UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

CONSTANT-pH MOLECULAR DYNAMICS SIMULATIONS: DEVELOPMENT AND APPLICATIONS

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

JASON AVERY WALLACE
Norman, Oklahoma
2012

Downloaded from SHAREOK. Persistent Handle URL: https://hdl.handle.net/11244/319485

CONSTANT-pH MOLECULAR DYNAMICS SIMULATIONS: DEVELOPMENT AND APPLICATIONS

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF CHEMISTRY AND BIOCHEMISTRY

BY

Dr. Jana K. Shen, Chair

Dr. Paul Cook

Dr. Helen Zgurskaya

Dr. Wai Tak Yip

Dr. Takumi Hawa

© Copyright by JASON AVERY WALLACE 2012
All Rights Reserved.

Table of Contents

List of Tables vii

List of Figures viii

List of Abbreviations x

Abstract xi

Chapter

1. General Introduction 1

1.1 Background and significance 1
1.2 Constant-pH molecular dynamics 4
1.2.1 Methods based on discrete protonation states 7
1.2.2 Methods based on continuous protonation states . . 11
1.3 Theoretical background 12
1.3.1 Generalized Born implicit solvent 12
1.3.2 Continuous constant-pH molecular dynamics 14
1.4 Overview of dissertation 17
1.4.1 Hypothesis and proposal 17
1.4.2 Description of content 18
1.5 Summary 20

2. Toward accurate prediction of pK_a values for internal protein residues: The importance of conformational relaxation and desolvation energy 21

2.1 Abstract 21
2.2 Introduction 22
2.3 Methods 25
2.3.1 pK_a calculation 25
2.3.2 Simulation details 26
2.4 Results and Discussion 27
2.4.1 Performance of continuous constant-pH molecular dynamics for pK_a predictions 27
2.4.2 Strengths of continuous constant-pH molecular dynamics for pK_a predictions 32
2.4.3 Sources of prediction error 34
2.5 Conclusion 39

3. Improving protonation state sampling of continuous constant-pH molecular dynamics 42

3.1 Abstract 42
3.2 Introduction 43
3.3 Methods 45
3.3.1 Langevin dynamics 45
3.3.2 pH-replica exchange 45
3.3.3 Analysis 46
3.3.4 Simulation details 47
3.4 Results and Discussion 48
3.4.1 Langevin titration 48
3.4.2 pH-replica exchange with deterministic titration . . 51
3.4.3 pH-replica exchange with Langevin titration 52
3.5 Conclusion 53

4. Continuous constant-pH molecular dynamics in explicit solvent with pH-based replica exchange 55

4.1 Abstract 55
4.2 Introduction 56
4.3 Methods 60
4.3.1 Continuous constant-pH molecular dynamics in explicit solvent 60
4.3.2 pH-replica exchange 61
4.3.3 Simulation details 62
4.4 Results and Discussion 66
4.4.1 Trajectory stability 66
4.4.2 Response of explicit solvent to titration 68
4.4.3 Convergence and accuracy of model compound titrations 70
4.4.4 Enhanced sampling of protonation and conformational states of proteins 75
4.4.5 Convergence and overall accuracy of protein titrations 75
4.5 Conclusion 91

5. Unraveling a trap-and-trigger mechanism in the pH-sensitive self-assembly of spider silk proteins 95

5.1 Abstract 95
5.2 Introduction 96
5.3 Methods 97
5.3.1 Structure preparation 97
5.3.2 Simulation details 98

5.3.3 Calculation of pK_a's and pH-dependent dimer stability 99
5.3.4 Poisson-Boltzmann calculations 100
5.3.5 Error estimates 100
5.4 Results and Discussion 101
5.5 Conclusion 112

6. Explicit-solvent continuous constant-pH molecular dynamics with reaction field electrostatics and charge leveling 114

6.1 Abstract 114
6.2 Introduction 115
6.3 Methods 119
6.3.1 Explicit-solvent continuous constant-pH molecular dynamics with charge leveling 119
6.3.2 Data analysis 123
6.3.3 Simulation details 124
6.3.4 Deriving reference compound potential of mean force 126
6.4 Results and Discussion 131
6.4.1 Dicarboxylic acids 131
6.4.2 Proteins 139
6.5 Conclusion 161

Bibliography 162

Appendices 173

List of Tables

Table

1.1	Standard pK_a values of amino acid side chains	4
2.1	Comparison of pK_{a} accuracy and conformational fluctuation of titratable residues	32
3.1	Correlation times using deterministic and Langevin titration.	50
4.1	Calculated and experimental pK_{a} values of model compounds	73
4.2	Calculated and experimental pK_{a} values of HP36, BBL, and NTL9	81
4.3	Effects of adding explicit ions on calculated pK_{a} values of NTL9	82
4.4	Calculated and experimental pK_{a} values of SNase	86
4.5	Calculated and experimental pK_{a} values of HEWL	89
5.1	Calculated pK_{a} values of the unbound (monomer) and bound (dimer)	102
6.1	Calculated and experimental pK_a values of amino acids from explicit-solvent continuous constant-pH molecular dynamics simulations.	130
6.2	Potential of mean force parameters and deprotonation free energy for azelaic acid	131
6.3	Experimental and calculated pK_{a}'s and pK_{a} shifts of dicarboxylic acids	132
6.4	Experimental and calculated pK_a values of HP36	140
6.5	Calculated and experimental pK_a values of HP36, BBL, and HEWL	142

List of Figures

Figure

1.1 Structures of the common titratable amino acids 3
2.1 Comparison of predicted and experimental pK_a values of SNase mutants 29
2.2 Histogram of SNase mutants pK_a prediction absolute error 31
2.3 Location of mutation site in SNase three-dimensional structure . . . 31
2.4 Time series of unprotonated fractions of T41 mutants 34
2.5 Comparison of degree of burial and relative Born radius calculated using two GB models 36
3.1 Standard deviations of unprotonated fractions using deterministic and Langevin titration 49
3.2 Autocorrelation functions of unprotonated fractions using deterministic and Langevin titration 50
3.3 Standard deviations of unprotonated fractions from deterministic titration with and without pH-replica exchange 51
3.4 Block error analysis of unprotonated fractions from deterministic titration with and without pH-replica exchange 52
3.5 Standard deviations of unprotonated fractions using Langevin titration with and without pH-replica exchange 53
3.6 Block error analysis of unprotonated fractions using Langevin titration with and without pH-replica exchange 53
4.1 Comparison of pressure, energy, and temperature from fixed charged and continuous constant-pH molecular dynamics in explicit solvent simulations 67
4.2 Response of explicit solvent molecules to titration using continuous constant-pH molecular dynamics in explicit solvent 69
4.3 Titration curves for the blocked model compounds from continuous constant-pH molecular dynamics in explicit solvent simulations . . . 72
4.4 Enhancement of protonation-state and conformational sampling of protein using pH-replica exchange 76
4.5 Convergence of pK_a values of BBL using continuous constant-pH molecular dynamics in explicit solvent 77
4.6 Comparison between calculated and experimental pK_a values and pK_a shifts relative to model values 78
4.7 Comparison of BBL conformations from explicit- and implicit-solvent simulation 83
4.8 Comparison of Lys70-Asp95 of SNase salt-bridge distribution from explicit- and implicit-solvent simulation 85
4.9 Comparison of Asp77 of SNase backbone hydrogen bond from explicit- and implicit solvent simulation 87

