parameter in the continuum electrostatic model and will be explained later this chapter. In fact, the direct potentiometric measurements were first carried out by the Garcia-Moreno group on PHS and PHS/V66K.[18] A pK_a value of 6.38 was found for Lys66, while the pK_a value of lysine model compound is 10.4.

Recent site-directed mutagenesis studies on PHS have extended to Leu38.[22] Mutations to aspartate, glutamate and lysine were conducted. Similar to their treatment on Val66 mutations, potentiometric titration and protein denaturation experiments were conducted to determine pK_a values by the Garcia-Moreno group. For the PHS/L38E,

NMR technique was employed to facilitate Glu38 pK_a measurement. PHS/L38K has shown a pK_a value close to the intrinsic value of lysine. After mutation, lysine was found to adjust its side-chain to let water molecules penetrate. However, L38D and L38E have shown elevated pK_a values. Both Asp38 and Glu38 were still inaccessible to water, although structural rearrangement was also observed. Their pK_a values were further perturbed by electrostatic interactions with surface carboxylic groups. Their investigations have unveiled how conformational changes, desolvation and electrostatic interactions affect pK_a values.

1.5 Molecular Modeling

Experimental techniques such as spectroscopy are fundamental to the study of protein structure and function. For example, NMR spectroscopy is frequently employed in biological science, X-ray crystallography can be applied to resolve protein structures and circular dichroism (CD) spectrometry is employed to determine the secondary structure of a protein. However, the advances in computational power combined with the leap in theory make experiments not the only way to understand biological molecules.

Molecular modeling offers another way to investigate structures and properties of biological molecules. It combines theories developed in the fields of physics, chemistry and biology with the computer resources to simulate the behaviors of molecules. Results from simulations are often compared to experimental observations in order to validate the method and understand the behavior of biological molecules from an atomistic level.

1.6 Potential Energy Surface

Molecules possess more than one stable configuration in general. In principle, all possible molecular configurations need to be considered in order to simulate a molecule correctly. A potential energy surface (PES), which is a surface defined by the potential energies of all possible configurations, can be utilized to fulfill this requirement. The local minima of a PES indicate stable conformations of a molecule. There are multiple ways to generate a PES. Quantum mechanical calculations offer the most accurate way to construct a PES. By solving the Schrodinger Equation, one can obtain energies and wave function of the molecule. In the field of chemistry, electronic structure theory utilizes quantum mechanics to describe the motion of electrons, in the framework of Born-Oppenheimer approximation. The Born-Oppenheimer approximation states that the electronic relaxation caused by nuclear motion is instantaneous because of the huge difference in the masses of electrons and nuclei. Thus, electronic motion and nuclear motion are decoupled. The eigenvalue of the electronic Schrodinger equation at each nuclear configuration is the potential energy of nuclei at that geometry. Solving Schrodinger equation at different configurations will yield the PES of a molecule. However, the cost of electronic structure calculations is very expensive, which hinders the use of high-level of theory when studying large biological molecules.

Due to the cost of electronic structure methods, an alternative way to describe a PES is to use a classical mechanical model. One of the commonly used algorithms is the all-atom force field in which the PES is computed without solving the Schrodinger equation. In an all-atom force field model, no electrons are present and each atom is represented by a single particle (in contrast to the united-atom force field model where a functional group is represented by a particle). Atoms interact with each other via bonded

and non-bonded potential energy terms. Equation 1-9 shows an example of all-atom force field model that is frequently adopted in the simulations of proteins:

$$U(q^N) = \sum_{bonds} \frac{1}{2} k_b (r - r_0)^2 + \sum_{angles} \frac{1}{2} k_a (\theta - \theta_0)^2 + \sum_{dihedrals} \sum_{\mathrm{n}=1}^{3} \frac{V_n}{2} [1 + cos(n\phi - \gamma)] + \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left\{ \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} + 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \right\} \quad (1\text{-}9)$$

The first three summations are bonded terms and they represent interactions of bond stretching, valence angle bending, and torsions, respectively. In Eq. 1-9, bond stretching and angle bending are considered by a harmonic potential. The torsion term is expressed as Fourier series due to the periodic nature of a dihedral angle. The latter two summation terms are the non-bonded interaction terms. The two components in the double summation represent electrostatic interactions and van der Waals interactions, respectively. Electrostatic potential is represented by Coulomb interaction. q_i and q_j are partial charges on atom i, and j, respectively. r_{ij} is the distance between the two atoms. In Eq. 1-9, van der Waals interaction is calculated by the Lennard-Jones potential, in which ε_{ij} is the well depth and σ_{ij} is the distance when repulsive and attractive potentials are equal. Solvent effect is also considered when implicit solvent such as the Generalized Born (GB) model[39,40] is adopted (solvent models will be briefly described in the next chapter). The cost of all-atom force field model is low compared with ab-initio methods because it utilizes pre-defined parameters when calculating potential energies. The strategy of generating those parameters is via fitting to experimental data and quantum mechanical calculations. One must notice that the parameters are often internally consistent which means parameters of different force fields are in general non-transferrable. The all-atom force field models are utilized much more frequently

than the quantum mechanical methods when simulating large systems such as proteins. However, force fields such as Eq. 1-9 do not allow bond breaking or forming. Thus, they are not able to study reactions. Nowadays, linear scaling techniques in electronic structure theory are developed in order to fill the gap between force fields and the high accuracy ab-initio methods.[41,42] One example of the linear scaling algorithm is the DivCon program developed by the Merz group.[43]

The balance between computational accuracy and cost is the main theme in the computational chemistry.[44] One category of schemes attempting to achieve this balance is the so-called hybrid quantum mechanical molecular mechanical (QM/MM) methods.[41,45-47] The basic idea of the QM/MM methods is that different regions of a system may play different roles. For example, if one wants to study an enzymatic reaction, the potential energy calculation involving the active site should be done by a quantum mechanical model because the classical force field is not able to describe bond forming/breaking. On the other hand, the bulk water (assuming no water molecule participates in enzymatic reaction) and the protein environment of the enzyme can be represented by the force field in order to save simulation time. In the QM/MM methodology, different regions of a system are treated by different level of theory and interact with each other. The QM/MM approaches have become a key area in the simulation of proteins.[48,49]

1.7 Molecular Dynamics, Monte Carlo Methods and Ergodicity

Accurately simulating the behavior of a molecule requires more than knowing the PES. A molecule often has more than one minimum on the PES. Finding the correct probability distribution of molecular conformations is also important because the majority of experiments measure molecular properties as averages over molecular

structures. Sampling algorithms such as molecular dynamics (MD) and the Metropolis Monte Carlo (MC) method are crucial to molecular modeling.

For a system containing *N* number of particles, there are 6*N* degrees of freedom (DOF). Half of the DOF comes from coordinates and the other half represents the momentum of all particles. The 6*N*-dimensional space defined by those DOF is called the phase space. Both MD and MC methods sample the molecular phase space. Over time, the system will generate a trajectory in the phase space.

MD utilizes the equation of motion to propagate a system in the phase space (The details of molecular dynamics will be presented in the next chapter). Each particle in the system has velocity and position and Newton's second law (Eq. 1-10) is applied to control the dynamics:

$$\vec{F} = m \cdot \vec{a} = -\vec{\nabla} U \tag{1-10}$$

The force on any particle in the system is given by the negative gradient of the potential energy. The equation of motion is usually solved numerically. By propagating the equation of motion, the phase space will be explored and a probability distribution for DOFs will be obtained. Therefore, molecular properties are able to be computed by averaging over times:

$$\langle A \rangle_{MD} = \lim_{t \to \infty} \frac{1}{N} \sum_{t=0}^{t} A(t_i) \tag{1-11}$$

In Eq. 1-11, *A* is the property of interest. *t* is the total simulation time. N is the size of the sample taken during the entire simulation. The bracket stands for taking average. $A(t_i)$ is the value of *A* at time t_i in the simulation.

In contrast to MD, the Metropolis MC method (from now on, we will call the Metropolis MC method as MC method unless otherwise mentioned) does not utilize the

equation of motion. MC method samples the phase space through a Markov chain (the details of Monte Carlo method will be presented in the next chapter). In MC algorithm, a new state (for example, a new molecular configuration) is randomly selected and the transition probability relationship between the current state and the new state is calculated by the detailed balance equation. Then a Metropolis criterion[50] is applied to accept or reject the transition to the new state. The Markov chain can be applied because the system is assumed to be at equilibrium. Likewise, after a sufficient number of transitions, the phase space will be explored and molecular properties can be computed by averaging over ensemble:

$$\langle A \rangle_{MC} = \int A(\vec{x})\, p(\vec{x}) d\vec{x} \tag{1-12}$$

Here $A(\vec{x})$ is the value of *A* in state $\vec{x}$. $p(\vec{x})$ is the normalized probability density of state $\vec{x}$.

The MD and the MC methods represent two different ways of sampling phase space and computing average molecular properties. According to the ergodic hypothesis, the time average is equal to the ensemble average:

$$\langle A \rangle_{MC} = \int A(\vec{x})\, p(\vec{x}) d\vec{x} = \lim_{t\to\infty} \frac{1}{N} \sum_{t=0}^{t} A(t_i) = \langle A \rangle_{MD} \tag{1-13}$$

The ergodic hypothesis is often assumed to be true in molecular simulations. This hypothesis makes MD and MC methods equivalent in sampling phase space. If the system is ergodic, the phase spaces generated by MD and MC should be the same because the phase space does not depend on sampling technique. The same behavior should also extend to any observable properties.

Conformational sampling in a MD or MC simulation is essential in the study of complex systems such as polymers and proteins. One major concern is that the PES of

a complex system is very rugged and contains a lot of local energy minima.[51] Thus, kinetic trapping would occur as a result of the low rate of potential energy barrier crossing, especially when the barrier is high. In order to overcome this kinetic trapping behavior, generalized ensemble methods (advanced sampling methods)[52,53] are frequently employed in molecular simulations. Popular generalized ensemble methods include multicanonical algorithm,[54,55] simulated tempering method,[56,57] parallel tempering method,[58-60] and replica exchange molecular dynamics (REMD) method.[61,62] A more thorough description of MD, MC and the advanced sampling methods will be presented in the next chapter.

1.8 Theoretical Protein Titration Curves and pK_a Calculations Using Poisson-Boltzmann Equation

Studying protein titration curves theoretically has a long history. As early as 1957, Tanford and Kirkwood presented their study of protein titration curve.[63] In their model, proteins were considered to be low-dielectric spheres with discrete unit charges on ionizable residues. They proposed that the pK_a value of an ionizable residue can be calculated from its intrinsic pK_a value and pair-wise electrostatic interactions with other ionizable residues. Calculating the pair-wise electrostatic interactions involves using empirical parameters. A protein titration curve showing average charge as a function of pH value was plotted. The Tanford-Kirkwood model was further extended and utilized to study lysozyme by Tanford and Roxby.[64] The equations used to generate a titration curve in the Tanford and Roxby paper were the same as those Tanford and Kirkwood used. However, they employed an iterative approach to generate titration curves and pK_a values for all ionizable residues. In their approach, each ionizable residue was initially assigned a pK_a value that is equal to its intrinsic value. At a given pH, the

average charge on each site (representing fraction of deprotonation/protonation) can be computed. Those average charges were then employed to update pK_a values. This process was repeated until self-consistent average charge and pK_a value of a site was obtained. Therefore, a titration curve can be produced by plotting average charge as a function of pH value.

In 1990, Bashford and Karplus utilized the finite difference Poisson-Boltzmann (FDPB) equation in the calculation of pK_a values.[65] A detailed description of the FDPB method will be present in the next chapter. The pK_a shift of an ionizable residue relative to a model compound is calculated (in their paper, intrinsic pK_a is a quantity defined as the pK_a value of an ionizable residue when other sites are neutral, that is, no interactions between ionizable sites). Given a molecular configuration, three terms are calculated by FDPB equation for each ionizable site: the Born solvation free energy, the pair-wise electrostatic interactions with non-ionizable residues (represented by partial charges), and the pair-wise electrostatic interactions between ionizable sites. Summing the three terms yields the electrostatic work of charging the ionizable side-chain, and hence yields the pK_a shift.

A protein titration curve is represented by plotting fraction of protonation vs pH value. Considering a protein with N ionizable sites and each site can have two states (protonated and deprotonated), there are 2^N possible macro-states and each macro-state can be represented by an N-dimensional vector. Once the FDPB equation is solved, free energy differences of each vector relative to completely deprotonated are computed. Thus, the fraction of protonation of an ionizable site can be calculated by taking the Boltzmann weighted average of the 2^N macro-states.

The FDPB method forms the foundation of the continuum electrostatic (CE) models, which are frequently utilized when studying protein pK_a values.[16,65-71] The FDPB method has been implemented into many modeling software packages such as UHBD[72] and DELPHI.[73] Many modifications have been done to improve its performance. In 1991, Beroza *et al.* employed the Metropolis MC method to sample 2^N numbers of protonation states, instead of calculating the protonation fraction at a given pH value directly.[74] After using MC sampling of protonation states, the number of ionizable residues included in the simulation can increase dramatically.

Solving the FDPB equation requires the dielectric constant in a protein as an input parameter and the dielectric constant is very important because the electrostatic energy is inversely proportional to it. It is considered as the most important adjustable parameter in FDPB-based pK_a calculations.[16] Thus, one question arisen from utilizing FDPB method is how to choose dielectric constant for proteins. The values between 4 and 20 are typically adopted in the FDPB calculations.[67] Direct experimental determination of the interior dielectric constant is extremely difficult. In practice, the protein dielectric constants are measured utilizing protein powders, which will cause problems in interpreting the resulting dielectric constants.[18,75,76] Research has been performed to find an optimal interior dielectric constant for protein pK_a predictions. However, considering the difference in protein environment, no single dielectric constant can yield experimental pK_a values for both internal and surface residues in a protein.[77] In 1996, Simonson and Brooks studied charge screening effect and protein dielectric constant by MD simulations.[78] What they found was that protein dielectric constant can range from ~4 in the interior of protein to a much higher value (~30) in the region near

the surface. As mentioned in section 1.4, the Garcia-Moreno group conducted site-directed mutagenesis experiments in the deep pocket of a protein where water is inaccessible and measured the pK_a value of mutated ionizable residue.[18,19,21-23,77] Then, the experimental pK_a value was put back into FDPB equation in order to examine protein interior dielectric constant. The protein interior dielectric constants were found to be ~11.[18] Mehler and his co-worker employed a sigmoidal screened electrostatic interaction to treat the protein dielectric environment.[79,80] Their method had been applied to Glu35 and Asp66 in hen egg white lysozyme and had obtained satisfactory results.[80]

Another problem in the FDPB-based pK_a calculation is that the FDPB equation is often solved on the basis of one structure such as X-ray crystal structure. The entropic effect is missing when a single structure is used. To improve the performance of the CE model in pK_a calculations, protein conformational sampling is also considered in order to incorporate conformational flexibility into pK_a calculations.[81-86] In the 1990s, You and Bashford developed an algorithm in which 36 side-chain conformations of ionizable residues are adopted in the calculation of pK_a values.[86] In 1997, Alex and Gunner proposed to use Monte Carlo method to sample $(2M)^N L^K$ possible states instead of just 2^N protonation states.[81] Here *N* is the number of ionizable residues and each one can have *M* possible conformations. Furthermore, each one of the *K* non-ionizable residue possesses *L* number of possible conformations. The Gunner group further extends this algorithm to the so-called multiconformation continuum electrostatic method (MCCE).[83] Recently, Barth *et al.* proposed a rotamer repacking technique combined with FDPB method and was given the name FDPB_MF.[82] In the FDPB_MF method, the

conformational space of side-chain of ionizable residues was defined by a rotamer probability distribution. Each rotamer was given a weight and was interacting with other ionizable residues in a mean-field scheme.

1.9 Computing pK_a Values by Free Energy Calculations

MD-based free energy (MDFE) calculations[87,88] have also been employed to predict pK_a values. MDFE calculations combine free energy calculation algorithms with MD propagations. MD propagations sample phase space and generate a conformational ensemble. Free energy calculation methods calculate the free energy difference between two states on the basis of the phase space sampled by MD. Free energy perturbation (FEP) and thermodynamic integration (TI) are two frequently employed free energy calculation methods and will be explained with more details in the next chapter. Free energy calculation algorithms such as FEP and TI methods can be used to compute pK_a because K_a is associated with the free energy of reaction.

Early pK_a calculations utilizing free energy calculations were conducted by the Warshel *et al.*,[89,90] Jorgensen *et al.*,[91] and Merz[92] with the FEP method and classical force fields. In the 1980s, Warshel *et al.* proposed a protein dipole Langevin dipole (PDLD) model for the pK_a calculations.[90] In the PDLD model, proteins were treated as particles having partial charges and polarizable dipoles, while the solvent molecules nearby were viewed as Langevin dipoles. The bulk water that is far away from ionizable residues was still treated as dielectric continuum. Electrostatic interactions between charges and dipoles, and dipoles and dipoles were computed.

Jorgensen *et al.* combined ab-initio quantum mechanical calculations and classical FEP calculations in 1989.[91] Jorgensen *et al.* calculated the pK_a difference between two acids, AH and BH. The gas-phase dissociation free energy of AH and BH were

computed by quantum mechanical methods. The solvation free energy calculations were conducted using MC FEP method for the neutral molecules and the anions. One shortcoming of their calculations is that only small organic molecules were investigated due to the computational cost of quantum mechanical methods.

In 1991, Merz performed classical FEP calculations for three glutamate residues in two proteins (HEWL and human carbonic anhydrease II).[92] The glutamate dipeptide was utilized as a model compound to eliminate the gas-phase dissociation free energy calculations.

When MDFE calculations utilizing the classical force fields are performed, quantum effects such as bond forming/breaking cannot be simulated. Thus, the pK_a shift of an ionizable residue relative to its intrinsic pK_a value (pK_a value of the reference compound which is defined in section 1.3 of this dissertation) is computed by the free energy calculations. A diagrammatic explanation of pK_a shift calculation utilizing the MDFE method is demonstrated in Figure 1-7 and Figure 1-8.

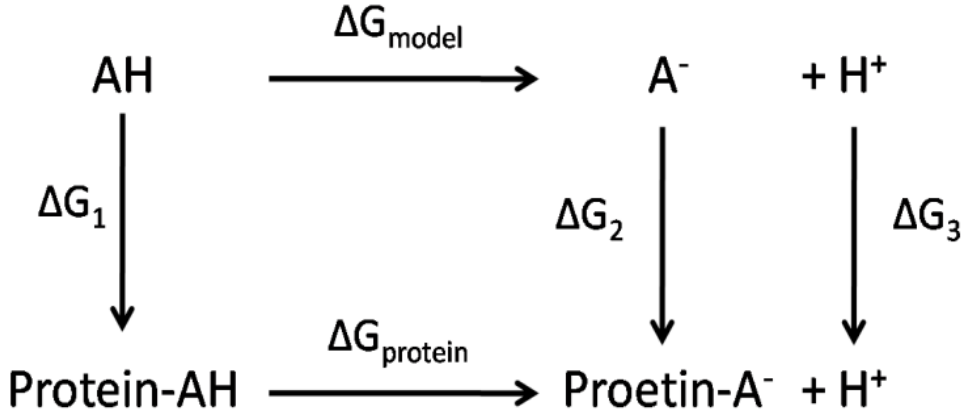


Figure 1-7. Thermodynamic cycle used to compute pK_a shift. Both acid dissociation reactions occur in aqueous solution. A thermodynamic cycle is a series of thermodynamic processes that eventually returning to the initial state. A state function, such as reaction free energy in this case, is path-independent and hence, unchanged through a cyclic process.

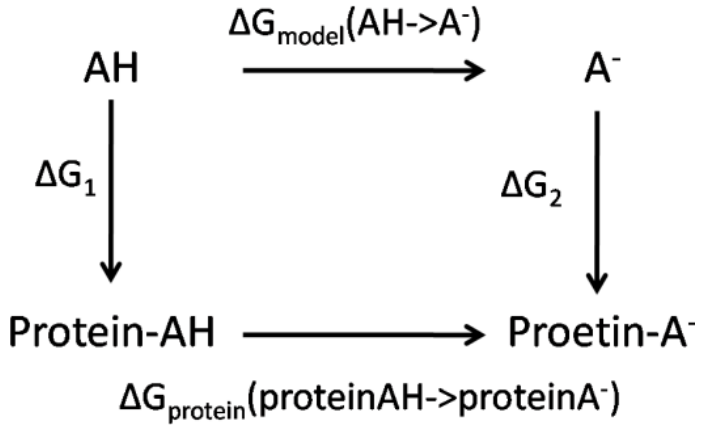


Figure 1-8. Thermodynamic cycle utilized to calculate the difference between ΔG1 and ΔG2. In Figure 1-7 and Figure 1-8, protein-AH represents the ionizable residue in protein environment. AH represents the reference compound which is usually the ionizable residue with two termini capped. In practice, a proton does not disappear but instead becomes a dummy atom. The proton has its position and velocity. The bonded interactions involving the proton are still effective. However, there is no non-bonded interaction for that proton. The change in protonation state is reflected by changes of partial charges in the ionizable residue.

Equations 1-14 to 1-20 explain how pK_a values will be computed from free energy calculations using force fields:

$$pK_{a,protein} = \frac{1}{2.303 k_B T} \Delta G_{protein} \quad (1\text{-}14)$$

$$pK_{a,model} = \frac{1}{2.303 k_B T} \Delta G_{model} \quad (1\text{-}15)$$

In Eq. 1-14 and 1-15, $\Delta G_{protein}$ and ΔG_{model} are the acid dissociation reaction free energy of the ionizable residue in protein and the reference compound, respectively. Therefore, the pK_a shift between ionizable residue in protein environment and the reference compound can be calculated as $pK_{a,protein} - pK_{a,model} = \frac{1}{2.303 k_B T}(\Delta G_{protein} - \Delta G_{model})$. According to the thermodynamic cycle shown in Figure1-6A, $(\Delta G_{protein} - \Delta G_{model}) = -\Delta G_1 + \Delta G_2 + \Delta G_3$. Here, ΔG_1 and ΔG_2 are the free energy difference between two protonated species, and between two deprotonated species, respectively.

ΔG_3 is equal to zero because the free energy difference between two protons that are in the same environment is zero. However, calculating ΔG_1 and ΔG_2 directly utilizing MDFE calculations is not preferable because the difference between the reference compound and the protein system is very large. A simple way to determine the difference between ΔG_1 and ΔG_2 is needed. Therefore, the thermodynamic cycle shown in Figure1-6B is employed. By utilizing that thermodynamic cycle, $(-\Delta G_1 + \Delta G_2)$ can be expressed as $(\Delta G(proteinAH \rightarrow proteinA^-) - \Delta G(AH \rightarrow A^-))$, where $\Delta G(proteinAH \rightarrow proteinA^-)$ and $\Delta G(AH \rightarrow A^-)$ are the free energy difference between the protonated and deprotonated ionizable residue in protein and the reference compound, respectively. $\Delta G(proteinAH \rightarrow proteinA^-)$ and $\Delta G(AH \rightarrow A^-)$ can be further expressed as:

$$\Delta G(proteinAH \rightarrow proteinA^-) = \Delta G_{QM}(proteinAH \rightarrow proteinA^-) + \Delta G_{MM}(proteinAH \rightarrow proteinA^-) \quad (1\text{-}16)$$

And

$$\Delta G(AH \rightarrow A^-) = \Delta G_{QM}(AH \rightarrow A^-) + \Delta G_{MM}(AH \rightarrow A^-) \quad (1\text{-}17)$$

In Eq. 1-16 and Eq. 1-17, the *MM* in the subscripts stands for the free energy differences which are calculated by classical force fields. The quantum mechanical contributions (labeled by *QM* in the subscripts) to the free energy difference of an ionizable residue in protein environment and its reference compound are assumed to be the same:

$$\Delta G_{QM}(proteinAH \rightarrow proteinA^-) = \Delta G_{QM}(AH \rightarrow A^-) \quad (1\text{-}18)$$

Combining all derivations and assumption, the difference between two acid dissociation reaction free energies can be written as:

$$\Delta G_{protein} - \Delta G_{model} = \Delta G_{MM}(proteinAH \rightarrow proteinA^-) - \Delta G_{MM}(AH \rightarrow A^-) \quad (1\text{-}19)$$

Thus, subtracting Eq. 1-15 from Eq. 1-14 yields:

$$pK_{a,protein} = pK_{a,model} + \frac{1}{2.303k_BT}\left(\Delta G_{MM}(proteinAH \rightarrow proteinA^-) - \Delta G_{MM}(AH \rightarrow A^-)\right) \quad (1\text{-}20)$$

$\Delta G_{MM}(proteinAH \rightarrow proteinA^-)$ and $\Delta G_{MM}(AH \rightarrow A^-)$ are are computed by MDFE calculations (for example, TI). A more detailed description of the MDFE methodology and how to compute $\Delta G(proteinAH \rightarrow proteinA^-)$ and $\Delta G(AH \rightarrow A^-)$ will be explained in the next chapter.

An example of using classical force field MDFE calculations to study pK_a values is given by Simonson *et al.*[15] The pK_a values of Asp20 (experimental pK_a of 2, which is lower than the intrinsic Asp pK_a value), Asp26 (experimental pK_a of 7.5) in thioredonxin, and Asp14 (with an experimental pK_a around 4) in ribonuclease A were evaluated by TI calculations. The aspartate dipeptide was taken as the model compound; both explicit and implicit water models were used in their simulations. Proton dissociation was represented by changes in the partial charges of carboxylic group only. The free energy change caused by the disappearance of the proton van der Waals interaction was not considered because the van der Waals radius of the proton in aspartate is zero in the AMBER force field. Correct protonation free energies have been obtained. Entropic and enthalpic effects are also correctly obtained. However, several problems have also been found with the MDFE-based pK_a calculations. For example, interactions between ionizable sites are not able to be incorporated directly. Furthermore, their free energy differences have shown dependence on the force fields and solvation models.

Hybrid quantum mechanical/molecular mechanical (QM/MM) methods can be coupled with free energy calculation simulations.[48,93] Recently, the Cui group has

conducted pK_a calculations using FEP calculations coupled with SCC-DFTB method.[94,95] A detailed description of QM/MM free energy calculations of pK_a values can be found in a recent review by Kamerlin *et al.*[48]

1.10 pK_a Prediction Using Empirical Methods

Empirical models are also employed to study protein pK_a values. According to Lee and Crippen,[16] the seemingly most accepted empirical method is PROPKA which is developed by the Jensen group.[96-101] The PROPKA method involves using 30 parameters obtained from 314 residues in 44 proteins. QM calculations and the effective fragment potential (EFP) method,[102,103] which is a QM/MM method, are employed to generate those parameters. In the PROPKA method, a pK_a value is calculated by adding "perturbations" to its intrinsic pK_a values. Three types of perturbations are considered: the hydrogen bonding, desolvation effect and charge-charge interactions. A detailed description of the PROPKA method can be found in a review by Jensen *et al.*[97]

1.11 Constant-pH Molecular Dynamics (Constant-pH MD) Methods

Traditionally, MD simulations have been performed in a manner of constant protonation state. The protonation state of an ionizable residue is assigned before a MD simulation is started. Moreover, the protonation states are not allowed to change during MD propagations. Performing constant protonation state MD simulations requires knowing the pK_a values of all ionizable residues beforehand. Not knowing the pK_a value may result in wrong assignment of protonation state. In addition, if pK_a values are near the solution pH values, constant protonation state MD simulations are not able to reflect this situation. More importantly, constant protonation state MD simulations cannot be employed to study the coupling between conformations and protonation states. Thus,

constant-pH MD algorithms were developed in order to correlate protein conformation and protonation state.[104] The purpose of constant-pH MD is to describe protonation equilibrium correctly at a given pH value. Therefore, its applications include pK_a predictions and studying pH effects. One category of constant-pH MD methods uses a continuous protonation parameter.[105-115] Earlier models include a grand canonical MD algorithm developed by Mertz and Pettitt in 1994[115] and a method introduced by Baptista *et al.* in 1997.[106] In the Mertz and Pettitt model, protons are allowed to be exchanged between a titratable side chain and water molecules. Baptista *et al.* used a potential of mean force to treat protonation and conformation simultaneously. Later, Börjesson and Hünenberger developed a continuous protonation variable model in which the protonation fraction is adjusted by weak coupling to a proton bath, using an explicit solvent.[107,108] More recently, the continuous protonation state model has been further developed by the Brooks group.[109-114] They developed a constant-pH MD algorithm by the name of continuous constant-pH molecular dynamics (CPHMD). In the CPHMD method, Lee *et al.*[114] applied λ-dynamics[116] to the protonation coordinate and used the Generalized Born (GB)[40,117] implicit solvent model. They chose a λ variable to control protonation fraction and introduced an artificial potential barrier between protonated and deprotonated states. The potential is a biasing potential to increase the residency time close to protonation/deprotonation states and it centered at half way of titration (λ=1/2). The CPHMD method was then extended by incorporating improved GB model and REMD algorithm for better sampling. The applications of CPHMD and replica exchange CPHMD included predicting pK_a values of various proteins,[110,114] studying

proton tautomerism[109] and pH-dependent protein dynamics such as folding[112,113] and aggregation.[111]

In addition to continuous protonation state models, discrete protonation state methods have also been developed to study pH-dependence of protein structure and dynamics.[118-131] The discrete protonation state models utilize a hybrid molecular dynamics and Monte Carlo (hybrid MD/MC) method. Protein conformations are sampled by molecular dynamics and protonation states are sampled using a Monte Carlo scheme periodically during a MD simulation. A new protonation state is selected after a user-defined number of MD steps and the free energy difference between the old and the new state is calculated. The Metropolis criterion is used to accept or reject the protonation change. Various solvent models and protonation state energy algorithms were used in discrete protonation state constant pH MD simulations.

Burgi *et al.*[130] presented their constant-pH MD method using discrete protonation state model and applied it to hen egg white lysozyme (HEWL). The lysozyme was dissolved into explicit water. Short TI calculations (20 ps of dynamics) were carried out to provide classical free energy difference between old and new protonation states at each MC attempt. The MC move is evaluated based on the following free energy difference:

$$\Delta G = k_B T \ln 10 \left(pH - pK_{a,ref}\right) + \Delta G_{prot,MM} - \Delta G_{ref,MM} \quad (1\text{-}21)$$

In the above equation, pH is a parameter and represent the pH value of the solution, $pK_{a,ref}$ is the pK_a value of the model compound (reference compound), $\Delta G_{prot,MM}$ and $\Delta G_{ref,MM}$ is the classical force field proton dissociation free energy given by TI for the protein and reference compound, respectively. One pitfall of the method

developed by Burgi *et al.* is the choice of simulation time of TI. The 20 ps TI calculation represents neither single-structure protonation free energy nor an average of the entire ensemble.

The Baptista group proposed their constant-pH MD method using the FDPB method to calculate protonation energies and their MD was done in explicit solvent.[118,123-126] The MD propagations are conducted at fixed protonation states. The MC moves in the protonation states are performed at fixed molecular configurations. The MD propagation is able to generate a conditional PDF of coordinates and momenta given protonation states, while the MC sampling is able to yield a conditional PDF of protonation states given molecular configurations. Baptista *et al.* proved that the hybrid MD and MC method is able to generate an ergodic Markov chain.[118] Hence, conditional probability distributions yielded by MD and MC are able to generate a joint probability distribution satisfying semigrand canonical ensemble. The work done by Baptista *et al.* provides the theoretical justification for combined MD and MC sampling in the discrete protonation state constant-pH methods. In practice, MD simulations are conducted in explicit water to sample conformational space. A new protonation state is selected and the free energy difference is calculated using the structure at that moment and the continuum electrostatic model. The MC transition is evaluated and if the move is accepted, a short MD run is performed to relax the solvent. After solvent relaxation, MD steps continue for solute and solvent. The Baptista group applied their constant-pH MD method to the study of protonation-conformation coupling effect,[123] the pH-dependent conformation states of kyotorphin,[124] pK_a predictions of the HEWL[125] and the redox titration of cytochrome c_3.[126]

Walczak and Antosiewicz also employed the FDPB method to determine protonation energy but they used Langevin Dynamics to propagate coordinates between MC steps.[128] This method is further extended by Dlugosz and Antosiewicz.[119-122,128] The extended method combines conventional MD simulation using the analytical continuum electrostatic (ACE)[132] scheme to sample conformations with the FDPB method for the MC moves. Succinic acid[119] and a heptapeptide derived from ovomucoid third domain (OMTKY3)[122] have been studied by Dlugosz and Antosiewicz. This heptapeptide corresponds to residues 26-32 of OMTKY3 and has the sequence of acetyl-Ser-Asp-Asn-Lys-Thr-Tyr-Gly-methylamine. Nuclear magnetic resonance (NMR) experiments indicated the pK_a of Asp is 3.6,[122] 0.4 pK_a unit lower than the value of blocked Asp dipeptide. In their studies, the conventional molecular dynamics (MD) simulations were carried out to sample peptide conformations. Their method predicted the pK_a to be 4.24.

Mongan *et al.* developed a method combining the GB model and the discrete protonation state model and implemented it into the AMBER simulation suite.[127] In Mongan's method, the GB model was used in protonation state transition energy as well as solvation free energy calculations. Therefore, solvent models in conformational and protonation state sampling are consistent and the computational cost is small. More recently, the accelerated molecular dynamics (AMD)[133,134] method was combined with Mongan's constant-pH algorithm to enhance conformational sampling.[129] This model has been utilized to calculate pK_a values of an enzyme and to explore the protonation-conformation coupling. The continuous protonation state model developed by the

Brooks group, the discrete protonation state model proposed by Baptista *et al.* and by Mongan *et al.* will be further explained in chapter 2.

CHAPTER 2
THEORY AND METHODS IN MOLECULAR MODELING

Molecular Modeling or molecular simulation is a way to study molecules using theories developed in the fields of physics, chemistry and biology coupled with the computer resources. With the development of computer power and parallel computation, molecular modeling is more and more often involved in the research of biology, chemistry and physics.[42] Understanding the underlying theory and methods of molecular modeling is necessary in order to perform simulations and analyze the data generated. In this chapter, the basic theory and methods of constant-pH replica exchange molecular dynamics method and protein pK_a calculations methods are described.

2.1 Potential Energy Functions and Classical Force Fields

2.1.1 Potential Energy Surface

Molecular modeling studies molecules, which in general possess more than one configuration for a chemical formula in general. In principle, all possible molecular configurations need to be considered in order to simulate a molecule correctly. A potential energy surface (PES), which is a surface defined by the potential energies of all possible configurations, can be utilized to fulfill this requirement. The concept of PES is a result of the Born-Oppenheimer approximation. The Born-Oppenheimer approximation states that the electronic relaxation caused by nuclear motion is instantaneous because of the huge difference in the masses of electrons and nuclei. Thus, electronic motion and nuclear motion are decoupled. Electronic energy, which is computed at a fixed nuclear geometry (molecular structure), is the potential energy of nuclei at that structure. Local minima on the PES indicate stable conformations of a

molecule. Quantum mechanics forms the foundation of understanding the molecular behaviors and offers the most accurate way to construct a PES. Ideally, the Schrodinger equation is solved for electronic energy at all possible nuclear configurations and hence, yields the PES of a molecule.

2.1.2 Force Field Models

Although quantum mechanical calculations generate very accurate energies, performing a molecular simulation using quantum mechanical method is too time-consuming even through the use of parallel computation, especially for large systems such as polymers and proteins. Force field (equivalent to molecular mechanics) models have been designed to solve this problem. Force field models ignore electrons and calculate the potential energy of a system based on nuclear geometry only. Force field calculations are fast because the potential energy functions are simple and parameterized.

In a force field model, the potential energy of a system has the following contributions in general: bond stretching (vibration), angle bending, bond rotation (torsion), electrostatic interaction, and the van der Waals interaction. The former three contributions are often called the bonded interactions and the last two belong to non-bonded interactions.

In many force field models, such as the AMBER force field,[135] bond stretching energy between atoms i and j is the second order truncation of the Taylor expansion of potential energy function about equilibrium distance and hence, can be formulated as a harmonic potential:

$$U_{bond} = \frac{1}{2} k_{ij} \left(r_{ij} - r_{ij,eq} \right)^2 \quad (2\text{-}1)$$

where k_{ij} is the force constant, r_{ij} is the distance between two atoms and $r_{ij,eq}$ is the equilibrium distance between the two atoms. One drawback of this function is that a bond cannot be broken and has infinite energy when two atoms are infinitely apart. Therefore, such a potential energy can be applied to bond stretching near equilibrium distance only. A simplest remedy is to include higher order Taylor expansion terms but this increases the computation time. For example, expansions up to the fourth-order are adopted in the general organic force field MM3.[136] This Taylor expansion strategy is also employed in deriving angle-bending potential functions. Torsions (or dihedral angles) are periodic and hence, Fourier series is adopted as torsion potential energy function. One example of the formula of torsion potential energy is displayed in Eq. 1-9.

The van der Waals interaction in a force field model should be able to reproduce the repulsion and attraction between two particles having no permanent charges. This attractive interaction is generally called dispersion. Quantum mechanics indicates that the dispersion energy is inversely proportional to the sixth-power of the distance between two particles (say atoms) i and j (under the dipole-dipole interaction approximation):[137]

$$U_{dispersion} = -\frac{b_{ij}}{{r_{ij}}^6} \quad (2\text{-}2)$$

where b_{ij} is a constant specific to i and j and r_{ij} is the distance between i and j. There is no theoretical derivation for the repulsive interaction. However, for computational simplicity, the repulsive energy is taken to be inversely proportional to the twelfth-power of the distance. A simple way to combine repulsive and attractive potentials is just adding up the two potentials. Thus, van der Waals interaction is governed by the Lennard-Jones potential shown in Eq. 1-9. Due to the fact that van der Waals

interaction decays very fast as a function of inter-particle distance, it is often called "short-range interaction".

Electrostatic interaction is often considered as the "long-range interaction". The simplest model of electrostatic interaction is the point-charge model which is adopted in the AMBER force field. Partial charges are assigned to each atom and Coulomb's law is applied to calculating interaction energy. More complicated models such as calculating electrostatic energy through dipole moment-dipole moment interaction have also been employed.[137]

Bond, angle and torsion interactions are coupled. Thus, the coupling effects (cross terms) should be incorporated into force fields. Mathematically, cross terms are generated from multi-dimensional Taylor expansions. For example, the angle-bending accompanied by two bond-stretching motions (shown in Figure 2-1) is formulated to be (as in MM3):

$$U_{bond-angle} = \frac{1}{2} k_{ijk} \left[\left(r_{ij} - r_{ij,eq} \right) + \left(r_{ik} - r_{ik,eq} \right) \right] \left(\theta_{ijk} - \theta_{ijk,eq} \right) \quad (2\text{-}3)$$

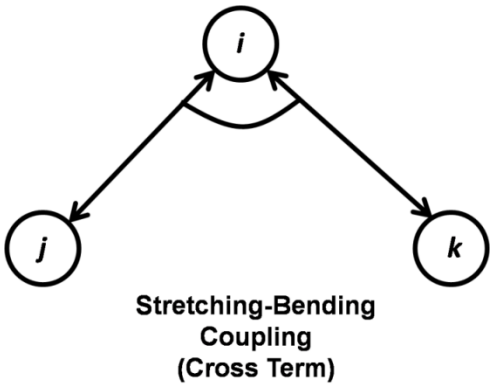


Figure 2-1. A diagram showing bond-stretching coupled with angle-bending. A cross term calculating coupling energy is adopted when evaluating the total potential energy.

The force field is simply a function and corresponding parameters. Thus, obtaining parameters is crucial for force field development. Given a potential energy function,

parameters are required to reproduce experimental data or quantum mechanical calculation results as much as possible.

2.1.3 Protein Force Field Models

Computer simulations of biological molecules often involve thousands of atoms or even more,[138] especially when using explicit solvent models. Many simulations on proteins choose to use force fields to reduce computational cost. Popular protein force fields include (but are not limited to) AMBER99SB,[139] CHARMM22,[140] GROMOS96,[141] and OPLS force fields.[142] In general, a simple potential energy function like Eq.1-9 is employed in the protein force fields. Protein force field parameters are in general optimized on the basis of small molecules. Take the AMBER force field (Eq. 1-9) as an example; there are bonded and non-bonded terms in it. In the non-bonded terms, the partial charges are fitted to quantum mechanical calculation using Hartree-Fock/6-31G* level of theory in vacuum. This level of theory typically overestimates dipole moment, and hence the resulting partial charges can satisfactorily approximate the condensed-phase charge distribution. The Lennard-Jones parameters have been obtained from reproducing liquid properties following the work of Jorgensen *et al*.[142] After the partial charges are assigned, the Lennard-Jones parameters are fitted to reproduce experimental data such as heat capacity, liquid density, and the heat of vaporization.

The bond stretching and angle bending parameters are derived by fitting to structural and vibrational experimental data of small molecules that make up proteins. The bond and angle parameters should ensure that the geometries of simple protein fragments are close to experimental data. The torsion (dihedral angle) parameters can be obtained from quantum mechanical conformational energy calculations. Determining torsion parameters is often the last step of force field parameter optimizations. Given

the previous obtained individual energy term parameter sets, the torsion parameters are adjusted to best fit quantum mechanical conformational energies, for example, the Ramachandran plot of a model compound. Detailed description of the protein force field parameter determinations can be found in the paper of Cornell *et al.*,[143] MacKerell *et al.*,[140] and Hornak *et al*.[139]

2.2 Molecular Dynamics (MD) Method

2.2.1 MD Integrator

As mentioned in the introduction, MD samples the phase space utilizing the equation of motion. A trajectory in the phase space will be generated over time. The ergodic hypothesis is assumed to be true, that is, the time average of any property at equilibrium is equivalent to the ensemble average. Thus, given a set of initial positions and momenta and a method to compute forces, a MD simulation can be applied to any system. For a simple system such as a harmonic oscillator moving along one axis, there exists an analytical solution of the trajectory (the coordinate and momentum as a function of time can be expressed analytically). However, it's almost impossible to know the analytical solution of complex systems such as polymers or proteins. Therefore, numerical integrators are implemented to propagate positions and velocities of particles. One of the frequently used integrator is the leap-frog algorithm:[41,144]

$$\vec{q}(t+\Delta t) = \vec{q}(t) + \vec{v}\left(t+\frac{1}{2}\Delta t\right)\Delta t \tag{2-4}$$

$$\vec{v}\left(t+\frac{1}{2}\Delta t\right) = \vec{v}\left(t-\frac{1}{2}\Delta t\right) + \vec{a}(t)\Delta t \tag{2-5}$$

$$\vec{a}(t) = -\frac{\vec{F}(t)}{m} = -\frac{\vec{\nabla}U(t)}{m} \tag{2-6}$$

Here, *q* and *v* stand for the position and velocity of a particle respectively; *a*(t), *F*(t) and *U*(t) represent the acceleration, the force and the potential energy at time *t*; and Δt

is the time step used in MD simulation. One frequently employed potential energy function is the force field model introduced in the previous section. According to Eq. 2-4, 2-5 and 2-6, the leapfrog algorithm propagates positions and velocities in a coupled way. The velocity at time t can be calculated by velocities at $t+\frac{1}{2}\Delta t$ and $t-\frac{1}{2}\Delta t$ by the following equation:

$$v(t)=\frac{1}{2}\left[v\left(t+\frac{1}{2}\Delta t\right)+v\left(t-\frac{1}{2}\Delta t\right)\right] \quad (2\text{-}7)$$

One important issue in the MD propagation is choosing a proper time step that optimizes speed of propagation and accuracy of the simulation. A too small time step will waste simulation time in sampling the same conformation, whereas a too large time step can bring two atoms too close and hence cause instability of the trajectory. In general, a time step is a tenth of the period of fastest motion. In biological molecules, the fastest motion is the bond stretching and bonds with hydrogen atoms in particular. Thus, one way to increase time step without reducing accuracy is to remove the degree of freedom having highest frequency. One commonly employed algorithm to achieve this goal is the SHAKE algorithm.[145] When using the SHAKE algorithm to remove heavy-atom-to-hydrogen DOF, the heavy-atom-to-hydrogen bond length is fixed. The fixed bond lengths act as distance constraints between heavy and hydrogen atoms. Lagrangian multipliers have been utilized to keep the bond lengths constant. By employing the SHAKE algorithm, a large time step such as 2 fs could be used. Methods that can integrate the equation of motion more efficiently are popular area of research.

2.2.2 Thermostats in MD Simulations

Before describing thermostats in MD simulations, the concept of thermodynamic ensemble (statistical ensemble) should be introduced first. An ensemble is a large

amount of replicas of the system of interest (it may contain infinite number of replicas). All replicas in an ensemble are considered at once. Each replica represents the system in one possible state. Thermodynamic ensembles are characterized by macroscopic thermodynamic properties. Several frequently employed thermodynamic ensembles are microcanonical ensemble (NVE ensemble), canonical ensemble (NVT ensemble), isothermal-isobaric ensemble (NPT ensemble), and grand canonical ensemble.

MD simulations are controlled by Newton's second law. This makes a MD simulation conserve the total energy and represent a system in the microcanonical (NVE) ensemble, where number of particles (N), volume (V), and total energy (E) are constant. However, our system of interest is in the canonical (NVT) ensemble, in which number of particles (N), volume (V), and temperature (T) are constant. Therefore, maintaining a constant temperature in a MD simulation is necessary. Any algorithm that can maintain constant temperature and approximate the NVT ensemble is called a thermostat. Popular thermostats include Berendsen thermostat,[146] Langevin dynamics[147] and Nose-Hoover thermostat.[148] The Berendsen thermostat and Langevin dynamics are utilized in our MD simulations and thus explained here.

In a MD simulation, the temperature can be written as:

$$T = \frac{1}{(3N-n)k_B}\sum_{i=1}^{N}\frac{m_i\vec{v}_i^2}{2} \qquad (2\text{-}8)$$

Here *N* is the number of particles, *n* is number of constrained degree of freedom, m_i and $\vec{v}_i$ are the mass and velocity of particle *i*. Thus, temperature is a function of velocities of all particles. The simplest way to control temperature is to rescale velocity at each time step. However, this will cause discontinuity in the momentum trajectory in phase space.

Berendsen *et al*. introduced a weak coupling method to an external heat bath to MD simulations. The heat bath can add or remove heat from the system in order to maintain a constant temperature. The rate of temperature change is governed by Eq. 2-9:

$$\frac{dT(t)}{dt} = \frac{1}{\tau_T}\left(T_0 - T(t)\right) \tag{2-9}$$

where T_0 is the temperature of the bath and τ_T is the coupling time which indicates the time scale a system relaxes to target value. By employing a coupling time, the MD propagation can avoid sudden change in velocities.

Since temperature is computed from velocities of all the atoms, what the Berendsen thermostat really does is to multiply all velocities with a scaling factor λ (shown in Eq. 2-10) in order to rescale the current temperature *T* to the target value T_0.

$$\lambda = \left[1 + \frac{\Delta t}{\tau_T}\left(\frac{T_0}{T} - 1\right)\right]^{1/2} \tag{2-10}$$

By rescaling velocities, the Berendsen thermostat controls the temperature in MD simulations. As mentioned before, the coupling time τ_t determines how tightly the system and the heat bath coupled together. A large τ_t means the coupling is weak. It takes long time for the system to relax from current temperature to target temperature. As $\tau_t \to \infty$, the internal energy will be conserved and the microcanonical ensemble will be restored. If τ_t is small, the coupling between the system and the heat bath is strong and the velocity scaling factor is large. However, large velocity scaling factor will cause large disruption in the momentum part of the phase space trajectory. The larger the scaling factor is, the less natural the trajectory is.

Langevin dynamics belongs to the category of stochastic thermostat.[137] It mimics the Brownian motion of a particle. Instead of Newton's second law, the equation of motion of MD method when using stochastic thermostat becomes:

$$\frac{d\vec{v}_i}{dt} = -\frac{1}{m_i}\frac{\vec{\nabla}U}{d\vec{q}_i} - \gamma\vec{v}_i + A(t) \quad (2\text{-}11)$$

In Eq. 2-11, $\vec{v}_i$, $\vec{q}_i$ and m_i are the velocity, position and mass of particle *i* respectively, *U* is the potential energy, *γ* is the friction coefficient and *A(t)* is a random force at time *t*. The amplitude of this force is determined by fluctuation-dissipation theorem (Eq. 2-12).

$$\langle A_i(t_1)A_j(t_2)\rangle = 2\gamma k_B T\delta_{ij}\,\delta(t_1 - t_2) \quad (2\text{-}12)$$

$\langle A_i(t_1)A_j(t_2)\rangle$ is the time correlation of *A* on particle *i* at time t_1 with *A* on particle *j* at time t_2. γ is the friction coefficient, k_B is the Boltzmann constant, *T* is the temperature, δ_{ij} is the Kronecker delta function and $\delta(t_1 - t_2)$ is the Dirac delta function. Langevin dynamics can be used as thermostat because the equation of motion is temperature dependent via the random force term.