4.10 Comparison of conformational states of HEWL from explicit- and implicit-solvent simulations 91
5.1 Time series of pK_a values of NTD monomer and dimer 103
5.2 pH-dependent change in total charge of NTD monomer and dimer and stability of the NTD dimer 104
5.3 pH-dependent hydration of Glu79 and Glu119 in NTD dimer 105
5.4 pH-dependent solvent exposure of NTD dimer interface 106
5.5 Difference (pH 4 less pH 8) contact probability map of NTD monomer-monomer side chain interactions 108
5.6 Probability distribution of Glu79 and Glu119 distance to dimer center 108
5.7 Probability distribution of Glu119 side chain orientation 109
5.8 pH-dependent conformational rearrangement of NTD dimer 109
5.9 Probability distributions of NTD inter-monomeric salt-bridges . . . 111
5.10 pH-dependent electrostatic potential of NTD dimer subunits 112
6.1 Potentials of mean force for co-ions and azelaic acid from explicit-solvent continuous constant-pH molecular dynamics simulations . . 127
6.2 Average forces for Lys deprotonation from explicit-solvent continuous constant-pH simulations. 128
6.3 Amino acid titration curves from pH-replica exchange explicit-solvent continuous constant-pH simulations 130
6.4 Pimelic acid carboxyl-oxygen to sodium radial distribution functions 133
6.5 First pK_a's and pK_a shifts of dicarboxylic acids 135
6.6 Correlation between glutaric acid protonation states. 137
6.7 Populations and exchange ratio of azelaic acid 138
6.8 Conformation and solvent distributions of azelaic acid at different pH conditions 139
6.9 Experimental versus calculated pK_a values of proteins 143
6.10 Time series of pK_a values and exchange ratio for HP36 144
6.11 Time series of pK_a values and exchange ratio for BBL 144
6.12 Time series of pK_a values and exchange ratio for HEWL 145
6.13 pH-dependent environment of His166 of BBL 146
6.14 Correlation between Asp48 protonation state and Ser50 hydrogen bond 148
6.15 pH-dependent orientation of Asp48, Ser50, and Arg61 of HEWL . . 149
6.16 Correlation between Asp87 protonation state and Thr89 hydrogen bonding 150
6.17 pH-dependent orientation of Asp87, Thr89, and His15 of HEWL . . 151
6.18 Solvent accessible surface area of Glu35 and Asp52 of HEWL 152
6.19 NMR structure of HP36 153
6.20 Probability distributions of Asp44-Arg55 distances in HP36 155
6.21 Free-energy surfaces of HP36 158
6.22 Secondary structure of HP36 159
6.23 Distributions of HP36 Lys71 to Glu72 distances from native and non-native basins 160

List of Abbreviations

MD molecular dynamics

MC Monte-carlo

TI thermodynamic integration

CE continuum electrostatics

ACE analytical continuum electrostatics

PB Poisson-Boltzmann

GB generalized Born

pHMD constant-pH molecular dynamics

CpHMD continuous constant-pH molecular dynamics

ECpHMD explicit-solvent continuous constant-pH molecular dynamics

LJ Lennard-Jones

PMF potential of mean force

PME particle mesh Ewald

GRF generalized reaction feild

HEWL hen egg-white lysozyme

pHREX pH-replica exchange

TREX temperature-replica exchange

GBSW generalized Born with simple switching function

BSE block standard error

RDF radial distribution function

RMSD root-mean-squared deviation

RMSF root-mean-squared fluctuation

SASA solvent accessible surface area

vdW van der Waals

Abstract

Constant-pH molecular dynamics has recently emerged as a useful technique for studying the microscopic details underlying pH dependent properties of proteins. We further develop continuous constant-pH molecular dynamics (CpHMD) in several ways. First, we benchmark the implicit-solvent based CpHMD approach by calculating p$K_{\rm a}$ values for a set of over 100 engineered mutants of hyper-stable variants of staphylococcal nuclease which have titratable residues placed in the hydrophobic interior of the protein and comparing our results to experiment. We present the correlation between the calculated and experimental p$K_{\rm a}$ values and correlations of the calculated p$K_{\rm a}$ error with structural and dynamic quantities of the titratable residues. This analysis allows us to discern the strengths and limitations of implicit-solvent CpHMD.

Secondly, we implement the Langevin algorithm to propagate titration coordinates and develop a pH-based replica exchange protocol to accelerate protonation state sampling in CpHMD. We test the effects these methods have on the convergence of the unprotonated fraction of titratable amino acids. We present statistical tests which allow us to quantify the sampling enhancement. We find that both approaches speed-up protonation-state sampling significantly.

Next, we develop hybrid-solvent CpHMD to eliminate conformational biases of the generalized Born (GB) implicit-solvent model. In this method, conformational dynamics are governed by the explicit-solvent force-field, but protonation state energetics are determined by the GB implicit-solvent model. We calculate p$K_{\rm a}$ values for a series of proteins using both the GB and hybrid-solvent approaches and compare the results to experimental values. We compare the conformational states observed using the GB and the hybrid-solvent approaches and correlate this information with the accuracy of the calculated p$K_{\rm a}$ values. The results indicate that running dynamics

in explicit solvent, while using the implicit-solvent model to evaluate protonation-state energetics, yields more realistic conformational sampling which leads to more accurate pK_{a} calculation.

We then apply hybrid-solvent CpHMD to shed light on the microscopic origins of the pH-dependent assembly of spider dragline silk. We are able to calculate the pH-dependent free energy of N-terminal domain dimerization by calculating pK_{a} values of the N-terminal domain monomer and dimer, and applying linkage thermodynamics. Combining this with pH-dependent conformational changes of the intact dimer allows us to rationalize the experimentally observed pH-dependent dimer formation which is a critical step in silk assembly.

Lastly, we combine the generalized reaction field treament of long-range electrostatics and a charge-neutralization procedure which together allow fully explicit-solvent CpHMD (ECpHMD) to deliver pK_{a} values that are in good agreement with experiment. We test our ECpHMD method on a series of dicarboxylic acids and proteins. We find that the calculated pK_{a} values of dicarboxylic using ECpHMD are more accurate than those from GB-based CpHMD. Overall protein pK_{a} accuracy is on-par with the hybrid-solvent approach, but difficulty in sampling conformational states separated by high-energy barriers for residues that participate in strong hydrogen-bond or salt-bridge interactions can reduce pK_{a} accuracy. Initial data suggest this limitation can be overcome by combining the method with more effective sampling techniques. This work paves the way for future application of ECpHMD to study pH-modulated structure and function in chemistry and biology.

Chapter 1

General Introduction

Computer simulations, in particular molecular dynamics, of proteins and other biological macromolecules have become increasingly realistic over the last several decades. Modern simulation techniques are now able to offer information about molecular motion and energetics at a level of detail not possible with traditional experimental techniques; however, methods which explicitly include solution pH, a key factor in determining stability and activity of proteins, are just now emerging as practical tools. These methods are known as constant-pH molecular dynamics. We review the importance of pH for protein structure and function, the different variants of constant-pH molecular dynamics which have previously been proposed, and outline our progress in enhancing the accuracy and efficiency of the continuous constant-pH molecular technique and our applications of these newly developed methods.

1.1 Background and significance

Proteins are arguably the most important molecules in nature, serving as both nature's toolkit and building blocks. Proteins catalyze chemical processes that make life possible and they serve as structural components giving form to cells, tissues, and organs. Proteins act as messenger and transport molecules allowing information and material to pass from one location to another. Also, proteins are important for cell defense. Specialized antibody proteins allow cells to recognize and respond to foreign matter. The crucial roles proteins play in the proper maintenance and operation of cells makes understanding how, and under what conditions, they carry out their specific functions a desirable goal.

To understand how proteins work in nature, it is necessary to study them in a

biological context: life does not exist in neat water. The cellular environment is heterogeneous and crowded[1]. Molecular crowding in a biological context can hinder protein refolding[2] or can guide protein folding toward more expedient pathways and lead to an increase in the folding rate[3]. Crowding may induce shape changes in native proteins[4] and the presence of inert co-solutes can stabilize compact states of proteins[5,6]. Small, naturally occurring osmolytes, such as trimethylamine *N*-oxide, can stabilize the folded state of proteins and the addition of chemical denaturants can destabilize the native state[7]. Ionic strength and temperature can also dramatically affect protein stability[8]. Another environmental factor that can drastically alter the structure and function of proteins is the concentration of protons, or pH.

Solution pH is not consistent in all cellular compartments. For example, the pH of the cytoplasm, the endoplasmic reticulum and mitochondria is near neutral while it is acidic in lysosomes, vacuoles, and Golgi and slightly basic in the nucleus and peroxisomes[9]. There are many examples of proteins that can respond to local pH, which triggers them to carry out their biological functions. The most well known example of the modification of protein function local proton concentration may be the alkaline Bohr effect, where hemoglobin O_2 binding is altered by tissue acidity[9]. Other examples include the activation of influenza virus in lysozomes by increased acidity[10,11] and acid induced unfolding of bacteria effector proteins which is required for the proteins to enter host cells[12]. Another example is spider-silk formation which requires acidification[13,14] as part of the chemical processing that transforms the soluble protein micro-emulsion into the fibrous product. Given the numerous and varied examples of proteins recognizing changes in pH as a cue for altered activity, it is clear that to fully understand protein function, we must understand under what circumstances and by what means pH plays a role.

Proteins can gain or loose protons in response to the pH of solution either at titratable side chains or the C(carboxy)-terminus or the N(amino)-terminus. The

common titratable amino acids are shown in Figure 1.1 in their fully protonated states.

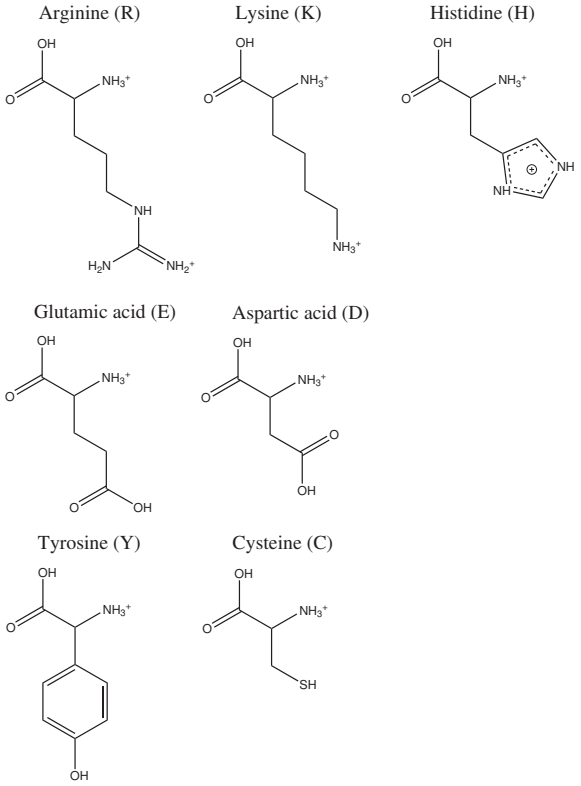

Figure 1.1: Chemical structures of the common titratable amino acids. All amino acid residues are shown in the fully protonated form.

A change in the number of bound protons alters the net charge of the protein, as well as how charge is distributed, and affects protein structure, interactions, and enzymatic activity. Protein stability[15], protein-protein[16], and protein-ligand[17] interactions are affected by changes in solution pH. These effects are manifestations of linkage thermodynamics[18,19]. Linkage thermodynamics allows one to quantitatively understand how chemical forces exerted on one equilibrium process, such as side chain titration, can have far reaching impacts on structural equilibria, such as protein-protein binding[20] or denaturation[21].

Protein catalysis is often made possible by titratable amino acids which reside at the active site; thus, enzymatic activity can be modulated by pH[22]. Although the

pK_{a}'s (p$K_{\mathrm{a}} = -log_{10}\{[\mathrm{H}^+][\mathrm{A}]/[\mathrm{HA}]\}$ for the chemical reaction $\mathrm{HA} \rightleftharpoons \mathrm{H}^+ + \mathrm{A}$) of the titratable side chains, excluding histidine, are far from the pH normally found within the cell, protein active-site structure is such that hydrophobic and electrostatic micro-environments often lead to large pK_{a} shifts (ΔpK_{a}= p$K_{\mathrm{a}}{}^{protein}$-p$K_{\mathrm{a}}{}^{solution}$)[23–27]. The ability of enzymes to tailor active site pK_{a} values to mechanistic specifications, through the evolution of specialized active-site micro-environments, allows proteins a much wider palette of amino acids to incorporate in the active site than one would suspect by a naive inference from standard amino acid pK_{a} values (see Table 1.1).

Table 1.1: Standard pK_a values of amino acid side chains

Residue	pK_{a}[28]
Arg	12.10
Lys	10.67
His	6.04
Glu	4.15
Asp	3.71
Tyr	10.10
Cys	8.14

1.2 Constant-pH molecular dynamics

Over the decades, since the pioneering work of Karplus and co-workers[29] who carried out the first gas-phase simulations of bovine pancreatic trypsin inhibitor, molecular dynamics (MD) simulation has developed into a mature tool commonly used to study protein dynamics and energetics. An exhaustive discussion of MD force field development and simulation techniques is beyond the scope of this work, but we give a brief introduction for the sake of completeness.

MD simulation attempts to model the motion of a molecule on the electronic ground state energy surface. Since the direct calculation of electronic ground state energy surfaces is computationally demanding for macromolecules in solution, we in-

stead try to mimic these surfaces by a simple empirically derived energy function[30]. There are several modern force fields that have slight variations in the energy function and the way experimental data and quantum calculations were combined during parametrization; to mention a few there are the AMBER[31], OPLS[32], Merck[33], GROMOS[34], and CHARMM[35,36] force fields. Despite small differences, all of these force fields calculate the total system energy by summing bonded and non-bonded energy terms. The bonded energy terms include stretching, angle bending, and rotations about torsion angles while the non-bonded energy terms comprise electrostatic and van der Waals energies. Thus, the total force field can be written as follows

$$E_{tot} = E_b + E_\theta + E_\phi + E_\omega + E_{vdW} + E_{elec} \tag{1.1}$$

where the individual bonded terms are expressed below.