2.2.3 Pressure Control in MD Simulations

Most biological experiments are performed in a constant pressure and constant temperature situation (NPT ensemble). Therefore, pressure control techniques (barostats) should be used in simulations to maintain system pressures and it is done by adjusting the system volumes. Since the number of particles is constant during a simulation, another application of maintaining pressure is to regulate system density which should be at certain appropriate value. A generally employed barostat is the Berendsen barostat.[146]

The pressure of a system in a simulation is calculated using the virial theorem of Clausius and can be expressed as:

$$P = \frac{1}{V}\left[Nk_BT - \frac{1}{3}\sum_{i=1}^{N}\sum_{j=i+1}^{N} r_{ij}\,\frac{dv(r_{ij})}{dr_{ij}}\right] \quad (2\text{-}13)$$

In the above equation, *P* is pressure, *N* is the number of particles, and *T* is the temperature. r_{ij} and $v(r_{ij})$ are the distance and interaction energy between atoms i and j, respectively.

Analogous to temperature control, the pressure can be maintained simply by rescaling volume at each time step although the system volume will be disrupted too much. Berendsen barostat was developed in order to smooth the change in volume. The Berendsen barostat, in which the algorithm is the same as Berendsen thermostat, utilizes a pressure bath. The rate of pressure change is governed by following equation:

$$\frac{dP(t)}{dt} = \frac{1}{\tau_P}\left(P_0 - P(t)\right) \quad (2\text{-}14)$$

where τ_P is the coupling constant and P_0 is the pressure of the bath.

The change in pressure is reflected by adjusting system volume. The coordinates of all particles in the system are scaled by a factor $\lambda^{1/3}$ and λ is formulated as:

$$\lambda = 1 - \kappa\frac{\Delta t}{\tau_P}(P - P_0) \quad (2\text{-}15)$$

The κ in the above equation is the isothermal compressibility. It represents the volume fluctuation caused by pressure change:

$$\kappa = -\frac{1}{V}\left(\frac{\partial V}{\partial P}\right) \quad (2\text{-}16)$$

2.3 Monte Carlo (MC) Method

2.3.1 Canonical Ensemble and Configuration Integral

In statistical mechanics, an ensemble is a collection of a very large number of systems and each system is a replica (on a thermodynamic level) of a particular thermodynamic system of interest. If the thermodynamic system of interest has a volume of *V*, *N* number of particles and temperature *T*, then an ensemble containing a very large number of such systems is called the canonical ensemble. The canonical ensemble is important because it best represents systems of interest in practice. Because each system of the canonical ensemble is not isolated, the energy of each system is not fixed. Thus, there is a probability of finding a system with energy E_i and the probability distribution of systems in the canonical ensemble is the so-called Boltzmann distribution (Eq. 2-17).

$$P_i = \frac{1}{Q} e^{-E_i/(k_B T)} \tag{2-17}$$

Here *Q* is the partition function and is essentially a normalization factor. E_i is the quantum energy of a system.

$$Q = \sum_i e^{-E_i/(k_B T)} \tag{2-18}$$

In classical mechanics, the Hamiltonian function *H* is employed to describe the total energy of a system and can be expressed as $H(p, q)$ where *p* and *q* are momenta and positions respectively. In general, the Hamiltonian can be separated into kinetic energy which depends only on momenta and potential energy which depends only on positions. In addition to using the Hamiltonian instead of quantum energy, the energy levels become continuous in the classical limit. Thus, the partition function will be written as an integral.

$$Q = \iint e^{-\beta H(p,q)} dpdq \quad (2\text{-}19)$$

Here $\beta = 1/(k_B T)$. After integrating the kinetic energy term, the partition function has the form of Eq. 2-20 and is called configuration integral.

$$Z = \int e^{-\beta U(q)}\, dq \quad (2\text{-}20)$$

Thus, the Boltzmann distribution in the classical limit is given by Eq. 2-21:

$$P = \frac{1}{Z} e^{-\beta U} \quad (2\text{-}21)$$

2.3.2 Markov Chain Monte Carlo (MCMC)

The definition of Markov chain is crucial to the MCMC methods, so it will be explained first in this section. Consider a stochastic process at discrete steps ($t_1, t_2, \ldots$) for a system that has a set of states ($S_1, S_2, \ldots$) with finite size. We define that the system is in state X_t at step t. The conditional probability of $X_{t_n} = S_j$ given that $X_{t_{n-1}}$ is in state S_i, etc, is:

$$P\left(X_{t_n} = S_j \,|X_{t_{n-1}} = S_i, X_{t_{n-2}} = S_k, \ldots, X_{t_1} = S_h\right) \quad (2\text{-}22)$$

A Markov process is defined in Eq. 2-22 with the property that the conditional probability of $X_{t_n} = S_j$ only depends on its previous state $X_{t_{n-1}} = S_i$:

$$P\left(X_{t_n} = S_j \,|X_{t_{n-1}} = S_i, X_{t_{n-2}} = S_k, \ldots, X_{t_1} = S_h\right) = P\left(X_{t_n} = S_j \,|X_{t_{n-1}} = S_i\right) \quad (2\text{-}23)$$

The corresponding sequence of states ($X_1, X_2, \ldots$) is called a Markov chain. The conditional probability $P\left(X_{t_n} = S_j \,|X_{t_{n-1}} = S_i\right)$ is essentially the transition probability from state S_i to S_j and is denoted as $w(i \to j)$. Based on the probability theory, a transition probability has the properties $w(i \to j) \geq 0$ and $\sum_j w(i \to j) = 1$. Thus, the probability of $X_{t_n} = S_j$ can be written as:

$$P\left(X_{t_n} = S_j\right) = P\left(X_{t_n} = S_j \,|X_{t_{n-1}} = S_i\right) \cdot P\left(X_{t_{n-1}} = S_i\right) = w(i \to j) P\left(X_{t_{n-1}} = S_i\right) \quad (2\text{-}24)$$

A change in $P\left(X_{t_n} = S_j\right)$ with respect to step is governed by the master equation:

$$\frac{dP\left(X_{t_n}=S_j,t\right)}{dt} = -\sum_i w(j \to i)P\left(X_{t_n} = S_j\right) + \sum_j w(i \to j)P\left(X_{t_n} = S_i\right) \quad (2\text{-}25)$$

At equilibrium (or under steady-state approximation), it is clear that $P\left(X_{t_n} = S_j\right)$ should not change with steps. This leads to:

$$\sum_i w(j \to i)P\left(X_{t_n} = S_j\right) = \sum_j w(i \to j)P\left(X_{t_n} = S_i\right) \quad (2\text{-}26)$$

Since the Markov chain introduced above possesses discrete and finite number of states, the transition probability can be described as a matrix, which is called the transition matrix. The $(i,j)^{\text{th}}$ element of the transition matrix represents $w(i \to j)$. The probability distribution can be represented by a row vector. Multplying a probability distribution with transition matrix will generate a new probability distribution. If a Markov chain is time-homogeneous (the definition of time is essentially a step due to the stochastic nature of a Markov chain), the elements of transition matrix are constants (time-independent). When a probability distribution vector is not changed by multiplying with the transition matrix, the distribution is said to be stationary.

At equilibrium, the elements of the transition matrix are independent of time. The equilibrium distribution is an eigenvector of the transition matrix with an eigenvalue of 1. Hence, multiplying equilibrium probability distribution with transition will not change it.

Properties of a Markov chain include: a Markov chain is irreducible, if all states communicate with each other; a Markov chain is called aperiodic, if number of steps needed to move between two states is not periodic; it is positive recurrent, if the expectation value of the return time to a state is finite. These properties are closely related to the ergodicity of a Markov chain.

The MCMC methods are Monte Carlo samplings from a probability distribution by employing a Markov chain whose equilibrium probability distribution is the intended probability distribution. States sampled by Monte Carlo method form a Markov chain. The transitions in MCMC must satisfy the detailed balance equation:

$$w(j \to i)P\left(X_{eq} = S_j\right) = w(i \to j)P\left(X_{eq} = S_i\right) \quad (2\text{-}27)$$

A Markov chain is said to be reversible when it satisfies the detailed balance equation.

2.3.3 The Metropolis Monte Carlo Method

In 1953, Metropolis *et al*.[50] proposed an algorithm to sample the phase space of a system at equilibrium by the MC method. According to the Metropolis algorithm, at configuration *i*, a new configuration *j* is chosen, both configurations are weighted by Boltzmann distribution (Eq. 2-21) and the detailed balance condition (Eq. 2-27) is employed to evaluate the transitions (MC moves) between configurations,

$$P(i)w(i \to j) = P(j)w(j \to i) \quad (2\text{-}28)$$

In the above equation, $P(i)$ is the Boltzmann weight of configuration *i* and $w(i \to j)$ is the transition probability from configuration *i* to *j*. Inserting Eq. 2-21 into Eq. 2-28 and rearranging Eq. 2-28 yields:

$$\frac{w(i \to j)}{w(j \to i)} = \frac{P(j)}{P(i)} = e^{-\beta(U(j)-U(i))} = \mathrm{e}^{-\Delta} \quad (2\text{-}29)$$

And the transition probability from configuration *i* to *j* can be written as:

$$w(i \to j) = min\{1, e^{-\Delta}\} \quad (2\text{-}30)$$

In practice, the new configuration is accepted if $\Delta \leq 0$. However, if $\Delta > 0$, a random number between zero and one is generated and is compared with $e^{-\Delta}$. If the random number is less than or equal to $e^{-\Delta}$, then the new configuration is accepted. Otherwise,

the current configuration is kept and is added to the configuration ensemble. This accept/reject criterion is the so-called Metropolis criterion. The MC sampling with the Metropolis criterion generates a Markov chain whose equilibrium PDF is the Boltzmann distribution. Compare the Metropolis MC with MD, MC method simulates a system in the canonical ensemble without controlling temperatures; the bottleneck of MC sampling is the potential energy difference while the bottleneck of MD is the energy barrier.

2.3.4 Ergodicity and the Ergodic Hypothesis

In statistical mechanics, ergodic (adjective of ergodicity) is a word used to describe a system which satisfies the ergodic hypothesis. The ergodic hypothesis states that over a long period of time, the time average and the ensemble average of a property should be the same. In our simulations, the ergodic hypothesis is often assumed to be true. Ergodicity breaking (the ergodic hypothesis does not hold) often means that the system is trapped in a local region of the phase space. One example when the ergodic hypothesis does not hold is the spontaneous magnetization of a ferromagnetic system below Curie temperature. The ensemble average of net magnetization is zero since spin up and spin down are degenerate states and the population of either states should be the same. However, a net magnetization exists when temperature is below Curie temperature. Ergodicity is often discussed in a Markov chain. A Markov chain is called ergodic when all its states are irreducible, aperiodic and have positive recurrent.

2.4 Solvent Models

Because proteins are stable and perform their functions in condensed phase, especially in aqueous solution, representing the solvation effect is of great importance. One frequently used solvent model in MD simulations is the water model. Two ways of representing aqueous solution are present here: the explicit and the implicit solvent

models. As its name indicates, the explicit water model employs water molecules in the simulation and the implicit water model treats water as a dielectric continuum.

2.4.1 Explicit Solvent Model

Different types of water molecules such as SPC/E,[149] TIP3P,[150] and TIP4P[150] are developed. Water molecules parameters are fitted to bulk water properties such as density, heat of vaporization, and dipole moment.[150] The density of liquid water is an important physical quantity to check the water models. The density of liquid water shows a maximum at 4° C and water models should correctly reflect this. TIP3P failed to achieve that, while TIP4P and TIP5P[151] and their variants were able to reproduce this trend. Take the TIP3P and TIP4P water models as examples. A simple diagrammatical description of TIP4P and TIP4P water models are shown in Figure2-2. The TIP3P water model has one oxygen atom and two hydrogen atoms. The geometry of TIP3P water is the same as experimental geometry with OH bond length of 0.9572 Å and HOH angle of 104.52°. Only oxygen atom has a van der Waals radius. Thus, the van der Waals interactions only occur among oxygen atoms. Partial charges are placed on oxygen atom and hydrogen atoms. The partial charge on the oxygen atom is -0.834e and the partial charge on each hydrogen atom is 0.417e, where e is the charge of an electron. When computing interactions (Coulomb interaction and Lennard-Jones interaction) between two TIP3P water molecule, there are 3×3=9 distances needed to be calculated. The TIP4P water model, as its name implies, has four sites. Similar to the TIP3P water model, experimental geometry (bond length and bond angle) is also adopted in the TIP4P model. The only atom, in the TIP4P molecule, having the van der Waals interaction is oxygen too. However, for the TIP4P model, the negative partial

charge is located on the fourth site, instead of being placed on the oxygen atom, as in the TIP3P model. The use of the fourth site carrying negative charge is able to improve electrostatic properties of water such as dipole moment. The positive partial charges are still placed on hydrogen atoms. The new partial charges are -1.04e and 0.52e. New Lennard-Jones potential parameters have also been employed for the TIP4P water model to achieve better fitting results. Computing the interactions between a pair of the TIP4P molecules requires knowing 9 distances for electrostatic interactions and 1 distance for the Lennard-Jones potential. Therefore, using TIP4P model in a simulation will be computationally more expensive than using TIP3P model. For a five-site water model such as TIP5P, 17 distances are needed in order to calculate water-water interactions.

When simulating a molecule with explicit water molecules, the periodic boundary condition (PBC) is utilized in order to mimic reality.[152] Otherwise, water molecules evaporate into vacuum. Ewald summation[153] or Particle-Mesh Ewald (PME) summation[154] is employed to compute the long-range electrostatics efficiently when the PBC is employed.

One advantage of employing the explicit water model is that the solvent-solute interaction can be represented. For example, studying the hydrogen bonding between water molecules and proteins requires using the explicit water model. However, it suffers from computational cost. CPU time is approximately proportional to number of inter-atomic interactions.

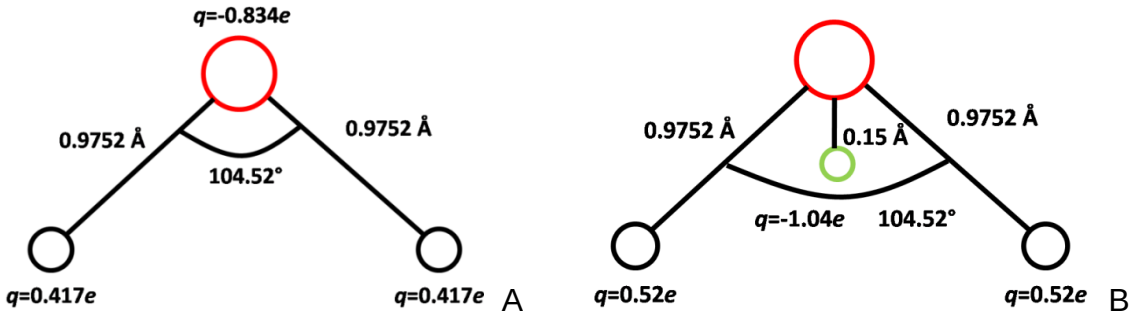


Figure 2-2. A diagrammatic description of TIP3P and TIP4P water models. A) TIP3P model. The red circle is oxygen atom and the black circles are the hydrogen atoms. Experimental bond length and bond angle are adopted. B) TIP4P model. Oxygen and hydrogen atoms are labeled with same color as in the TIP3P model. TIP4P model also employs the experimental OH bond length and HOH bond angle. Clearly, the fourth site (green circle) which carries negative partial charge has been added to the TIP4P model.

2.4.2 The Poisson-Boltzmann (PB) Implicit Solvent Model

An alternative way of representing solvation effect is to reproduce the PES after a molecule is dissolved in solvent. The solution-phase potential energy of a molecule can be computed by adding solvation free energy to the gas-phase potential energy. Given the correct solution-phase PES, correct forces can be generated for the equation of motion. Thus, the key issue is finding the accurate free energy of solvation. A dielectric continuum model can be employed to calculate free energy of solvation. In the dielectric continuum model, the free energy (work) of assembling a charge distribution is expressed as:

$$G = -\frac{1}{2}\int \rho(\vec{r})\phi(\vec{r})dq \qquad (2\text{-}31)$$

Here $\rho(\vec{r})$ is the charge density of the molecule and $\phi(\vec{r})$ is the electrostatic potential.

The Poisson-Boltzmann model utilizes the Poisson-Boltzmann equation to describe the electrostatic potential as a function of charge density. In practice, the

linearized PB equation (Eq. 2-32), which utilizes the first order truncation of Taylor series expansion of the hyperbolic sine, is often employed.

$$\vec{\nabla} \cdot \left[\epsilon(\vec{r})\vec{\nabla}\phi(\vec{r})\right] = -4\pi\rho(\vec{r}) + \epsilon(\vec{r})\lambda(\vec{r})\kappa^2\phi(\vec{r}) \quad (2\text{-}32)$$

In the above equation, ϵ is the dielectric constant, λ is a switching function which is zero when electrolyte is inaccessible and otherwise one, and κ^2 is the Debye-Hückel parameter.

For simple cases such as spherical charge distributions, the solutions to PB equation are analytical and simple. Consider dissolving a sphere with charge q and radius a and the charge is uniformly distributed on the surface. The charge density on the surface can be expressed as:

$$\rho(x) = \frac{q}{4\pi a^2} \quad (2\text{-}33)$$

Here x is any point on the surface. From outside of the sphere, the electrostatic potential at $\vec{r}$ is calculated by:

$$\phi(\vec{r}) = -\frac{q}{\epsilon|r|} \quad (2\text{-}34)$$

Integrating the right-hand side of Eq. 2-31 from infinity to a with Eq. 2-33 and Eq. 2-34 will yield $G = \frac{q^2}{2\epsilon a}$. The free energy of solvation is the difference between gas-phase and solution-phase free energies. Thus, it can be written as:

$$\Delta G_{sol} = -\frac{1}{2}\left(1 - \frac{1}{\epsilon}\right)\frac{q^2}{a} \quad (2\text{-}35)$$

This is the so-called Born equation and is the basis of the generalized Born (GB) method which will be introduced later.

For complex systems such as proteins, there is no analytical solution to the linearized PB equation.[73] Therefore, this equation is solved iteratively until self-consistent is achieved for the charge density and electrostatic potential.

2.4.3 The Generalized Born (GB) Implicit Solvent Model

Solving the linearized PB equation is computationally expensive. An approximate method to the PB implicit solvent model is proposed as the GB method.[39,117] Using the GB implicit solvent can greatly shorten the simulation time, which makes the GB frequently employed in molecular simulations. Similar to Eq. 2-35, the free energy of solvation in the GB method is given by:

$$\Delta G_{sol} = \frac{1}{2}\left(\frac{1}{\epsilon} - 1\right)\sum_i \sum_j \frac{q_i q_j}{f_{GB}} \tag{2-36}$$

Here q_i and q_j are charges on nuclei i and j. f_{GB} is calculated by:

$$f_{GB} = \left({r_{ij}}^2 + a_i a_j e^{-r_{ij}^2/4a_i a_j}\right)^{\frac{1}{2}} \tag{2-37}$$

Here a_i is the effective Born radius of charge q_i, and r_{ij} is the distance between the two charges.

Another approximation in the GB method is the Coulomb field approximation.[40] This approximation estimates the effective Born radius by integrating the energy density of a Coulomb field over the molecular volume. The integral is often evaluated numerically. One should notice that the GB theory involves two approximations to reproduce the PB results. The first approximation contains Eq. 2-36 and 2-37. The second one is the Coulomb field approximation. Further approximations are often introduced to reduce the time computing the effective Born radii in practice. The pair-wise approximation[155] is often applied. In this approximation, the van der Waals radius

of an atom and a function dependent on positions and the van der Waals radii of atom pairs are utilized to compute the effective Born radius.

2.5 pK_a Calculation Methods

2.5.1 The Continuum Electrostatic (CE) Model

The basic idea of the CE model is also given in Figure1-6. Since computing the pK_a value of an ionizable residue in a protein directly is difficult (breaking a bond plus dissolving all species into water), a model compound is utilized and the pK_a shift is calculated via the thermodynamic cycles shown in Figure 1-7 and Figure 1-8. Like the MDFE calculations, the CE model also computes the pK_a value of an ionizable residue relative to its intrinsic value (or model compound value according to the definition of Bashford and Karplus; the definition of the intrinsic pK_a can be found in section 1.3). The pK_a value of an ionizable residue is written as:

$$pK_a(protein) = pK_a(model) + \frac{1}{2.303 k_B T}\left(\Delta G_{AH\to A}(protein) - \Delta G_{AH\to A}(model)\right) \quad (2\text{-}38)$$

In the above equation, $pK_a(model)$ is the intrinsic pK_a value of an ionizable residue and can be found in Table 1-1. $\Delta G_{AH\to A}(protein)$ and $\Delta G_{AH\to A}(model)$ is the free energy difference between protonated and deprotonated species for that ionizable residue and its reference compound (the reference compound utilized in the CE model is an isolated ionizable residue with two ends capped and fully exposed to aqueous environment.), respectively. Eq. 2-38 is essentially the same as Eq. 1-20. The difference between MDFE methods and the CE model is how the free energy differences between the protonated and deprotonated species on the right-hand side of Eq.1-20 are generated. MDFE methods compute the two free energy differences via free energy calculation algorithms while the CE model calculates them via FDPB method. In this

continuum electrostatic model, proteins are considered as low-dielectric regions surrounded by high-dielectric continuum representing water. Protonation is represented by adding a unit charge to the ionizable site.

In the continuum electrostatic model, $\Delta G_{AH\rightarrow A}(protein)$ and $\Delta G_{AH\rightarrow A}(model)$ are assumed to differ only in their electrostatic contributions. This assumption will result in the cancellation of non-electrostatic free energy contributions. Thus, calculating the electrostatic work of charging a site in the ionizable residue and in the reference compound from zero to unit charge is required. This electrostatic work can be further decomposed into three terms. For any ionizable site in a fixed protein structure, the electrostatic work consists of three terms: the Born solvation free energy (ΔG_{Born}), the background free energy which is the interaction of the ionizable site with non-ionizable charges (ΔG_{back}), and the interaction with other ionizable sites ($\Delta G_{interact}$). For the reference compound, only the first two terms exist. Thus, $\Delta G_{AH\rightarrow A}(protein)$ can be written as:

$$\Delta G_{AH\rightarrow A}(protein) = \Delta G_{Born}(protein) + \Delta G_{back}(protein) + \Delta G_{interact}(protein) \quad (2\text{-}39)$$

And $\Delta G_{AH\rightarrow A}(model)$ can be written as:

$$\Delta G_{AH\rightarrow A}(model) = \Delta G_{Born}(model) + \Delta G_{back}(model) \quad (2\text{-}40)$$

Linearized PB equation (described in Section 2.3.2) is solved for electrostatic potentials using finite difference method. For an ionizable site i, the Born solvation is determined by Eq. 2-35. The background free energy is calculated using Eq. 2-41:

$$\Delta G_{back} = \sum_k q_i\, q_k \phi(r_i, r_k) \quad (2\text{-}41)$$

Here q_k is non-ionizable partial charge and $\phi(r_i, r_k)$ stands for the electrostatic potential produced at r_k by a unit charge place at r_i. The electrostatic interaction with

other ionizable sites can also be evaluated by Eq. 2-41 except that charges on ionizable sites must be used. After computing all components on the right-hand sides of Eq. 2-38 and Eq. 2-39, the pK_a of ionizable residue i will be obtained.

To produce a titration curve, a protein containing *N* ionizable residues is considered here. Each ionizable residue has two states: protonated and deprotonated. Thus, there are 2^N numbers of macro-states for that protein. Each macro-state can be represented by a vector $\vec{x}$=(x_1, x_2, …, x_N), whose elements x_i is 0 or 1 according to whether ionizable site i is deprotonated or protonated. The free energy of $\vec{x}$ relative to the vector whose components are all zero (this is equivalent to the free energy change when charging the non-zero components in the vector) is given by Eq. 2-42:

$$\Delta G(\vec{x}) = \sum_{i=1}^{N} \Delta G_i \cdot x_i + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} W_{i,j}(q_i^0 + x_i)(q_j^0 + x_j) \quad (2\text{-}42)$$

Here $\Delta G_i = \Delta G_{Born}(protein) + \Delta G_{back}(protein) - \Delta G_{Born}(model) - \Delta G_{back}(model)$ for ionizable site i, $W_{i,j}$ is the electrostatic interaction between unit charges at ionizable site i and j, and q_i^0 is the charge of site i when it is in the deprotonated state. Thus, θ_i, which is the fraction of protonation of site i, can be written as (Eq. 2-43):

$$\theta_i = \frac{\sum_{\vec{x}} x_i e^{-\beta \Delta G(\vec{x}) - 2.303\, \nu(\vec{x}) pH}}{\sum_{\vec{x}} e^{-\beta \Delta G(\vec{x}) - 2.303\, \nu(\vec{x}) pH}} \quad (2\text{-}43)$$

Here $\beta = 1/k_B T$ and $\nu(\vec{x})$ is the number of non-zero components in $\vec{x}$. Summing up individual θ_i will generate a titration curve of the entire protein.

2.5.2 Free Energy Calculation Methods

As mentioned previously, the pK_a value is proportional to the standard free energy of reaction. Therefore, free energy calculation methods can be employed to compute the pK_a value of ionizable residue one is interested in. In this section, two frequently

used free energy calculation methods: thermodynamic integration (TI)[156,157] and free energy perturbation (FEP)[158] are described. Both TI and FEP belong to the so-called "slow-growth" or equilibrium method and can be employed to compute the free energy difference between two states. In other words, each transition should be reversible.

In the TI method, initial state *A* (having potential energy $U_A(q)$, where q is the molecular structure) and final state *B* (having potential energy $U_B(q)$) are connected by a reaction coordinate λ (this reaction coordinate does not necessarily have any physical significance). The simplest scheme of constructing the potential energy as a function of λ is:

$$U(\lambda) = (1-\lambda)U_A + \lambda U_B \tag{2-44}$$

Slowly transforming λ from zero to one converts state *A* to *B*; the intermediate values of λ correspond to a mixed system without physical meaning.

The Helmholtz free energy A in the canonical ensemble (or the Gibbs free energy G in the isothermal-isobaric ensemble) is formulated as:

$$A = -k_B T \ln Q = -k_B T \ln Z \tag{2-45}$$

where Q is the partition function and Z is the configuration integral. From now on, our derivation will focus on the canonical ensemble and the Helmholtz free energy but can be extended to isothermal-isobaric ensemble and the Gibbs free energy in the same manner (this statement also holds when the free energy perturbation method is described later). Following Eq. 2-45, the Helmholtz free energy as a function of λ is:

$$A(\lambda) = -k_B T \ln Z(\lambda) = -k_B T \int e^{-U(q,\lambda)/k_B T} dq \tag{2-46}$$

Here, U is the potential energy function and q is molecular structure.

The free energy difference can be written as:

$$\Delta A_{A\to B} = A_B - A_A = \int_0^1 \partial A/\partial\lambda \, d\lambda \tag{2-47}$$

Then,

$$\frac{\partial A(\lambda)}{\partial\lambda} = -k_B T \frac{\partial \ln Z(\lambda)}{\partial\lambda} = -k_B T \frac{1}{Z(\lambda)} \frac{\partial Z(\lambda)}{\partial\lambda} \tag{2-48}$$

Plugging the explicit form of configuration integral into the derivative leads to:

$$\frac{\partial Z(\lambda)}{\partial\lambda} = \frac{\partial}{\partial\lambda} \int e^{-U(q,\lambda)/k_B T} dq = \int \frac{\partial \left(e^{-U(q,\lambda)/k_B T}\right)}{\partial\lambda} dq \tag{2-49}$$

$$\frac{\partial \left(e^{-U(q,\lambda)/k_B T}\right)}{\partial\lambda} = e^{-U(q,\lambda)/k_B T} (-1/k_B T) \frac{\partial U(\lambda)}{\partial\lambda} \tag{2-50}$$

Therefore,

$$-k_B T \frac{1}{Z(\lambda)} \frac{\partial Z(\lambda)}{\partial\lambda} = -k_B T \frac{1}{Z(\lambda)} \int e^{-U(q,\lambda)/k_B T} (-1/k_B T) \frac{\partial U(\lambda)}{\partial\lambda} dq \tag{2-51}$$

Since the integration is over coordinate space, the configuration integral can be moved into the integral. Eq. 2-51 now becomes:

$$\frac{\partial A(\lambda)}{\partial\lambda} = -k_B T \frac{1}{Z(\lambda)} \frac{\partial Z(\lambda)}{\partial\lambda} = \int \frac{e^{-U(q,\lambda)/k_B T}}{Z(\lambda)} \cdot \frac{\partial U(\lambda)}{\partial\lambda} dq \tag{2-52}$$

The first term in the integrand is the Boltzmann weight factor $P(q,\lambda)$. Rewriting Eq. 2-51 yields:

$$\frac{\partial A(\lambda)}{\partial\lambda} = \int P(q,\lambda) \frac{\partial U(\lambda)}{\partial\lambda} dq = \langle \frac{\partial U(\lambda)}{\partial\lambda} \rangle_\lambda \tag{2-53}$$

Thus, the final form of $\Delta A_{A\to B}$ is:

$$\Delta A_{A\to B} = \int_0^1 \partial A/\partial\lambda \, d\lambda \; = \int_0^1 \langle \frac{\partial U}{\partial\lambda} \rangle_\lambda \, d\lambda \tag{2-54}$$

In both Eq. 2-53 and 2-54, the bracket represents an ensemble average generated at λ.

In pK_a calculations, state *A* (or *B*) represents the protonated species and the other represents the deprotonated species. Each intermediate value λ corresponds to a mixed protonated and deprotonated state, without any physical meaning. When

classical force fields are applied, the proton becomes a dummy atom in the deprotonated state but retains its position and velocity in the protein (or model compound). Furthermore, state *A* and *B* only differ in charge distributions. Dissociation free energy can be computed using methods of numerical integration (such as trapezoidal rule or Gaussian quadrature) to treat Eq. 2-54. As explained in the previous chapter, the quantum mechanical contributions to the proton dissociation free energy are assumed to be the same for protein and the model compound. Therefore, subtracting dissociation free energy of model compound from that of protein will yield the pK_a shift relative to the pK_a value of the model compound.

The FEP method, which was initially introduced by Zwanzig in 1954,[158] is another frequently employed free energy calculation method. Consider two states (*A* and *B*) with partition functions Q_A and Q_B, respectively, and the Helmholtz free energy A_A and A_B, respectively. The free energy difference from *A* to *B* can be expressed as:

$$\Delta A_{A\to B} = A_B - A_A = -k_B T \ln(Q_B/Q_A) \tag{2-55}$$

Suppose the configuration integrals Z are adopted instead of partition functions. The potential energy function of state *A* and *B* is $U_A(q)$ and $U_B(q)$, respectively, where q is the molecular structure. Thus,

$$\Delta A_{A\to B} = -k_B T \ln(Z_B/Z_A) = -k_B T \ln\left(\int \left(e^{-U_B(q)/k_B T}/Z_A\right) dq\right) \tag{2-56}$$

According to Zwanzig, $U_B(q)$ can be written as the sum of $U_A(q)$ and a perturbation term $\Delta U(q)$.

$$U_B(q) = U_A(q) + \Delta U(q) \tag{2-57}$$

$$\Delta A_{A\to B} = -k_B T \ln\left(\int \left(e^{-\left(U_A(q)+\Delta U(q)\right)/k_B T}/Z_A\right) dq\right) \tag{2-58}$$

$$\Delta A_{A\to B} = -k_B T \ln\left(\int \frac{e^{-U_A(q)/k_B T} \cdot e^{-\Delta U(q)/k_B T}}{Z_A} dq\right) \tag{2-59}$$

The Boltzmann weight factor of state *A* has the form:

$$P_A(q) = e^{-U_A(q)/k_BT}/Z_A \tag{2-60}$$

Therefore,

$$\Delta A_{A\to B} = -k_BT\ln\left(\int P_A(q)e^{-\Delta U(q)/k_BT}dq\right) = -k_BT\ln\langle e^{-\Delta U(q)/k_BT}\rangle_A \tag{2-61}$$

The bracket with subscript *A* stands for the ensemble average performed on the structural ensemble generated from state *A*. Substituting $\Delta U(q)$ with $U_B(q) - U_A(q)$, Eq. 2-61 becomes:

$$\Delta A_{A\to B} = -k_BT\ln\langle e^{-\left(U_B(q)-U_A(q)\right)/k_BT}\rangle_A \tag{2-62}$$

In order to compute $\Delta A_{A\to B}$, one simulation of state *A* is performed. Once a configuration q is generated, the potential energy difference at configuration q is computed. The ensemble average of $e^{-\left(U_B(q)-U_A(q)\right)/k_BT}$ can be calculated easily and hence, $\Delta A_{A\to B}$ is obtained. According to Eq. 2-62, if the potential energy difference between the two states (perturbation) is too large, the free energy difference given by FEP calculation can be unreasonably large. Thus, FEP calculations cannot accurately reflect the true free energy difference of large changes in Hamiltonian (basically, potential energy). Only similar Hamiltonians contributes to the free energy difference.

In order to compute the free energy difference between two very different systems (such as calculating free energy difference from benzene to toluene), intermediate systems mixing the two very different systems (end points) are adopted in such a way that the differences between neighbors can be treated as perturbations. To be specific, a coupling parameter can be adopted in the same fashion as TI. The sum of free energy difference between intermediate systems (each intermediate state has a specific coupling parameter λ_i) will be the targeted free energy difference.

In practice, computing $\Delta A_{A\to B}$ (forward free energy difference) is equally easy (or hard) as computing $\Delta A_{B\to A}$ (backward free energy difference) and one is exactly the opposite of the other in principle. Evaluation of forward and backward free energy differences provides an indication of convergence. The Bennett Acceptance Ratio (BAR) method[159,160] is a frequently employed scheme to reduce sampling bias and statistical error.

In 1985, Jorgensen *et al*.[161] proposed a "double-wide" scheme to perform FEP calculations in order to reduce the computational cost. The double-wide FEP can be explained by the following example. Suppose $\Delta A(\lambda_i \to \lambda_j)$ is to be computed. Instead of performing two MD simulations at λ_i and λ_j, only one MD simulation at $\lambda_{\left(\frac{i+j}{2}\right)}$ is conducted. The $\Delta A\left(\lambda_i \to \lambda_{\left(\frac{i+j}{2}\right)}\right)$ and $\Delta A\left(\lambda_j \to \lambda_{\left(\frac{i+j}{2}\right)}\right)$ are calculated then the objective free energy difference can be obtained. If N configurations of each MD simulation are taken in order to compute $\Delta A(\lambda_i \to \lambda_j)$, the conventional FEP scheme requires $4N$ potential energy calculations, while double-wide FEP only requires $3N$.

2.5.3 Constant-pH MD Methods

As described in the previous chapter, the constant-pH MD methods want to describe protonation equilibrium correctly at a given pH value. The constant-pH MD models sample protonation state space explicitly, along with the sampling of conformational space. In practice, two protonation state sampling schemes have been developed. One scheme utilizes a binary protonation state space: only the protonated and deprotonated states are defined. MC steps have been performed periodically during MD propagations, which sample the conformational space. At each MC step, a new

protonation state is selected and the free energy difference between the old and new states is computed. The Metropolis criterion is the applied to evaluate the MC move. Since a binary protonation state space is adopted, this scheme is generally called the discrete protonation state model. The other scheme employs a continuous protonation state space. Not only the completely protonated and deprotonated species are defined, fractional protonation states also exist in the simulation. The MD propagations sample both conformational and protonation state space. The latter scheme is named continuous protonation state model. In this section, the CPHMD model developed by the Brooks group and two discrete protonation state constant-pH MD methods developed by Baptsta *et al*. and by Mongan *et al.* are described to provide a brief overview.

In the CPHMD method, Lee *et al*.[114] applied λ-dynamics[116] to the protonation coordinate and used the Generalized Born (GB) implicit solvent model. They chose a λ variable, which is bound between 0 and 1, to control protonation fraction. $\lambda = 0$ represents an ionizable residue in its protonated state, while $\lambda = 1$ corresponds to the deprotonated ionizable residue. Due to its continuous nature, $\lambda = 0$ and $\lambda = 1$ are rarely sampled. Thus an arbitrary value λ_p is adopted such that any λ value smaller than λ_p is defined to be protonated, while any λ is greater than $1 - \lambda_p$ is set to be deprotonated. To ensure an unbounded reaction coordinate is practically used, a new coordinate θ is introduced and is propagated in a MD simulation. λ is expressed as:

$$\lambda = sin^2(\theta). \tag{2-63}$$

An artificial potential barrier between the protonated and deprotonated states has been introduced. The potential is a biasing potential to increase the residency time

close to protonation/deprotonation states and it is centered at half way point of titration (λ=1/2). The formula of this biasing potential used by Lee *et al.* is

$$U_{bias} = -4\beta\left(\lambda - \frac{1}{2}\right)^2 \quad (2\text{-}64)$$

where β is an adjustable parameter controlling the height of the biasing potential. A value of 1.25 kcal/mol is found enough to provide occupation time in the protonated and deprotonated states.

The total potential of the system, which provides the forces for MD propagation, has the form:

$$U_{total} = U_{bond} + U_{angle} + U_{torsion} + U_{elec}(\theta) + U_{vdw}(\theta) + U_{GB}(\theta) + U_{nonpolar} + \sum_{i=1}^{n}\left[-U^i_{model}(\theta_i) + U^i_{pH}(\theta_i) + U^i_{bias}(\theta_i)\right] \quad (2\text{-}65)$$

Here, the first five terms are essentially defined by Eq.1-9. U_{GB} is the GB solvation free energy which will be explained in the next chapter. $U_{nonpolar}$ is the energy related to surface accessible areas. i in Eq. 2-65 represents an ionizable residue. U_{model} is a potential of the mean force (PMF) in the titration coordinate for a model compound. The $\Delta G_{MM}(AH \rightarrow A^-)$ shown in Eq. 1-17 can be represented by $U_{model}(\lambda = 0) - U_{model}(\lambda = 1)$. The U_{model} in Eq. 2-65 is fit to a two-parameter parabolic function having the form $U^i_{model} = -A_i(sin^2(\theta_i) - B_i)^2$. $U^i_{pH}(\theta_i) = 2.303k_BT \cdot sin^2(\theta_i) \cdot \left(pK^i_a - pH\right)$, which is the chemical potential of adding a fractional proton to the solution at pH. The term $-U^i_{model}(\theta_i) + U^i_{pH}(\theta_i)$ is essentially the quantum mechanical dissociation free energy of a fractional proton. The CPHMD method also assumes Eq. 1-18 is true.

Another feature of the CPHMD method is using an extended Hamiltonian. A kinetic energy term of titration coordinate θ is employed in CPHMD:

$$K_\theta = \frac{1}{2}\sum_{i=1}^{n} M_i \dot{\theta}_i^2 \quad (2\text{-}66)$$

The fictitious mass M_i controls the speed of response of the protonation state change to the force on it.

Baptista *et al*.[118] proposed that MD simulations incorporating protonation state change is essentially a semigrand canonical ensemble. The joint PDF can be written as:

$$P(p, q, p_s, q_s, \vec{n}) = \frac{exp\,[\beta\mu n - \beta H(p,q,p_s,q_s,\vec{n})]}{\sum_{\vec{n}} \int exp\,[\beta\mu n - \beta H(p,q,p_s,q_s,\vec{n})] dpdqd\ p_s d q_s} \quad (2\text{-}67)$$

Here, p, q is the momenta and coordinates of solute, respectively. p_s and q_s is the momenta and coordinates of solvent, respectively. $\vec{n}$ is the vector containing protonation state information of each ionizable residue. The details of $\vec{n}$ is explained in the continuum electrostatic model. n is essentially the number of protonated ionizable residues. μ is the chemical potential of protons and $\beta = 1/k_B T$. The Hamiltonian contains quantum mechanical and classical force field terms. The quantum mechanical part in their model is assumed not to depend on coordinates and momenta. The introduction of dummy atom to replace the proton in a deprotonated residue makes kinetic energy only a function of momenta.

Two conditional samplings have been considered by Baptista *et al.*: one is conformational sampling under a fixed protonation state, the other one is protonation state samping under a fixed structure. The PDF of conformations at fixed protonation state is:

$$P(p, q, p_s, q_s|\vec{n}) = \frac{exp\,[-\beta H_c(p,q,p_s,q_s,\vec{n})]}{\int exp\,[-\beta H_c(p,q,p_s,q_s,\vec{n})] dpdqd\ p_s d q_s} \quad (2\text{-}68)$$

where H_c is the classical Hamiltonian. Due to the fact that quantum mechanical Hamiltonian depends only on protonation state, which is fixed in conformational

sampling, the quantum contribution is a constant and is canceled. The PDF of protonation states at fixed coordinates is given in Eq. 2-69:

$$P(\vec{n}|q) = \frac{exp\,[-2.303npH - \beta\,\Delta G(q,\vec{n})]}{\sum_{\vec{n}} exp\,[-2.303npH - \beta\,\Delta G(q,\vec{n})]} \tag{2-69}$$

where ΔG is the free energy of a protonation state relative to the completely deprotonated state. In their model, FDPB-based method is executed to calculate free energy difference.

Combining the two conditional samplings, one is able to generate an ensemble satisfying Eq.2-67. In order to prove the above statement, one must show the Markov chain constructed by transition matrix and the two conditional probabilities satisfies the following condition,

$$\boldsymbol{\rho} = \lim_{n\to\infty} \boldsymbol{p}\boldsymbol{W}^n \tag{2-70}$$

In the above equation, $\boldsymbol{\rho}$ is the joint PDF as defined in Eq. 2-67, $\boldsymbol{p}$ is a joint PDF depend on the same variables as $\boldsymbol{\rho}$, and $\boldsymbol{W}$ is transition matrix. Proving Eq. 2-70 holds means that one must prove the Markov chain defined by $\boldsymbol{p}$ and $\boldsymbol{W}$ is ergodic.

In order to prove a Markov chain is ergodic, one needs to prove (a) the Markov chain is irreducible; (b) the chain needs to be aperiodic; (c) the transition matrix elements are time-independent; and (d) the limiting distribution should be stationary. The detailed proof is given by Baptista *et al.* in their 2002 paper. Their proof justified the discrete protonation state constant-pH method which samples conformational space at fixed protoation state and samples protonation state at fixed structure.

In 2004, Mongan *et al*.[127] proposed a constant-pH MD method and implemented in the AMBER suite. This algorithm follows the scheme proposed by Baptista *et al*.[118] but employs the GB model in both MD and MC. Given a protein with *N* titratable sites, the

discrete protonation state model means protonation states of a protein are described by a vector $\vec{x}$=(x_1, x_2, …, x_N) where each x_i is some integer representing the protonation state of titratable residue *i*. In AMBER, five amino acids are designed to be titratable: aspartate, glutamate, histidine, lysine and tyrosine. For each titratable residue, different protonation states have different partial charges on the side chain. This model also includes syn and anti forms of protons for the aspartate and glutamate side chains as well as the δ and ε proton locations for histidine.

At each Monte Carlo step, a titratable site and a new protonation state for that site are chosen randomly and the transition free energy at this fixed configuration is used to evaluate the MC move.

Considering a titratable site A in a protein environment, its protonated form is protA-H and deprotonated form is protA-. The equilibrium between the two forms is governed by their free energy difference. This free energy difference is the ensemble average of different configurations. However, the free energy difference cannot be computed by a molecular mechanics (MM) model since the transition between two forms deals with bond breaking/forming and solvation of a proton which involves quantum mechanical effects.

The above problems can be solved by using a reference compound. The reference compound has the same titratable side chain as protA-H but with known pK_a value ($pK_{a,ref}$). Following Mongan et al., we assume the transition free energy can be divided into the quantum mechanics (QM) part and the molecular mechanics (MM) part. We further assume that the quantum mechanical energy components are the same between the reference compound and the protA-H. Since the pK_a of the reference

compound is known, its transition free energy from deprotonated form to protonated form at a given pH is:

$$\Delta G_{ref} = k_B T \ln 10 \left(pH - pK_{a,ref}\right) \quad (2\text{-}71)$$

So the QM component of the transition free energy can be expressed as:

$$\Delta G_{ref,QM} = \Delta G_{ref} - \Delta G_{ref,MM} \quad (2\text{-}72)$$

Here $\Delta G_{ref,MM}$ is the molecular mechanics contribution to the free energy of protonation reaction for that reference compound. In practice, the QM component of the transition free energy also contains errors from MM calculations so it's actually a non-MM component. Since the approximation of the QM component of the transition free energy is:

$$\Delta G_{ref,QM} = \Delta G_{protein,QM} \quad (2\text{-}73)$$

Then the transition free energy from protA- to protA-H can be calculated as:

$$\Delta G = k_B T \ln 10 \left(pH - pK_{a,ref}\right) + \Delta G_{MM} - \Delta G_{ref,MM} \quad (2\text{-}74)$$

Here, ΔG_{MM} is the molecular mechanics contribution (electrostatic interactions in nature) to the free energy of the protein titratable site. Hence, by using a reference compound, the QM effects are not needed. Effectively, we compute ΔpK_a relative to the reference compound. Computing ΔpK_a can also help canceling some error introduced by GB solvation model through the use of $\Delta G_{ref,MM}$. In AMBER, a reference compound is a blocked dipeptide amino acid possessing titratable side chain (for example, acetyl-Asp-methylamine). Five reference compounds were constructed corresponding to five titratable residues. The values of $\Delta G_{ref,MM}$ for each reference compound are obtained from thermodynamic integration calculations at 300 K and set as internal parameters in AMBER. The ΔG_{MM} is calculated by taking the difference between the potential energy

with the charges of the current protonation state and the potential energy with the charges of the new protonation state. If the transition is accepted, MD steps are carried out to sample conformational space in the new protonation state. If the MC attempt is rejected, MD steps are also carried out with no change to the protonation state.

2.6 Advanced Sampling Methods

Conformational sampling in a MD or MC simulation is essential in the study of complex systems such as polymers and proteins. One major concern is that the PES of a complex system is very rugged and contains a lot of local energy minima. Thus, kinetic trapping would occur as a result of the low rate of potential energy barrier crossing, especially when the barrier is high. To overcome this kinetic trapping behavior, generalized ensemble methods can be employed in molecular simulations. As its name implies, a generalized ensemble method differ from the canonical ensemble method in the weight factor of a state. The weight factor in the canonical ensemble is Boltzmann weight. However, a non-Boltzmann weight factor can be used in a generalized ensemble method (This does not mean that Boltzmann factor is prohibited in a generalized ensemble method. In fact, parallel tempering which belong to the family of generalized ensemble method, does adopt Boltzmann factor.). By choosing a non-Boltzmann weight factor, the system is able to perform a random walk in the potential energy space. Thus, potential energy barriers will be overcome easily and more conformations will be visited. Frequently utilized generalized ensemble algorithms include the multicanonical (MUCA) method and replica exchange molecular dynamics (REMD) method. In this section, the MUCA and parallel tempering will be introduced briefly. Due to the importance of REMD method to this dissertation, the details of REMD method will be explained in the next section.

2.6.1 The Multicanonical Algorithm (MUCA)

In canonical ensemble, the probability of visiting a state in the energy space is:

$$P_{canonical}\ (T, E) \propto n(E)e^{-E/k_B T} \tag{2-75}$$

Here, $n(E)$ is the density of states (DOS), which means the number of states between E and $E + dE$. $e^{-E/k_B T}$ is the Boltzmann factor. As potential energy increases, the Boltzmann factor decreases but the DOS increases rapidly. A bell-shaped probability distribution function (PDF) of E can be observed. However, in the MUCA method,[54,55,137] the PDF is designed to be flat (a constant), although it still can be written in the form of Eq. 2-76:

$$P_{MUCA}(E) \propto n(E)w_{MUCA}(E) = c \tag{2-76}$$

where $w_{MUCA}(E)$ is the multicanonical weight factor and $n(E)$ is DOS. The multicanonical weight factor needs to be inversely proportional to the DOS in order to generate a flat PDF. However, the DOS of a system is in general unknown, which makes the multicanonical weight a-priori unknown. Generating correct distribution of $n(E)$ is the central task of a MUCA simulation. In practice, short simulations are performed in order to determine the DOS in an iterative manner. Details of determining the DOS can be found in the paper of Okamoto and Hansmann published in 1995.[162] After the DOS is resolved, the canonical ensemble PDF will be ontained. Thus, the average of any quantity can be determined by Eq. 1-11 or Eq. 1-12, depending on either MD or MC simulation is performed.

Another way to explore the DOS is by using the Wang-Landau algorithm.[163,164] In the Wang-Landau algorithm, the DOS is recorded by a histogram $g(U)$ and initially set to unity for all its elements. Another histogram which is called visit histogram is also

constructed with initial values set to zero. The visit histogram represents the number of visits to each energy level. Monte Carlo moves are made. Instead of being evaluated by the Metropolis criterion, they are evaluated by the DOS,

$$w(i \to j) = min\left\{1, \frac{g(U_i)}{g(U_j)}\right\} \tag{2-77}$$

where $w(i \to j)$ is the transition probability from state i to state j. Each time an energy level is visited, the corresponding element of the DOS histogram is updated by multiplying the current value with a modification coefficient that is greater than 1. The initial value of the modification coefficient is $f_0 = e \approx 2.71828$. Every time a MC move is performed, the corresponding element of the visit histogram is also updated. The MC moves will continue until the visit histogram is flat. At this stage, the DOS are converged. In order to achieve a finer convergence, a second round of the above process will be performed. This time, the modification coefficient f_1 in the second round is given by $f_1 = \sqrt{f_0}$. The visit histogram is then reset to zero. This process will iterate until a modification coefficient that is approximately 1 is achieved (in the paper of Wang and Landau, the final value of the modification coefficient is 1.00000001). By utilizing Wang-Landau algorithm, the DOS will be obtained and a random walk in the potential energy space will be achieved.

2.6.2 Parallel Tempering

In 1986, Swendsen and Wang firstly performed parallel tempering (replica exchange MC) simulations to investigate spin glass.[59] Multiple non-interacting copies (replicas) of the system are simulated at different temperatures. At each temperature, MC simulation is conducted to sample the conformational space. Structures or temperatures of the two replicas are attempted to be exchanged periodically. The

detailed balance condition is applied and the weight factor of the state is the Boltzmann weight factor. The Metropolis criterion has been utilized to accept or reject the move. Hansmann *et al*.[58] first utilized the parallel tempering algorithm in the study of a biomolecule (7-residue Ket-enkephalin). Other applications of the parallel tempering algorithm include X-ray structures determination performed by Falcioni and Deem.[165]

A MC simulation at a high temperature accepts the transition attempts more often than doing that at a low temperature. Thus, the simulation at high temperatures tends to visit more conformations in conformational space. Exchanging structures with replicas at lower temperatures can help them avoid getting trapped in the conformational space.

The acceptance ratio, which is the averaged fraction of successful exchange attempts, is an important issue in the parallel tempering method. It is correlated with temperature distribution of replicas. According to Kofke,[166] the acceptance ratio is the area of overlap between the potential energy PDF at two temperatures. Given the number of replicas, if the temperatures of the two replicas are too different, the overlap between the two potential energy PDFs will be small. Therefore, accepting an exchange attempt is unlikely, which makes parallel tempering simulation inefficient. However, if the temperatures of the two adjacent replicas are too close, the overlap between two PDFs will be large, and hence the acceptance ratio will be large. But the conformational space sampled by two adjacent replicas will be too close. More replicas than actually needed are utilized to achieve the same goal, and hence computer resource is wasted.

2.7 Replica Exchange Molecular Dynamics (REMD) Methods

Due to the correlation between conformation and protonation sampling, correct sampling of protonation states requires accurate sampling of protein conformations. Hence, generalized ensemble methods such as multicanonical algorithm and REMD

should be used to avoid kinetic trapping which comes from low rates of barrier crossing in constant temperature MD simulations. REMD has been applied to the continuous protonation state constant-pH method (REX-CPHMD) by Khandogin *et al*.[110-113] They have performed REX-CPHMD simulations to predict pK_a values[110] and to explore pH-dependent protein dynamics.[111-113]

The REMD, which is the MD version of parallel tempering, have been developed by Sugita and Okamoto in 1999.[62] The theory of REMD is essentially the same as parallel tempering. In their method, temperatures are attempted to be exchanged. This leads to the unique part of REMD: the treatment of velocities after accepting an exchange attempt, because the velocities must reflect the temperature correctly. Sugita and Okamoto proposed to rescale the velocities in order to recover the desired temperature when temperatures are swapped. Similar to other generalized-ensemble methods, REMD algorithm wants to make the system perform a random walk in either temperature or potential energy space, and hence avoid kinetic trapping. The advantage of REMD over other generalized-ensemble method is that the weight factor is Boltzmann weight which is a-priori known. This advantage makes REMD very frequently employed in the MD simulations of complex systems. The REMD algorithm has been applied to studies of peptides, proteins, protein-membrane system in order to describe free energy landscape, amyloid formation, structure prediction and binding. Many extended versions such as solute-tempering REMD[167] and structure-reservoir REMD[168-170] have been proposed to improve the performance of REMD algorithm. The REMD variants will be briefly explained later in this section.

2.7.1 Temperature REMD (T-REMD)

A thorough description of the T-REMD algorithm can be found in the original paper of Sugita and Okamoto.[62] In T-REMD, *N* non-interacting copies (replicas) of a system are simulated at *N* different temperatures (one each). Regular MD is performed and periodically an exchange of configurations between two (usually adjacent) temperatures is attempted. Suppose replica *i* at temperature T_m and replica *j* at temperature T_n are attempting to exchange; the following satisfies the detailed balance condition:

$$P_m(i)P_n(j)w(i \to j) = P_n(i)P_m(j)w(j \to i) \tag{2-78}$$

Here $w(i \to j)$ is the transition probability between two states *i* and *j* and $P_m(i)$ is the population of state *i* at temperature *m* (in REMD assumed Boltzmann weighted).

Since,

$$P_m(i) \propto e^{-H(p_i,q_i)/k_B T_m} \tag{2-79}$$

where H is the Hamiltonian of the state, q represents the molecular structure, and p stands for momentum. The Hamiltonian consists of kinetic energy (*K*) and potential energy (*U*) terms and can be written as:

$$H(p,q) = K(p) + U(q) \tag{2-80}$$

In the original derivation of exchange probability, Sugita and Okamoto mentioned that exchanging two replicas (states) is equivalent to exchanging temperatures. The momenta of each replica after exchange attempt need to be rescaled:

$$p_n(i) = \sqrt{T_n/T_m}\,p_m(i) \tag{2-81}$$

$$p_m(j) = \sqrt{T_m/T_n}\,p_n(j) \tag{2-82}$$

After inserting Eq. 2-79 and Eq. 2-80 into Eq. 2-78, the detailed balance equation becomes:

$$exp\{-[K(p_m(i)) + U(q_i)]/k_B T_m\} \cdot exp\{-[K(p_n(j)) + U(q_j)]/k_B T_n\} \cdot w(i \to j) =$$

$$exp\{-[K(p_n(i)) + U(q_i)]/k_B T_n\} \cdot exp\{-[K(p_m(j)) + U(q_j)]/k_B T_m\} \cdot w(j \to i) \quad (2\text{-}83)$$

According to Eq. 2-81 and Eq. 2-82,

$$K(p_m(i)) = (T_m/T_n)K(p_n(i)) \quad (2\text{-}84)$$

$$K(p_n(j)) = (T_n/T_m)K(p_m(j)) \quad (2\text{-}85)$$

Therefore, kinetic energy contributions on both sides of Eq. 2-83 will be canceled out, leaving only potential energy terms contribute to exchange probability.

$$\frac{w(i \to j)}{w(j \to i)} = \frac{exp\,[-U(q_i)/k_B T_n] \cdot exp\,[-U(q_j)/k_B T_m]}{exp\,[-U(q_i)/k_B T_m] \cdot exp\,[-U(q_j)/k_B T_n]} \quad (2\text{-}86)$$

Further manipulation of Eq. 2-86 yields:

$$\frac{w(i \to j)}{w(j \to i)} = exp\left[\left(\frac{1}{k_B T_m} - \frac{1}{k_B T_n}\right)\left(U(q_i) - U(q_j)\right)\right]$$

If the Metropolis criterion is applied, the exchange probability is obtained as:

$$w(i \to j) = min\left\{1, exp\left[\left(\frac{1}{k_B T_m} - \frac{1}{k_B T_n}\right)\left(U(q_i) - U(q_j)\right)\right]\right\} \quad (2\text{-}87)$$

If the exchange attempt between two replicas is accepted, the temperatures of the two replicas will be swapped and velocities rescaled to the new temperatures by multiplying all the old velocities by the square root of the new temperature to old temperature ratio:

$$v_{new} = v_{old}\sqrt{\frac{T_{new}}{T_{old}}} \quad (2\text{-}88)$$

Here, v_{new} and v_{old} are the new and old velocities, respectively. T_{new} and T_{old} are the temperatures after and before an exchange is accepted, respectively.

The acceptance ratio is the average value of the exchange probabilities between two temperatures:

$$P_{acc} = \langle min\left\{1, exp\left[\left(\frac{1}{k_B T_m} - \frac{1}{k_B T_n}\right)\left(U(q_i) - U(q_j)\right)\right]\right\}\rangle \quad (2\text{-}89)$$

For a given system, the potential energy function is independent of temperature but the potential energy PDF in a canonical ensemble depends on temperature. The potential energy PDF can be considered as a Gaussian function (to the second order truncation of the Taylor expansion of the PDF at the potential energy value corresponding to maximum probability). The Gaussian is centered at mean potential energy of the system with a variance $\sigma^2 = k_B T^2 C_v$, where C_v is the heat capacity. At this stage, the Gaussian function expression of the potential energy PDF is not adopted. It will be employed later in this section. The potential energy PDF at temperature T_m is currently written as:

$$P_m(U) = \frac{1}{Q_m} n(U) exp(-U/k_B T) \quad (2\text{-}90)$$

where $n(U)$ is the DOS and the exponential term is the Boltzmann weight factor as a function of potential energy. Recall that in the probability theory, the average quantity can be expressed as:

$$\langle A \rangle = \int P(A) \cdot A \cdot dA \quad (2\text{-}91)$$

Extend Eq. 2-91 to the bivariate case and notice that the two PDFs are independent. The acceptance ratio can be rewritten as,

$$P_{acc} = \iint_{-\infty}^{+\infty} P_m(U) \cdot P_n(U') \cdot min\left\{1, exp\left[\left(\frac{1}{k_B T_m} - \frac{1}{k_B T_n}\right)(U - U')\right]\right\} dU dU' \quad (2\text{-}92)$$

Let a function $g(U, U')$ to denote $min\left\{1, exp\left[\left(\frac{1}{k_B T_m} - \frac{1}{k_B T_n}\right)(U - U')\right]\right\}$, $\beta_m = 1/k_B T_m$ and $\beta_n = 1/k_B T_n$, then,

$$g(U, U') = min\{1, exp[(\beta_m - \beta_n)(U - U')]\} \quad (2\text{-}93)$$

Without loss of generality, we can assume that $\beta_m > \beta_n$, which means $T_m < T_n$. Therefore, another way of writing Eq. 2-93 is $g(U, U') = 1$ when $U > U'$ and $g(U, U') = exp[(\beta_m - \beta_n)(U - U')]$ when $U < U'$. Inserting $g(U, U')$ into Eq. 2-92 will lead to:

$$P_{acc} = \int_{-\infty}^{\infty} P_m(U)dU \int_{-\infty}^{U} 1 \cdot P_n(U')dU' + \int_{-\infty}^{\infty} P_m(U)dU \int_{U}^{\infty} exp[(\beta_m - \beta_n)(U - U')] \cdot P_n(U')dU' \quad (2\text{-}94)$$

For simplicity, we denote $\int_{-\infty}^{\infty} P_m(U)dU \int_{U}^{\infty} exp[(\beta_m - \beta_n)(U - U')] \cdot P_n(U')dU'$ as $h(U, U')$. Inserting Eq. 2-90 into $h(U, U')$,

$$h(U, U') = \int_{-\infty}^{\infty} \frac{1}{Q_m} n(U) e^{-\beta_m U} dU \int_{U}^{\infty} e^{\beta_m U} \cdot e^{-\beta_m U'} \cdot e^{-\beta_n U} \cdot e^{\beta_n U'} \cdot \frac{1}{Q_n} n(U') e^{-\beta_n U'} dU' \quad (2\text{-}95)$$

Since U and U' are independent, Eq. 2-95 can be rewritten as:

$$h(U, U') = \int_{-\infty}^{\infty} \frac{1}{Q_m} n(U) e^{-\beta_m U} \cdot e^{\beta_m U} \cdot e^{-\beta_n U} dU \int_{U}^{\infty} \frac{1}{Q_n} n(U') e^{-\beta_n U'} \cdot e^{-\beta_m U'} \cdot e^{\beta_n U'} dU' \quad (2\text{-}96)$$

Simplifying Eq. 2-96 will formulate $h(U, U')$ as:

$$h(U, U') = \int_{-\infty}^{\infty} \frac{1}{Q_m} n(U)\, e^{-\beta_n U} dU \int_{U}^{\infty} \frac{1}{Q_n} n(U')\, e^{-\beta_m U'} dU' \quad (2\text{-}97)$$

Recall that a partition function is just a normalizing constant. Q_m and Q_n in Eq. 2-97 can switch their positions in the integrand. Thus, Eq. 2-97 becomes:

$$h(U, U') = \int_{-\infty}^{\infty} P_n(U)dU \int_{U}^{\infty} P_m(U')dU' \quad (2\text{-}98)$$

Inserting Eq. 2-98 into Eq. 2-94,

$$P_{acc} = \int_{-\infty}^{\infty} P_m(U)dU \int_{-\infty}^{U} P_n(U')dU' + \int_{-\infty}^{\infty} P_n(U)dU \int_{U}^{\infty} P_m(U')dU' \quad (2\text{-}99)$$

Each term on the right-hand side of Eq. 2-99 can be interpreted as an overlap between two PDFs. The sum is the entire overlap between two PDFs. Therefore, the

average exchange probability is just the overlap between potential energy PDFs at two temperatures.

Next, let us consider the temperature distribution in the simplest case, in which the heat capacity is a constant. As mentioned earlier, a potential energy PDF of a canonical ensemble can be written as a Gaussian function,

$$P(U) = P(\overline{U})exp\left\{\frac{-(U-\overline{U})^2}{2k_BT^2C_v}\right\} \quad (2\text{-}100)$$

where $\overline{U}$ is the average potential energy, $P(\overline{U})$ is the probability density of finding $\overline{U}$ at temperature T, and C_v is the heat capacity. Since the PDF should be normalized, it is easy to find the relationship between $P(\overline{U})$ and the standard deviation of the Gaussian function:

$$P(\overline{U}) = 1/\sqrt{2\pi k_BT^2C_v} \quad (2\text{-}101)$$

For simplicity in the derivation of the acceptance ratio, the Gaussian PDF at temperature T_i will be written as Eq. 2-102 from now on:

$$P(U) = \frac{1}{\sqrt{2\pi\sigma_i{}^2}}exp\left\{\frac{-(U-\overline{U}_i)^2}{2\sigma_i{}^2}\right\} \quad (2\text{-}102)$$

Recall that one assumption to distribute temperatures is to maintain a random walk in temperature space. Hence, a constant acceptance ratio should be achieved for any two adjacent temperatures. As shown previously, the acceptance ratio is the overlap between two potential energy PDFs. Consider two potential energy PDFs at temperatures $T_m < T_n$. The PDF at T_m will be to the left of the PDF at T_n. After finding the potential energy $U_{intersect}$ where the two Gaussian PDFs intersect, the overlap between two PDFs can be computed by integrating the left Gaussian PDF from

$U_{intersect}$ to infinity and the Gaussian on the right from minus infinity to $U_{intersect}$ and adding them up,

$$P_{acc} = \int_{U_{intersect}}^{\infty} \frac{1}{\sqrt{2\pi\sigma_m{}^2}} exp\left\{\frac{-(U-\bar{U}_m)^2}{2\sigma_m{}^2}\right\} dU + \int_{-\infty}^{U_{intersect}} \frac{1}{\sqrt{2\pi\sigma_n{}^2}} exp\left\{\frac{-(U-\bar{U}_n)^2}{2\sigma_n{}^2}\right\} dU \quad (2\text{-}103)$$

Complementary error functions will be utilized and Eq. 2-103 will become,

$$P_{acc} = \frac{1}{2} erfc\left(\frac{U_{intersect} - \bar{U}_m}{\sqrt{2}\sigma_m}\right) + \frac{1}{2} erfc\left(\frac{\bar{U}_n - U_{intersect}}{\sqrt{2}\sigma_n}\right) \quad (2\text{-}104)$$

According to Rathore *et al*.,[171] the acceptance ratio can be approximate to:

$$P_{acc} \cong erfc\left(\frac{\bar{U}_n - \bar{U}_m}{2\sqrt{2}\sigma}\right) \quad (2\text{-}105)$$

where $\sigma = (\sigma_m + \sigma_n)/2$.

For a geometric distribution of temperatures where $T_n = cT_m$, $\sigma_m + \sigma_n = \sqrt{k_B C_v} T_m (c+1)$. The average potential energy difference can be computed as,

$$(\bar{U}_n - \bar{U}_m) \approx C_v \Delta T = C_v(T_n - T_m) = C_v(c-1)T_m \quad (2\text{-}106)$$

Thus, if the heat capacity does not change with temperature, the temperature term in the numerator and denominator in Eq. 2-105 will be canceled, which means the acceptance ratio will be a constant. Furthermore, Eq. 2-105 also signals the number of replicas needed to cover a temperature range as a function of system size. In order to have a non-zero P_{acc}, $(\bar{U}_n - \bar{U}_m)/\sigma \approx 1$. This leads to, $C_v \Delta T / \left(\sqrt{k_B C_v} T_m (c+1)\right) \approx 1$. Further simplifications lead to:

$$\Delta T_m \approx \sqrt{\frac{k_B}{C_v}} T_m \quad (2\text{-}107)$$

Since the heat capacity is $O(N)$, where N is the number of particles, the number of replicas to cover a temperature range is $O\left(N^{1/2}\right)$.

2.7.2 Hamiltonian REMD (H-REMD)

Instead of preparing replicas with different temperatures, another way to overcome potential energy barriers is simply changing the PES to reduce potential energy barriers.[61] And this is the basic idea of H-REMD. In H-REMD algorithm, replicas differ in their Hamiltonians but have the same temperature. Likewise, regular MD is performed and an exchange of configurations between two neighboring replicas is attempted periodically. Let us consider replica *i* with Hamiltonian H_n and replica *j* with Hamiltonian H_m are attempting to exchange. By employing the detailed balance equation (Eq. 2-78) and Boltzmann weight of a molecular structure, the transition probability can be written as:

$$w(i \to j) = min\{1, exp[-\beta(H_n(i) + H_m(j) - H_m(i) - H_n(j))]\} \quad (2\text{-}108)$$

2.7.3 Technical Details in REMD Simulations

Temperature distributions have been explored in order to optimize the performance of REMD method. For systems having constant heat capacity, a geometrical distribution of temperatures has been adopted. Sugita and Okamoto,[62] and Kofke[166] believed that the most efficient way to exploit REMD algorithm is letting each replica spend the same amount of simulation time at each temperature (a random walk in temperature space). In practice, this is achieved by producing the same acceptance ratio for each replica, given that each replica only attempts to exchange with its neighbors in temperature space. Under the condition that the system has a constant heat capacity, a geometrical distribution of temperatures ($T_j/T_i = c$) is achieved. Sanbonmatsu and Garcia suggested an iterative method to distribute temperatures for replicas in 2002.[172] They have chosen the averaged values of potential energy as a function of temperature to maintain a random walk in the temperature space. In 2005,

Rathore *et al*.[171] suggested that an acceptance ratio of 0.2 yields the best performance, based on constant heat capacity assumption. They have chosen Go-type model of protein A and the Lennard-Jones liquid to study the deviation of heat capacity relative to the final value as a function of acceptance ratio. A minimum of deviation at acceptance ratio around 0.2 has been observed. Kone and Kofke[173] have performed similar study for the parallel tempering simulations. They also considered a random-walk model in temperature space through replica exchange moves. The acceptance ratio is given by:

$$\bar{P}_{acc} = erfc\left(\frac{1-B}{1+B}C^{1/2}\right) \quad (2\text{-}109)$$

where $B = \beta_1/\beta_0$, β_i is the Boltzmann weight factor, and C is the heat capacity which is assumed to be constant in their study. Without loss of generality, β_0 is greater than β_1. The mean-square displacement of this random-walk (Eq. 2-110) has been maximized with respect to acceptance ratio. The maximum is shown near an acceptance ratio of 20%.

$$\sigma^2 \propto (\ln B)^2 \bar{P}_{acc}(B) \quad (2\text{-}110)$$

where σ^2 is the mean-square displacement, B and $\bar{P}_{acc}$ are shown in Eq. 2-109.

Temperature distributions in parallel tempering simulation of villin headpiece subdomain HP-36 have been investigated by Trebst *et al*.[174] HP-36 will undergo helix-coil transition at high temperatures and hence, the heat capacity will not be held constant. The diffusion of a replica in temperature space has been introduced to judge the performance. In their method, a replica is labeled “up” when its previous visit of the extreme temperature is the highest temperature; it is labeled “down” when its previous visit of the extreme temperature is the lowest. For each temperature T_i, two histograms $n_{up}(T_i)$ and $n_{down}(T_i)$ are recorded. The two histograms keep the record of the number

of visits from replicas with label “up” and “down”, respectively. The average fraction of replicas traveling from the lowest to highest temperature can be calculated as:

$$f(T_i) = \frac{n_{up}\,(T_i)}{n_{up}\,(T_i)+n_{down}\,(T_i)} \tag{2-111}$$

The diffusivity $D(T)$ is adopted and has the form:

$$D(T) \propto \frac{\Delta T}{df/dT} \tag{2-112}$$

They have pointed out that the diffusivity is temperature dependent, a minimum of diffusivity has been observed around the temperature where heat capacity is at maximum. The plot showing diffusivity vs temperature indicates that random walk is suppressed the most when phase transition occurs. The numbers of round-trip between temperature extremes of each replica has been maximized to generate an optimal temperature distribution. More recently, Nadler and Hansmann[175-177] suggested that the optimal number of replicas between the lowest and highest temperatures in explicit-solvent simulation has the following formula: $N_{optimal} = 1 + 0.594\sqrt{C}\ln(T_{max}/T_{min})$, where the C is the heat capacity, and T_{max} and T_{min} is the highest and lowest temperature, respectively. They also proposed that the optimal temperature distribution can be formulated as: $T_{optimal}\,(i) = T_{min}\left(\frac{T_{max}}{T_{min}}\right)^{\frac{i-1}{N-1}}$.

In addition to replica temperature distribution, exchange attempt frequency (EAF) is also an important issue in parallel tempering and REMD sampling efficiency. In 2001, Opps and Schofield[178] investigated the effect of EAF for parallel tempering. Two-dimensional spin system and a polypeptide in vacuum have been selected to test the effect of EAF on the properties such as order parameter and radius of gyration of the polypeptide. They suggested that the most efficient scheme is to attempt after a few MC

steps. The situation is more complicated in the case of REMD. In general, thermostats are used in MD propagations to maintain a canonical ensemble is satisfied. It is argued that exchanges in REMD should happen when system temperature stabilizes.[179] Attempting to exchange frequently may prevent the system from heat dissipation. This argument was supported by studies of a peptide Fs21 performed by Zhang *et al*.[179] They have suggested that 1 ps of exchange attempt interval is desirable for REMD. However, Sindhikara *et al*.[180] have later shown that small exchange attempt interval (even as small as a few MD steps) does not affect heat dissipation, given that REMD exchange is done properly. Conformational sampling deviation relative to long simulation time reference calculation as a function of EAF has been investigated. They have pointed out that large EAF (small exchange attempt time interval) is preferred. Abraham and Gready[181] studied the effect of EAF based on a 23-residue peptide in explicit water. By examining the potential energy autocorrelation time, they argued that an exchange period below 1 ps is too short for replica exchange attempts to be independent, and hence reduce the tempering efficiency. However, their conclusion was not supported by an investigation of tempering efficiency performed by Zhang and Ma.[182] Zhang and Ma utilized the transition matrix and its correlation functions. The autocorrelation function of transition probability can be written as a function of eigenvalues of transition matrix. The decay time has been explored in order to understand the tempering efficiency. Zhang and Ma found that tempering efficiency increases monotonically as EAF increases.

Thermostat effects on the performance of REMD have also been explored. Earlier work has been done by the Garcia group.[172] They have studied if the potential energy

PDFs satisfy the Boltzmann distribution: $\ln[P(U,T_1)/P(U,T_2)] = \left(\frac{1}{k_B T_2} - \frac{1}{k_B T_1}\right) \cdot U + c$, where $P(U,T)$ is the potential energy PDF at temperature T and c is a constant. They have found that Nose-Hoover and the Anderson thermostats satisfy the above condition, while the Berendsen thermostat does not. Rosta *et al*.[183] investigated the thermostat artifact in the REMD simulations in 2009. The current REMD exchange scheme assumes Boltzmann distribution (canonical ensemble) in the calculation of exchange probability. However, the Berendsen thermostat cannot preserve the Boltzmann distribution. Thus REMD simulations of bulk water and protein folding are performed and the temperature is controlled by Berendsen thermostat and Langevin dynamics. They have studied the potential energy PDFs and thermal unfolding under the two thermostats. The Berendsen thermostat has been shown to produce a shift average potential energy and prolonged tails for potential energy PDF for bulk water, while no such effect has been seen when Langevin dynamics is employed. An increased probability of folding at low temperatures has been reported by Berendsen thermostat, whereas the probability of folding is decreased at high temperatures. The authors proposed that REMD simulations performed with thermostats that can generate a Boltzmann distribution, such as Langevin dynamics, Andersen and Nose-Hoover thermostats.

In a REMD simulation, the number of replicas needed to cover a temperature range scales as $O\left(f^{1/2}\right)$, where f is the degree of freedom of the system. Given a large system, the number of replicas needed is large. For example, 64 replicas have been used in a REMD study of β-hairpin surrounded by explicit water molecules (4342 atoms in each replica) to cover the temperature range from 270 K to 695 K.[184] A number of

methods have been developed to reduce the number of replicas needed in REMD simulations. In 2002, Fukunishi *et al*.[61] proposed Hamiltonian-REMD (H-REMD). In the H-REMD scheme, replicas differ in their Hamiltonians but have the same temperature. The exchange strategies in the paper of Fukunishi were to scale hydrophobic interactions and to scale van der Waals interactions. In 2005, Liu *et al*.[167] published a method with the name replica exchange with solute tempering. In the replica exchange with solute tempering algorithm, the protein-water interactions and water-water interactions are scaled such that the exchange probability does not depend on the number of explicit water molecules. The number of replicas in replica exchange with solute tempering simulation to cover the same temperature range is significantly reduced when comparing with original REMD algorithm. Lyman *et al*.,[185] and Liu and Voth later,[186,187] have developed resolution exchange schemes to improve the performance of REMD. Coarse-grained models (low resolution) are employed to replace the role of high-temperature replicas. The Simmerling group has contributed the hybrid explicit/implicit solvation model[188] in order to reduce the number of replicas needed in REMD simulations with explicit water molecules. Each replica is propagated in an explicit water box. At an exchange attempt, the solute and its solvation shell, which is calculated on-the-fly, are placed in dielectric continuum. Exchange probabilities are calculated based on the potential energies of the solute and the hybrid solvent. The usage of a hybrid solvent can shrink the number of replicas from 40 to 8, in a test case of polypeptide Ala_{10} simulated at temperatures from 267 K to 571 K. Structural reservoir techniques[168-170] have also been incorporated into REMD algorithm. High temperature MD simulations are performed first to generate a structural reservoir. Structures in the

reservoir will be brought to replicas via exchanges. One advantage of using structural reservoir is that non-Boltzmann weight factors can be chosen in the calculation of exchange probabilities.[170] Recently, Ballard and Jarzynski[189] proposed to use non-equilibrium work simulations to accept exchange attempts. Kamberaj and van der Vaart[190] developed a new scheme to perform exchanges, in which the generalized canonical PDF have been employed to achieve a flat potential of the mean force in temperature space. The Wang-Landau algorithm[163,164] has been adopted in order to estimate the DOS in temperature space and the round-up time between extreme temperatures has been minimized. More recently, solvent viscosity has been selected as a parameter in addition to temperature for REMD method.[191] This method is named V-REMD and it is essentially a two-dimensional REMD method. The motivation of choosing viscosity as a parameter is that the lower the viscosity, the faster a protein will diffuse, and sample the conformational space. In this algorithm, one replica is selected to have normal viscosity, others use reduced viscosities. The mass of solvent molecules is scaled by a factor of λ^2 when the viscosity is scaled by a factor of λ. Changing the mass of solvent molecules does not affect the potential energy at an exchange attempt. Thus, the exchange probability of the V-REMD is the same as conventional T-REMD. The author applied V-REMD to the study of trialanine, deca-alanine, and a 16-residue β-hairpin peptide. By using the V-REMD, replica numbers are reduced by a factor of 1.5 to 2.

The replica exchange method (REM) can be coupled with other generalized-ensemble methods in order to enhance conformational sampling. The Okamoto group have coupled REM with MUCA and simulated tempering. The two new schemes are

called multicanonical replica exchange method,[192] and replica exchange simulated tempering,[193,194] respectively. The details of coupled REM and generalized-ensemble methods can be found in a review by Mitsutake *et al*.[53]

Due to its stochastic nature, the REMD algorithm has been employed to investigate thermodynamics rather than kinetics.[195] However, a properly designed scheme of analyzing the REMD trajectory in phase space can yield information about kinetics. In 2005, Levy and his coworkers[195] designed a kinetic network and used master equation to solve for the transition rate from REMD simulations. The structures at all temperatures are grouped into states based on their structural similarity (they selected a 42 dimensional Euclidean distance space based on C_α-C_α distances, instead of clustering, to group their structures). A state is denoted as a node and an edge stands for a transition between two nodes. A total of 800,000 nodes and 7.347×10^9 edges were obtained. The master equation has been utilized to describe the transitions between two states. Since they discretized the conformational space into states, the master equation is written in a matrix notation, $\frac{d\vec{P}(t)}{dt} = K\vec{P}(t)$, where K is the transition matrix and $\vec{P}(t)$ is probability distribution of states at time t. Instead of solving for eigenvalues of the transition matrix or solving the differential equation numerically, the authors actually simulated the path satisfying the master equation. Likewise, this Markov state model has been employed in the study of protein folding too.

In 2006, van der Spoel and Seibert[196] studied protein folding rate based on Arrhenius equation. The folding mechanism in their investigation has been assumed to be two-state. A binary folding indicator, which is the RMSD relative to the native state, has been adopted by the authors. Hence, the first-order reaction rate equation has been

set up. Then, the rate equation was integrated and averaged over all trajectories in order to generate an derived fraction of folded structures. A fitting parameter χ^2, which is equal to the difference between derived and actual fraction of folded structures, was minimized numerically with respective to energy barriers and pre-exponential factors. In this manner, the Arrhenius reaction rate will be resolved from REMD simulations.

Yang *et al*.[197] proposed to use diffusion equation to extract kinetics from REMD simulation in 2007. The Fokker-Planck equation has been employed to extract local drift velocity and diffusion coefficient from REMD simulations. Langevin dynamics on the reaction coordinate is performed using drift velocity and diffusion coefficient. The free energy landscape will be reconstructed based on drift velocity and diffusion coefficient.

In 2008, Buchete and Hummer[198] demonstrated that both local conformational transition rate as well as globally folding rates can be accurately extracted from REMD simulations, without any assumption in temperature dependence of the kinetics (Arrhenius and non-Arrhenius). Similar to Levy and coworkers, Buchete and Hummer have also adopted the master equation operating on discretized space to describe transitions. Conditional probability of state j at time t, given the initial state i, was computed by the master equation. The likelihood of seeing N_{ji} number of transitions in a time interval has been maximized with respective to the natural log of transition rate constant (transition matrix elements) and the natural log of equilibrium population of state i. Thus, the rate constants will be generated. A detailed description can be found in the paper of Buchete and Hummer.

CHAPTER 3
CONSTANT-pH REMD: METHOD AND IMPLEMENTATION*

3.1 Introduction

In this chapter, the constant-pH REMD algorithm used in the AMBER simulation suite is presented and is employed to study model systems. We first tested our method based on five dipeptides and a model peptide having the sequence Ala-Asp-Phe-Asp-Ala (ADFDA). The two ends of model peptide ADFDA were not capped so the two ionizable side chains would have different electrostatic environment. The pK_a values of the two Asp residues are expected to be different due to the difference in electrostatic environment.

Then our constant-pH REMD method is applied to a heptapeptide derived from OMTKY3, the same heptapeptide as Dlugosz and Antosiewicz studied in their paper. NMR experiments indicated the pK_a of Asp is 3.6,[122] 0.4 pK_a unit lower than the value of blocked Asp dipeptide. Dlugosz and Antosiewicz performed constant-pH MD simulations and their method predicted the pK_a to be 4.24.[122] Our purpose is to show that the REMD algorithm coupled with a discrete protonation state description can greatly improve pH-dependent protein conformation and protonation state sampling.

3.2 Theory and Methods

3.2.1 Constant-pH REMD Algorithm in AMBER Simulation Suite

In the case of constant pH molecular dynamics, the potential energy of the system depends not only on the protein structure but also on the protein protonation state.

---

* Reproduced in part with permission from Meng, Y.; Roitberg, A.E. Constant pH Replica Exchange Molecular Dynamics in Biomolecules Using a Discrete Protonation Model, J. Chem. Theory. Comput. 2010, 6, 1401-1412. Copyright 2010 American Chemical Society.

Likewise, when coupling REMD algorithm with constant-pH MD, one can either attempt to exchange molecular structures only or swap both structures and protonation states at the same time. For simplicity, let us consider two replicas where replica 0 has temperature T_0, protein structure q_0 and protonation state n_0, while replica 1 has temperature T_1, structure q_1 and protonation state n_1. A diagrammatic description of the two exchange algorithms is shown in Figure 3-1.

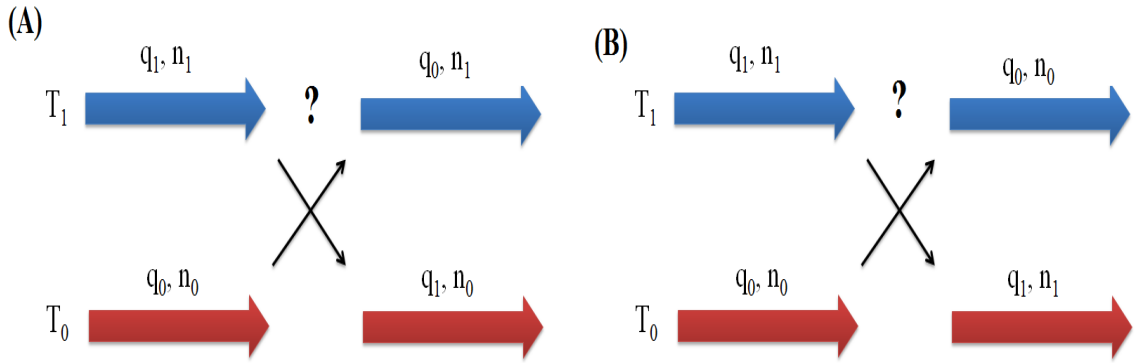


Figure 3-1. Methods to perform exchange attempts. A) Only molecular structures are attempted to exchange. The protonation states are kept the same. B) Both molecular structures and protonation states are attempted to exchange.

The first way of performing an exchange attempt is that replica 0 tries to jump from state (q_0, n_0) to state (q_1, n_0) at temperature T_0 in one Monte Carlo step. Similarly, replica 1 attempts to transit from state (q_1, n_1) to state (q_0, n_1) at temperature T_1. Protonation states are kept at exchange attempts and only change during dynamics. Therefore, the detailed balance equation now becomes:

$$\frac{w(\beta_0 q_0 n_0, \beta_1 q_1 n_1 \rightarrow \beta_0 q_1 n_0, \beta_1 q_0 n_1)}{w(\beta_0 q_1 n_0, \beta_1 q_0 n_1 \rightarrow \beta_0 q_0 n_0, \beta_1 q_1 n_1)} = \frac{\exp(-\beta_0 E(q_0, n_0)) \cdot \exp(-\beta_1 E(q_1, n_1))}{\exp(-\beta_0 E(q_1, n_0)) \cdot \exp(-\beta_1 E(q_0, n_1))} \quad (3\text{-}1)$$

Here $w(\beta_0 q_0 n_0, \beta_1 q_1 n_1 \rightarrow \beta_0 q_1 n_0, \beta_1 q_0 n_1)$ is the transition probability of swapping structures. If Metropolis criterion is used, this exchange probability can be written as:

$$w(\beta_0 q_0 n_0, \beta_1 q_1 n_1 \rightarrow \beta_0 q_1 n_0, \beta_1 q_0 n_1) = min\{1, exp(-\Delta)\} \quad (3\text{-}2)$$

In Eq. 3-2, Δ has the form:

$$\Delta = \beta_0[E(q_0, n_0) - E(q_1, n_0)] - \beta_1[E(q_0, n_1) - E(q_1, n_1)] \quad (3\text{-}3)$$

Here $\beta_0 = 1/k_B T_0$, $\beta_1 = 1/k_B T_1$ and *E* is the potential energy. If the protonation states of two adjacent replicas at an exchange attempt are the same, the exchange probability of our constant pH REMD will be equivalent to the conventional REMD exchange probability. However, if it is not the case, four potential energy terms are needed to calculate exchange probability. Under this circumstance, the constant-pH REMD becomes a REMD algorithm that combines both temperature and Hamiltonian REMD algorithms.

One possible concern of exchanging only structures would be the role of kinetic energy, especially when n_0 and n_1 are different. In the REMD algorithm developed by Sugita and Okamoto, the kinetic energy terms in the Boltzmann factors cancel each other on average through velocity rescaling (Eq. 2-88). Only potential energies are required to compute exchange probabilities. There is a problem in canceling kinetic energy terms when the numbers of particles of two systems attempting to exchange are not the same. However, according to the constant-pH MD algorithm proposed by Mongan et al., a proton does not leave the molecule but becomes a dummy atom when an ionizable side chain is in deprotonated state. Furthermore, that dummy atom retains its position and velocity which are controlled by molecular dynamics. Hence, the kinetic energy contributions to the Boltzmann weight will be cancelled out during exchange probability calculation, leaving only potential energy useful for the calculation.

The second possibility consists of exchanging protonation states as well as molecular structures at REMD Monte Carlo moves. For instance, replica 0 attempts to

move from state (q_0, n_0) to state (q_1, n_1) at temperatures T_0 in one MC move and replica 1 attempts to jump from state (q_1, n_1) to state (q_0, n_0) at temperature T_1. The detailed balance equation now can be written as:

$$\frac{w(\beta_0 q_0 n_0, \beta_1 q_1 n_1 \rightarrow \beta_0 q_1 n_1, \beta_1 q_0 n_0)}{w(\beta_0 q_1 n_1, \beta_1 q_0 n_0 \rightarrow \beta_0 q_0 n_0, \beta_1 q_1 n_1)} = \frac{w(\beta_1 q_1 n_1 \rightarrow \beta_1 q_0 n_0)}{w(\beta_1 q_0 n_0 \rightarrow \beta_1 q_1 n_1)} \cdot \frac{w(\beta_0 q_0 n_0 \rightarrow \beta_0 q_1 n_1)}{w(\beta_0 q_1 n_1 \rightarrow \beta_0 q_0 n_0)} \quad (3\text{-}4)$$

This equation states that the exchange probability is the product of MC transition probabilities at temperature T_0 and T_1. If the protonation states of two adjacent replicas are the same at an exchange attempt, the exchange probability of constant-pH REMD becomes the exchange probability of conventional temperature-based REMD. If n_0 and n_1 are different, then each MC transition is essentially the protonation state change step in constant-pH MD plus a structural transition. For example, consider the MC transition at temperature T_0,

$$w(\beta_0 q_0 n_0 \rightarrow \beta_0 q_1 n_1) = min\{1, exp(-\Delta_1)\} \quad (3\text{-}5)$$

In Eq. 3-5, Δ_1 has the form:

$$\Delta_1 = \beta_0[E(q_1, n_0) - E(q_0, n_0)] + \left(pH - pK_{a,ref}\right) + \beta_0[E_{elec}(q_1, n_1) - E_{elec}(q_1, n_0)] - \beta_0 \cdot \Delta G_{ref,MM} \quad (3\text{-}6)$$

The first term in Δ_1 derives from the transition in configuration at fixed protonation state n_0, and the rest corresponds to protonation state change at fixed structure q_1. E_{elec} represents the electrostatic component of potential energy. Similarly, the transition probability of MC jump at T_1 can be expressed as:

$$w(\beta_1 q_1 n_1 \rightarrow \beta_1 q_0 n_0) = min\{1, exp(-\Delta_2)\} \quad (3\text{-}7)$$

And

$$\Delta_2 = \beta_1[E(q_0, n_1) - E(q_1, n_1)] - \left(pH - pK_{a,ref}\right) - \beta_1[E_{elec}(q_0, n_1) - E_{elec}(q_0, n_0)] + \beta_1 \cdot \Delta G_{ref,MM} \quad (3\text{-}8)$$

Therefore, similar to Eq. 3-2, the exchange probability can be written as:

$$w(\beta_0 q_0 n_0, \beta_1 q_1 n_1 \rightarrow \beta_0 q_1 n_1, \beta_1 q_0 n_0) = min\{1, exp(-\Delta')\} \quad (3\text{-}9)$$

And

$$\Delta' = \Delta + \beta_0[E_{elec}(q_1, n_1) - E_{elec}(q_1, n_0)] - \beta_1[E_{elec}(q_0, n_1) - E_{elec}(q_0, n_0)] + (\beta_0 - \beta_1) \cdot \Delta G_{ref,MM} \quad (3\text{-}10)$$

In Eq. 3-10, Δ is the same quantity as in Eq. 3-3.

The exchange probability calculation in the second method of coupling REMD and constant-pH MD utilizes the same energy terms required by the first method since obtaining electrostatic potential energies does not require extra energy calculations. The advantage of implementing the second exchanging protocol over the first one should not be significant because it is the conformational sampling at higher temperature that greatly improves conformational sampling at lower temperatures. Allowing protonation states to change at exchange attempts does not provide extra gains in conformational sampling. In addition, one can always choose to sample protonation state space during the MD propagation. Therefore, only the first method of performing exchanges was implemented.

3.2.2 Simulation Details

Constant pH REMD simulations were carried out first on five reference compounds: blocked Aspartate, Glutamate, Histidine, Lysine and Tyrosine to test our method and implementation. The experimental pK_a values of those reference compounds are known and listed in Table 3-1. We later performed constant pH REMD

simulations on a model peptide ADFDA (Ala-Asp-Phe-Asp-Ala, unblocked termini) and the heptapeptide derived from OMTKY3 (residues 26 to 32 with blocked termini). Four replicas were used in the reference compounds and ADFDA REMD simulations. The temperatures were 240, 300, 370 and 460 K for all six molecules. The pH range for the study of acidic side chains was sampled from 2.5 to 6 and the pH range of histidine-δ is from 5.5 to 8. The basic side chains were titrated from pH 9 to 12. An interval of 0.5 was chosen for all titrations.

Eight replicas were chosen for the heptapeptide with a temperature range from 250 to 480 K. 10 ns were used for each replica in all REMD simulations and an exchange was attempted every 2 ps. A MC move to change protonation state was attempted every 10 fs. A second set of REMD runs was done with the same overall conditions but different initial structures in order to check simulation convergence.

To compare conformational and protonation state sampling, 100 ns of constant pH MD simulations were carried out for aspartate reference compound and ADFDA at the same pH values as in the REMD runs. For the heptapeptide, one set of 10 ns constant pH MD simulations were done at all pH values simulated by REMD method.

Constant pH REMD and MD simulations were done using the AMBER 10 molecular simulation suite.[199] The AMBER ff99SB force field[139] was used in all the simulations. The SHAKE algorithm[145] was used to constrain the bonds connecting hydrogen atoms with heavy atoms in all the simulations which allowed use of a 2 fs time step. OBC Generalized Born implicit solvent model[200] was used to model water environment in all our calculations. The Berendsen thermostat,[146] with a relaxation time of 2 ps, was used to keep the replica temperature around their target values. Salt

concentration (Debye-Huckel based) was set at 0.1M. The cutoff for non-bonded interaction and the Born radii was 30 Å.

3.2.3 Global Conformational Sampling Comparison Using Cluster Analysis

In our study, global conformational samplings have been compared utilizing cluster analysis.[169,188] Cluster analysis is a technique to group "similar" structures and each group is called a cluster. A cluster analysis measures the similarity between two objects. In the cluster analysis we performed, protein backbone similarity (measured by backbone RMSD) is considered and the hierarchical agglomerative clustering algorithm is employed. Hierarchical algorithm basically creates a hierarchy of clusters and a hierarchical algorithm can be agglomerative or divisive. The hierarchical agglomerative algorithm starts with considering every object as a cluster and combines similar clusters into one cluster, while the divisive algorithm starts with one cluster containing all objects and divides it into more groups.

In our work, the cluster analysis was done using the Moil-View program.[201] The MD and REMD trajectories (having same number of frames) at 300 K and under the same solution pH value were first combined. The ptraj module of the AMBER package has been utilized to create the combined trajectory. The "trajin" keyword was used to read in two trajectories and the "trajout" command generated the trajectory we need.

The combined trajectory was clustered based on peptide backbone atoms root-mean-square deviations (RMSDs). A cluster cutoff RMSD of 1.5 Å is chosen for both ADFDA and the heptapeptide during our analysis. By clustering the combined trajectory, the MD and REMD conformational samplings will populate the same clusters. The fraction of the conformational ensemble corresponding to each cluster (fractional population of each cluster) was calculated for MD and REMD run, respectively. Two

sets of fractional population of clusters were generated. One must note that the fractional population of each cluster from MD and REMD trajectory may not be the same. Therefore, the correlation between the two sets of fractional population can be investigated by plotting one set against the other and doing a linear fitting.

The Moil-View program will generate a file pointing out which cluster a snapshot in the combined trajectory belongs to. Thus, the fractional population of each cluster was obtained for MD and REMD simulation. If the MD and REMD simulations produced the same structural ensemble, the fractional population of a cluster from MD simulation will be the same as that from REMD simulation. Cluster population fraction from REMD simulation vs that from MD simulation was plotted (see Figure 3-7A). The correlation coefficient values which represent the correlations between MD and REMD cluster population were calculated at each solution pH value by doing linear regression.[169,188] A high correlation between MD and REMD cluster population indicates that the structure ensembles are similar to each other. This method provides a direct comparison of global conformational sampling between MD and REMD simulations.

The same technique was used when studying convergence of constant pH REMD and MD trajectories (see Figure 3-7B and Figure 3-12). When investigating convergence of conformational sampling, snapshots from two constant-pH REMD simulations (or two constant-pH MD simulations) were combined. The two constant-pH simulations should have the same temperatures and solution pH values. They only differ in initial structures. A high correlation coefficient indicates the two structural ensembles are similar and two conformational samplings are converged, while a poor

correlation means the structural ensembles are different and the conformational sampling depends on initial condition.

3.2.4 Local Conformational Sampling and Convergence to Final State

In our study, the local conformational sampling was examined by comparing the probability distribution of backbone dihedral angle pair (φ, ψ). Essentially, we are comparing the Ramachandran plot of a residue. Each (φ, ψ) probability density was computed by binning φ and ψ angle pairs 10°×10°, which would lead to a 36×36 histogram. These two dimensional histograms were normalized into populations and the contours were plotted. The metric used to evaluate (φ, ψ) probability density convergence was the root-mean-squared deviation (RMSD) between the cumulative (φ, ψ) histogram and the one produced by using all configurations. Each cumulative histogram was constructed by using (φ, ψ) pairs up to current time and following the same algorithm mentioned earlier in this section. Essentially, we were computing the RMSD between two matrices. The RMSD between the cumulative probability density at time t and the final probability density (all configurations were utilized to compute final probability density) is given by,

$$RMSD(t) = \sqrt{\sum_{i=1}^{36}\sum_{j=1}^{36}\left[P_{ij}(t) - P_{ij,final}\right]^2 / 36 \times 36} \quad (3\text{-}11)$$

where $P_{ij}(t)$ is the ijth element of the cumulative probability density of the (φ, ψ) pairs at time t and $P_{ij,final}$ is the corresponding element in the final probability density matrix.