$$E_b = \sum k_b(r - r_0)^2 \tag{1.2}$$

$$E_\theta = \sum k_\theta(\theta - \theta_0)^2 \tag{1.3}$$

$$E_\phi = \sum k_\phi(1 + cos(n\phi - \delta)) \tag{1.4}$$

$$E_\omega = \sum k_\omega(\omega - \omega_0)^2 \tag{1.5}$$

The bond (Eq. 1.2), angle (Eq. 1.3), and improper terms (Eq. 1.5) are harmonic potentials that model the deviation from the minimum energy point, while the torsion term (Eq. 1.4) is a periodic, four-atom, potential describing the energetics of rotation of the outer atoms about the axis connecting the two central ones. In the most recent generation of the CHARMM force field there is also a grid-based correction term to allow more accurate backbone ϕ/ψ angle distributions[36,37]. The non-bonded electrostatic energy is treated by Coulomb's law and a Lennard-Jones (LJ) potential[38]

models the van der Waals energy. Together they are expressed as

$$E_{nb} = E_{vdW} + E_{elec} = \sum_{\text{non-bonded pairs}} \left\{ \epsilon_{ij}^{min} \left[\left(\frac{R_{ij}^{min}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{ij}^{min}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_0 \epsilon r_{ij}} \right\} \tag{1.6}$$

where ϵ_{ij}^{min} is the well depth, R_{ij}^{min} is the separation distance where the LJ term is minimum, q_i and q_j are atomic charges for a non-bonded pair, ϵ_0 is the permittivity of free space, and ϵ is the relative dielectric. Non-bonded interactions are calculated for all pairs of atoms separated by a distance less than a specified cut-off radius and pairs of atoms that are directly connect via a chemical bond (1,2 interaction), or bond-angle potential (1,3 interaction) are also excluded from the non-bonded energy. Modern MD software packages include methods to treat long-range electrostatic interactions efficiently such as particle mesh Ewald (PME)[39], extended electrostatics[40], and generalized reaction feild (GRF)[41,42]. Many packages also include implicit solvent models (generalized Born (GB)[43] and analytical continuum electrostatics (ACE)[44] for example) which can be viewed as methods for predefining or calculating the relative dielectric (ϵ in Eq. 1.6).

With an initial set of coordinates and the energy function, we calculate the force on each particle by making use of Newton's second law of classical mechanics

$$F_i = \frac{\partial U(x_1, ..., x_i, ..., x_N)}{\partial x_i} = m_i a_i \tag{1.7}$$

where m_i is the mass, and a_i is the acceleration. From the acceleration and position of each particle at a particular time, there are many methods to solve the equations of motion numerically[39]. The force field, combined with finite-difference techniques, is used to propagate the system over time and create a simulated trajectory of molecular motion. By including counter ions and making use of pressure and temperature coupling methods, proteins, DNA, RNA, or other molecules of interest

can be simulated under conditions which mimic those used in experiment allowing MD simulation to provide microscopically detailed information that can complement information gleaned from experiment.

MD is often used to refine models derived from X-ray diffraction data, build models based on NMR constraints, and to calculate free energy differences that occur upon side chain mutation[45]. MD can also be used to calculate the free energy of ligand binding for structure-based drug design[46]. In more recent years, with greater computational resources available, all-atom folding simulations of mini-proteins in explicit solvent have been performed[47,48], and simulated folding of larger proteins using implicit-solvent models[49] has also been carried out.

Traditional MD techniques easily handle environmental conditions such as temperature, pressure, and ion concentration, but solution pH is typically ignored. Although pH is usually not explicitly considered in MD simulations, there has been considerable effort over the years towards the goal of conducting simulations at constant pH. Possibly the earliest effort in this direction was put forth by Mertz and Pettitt[50]. For the following decade there was very little advancement of constant-pH molecular dynamics (pHMD), but with recent renewed interest several different pHMD simulation methods have been proposed.

1.2.1 Methods based on discrete protonation states

The most straight-forward procedure for realizing MD simulations at constant pH is to carry out a standard MD simulation and periodically interrupt the simulation and attempt to change the protonation state based on the Monte-carlo (MC) criteria[51]. In MC sampling, instead of calculating forces, velocities, and generating an MD trajectory, one simply generates a new state arbitrarily (by changing atomic positions or, in this case, protonation state) and calculates the energy difference between the previous and new state using the energy function (see equation 1.1). The current

state is accepted or rejected based on the change in energy with a probability that is given by

$$P = min[1, exp(-\beta \Delta E)] \tag{1.8}$$

where $\beta = 1/RT$. By accepting the move with a probability given by the Boltzmann factor, after many attempts the average will converge to the desired ensemble, in this case the canonical ensemble[51]. This mixed sampling approach, running MD to generate protein conformations and MC to sample different protonation states, has been proposed several times over the years, with only slight variation. The main differences being the choice of energy function used for MD (conformation) and MC (protonation state) sampling. We briefly review these stochastic titration pHMD methods.

In an early attempt at allowing variable protonation states in an MD simulation, Baptista and co-workers fashioned a scheme where explicit-solvent MD simulation was intermittently interrupted and continuum electrostatics (CE), in the form of the Poisson-Boltzmann (PB) equation, is used to estimate the change in energy upon an update in the protein protonation state. This energy gap is used to calculate the probability of accepting the new protonation state. After a MC move is accepted, a short simulation with rigid solute is conducted to allow solvent to relax and accommodate the new protonation state[52]. This method was latter extended by including an approximate number of counter ions to keep the system close to neutrality and obtain a salt concentration near experimental conditions. The number of counter ions needed was estimated by running a short test simulation at a given pH, calculating the protein net charge, and adding the corresponding counter ions and additional ions to obtain the desired salt concentration[53]. More recently the method has shown promise in tests using different methods for treating long-range electrostatics (PME and GRF) in the MD step, and results for acidic range pK_a values of hen egg-white lysozyme (HEWL) were reported to have root-mean-squared deviation (RMSD) of

approximately 1 pK unit[54]. In a recent study, using this method and different versions of the GROMOS force field, it was proposed that different force fields strongly influence the accuracy of calculated pK_a values and that pHMD may be limited by underlying force field inaccuracies[55]. Another similar attempt was made by Bürgi, Kollman, and van Gunsteren[56]. Their method did not rely on CE energetics, but instead used thermodynamic integration (TI) to calculate the free energy change upon a protonation state update.

Thermodynamic integration is a method to calculate free energy difference between two states, for instance the free energy change upon side chain ionization, from MD simulations. In thermodynamic integration (TI), the total system energy is defined as the sum of the energies of the two end points

$$U_{total} = \lambda U_{initial} + (1 - \lambda) U_{final} \tag{1.9}$$

where λ is the variable of interpolation that allows one to follow a path between the end points, and $U_{initial}$ and U_{final} correspond to the initial and final state energy functions[57]. Simulations are conducted at several λ values, under the combined energy function, and the forces on the interpolation variable (-$dU_{total}/d\lambda$) are recorded. The resulting forces from each trajectory are averaged, and the free energy change is calculated as

$$\Delta G = \int_0^1 \left\langle \frac{\partial U(\lambda)}{\partial \lambda} \right\rangle d\lambda \tag{1.10}$$

where the angled brackets indicate averaging. The integration is carried out by fitting the data to an assumed functional form, which can be integrated analytically, or using numerical integration techniques.

In the method of Bürgi, if the protonation state change is rejected, the simulation is restarted from the previous MD round prior to beginning the TI energy evaluation[56]. Due to the high cost of each TI calculation, very long simulations were

required and few changes in protonation states were observed. In addition, the calculated pK_{a} values showed large fluctuations (on the order of several pK units) over the course of the simulation.