3.3 Results and Discussion

3.3.1 Reference Compounds

We first applied our constant pH REMD method to the reference compounds. Table 3-1 shows the pK_a values predicted by REMD simulations (10 ns for each replica)

as well as the reference pK_a values. All our pK_a values were calculated by fitting to the HH equation. Agreement between constant pH REMD predictions and the reference values can be seen.

Table 3-1. The REMD pK_a predictions of reference compounds.

pKa	Aspartate	Glutamate	Histidine	Lysine	Tyrosine
REMD	3.97(0.01)	4.41(0.01)	6.40(0.03)	10.42(0.01)	9.61(0.01)
Reference	4.0	4.4	6.5	10.4	9.6

The numbers in parenthesis are the standard errors.

The pH titration curves of the same reference compounds showed agreement between MD (100 ns) and REMD simulations. Figure 3-2 demonstrates the REMD and MD titration curves of aspartic acid reference compound as an example.

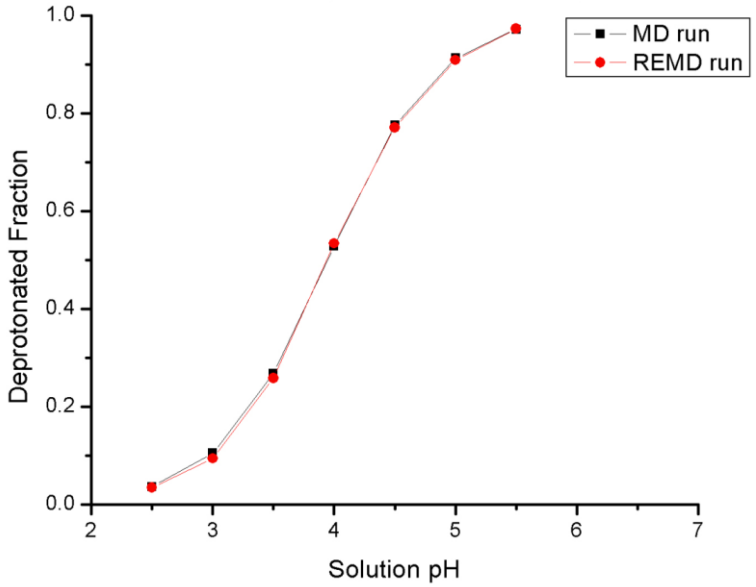


Figure 3-2. Titration curves of blocked aspartate amino acid from 100 ns MD at 300K and REMD runs. Agreement can be seen between MD and REMD simulations.

We further studied the convergence of protonation states sampling. REMD and MD protonation fraction (cumulative protonation fraction) were plotted with respect to MC attempts for aspartate reference compound at all pH values. Figure 3-3 demonstrated the protonated fraction versus time at pH 4 as one example. According to

Figure 3-3, it suggests that although the final pK_a predictions are the same between REMD and MD simulations, the protonation state sampling during REMD simulations clearly converges faster than that in a MD run.

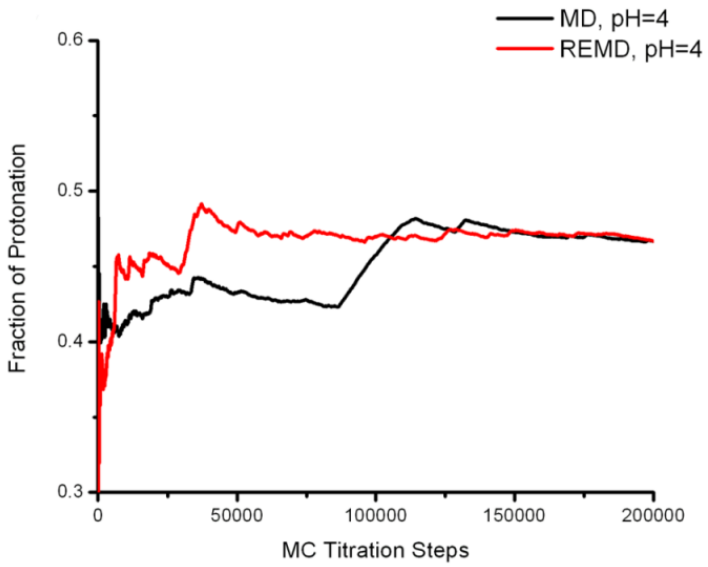


Figure 3-3. Cumulative average protonation fraction of aspartic acid reference compound vs Monte Carlo (MC) steps at pH=4.

3.3.2 Model peptide ADFDA

The model peptide ADFDA (as zwitterion) was chosen as a more stringent test of our constant pH REMD method. The charged termini will provide different electrostatic environment for each titratable Asp residue and hence a correct constant pH REMD model should reflect this difference between titration curves of the two Asp residues. The Asp2 residue is closer to the NH_3^+, so the deprotonated state is favored and the pK_a value of Asp2 residue should shift below 4.0 (which is the pK_a value of the reference aspartic dipeptide). The Asp4 residue is closer to the COO^- negative charge and hence the pK_a value should shift above 4.0.

The titration curves of the model peptide ADFDA from REMD simulations are shown in Figure 3-4. We can clearly see that Asp2 and Asp4 have different titration

curves from each other and from the reference compound. The pK_a value and Hill coefficient for each Asp residue were obtained by fitting titration curves to a Hill plot. The results are shown in Table 3-2. The REMD pK_a predictions reflect the difference between Asp2 and Asp4 due to different peptide electrostatic environments. We also displayed the MD titration curves of Asp2 and Asp4 in Figure 3-4 and listed the MD pK_a predictions and corresponding Hill coefficients in Table 3-2. The titration curve of Asp2 residue only showed a small difference between MD and REMD simulation. But we can see differences in titration behaviors of Asp4 between MD and REMD calculations when solution pH is below 5. Interestingly, Lee *et al.* studied blocked Asp-Asp peptide using CPHMD method, reporting different Hill coefficient for each of the two Asp residues.

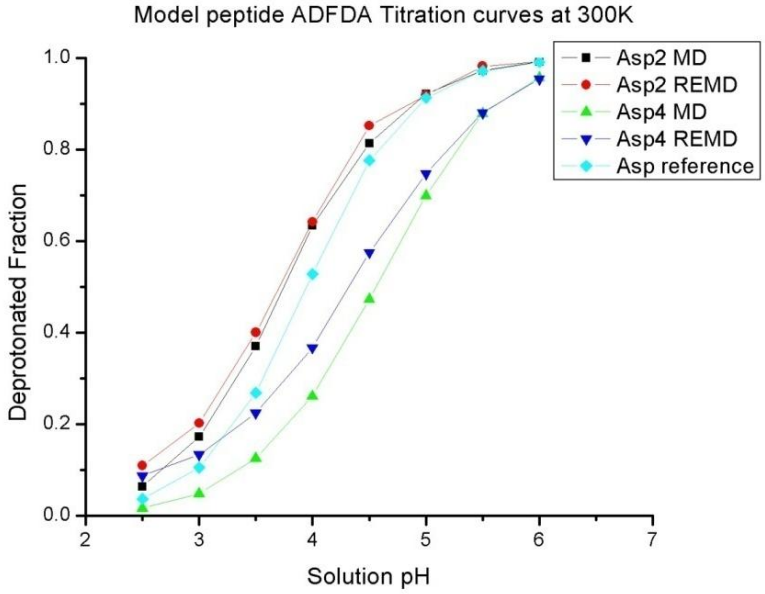


Figure 3-4. The titration curves of the model peptide ADFDA at 300K from both MD and REMD simulations. MD simulation time was 100 ns and 10 ns were chosen for each replica for REMD runs.

Table 3-2. pK_a predictions and Hill coefficients fitted from the Hill's Plot

	Asp2		Asp4	
	pK_a	Hill Coefficient	pK_a	Hill Coefficient
REMD	3.74	0.87	4.38	0.67
MD	3.76	0.89	4.54	0.85

Convergence rates of Asp2 titration behavior were compared between REMD and MD calculations due to the fact that Asp2 titration curves are very close. The cumulative protonated fractions versus MC attempts at pH 4 are shown in Figure 3-5. Likewise, faster convergence in protonation state sampling can be seen for REMD simulation even though both REMD and MD calculations resulted in the same final protonated fraction. Clearly, our constant pH REMD method accelerates the convergence of sampling of protonation states.

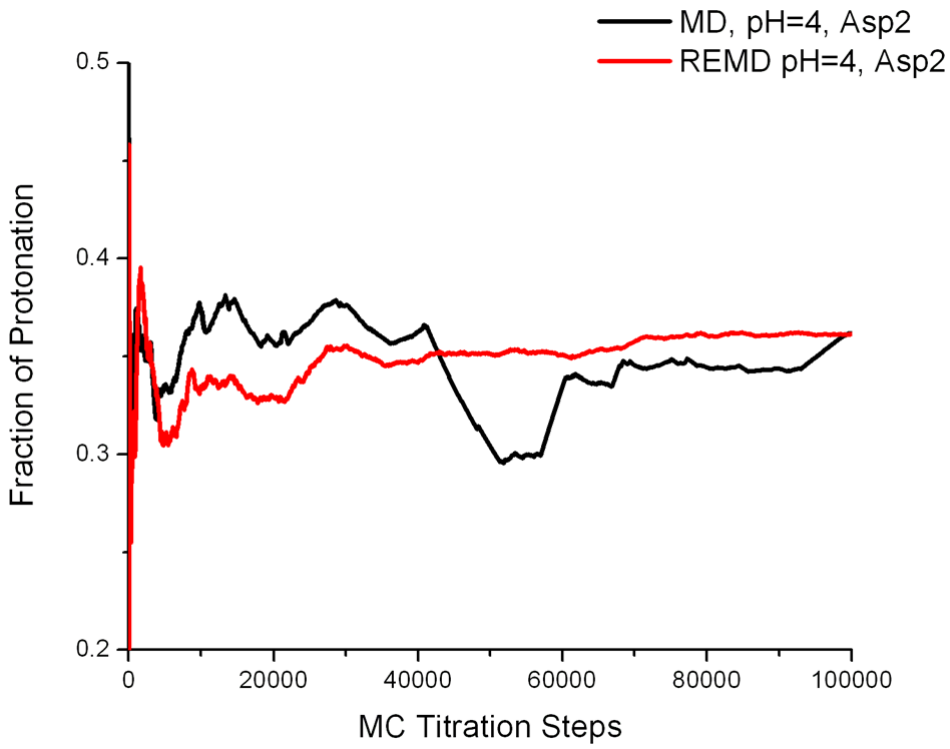


Figure 3-5. Cumulative average protonation fraction of Asp2 in model peptide ADFDA vs Monte Carlo (MC) steps at pH=4.

In addition to protonation state sampling, we also evaluated the conformational sampling in constant pH MD and REMD simulations. First, distribution of backbone φ and ψ angle pairs (Ramachandran plots) of residue Asp2, Phe3 and Asp4 in ADFDA at each solution pH were studied. The regions in Ramachandran plots sampled by MD and

REMD simulations are the same at all pH values. Ramachandran plots for residue Asp2 at pH 4 are shown in Figure 3-6 as an example.

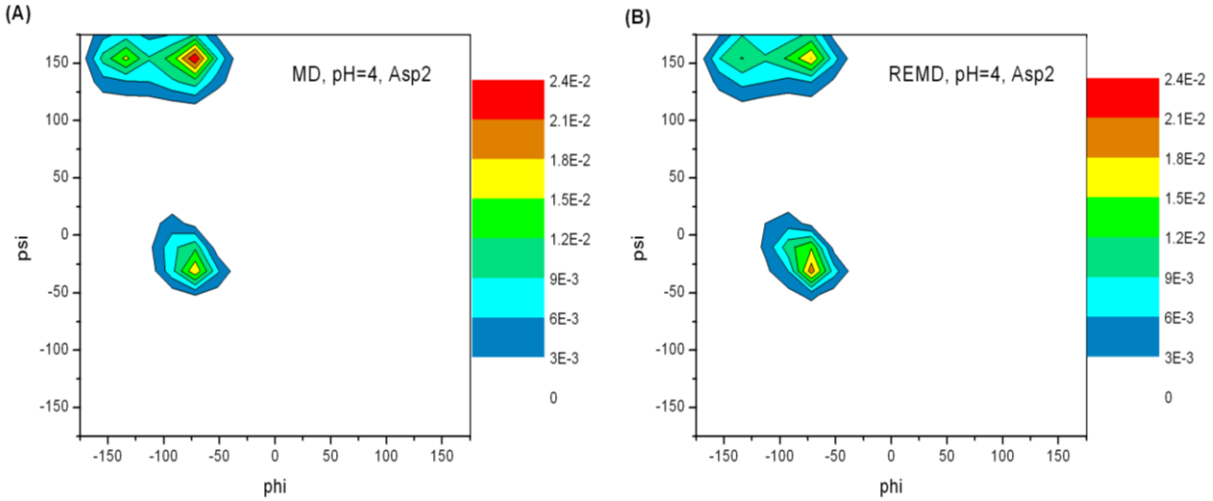


Figure 3-6. Backbone dihedral angle (φ, ψ) normalized probability density (Ramachandran plots) for Asp2 at pH 4 in ADFDA. Ramachandran plots at other solution pH values are similar. For Asp2, constant-pH MD and REMD sampled the same local backbone conformational space. Phe3 and Asp4 Ramachandran plots also display the same trend.

Since the Ramachandran plot only represented local conformational sampling, we also evaluated global conformational sampling by clustering MD and REMD trajectories and comparing the cluster populations. The MD and REMD cluster population R^2 values are listed in Table 3-3. A plot of cluster populations from MD and REMD trajectories at solution pH of 4 is shown in Figure 3-7A as an example. The large R^2 values indicate that the MD and REMD sampled the same conformational space and generated the same structure ensemble. The small size of ADFDA and simple structure of each residue make 100 ns long enough for MD to sample the relevant conformations.

We further studied the convergence of REMD simulations by comparing global conformation distribution between two REMD simulations starting from two different structures. Cluster populations of the two REMD simulations at solution pH 4 are

displayed in Figure 3-7B. The R^2 value is 0.959 at pH 4. This large correlation tells us that the two REMD simulations provide the same structure ensemble and hence the two simulations are converged.

Table 3-3. Correlation coefficients between MD and REMD cluster populations.

	pH=2.5	pH=3	pH=3.5	pH=4
R^2	0.94	0.90	0.79	0.93
	pH=4.5	pH=5	pH=5.5	pH=6
R^2	0.85	0.98	0.92	0.96

The R^2 values were calculated by linear regression.

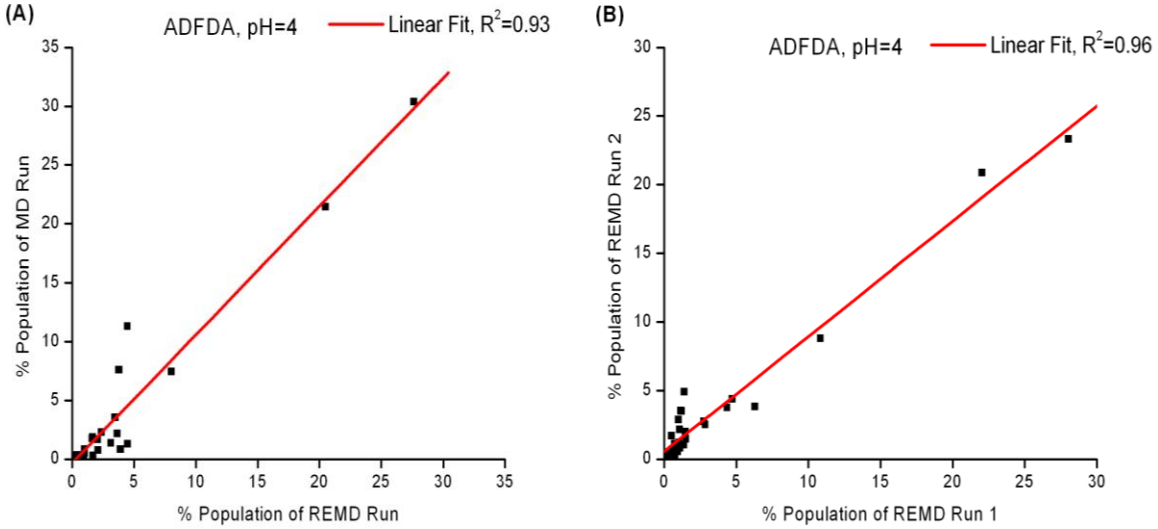


Figure 3-7. Cluster populations of ADFDA at 300K. A) MD vs REMD at pH 4. Trajectories from MD and REMD simulations are combined first. By clustering the combined trajectory, the MD and REMD structural ensembles will populate the same clusters. The fraction of the conformational ensemble corresponding to each cluster (fractional population of each cluster) was calculated for MD and REMD simulation, respectively. Two sets of fractional population of clusters were generated, and hence plotted against each other. B) Two REMD runs from different starting structures at pH 4. Large correlation shown in Figure 3-7B suggests that the REMD runs are converged. Large correlations between two independent REMD runs are also observed at other solution pH values. Correlations between MD and REMD simulations can be found in Table 3-3.

3.3.3 Heptapeptide derived from OMTKY3

We first compared the protonation state sampling between constant pH REMD and MD simulations. Titration curves of Asp3, Lys5 and Tyr7 from two sets of

simulations are plotted in Figure 3-8A and 3-8B. For each titratable residue, titration curves generated by constant pH REMD and MD are close to each other. Since the pK_a value of Asp3 in this heptapeptide is experimentally determined to be 3.6, it will be interesting to evaluate how our predicted values compare to the experimental result. The pK_a values of Asp3 were calculated based on Hill's plots which are displayed in Figure 3-8C. The predicted pK_a value is 3.7 for both REMD and MD simulations and they are in excellent agreement with the experimental pK_a value. Following the same procedures, our predicted pK_a values of Lys5 and Tyr7 from constant pH REMD and MD simulations were obtained. Not surprisingly, the REMD and MD schemes yielded essentially the same predicted pK_a values for Lys5 and Tyr7.

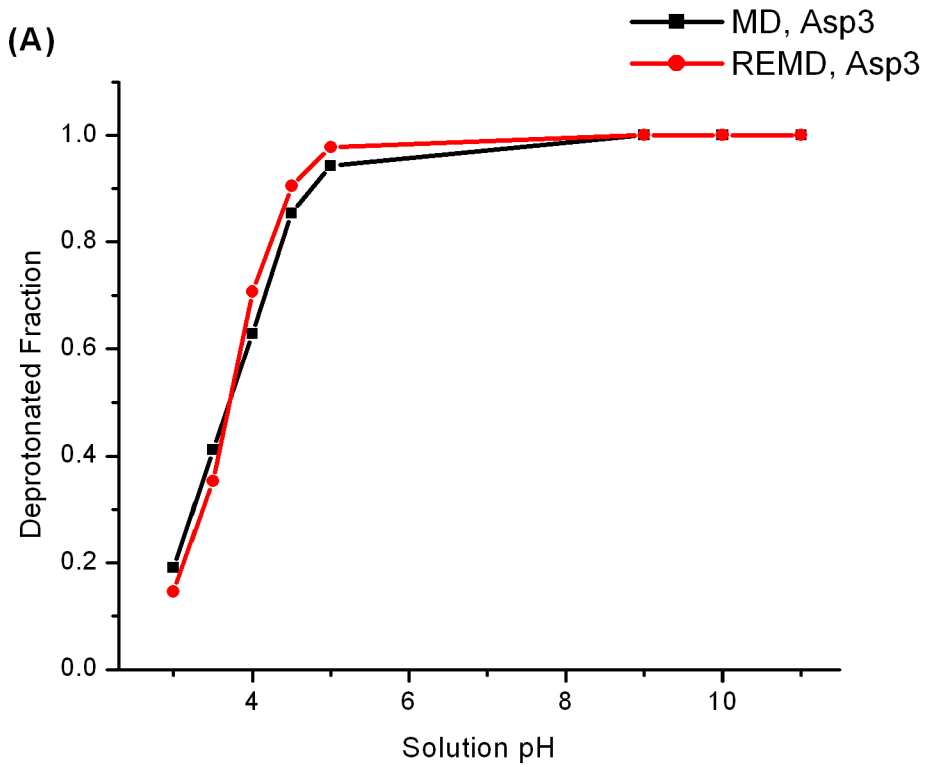


Figure 3-8. A) Titration curves of Asp3 in the heptapeptide derived from protein OMTKY3. B) Titration curves of Lys5 and Tyr7 in the heptapeptide derived from protein OMTKY3. C) shows the Hill's plots of Asp3. The pK_a values of Asp3 are found through Hill's plots.

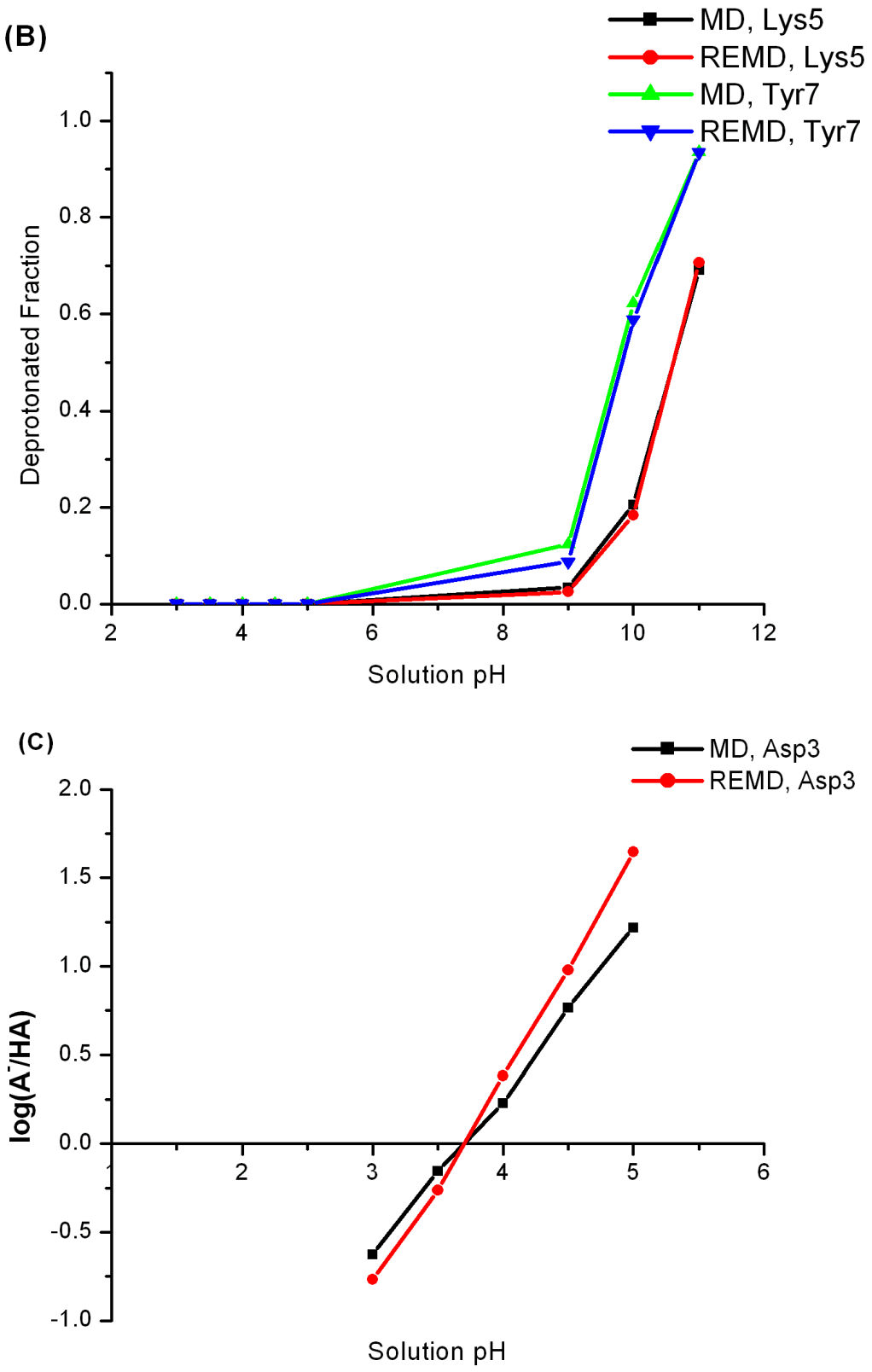


Figure 3-8. Continued

Although the final pK_a predictions are the same for constant pH REMD and MD simulations, constant pH REMD showed clear advantage in the convergence of protonation state sampling. Again, we chose the cumulative average protonation fraction vs MC steps to reflect protonation state sampling convergence for all three titratable residues. Several representative plots are shown in Figure 3-9. The trend that constant pH REMD simulations produce faster convergence in protonation fraction is universal. Therefore, it is very clear that constant pH REMD method is better than constant pH MD in protonation state sampling.

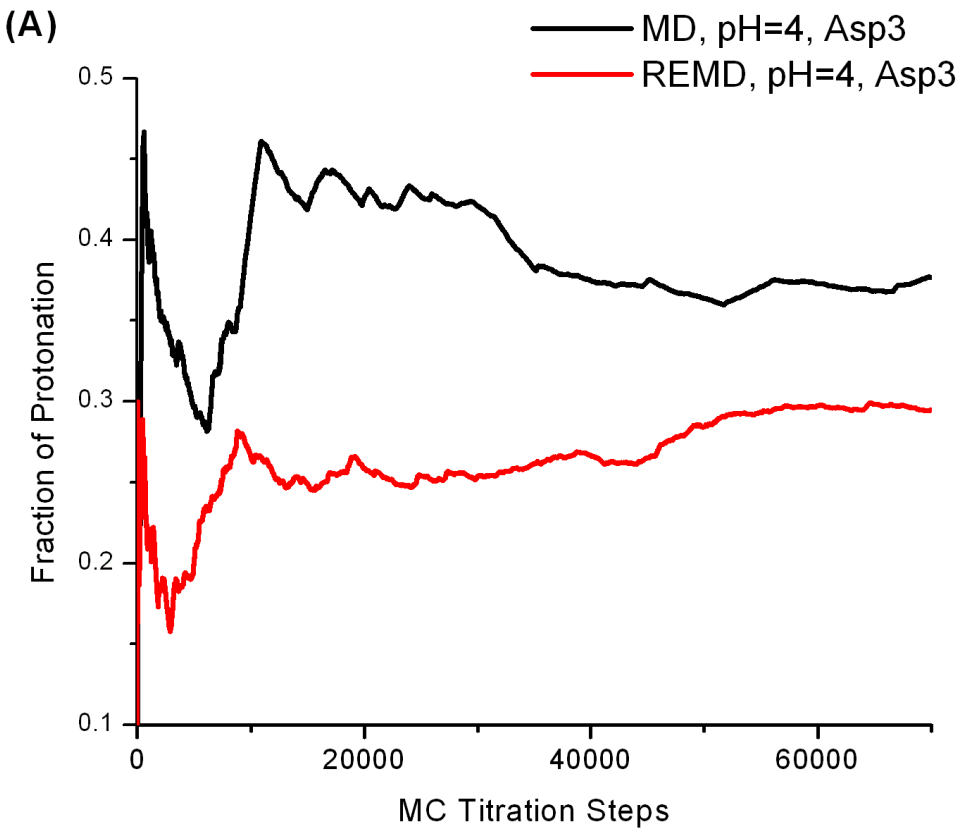
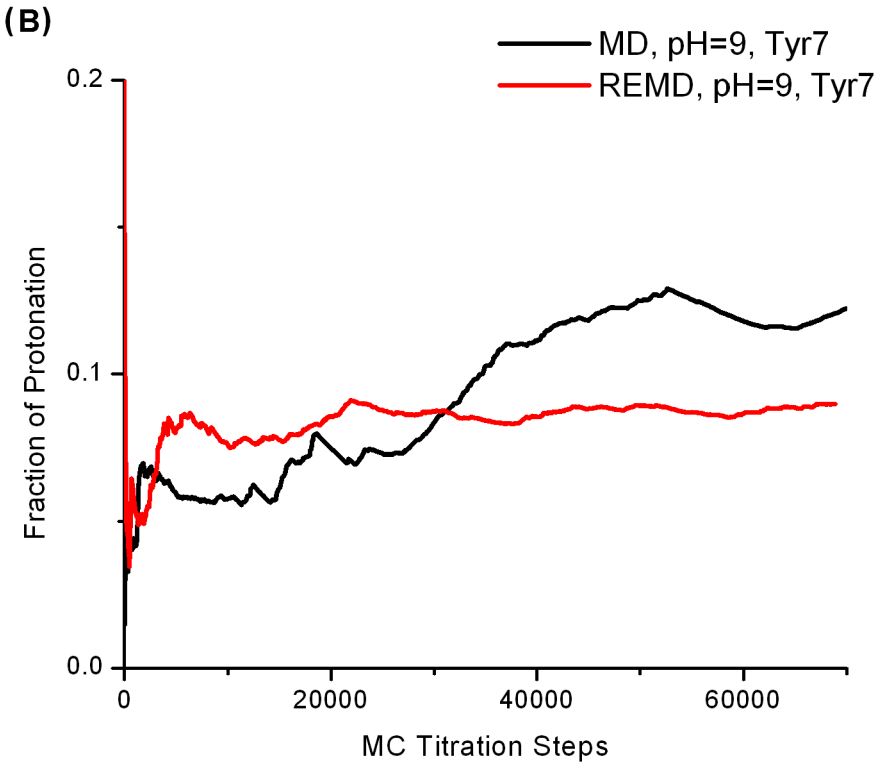


Figure 3-9. A) Cumulative average protonation fraction of Asp3 of the heptapeptide derived OMTKY3 vs MC steps. B) and C) is cumulative average protonation fraction of Tyr7 and Lys5 in the heptapeptide vs MC steps, respectively. Clearly, faster convergence is achieved in contant-pH REMD simulations.



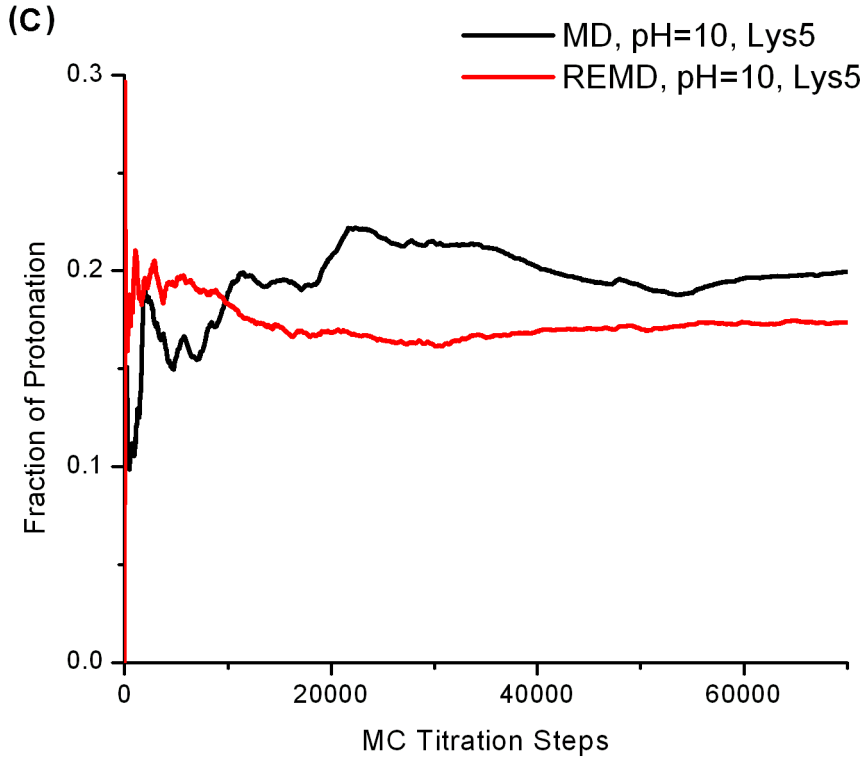


Figure 3-9. Continued

Conformational sampling is an important issue in constant pH studies. We first looked at the conformational sampling on peptide backbones. We evaluated backbone conformational sampling through Ramachandran plots. Six residues (from Ser2 to Tyr7) are studied here. Not surprisingly, Ramachandran plots from constant pH REMD and MD simulations are very close, suggesting that the overall local conformational samplings are similar. The Ramachandran plots of Asp3 at pH 4 are shown in Figure 3-10 as examples. The only exception is Tyr7 in acidic pH values. Tyr7 can visit the left-handed alpha helix conformation during constant pH REMD runs but is not able to do that in constant pH MD runs. In general, constant pH REMD and MD yielded the same Ramachandran plots for the heptapeptide.

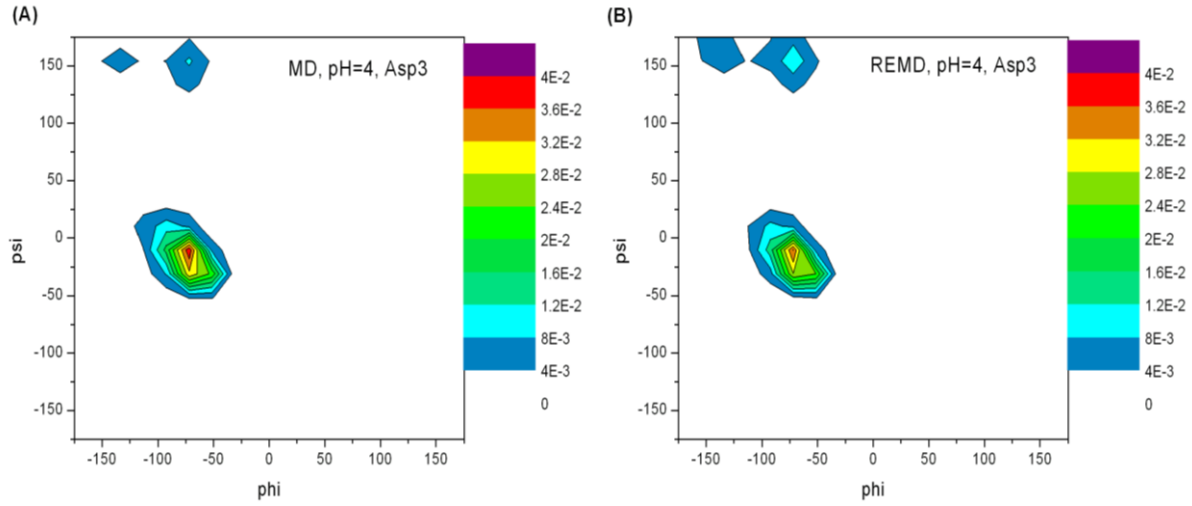


Figure 3-10. Dihedral angle (φ, ψ) probability densities of Asp3 at pH 4. A) Constant-pH MD results. B) Constant-pH REMD results. The two probability densities are almost identical, indicating that constant-pH MD and REMD sample the same local conformational space. All others also show very similar trend.

As demonstrated earlier, the overall samplings of (φ, ψ) distribution by constant pH REMD and MD are similar for Ser2 to Thr6. It is interesting to determine how fast each sampling scheme reaches the final distribution. We studied evolution of backbone conformational sampling based on cumulative data as what we did in the case of

protonation state sampling convergence. As described in the METHOD section, the RMSD between the (φ, ψ) distribution up to current time versus total simulation time was calculated. The smaller a RMSD is, the closer a probability distribution reaches to the final distribution. Deviations were calculated starting from the second nanosecond with time intervals incremented by 100 ps. The cumulative time-dependence RMSD of Asp3 and Lys5 are also shown in Figure 3-11 as examples. As seen in the figures, these curves decrease faster in constant pH REMD simulations. Figure 3-11 suggests that although the final (φ, ψ) probability distributions are similar between constant pH REMD and MD simulations, the constant pH REMD simulation clearly reaches the final state faster.

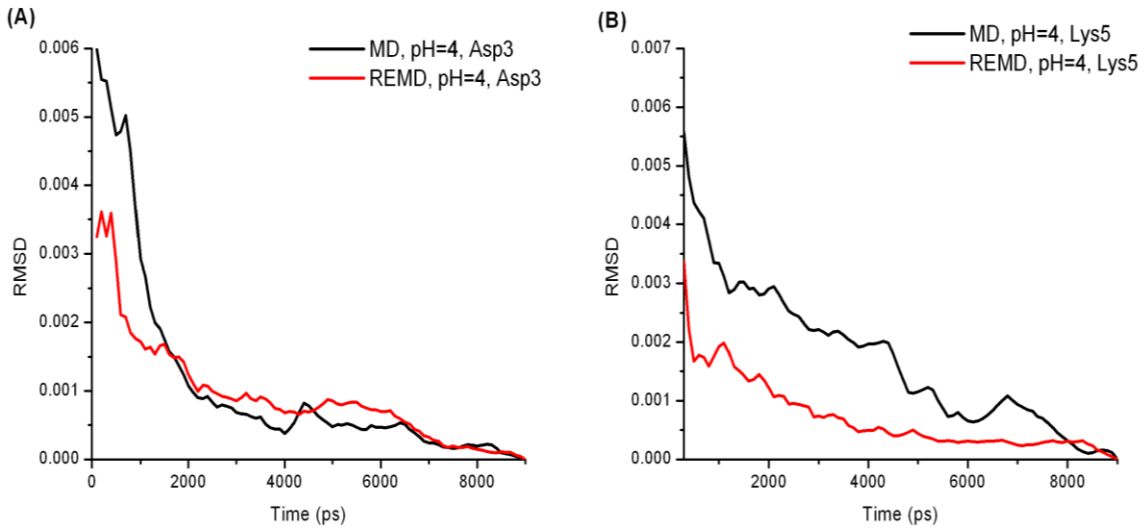


Figure 3-11. The root-mean-square deviations (RMSD) between the cumulative (φ, ψ) probability density up to current time and the (φ, ψ) probability density produced by entire simulation. (φ, ψ) probability density convergence behaviors at other pH values also show that REMD runs converge to final distribution faster.

Cluster analysis was also applied to study the convergence of conformation sampling in the heptapeptide. By comparing cluster populations between the first and

second half of one trajectory, one could check the convergence of that simulation. The two halves of a structural ensemble should yield the same populations in each cluster if convergence is reached. For example, simulations at pH 4, both constant pH REMD and MD yield about 20 clusters and the correlations coefficients are calculated through a linear regression. Cluster population plots and correlation coefficients are shown in Figure 3-12. A much higher correlation coefficient can be seen in constant pH REMD simulation, suggesting the two halves of the constant pH REMD simulation at pH 4 populate each cluster much more similarly than the corresponding constant pH MD does. Hence, much better convergence is achieved by the constant pH REMD run.

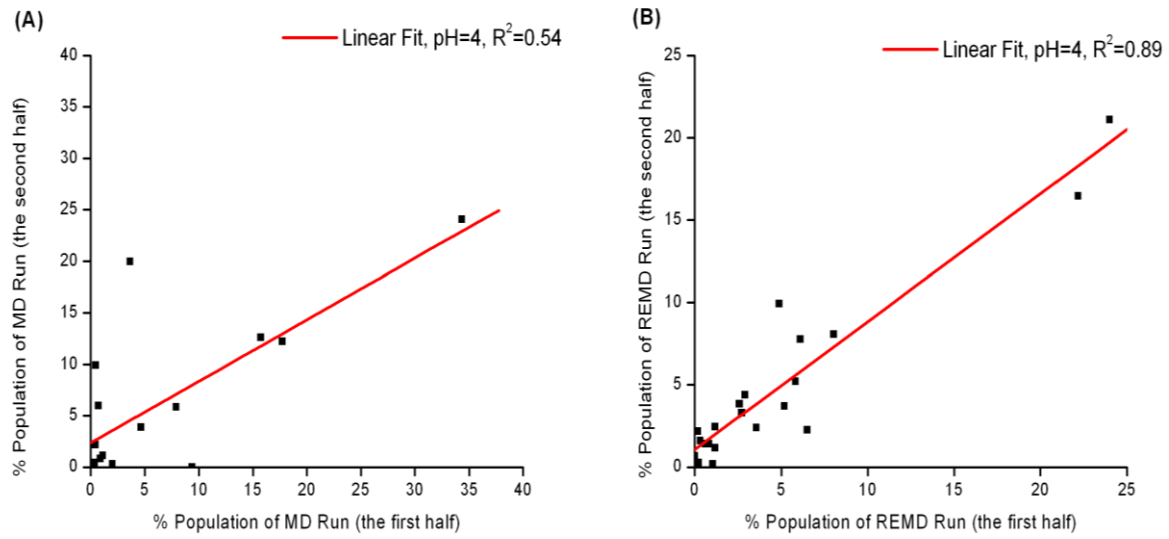


Figure 3-12. Cluster population at 300 K from constant pH MD and REMD simulations at pH=4. Cluster analysis is performed using the entire simulation. The populations in each cluster from the first and second half of the trajectory are compared and plotted. Ideally, a converged trajectory should yield a correlation coefficient to be 1. A) Constant pH MD. B) Constant pH REMD. Much higher correlation coefficient can be seen in constant pH REMD simulation, suggesting much better convergence is achieved by the constant pH REMD run.

3.4 Conclusions

In our work, we have applied replica exchange molecular dynamics (REMD) algorithm to the discrete protonation state model developed by Mongan *et al.* in order to study pH-dependent protein structure and dynamics. Seven small peptides were selected to test our constant pH REMD method. Constant pH molecular dynamics (MD) simulations were ran on the same peptides for comparison. The constant REMD method results are encouraging. The constant REMD method can predict pK_a values in agreement with literature and experimental results. Constant pH REMD method also displays advantage in convergence behaviors during protonation states and conformational sampling.

The REMD algorithm has been proven beneficial to study pH-dependent protein structures. Our future work will include studies of pH-dependent protein dynamics and application of this constant pH REMD to large proteins.

CHAPTER 4
CONSTANT-pH REMD: STRUCTURE AND DYNAMICS OF THE C-PEPTIDE OF RIBONUCLEASE A

4.1 Introduction

The protein and peptide folding problem[202] is an important aspect of protein science and biophysical chemistry.[203] In 1961, Anfinsen studied the refolding of denatured ribonuclease (RNase).[204] He first increased the temperature of the protein and the protein lost its functional three-dimensional shape (native state). When Anfinsen lowered the temperature, he found that the RNase was able to refold into its normal shape, without any other help. His experiment raised questions about protein folding. In general, people are interested in the thermodynamics (such as free energy landscape, folding pathway, and interactions in a protein), folding kinetics (such as how fast a protein folds), and native state prediction for a given sequence in protein folding.[202] Both experimental and theoretical approaches have been employed to understand protein folding.[205,206]

From now on, our introduction to protein folding will focus on computer simulations. In a protein folding simulation, the concept of free energy landscape always plays an important role.[202,207] Many questions can be answered once the free energy landscape is obtained. Levinthal,[208] in 1968, proposed that it is impossible for a protein to search all its conformations during folding process because the time taken to visit all conformations will be much longer than the folding time observed. His argument is well known as the “Levinthal’s paradox”. Thus, proteins must fold to their native states along some well-defined folding pathways. The “new” view of protein folding is the free energy landscape theory, which provides a statistical view of the folding landscape.[202,203,207] The folding process does not require chemical-reaction-like steps between specific

states. Basically, a protein folds on a funnel-shaped free energy landscape, which is defined by the amino acid sequence of the protein. Folding process is a directed visit of conformations on a landscape in order to reach the native state, which is the most thermodynamically stable conformation. Changing temperature, adding denaturant to the protein solution, or changing solution pH value of the protein system is able to change the free energy landscape, and hence affect protein folding. The free energy landscape of a protein is often rugged[51] and requires advanced sampling techniques such as REMD method to sample the conformational space. Due to the visual limitation, a free energy landscape is frequently projected onto one or two reaction coordinates. In practice, the free energy landscape is often projected onto several important reaction coordinates such as the radius of gyration of a protein, the number of backbone hydrogen bonds, and native contacts. Principal component analysis has also been carried out to generate the folding free energy landscape. The relative free energy (potential of the mean force, PMF) can be calculated by the following,

$$\Delta F(B \to A) = F(A) - F(B) = -k_B T \ln\big(P(A)/P(B)\big) \quad (4\text{-}1)$$

where $\Delta F(B \to A)$ is the relative PMF between state A and state B defined by reaction coordinate(s), $P(A)$ and $P(B)$ are the probability density of find state A, and B along the reaction coordinate(s), respectively.

Knowing the free energy landscapes can help people understand folding mechanisms. Transition states, intermediates, and folding pathways can be obtained from a folding free energy landscape. For example, when the free energy barrier between folded and unfolded state is disappeared, the folding is called downhill folding, in which the folding time is determined by diffusion rate on the free energy landscape.

One example of the protein folding free energy landscape studies is simulating the folding of C-terminal β-haripin of protein G, performed by Zhou *et al*. in 2001.[184] The OPLSAA force field, SPC explicit water model, and REMD algorithm have been employed in their simulation. The free energy landscape has been projected onto seven different reaction coordinates such as radius of gyration, number of hydrogen bonds, and fraction of native contacts. Two-dimensional free energy landscapes along those reaction coordinates were generated in order to elucidate the folding pathway. Four different states were found in the folding landscape, native state, unfolded state, and two intermediate states. Structural features of each state were also characterized. The formation of hydrophobic core and hydrogen-bonding in the folding process has been investigated. They have found that the hydrophobic core and hydrogen bonds formed almost simultaneously after initial collapse.

Although not investigated in this chapter, protein folding kinetics is also an important aspect of protein folding.[209] One example of the folding kinetics study is seeking the speed of protein folding.[210] Computer simulations have been performed to elucidate folding kinetics.[211] The Pande group at Stanford University pioneered computer simulations of folding kinetics.[206,211-213] When studying protein folding kinetics, the Pande group conducted multiple independent MD simulations starting from different initial conditions. The probability of the native state in the structure ensemble was computed after a pre-defined simulation time. Assuming the folding mechanism is two-state folding and follows the first-order reaction kinetics, and the transition time is much shorter than staying time in either state, the probability of barrier-crossing can be given by,

$$P(t) = 1 - e^{-kt} \quad (4\text{-}2)$$

where t is simulation time and k is the folding rate. In the limit of $t \ll 1/k$, Eq. 4-2 can be simplified to $P(t) \cong kt$, according to the Taylor expansion. The probability of barrier-crossing can be computed by using the fraction of simulations that crossed the barrier. Other methods utilized to explore folding kinetics include Markov state models.[195,198,214-217]

One example of predicting folding time is given by studying the C-terminal β-hairpin of protein G. In their studies, Pande and co-workers[213] utilized the OPLSAA force field and the GB implicit solvent model using water-like viscosity via Langevin collision coefficient. A total simulation time of 38 μs has been accumulated through 2700 independent simulations, among which 8 completely folded trajectories were found. Thus, a folding time of 4.7±1.7 μs can be derived from Eq. 4-2, which is in agreement with the experimental result of 6 μs. Furthermore, the folding free energy landscape has been generated and the folding pathway and folding intermediates etc have also been probed.

Another area of protein folding simulation is to probe protein folding through the unfolding simulations. The unfolding simulations adopt the assumption that folding processes follow the reverse pathways of unfolding processes. Both temperatures and denaturants can be employed to denature proteins. Levitt and Daggett have been performed unfolding simulations extensively.[218-220]

The C-peptide, residues 1 to 13 from the N-terminus of RNase A, is a peptide well studied by experiments.[5,7,221-226] In 1971, Brown and Klee[223] first observed the presence of α-helix of C-peptide through circular dichroism (CD) spectroscopy. This peptide was

further studied extensively by the Baldwin group.[5,7,222,224,226] CD spectroscopy showed that the C-peptide demonstrated pH-dependent α-helix formation. The mean residue ellipticity at 222 nm of the C-peptide showed a bell-shaped pH profile, having a maximum at pH value of 5. Mutation experiments indicated that the Glu2 and His12 in the C-peptide were crucial to the pH-dependent helix formations.[5,7,224,226] Maximal mean residue ellipticity occurred at pH 5 because both the glutamate and histidine residues are charged at that pH. NMR experiments on an analog of the C-peptide (RN-24) by the Wright group also confirmed the formation of complete and partial helix.[225] Two side chain interactions were believed to stabilize the partial helix formation in the C-peptide and its analogs in the mutation experiments and NMR studies.[7,224-226] A salt-bridge between Glu2 and Arg10 side chains was proposed to improve the helix formation as the pH values increased to 5. The interaction between Phe8 and His12 was also believed to improve helix formation as the pH values reduced to pH of 5.