In an attempt to reduce the convergence problems plaguing the methods of Baptista and Bürgi, the idea of using MD simulations followed by MC sampling of protonation states was followed by Dlugosz and Antosiewiczs who used an early implicit-solvent model, ACE[58], for the MD step, but again relied on titration energetics provided by PB[59]. This departure from explicit-solvent MD helped to alleviate convergence problems of the TI based MC moves of Bürgi, and inherent problems of mixing explicit-solvent dynamics with CE protonation state changes as in the method proposed by Baptista.

A similar protocol was proposed by Mongan and Case[60] that used the more accurate generalized Born implicit-solvent model for both the MD and MC steps. The success of this method, like all constant-pH techniques, was hindered by sampling inefficiency, but the recent combination of the method with the commonly used temperature-replica exchange (TREX) protocol has accelerated convergence significantly[61].

Another attempt at mixed MD/MC-pHMD was made by Stern[62]. This method accepts protonation state updates by considering the change in the total Hamiltonian (including the kinetic energy), a model compound potential of mean force (PMF) to offset bonding energy that is not captured by MD, and a pH bias, typical of all constant-pH techniques, that is taken from simple isolated compound equilibrium considerations. This method was successful for modeling the protonation equilibrium of a molecule with a single titratable group, acetic acid, and appealing due to it simplicity, but application to more complex systems has not been reported.

1.2.2 Methods based on continuous protonation states

The other approach to include solution pH explicitly in MD simulation, and allow protonation states to fluctuate over the course of an MD trajectory, is to follow the route first pioneered by Mertz and Pettitt[50]. Instead of stopping MD simulation periodically and attempting to abruptly switch protonation states via MC sampling, the titration event is treated as an additional continuous degree of freedom. This additional degree of freedom takes the form of a coupling scheme between the two force field end points in the same spirit of TI (see equation 1.9), but instead of calculating the average force, a mass is assigned to the additional titration variable, forces on the titration coordinate are calculated and used to propagated it alongside spatial degrees of freedom. In the method of Mertz and Pettitt, a chemical potential term was added to calibrate a chemical-potential difference to a particular pH value. This general approach, as will be discussed, was later extended by others.

One attempt at continuous constant-pH molecular dynamics (CpHMD) was made by Börjesson and Hünenberger[63]. In their method, the protonation variable was relaxed to a pre-assumed equilibrium value, depending on the pH. Since the protonation state was pre-assumed this method cannot be used to predict pK_{a} values.

At the same time the stochastic GB-based pHMD method of Mongan was published, Brooks and co-workers developed a continuous titration analog. This method is based on the GB implicit-solvent model, and draws from ideas of previous continuous constant-pH molecular dynamics (CpHMD) approaches: titration is treated as an additional degree of freedom that is propagated on the same footing as conformational dynamics and the pH is included to reproduce the protonation equilibria of model compounds. The PMF of a model compound is included to offset missing bonding energies and calibrate model compound titration equilibria to experimental data[64]. In addition, a biasing potential is added to reduce the population of intermediate λ values, which correspond to unphysical mixed states. This added barrier

also allows the transition rate to be adjusted. In follow up papers, the method was refined to include dependence on solution salt concentration via a Debye-Hückel screening term[65]. Tautomeric inter-conversion of titrating residues was included to allow titration to occur at either of two quasi-degenerate proton binding sites[66] and the method was combined with TREX to enhance protonation state sampling[65]. With these improvements, CpHMD became a viable tool for studying protein ionization equilibria[65] and the method was subsequently applied the investigate pH-coupled conformational propensities of small peptides[67,68].

More recently, similar methods have emerged which do not rely on an implicit-solvent model for driving protonation events. These methods have produced reasonable results for small peptides[69] and nucleic acids[70]; however, the accuracy of these explicit-solvent CpHMD approaches have not yet been tested for complex systems such as protein or RNA.

1.3 Theoretical background

The central theme of this work is the improvement of CpHMD. We investigate limitations of the underlying GB implicit-solvent model and move CpHMD beyond the implicit representation of solvent. Here, we present the GB model and outline the CpHMD framework.

1.3.1 Generalized Born implicit solvent

The widely used GB solvation model is an attempt to cast the PB equations, which describes electrostatic interactions in a heterogeneous dielectric environment, into a much simpler form that is fast and can be readily applied to MD simulations. A full derivation of the GB formalism is not appropriate, but the reader is referred to the review by Bashford and Case[71]. In the GB theory the electrostatic solvation

energy is given by

$$\Delta G_{solv} = -\frac{1}{2}\left(1 - \frac{1}{\epsilon_w}\right)\sum_{i,j}\frac{q_i q_j}{f_{GB}} \tag{1.11}$$

where q_i and q_j are the atomic charges, ϵ_w is the solvent dielectric, and f_{GB} is an interpolation formula that behaves as a radially dependent dielectric function. At large separation f_{GB} tends toward r_{ij}, but at small separation distances f_{GB} approaches the "effective Born radii". The most commonly used form of f_{GB}, proposed by Still[43], is

$$f_{GB}(r_{ij}) = \sqrt{r_{ij}^2 + R_i R_j exp(-r_{ij}^2/4R_i R_j)} \tag{1.12}$$

where R_i and R_j are the "effective Born radii" and r_{ij} is the separation distance. In most implementations, the "effective Born radii" are calculated by numerically integrating the volume within the molecular surface, the definition of which depends on the particular implementation, excluding the volume of the atom itself. This can be written as

$$\frac{1}{R_i} = \frac{1}{a_i} - \frac{1}{4\pi}\int\limits_{in, r>a_i}\frac{1}{r^4}dV. \tag{1.13}$$

The effects of solution salt concentration can be incorporated into the Born formalism by replacing

$$\left(1 - \frac{1}{\epsilon_w}\right) \tag{1.14}$$

with

$$\left(1 - \frac{exp(-\kappa r_{ij})}{\epsilon_w}\right) \tag{1.15}$$

where

$$\kappa = \sqrt{2IF^2}\epsilon_0 \epsilon RT \tag{1.16}$$

and I is the ionic strength, F is Faraday's constant, R is the ideal gas constant, T is the absolute temperature, ϵ_0 is the permittivity of free space, and ϵ is dielectric constant of the surrounding medium.

1.3.2 Continuous constant-pH molecular dynamics

CpHMD makes use of an extended Hamiltonian formalism to allow the simultaneous propagation of spatial and alchemical degrees of freedom. In addition to "real" particle potential and kinetic energy terms, there is an additional potential energy term that couples spatial and alchemical coordinates, the kinetic energy of "virtual" alchemical particles, and biasing energies applied only to the "virtual" particles. The total Hamiltonian can be written as

$$H(\{r_a\},\{\theta_i\}) = \sum_a \frac{m_a \dot{r}_a^2}{2} + \sum_i \frac{m_i \dot{r}_i^2}{2} + U^{int}(\{r_a\}) + U^{hybr}(\{r_a\},\{\theta_i\}) + U^*(\{\theta_i\}) \tag{1.17}$$

where a is the index for spatial coordinates and i is the index for alchemical coordinates. By making use of a change of variables from θ to λ where

$$\lambda_i = sin^2(\theta_i) \tag{1.18}$$

the titration coordinate λ is restricted to the bounds of $0 \leq \lambda \leq 1$.

The term U^{int}, which does not depend on titration coordinates, includes all bonded energy terms as well as non-bonded interactions between pairs of atoms which are not titrating. The hybrid energy term (U^{hybr}) includes any non-bonded interaction involving a titrating atom, which depends largely on the treatment of solvent. For the GB-based CpHMD method, (U^{hybr}) includes Coulomb, vdW, and generalized Born energies, while for our fully explicit-solvent CpHMD method (as described in Chapter 6) we the GB-term with explicit solvent and GRF electrostatics. These energies are computed by interpolating between the protonated and deprotonated electrostatic and van der Waals (vdW) interactions. The charge on atom j of titrating residue i is computed as

$$q_j(\lambda) = (1-\lambda_i)q_j^{prot} + \lambda_i q_j^{unprot} \tag{1.19}$$

and the vdW interaction for a titrating hydrogen j on titrating residue i is similarly given by

$$U_j^{vdW} = (1 - \lambda_i)U_j^{vdW}. \tag{1.20}$$

In addition, biasing energies are added that only affect the titration coordinates, and are written in 1.17 as $U^*(\{\theta_i\})$ which is a sum of a model compound PMF, a pH biasing energy, and a barrier potential, together written as

$$U^*(\{\theta_i\}) = -U^{model}(\lambda_i) + U^{barr}(\lambda_i) + U^{pH}(\lambda_i). \tag{1.21}$$

The pH-biasing energy is taken from simple single model equilibrium considerations and is given by

$$U^{pH}(\lambda_i) = k_b T ln(10)(pH - \mathrm{p}K_a{}^{mod})\lambda_i \tag{1.22}$$

where $\mathrm{p}K_a^{mod}$ is the experimentally determined model compound $\mathrm{p}K_a$. The barrier energy (U^{barr}) is a harmonic potential

$$U^{barr}(\lambda_i) = -4\beta\left(\lambda_i - \frac{1}{2}\right)^2 \tag{1.23}$$

and the model compound PMF, determined by TI (see equation 1.10), can be approximated by a harmonic potential

$$U^{model} = A(\lambda_i - B)^2 \tag{1.24}$$

where A and B are fitting parameters. To determine the parameters A and B, we fit the derivative of Eq. 6.5 to the mean force calculate at several values of $\lambda(\theta_i)$. For residues with two possible proton binding sites, the two-dimensional PMF, having a titration degree of freedom, λ, and a tautomeric degree of freedom, x, can be

approximated by a second-order bivariate polynomial of the general form

$$U^{mod}(\lambda_i, x_i) = a_0\lambda_i^2 x_i^2 + a_1\lambda_i^2 x_i + a_2\lambda_i x_i^2 + a_3\lambda_i^2 + a_4 x_i^2 + a_5\lambda_i x_i + a_6\lambda_i + a_7 x_i + a_8 \quad (1.25)$$

In the CpHMD approach, the deprotonation free energy of a titration site in a specific chemical environment (ΔG^{exp}_{env} e.g. a side chain of a protein) is obtained by calculating the difference between the deprotonation of the titratable site in the environment of interest (ΔG^{sim}_{env}) and in solution (ΔG^{sim}_{sol}). We can relate the experimental and calculated deprotonation free energy differences by

$$\Delta G^{exp}_{env} - \Delta G^{exp}_{sol} \approx \Delta G^{sim}_{env} - \Delta G^{sim}_{sol}. \quad (1.26)$$

If the two sides of Eq. 1.27 were equivalent, the calculated and experimental pK_{a} values would exactly match. Adding the pH-dependent free energy of the reference compound (ΔG^{exp}_{sol}) to both sides gives the expression

$$\Delta G^{exp}_{env} \approx \Delta G^{sim}_{env} - \Delta G^{sim}_{sol} + \Delta G^{exp}_{sol} \quad (1.27)$$

where ΔG^{sim}_{sol} is our calculated model compound PMF (U^{mod}), ΔG^{exp}_{sol} is the pH-bias (U^{pH}), and ΔG^{sim}_{env} arises as a result of the hybrid-energy term (U^{hybr}).

The total deprotonation energy can be considered to have two separate components: the energy of breaking bonds and the accompanying electronic reorganization (ΔG^{quant}) which cannot be captured by MD simulations, and the energy difference that arises due to classical interactions of the titratable site with the environment. We assume $\Delta G^{quant}_{env} \approx \Delta G^{quant}_{sol}$; therefore, they cancel and are omitted.

Instead of calculating ΔG^{deprot} directly, we make use of the fact that this free

energy difference is related to the p$K_{\rm a}$ via

$$p\mathrm{K_a} = \frac{1}{\ln(10)\mathrm{RT}}\Delta\mathrm{G}^{\mathrm{deprot}} \tag{1.28}$$

and calculate the $p\mathrm{K_a}$ by running simulations at several pH conditions. We then calculate the fraction of time spent in each protonation state, and fit the unprotonated fraction to an appropriate titration model.

1.4 Overview of dissertation

1.4.1 Hypothesis and proposal

Given the importance of pH, the success of the implicit-solvent based CpHMD approach in calculating p$K_{\rm a}$ values[65] and pH-dependent conformational propensities[67], and the emergence of explicit-solvent CpHMD, further development of CpHMD is expected to yield more accurate and informative results in the future. Considering the need to understand the role of pH and how changes in protonation state can affect protein dynamics and function, the aim of this work is to further develop CpHMD by enhancing the level of realism and accuracy. We have tested the limitations of the implicit-solvent based approach to learn where improvements could be made and systematically push the method forward toward a fully atomistic representation. By improving CpHMD and implementing our newly developed methods in widely used biomolecular software packages, other members of the scientific community interested in exploring pH-dependent processes in chemistry and biology via computer simulation will have access to a more robust and accurate CpHMD method. The dissemination of these methods, and use by others, will serve to enhance our knowledge of specific problems in chemistry and biology. As will be described later, an example of the broad impact these methods may have, by informing about specific problems, is our application of enhanced CpHMD techniques toward the understanding of pH-

dependent energetics of spider-silk assembly.

1.4.2 Description of content

The content of next five chapters covers a range of topics related to our interest in understanding the shortfalls of the current GB-based CpHMD technique and using that knowledge to make systematic improvement to the method. First, In Chapter 2, we describe our benchmark study of the GB-based CpHMD technique. In an attempt to understand the strengths and limitations, we calculated the pK_{a} values for over 100 mutants with titratable side chains introduced into the hydrophobic core of hyperstable variants of staphylococcal nuclease (PHS and Δ+PHS). This work was part of a blind pK_{a} prediction exercise. We find that GB-based CpHMD pK_{a} predictions for this challenging data set were generally within 1 pK unit of the experimental result. Analysis of the outliers and correlating the position of the mutation sites with the overall error at that site indicated that the GB model provides more accurate results for residues located near the proteins surface than those that are more deeply buried. We observed that poor convergence of pK_{a} values may limit the pK_{a} prediction accuracy. Another limiting factor may be an inadequate level of realism in describing protein conformational dynamics.

Next, to address the issues noted in Chapter 3 related to poor pK_{a} convergence, we implemented a Langevin integration algorithm for protonation state propagation and developed a pH-replica exchange (pHREX) technique. We find that each of these advancements was able to independently provide a significant speed-up of protonation sampling and, when combined, the uncertainty of the unprotonated fraction for model compounds was reduced significantly. After addressing the problem of slow pK_{a} value convergence, we next turn to correcting the energetic deficiencies CpHMD inherited from the GB implicit-solvent model.