The folding and side chain interactions of C-peptide and its analogs were also extensively studied by molecular simulations.[227-235] Schaefer *et al*.[232] studied the helical conformations and folding thermodynamics. The Okamoto group[228-230,233-235] has performed thorough investigations of the C-peptide using a multicanonical algorithm (MUCA) and the replica exchange method (REM) in both implicit solvent and explicit solvent. They have studied secondary structures of the C-peptide, roles of Glu2 and His12 in the C-peptide, helix-coil transition, and dielectric effect in the implicit solvent. Ohkubo and Brooks[231] utilized REMD simulations with the GB model to explore the helix-coil transition of short peptides including the C-peptide. Conformational entropy as a function of temperature has been explored for the C-peptide and its analogues

(different chain length). The conformational entropy has been found to be proportional to chain length over a wide range of temperatures. Felts and co-workers[227] carried out REMD simulations with the AGBNP implicit solvent model to study the folding free energy landscape of the C-peptide. The free energy landscape was projected onto radius of gyration and helical length. The possible interaction between Glu2-Arg10 was also explored. Dielectric effects of AGBNP solvation model on helical length and salt-bridge has been investigated too. In 2005, Sugita and Okamoto[233] performed replica exchange multicanonical algorithm simulations in explicit solvent to explore the folding mechanism and side-chain interactions such as Glu2-Arg10 and Phe8-His12. They constructed folding free energy landscape along the principal component axes. The correlations between Glu2-Arg10 and Phe8-His12 interactions and the C-peptide conformations have been elucidated. They have found that the minimum free energy conformation possess both interactions. They have also suggested that the purpose of Glu2-Arg10 salt-bridge is to prevent α-helix extending to N-terminus of the C-peptide and the Phe8-His12 stabilizes the alpha-helix conformation toward the C-terminus. More importantly, Khandogin *et al*.[112] studied the pH-dependent folding of the C-peptide with REX-CPHMD. Important electrostatic interactions such as the Lys1-Glu9, Glu2-Arg10 and Phe8-His12 interactions were also investigated.

The C-peptide has also been selected to test the effect of force fields on protein folding simulations and simulation convergence. In 2004, Yoda *et al*.[234,235] tested six commonly employed force fields (AMBER94, AMBER96, AMBER99, CHARMM22, OPLS-AA/L, and GROMOS96) on the C-peptide as well as the C-terminal fragment from the B1 domain of the G-peptide in explicit water using generalized-ensemble

method. Melting curves have been studied. Secondary structures of both peptides were also computed and compared with experimental data. AMBER99 and CHARMM22 were found showing best agreement for the C-peptide.

In this chapter, we present a study of the C-peptide using constant-pH REMD method introduced in the previous chapter. The effect of pH on the folding of C-peptide and the structural ensemble is studied. We compare directly with experimental measurements of helicity, namely the mean residue ellipticity at 222 nm. Important electrostatic interactions such as Glu2-Arg10 salt-bridge and Phe8-His12 interaction are also examined.

4.2 Methods

4.2.1 Simulation Details

The C-peptide we simulated has the sequence: KETAAAKFERQHM. The N-terminus of the C-peptide (lysine) is charged while the C-terminus (methionine) is capped with an amide. For our study, constant-pH REMD simulations were performed starting from a completely extended structure at pH values 2, 3, 4, 5, 6.5 and 8. Eight replicas were chosen with a temperature range from 260 to 420 K. A simulation time of 44 ns were used for each replica in all REMD runs and an exchange was attempted every 2 ps. The structures obtained from the first 4 ns were discarded, resulting in a 40 ns of production time for each replica. Glu2, Lys7, Glu9 and His12 are selected to be titratable. A MC move to change protonation state was attempted every 10 fs. A second set of REMD runs was done at pH values of 2, 5 and 8 starting from a fully helical initial structure in order to check simulation convergence. The three pH values are selected to represent low pH, pH where maximum helicity was observed experimentally and high pH, respectively.

AMBER 10 molecular simulation suite[199] was used to simulate the C-peptide. The AMBER ff99SB force field[139] was used in all the simulations. The SHAKE algorithm[145] was used in all the simulations which allowed use of a 2 fs time step. OBC Generalized Born implicit solvent model[200] was used to model water environment in all our calculations. The Berendsen thermostat,[146] with a relaxation time of 2 ps, was used to keep the replica temperature around their target values. Salt concentration (Debye-Huckel based) was set at 0.1 M. The cutoff for non-bonded interaction and the Born radii was 30 Å (this cutoff is longer than the peptide).

4.2.2 Cluster Analysis

When studying the folding of C-peptide, the roles of cluster analysis are two-fold. One role is to compare structural ensembles and check convergence at particular temperature and solution pH value, while the other is to analyze a single ensemble of structures to investigate protein structures and interactions. As described in the previous chapter, cluster analysis was done using the Moil-View program[201] and the C_α RMSD has been chosen to measure structure similarity.

When comparing conformational sampling, two different ways of comparisons have been adopted. The first way is to compare the first and the second halves of one trajectory. In this case, cluster analysis was performed on a single trajectory and the cluster information can be utilized to study folding thermodynamics and interactions in the C-peptide. The second way is to compare the structural ensembles produced by simulations starting from the fully extended and fully helical structures. In the second case, the two trajectories (having same number of frames) at 300 K and under the same solution pH value were first combined. Then the combined trajectory was clustered on the basis of peptide backbone atoms root-mean-square deviations (RMSDs). The

population fraction corresponding to each cluster was obtained for both trajectories. The correlation coefficient, which represents the correlation between the cluster populations of the two trajectories, was calculated at each solution pH value by doing linear regression. A high correlation indicates that the structure ensembles are close to each other. This method provides a direct comparison of global conformational sampling between the two trajectories. A cluster cutoff RMSD of 2.0 Å is chosen during our analysis.

4.2.3 Definition of the Secondary Structure of Proteins (DSSP) Analysis

The secondary structures of the C-peptide have been explored by DSSP algorithm,[236] which is proposed by Kabsch and Sander. The DSSP algorithm identifies the secondary structure of a residue by hydrogen bond calculations. The calculation is based on electrostatic energy between backbone carbonyl group and amide group,

$$U = q_1 q_2 \left(\frac{1}{r_{ON}} + \frac{1}{r_{CH}} - \frac{1}{r_{OH}} - \frac{1}{r_{CN}}\right) \cdot 332\ kcal/mol \quad (4\text{-}3)$$

In the above equation, q_1 and q_2 are the partial charges on each atoms. If the electrostatic energy is below -0.5 kcal/mol, then a hydrogen bond will assigned to corresponding carbonyl and amide groups. The secondary structure of a residue is labeled by one letter: G for 3_{10} helix, H for alpha-helix, I for pi-helix, B for antiparallel beta-sheet, b for parallel beta-sheet, and T for turns.

4.2.4 Computation of the Mean Residue Ellipticity

CD spectroscopy is one of the most commonly used techniques to study protein secondary structures and folding.[237] Chiral molecules absorb left circularly polarized light (LCPL) and right circularly polarized light (RCPL) differently. CD spectroscopy

measures the difference in absorbance of LCPL and RCPL of a chiral molecule. It can provide information of protein secondary structures.

Electromagnetic waves contain oscillating electric and magnetic fields perpendicular to each other and to the propagating directions. A circularly polarized light (CPL) has an electric field vector rotating along its propagation direction but maintains its magnitude. This is in contrast to linearly polarized light which has an electric field vector oscillating in one plane but change its magnitude. When a LCPL is propagating toward an observer, the electric field vector rotates counterclockwise, while the RCPL rotates clockwise.

When a circularly polarized light passes through chiral molecules, the difference in the absorption of LCPL and RCPL is given by:

$$\Delta\varepsilon(\lambda) = \varepsilon_L(\lambda) - \varepsilon_R(\lambda) \tag{4-4}$$

where ε_L and ε_R is extinction coefficient of LCPL and RCPL, respectively and λ is wavelength. $\Delta\varepsilon$ has the dimensions of $(cm \cdot M)^{-1}$ or $cm^2 \cdot dmol^{-1}$. The extinction coefficient ε can be calculated by Beer-Lambert law: $\varepsilon = A/c \cdot l$ where A is the absorbance, c is the concentration, and l is the width of the cuvette. This difference gives CD spectroscopy. Many CD instruments record signal in ellipticity, θ, which is measured in degrees. The ellipticity can be calculated as: $\theta = 32.98(A_L - A_R) = 32.98 \cdot c \cdot l \cdot \Delta\varepsilon$, where 32.98 has unit of degree. A more frequently adopted measurement of CD is the molar ellipticity $[\theta]$,[238]

$$[\theta] = \frac{100\theta}{c \cdot l} = 3298 \cdot \Delta\varepsilon(\lambda) \tag{4-5}$$

Here, the molar ellipticity has units of $deg \cdot cm^2 \cdot dmol^{-1}$.

The integrated intensity of a CD band is called rotational strength. Theoretically, for a electronic transition from ground state (0) to excited state (i), the rotational strength can be calculated as,

$$R_{oi} = Im(\langle\psi_0|\hat{\mu}_e|\psi_i\rangle \cdot \langle\psi_i|\hat{\mu}_m|\psi_0\rangle) \quad (4\text{-}6)$$

where ψ_0 and ψ_i is the wavefunction of electronic ground and excited state, respectively; $\hat{\mu}_e$ and $\hat{\mu}_m$ is the electronic transition and magnetic transition dipole moment operator, respectively; and Im stands for the imaginary part. Eq. 4-6 suggests that the frequently adopted units of rotational strength are Debye-Bohr magnetons (DBM, 1 DBM=$9.274 \times 10^{-39}\ erg \cdot cm^3$, where erg is the cgs unit of energy). Eq. 4-6 is origin-dependent because the magnetic transition dipole moment operator is origin-dependent. In order to avoid this origin-dependence, the dipole-velocity formulation can be employed,

$$R_{oi} = -(eh/2\pi m v_{0i}) Im\left(\langle\psi_0|\vec{\nabla}|\psi_i\rangle \cdot \langle\psi_i|\hat{\mu}_m|\psi_0\rangle\right) \quad (4\text{-}7)$$

Here, e is the charge of an electron, m is the mass of an electron, and v_{0i} is the frequency of the transition.

According to the paper of Sreerama and Woody,[238] CD spectrum can be calculated as, assuming each CD band (CD transition) is a Gaussian function of wavelength,

$$\Delta\varepsilon_k = 2.278 R_k \lambda_k / \Delta_k \quad (4\text{-}8)$$

where $\Delta\varepsilon_k$, R_k, λ_k, and Δ_k is the CD, rotational strength, wavelength and half-bandwidth (one half of the width at $1/e$ of its maximum) of the kth transition, respectively. In Eq. 4-8, the constant 2.278 has the dimensions of $DBM^{-1} \cdot cm^2 \cdot dmol^{-1}$.

The far ultraviolet (far UV, with a wavelength smaller than 250 nm) CD spectra of proteins can yield important information about the secondary structures of proteins.[238] In the far UV range, peptide bonds in a protein are the main chromophores. Thus, the CD spectra in the far UV range are reported on a residue basis (mean residue ellipticity). In a protein CD spectrum, a positive band at ~190 nm and two negative bands at 208 nm and 222 nm can be found for α-helix.[239] In particular, a strong negative band at 222 nm is a leading indication of the presence of helical structures. Structures containing β-sheet will show two bands in CD spectra: a positive band at ~198 nm and a negative band at ~215 nm.[240]

Computing protein CD spectra using quantum mechanical methods combining with Eq. 4-7 is only possible in principle due to the size and complexity of protein structures. The matrix method[241] using pre-determined parameters has been adopted to tackle this problem. In the matrix method, a secular matrix is constructed based on transition energies and interactions between transitions. A protein is considered as a set of independent chromophores. Each local transition energies and interactions between transitions in different chromophores are utilized to construct the secular matrix. A transition on a local chromophore is represented by a charge distribution. The charge distributions, as parameters, are determined from quantum mechanical wavefunctions or experiments or a combination of both.[242-244] The off-diagonal elements of the secular matrix, which represent the interactions between transitions in different chromophore, are further simplified by charge-charge (monopole-monopole) electrostatic interaction,[238]

$$V_{ij,kl} = \sum_m \sum_n q_{ijm}\, q_{kln} / r_{ijm,kln} \tag{4-9}$$

Here, $V_{ij,kl}$ is the electrostatic energy between transition j on chromophore i and transition l on chromophore k. m sums over the point charges of transition j on chromophore i and n sums over the point charges of transition l on chromophore k, and r denotes for the distance between two charges.

Diagonalization of the secular matrix using a unitary transformation will yield the eigenvalues and eigenvectors corresponding to all transitions of the protein. Eigenvalues provide information about transition energies and the eigenvectors describe the mixing of local transitions. The rotational strength can be obtained from eigenvectors.

In this work, the algorithm developed by the Woody group[238,244] was used to compute the mean residue elliptcity. Detailed description of their algorithm can be found in the paper of Sreerama and Woody. The peptide transitions (two $\pi\pi^*$ transitions at 140 and 190 nm, respectively and one $n\pi^*$ transition at 220 nm) were computed using the Matrix method[241] in the origin-independent form.[245] Transition charge distributions (monopole charges) are obtained from INDO/S[246] semi-empirical electronic structure calculations. Side chain transitions of phenylalanine, tyrosine and tryptophan were also included in the calculations. The α-helix formation can be characterized by two negative bands at 208 and 222 nm, and a positive band at 192 nm. Following the experiments performed by the Baldwin group, the mean residue ellipticity at 222 nm ($[\theta]_{222}$) was calculated to generate the pH profile.

In practice, Woody's program reads in one protein structure in PDB format and yields the mean residue ellipticity and the rotational strength as a function of wavelength. Therefore, the ptraj module of the AMBER 10 package has been utilized to

generate a protein structural ensemble in order to find out an ensemble average of the mean residue ellipticity at 222 nm.

4.3 Results and Discussion

4.3.1 Testing Structural Convergence

Conformational sampling convergence is investigated utilizing cluster analysis, as described earlier. Two ways of checking conformational sampling of the simulations from the fully extended structure are utilized. One way is to compare the first and the second halves of the trajectory and the other way is to compare to the structural ensembles produced by simulations starting from a fully helical structure. The R^2 values of the cross clustering are listed in Table 4-1. Plots demonstrating the cluster population correlations from both ways at pH 2 are showed in Figure 4-1 as an example. The large R^2 values indicate that converged structural ensembles are achieved through 40 ns simulations.

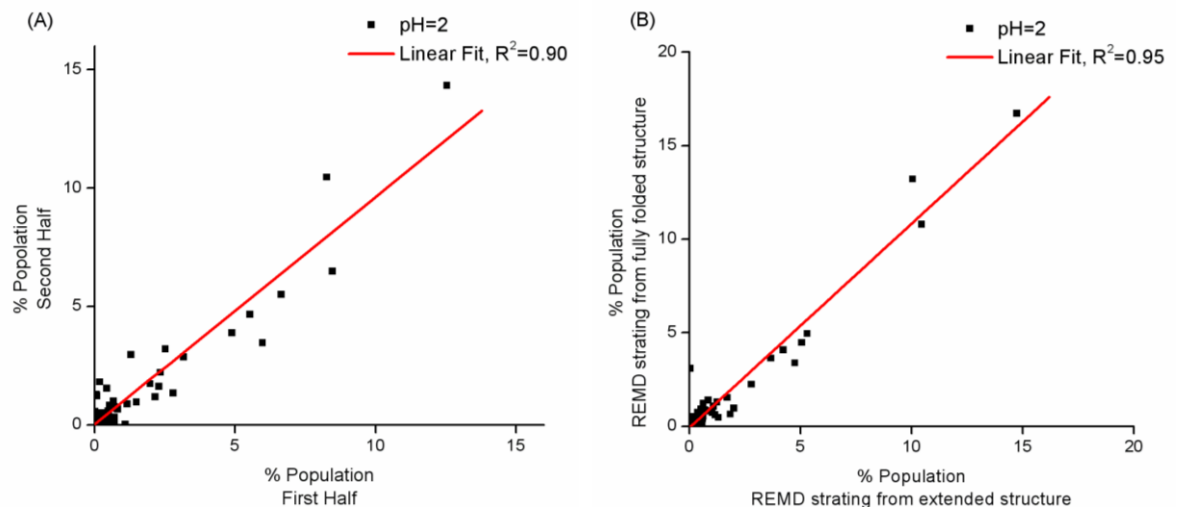


Figure 4-1. Cluster population at 300 K from constant pH REMD simulations at pH 2. A) Cluster analysis is performed on the trajectory initiated from fully extended structure. The populations in each cluster from the first and second half of the trajectory are compared and plotted. B) Two REMD runs from different starting structures at pH 2. Correlation coefficients at other pH values can be found in Table 4-1.

Table 4-1. Correlation coefficients between two sets of cluster populations.

	pH = 2	pH = 3	pH = 4	pH = 5	pH = 6.5	pH = 8
R^2 (E vs E)	0.90	0.92	0.90	0.94	0.93	0.85
R^2 (E vs H)	0.95	---	---	0.88	---	0.84

E vs E means comparing the first and the second halves of the trajectories starting from the fully extended structure. E vs H stands for comparing structural ensemble given by simulations starting from fully extended and fully helical structures, respectively.

4.3.2 pK_a Calculation and Convergence

Four residues of the C-peptide are titratable in our constant-pH REMD simulations: Glu2, Lys7, Glu9 and His12. Lys7 is always protonated in the pH range of 2 to 8, as expected. Thus, only the data from glutamate and histidine residues are analyzed. For each glutamate and histidine residue, the fraction of deprotonation at each pH value is obtained and a Hill's plot is utilized to find out the pK_a value. The pK_a values are 3.1, 3.7 and 6.5 for Glu2, Glu9 and His12 respectively.

The cumulative average fraction of protonation vs constant-pH MC attempts is chosen to study the convergence of the pK_a calculation. The cumulative average fraction of protonation represents the time evolution of the protonation state sampling. As shown in Figure 4-2, a stabilized fraction of protonation is achieved through 40 ns simulations.

4.3.3 The Mean Residue Ellipticity of the C-peptide

The mean residue ellipticity of the C-peptide at each pH value and at 300 K was computed. The pH-profile of the $[\theta]_{222}$ (Figure 4-3) is clearly a bell-shaped curve, in agreement to the experimental pH-profile of the $[\theta]_{222}$. The maximum of our calculated

$[\theta]_{222}$ is at pH value of 5, with a numerical value of ~ -6400 deg·cm^2·dmol^{-1}. However, the computed values of $[\theta]_{222}$ at the ends (pH = 2, 3, and 8) suggest that the helix is more populated in the simulations than in experiments at those pH values.

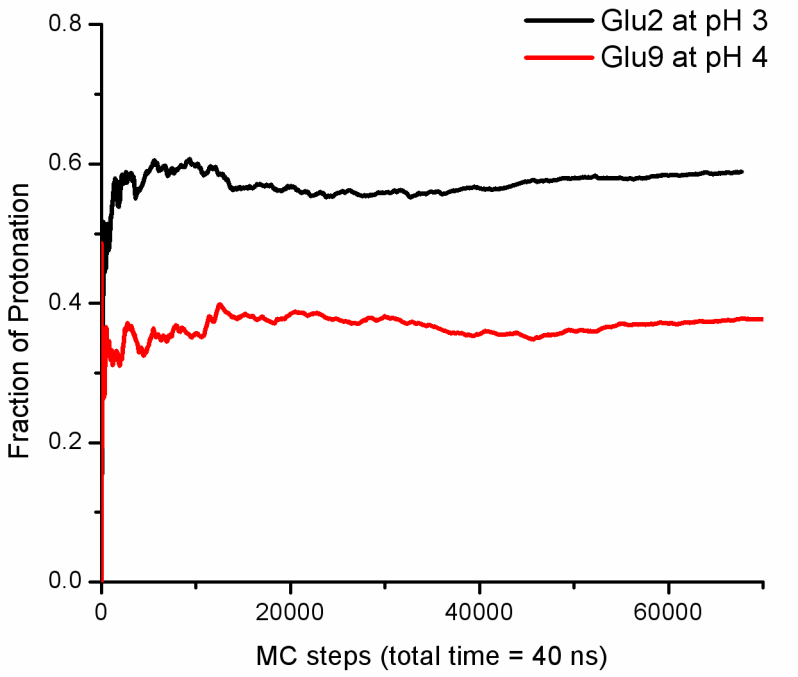


Figure 4-2. Cumulative average fraction of protonation vs Monte Carlo (MC) steps. Only the two glutamate residues are shown here and the histidine residue is found to show the same trend. The pH values are selected such that the overall average fraction of protonation is close to 0.5.

As mentioned in the section 2.2.2, the protonation state model involves using parameters fitted at 300 K, thus results obtained at temperatures other than 300 K should be viewed qualitatively, not quantitatively. C-peptide at a temperature lower than 300 K shows a more negative $[\theta]_{222}$ (more helical), while the $[\theta]_{222}$ becomes less negative (less helical) when the temperature is higher than 300 K. Experiments showed that the pH-profile becomes flat at high temperatures.[5] Our results also reflect the same trend: pH profile of the $[\theta]_{222}$ at 420 K is flat and less negative than those at 300 K, while the pH profile at 280 K is still bell-shaped and more negative.

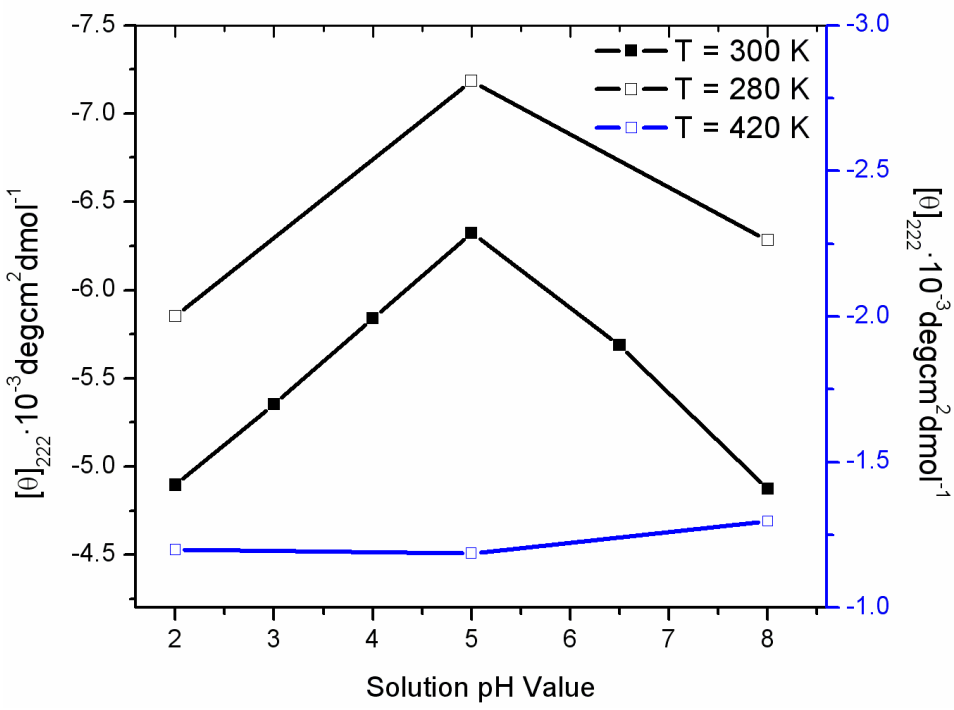


Figure 4-3. Computed the mean residue ellipticity at 222 nm as a function of pH values. A bell-shaped curve at 300 K is obtained with a maximum at pH 5. The effect of temperature on mean residue ellipticity at 222 nm is also demonstrated.

4.3.4 Helical Structures in the C-peptide

In order to examine the helical conformations in different environments, constant-pH REMD at pH values 2, 5, and 8 are selected to represent the pH range. The secondary structures of the C-peptide were computed utilizing the DSSP algorithm.[236] Any residue which according to the DSSP algorithm belongs to the 3_{10}-helix or α-helix conformation is called helical. The helical percentages of each residue are shown in Figure 4-4. The maximum helical percentage of a residue is ~ 55% at pH 2 and 5, and the maximum helical percentage is ~ 40% at pH 8. The averaged helical percentage at pH 5 is around 30%, which is in good agreement with experiments (29±2%). Figure 4-4 suggests that the C-peptide contains a lot of non-helical structures, even at pH 5 where the helical content is maximal.

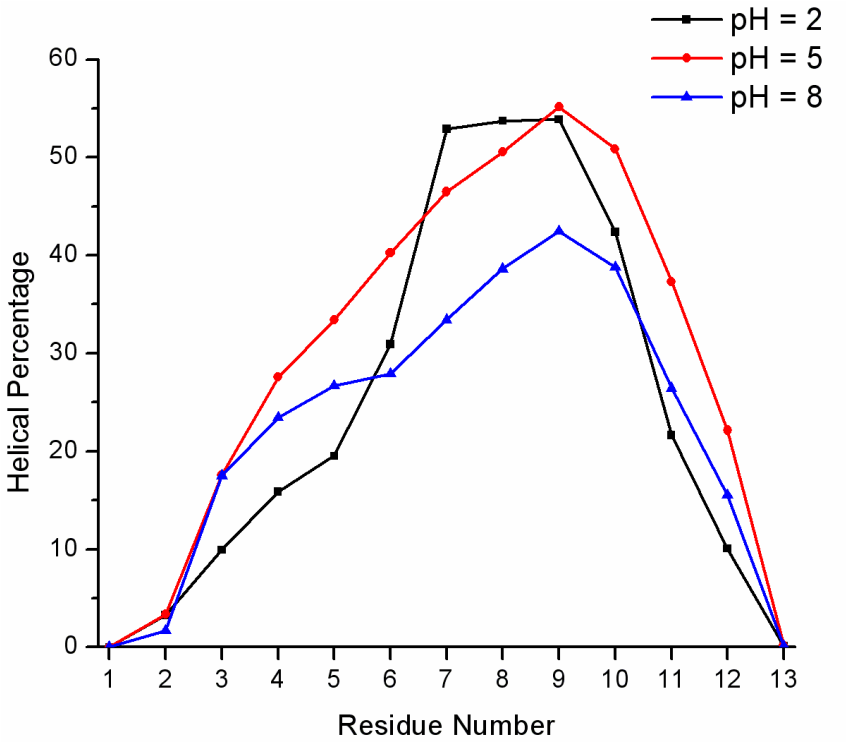


Figure 4-4. Helical Content as a function of residue number.

We calculated the C_α RMSD vs the fully folded structure (the fully helical structure has a C_α RMSD of 0.8 Å relative to the ribonuclease A X-ray structure, Thr3 to His12 are chosen to calculate C_α RMSD) and the C_α radius of gyration (R_g) of the C-peptide. The time series and the probability density of RMSDs and R_g are illustrated in Figure 4-5. According to Figure 4-5B, two conformations can be seen at all three pH values. The conformation with the smaller RMSD represents structures closer to the fully helical structure and the structural ensemble at pH 5 possesses more such kind of structures than the other two structural ensembles. Figure 4-5D demonstrate the probability density of the R_g, and it suggests that the C-peptide is more compact at pH 5 than at pH 2 and 8. The results of R_g agree with the results of RMSDs because the helical structures are more compact.

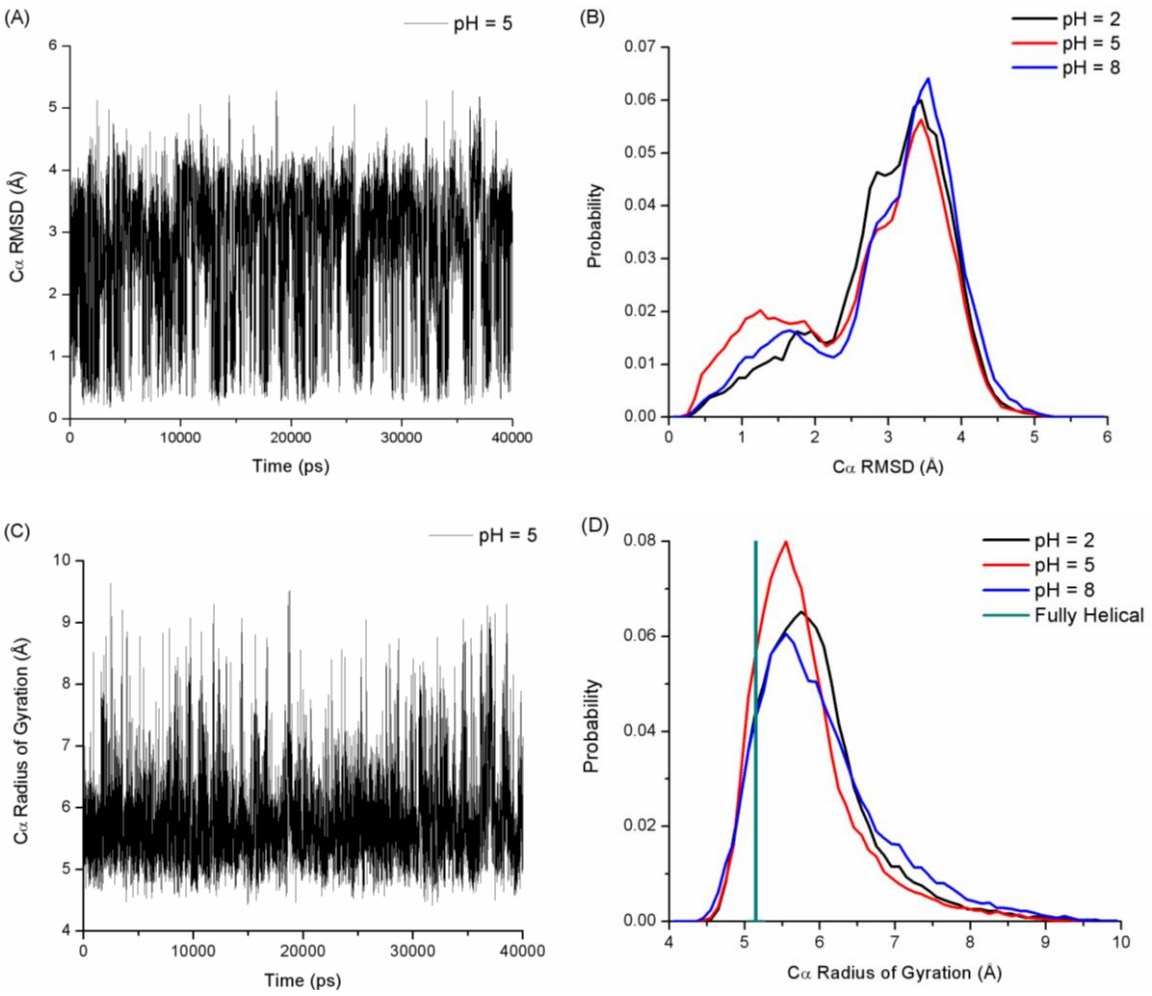


Figure 4-5. A) Time series of C_α RMSDs vs the fully helical structure at pH 5. The first two residues at each end are not selected because the ends are very flexible. B) Probability densities of the C_α RMSDs. Clearly, the structural ensemble at pH 5 contains more structures similar to the fully helical structure. C) Time series of C_α radius of gyration at pH 5. D) Probability density of the C_α radius of gyration. More compact structures are found at pH 5.

We further studied the details of the C-peptide structural ensemble with respect to pH values. The studies of helical structure were on the basis of our DSSP results. We first show the probability density of total number of helical residues at pH 2, 5 and 8 in Figure 4-6A. As expected, simulations at pH 5 generated the smallest number of non-helical structures and the percentage is ~ 25%. Simulation at pH 8 generated the most non-helical structures and ~ 37% of the structural ensemble possesses no helical

residue. For those structures possessing helical residues, structures having four helical residues are the most probable and structures containing three helical residues are also common at all three pH values. Besides, structures possessing six helical residues are also found. Furthermore, simulation at pH 5 yielded more configurations possessing seven-residue and longer helices. Thus, longer helical chains are formed more often at pH 5.

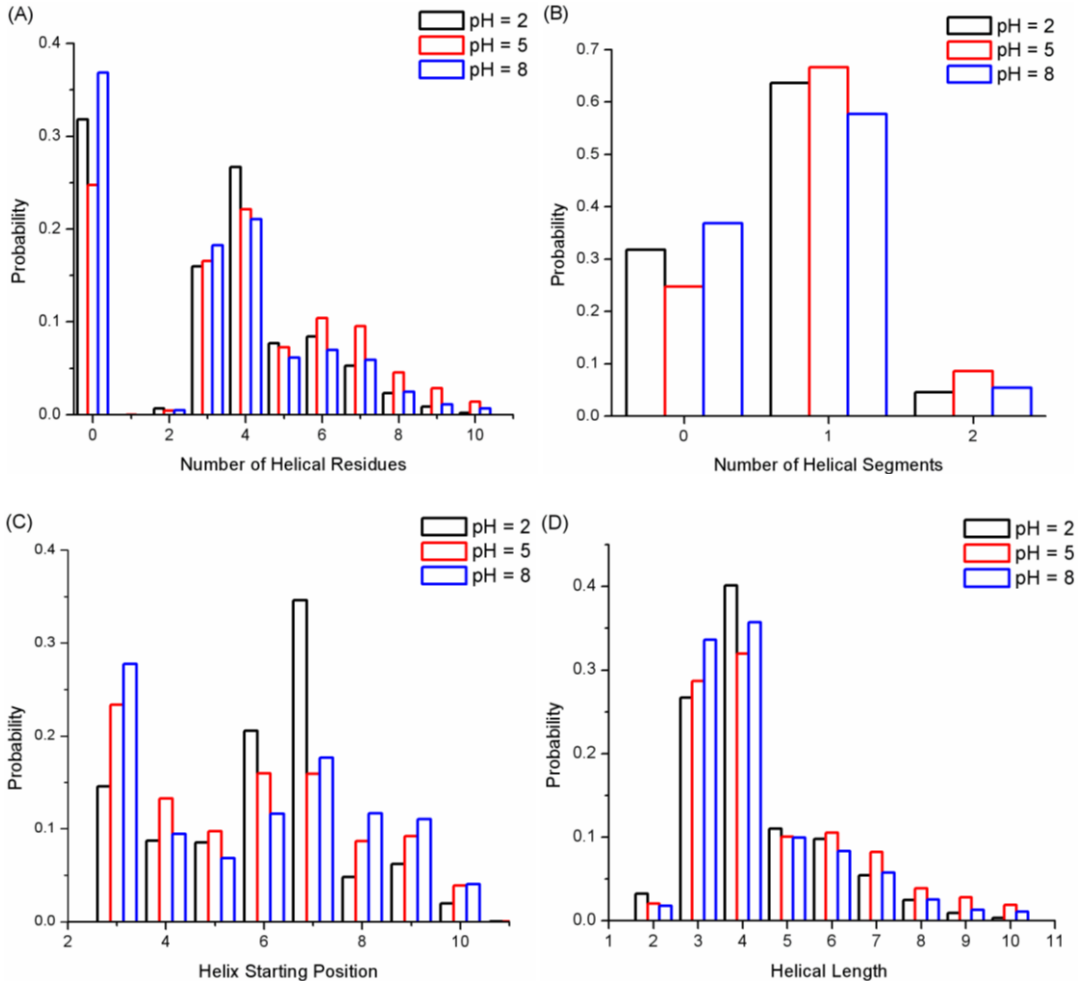


Figure 4-6. A) Probability densities of number of helical residues in the C-peptide. B) Probability densities of the number of helical segments in the C-peptide. A helical segment contains continuous helical residues. The probability of forming the second helical segment is very low at all three pH values, thus only the first helical segment is further studied. C) Probability densities of the starting position of a helical segment. D) Probability densities of the length of a helical segment (number of residues in a helical segment).

Next, the number of helical segments (a helical segment contains continuous helical residues) is studied and shown in Figure 4-6B. The number of helical segment ranges from zero to two at all three pH values. However, C-peptide structures having two helical segments are really rare. The probability densities of having two helical segments at pH 2 and 8 are ~ 0.05, while that at pH 5 is ~ 0.1.

Due to the small population of the second helical segment, the analysis of the helical length (number of helical residues in a segment) and the helix starting position (residue number of the amino acid initiating a helical segment) is focused on the first helical segment. Figure 4-6C demonstrates the probability density of helix starting position in the C-peptide. The helix starting position is affected by pH. The most probable starting position is affected by solution pH. At pH 2, Lys7 is the most favorable position to start a helix but the most probable place to initiate a helix is Thr3 at pH 5 and 8. At pH 2 and 5, Thr3, Ala6 and Lys7 are favorable positions to start a helix, while Thr3 and Lys7 are the favorable place to start a helix at pH 8. However, the effect of solution pH on the helical segment length is not as significant as the effect on helix starting position. Figure 4-6D shows that the three-residue or four-residue helices are dominant at all three pH values.

4.3.5 The Two-Dimensional Probability Densities

Two-dimensional (2D) probability density can be employed to study the correlations between important variables. The peaks in the plots indicate the coupling between two variables and represent stable conformations. The more populated a region is, the more stable the corresponding conformation is. The 2D probability densities between helix starting position and helical length are illustrated in Figures 4-7 to 4-9. Helices consisting of Thr3-Ala5, Lys7-Arg10 and Glu9-His12 are present at all

three pH values, while the number of helical conformations is more at pH 5 and 8. At pH 2 and 5, the most probable helix formation is the four-residue helix starting from Lys7 (Lys7-Arg10). The 2D-probability densities reveal that the six-residue (Lys7-His12) helix and the seven-residue (Ala6-His12) helix are found stable at pH 5. At pH 8, Thr3-Ala5 becomes the most favorable helical formation. Lys7-Arg10 and Lys7-His12 are also favorable. At pH 8, a new seven-residue helix (Thr3-Glu9) is found.

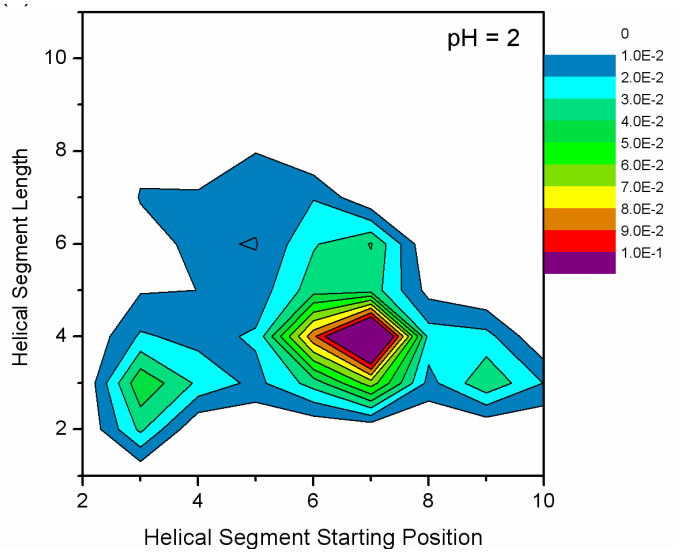


Figure 4-7. 2D probability density of helical starting position and helical length, pH = 2.

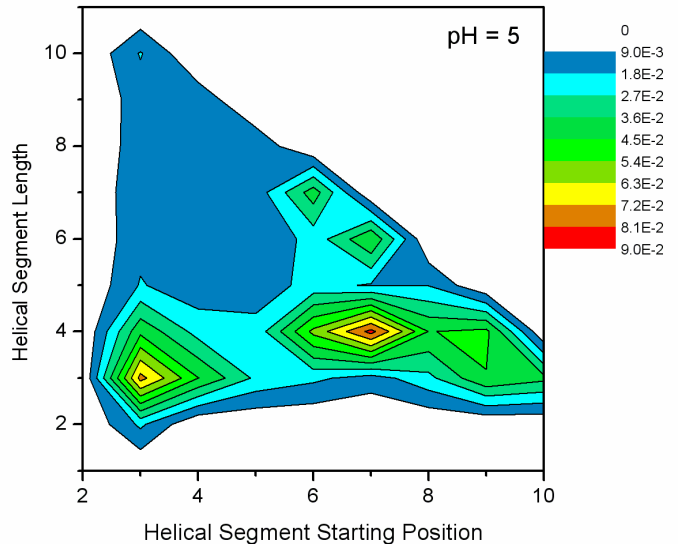


Figure 4-8. 2D probability density of helical starting position and helical length, pH=5.

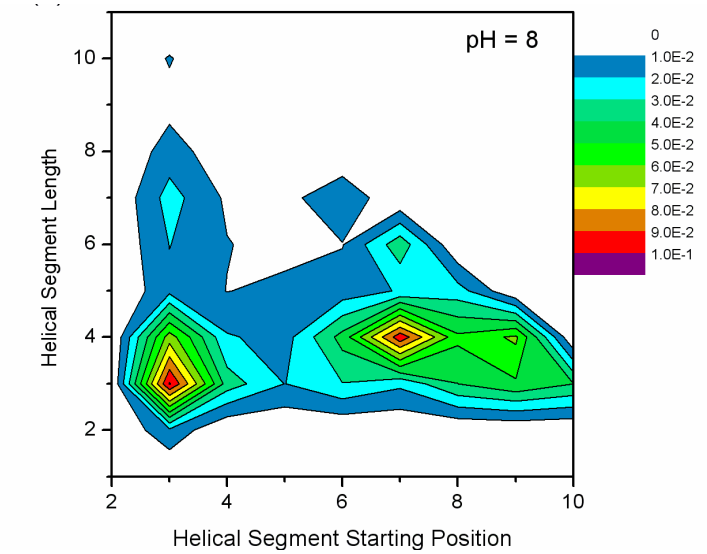


Figure 4-9. 2D probability density of helical starting position and helical length, pH=8.

2D-probability densities correlating helical length and C_α RMSDs relative to fully helical structure are shown in Figures 4-10 to 4-12. As expected, structures having long helices (helical length > 7) correspond to the conformations with RMSDs smaller than 2.2 Å and this region is more populated at pH 5. Interestingly, configurations possessing four-residue helix can also yield RMSDs smaller than 2.2 Å, suggesting that structures having partial helix can be similar to the fully helical too.

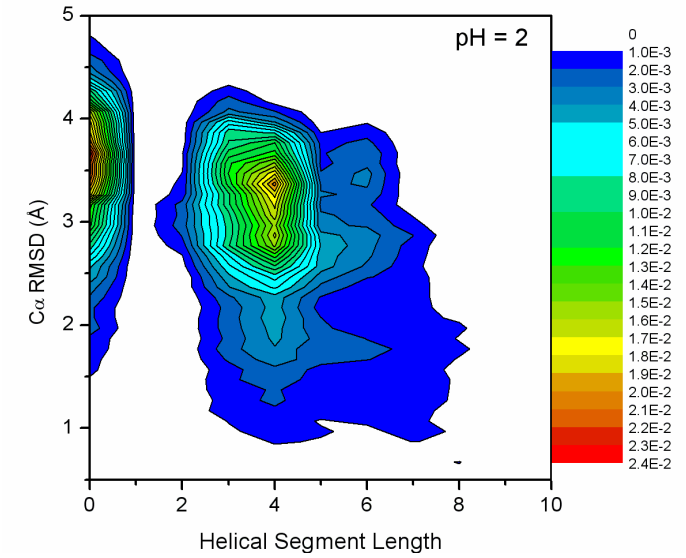


Figure 4-10. 2D probability density of helical length and C_α-RMSD at pH = 2.

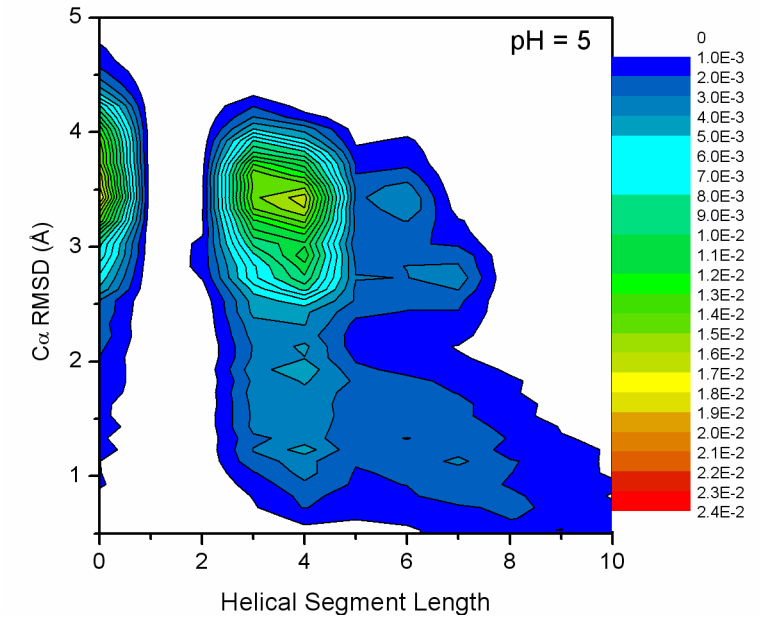


Figure 4-11. 2D probability density of helical length and C_α-RMSD at pH = 5.

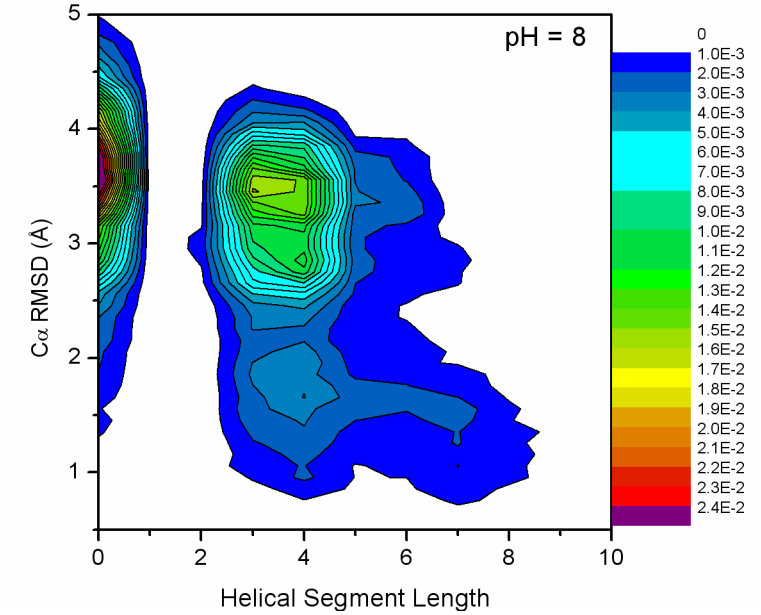


Figure 4-12. 2D probability density of helical length and C_α-RMSD at pH = 8.

4.3.6 Important Electrostatic Interactions: Lys1-Glu9 and Glu2-Arg10

The salt-bridge between Glu2 and Arg10 was found in the X-ray structure of RNase A.[247] Amino acid substitution experiments on the C-peptide indicated this salt-bridge is crucial to the increase in helical content when the pH value is increasing to pH

5.[7,224] Proton NMR experiments done by Osterhout *et al*.[225] suggested that this salt-bridge stabilizes partial helix instead of complete helix. They proposed that the RN-24 structural ensemble contains three major conformations: unfolded, complete folded and partial helix with Glu2-Arg10 interaction. Hansmann *et al*.[229] also proposed that the salt-bridge stabilizes partial helix by performing multicanonical simulations. Felts *et al*.[227] found that the salt-bridge is only significantly found in the globular non-helical C-peptide structures. Sugita and Okamoto[233] studied the C-peptide using multicanonical REM and explicit solvent. They found that Glu2-Arg10 salt-bridge does not stabilize helix directly, but to stop the helix extending to the N-terminus. In the REX-CPHMD study performed by Khandogin *et al.*, they found that Lys1-Glu9, instead of Glu2-Arg10, contributes to the helix formation.

The Lys1-Glu9 and Glu2-Arg10 interactions are studied in our work. Figure 4-13A and 4-13B show the probability density vs charge distance of the two interactions at pH 2, 5 and 8. At pH 2, neither Lys1-Glu9 nor Glu2-Arg10 salt-bridge is formed, consistent with mostly protonated glutamate. At pH 5 and 8, Glu2-Arg10 salt-bridge is clearly formed (Figure 4-13A) while the Lys1-Glu9 salt-bridge is formed in a much less extent (Figure 4-13B). Figure 4-14 shows the correlation between the two salt-bridges at pH 5. Clearly, the two salt-bridges cannot be formed at the same time. The effect of Glu2-Arg10 salt-bridge on helical structure formation can be reflected by conditional probabilities. The probabilities of finding helical residue(s) given that the Glu2-Arg10 salt-bridge is formed are calculated at pH 2, 5 and 8. The conditional probabilities are 0.64, 0.73 and 0.63, respectively. Although at pH 2, the probability of forming Glu2-

Arg10 salt-bridge is low (~ 1%), the chance of having a helical structure is 63% once it is formed. This clearly shows the stabilizing effect of Glu2-Arg10 on helix formation.