In Chapter 4, we describe our effort to move CpHMD into an explicit represen-

tation of solvent. Toward this end, we developed and tested a hybrid-solvent scheme where the explicit-solvent force field was used to drive conformational sampling, while the GB model was used to evaluate protonation state energetics. We show that by running dynamics in explicit solvent we are able to eliminate conformational biases of the GB model. At the same time, by making use of the efficiency of GB in estimating the protonation-state electrostatic energies, in combination with pHREX, we are able to obtain greater accuracy than the GB-based CpHMD with short 1 ns trajectories.

After demonstrating the pK_{a} calculation accuracy of hybrid-solvent CpHMD, we next turn to applying the technique to an interesting problem in biology: this work is described in Chapter 5. We sought to understand the molecular origin of the pH-dependence of spider dragline-silk formation. The protein responsible for this pH-dependent phenomenon had been previously identified, and it had been shown that dimerization of the NT-domain of the MaSp1 protein in spider silk was linked to fiber formation; however, the detailed mechanism by which a drop in pH leads to NT-domain dimerization was not understood. We hypothesized that the pH-dependent dimerization is a result of linkage thermodynamics and that burial of acidic residues at the hydrophobic protein-protein interface may be responsible for the observed acid-induced dimerization. To test this, we applied hybrid-solvent CpHMD to calculate pK_{a} values of the NT-domain dimer and monomer. Utilizing thermodynamic linkage relations, we are able to demonstrate that there is a complex network of electrostatic interactions at the protein dimerization interface that responds to acidification and can cause the observed pH-dependent dimerization.

Lastly, in Chapter 6, we again turn toward method development. We attempt to eliminate all dependence on the GB implicit-solvent model in CpHMD simulation. To accomplish this, we implemented the GRF method to handle truncated electrostatic interactions. We also employ a novel procedure to automatically neutralize the net charge of the system despite protonation state fluctuations. This was accomplished by

coupling ionization of titratable sites to the simultaneous charging and neutralization of like charged co-ions. We tested this approach by calculating pK_{a} values of a series of dicarboxylic acids and proteins. Our results indicate that charge neutralization is necessar in order to obtain accurate pK_{a} values. The overall accuracy of protein pK_{a} values using our explicit-solvent CpHMD method is on par with the hybrid-solvent approach. However, in certain cases, accuracy is limited by slow conformational rearrangement coupled to protonation events.

1.5 Summary

Solution pH is an extremely important factor that has profound effects on diverse systems in chemistry and biology. Traditional MD is limited to simulation at constant protonation state, which in some cases is a gross distortion of reality. The development of advanced simulation methodologies that explicitly and seamlessly include pH in the mathematical formulation of the model is extremely important. These methods allow the dissection and understanding of pH-dependent phenomena at the microscopic level with atomic resolution, in a quantitative fashion. Described in this work are methods that push the boundaries of CpHMD simulation techniques. We rigorously tested the previous GB-based method, found weaknesses and made systematic improvement to protonation-state convergence, as well as conformational and protonation state energetics. We developed a hybrid-solvent method which, for the first time, allowed CpHMD to be combined with explicit-solvent MD, while simultaneously delivering accurate pK_{a} values. We then used our hybrid-solvent approach to understand the origin of the pH-dependence of spider silk assembly. Finally, we employed a charge-neutralization procedure and show for the first time that fully explicit-solvent CpHMD can deliver accurate pK_{a} values of complex systems. The studies described in what follows open the door to a deeper understanding of pH-dependent phenomena at atomic resolution.

Chapter 2

Toward accurate prediction of pK_a values for internal protein residues: The importance of conformational relaxation and desolvation energy

In order to improve continuous constant-pH molecular dynamics, it is necessary to understand it's limitations. To evaluate the methods strengths and weaknesses, we tested the method by predicting the pK_a values of nearly 100 pK_a values of titratable residues introduced into the hydrophobic core of a protein. This allowed us to identify several factors which hinder the accuracy of the generalized-Born-based continuous constant-pH molecular dynamics.

The following content was published in :

Proteins: structure, function, and bioinformatics

volume 79, pages 3364-3373, 2011

2.1 Abstract

Proton uptake or release controls many important biological processes such as energy transduction, virus replication, and catalysis. Accurate pK_a prediction informs about proton pathways, thereby revealing detailed acid base mechanisms. Physics-based methods, in the framework of molecular dynamics simulations, not only offer pK_a predictions but also inform about the physical origins of pK_a shifts and provide details of ionization-induced conformational relaxation and large-scale transitions. One such method is the recently developed continuous constant-pH molecular dynamics (CpHMD) approach, which has been shown to be an accurate and robust pK_a prediction tool for naturally occurring titratable residues. In order to further

examine the accuracy and limitations of CpHMD, we predicted p$K_{\rm a}$ values of 87 titratable residues introduced in various hydrophobic regions of staphylococcal nuclease and variants. The predictions gave an root-mean-squared deviation (RMSD) of 1.69 pK units from experiment and there were only two p$K_{\rm a}$'s with errors greater than 3.5 pK units. Analysis of the conformational fluctuations of titrating side chains in the context of the error of calculated p$K_{\rm a}$ values indicates that explicit treatment of conformational flexibility and the associated dielectric relaxation gives CpHMD a distinct advantage. Analysis of the sources of error suggests that more accurate p$K_{\rm a}$ predictions can be obtained for the most deeply buried residues by improving the accuracy in calculating desolvation energies. Furthermore, it is found that the generalized Born implicit-solvent model underlying the current CpHMD implementation slightly distorts the local conformational environment such that the inclusion of an explicit-solvent representation may offer improved of accuracy.

2.2 Introduction

Many important biological processes are driven by proton translocation. For example, ATP synthesis is driven by a transmembrane proton gradient [72], while replication of influenza virus requires proton conductance of the M2 proteins [73]. Thus, knowledge of the protonation states of these biological assemblies is critical for unraveling detailed mechanisms. In order to gain a deeper understanding of such pH-driven processes it is often desirable not only to predict p$K_{\rm a}$ values correctly, but also to identify the underlying physical principles guiding these processes. The ability to report on physical origins of p$K_{\rm a}$ shifts relative to model or solution values gives physics-based methods a distinct advantage over empirical approaches, although the latter are more computationally efficient and thus useful in certain applications. In this paper physics-based methods are referred to those that do not employ parameters derived from experimental protein p$K_{\rm a}$ values.

In recent years a class of methods based on molecular dynamics (MD) have been developed that offer simultaneous description of conformational dynamics and proton titration at a specified pH condition. These methods assume an infinite proton bath and are commonly referred to as constant-pH molecular dynamics (pHMD) [74,75]. pHMD allows pK_a values to be calculated by naturally incorporating dielectric relaxation due to intrinsic motion of the protein and ionization-induced conformational changes. The ability to explicitly account for dielectric relaxation of the protein, a phenomenon which can only be approximately captured by the use of an *effective* internal dielectric constant in Poisson-Boltzmann (PB) calculations [76], makes pHMD approaches, in principle, less dependent on the initial structure since possible side chain rearrangement as well as the more dramatic conformational changes upon uptake or release of protons are sampled in the simulation. Although the ability to include the dynamic nature of proteins is attractive, it also introduces difficulties. Not only is it necessary to evaluate the free energy of side chain ionization, pHMD approaches must also produce accurate conformational ensembles at a given pH which introduces issues with convergence, as well as dependence on a particular molecular mechanics force field and treatment of solvent. On the other hand, this dependence comes with the added benefit of providing a means to test and validate force fields, solvent models, and sampling convergence.