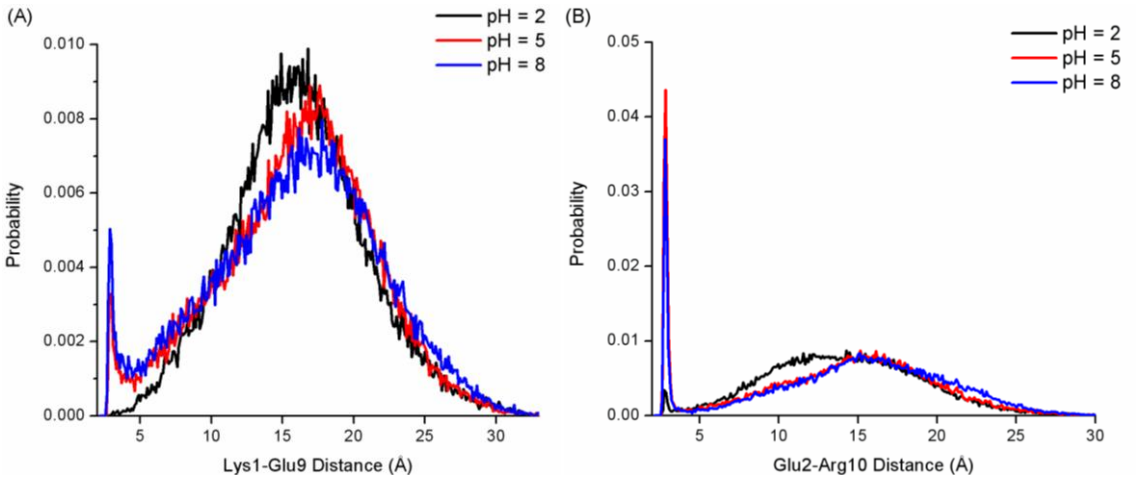


Figure 4-13. A) Probability density of Lys1-Glu9 distance (Å). The distance is the minimum distance between the side-chain nitrogen atom of Lys1 and the side-chain carboxylic oxygen atoms of Glu9. B) Probability density of Glu2-Arg10 distance (Å). The distance is the minimum distance between side-chain carboxylic oxygen atoms of Glu2 and guanidinium nitrogen atoms of Arg10.

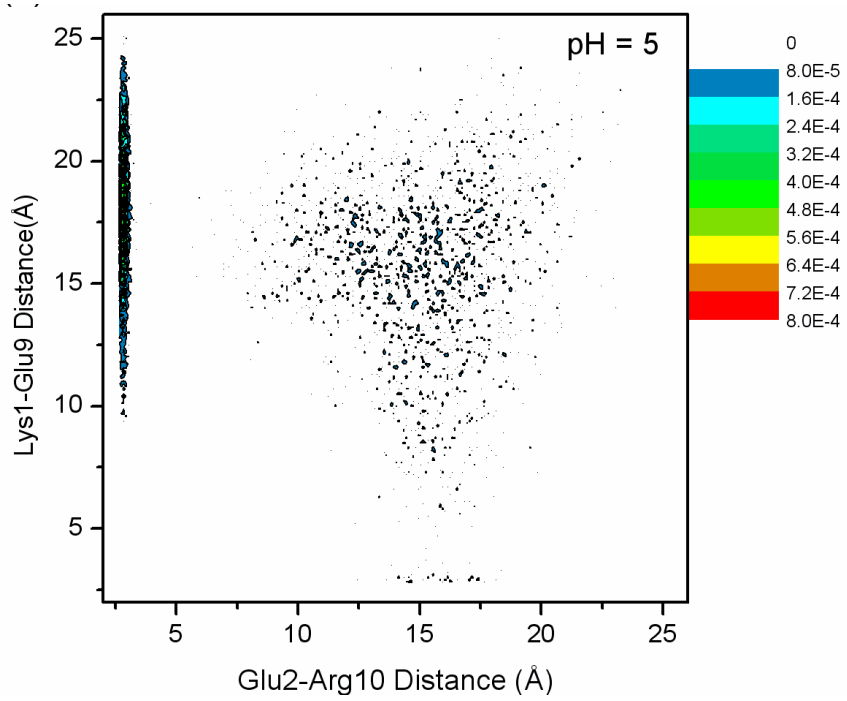


Figure 4-14. Two-dimensional probability density of Lys1-Glu9 and Glu2-Arg10 at pH 5. Apparently, Lys1-Glu9 and Glu2-Arg10 salt-bridges cannot be formed simultaneously.

The correlation between Glu2-Arg10 salt-bridge and helical length, and helix starting position are further studied. Figure 4-15A shows that the Glu2-Arg10 salt-bridge can be found in non-helical configurations, four-residue and six-residue helices at pH 5. Moreover, in the six-residue helix, the Glu2-Arg10 salt-bridge is always formed. The same pattern is obtained at pH 8, thus the pH 8 results are not shown here. Figure 4-15B shows the correlation between the salt-bridge and helix starting position at pH 5. When a helix is initiated at Thr3, the salt-bridge is not formed. When a helix begins at Ala4, Lys7 and residues behind Lys7, only the salt-bridge is seen. However, in the non-helical configurations and helices begin at Ala6, both states are found. Besides, Lys7 is the most probable place to initiate a helix when the salt-bridge is formed. Again, no salt-bridge is found when a helix starts at Thr3. Combining the correlations between Glu2-Arg10 and helical length, and Glu2-Arg10 and helix starting position, the salt-bridge clearly has the effect that preventing forming helices near the N-terminus and stabilizing partial helix near the C-terminus (Lys7-Arg10 and Lys7-His12).

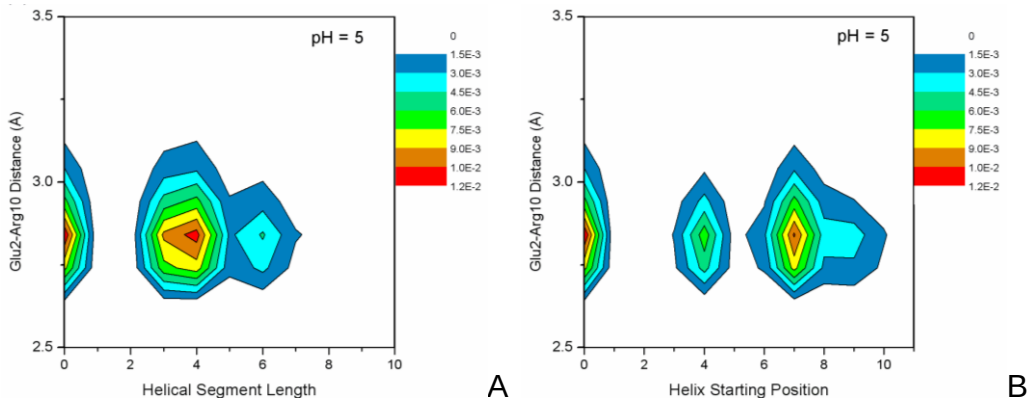


Figure 4-15. A) Two-dimensional probability density of Glu2-Arg10 salt-bridge formation and helical length at pH 5. According to the plot, the Glu2-Arg10 salt-bridge can be found in four-residue, six-residue and non-helical structures. B) Two-dimensional probability density of Glu2-Arg10 salt-bridge and the helix starting position at pH 5. If a helix begins from Thr3, it cannot have a Glu2-Arg10 salt-bridge. Thus, one role of the Glu2-Arg10 salt-bridge is to prevent helix formation from Thr3.

4.3.7 Important Electrostatic Interactions: Phe8-His12

His12 is believed to be responsible for the decrease in helical content when solution pH values increase from 5 to 8.[226] His12 was found to interact with Phe8.[221] However, the nature of the Phe8-His12 interaction is not completely clear. A weak hydrogen bond between the charged side chain of His12 (proton donor) and the aromatic ring of Phe8 (proton acceptor) is supported by the configuration in RNase A X-ray structure[247] and ion screening experiments[222,226] but is in contrast to proton NMR experiments.[221] A contact between the aromatic ring of His12 and backbone carbonyl oxygen of Phe8 has been proposed to explain the proton NMR results. Sugita and Okamoto studied the interaction between the aromatic ring of Phe8 and the charged ring of His12.[233] They observed the contact between two rings has been made and stabilizes helix near the C-terminus. However, the REX-CPHMD results showed that the interaction between backbone carbonyl oxygen of Phe8 and the charged side-chain of His12 is responsible for the increased helical content at pH 5.[112]

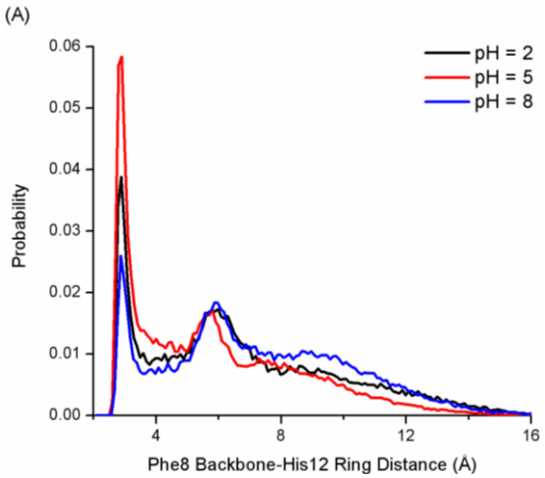


Figure 4-16. A) Probability density of Phe8 backbone to His12 ring distance. The distance is the minimum distance between Phe8 backbone carbonyl oxygen atom and His12 imidazole nitrogen atoms. B) Probability density of Phe8 ring to His12 ring distance. The distance is the minimum distance between Phe8 aromatic ring carbon atoms and His12 imidazole nitrogen atoms.

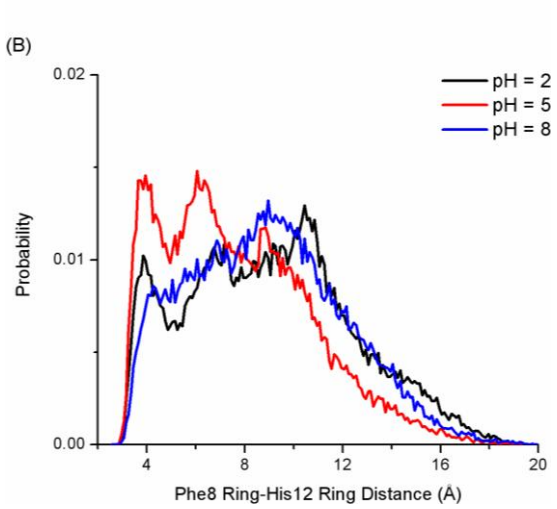


Figure 4-16. Continued

We also studied ring-ring and backbone-ring interactions between Phe8 and His12 at pH 2, 5 and 8. The ring-ring interaction is represented by minimum distance between aromatic atoms in Phe8 and the two side-chain nitrogen atoms of His12. The backbone-ring interaction is represented by minimum distance between backbone carbonyl oxygen atom of Phe8 and the two side-chain nitrogen atoms of His12. Figure 4-16A and 4-16B show the probability densities of each distance at three pH values. We found that the backbone-ring contact is made at all three pH values. However, forming such a contact at pH 8 is much less favorable than doing that at pH 5. Interestingly, Phe8 backbone and His12 ring close contact and Glu2-Arg10 salt-bridge formation are coupled (Figure 4-17). The ring-ring contact is observed at pH 5 but not at pH 8. At pH 2, the ring-ring contact is formed but is much less probable. More importantly, the integrated probability of making a backbone-ring contact is larger than the integrated probability of forming a ring-ring contact at pH 2 and 5. In order to separate configurations making a contact from the rest, a cutoff distance of 4.0 Å and 5.0 Å is adopted, in the case of backbone-ring and ring-ring contact, respectively. The integrated

probability (area under the curve) of making backbone-ring contact and ring-ring contact is 0.34 and 0.22, respectively, at pH 5. The integrated probability is 0.23 and 0.14, respectively, at pH 2. Thus, the Phe8 backbone-His12 ring interaction is the major form of the contact.

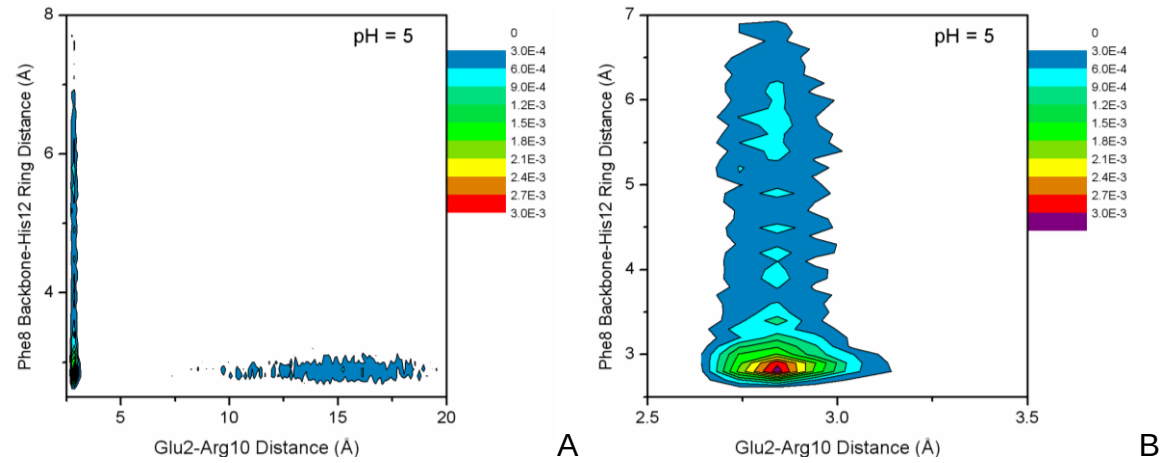


Figure 4-17. A) Two-dimensional probability density of Glu2-Arg10 distance and Phe8-His12 backbone-to-ring distance at pH 5. B) Correlations between Glu2-Arg10 salt-bridge and Phe8-His12 contact at pH 5.

We further examine the correlation between the Phe8 backbone-His12 ring contact and helical properties such as helical length and helix starting position. The backbone-ring contact is found in the four-residue and six-residue helices at pH 2 and 5. At pH 8, it can be seen in the four-residue helix. The 2D probability densities are similar at the three pH values, thus only the plot at pH 5 is shown as an example (Figure 4-18A and 4-18B). Similar to the Glu2-Arg10 salt-bridge, Lys7 is the most favorable place to initiate a helix with a contact between Phe8 and His12. Thus, the Phe8-His12 backbone-ring contact stabilizes the helix formation near the C-terminus (Lys7 to Arg10 and Lys7 to His12). However, unlike the Glu2-Arg10 interaction, helix formation initiated from Thr3 is able to form a contact between Phe8 and His12. Phe8-His12 contact does not affect helix formation near the N-terminus.

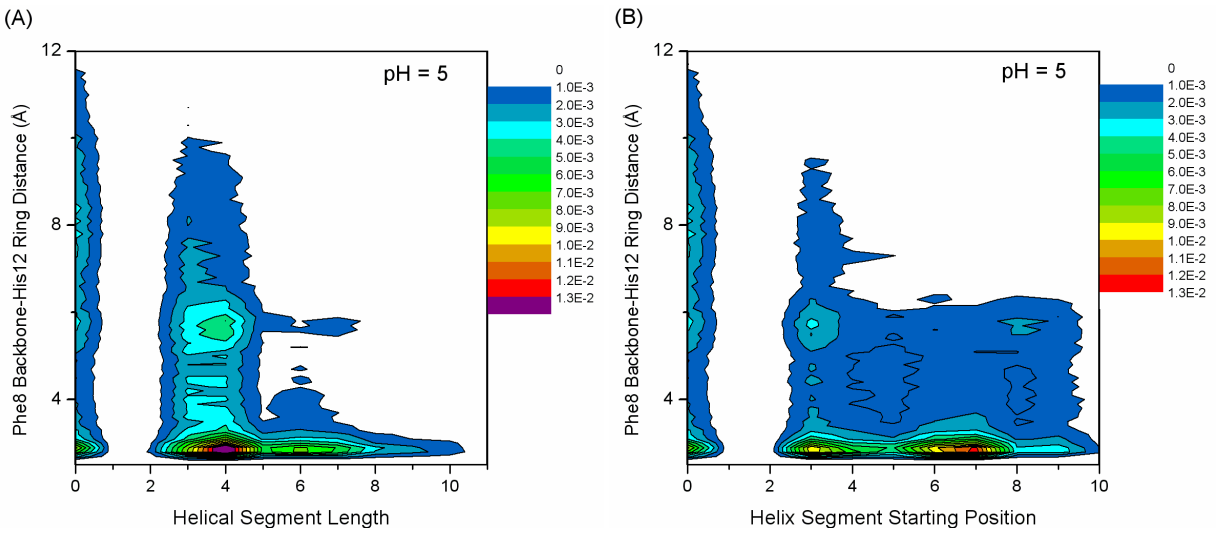


Figure 4-18. A) Two-dimensional probability density of helical segment length and Phe8-His12 interaction. B) Two-dimensional probability density of helical segment starting position and Phe8-His12 interaction. Phe8-His12 also stabilizes four-residue and six-residue structures. Helices begin at Lys7 and Phe8-His12 is coupled. Unlike Glu2-Arg10, Phe8-His12 stabilizes helices starting from Thr3.

4.3.8 Cluster Analysis Results

Cluster analysis is performed to find out significant conformations and to examine important electrostatic interactions. The structures at pH 5 are clustered because both Glu2-Arg10 and Phe8-His12 contacts are more probable than at pH 2 or 8 so that the contacts can be studies in clusters. The top 20 populated clusters and their average helical percentage is plotted in Figure 4-19A. The most populated cluster shows the largest average helical content and the second most populated cluster shows a much lower helical content (close to the lowest among 20 clusters). The most populated cluster corresponds to the conformation yielding small C_α-RMSDs (< 2.2 Å) relative to the fully helical structure (Figure 4-19B). Interestingly, the plot showing helical percentage vs the residue number (Figure 4-19C) reveals that the second most populated cluster only shows helical structures between Lys7 and His12. Thus, helices are only formed near the C-terminus. Figure 4-19D demonstrates the probability density

of the Glu2-Arg10 and Phe8-His12 interactions. Compare with the corresponding probability densities on the basis of the entire structural ensemble, forming a contact between Glu2-Arg10, and Phe8-His12 is more probable in the structures belong to the second most populated cluster than in the entire structural ensemble. This is especially obvious for the Glu2-Arg10 interaction. Results obtained from the second most populated cluster confirm that Glu2-Arg10 and Phe8-His12 contacts, especially the Glu2-Arg10 contact, stabilize partial helix formation near the C-terminus.

4.4 Conclusions

In this chapter, we have studied the pH-dependent helix formation of the C-peptide of ribonuclease A using constant-pH REMD simulations. The mean residue ellipticity at 222 nm at each pH value is computed and utilized to gauge helical content. The pH profile clearly demonstrates a bell-shaped curved with a maximal helicity at pH 5, in good agreement with experimental results. The pH effect on the C-peptide structural ensembles is studied at three representative pH values: 2, 5 and 8, representing the two ends in the pH profile and the pH value yielding the maximum helical content. At pH 2, helices consisting of Thr3-Ala5, Lys7-Arg10 and Glu9-His12 are formed and the Lys7-Arg10 is the most stable one. At pH 5, additional six-residue (Lys7-His12) and seven-residue (Ala6-His12) helices are stable helices but the most probable helix is the same as that at pH 2. At pH 8, the most favorable helix switched to Thr3-Ala5. Lys7-His12 and a new seven-residue helix (Thr3-Glu9) are also present.

Glu2-Arg10 salt-bridge formation and its role in the helix formation are studied. We find that the salt-bridge is formed and is more probable at pH 5. The Glu2-Arg10 salt-bridge is found to stabilize helix formation near the C-terminus. The nature of Phe8-His12 interaction and its role in helix formation are also explored. Backbone carbonyl

oxygen of Phe8 and side-chain charge of His12 contact is the major form. The role of Phe8 and His12 contact is similar to that of the Glu2-Arg10 salt-bridge. Results from cluster analysis on trajectory generated at pH 5 confirmed the effects of Glu2-Arg10 and Phe8-His12 interactions.

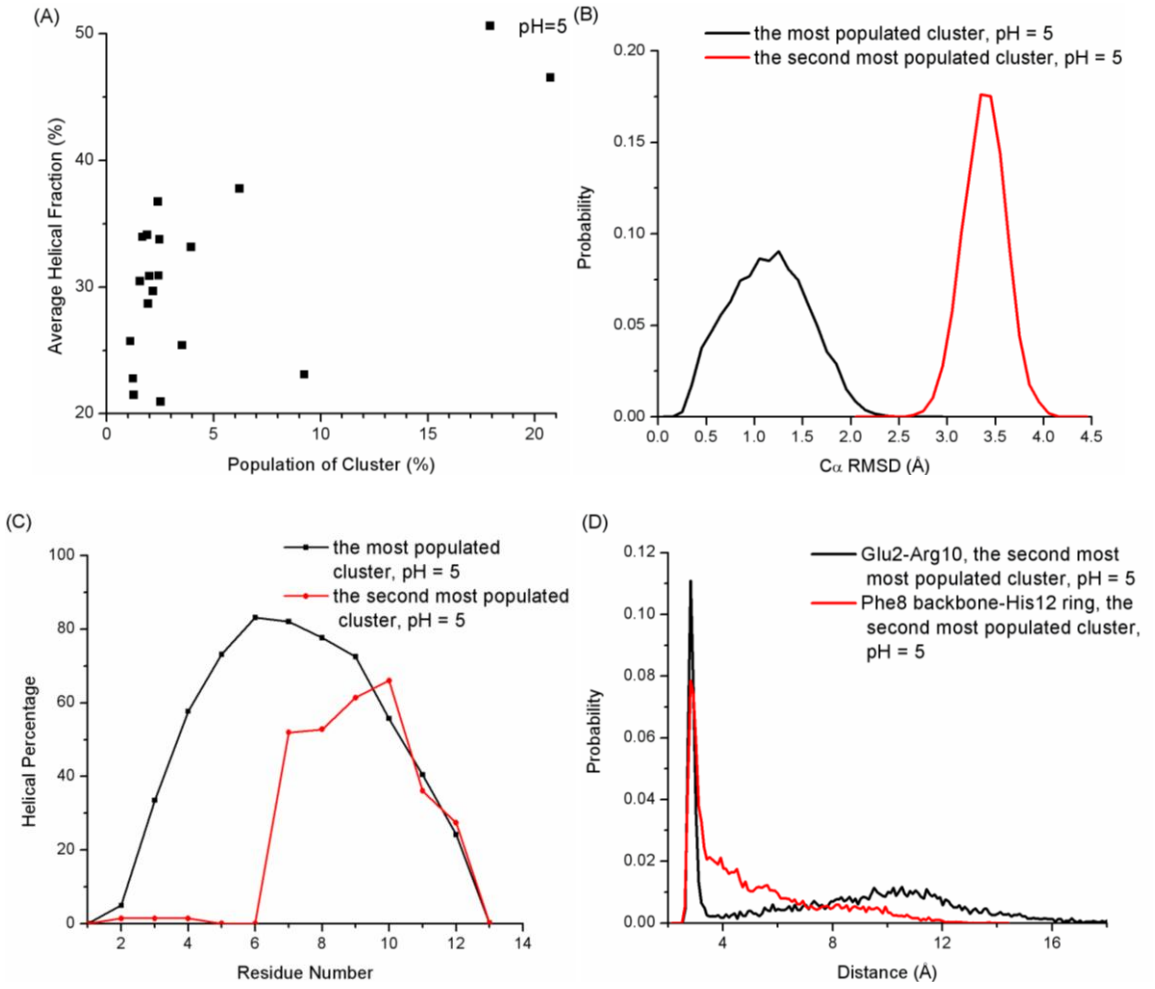


Figure 4-19. A) Top 20 populated clusters and average helical percentage. B) Probability densities of the C_α-RMSD vs the fully helical structure of the top 2 populated clusters. C) Helical Percentage as a function of residue number of the top 2 populated clusters. D) Probability density of the Glu2-Arg10 and Phe8 backbone-His12 ring interactions in the second most populated cluster.

CHAPTER 5
CONSTANT-pH REMD: pK_a CALCULATIONS OF HEN EGG WHITE LYSOZYME

5.1 Introduction

Hen egg white lysozyme (HEWL, shown in Figure 5-1) has been selected to test pK_a prediction methods or constant-pH methods for a long time.[125] This protein is a 129 amino acids enzyme and is the first enzyme to have its three-dimensional structure determined by X-ray crystallography.[248,249] Lysozyme can be found in the secretions such as tears and saliva. The function of this enzyme is to catalyze the hydrolysis of a polysaccharide and the reaction has an optimal pH around 5.[125] By hydrolyzing polysaccharides, lysozyme can damage the cell walls of certain bacteria. HEWL is a monomeric single-domain enzyme whose active site is situated in a cleft between two regions. Two residues are crucial to the catalysis, Glu35 and Asp52. During the hydrolysis, a covalent enzyme-substrate intermediate is formed.[249] In this process, Glu35 acts as the proton donor and Asp52 becomes the nucleophile.[249] The starting point of the catalytic mechanism is the donation of a proton from Glu35 to the substrate. Then, Asp52 will attack the anomeric carbon of the substrate and form a covalent bond with the substrate. In the final step, the enzyme-substrate complex is hydrolyzed by a water molecule and the initial protonation states of Glu35 and Asp52 are restored.

HEWL has been a good test system of pK_a prediction studies for several reasons. First, accurate predicting the pK_a values of both ionizable residues in active site can help people identify proton donor and nucleophile in HEWL according to a simple criterion proposed by Nielsen and McCammon in 2003.[250] They proposed that if catalytic mechanism involves two acidic residues, then the proton donor should have a pK_a value of at least 5.0 and the pK_a of nucleophile should be at least 1.5 pH units lower

than that of proton donor. Second, the pK_a values of HEWL acidic residues were determined by Bartik *et al*.[251] using two-dimensional proton NMR. It shows several ionizable residues having pK_a values much different from their intrinsic pK_a values. Furthermore, there are more than 100 PDB entries of the wild-type HEWL structure, the effect of structural variation can be tested for pK_a calculation methods, especially for the FDPB method.[250] Thus, our constant-pH REMD method will be tested on HEWL.

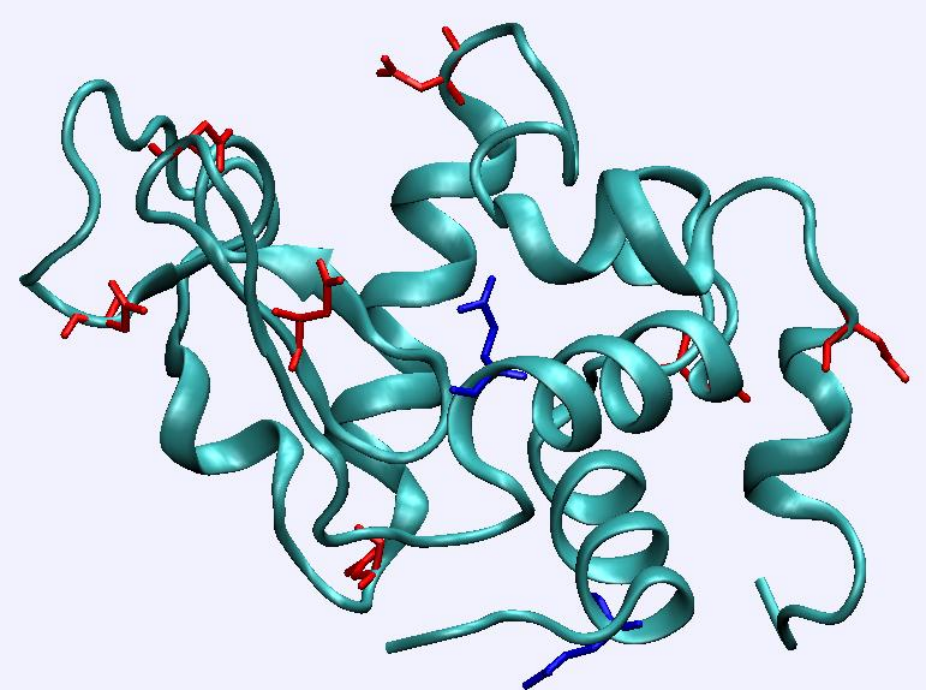


Figure 5-1. Crystal structure of HEWL (PDB code 1AKI). Residues in red represent aspartate and residues in blue are glutamate.

Various constant-pH methods have been tested on HEWL. Burgi *et al*.[130] utilized their constant-pH method to predict pK_a values of HEWL. The RMS error between predicted and experimental pK_a values was determined to be from 2.8 to 3.8 pH units. In 2004, Lee *et al*.[114] applied their CPHMD method to four proteins: turkey ovomocoid (PDB code 1OMT), bovine trypsin inhibitor (1BPI), HEWL (193L) and ribonuclease A (7RSA). The overall pK_a RMS error relative to experimental data was around 1 pH unit.

For HEWL, the average absolute error of all ionizable residues (including the termini) was 1.6 pH units, while the average absolute error of pK_a values of acidic ionizable residues relative to experimental data was 1.5 pH units. However, the pK_a values of Glu35 and Asp52 were both 5.8, indicating that CPHMD results were not able to predict proton donor and nucleophile. In the same year, Mongan *et al*.[127] published their discrete protonation state constant-pH MD method. HEWL was also selected as the test system. In the study of performed by Mongan *et al*., four different crystal structures of HEWL were utilized (1AKI, 1LSA, 3LZT, and 4LYT). The RMSD of pK_a values of all ionizable residues relative to experimental results were 0.86, 0.77, 0.88, and 0.95 for 1AKI, 1LSA, 3LZT, and 4LYT, respectively. In addition to pK_a predictions, Mongan *et al*. also studied protonation-conformation correlation. Principal component analysis of a trajectory was conducted and projected onto the first two (largest eigenvalues) eigenvectors and association between conformation and protonation was observed. In 2006, Khandogin and Brooks[110] utilized REX-CPHMD method to predict pK_a values of 10 proteins. The RMS error values between REX-CPHMD and experimental pK_a values ranged from 0.6 to slightly greater than 1 pH unit. For HEWL, the RMS error between predicted and experimental pK_a values was 0.6 pH unit and the maximum absolute error is 1.0 pH unit. So far, their HEWL pK_a prediction RMS error is the smallest among constant-pH pK_a calculations on HEWL. Machuqueiro and Baptista presented HEWL pK_a predictions from their stochastic titration constant-pH MD with explicit water model in 2008.[125] The RMS error between predicted and experimental pK_a values were 0.82, and 1.13 for generalized reaction field,[252] and PME[154] treatment of long-range electrostatics, respectively. A comparative FDPB calculation (single crystal structure,

which is the same as that utilized in constant-pH MD, and a protein dielectric constant of 2) was also conducted and the RMS error was found to be 2.76. Since the constant-pH method proposed by Baptista requires FDPB calculation, the selection of dielectric constant inside the protein was crucial. Machuqueiro and Baptista performed constant-pH MD utilizing three different dielectric constants (ε=2, 4, and 8) combined with PME treatment of long-range electrostatics. The pK_a RMS error values were 1.13, 1.02, and 1.12 for ε =2, 4, and 8, respectively. More recently, the constant-pH MD proposed by Mongan *et al*.[127] was coupled with accelerated molecular dynamics (AMD)[133,134] and tested on HEWL by Williams *et al.*[129] Constant-pH AMD and MD simulations of 5 ns in length have been performed. Only acidic ionizable residues in HEWL were taken into consideration by constant-pH scheme. RMS error values between predicted and experimental pK_a values were calculated. The constant-pH AMD yielded an overall RMS error value of 0.73, while the original constant-pH MD pK_a RMS error was 0.80. The pK_a RMS error of aspartates were 0.75, and 1.46 from constant-pH AMD, and MD, respectively. The pK_a RMS error of glutamates were 0.85, and 1.04 from constant-pH AMD, and MD, respectively. In general, recent works utilizing various constant-pH schemes have achieved RMS error values in the range of 0.6~1.13 for HEWL.

In this chapter, we present a study of HEWL using constant-pH REMD algorithm. Both structural restrained and unrestrained simulations were done. pK_a values from constant-pH REMD are compared with experimental values. We also investigated the pK_a convergence, effect of structural restraint and conformation-protonation correlations.

5.2 Simulation Details

Crystal structure 1AKI (PDB code) has been taken as HEWL starting structure in our study. Water molecules in the crystal structure were striped first. Only aspartate and glutamate residues were studied so there are nine ionizable residues selected. Hydrogen atoms were added by the LEaP module in the AMBER suite. The post-processed crystal structure was then minimized and heated from 0 K to 300 K. The restart structure from the heating process was taken as the initial structure for our constant-pH REMD simulations. In this chapter, all REMD runs refer to constant-pH REMD simulations for simplicity. The pH range was from 2 to 6 in an increment of 0.5 pH unit.

Two sets of REMD simulations were performed: the unrestrained ones (ntr=0 in AMBER) and the restrained ones (ntr=1 in AMBER). In each REMD run, an exchange of structures was attempted every 500 MD steps. 1000 exchange attempts were intended to use for both sets. Thus Simulation time of each replica in each set is 1 ns. In the unrestrained REMD runs, we chose the highest temperature to be 320 K in the hope that HEWL will not unfold at all temperatures. In the restrained REMD runs, C_α atoms from residue 3 to 126 were restrained by harmonic potentials. The restraining harmonic potential has the following form: $U_{res} = \frac{1}{2}k\left(\vec{q}_t - \vec{q}_{ref}\right)^2$, where $\vec{q}_t$ and $\vec{q}_{ref}$ are Cartesian coordinates at current time and Cartesian coordinates of the reference structure, respectively, k is the force constant of the harmonic potential which determines the strength of a restraint. In our simulations, the reference coordinates are the initial C_α atoms coordinates. By putting restraining harmonic potential on C_α atoms, the secondary structure of HEWL will be preserved and the highest temperature will be

increase to 420 K in order to achieve better side-chain conformational sampling. The force constant of the harmonic potentials was 1.0 kcal/mol·Å^2 (setting restraint_wt=1 in AMBER).

Several other REMD simulations were done according to results from the two sets of REMD runs. The general goal of those simulations was to test what we proposed from the two previous sets. First, another 1 ns constant-pH REMD simulation with restraint on C_α atoms was continued for all the pH values in order to check the pK_a convergence of the restrained simulations. Likewise, 1000 exchange attempts were conducted in those 1 ns simulations and the restraint strength is still 1.0 kcal/mol·Å^2. Second, a new set of constant-pH REMD simulations with restraint on C_α atoms was performed. The force constant adopted in the second set was 0.1 kcal/mol·Å^2 so that the effect of restraint strength can be tested. The details of constant-pH REMD simulations can be found in Table 5-1.

Table 5-1. Simulation details of constant-pH REMD runs

pH values	Restrained or not	Restraint Strength	Number of Replicas	Temperature (K)	Simulation Time (ns)	Exchange Attempts
2~6	No	0	4	280~320	1	1000
2~6	Yes	1	8	280~420	2	2000
3, 4, 4.5	Yes	0.1	8	280~420	2	2000

The restraint strength was represented by the force constant of a harmonic potential. The unit of force constant is kcal/mol·Å^2. For the REMD simulation with 1 kcal/mol·Å^2 restraint, it was actually performed in two stages. Each stage lasted for 1 ns and the purpose of the second stage was to check the pK_a convergence.

All simulations were done using the AMBER 9 molecular simulation suite[253] with the AMBER ff99SB force fields.[139] The SHAKE algorithm[145] was used to allow a 2 fs time step. OBC Generalized Born implicit solvent model[200] was used to model water

environment in all our calculations. The Berendsen thermostat,[146] with a relaxation time of 2 ps, was used to keep the replica temperature around their target values. Salt concentration (Debye-Huckel based) was set at 0.1M. The cutoff for nonbonded interaction and the Born radii was 30 Å.

5.3 Protein Conformational and Protonation State Equilibrium Model

Suppose an ionizable side chain has only two conformations in equilibrium and each conformer has its own equilibrium in protonation state. We can use 1p, 1d, 2p and 2d to label conformer 1 in protonated form, conformer 1 in deprotonated form, conformer 2 in protonated form, and conformer 2 in deprotonated form, respectively. The equilibrium among all species is demonstrated in Figure 5-2.

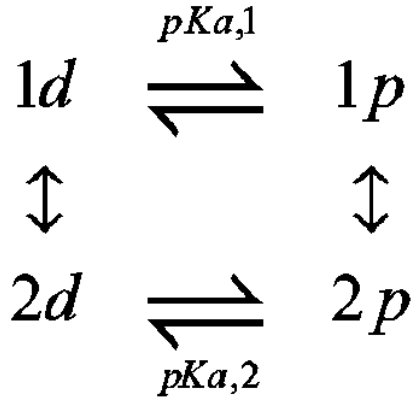


Figure 5-2. A simple schematic view of the conformation-protonation equilibrium in a constant-pH simulation.

Then, K_{12}, the equilibrium constant between conformation 1 and 2 is

$$K_{12} = \frac{[1p]+[1d]}{[2p]+[2d]} \tag{5-1}$$

In the above model, p$K_{a,1}$ and p$K_{a,2}$ represent protonation equilibrium within each conformation. They can be expressed as:

$$pKa,1 = pH - \log(\frac{[1p]}{[1d]}) \tag{5-2}$$

and

$$pKa,2 = pH - \log(\frac{[2p]}{[2d]}). \quad (5\text{-}3)$$

So, the pK_a of that ionizable residue is

$$pKa = pH - \log(\frac{[1p]+[2p]}{[1d]+[2d]}) \quad (5\text{-}4)$$

5.4 NMR Chemical Shift Calculations

Theoretical NMR chemical shift titration curve was generated. Due to the limitation of system size, full quantum mechanical NMR calculations were performed only on ionizable residue dipeptide (ionizable residue with two ends blocked). The structure of ionizable dipeptide was extracted from the representative structures (representing different side chain conformations) generated from cluster analysis. Proper protonation states were assigned for each structure. All full quantum mechanical NMR calculations were done in Gaussian03 software package[254] using B3LYP functional and 6-311++G** basis set. Isotropic magnetic shielding constants were computed in vacuum using GIAO method.[255] Tetramethylsilane (TMS) was used as reference in order to obtain the chemical shift.

Recently, Merz and co-workers[256] developed an automated fragmentation quamtum mechanical/molecular mechanical (AF-QM/MM) approach to study protein properties. They have applied their method to compute protein chemical shift of Trp-Cage. In this AF-QM/MM model, one residue and the atoms near it (less than 4 Å) are assigned to the QM region and the rest of a protein will be put into the MM region. During NMR calculations, all atoms in the MM region will be viewed as point charges.

We applied this AF-QM/MM method to 1AKI to calculate chemical shift as well. Again, all AF-QM/MM calculations were based on representative structures.

5.5 Results and Discussions

5.5.1 Structural Stability and pK_a Convergence

Since changing protonation state during simulation will cause discontinuity in force and energy, structural stability in our simulations is important. We chose to use C_α atoms root-mean-square deviation (RMSD) vs 1AKI structure as our metric. Figure 5-3A shows us the C_α RMSD vs time in unrestrained REMD runs. In Figure 5-3A, HEWL is instable at all the pH simulated. The RMSD can reach a very high value (~ 18 Å) during simulations. Even at pH=4 where C_α RMSD values are small relative to the rest, the C_α RMSD can still go beyond 3 Å. pK_a predictions from unrestrained REMD runs shouldn't be used.

Figure 5-3B shows the RMSDs in the restrained REMD runs. Although the RMSD values are small and stable throughout 2 ns simulations, the restrained REMD simulations still reveal problems, according to Figure 5-3B. Our simulations use 1AKI which is resolved at pH=4.5 as starting structure. As pH is moving away from 4.5, one may expect HEWL will adopt conformations a little bit different from 1AKI. So a bigger RMSD should be expected where the pH value is far away from 4.5. This behavior has been confirmed in the work of Mongan *et al*. However, putting restraint on C_α atoms results in the same RMSDs in the entire pH range. This may have negative effect on pK_a predictions at pH values far away from 4.5.

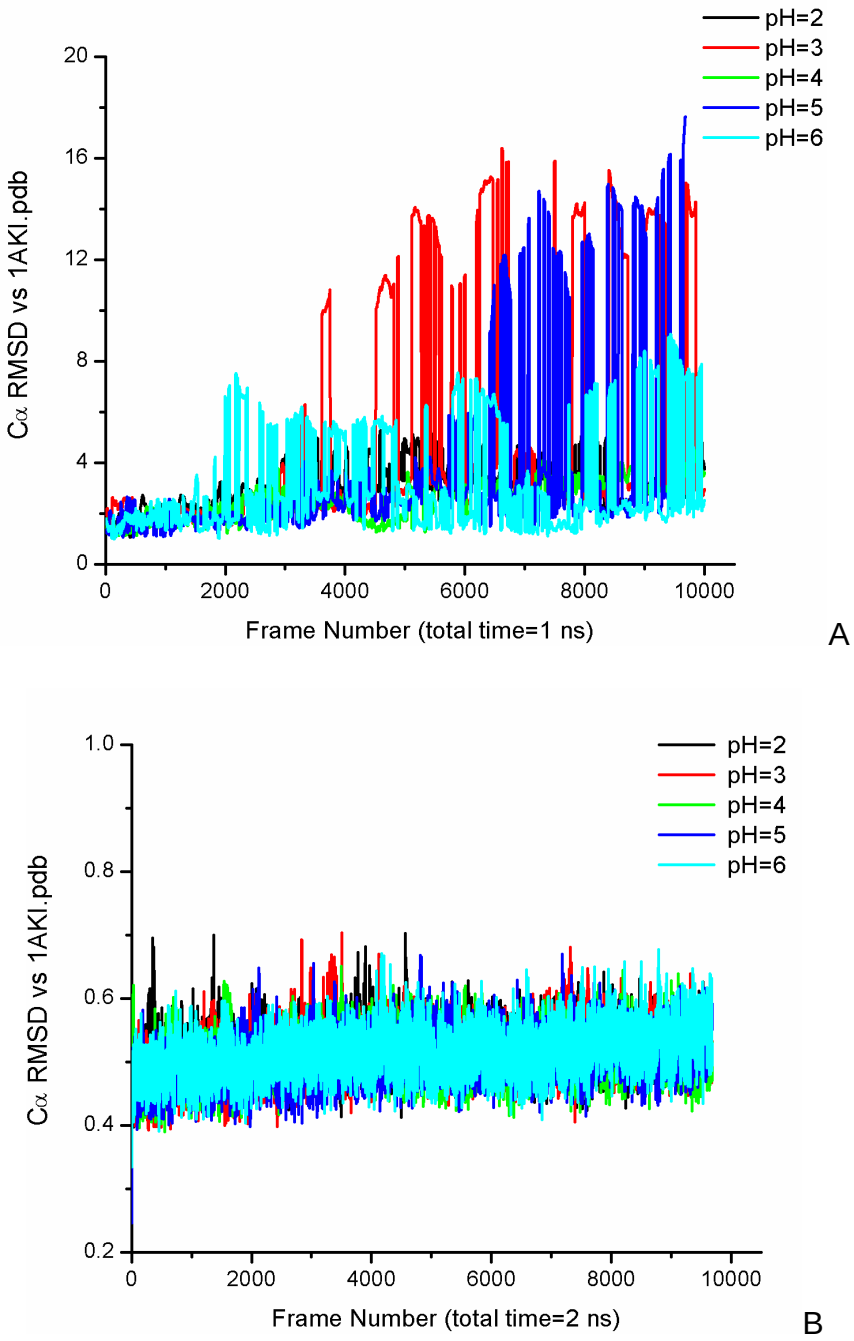


Figure 5-3. C_α RMSD vs crustal structure (PDB code: 1AKI). A) C_α RMSD vs 1AKI from REMD without restraint on C_α. B) C_α RMSD vs 1AKI from REMD with restraint on C_α. The restraint strength is 1 kcal/mol·A^2.

In order to check protonation state sampling convergence from the restrained REMD simulations, pK_a prediction error (predicted value minus experimental value) against time as well as time evolution of prediction deviation (predicted pK_a value at

current time minus the final predicted pK_a value) are followed and demonstrated in Figure 5-4 and 5-5. According to those plots, stabilizations in pK_a predictions are seen after a few hundred picoseconds of simulations. Increasing simulation time wouldn't change average pK_a predictions and their errors relative to experimental values. In order to show convergence in protonation state sampling is reached in a wide range of pH, a representative plot of Asp52 pK_a deviations are shown in Figure 5-5B. Convergence is clearly seen over the pH range.

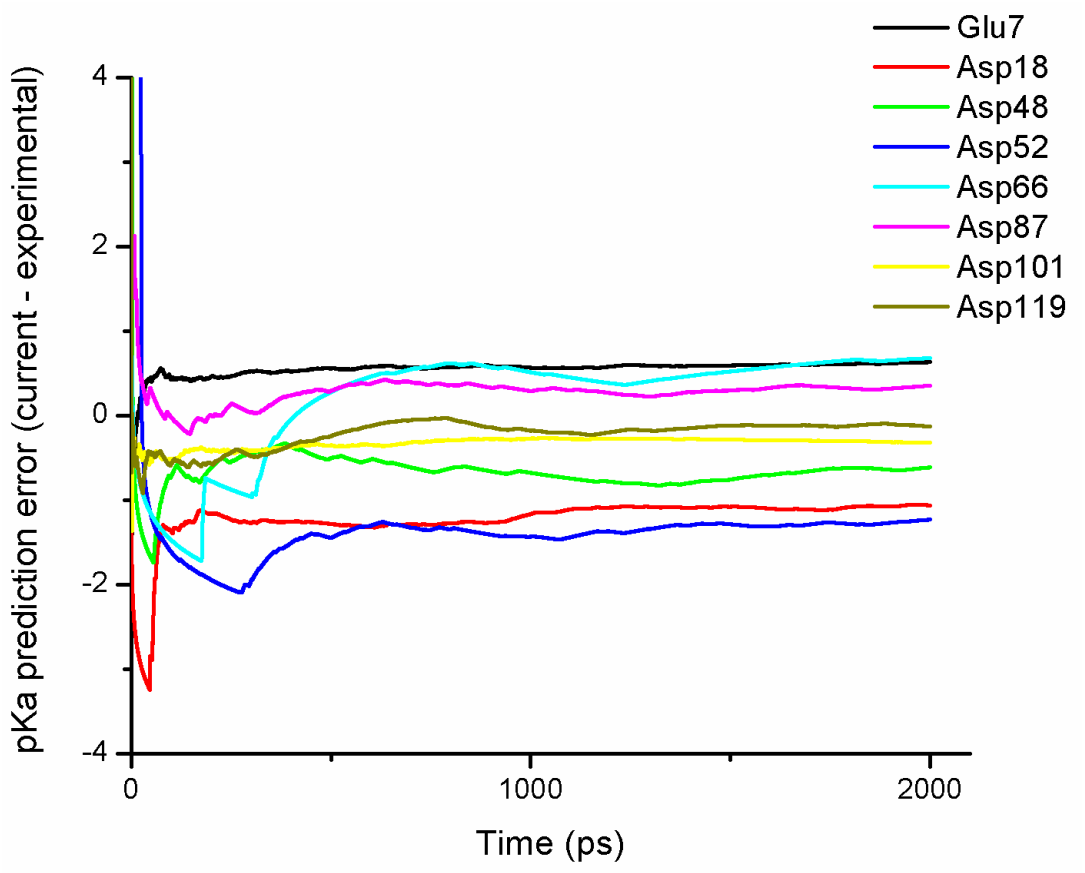
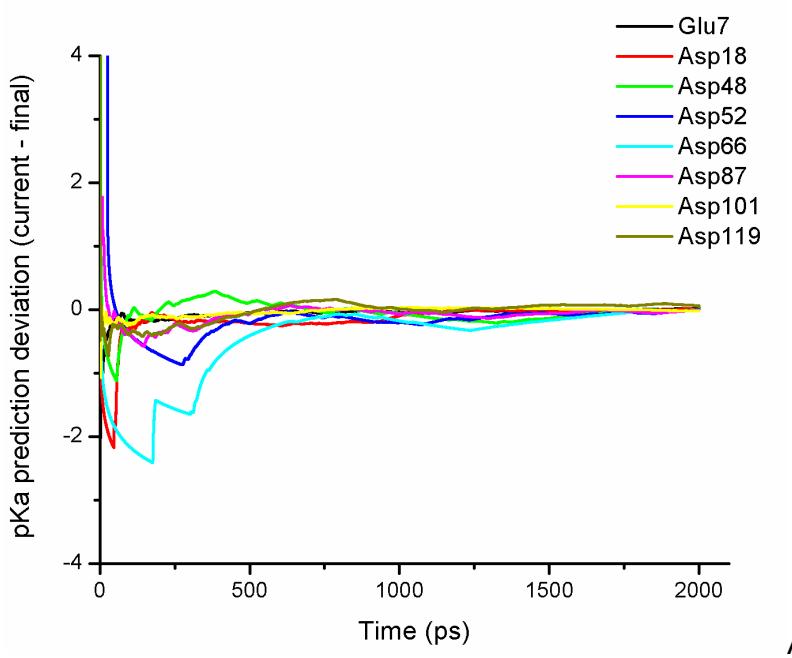
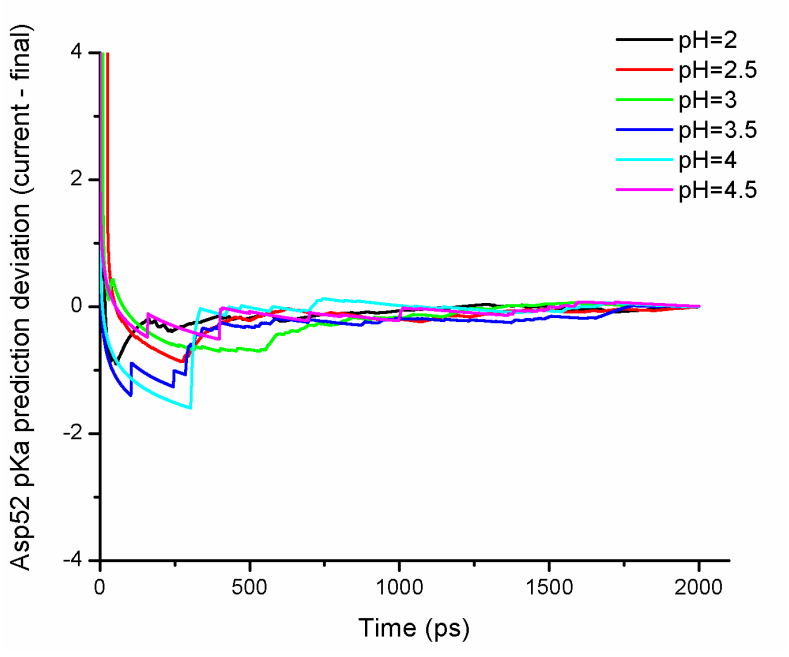


Figure 5-4. pK_a prediction error as a function of time. The predicted pK_a at a given time is a cumulative result. For each ionizable residue, the time series of its pK_a error is generated at a pH where the average predicted pK_a is closest to that pH value. In this way, we try to eliminate any bias toward the energetically favored state. A flat line is an indication of convergence. Glu35 is not shown here due to poor convergence.



A



B

Figure 5-5. A) pK_a prediction convergence to its final value. Similarly, the pK_a value at a given time is a cumulative average. A flat line having y-value of 0 is expected when pK_a calculation convergence is reached. The same pH values are chosen for each ionizable residue as in Figure 5-4. B) Asp52 pK_a prediction convergence to its final value at multiple pH values. The pH values are selected in such a way that the pK_a calculated at this pH will be used to compute composite pK_a.

5.5.2 pK_a Predictions

A popular way to study the accuracy of pK_a prediction is to look at the pK_a RMS error relative to experimentally measured pK_a values. In general, a Hill's plot is used to generate pK_a for each ionizable residue because Hill's plot can combine results from all simulations. Mongan *et al.* proposed a way to calculate pK_a without using Hill's plot in their constant-pH MD paper. They called pK_a values calculated in their way composite pK_a values. A composite pK_a is an average of all pK_a values having an absolute offset less than 2 pH units. Here an offset means the difference between predicted pK_a and its corresponding pH values.

Table 5-2 shows pK_a values and the pK_a RMS error values from the 2-ns restrained REMD runs. Composite pK_a values, pK_a values obtained from Hill's plots and their RMS error values relative to experimental measurements are also listed in Table 5-2. We used the same experimental pK_a values as Mongan *et al.* did to calculate pK_a RMS error. In our work, the pK_a predictions from Hill's plots yield a RMSD value of 0.84, while utilizing composite pK_a values produces a RMS error value of 0.87. According to constant-pH simulation literatures, the RMS error values of HEWL pK_a prediction are around 0.8 for acidic ionizable residues. So there is no significant improvement in pK_a prediction from our simulations.

However, as we mentioned in the structural stability discussion, putting a restraint on C_α atoms of a protein lowers the ability to adjust its conformations. The further a pH value is away from crystal pH, the more a structure ensemble is skewed from the correct one. Simulations performed around pH 4.5 are less affected by the restraint than simulations done at pH values far away from 4.5. Since the less a structural ensemble is skewed, the less human error in pK_a predictions. So one may expect smaller pK_a RMS

error relative to experimental values will be seen around pH 4.5. pK_a prediction RMS error relative to experimental values are plotted against pH values in Figure 5-6. As expected, a minimum having RMS error of 0.74 at pH 4.5 can be found. An RMS error of 0.74 is among the best published HEWL predictions.

Table 5-2. Predicted pK_a values and their RMS errors relative to experimental measurements from the restrained REMD simulations.

	Exp[251]	pH 2	pH 2.5	pH 3	pH 3.5	pH 4	pH 4.5	pH 5	pH 6	Comp	Hill
Glu7	2.85	3.61	3.58	3.46	3.03	2.99	2.93	2.36	3.37	3.27	3.23
Asp18	2.66	1.59	1.54	1.51	1.61	1.91	2.35	2.5	3.69	1.63	1.4
Glu35	6.2	3.76	3.65	4.36	4.14	4.31	4.53	4.76	4.61	4.27	4.58
Asp48	2.5	1.88	1.98	2.14	2.34	2.6	2.45	1.96	2.9	2.23	2.01
Asp52	3.68	2.71	2.45	2.63	2.82	3.05	2.72	2.77	3.99	2.73	2.68
Asp66	2.0	2.5	2.69	2.86	2.92	3.12	2.72	3.09	4.04	2.8	2.73
Asp87	2.07	2.32	2.43	2.64	2.49	2.54	2.64	2.79	3.62	2.51	2.42
Asp101	4.09	4.52	4.4	4.14	4.03	3.79	3.55	3.44	3.96	3.89	3.85
Asp119	3.2	2.71	2.78	3.01	3.01	3.25	3.01	2.89	3.97	2.96	2.9
RMS Error		1.04	1.1	0.91	0.89	0.83	0.74	0.79	1.12	0.87	0.84

In this table, “Exp” stands for experimental pK_a values. “Comp” stands for the composite pK_a value of an ionizable residue (see Mongan’s paper for definition) and “Hill” stands for the pK_a value obtained from the Hill’s plot. The force constant of the harmonic potential used here is 1.0 kcal/mol·Å^2.

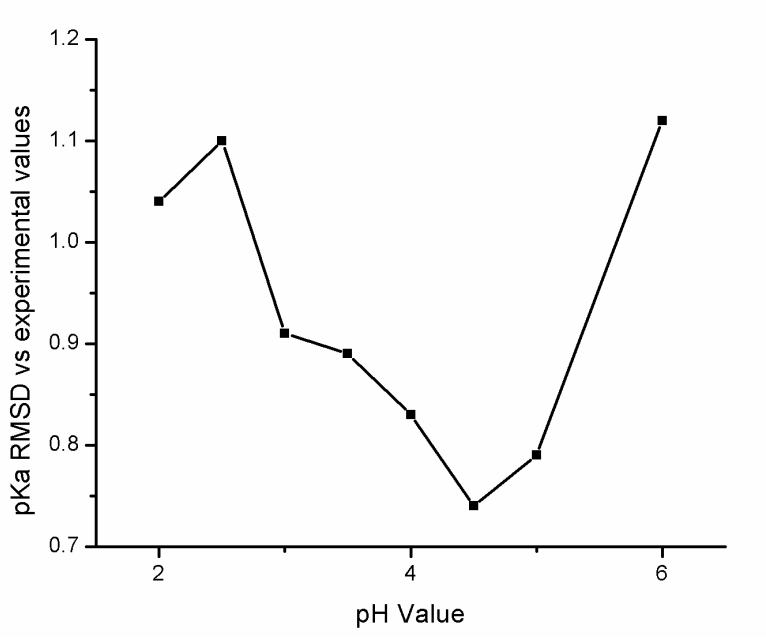


Figure 5-6. RMS error between predicted and experimental pK_a vs pH value. A minimum of pK_a RMS error can be found near the pH at which 1AKI crystal structure is resolved.

5.5.3 Constant-pH REMD Simulations with a Weaker Restraint

Based on what have been found so far, we propose that reducing restraint strength on C_α atoms will yield better pK_a predictions. This is because reducing restraint strength will increase degree of freedom in conformation sampling. HEWL can relax its structure further, even at pH 4.5. Thus a more accurate structure ensemble can be produced. This, in turn, will improve pK_a calculations. Constant-pH REMD simulations with a weaker restraint (harmonic potential on C_α atoms) of 0.1 kcal/mol·Å^2 were carried out at three different pH values to test our hypothesis. First, as shown in Figure 5-7A, all three simulations generate larger C_α RMSDs relative to 1AKI than those simulations with stronger restraint do. This means HEWL relaxes more when a weaker restraint is used. Besides, the C_α RMSD fluctuations in all three runs are bigger than those in the 1 kcal/mol·Å^2 REMD runs. This means more conformational space is visited. Another

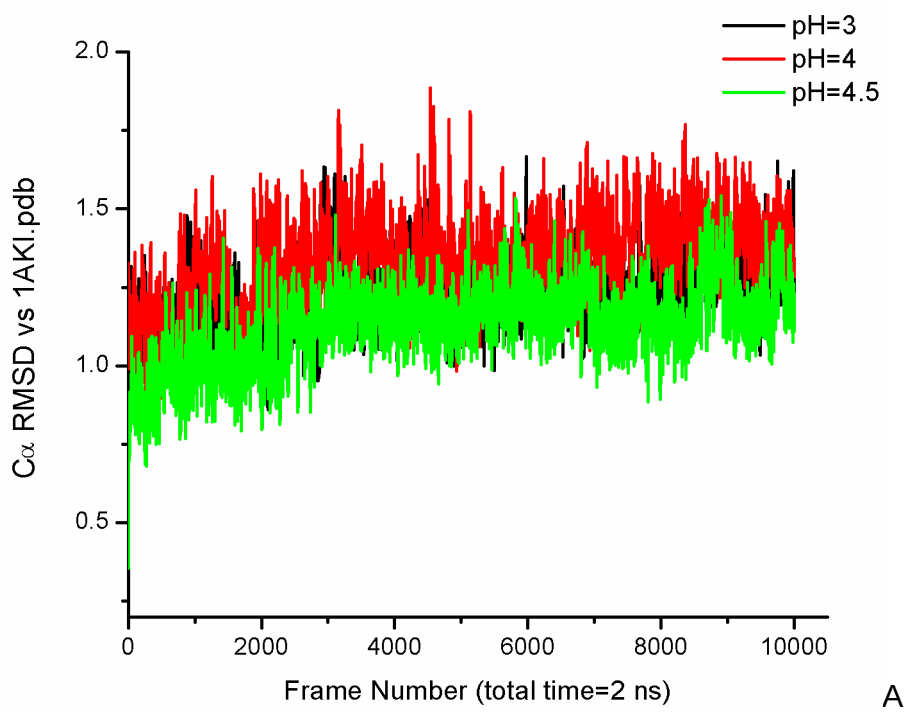
interesting point in the weaker-restrained REMD runs is that the C_α RMSDs at pH 3 and 4 are larger than those at pH 4.5. Simulations at pH 3 and 4 do tend to sample conformations that are different from at pH 4.5.

The pK_a prediction results are listed in Table 5-3. pK_a prediction deviation from the final value vs time at pH value of 4.5 is shown in Figure 5-7B to demonstrate protonation state sampling convergence. According to Table 5-3, nearly 0.1 pH unit improvement in the RMS error of predicted pK_a values can be seen at each pH for the weakly restrained REMD runs. However, among all three RMS error values, the best one is still obtained at pH 4.5 indicating that restraint is still favoring simulations near pH 4.5. After reducing the restraint strength, our best pK_a RMS error relative to experimental values is 0.62.

Table 5-3. Predicted pK_a values and their RMS errors relative to experimental measurements from weakly restrained REMD simulations.

	pH=3		pH=4		pH=4.5	
	1	0.1	1	0.1	1	0.1
Glu7	3.46	3.71	2.99	3.38	2.93	3.34
Asp18	1.51	1.57	1.91	1.76	2.35	2.23
Glu35	4.36	5.09	4.31	5.23	4.53	5.24
Asp48	2.14	2.27	2.6	2.48	2.45	2.71
Asp52	2.63	2.47	3.05	2.88	2.72	3.29
Asp66	2.86	2.63	3.12	2.66	2.72	2.93
Asp87	2.64	2.52	2.54	2.79	2.64	2.88
Asp101	4.14	3.82	3.79	3.77	3.55	3.54
Asp119	3.01	2.22	3.25	2.21	3.01	3.38
RMSE	0.91	0.84	0.83	0.72	0.74	0.62

In Table 5-3, the number 1 in the second row means the force constant of the restraining potential is 1 kcal/mol·Å^2, while 0.1 stands for 0.1 kcal/mol·Å^2. RMSE stands for RMS Error.



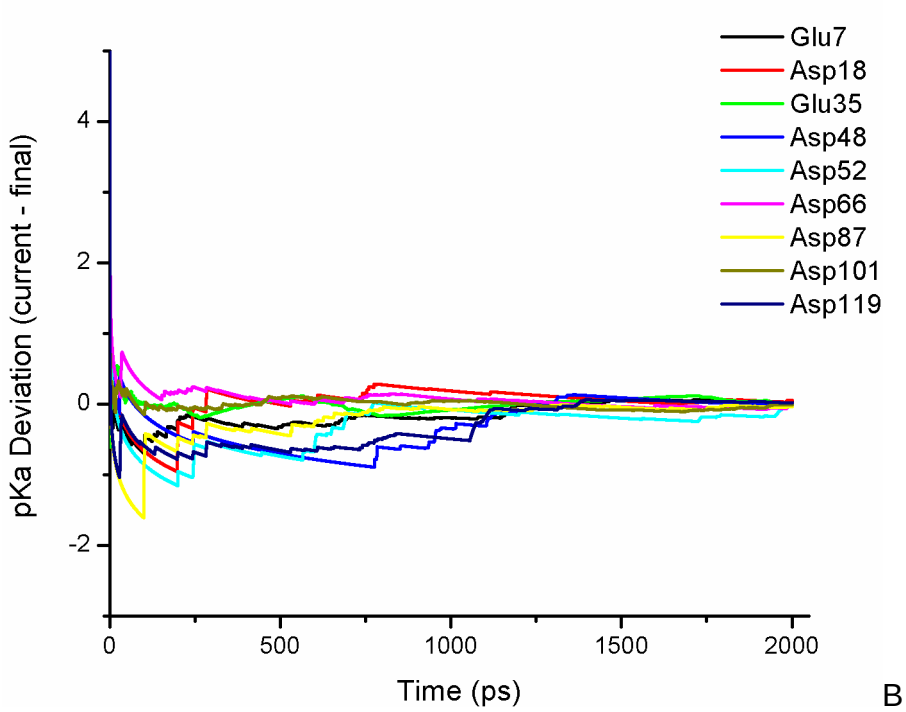


Figure 5-7. A) C_α RMSD of HEWL from weaker restraint REMD simulations. The RMSDs are larger than those with stronger restraints. When comparing RMSDs at different pH for simulations using weaker restraint, RMSDs are greater at pH 3 and 4 than those at pH 4.5. B) pK_a prediction deviation from final value at pH 4.5 from constant-pH REMD with 0.1 kcal/mol·Å^2.

5.5.4 Active Site Ionizable Residue pK_a Prediction: Asp52

Accurate calculations of the pK_a values of ionizable residues in active site are important because their protonation states are crucial in enzyme reactions. In the case of HEWL, Asp52 works as a nucleophile. This requires Asp52 to be deprotonated during reactions which has an optimal pH around 5. In both restrained REMD, Asp52 is indeed deprotonated around pH 5. However, the error of Asp52 relative to experimental value is about 1 pH unit. Mongan and co-workers also had the same trend except that a bigger error was obtained in their simulations. They claimed that Asp52-Asn46 hydrogen bond caused the very low predicted pK_a of Asp52.[127]

Asp52 and residues that strongly interact with it (three asparagine residues: Asn44, Asn46 and Asn59) in the crystal structure of 1AKI (hydrogen atoms are added and proper protonation state is chosen at pH 4.5) are shown in Figure 5-8. We studied those interactions which are represented by atom-to-atom distances in our REMD simulations. We find that Asp52 is closer to Asn59 and Asn44 rather than to Asn46, indicating that Asp52 has stronger interactions with Asn59 and Asn44 than with Asn46. Time series of Asp52 carboxylic oxygen atoms to Asn59 and Asn44 ND2 distances at pH 3 are shown in Figure 5-9. As can be seen from Figure 5-9A and 5-9B, Asp52 and Asn44, Asn59 stay within hydrogen-bonding distance for a long time at pH as low as 3. Furthermore, hydrogen-bonding distances between Asp52 and Asn44, and between Asp52 and Asn59 are coupled. Two oxygen atoms in the carboxylic group of Asp52 are able to work as proton acceptors simultaneously. This means that the deprotonated form of Asp52 is over-stabilized by hydrogen-bonding, even at low pH values.

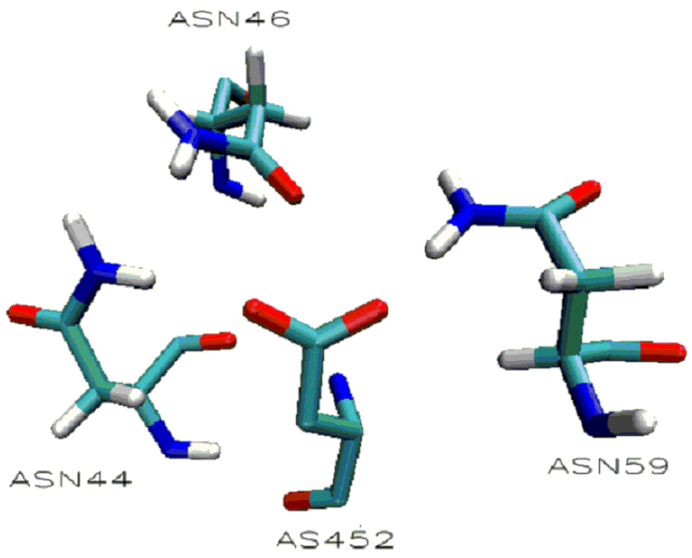


Figure 5-8. Asp52 in the crystal structure of 1AKI. Its neighbors that having strong electrostatic interactions are also shown.

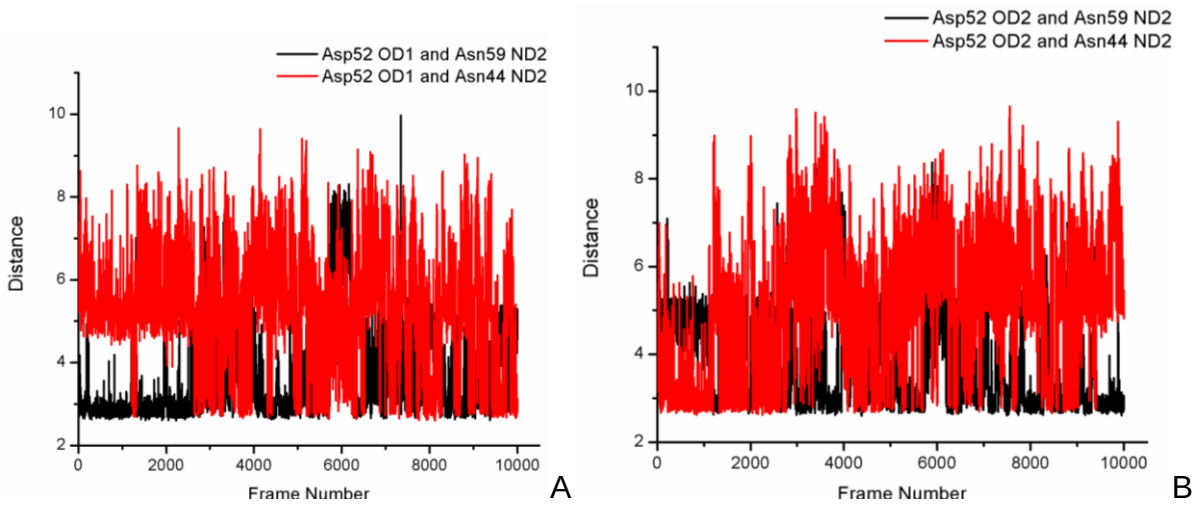


Figure 5-9. A) Time series of Asp52 carboxylic oxygen atom OD1 to Asn59 and Asn44 ND2 distances at pH 3 in the 1 kcal/mol·Å^2 constant-pH REMD run. B) Time series of Asp52 carboxylic oxygen atom OD2 to Asn59 and Asn44 ND2 distances under the same condition. Hydrogen bonds which are stabilizing deprotonated Asp52 are formed in a large extent even at a low pH.

Next, hydrogen bond analysis was conducted with PTRAJ module in the AMBER suite for both sets of restrained REMD simulations. Hydrogen bonds can be found between Asp52 and all three asparagines (Asn44, Asn46, and Asn59) in both sets. The occupation times of Asp52-Asn44 and Asp52-Asn59 hydrogen-bonding are longer than

that of Asp52-Asn46 hydrogen-bonding. Furthermore, the Asp52-Asn44 and Asp52-Asn59 hydrogen-bonding are coupled according to the distances demonstrated in Figure 5-9. Asp52 is protonated only when the entire carboxylic group is pointing away from Asn44 and Asn59. The Asp52-Asn44 and Asp52-Asn59 hydrogen-bonding, not the Asp52-Asn46 hydrogen-bonding, is responsible for low predicted pK_a value of Asp52. The hydrogen bond contents are similar in both strongly and weakly restrained REMD simulations. This indicates that the hydrogen-bonding effect on Asp52 in our simulations is too strong. Reducing restraint strength doesn't help the conformational sampling of Asp52.

5.5.5 Active Site Ionizable Residue pK_a Prediction: Glu35

Glu35 is another problematic case in our study. In the 1 kcal/mol·Å^2 runs, it's the largest single residue error: the error is almost 2 pH units. Excluding Glu35 will lower the pK_a RMS error value by nearly 0.2 pH unit. In the 0.1 kcal/mol·Å^2 runs, the pK_a value of Glu35 is improved, having an error around 1 pH unit. This is the main reason that smaller pK_a RMS errors relative to experimental data are found in all three 0.1 kcal/mol·Å^2 REMD simulations. Although the pK_a error of Glu35 in the weakly restrained REMD simulation is large, the good news for weakly restrained REMD simulations is that Glu35 can be correctly identified as proton donor based on the criterion proposed by Nielsen and McCammon: Glu35 has a pK_a value ~5.2 and the pK_a difference between Asp52 and Glu35 is greater than 1.5 pH units.

The predicted pK_a value of Glu35 was determined to be 5.32 in the study performed by Mongan *et al*. They claimed that a similar hydrogen-bonding effect as Asp52 demonstrated was responsible for the low predicted pK_a value of Glu35.[127] However, hydrogen-bonding analysis of our data does not show any significant

hydrogen-bonding is formed by Glu35, which is in contrary to what Mongan *et al.* claimed.

In the 1AKI crystal structure, Glu35 side-chain is in the vicinity of Gln57, Trp108 and Ala110 side-chains. Several key distances between Glu35 carboxylic group and Gln57, Trp108 and Ala110 side chains in the crystal structure are listed in Table 5-4. According to Table 5-4, Glu35 is in a hydrophobic region except that a close distance between Glu35 OE2 atom and Ala110 backbone amide nitrogen atom. The hydrophobic effect is the main reason of an elevated pK_a value of Glu35. However, when the carboxylic group is pointing toward the Ala110 amide group, the deprotonated form of Glu35 will be favored. If such a conformation is stable throughout simulations, the predicted pK_a value will be smaller than what it supposed to be. We think one reason of a low predicted pK_a value is that Glu35 is stuck in conformations stabilizing deprotonated form. But the weakly restrained simulations allow Glu35 to relax structure further and visit conformations stabilizing protonation more frequently.

Table 5-4. Distance between Glu35 carboxylic oxygen atoms and neighboring residue side-chain atoms in 1AKI crystal structure.

	Glu35 OE1	Glu35 OE2
Gln57 CB	3.56	5.25
Gln57 CG	3.85	5.84
Trp108 CB	5.36	3.43
Trp108 CG	5.43	3.94
Trp108 CD1	4.65	3.67
Ala110 N	4.65	3.09
Ala110 CB	4.19	3.48

The unit of all distances in Table 5-4 is Å.

Glu35 heavy-atom RMSD relative to 1AKI as well as cluster analysis on the basis of those RMSDs are chosen to study Glu35 conformational sampling. Distributions of

heavy-atom RMSD, which are shown in Figure 5-10, show that 2 conformations are found in the strongly restrained simulations: one centered at RMSD ~0.1 Å (we label that conformation as conformation 1) and the other centered at ~0.6 Å (it is labeled as conformation 2). However, an extra conformation (conformation 3) is visited by the weakly restrained REMD simulations. Cluster analysis is employed to separate those conformations. For conformation 2, the carboxylic group of Glu35 points toward the Ala110 amide group in both sets of the restrained REMD runs (Figure 5-11). The carboxylic group in conformation 1 also points toward the Ala110 amide group, although in a lesser extent. However, conformation 3 (shown in the weakly restrained runs only) contains configurations in which Glu35 carboxylic group is pointing away from Ala110 amide group (Figure 5-12B). In this conformation, the Glu35 side-chain is in the hydrophobic region and the protonated species is favored. A too-low percentage of conformation 3 is responsible for the low predicted pK_a value of Glu35.

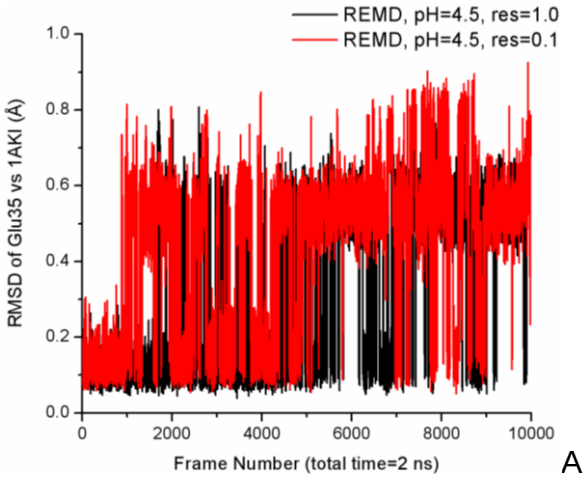


Figure 5-10. A) Time series of the Glu35 heavy atoms (excluding two carboxylic oxygen atoms) RMSD relative to crystal structure 1AKI. B) Probability distribution of the RMSD. The conformation centered at RMSD ~0.1 Å is labeled as conformation 1. The one centered at ~0.6 Å is named conformation 2. Apparently, an extra conformation (conformation 3) is visited by the weakly restrained REMD simulation.

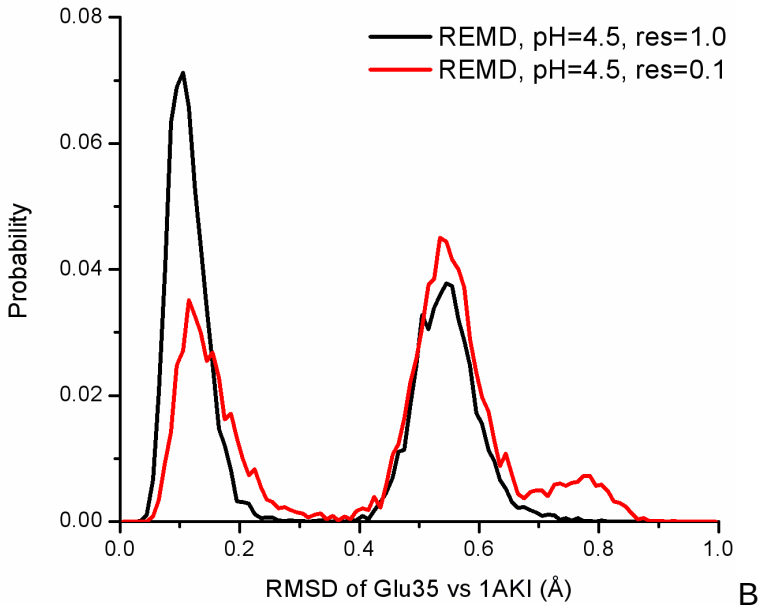


Figure 5-10. Continued

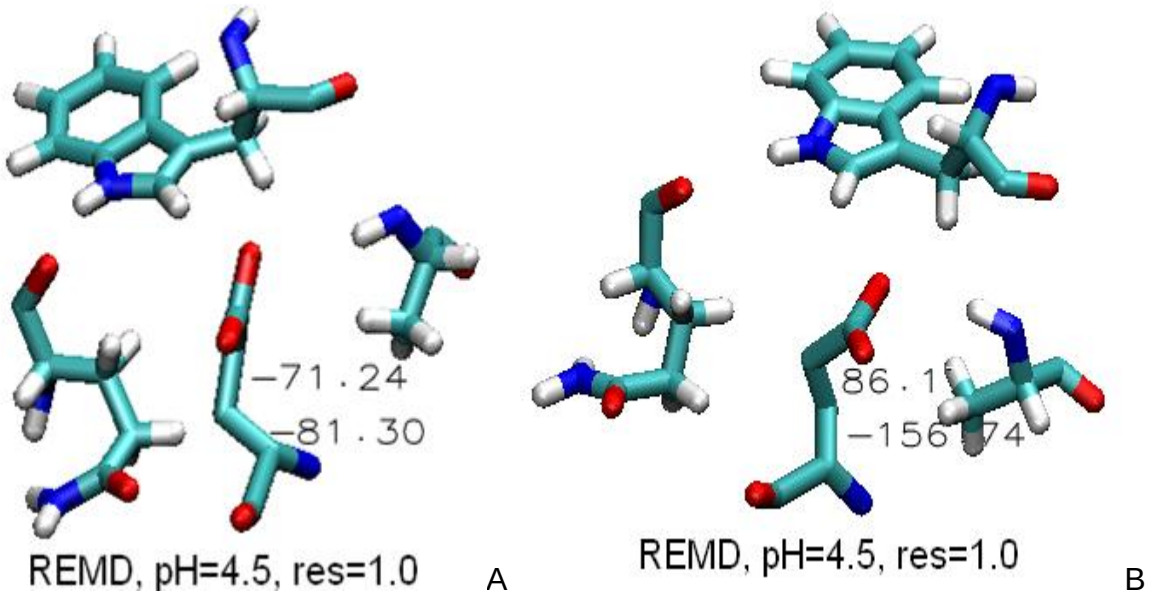


Figure 5-11. A) Representative Structure of conformation 1. B) Representative Structure of conformation 2. The structure ensemble is generated from REMD simulations with stronger restraining potential. The carboxylic group of Glu35 in conformation 2 is clearly pointing toward the amide group of Ala110. Deprotonated form of Glu35 tends to decrease the electrostatic energy. Furthermore, conformation 1 does not particularly favor the protonated Glu35. No significant stabilizing factor is found for the protonated Glu35.

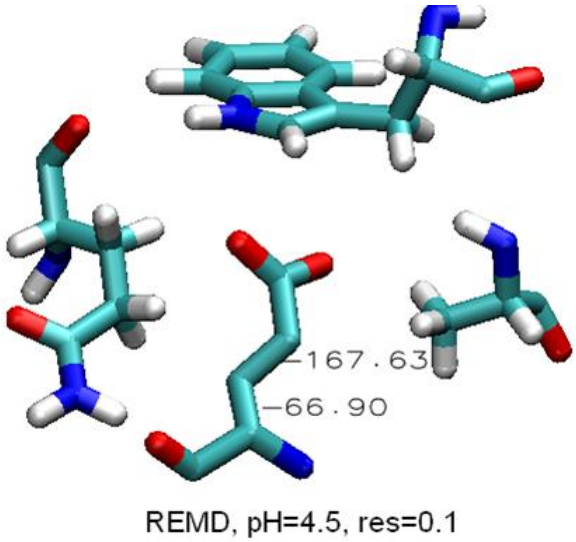


Figure 5-12. Representative Structure of conformation 3 from cluster analysis. Glu35 is in the hydrophobic region, consisting of Gln57, Trp108 and Ala110. Conformation 1 and 2 in the weakly restrained simulations are basically the same as those demonstrated in Figure 5-11.

Another possible reason of underestimating pK_a value of Glu35 is the use of implicit solvent in constant-pH MD and REMD simulations. Imoto et al. suggested that Glu35 and Asp52 were coupled by two water molecules through hydrogen-bonding. Glu35 carboxylic group acted as a proton donor in the hydrogen-bonding. Thus the protonated form of Glu35 was stabilized and contributed to the elevated pK_a value. Two water molecules are indeed found between Glu35 and Asp52 in the 1AKI crystal structure and they are within hydrogen-bonding distances to Glu35 and Asp52. If the hypothesis is true, the use of implicit solvent breaks this hydrogen-bonding network. Thus a stabilizing factor of protonated Glu35 is missing. A constant-pH algorithm employing explicit solvent is needed to study this effect.

5.5.6 Correlation between Conformation and Protonation

As described earlier, one advantage of utilizing constant-pH methods is that the conformational sampling and the protonation state sampling are directly coupled. In this

work, side-chain dihedral angles are chosen to study conformation-protonation coupling. Asp119 $\chi_1$and χ_2 dihedral angles at pH 3 will be shown as representatives. Two dimensional histograms between dihedral angles and protonation states are displayed in Figure 5-13. A two-dimensional (2D) histogram is generated by putting bins in dihedral angle and protonation state space (As explained in the second chapter, considering *syn* and *anti* configuration of protons will generate five protonation states in the case of ionizable aspartate in AMBER. They can be labeled as 0, 1, 2, 3 and 4 in which state 0 stands for deprotonated state and the rest represent protonated species).

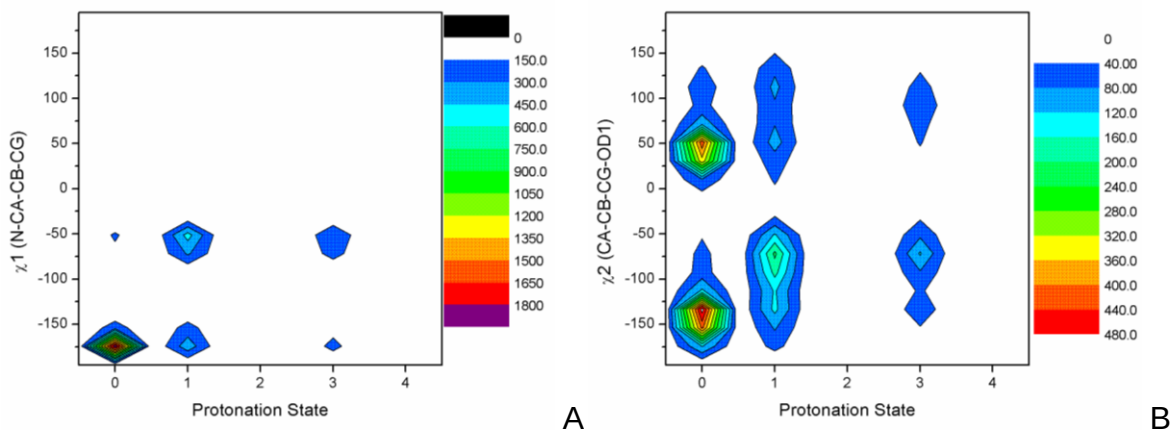


Figure 5-13. A) Correlation between side chain dihedral angle $\chi_1$and protonation states. B) Correlation between side chain dihedral angle $\chi_2$and protonation states.

Our 2D histograms can show the correlations between dihedral angle distribution and protonation state distribution. Two conformations are obtained in χ_1 space: conformation 1 having χ_1 angle around -60° while conformation 2 having χ_1 angle around -170°. In Figure 5-13A, we can clearly see that conformation 1 is coupled with protonated form and most structures in conformation 2 are in deprotonated state. According to Figure 5-13B, similar behavior can be seen in χ_2 space too. Most

deprotonated Asp119 are found having χ_2 near 40° and -140°, while configurations showing -75° and 100° of χ_2 are protonated.

A closer look at the 1AKI crystal structure reveals that side-chains of Asp119 and Arg125 are close to each other (the carboxylic group of Asp119 and the guanidinium group of Arg125 are in hydrogen bond distance). Since Arg125 has a positive charge on its guanidinium group, it stabilizes the deprotonated Asp119 when two side chains are close to each other. We calculated pK_a of Asp119 in 1AKI using H++ (H++ is a web-based FDPB server developed by Alexy Onufriev's group at Virginia Tech. The FDPB equation is solved on the basis of only one protein structure).[257,258] The calculated pK_a of Asp119 using FDPB method is -1.1, 0.7 and 1.3 when the internal dielectric constant is set to be 2, 4, and 6, respectively. All three pK_a values are much lower than experimental pK_a value of 3.2. This behavior agrees with what we just explained: Asp119-Arg125 side-chain coupling stabilizes the deprotonated form of Asp119. The single structure FDPB-based pK_a calculations yield such low pK_a values because only one conformation is visited by Asp119. Therefore, Asp119 must sample other conformations in order to yield accurate pK_a predictions. Time evolution of distance between Asp119 and Arg125 side chain is shown in Figure 5-14 to reflect that conformations other than crystal conformation are visited in our constant-pH REMD runs. In Figure 5-14, we can clearly see that the close contact between Asp119 and Arg125 side-chains can be broken during our simulations. Allowing side-chains to move will result in a pK_a value of 3.0 in our simulations. The comparison between constant-pH and single-structure FDPB algorithm clearly demonstrates the importance of conformational sampling in pK_a calculations.

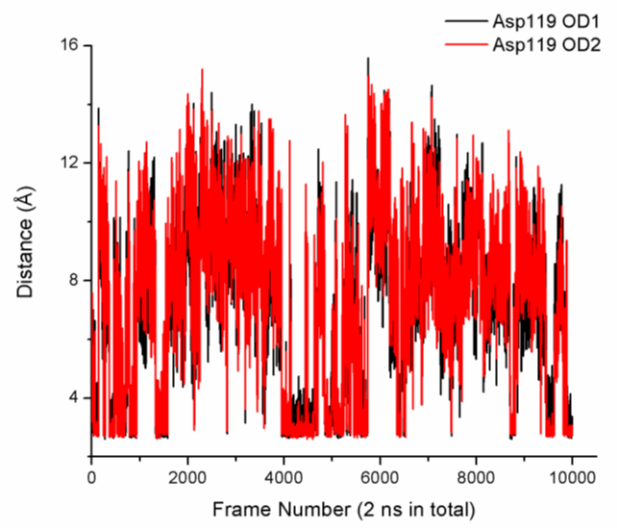
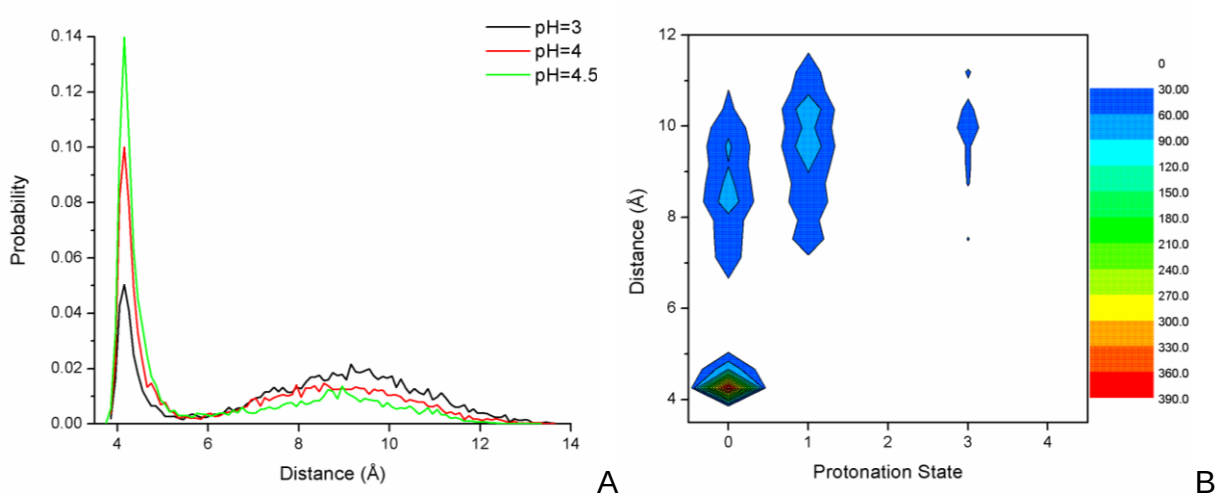


Figure 5-14. Minimal distance between Asp119 side chain carboxylic oxygen atoms (OD1 and OD2) and Arg125 guanidinium nitrogen atoms. Since guanidinium group has three nitrogen atoms, the minimal distance is the shortest distance between Asp119 OD1 (or OD2) and those three nitrogen atoms.

Therefore, another way to look at conformations is combining both Asp119 and Arg125. Now distances between Asp119 CG and Arg125 CZ atoms are selected to distinguish different conformations. Figure 5-15A shows the CG-CZ distance probability distribution. The probability distributions also reveal that two conformations exist. One conformation is centered at CG-CZ distance of 4.2 Å which represents the Asp119 and Arg125 coupling is on. The other conformation is actually representing all structures not belonging to the previous conformation. Based on the distance between Asp119 CG and Arg125 CZ, we can say the coupling is off. The 2D histogram between distance and protonation state at pH 3 is shown in Figure 5-15B. As can be seen in the 2D histogram contour plot, short distance conformation is indeed in the deprotonated state. The pK_a of shorter distance conformation is negative infinity. Although several snapshots possess both protonated state and short distance, 2D histogram doesn't reveal them as a stable conformation. So, the short distance conformation is purely coupled with deprotonated

form. We also obtain the pK_a value of the longer distance conformation is 3.3 according to Hill's plot.



A B

Figure 5-15. A) Probability distribution of Asp119 CG to Arg125 CZ distances. The Asp119 CG to Arg125 CZ distance is used to distinguish conformations. B) Coupling between conformations and protonation states.

5.5.7 Conformation-Protonation Equilibrium Model

Due to the coupling between conformation and protonation equilibrium, knowing the pH effect on conformational equilibrium will be interesting and important. Again, Asp119 is selected as the representative of our study. First, we want to show the derivation and the analytical form of K_{12} as a function of pH values in a general case. From now on, we will label conformation 1 in deprotonated form as 1d. The, 1p, 2d and 2p stand for conformation 1 in protonated form, conformation 2 in deprotonated form and conformation 2 in protonated form, respectively. According to eq. 2 and 3, [1p] = $[1d]\cdot 10^{(pKa,1-pH)}$ and [2p] = $[2d]\cdot 10^{(pKa,2-pH)}$. We can substitute [1p] and [2p] in eq. 1 with [1d] and [2d] so the conformational equilibrium constant will have the form:

$$K_{12} = \frac{[1d]}{[2d]} \cdot \frac{1+10^{(pKa,1-pH)}}{1+10^{(pKa,2-pH)}} . \qquad (5\text{-}5)$$

In Eq. 5-5, [1d]/[2d] is the equilibrium constant of conformation 1 and 2 in deprotonated form and it is equal to the K_{12} at high pH where both conformations are in the deprotonated form. So K_{12} has the final analytical formula:

$$K_{12} = K_{12,h} \cdot \frac{1+10^{(pKa,1-pH)}}{1+10^{(pKa,2-pH)}} \tag{5-6}$$

where $K_{12,h}$ stands for K_{12} at high pH. In our derivation, conformation 1 always has a smaller pK_a value than conformation 2. So the denominator always increases faster than the numerator when pH values going down. Considering that $K_{12,h}$ is a constant, then K_{12} is a sigmoid function. When pH is much greater than both pK_a values, K_{12} becomes $K_{12,h}$. When pH is much smaller than both pK_a values, K_{12} reaches its lower bound. In the case of Asp119, the pK_a value is minus infinity for conformation 1 when we use Asp119 CG and Arg125 CZ distance to distinguish two conformations. The ratios of K_{12} and $K_{12,h}$ from both analytical derivations and actual simulations are plotted in Figure 5-16. Close agreement between $K_{12}/K_{12,h}$ plots generated from simulations and conformation-protonation equilibrium model is seen in Figure 5-16A. The agreement shows that the model could represent conformational equilibrium in our constant-pH REMD simulations. So, further use of that model is possible. Different p$K_{a,1}$ and p$K_{a,2}$ values are also used in order to test how two pK_a values affect shape and inflection point of the sigmoid function. According to Figure 5-16B, 5-16C and 5-16D, if the difference between p$K_{a,1}$ and p$K_{a,2}$ is large (greater than 1 pH unit, approximately), the inflection point will appear at a pH value that equals to p$K_{a,2}$. p$K_{a,1}$ will affect the inflection point only when the difference is small. If we view a $K_{12}/K_{12,h}$ plot as a titration curve and the inflection point is the pK_a value, then the $K_{12}/K_{12,h}$ plot yields a pK_a value equals to p$K_{a,2}$ values, which is 3.3 in the case of Asp119.

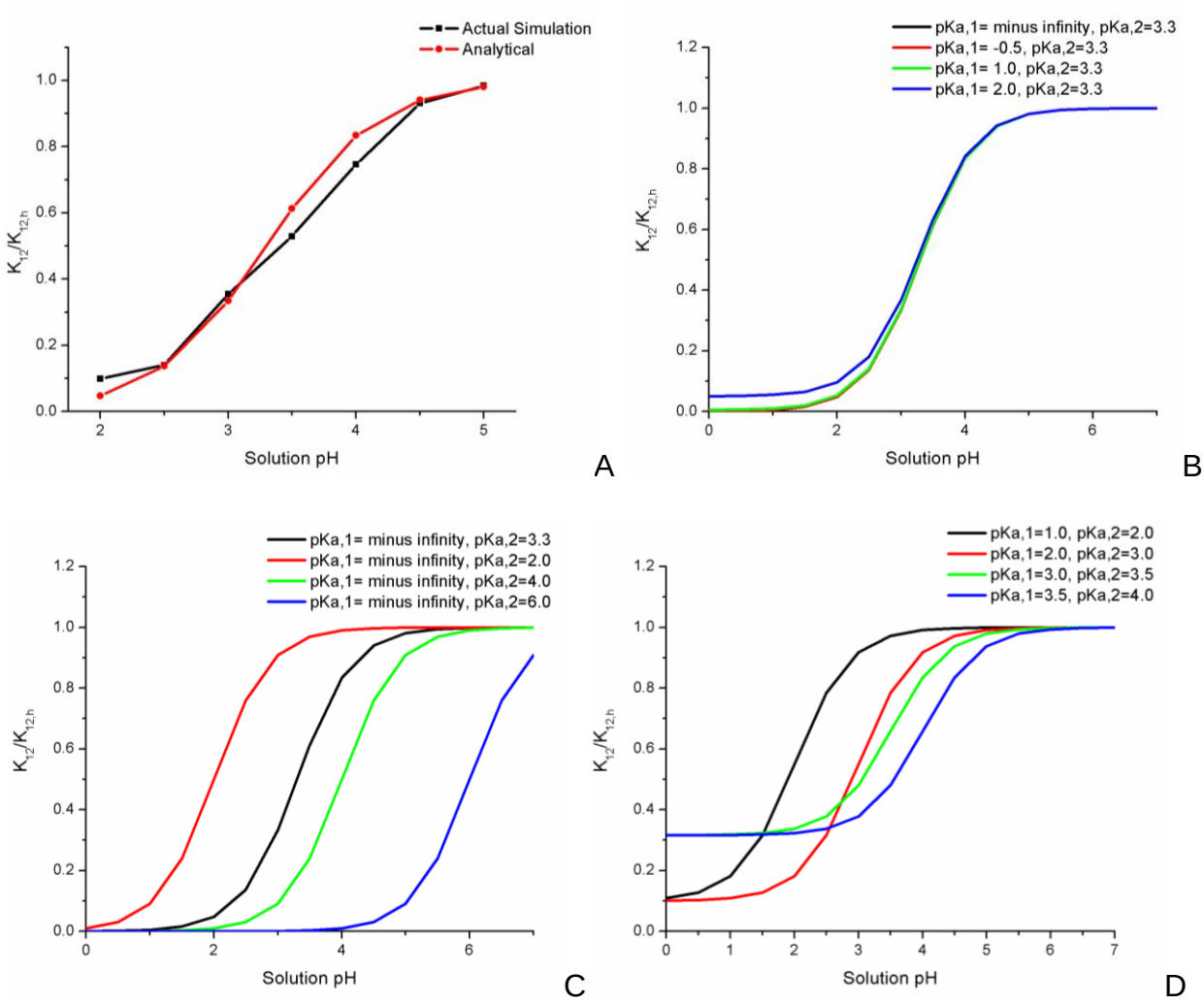


Figure 5-16. $K_{12}/K_{12,h}$ as a function of pH and its dependence on p$K_{a,1}$ and p$K_{a,2}$.

Since the analytical form of K12, pKa,1 and pKa,2 are known and the sum of all fractions is unity, we can figure out fractions of each species. The analytical expressions of each species are:

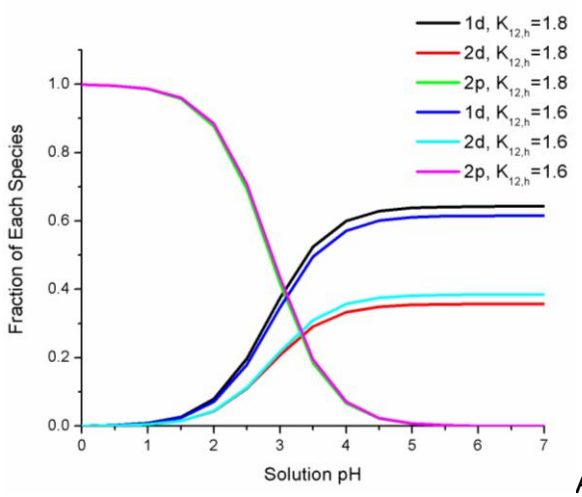
$$[1d] = \left(\frac{K_{12}}{K_{12}+1}\right)\left(\frac{1}{1+10^{(pK_{a,1}-pH)}}\right) \tag{5-7}$$

$$[1p] = \left(\frac{K_{12}}{K_{12}+1}\right)\left(\frac{10^{pK_{a,1}-pH}}{1+10^{pK_{a,1}-pH}}\right) \tag{5-8}$$

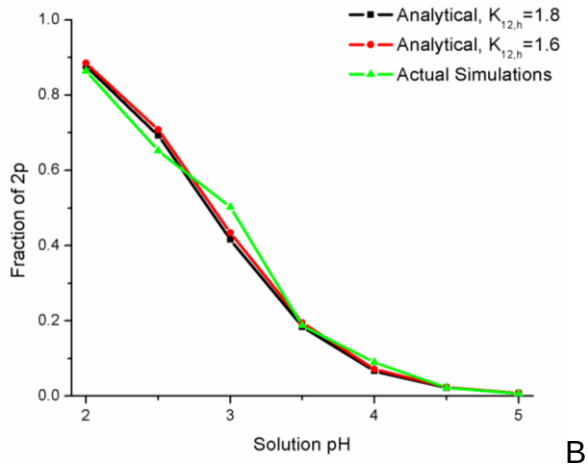
$$[2d] = \left(\frac{1}{K_{12}+1}\right)\left(\frac{1}{1+10^{pK_{a,2}-pH}}\right) \tag{5-9}$$

$$[2p] = \left(\frac{1}{K_{12}+1}\right)\left(\frac{10^{pK_{a,2}-pH}}{1+10^{pK_{a,2}-pH}}\right) \tag{5-10}$$

In our study of Asp119, p$K_{a,1}$ is minus infinity which lead to [1p] is equal to zero. $K_{12,h}$ is calculated as the average of all [1d]/[2d], which results in a K12,h of 1.6. Another $K_{12,h}$ of 1.8, which is the K_{12} at pH 5, is also tried. Then, fractions of each species from both analytical formula and actual simulations are shown in Figure 5-17.



A



B

Figure 5-17. A) Fraction of each species as a function of pH (titration curves) obtained from equations based on conformation-protonation equilibrium. The effect of $K_{12,h}$ is tested. B) Comparison of titration curves derived from actual simulations and from the equilibrium equations.

Firstly, the fraction of $2p$ vs pH plots are almost identical for two $K_{12,h}$ values. This means that although the fractions of 1d and 2d are affected, the sum of 1d and 2d is

not. Secondly, titration curves derived from analytical formula and actual simulations agree with each other very well. The agreement among titration curves leads to similar pK_a values. Both analytical titration curves using different $K_{12,h}$ yield pK_a values to be between 2.8 and 2.9 with negligible difference and the actual simulation titration curve gives a pK_a value of 3.0. The analysis demonstrates that the equilibrium model could represent protonation equilibrium in our simulations.

5.5.8 Theoretical NMR Titration Curves

Since the model can be used to simplify conformation-protonation equilibrium in our constant-pH REMD simulations, it is interesting to know whether it has some practical meanings. Reproducing experimental titration curves offers us a good objective. So, quantum mechanical calculations of NMR chemical shift (δ) are performed and their results are demonstrated and discussed in this part. As we have shown earlier, the dynamics of Asp119 generates two conformations indicating whether the Asp119-Arg125 electrostatic interaction is “on” or “off”. Our NMR calculations are based on the representative structures of each conformation, in proper protonation state. Due to the size of HEWL molecule, full quantum mechanical calculations are too expensive. So our first trial is using Asp119 dipeptide. Chemical shifts of the 1d, 2p and 2d are obtained and the fractions of each species at different pH can be calculated using eq. 7, 8 and 10. At each pH value, the theoretical chemical shift used to make a titration curve is calculated as follows: $\delta = \delta_{1d} \cdot [1d] + \delta_{2d} \cdot [2d] + \delta_{2p} \cdot [2p]$. The chemical shifts of 1d, 2d and 2p are 2.17, 2.48, 3.03 ppm respectively and the theoretical NMR titration curve is plotted in Figure 5-18. Compare theoretical titration curve with experimental one, the trend is correctly reproduced. At low pH, the theoretical and

experimental chemical shifts agree well: 3.03 ppm versus 3.13 ppm. However, the difference between calculated and experimental high pH chemical shifts is greater than 0.6 ppm. This makes our calculated ($\delta_{low\ pH}$-$\delta_{high\ pH}$) is 0.75 ppm while the experimental difference is only 0.21 ppm.

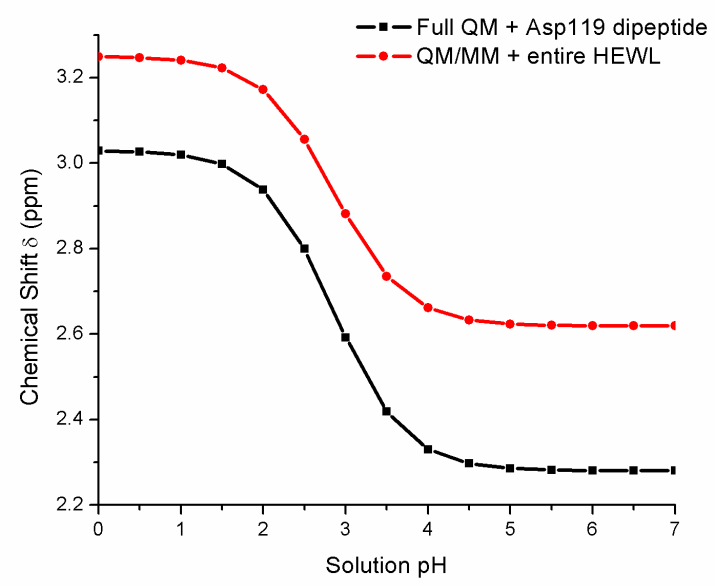


Figure 5-18. Theoretical NMR chemical shifts as a function of pH. It's plotted to see if the conformation-protonation equilibrium model can reproduce experimental titration curve based on NMR chemical shift measurements.

The problem at high pH could be that a dipeptide cannot accurately represent Asp119 and its environment especially we have known there is a strong Asp119-Arg125 Coulomb interaction. So a set of QM/MM calculations was conducted using the entire HEWL molecule. The new chemical shifts are 2.58, 2.69 and 3.25 ppm for 1d, 2d and 2p. Comparing chemical shifts based on dipeptide and the entire molecule, differences of 2p and 2d are 0.22 ppm and 0.21 ppm. More importantly, both 2p chemical shifts are similar to experimental $\delta_{low\ pH}$ (each one shows the difference near 0.1 ppm). The differences are small for 2p and 2d because there are no significant interactions for Asp119 in conformation 2. Unlike 2p or 2d, the chemical shift of 1d is improved by 0.41

ppm, telling that using the whole HEWL molecule does change 1d chemical shift a lot. After applying QM/MM method on the entire HEWL, the calculated ($\delta_{low\ pH}$-$\delta_{high\ pH}$) becomes 0.63 ppm. The theoretical titration curve using QM/MM technique is also displayed in Figure 5-18. But no matter whether a dipeptide or the entire HEWL is used in NMR calculations, the pK_a values are around 2.9 as expected. NMR titration curves yield the same pK_a value as protonation (deprotonation) fraction vs pH does. The NMR titration curve calculations validate the use of conformation-protonation equilibrium model and confirm its applicability. This model can be used to simplify a lot analysis involving further calculations.

5.6 Conclusions

In this chapter, constant-pH REMD simulations are performed to study the pK_a of hen egg white lysozyme. Three sets of constant-pH REMD simulations have been performed: one set of simulations are conducted without restraining potential, while a harmonic potential is put on the C_α atoms in the other two sets of REMD simulations. The force constants of the two harmonic potentials are 1, and 0.1 kcal/mol·Å^2, respectively, so that the effect of restraint strength on pK_a prediction accuracy can be studied.

In our constant-pH REMD simulations, the unrestrained ones are found to be structurally instable. The C_α atom RMSD relative to crystal structure can be as high as 18 Å. Due to the effect of restraining potential, HEWL in a restrained simulation is stable and similar to the crystal structure, according to the C_α atom RMSD values. In the restrained simulations with a force constant of 1 kcal/mol·Å^2, accurate pK_a predictions are achieved. The overall RMS errors between predicted and experimental pK_a values are 0.87 and 0.84, dependent of pK_a calculation methods. Unfortunately, those two

RMS errors are not better than constant-pH MD results obtained by Mongan *et al*. The advantage of incorporating REMD method is not observed. However, a plot showing RMS error as a function of pH value yields the smallest RMS error at pH 4.5, at which the crystal structure was resolved. Supported by the work of Mongan et al., we propose that the further away from crystal pH value, the stronger the biasing effect from the restraining potential. The biasing effect of conformational sampling will in turn affect pK_a predictions. As expected, reducing the strength of harmonic potential results in improved pK_a predictions. Likewise, the smallest pK_a RMS error of 0.62 is obtained at pH 4.5 in the weakly restrained constant-pH REMD simulations. An RMS error of 0.62 is among the best pK_a predictions generated from constant-pH simulations.

The pK_a predictions of catalytic ionizable residues are of particular interest in the case of HEWL. Constant-pH REMD simulations with stronger restraining potential failed to identify proton donor under the criteria proposed by Nielsen and McCammon in 2003. The weakly restrained constant-pH REMD simulations are able to predicted proton donor and nucleophile, although the errors of predicted pK_a values of Glu35 and Asp52 are among the largest in our simulations. Hydrogen-bonding is found to be responsible for the large error of Asp52. The hydrogen-bonding of Asp52 with Asn44 and Asn59 over-stabilizes the deprotonated form of Asp52, causing the pK_a value of Asp52 too small. For Glu35, conformational sampling also plays a role in underestimating its pK_a value. However, other factors such as the use of implicit solvent may affect the pK_a prediction of Glu35 too.

In this work, we also focused on conformation and protonation equilibrium in constant-pH REMD simulations. Correlations between protonation and side-chain

dihedral angles χ_1 and χ_2 are studied. Other representation of conformations such as whether an important electrostatic interaction is formed or not is also adopted. In both cases, the coupling between conformation and protonation is observed. The effect of conformation-protonation coupling is partially reflected by the comparison between constant-pH and single structure FDPB algorithms. Constant-pH REMD yields better pK_a values are seen because more conformation space is visited.

The conformation-protonation equilibrium is further studied. Equilibrium constants between conformations are derived in order to show how pH affects conformation equilibrium. The conformational equilibrium constant is shown to be pH dependent and it's a sigmoid function of pH values. The shape of the sigmoidal function is influenced by pK_a values of each conformation. Titration curves which are the means to obtain pK_a values are also derived from conformation-protonation equilibrium. All analytical results are in good agreement with our simulations. In addition, we apply this conformation-protonation equilibrium to reproduce experimental NMR titration curve by carrying out full QM and QM/MM calculations. First, we showed the importance of protein environment to chemical shift calculations. Calculation using isolated ionizable side chain can only qualitatively reproduce experimental NMR titration curve. The error mainly comes from the high pH end where an isolated side chain assumption fails. After adding protein environment, our theoretical titration curve is greatly improved and good agreement to experimental result is obtained. Our conformation-protonation equilibrium model can be used to represent our simulations and will simplify further calculations.

LIST OF REFERENCES

(1) Bettelheim, F. A. *Introduction to general, organic, and biochemistry*; 8th ed.; Thomson Brooks/Cole: Belmont, CA, 2007.

(2) Dey, A.; Verma, C. S.; Lane, D. P. *Br. J. Cancer* **2008**, *98*, 4-8.

(3) Vogelstein, B.; Lane, D.; Levine, A. J. *Nature* **2000**, *408*, 307-310.

(4) Matthew, J. B.; Gurd, F. R. N.; Garciamoreno, E. B.; Flanagan, M. A.; March, K. L.; Shire, S. J. *Crc Cr. Rev. Biochem.* **1985**, *18*, 91-197.

(5) Bierzynski, A.; Kim, P. S.; Baldwin, R. L. *Proc. Natl. Acad. Sci. U. S. A.* **1982**, *79*, 2470-2474.

(6) Ferguson, N.; Schartau, P. J.; Sharpe, T. D.; Sato, S.; Fersht, A. R. *J. Mol. Biol.* **2004**, *344*, 295-301.

(7) Shoemaker, K. R.; Kim, P. S.; Brems, D. N.; Marqusee, S.; York, E. J.; Chaiken, I. M.; Stewart, J. M.; Baldwin, R. L. *Proc. Natl. Acad. Sci. U. S. A.* **1985**, *82*, 2349-2353.

(8) Garcia-Mira, M. M.; Sadqi, M.; Fischer, N.; Sanchez-Ruiz, J. M.; Munoz, V. *Science* **2002**, *298*, 2191-2195.

(9) Hunenberger, P. H.; Helms, V.; Narayana, N.; Taylor, S. S.; McCammon, J. A. *Biochemistry* **1999**, *38*, 2358-2366.

(10) Demchuk, E.; Genick, U. K.; Woo, T. T.; Getzoff, E. D.; Bashford, D. *Biochemistry* **2000**, *39*, 1100-1113.

(11) Dillet, V.; Dyson, H. J.; Bashford, D. *Biochemistry* **1998**, *37*, 10298-10306.

(12) Harris, T. K.; Turner, G. J. *IUBMB Life* **2002**, *53*, 85-98.

(13) Laidler, K. J. *Chemical kinetics*; 3rd ed.; Harper & Row: New York, 1987.

(14) Fersht, A. *Structure and mechanism in protein science : a guide to enzyme catalysis and protein folding*; W.H. Freeman: New York, 1999.

(15) Simonson, T.; Carlsson, J.; Case, D. A. *J. Am. Chem. Soc.* **2004**, *126*, 4167-4180.

(16) Lee, A. C.; Crippen, G. M. *J. Chem. Inf. Model.* **2009**, *49*, 2013-2033.

(17) Langsetmo, K.; Fuchs, J. A.; Woodward, C. *Biochemistry* **1991**, *30*, 7603-7609.

(18) Garcia-Moreno, B.; Dwyer, J. J.; Gittis, A. G.; Lattman, E. E.; Spencer, D. S.; Stites, W. E. *Biophys. Chem.* **1997**, *64*, 211-224.

(19) Garcia-Moreno, B.; Fitch, C.; Karp, D.; Gittis, A.; Lattman, E. *Biophys. J.* **2002**, *82*, 300a-300a.

(20) Tanford, C. *Adv. Protein Chem.* **1962**, *17*, 69-165.

(21) Dwyer, J. J.; Gittis, A. G.; Karp, D. A.; Lattman, E. E.; Spencer, D. S.; Stites, W. E.; Garcia-Moreno, B. *Biophys. J.* **2000**, *79*, 1610-1620.

(22) Harms, M. J.; Castaneda, C. A.; Schlessman, J. L.; Sue, G. R.; Isom, D. G.; Cannon, B. R.; Garcia-Moreno, B. *J. Mol. Biol.* **2009**, *389*, 34-47.

(23) Mehler, E. L.; Fuxreiter, M.; Simon, I.; Garcia-Moreno, E. B. *Proteins: Struct., Funct., Genet.* **2002**, *48*, 283-292.

(24) Anderson, D. E.; Becktel, W. J.; Dahlquist, F. W. *Biochemistry* **1990**, *29*, 2403-2408.

(25) Dyson, H. J.; Jeng, M. F.; Tennant, L. L.; Slaby, I.; Lindell, M.; Cui, D. S.; Kuprin, S.; Holmgren, A. *Biochemistry* **1997**, *36*, 2622-2636.

(26) Bashford, D.; Case, D. A.; Dalvit, C.; Tennant, L.; Wright, P. E. *Biochemistry* **1993**, *32*, 8045-8056.

(27) Wang, Y. X.; Freedberg, D. I.; Yamazaki, T.; Wingfield, P. T.; Stahl, S. J.; Kaufman, J. D.; Kiso, Y.; Torchia, D. A. *Biochemistry* **1996**, *35*, 9945-9950.

(28) Dyson, H. J.; Tennant, L. L.; Holmgren, A. *Biochemistry* **1991**, *30*, 4262-4268.

(29) Jeng, M. F.; Dyson, H. J. *Biochemistry* **1996**, *35*, 1-6.

(30) Wilson, N. A.; Barbar, E.; Fuchs, J. A.; Woodward, C. *Biochemistry* **1995**, *34*, 8931-8939.

(31) Callis, P. R. *Methods Enzymol.* **1997**, *278*, 113-150.

(32) Callis, P. R.; Burgess, B. K. *J. Phys. Chem. B* **1997**, *101*, 9429-9432.

(33) Vivian, J. T.; Callis, P. R. *Biophys. J.* **2001**, *80*, 2093-2109.

(34) Inoue, M.; Yamada, H.; Yasukochi, T.; Kuroki, R.; Miki, T.; Horiuchi, T.; Imoto, T. *Biochemistry* **1992**, *31*, 5545-5553.

(35) Kajander, T.; Kahn, P. C.; Passila, S. H.; Cohen, D. C.; Lehtio, L.; Adolfsen, W.; Warwicker, J.; Schell, U.; Goldman, A. *Structure* **2000**, *8*, 1203-1214.

(36) Bartlett, G. J.; Porter, C. T.; Borkakoti, N.; Thornton, J. M. *J. Mol. Biol.* **2002**, *324*, 105-121.

(37) Jiang, Y. X.; Ruta, V.; Chen, J. Y.; Lee, A.; MacKinnon, R. *Nature* **2003**, *423*, 42-48.

(38) Luecke, H.; Richter, H. T.; Lanyi, J. K. *Science* **1998**, *280*, 1934-1937.

(39) Bashford, D.; Case, D. A. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129-152.

(40) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127-6129.

(41) Cramer, C. J. *Essentials of computational chemistry : theories and models*; J. Wiley: West Sussex, England ; New York, 2002.

(42) Raha, K.; Merz, K. M. In *Annual reports in computational chemistry*; Spellmeyer, D. C., Ed.; Elsevier: Amsterdam ; Boston, 2005; Vol. 1, p p113-130.

(43) Dixon, S. L.; Merz, K. M. *J. Chem. Phys.* **1996**, *104*, 6643-6649.

(44) Vreven, T.; Morokuma, K. In *Annual Reports in Computational Chemistry*; Spellmeyer, D., Ed.; Elsevier: Amsterdam ; Boston, 2006; Vol. 2, p p35-51.

(45) Field, M. J.; Bash, P. A.; Karplus, M. *J. Comput. Chem.* **1990**, *11*, 700-733.

(46) Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1986**, *7*, 718-730.

(47) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227-249.

(48) Kamerlin, S. C. L.; Haranczyk, M.; Warshel, A. *J. Phys. Chem. B* **2009**, *113*, 1253-1272.

(49) Monard, G.; Merz, K. M. *Acc. Chem. Res.* **1999**, *32*, 904-911.

(50) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087-1092.

(51) Wolynes, P. G.; Onuchic, J. N.; Thirumalai, D. *Science* **1995**, *267*, 1619-1620.

(52) Itoh, S. G.; Okumura, H.; Okamoto, Y. *Mol. Simul.* **2007**, *33*, 47-56.

(53) Mitsutake, A.; Sugita, Y.; Okamoto, Y. *Biopolymers* **2001**, *60*, 96-123.

(54) Berg, B. A.; Neuhaus, T. *Phys. Lett. B* **1991**, *267*, 249-253.

(55) Berg, B. A.; Neuhaus, T. *Phys. Rev. Lett.* **1992**, *68*, 9-12.

(56) Lyubartsev, A. P.; Martsinovski, A. A.; Shevkunov, S. V.; Vorontsovvelyaminov, P. N. *J. Chem. Phys.* **1992**, *96*, 1776-1783.

(57) Marinari, E.; Parisi, G. *Europhys. Lett.* **1992**, *19*, 451-458.

(58) Hansmann, U. H. E. *Chem. Phys. Lett.* **1997**, *281*, 140-150.

(59) Swendsen, R. H.; Wang, J. S. *Phys. Rev. Lett.* **1986**, *57*, 2607-2609.

(60) Earl, D. J.; Deem, M. W. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3910-3916.

(61) Fukunishi, H.; Watanabe, O.; Takada, S. *J. Chem. Phys.* **2002**, *116*, 9058-9067.

(62) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141-151.

(63) Tanford, C.; Kirkwood, J. G. *J. Am. Chem. Soc.* **1957**, *79*, 5333-5339.

(64) Tanford, C.; Roxby, R. *Biochemistry* **1972**, *11*, 2192-2198.

(65) Bashford, D.; Karplus, M. *Biochemistry* **1990**, *29*, 10219-10225.

(66) Gilson, M. K. *Proteins: Struct., Funct., Genet.* **1993**, *15*, 266-282.

(67) Antosiewicz, J.; Mccammon, J. A.; Gilson, M. K. *J. Mol. Biol.* **1994**, *238*, 415-436.

(68) Antosiewicz, J.; McCammon, J. A.; Gilson, M. K. *Biochemistry* **1996**, *35*, 7819-7833.

(69) Bashford, D.; Karplus, M. *J. Phys. Chem.* **1991**, *95*, 9556-9561.

(70) Yang, A. S.; Gunner, M. R.; Sampogna, R.; Sharp, K.; Honig, B. *Proteins: Struct., Funct., Genet.* **1993**, *15*, 252-265.

(71) Yang, A. S.; Honig, B. *J. Mol. Biol.* **1993**, *231*, 459-474.

(72) Madura, J. D.; Briggs, J. M.; Wade, R. C.; Davis, M. E.; Luty, B. A.; Ilin, A.; Antosiewicz, J.; Gilson, M. K.; Bagheri, B.; Scott, L. R.; Mccammon, J. A. *Comput. Phys. Commun.* **1995**, *91*, 57-95.

(73) Nicholls, A.; Honig, B. *J. Comput. Chem.* **1991**, *12*, 435-445.

(74) Beroza, P.; Fredkin, D. R.; Okamura, M. Y.; Feher, G. *Proc. Natl. Acad. Sci. U. S. A.* **1991**, *88*, 5804-5808.

(75) Bone, S.; Pethig, R. *J. Mol. Biol.* **1985**, *181*, 323-326.

(76) Harvey, S. C.; Hoekstra, P. *J. Phys. Chem.* **1972**, *76*, 2987-&.

(77) Garcia-Moreno, B.; Fitch, C. A. *Methods Enzymol.* **2004**, *380*, 20-51.

(78) Simonson, T.; Brooks, C. L. *J. Am. Chem. Soc.* **1996**, *118*, 8452-8458.

(79) Mehler, E. L.; Eichele, G. *Biochemistry* **1984**, *23*, 3887-3891.

(80) Mehler, E. L.; Guarnieri, F. *Biophys. J.* **1999**, *77*, 3-22.

(81) Alexov, E. G.; Gunner, M. R. *Biophys. J.* **1997**, *72*, 2075-2093.

(82) Barth, P.; Alber, T.; Harbury, P. B. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 4898-4903.

(83) Georgescu, R. E.; Alexov, E. G.; Gunner, M. R. *Biophys. J.* **2002**, *83*, 1731-1748.

(84) Gunner, M. R.; Alexov, E.; Torres, E.; Lipovaca, S. *J. Biol. Inorg. Chem.* **1997**, *2*, 126-134.

(85) Livesay, D. R.; Jacobs, D. J.; Kanjanapangka, J.; Chea, E.; Cortez, H.; Garcia, J.; Kidd, P.; Marquez, M. P.; Pande, S.; Yang, D. *J. Chem. Theory Comput.* **2006**, *2*, 927-938.

(86) You, T. J.; Bashford, D. *Biophys. J.* **1995**, *69*, 1721-1733.

(87) Kollman, P. *Chem. Rev.* **1993**, *93*, 2395-2417.

(88) Straatsma, T. P.; Mccammon, J. A. *Annu. Rev. Phys. Chem.* **1992**, *43*, 407-435.

(89) Warshel, A.; Sussman, F.; King, G. *Biochemistry* **1986**, *25*, 8368-8372.

(90) Russell, S. T.; Warshel, A. *J. Mol. Biol.* **1985**, *185*, 389-404.

(91) Jorgensen, W. L.; Briggs, J. M. *J. Am. Chem. Soc.* **1989**, *111*, 4190-4197.

(92) Merz, K. M. *J. Am. Chem. Soc.* **1991**, *113*, 3572-3575.

(93) Hu, H.; Yang, W. T. *Annu. Rev. Phys. Chem.* **2008**, *59*, 573-601.

(94) Li, G. H.; Zhang, X. D.; Cui, Q. *J. Phys. Chem. B* **2003**, *107*, 8643-8653.

(95) Riccardi, D.; Schaefer, P.; Cui, Q. *J. Phys. Chem. B* **2005**, *109*, 17715-17733.

(96) Bas, D. C.; Rogers, D. M.; Jensen, J. H. *Proteins: Struct., Funct., Bioinf.* **2008**, *73*, 765-783.

(97) Jensen, J. H.; Li, H.; Robertson, A. D.; Molina, P. A. *J. Phys. Chem. A* **2005**, *109*, 6634-6643.

(98) Li, H.; Hains, A. W.; Everts, J. E.; Robertson, A. D.; Jensen, J. H. *J. Phys. Chem. B* **2002**, *106*, 3486-3494.

(99) Li, H.; Robertson, A. D.; Jensen, J. H. *Proteins: Struct., Funct., Bioinf.* **2004**, *55*, 689-704.

(100) Li, H.; Robertson, A. D.; Jensen, J. H. *Proteins: Struct., Funct., Bioinf.* **2005**, *61*, 704-721.

(101) Minikis, R. M.; Kairys, V.; Jensen, J. H. *J. Phys. Chem. A* **2001**, *105*, 3829-3837.

(102) Day, P. N.; Jensen, J. H.; Gordon, M. S.; Webb, S. P.; Stevens, W. J.; Krauss, M.; Garmer, D.; Basch, H.; Cohen, D. *J. Chem. Phys.* **1996**, *105*, 1968-1986.

(103) Gordon, M. S.; Freitag, M. A.; Bandyopadhyay, P.; Jensen, J. H.; Kairys, V.; Stevens, W. J. *J. Phys. Chem. A* **2001**, *105*, 293-307.

(104) Mongan, J.; Case, D. A. *Curr. Opin. Struct. Biol.* **2005**, *15*, 157-163.

(105) Baptista, A. M. *J. Chem. Phys.* **2002**, *116*, 7766-7768.

(106) Baptista, A. M.; Martel, P. J.; Petersen, S. B. *Proteins: Struct., Funct., Genet.* **1997**, *27*, 523-544.

(107) Borjesson, U.; Hunenberger, P. H. *J. Chem. Phys.* **2001**, *114*, 9706-9719.

(108) Borjesson, U.; Hunenberger, P. H. *J. Phys. Chem. B* **2004**, *108*, 13551-13559.

(109) Khandogin, J.; Brooks, C. L. *Biophys. J.* **2005**, *89*, 141-157.

(110) Khandogin, J.; Brooks, C. L. *Biochemistry* **2006**, *45*, 9363-9373.

(111) Khandogin, J.; Brooks, C. L. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 16880-16885.

(112) Khandogin, J.; Chen, J. H.; Brooks, C. L. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 18546-18550.

(113) Khandogin, J.; Raleigh, D. P.; Brooks, C. L. *J. Am. Chem. Soc.* **2007**, *129*, 3056-3057.

(114) Lee, M. S.; Salsbury, F. R.; Brooks, C. L. *Proteins: Struct., Funct., Bioinf.* **2004**, *56*, 738-752.

(115) Mertz, J. E.; Pettitt, B. M. *Int. J. Supercomp. Appl.* **1994**, *8*, 47-53.

(116) Kong, X. J.; Brooks, C. L. *J. Chem. Phys.* **1996**, *105*, 2414-2423.

(117) Chen, J. H.; Brooks, C. L.; Khandogin, J. *Curr. Opin. Struct. Biol.* **2008**, *18*, 140-148.

(118) Baptista, A. M.; Teixeira, V. H.; Soares, C. M. *J. Chem. Phys.* **2002**, *117*, 4184-4200.

(119) Dlugosz, M.; Antosiewicz, J. M. *Chem. Phys.* **2004**, *302*, 161-170.

(120) Dlugosz, M.; Antosiewicz, J. M. *J. Phys. Chem. B* **2005**, *109*, 13777-13784.

(121) Dlugosz, M.; Antosiewicz, J. M. *J. Phys.: Condens. Matter* **2005**, *17*, S1607-S1616.

(122) Dlugosz, M.; Antosiewicz, J. M.; Robertson, A. D. *Phys. Rev. E* **2004**, *69*, 021915.

(123) Machuqueiro, M.; Baptista, A. M. *J. Phys. Chem. B* **2006**, *110*, 2927-2933.

(124) Machuqueiro, M.; Baptista, A. M. *Biophys. J.* **2007**, *92*, 1836-1845.

(125) Machuqueiro, M.; Baptista, A. M. *Proteins: Struct., Funct., Bioinf.* **2008**, *72*, 289-298.

(126) Machuqueiro, M.; Baptista, A. M. *J. Am. Chem. Soc.* **2009**, *131*, 12586-12594.

(127) Mongan, J.; Case, D. A.; McCammon, J. A. *J. Comput. Chem.* **2004**, *25*, 2038-2048.

(128) Walczak, A. M.; Antosiewicz, J. M. *Phys. Rev. E* **2002**, *66*, 051911.

(129) Williams, S. L.; de Oliveira, C. A. F.; McCammon, J. A. *J. Chem. Theory Comput.* **2010**, *6*, 560-568.

(130) Burgi, R.; Kollman, P. A.; van Gunsteren, W. F. *Proteins: Struct., Funct., Genet.* **2002**, *47*, 469-480.

(131) Meng, Y. L.; Roitberg, A. E. *J. Chem. Theory Comput.* **2010**, *6*, 1401-1412.

(132) Schaefer, M.; Karplus, M. *J. Phys. Chem.* **1996**, *100*, 1578-1599.

(133) Hamelberg, D.; Mongan, J.; McCammon, J. A. *J. Chem. Phys.* **2004**, *120*, 11919-11929.

(134) Hamelberg, D.; Mongan, J.; McCammon, J. A. *Protein Sci.* **2004**, *13*, 76-76.

(135) Ponder, J. W.; Case, D. A. *Adv. Protein Chem.* **2003**, *66*, 27-85.

(136) Allinger, N. L.; Yuh, Y. H.; Lii, J. H. *J. Am. Chem. Soc.* **1989**, *111*, 8551-8566.

(137) Leach, A. R. *Molecular modelling : principles and applications*; 2nd ed.; Prentice Hall: Harlow, England ; New York, 2001.

(138) MacKerell, A. D. In *Annual reports in computational chemistry* Spellmeyer, D. C., Ed.; Elsevier: Amsterdam ; Boston, 2005; Vol. 1, p p91~102.

(139) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins: Struct., Funct., Bioinf.* **2006**, *65*, 712-725.

(140) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586-3616.

(141) Daura, X.; Mark, A. E.; van Gunsteren, W. F. *J. Comput. Chem.* **1998**, *19*, 535-547.

(142) Jorgensen, W. L.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1988**, *110*, 1657-1666.

(143) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179-5197.

(144) Verlet, L. *Phys. Rev.* **1967**, *159*, 98.

(145) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327-341.

(146) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Dinola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684-3690.

(147) McQuarrie, D. A. *Statistical thermodynamics*; University Science Books: Mill Valley, Calif., 1973.

(148) Nose, S. *J. Chem. Phys.* **1984**, *81*, 511-519.

(149) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269-6271.

(150) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926-935.

(151) Mahoney, M. W.; Jorgensen, W. L. *J. Chem. Phys.* **2000**, *112*, 8910-8922.

(152) Allen, M. P.; Tildesley, D. J. *Computer simulation of liquids*; Clarendon Press ; Oxford University Press: Oxford [England] New York, 1987.

(153) Ewald, P. P. *Annalen Der Physik* **1921**, *64*, 253-287.

(154) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089-10092.

(155) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *Chem. Phys. Lett.* **1995**, *246*, 122-129.

(156) Kirkwood, J. G. *J. Chem. Phys.* **1935**, *3*, 300-313.

(157) Straatsma, T. P.; Mccammon, J. A. *J. Chem. Phys.* **1991**, *95*, 1175-1188.

(158) Zwanzig, R. W. *J. Chem. Phys.* **1954**, *22*, 1420-1426.

(159) Bennett, C. H. *J. Comput. Phys.* **1976**, *22*, 245-268.

(160) Shirts, M. R.; Chodera, J. D. *J. Chem. Phys.* **2008**, *129*, 124105.

(161) Jorgensen, W. L.; Ravimohan, C. *J. Chem. Phys.* **1985**, *83*, 3050-3054.

(162) Hansmann, U. H. E.; Okamoto, Y. *Nucl. Phys. B* **1995**, 914-916.

(163) Wang, F. G.; Landau, D. P. *Phys. Rev. E* **2001**, *64*, 056101.

(164) Wang, F. G.; Landau, D. P. *Phys. Rev. Lett.* **2001**, *86*, 2050-2053.

(165) Falcioni, M.; Deem, M. W. *J. Chem. Phys.* **1999**, *110*, 1754-1766.

(166) Kofke, D. A. *J. Chem. Phys.* **2002**, *117*, 6911-6914.

(167) Liu, P.; Kim, B.; Friesner, R. A.; Berne, B. J. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 13749-13754.

(168) Li, H. Z.; Li, G. H.; Berg, B. A.; Yang, W. *J. Chem. Phys.* **2006**, *125*, 144902.

(169) Okur, A.; Roe, D. R.; Cui, G. L.; Hornak, V.; Simmerling, C. *J. Chem. Theory Comput.* **2007**, *3*, 557-568.

(170) Roitberg, A. E.; Okur, A.; Simmerling, C. *J. Phys. Chem. B* **2007**, *111*, 2415-2418.

(171) Rathore, N.; Chopra, M.; de Pablo, J. J. *J. Chem. Phys.* **2005**, *122*, 024111.

(172) Sanbonmatsu, K. Y.; Garcia, A. E. *Proteins: Struct., Funct., Genet.* **2002**, *46*, 225-234.

(173) Kone, A.; Kofke, D. A. *J. Chem. Phys.* **2005**, *122*, 206101.

(174) Trebst, S.; Troyer, M.; Hansmann, U. H. E. *J. Chem. Phys.* **2006**, *124*, 174903.

(175) Nadler, W.; Hansmann, U. H. E. *Phys. Rev. E* **2007**, *76*, 065701.

(176) Nadler, W.; Hansmann, U. H. E. *Phys. Rev. E* **2007**, *75*, 026109.

(177) Nadler, W.; Hansmann, U. H. E. *J. Phys. Chem. B* **2008**, *112*, 10386-10387.

(178) Opps, S. B.; Schofield, J. *Phys. Rev. E* **2001**, *6305*, 056701.

(179) Zhang, W.; Wu, C.; Duan, Y. *J. Chem. Phys.* **2005**, *123*, 154105.

(180) Sindhikara, D.; Meng, Y. L.; Roitberg, A. E. *J. Chem. Phys.* **2008**, *128*, 024103.

(181) Abraham, M. J.; Gready, J. E. *J. Chem. Theory Comput.* **2008**, *4*, 1119-1128.

(182) Zhang, C.; Ma, J. P. *J. Chem. Phys.* **2008**, *129*, 134112.

(183) Rosta, E.; Buchete, N. V.; Hummer, G. *J. Chem. Theory Comput.* **2009**, *5*, 1393-1399.

(184) Zhou, R. H.; Berne, B. J.; Germain, R. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 14931-14936.

(185) Lyman, E.; Ytreberg, F. M.; Zuckerman, D. M. *Phys. Rev. Lett.* **2006**, *96*, 028105.

(186) Liu, P.; Shi, Q.; Lyman, E.; Voth, G. A. *J. Chem. Phys.* **2008**, *129*, 114103.

(187) Liu, P.; Voth, G. A. *J. Chem. Phys.* **2007**, *126*, 045106.

(188) Okur, A.; Wickstrom, L.; Layten, M.; Geney, R.; Song, K.; Hornak, V.; Simmerling, C. *J. Chem. Theory Comput.* **2006**, *2*, 420-433.

(189) Ballard, A. J.; Jarzynski, C. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 12224-12229.

(190) Kamberaj, H.; van der Vaart, A. *J. Chem. Phys.* **2009**, *130*, 074906.

(191) Nguyen, P. H. *J. Chem. Phys.* **2010**, *132*, 144109.

(192) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **2000**, *329*, 261-270.

(193) Mitsutake, A.; Okamoto, Y. *Chem. Phys. Lett.* **2000**, *332*, 131-138.

(194) Mitsutake, A.; Okamoto, Y. *J. Chem. Phys.* **2004**, *121*, 2491-2504.

(195) Andrec, M.; Felts, A. K.; Gallicchio, E.; Levy, R. M. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 6801-6806.

(196) van der Spoel, D.; Seibert, M. M. *Phys. Rev. Lett.* **2006**, *96*, 238102.

(197) Yang, S. C.; Onuchic, J. N.; Garcia, A. E.; Levine, H. *J. Mol. Biol.* **2007**, *372*, 756-763.

(198) Buchete, N. V.; Hummer, G. *Phys. Rev. E* **2008**, *77*, 030902.

(199) Case, D. A.; Darden, T. A.; T.E. Cheatham, I.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker, R. C.; Zhang, W.; Merz, K. M.; B.Wang; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I.; K.F.Wong; Paesani, F.; Vanicek, J.; X.Wu; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Kollman, P. A.; University of California, San Francisco: San Francisco, 2008.

(200) Onufriev, A.; Bashford, D.; Case, D. A. *J. Phys. Chem. B* **2000**, *104*, 3712-3720.

(201) Elber, R.; Roitberg, A.; Simmerling, C.; Goldstein, R.; Li, H. Y.; Verkhivker, G.; Keasar, C.; Zhang, J.; Ulitsky, A. *Comput. Phys. Commun.* **1995**, *91*, 159-189.

(202) Dill, K. A.; Ozkan, S. B.; Shell, M. S.; Weikl, T. R. *Annu. Rev. Biophys.* **2008**, *37*, 289-316.

(203) Dobson, C. M. *Nature* **2003**, *426*, 884-890.

(204) Anfinsen, C. B.; Haber, E.; Sela, M.; White, F. H. *Proc. Natl. Acad. Sci. U. S. A.* **1961**, *47*, 1309-1314.

(205) Mayor, U.; Johnson, C. M.; Daggett, V.; Fersht, A. R. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 13518-13522.

(206) Snow, C. D.; Nguyen, N.; Pande, V. S.; Gruebele, M. *Nature* **2002**, *420*, 102-106.

(207) Brooks, C. L. *Acc. Chem. Res.* **2002**, *35*, 447-454.

(208) Levinthal, C. *J. Chim. Phys. Phys.-Chim. Biol.* **1968**, *65*, 44-45.

(209) Gruebele, M. *Annu. Rev. Phys. Chem.* **1999**, *50*, 485-516.

(210) Kubelka, J.; Hofrichter, J.; Eaton, W. A. *Curr. Opin. Struct. Biol.* **2004**, *14*, 76-88.

(211) Snow, C. D.; Sorin, E. J.; Rhee, Y. M.; Pande, V. S. *Annu. Rev. Biophys. Biomol. Struct.* **2005**, *34*, 43-69.

(212) Snow, C. D.; Qiu, L. L.; Du, D. G.; Gai, F.; Hagen, S. J.; Pande, V. S. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 4077-4082.

(213) Zagrovic, B.; Sorin, E. J.; Pande, V. *J. Mol. Biol.* **2001**, *313*, 151-169.

(214) Jayachandran, G.; Vishal, V.; Pande, V. S. *J. Chem. Phys.* **2006**, *124*, 054118.

(215) Singhal, N.; Snow, C. D.; Pande, V. S. *J. Chem. Phys.* **2004**, *121*, 415-425.

(216) Swope, W. C.; Pitera, J. W.; Suits, F. *J. Phys. Chem. B* **2004**, *108*, 6571-6581.

(217) Swope, W. C.; Pitera, J. W.; Suits, F.; Pitman, M.; Eleftheriou, M.; Fitch, B. G.; Germain, R. S.; Rayshubski, A.; Ward, T. J. C.; Zhestkov, Y.; Zhou, R. *J. Phys. Chem. B* **2004**, *108*, 6582-6594.

(218) Daggett, V.; Levitt, M. *J. Mol. Biol.* **1993**, *232*, 600-619.

(219) Daggett, V.; Levitt, M. *J. Cell. Biochem.* **1993**, 223-223.

(220) Daggett, V.; Levitt, M. *Curr. Opin. Struct. Biol.* **1994**, *4*, 291-295.

(221) Dadlez, M.; Bierzynski, A.; Godzik, A.; Sobocinska, M.; Kupryszewski, G. *Biophys. Chem.* **1988**, *31*, 175-181.

(222) Baldwin, R. L. *Biophys. Chem.* **1995**, *55*, 127-135.

(223) Brown, J. E.; Klee, W. A. *Biochemistry* **1971**, *10*, 470-476.

(224) Fairman, R.; Shoemaker, K. R.; York, E. J.; Stewart, J. M.; Baldwin, R. L. *Biophys. Chem.* **1990**, *37*, 107-119.

(225) Osterhout, J. J.; Baldwin, R. L.; York, E. J.; Stewart, J. M.; Dyson, H. J.; Wright, P. E. *Biochemistry* **1989**, *28*, 7059-7064.

(226) Shoemaker, K. R.; Fairman, R.; Schultz, D. A.; Robertson, A. D.; York, E. J.; Stewart, J. M.; Baldwin, R. L. *Biopolymers* **1990**, *29*, 1-11.

(227) Felts, A. K.; Harano, Y.; Gallicchio, E.; Levy, R. M. *Proteins: Struct., Funct., Bioinf.* **2004**, *56*, 310-321.

(228) Hansmann, U. H. E.; Okamoto, Y. *J. Phys. Chem. B* **1998**, *102*, 653-656.

(229) Hansmann, U. H. E.; Okamoto, Y. *J. Phys. Chem. B* **1999**, *103*, 1595-1604.

(230) La Penna, G.; Mitsutake, A.; Masuya, M.; Okamoto, Y. *Chem. Phys. Lett.* **2003**, *380*, 609-619.

(231) Ohkubo, Y. Z.; Brooks, C. L. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 13916-13921.

(232) Schaefer, M.; Bartels, C.; Karplus, M. *J. Mol. Biol.* **1998**, *284*, 835-848.

(233) Sugita, Y.; Okamoto, Y. *Biophys. J.* **2005**, *88*, 3180-3190.

(234) Yoda, T.; Sugita, Y.; Okamoto, Y. *Chem. Phys.* **2004**, *307*, 269-283.

(235) Yoda, T.; Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **2004**, *386*, 460-467.

(236) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577-2637.

(237) Johnson, W. C. *Annu. Rev. Biophys. Biophys. Chem.* **1988**, *17*, 145-166.

(238) Sreerama, N.; Woody, R. W. *Methods Enzymol.* **2004**, *383*, 318-351.

(239) Gratzer, W. B.; Doty, P.; Holzwarth, G. M. *Proc. Natl. Acad. Sci. U. S. A.* **1961**, *47*, 1785-1791.

(240) Manning, M. C.; Illangasekare, M.; Woody, R. W. *Biophys. Chem.* **1988**, *31*, 77-86.

(241) Bayley, P. M.; Nielsen, E. B.; Schellma.Ja *J. Phys. Chem.* **1969**, *73*, 228-243.

(242) Clark, L. B. *J. Am. Chem. Soc.* **1995**, *117*, 7974-7986.

(243) Hirst, J. D. *J. Chem. Phys.* **1998**, *109*, 782-788.

(244) Woody, R. W.; Sreerama, N. *J. Chem. Phys.* **1999**, *111*, 2844-2845.

(245) Goux, W. J.; Hooker, T. M. *J. Am. Chem. Soc.* **1980**, *102*, 7080-7087.

(246) Ridley, J.; Zerner, M. *Theor. Chim. Acta* **1973**, *32*, 111-134.

(247) Wlodawer, A.; Svensson, L. A.; Sjolin, L.; Gilliland, G. L. *Biochemistry* **1988**, *27*, 2705-2717.

(248) Blake, C. C. F.; Koenig, D. F.; Mair, G. A.; North, A. C. T.; Phillips, D. C.; Sarma, V. R. *Nature* **1965**, *206*, 757-761.

(249) Vocadlo, D. J.; Davies, G. J.; Laine, R.; Withers, S. G. *Nature* **2001**, *412*, 835-838.

(250) Nielsen, J. E.; McCammon, J. A. *Protein Sci.* **2003**, *12*, 313-326.

(251) Bartik, K.; Redfield, C.; Dobson, C. M. *Biophys. J.* **1994**, *66*, 1180-1184.

(252) Tironi, I. G.; Sperb, R.; Smith, P. E.; Vangunsteren, W. F. *J. Chem. Phys.* **1995**, *102*, 5451-5459.

(253) Case, D. A.; Darden, T. A.; T.E. Cheatham, I.; Simmerling, C. L.; Wang, J.; Duke, R. E.; R.Luo; Merz, K. M.; Pearlman, D. A.; Crowley, M.; Walker, R. C.; Zhang, W.; Wang, B.; S.Hayik; Roitberg, A.; Seabra, G.; Wong, K. F.; Paesani, F.; Wu, X.; Brozell, S.; Tsui, V.; H.Gohlke; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Mathew, D. H.; C.Schafmeister; Ross, W. S.; Kollman, P. A.; University of California, San Francisco: San Francisco, 2006.

(254) Frisch, M. J. T., G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, Jr., J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; and Pople, J. A.; Gaussian, Inc.: Wallingford CT, 2004.

(255) Ditchfie.R *Mol. Phys.* **1974**, *27*, 789-807.

(256) He, X.; Wang, B.; Merz, K. M. *J. Phys. Chem. B* **2009**, *113*, 10380-10388.

(257) Anandakrishnan, R.; Onufriev, A. *J. Comput. Biol.* **2008**, *15*, 165-184.

(258) Gordon, J. C.; Myers, J. B.; Folta, T.; Shoja, V.; Heath, L. S.; Onufriev, A. *Nucleic Acids Res.* **2005**, *33*, 368-371.

BIOGRAPHICAL SKETCH

Yilin Meng was born in Jilin, Jilin Province, People's Republic of China. He went to the Dalian University of Technology at Dalian, Liaoning Province and studied chemical engineering. He graduated with a bachelor's degree in engineering in 2004. During his college, Yilin has developed an interest in the computational chemistry, especially the electronic structure theory and has worked in Dr. Ce Hao' group for a year.

In August 2004, Yilin came to the University of Florida and began his life as a graduate student. His original plan was to keep studying the electronic structure theory. However, he was impressed by the research of Dr. Adrian E. Roitberg. Later, he joined the Roitberg group and started his career in the molecular modeling.