The particular pHMD approach that we focus on here is the CpHMD [64,66] based on the λ-dynamics technique for free energy calculations [77] and the generalized Born (GB) implicit-solvent model [78,79]. CpHMD utilizes a set of fictitious λ particles to describe proton titration. The titration coordinates, bound between 0 and 1, are propagated simultaneously with the conformational degrees of freedom. Combined with the temperature replica exchange (TREX) conformational sampling protocol [80,81], CpHMD is a powerful tool not only for pK_a calculations [65,82], but also for atomically detailed simulations of pH-dependent conformational processes such as protein fold-

ing[67,68,83]. Previous benchmark studies based on a dozen proteins of various folds have demonstrated that TREX-CpHMD titrations can reliably predict pK_a's with simulation lengths of 1 ns per replica[65,82]. The root-mean-square deviation (RMSD) from experimental data is consistently below 1 pK unit for surface residues and 1.5 pK units for deeply buried ones. In a recent work we demonstrated the accuracy of TREX-CpHMD titrations for predicting pK_a's of intrinsically flexible proteins such as α-lactalbumin and deeply buried residues such as those in the designed mutants of staphylococcal nuclease (SNase)[82].

Blind pK_a prediction for internal residues offers perhaps the most stringent test of the ability of pK_a prediction methods to describe microscopic electrostatics in proteins. Let us consider the three energetic contributions to a pK_a shift: desolvation of the titrating site, and Coulomb interaction with the neutral protein background and other titratable sites[84]. In order for physics-based methods, to provide accurate prediction of the pK_a shift, all three terms, which are subject to the effects of dielectric relaxation, in principle, need to be calculated accurately. For an ionizable residue deeply buried in the hydrophobic core, desolvation energy favoring the charge-neutral form is very large. Due to the lack of solvent dielectric screening, the absolute energy due to interaction with nearby charged residues, if any, is also very large. In naturally occurring proteins, the latter interaction is stabilizing for the charged form, which at least partially offsets the pK_a shift due to desolvation and leads to error cancellation in the evaluation of the total energy. However, charge stabilizing interactions in the hydrophobic core are not always present in designed proteins, such as the single mutants of SNase and its hyper-stable variants from the current blind prediction set. In these designed mutants, a hydrophobic residue is substituted by a titratable one which may or may not have a partner for charge-charge interaction[85,86]. In many cases, as evident from experimental data[86] and which will be discussed in this work, the desolvation factor dominates, leading to an extremely large pK_a shift. In

these cases the pK_{a} of the buried residue is very challenging to predict, because a small percentage error in a large desolvation energy can result in a large error in the calculated pK_{a} shift.

The accuracy of existing electrostatic methods such as PB or GB to calculate the desolvation energy is limited because of the sensitivity to the location of the dielectric boundary and the need to account for the effects of dielectric relaxation. Thus, the current prediction data set is truly challenging for physics-based methods.

By allowing microscopic coupling between protonation equilibrium and conformational dynamics, TREX-CpHMD titrations offer pK_{a} predictions in very good agreement with experiment, typically within 1 pK unit, without the need for the use of an effective dielectric constant for protein interior and, in principle, without the need for a high-resolution structure. In this paper we will first give an overview of the performance of the CpHMD method in the context of the prediction set. We will then discuss the strengths of the method by examination of the prediction performance, local conformational relaxation and simulation convergence. We will demonstrate that the major source of error is related to the inaccuracy of the underlying GB model in the calculation of desolvation energy for deeply buried sites and the description of local conformational environment. Finally, we will outline future directions for the improvement of CpHMD for accurate pK_{a} calculations.

2.3 Methods

2.3.1 p$\boldsymbol{K}_{\mathrm{a}}$ calculation

CpHMD simulations performed at a specified pH condition result in the deprotonated fractions (S) for all titrating residues. By fitting the S values at a single pH to

the Henderson-Hasselbach equation

$$S = \frac{1}{1 + 10^{n(pK_a - \mathrm{pH})}} \tag{2.1}$$

or its generalized form, where n deviates from 1, at multiple pH values, the pK_{a} of the titrating residue of interest can be calculated.

2.3.2 Simulation details

The proteins studied in the prediction data set are single mutants of three parent proteins related to SNase, the wild type (PDB ID: 1STN[87]), PHS (PDB ID: 1EY8[88]) with three substitutions (P117G, H124L, S128A), and Δ+PHS (PDB ID: 3BDC[89]), which contains two more substitutions (G50F, V51N) than PHS and a deletion (residues 44–49). Initial structures were taken directly from the Protein Data Bank when available or were generated by computationally mutating the desired residues of the parent protein using the MOLDEN program[90]. The rest of the protocol utilized the CHARMM program[39] and MMTSB Tool Set[91]. Hydrogen atoms were added using the HBUILD facility in CHARMM followed by 50 steps of steepest descent and then 30 steps of adopted basis Newton-Ralphson energy minimization. All heavy atoms were constrained during minimization. Starting from the prepared structures, CpHMD titration simulations were performed employing the CHARMM22/CMAP force field[35,36]. To enhance sampling, the TREX protocol[80,81] was applied with 16 replicas occupying exponentially-spaced temperatures ranging from 298 to 400 K. Exchanges between adjacent temperature replicas were attempted every 2 ps. The actual exchange ratio was 40–50%. Each replica was subjected to Langevin dynamics with a collision frequency of 5 ps^{-1}. The SHAKE algorithm was applied to all bonds and angles involving hydrogen atoms to allow an integration step of 2 fs. Solvent was implicitly modeled by GBSW[79] with all param-

eters identical to those in the previous work[65,82]. The salt concentration was set to 100 mM in accord with experiment.

pK_a values were first estimated by running TREX-CpHMD simulations at several pH conditions for about 100 ps per replica. The TREX-CpHMD simulation at the pH condition which gave a deprotonated fraction closest to 0.5 was then extended to 1 ns per replica. Spatial and titration coordinates were saved every 2 ps, resulting in 1000 snapshots of conformational and protonation states.

The final pK_a's reported were obtained from the 298 K replica of the single 1-ns TREX-CpHMD simulation, by inserting the deprotonated fraction into the Henderson-Hasselbach equation (Eq. 2.1).

2.4 Results and Discussion

2.4.1 Performance of continuous constant-pH molecular dynamics for pK_a predictions

Using the GBSW based TREX-CpHMD titration simulations, we calculated the pK_a values for 95 titratable residues introduced at various interior hydrophobic sites of SNase and the stabilized variants, PHS and Δ-PHS. The data set consists of Asp, Glu, Lys, and Arg residues substituted for the wild type hydrophobic residues at 25 sites. Of these, 87 residues are blind predictions, without knowledge of experimental pK_a values at the time of calculation. The total data set (experimental and predicted pK_a values) are in the appendix. The total root-mean-square deviation (RMSD) of the calculated pK_a's from the experimental data for these residues is 1.69 pK units, which is comparable to the previous benchmark calculations for naturally occurring buried residues[65]. The RMSD by residue type is 1.63, 1.48, and 1.78 for Asp, Glu, and Lys, of which there are 22, 23, and 19 residues, respectively. Thus, the performance of CpHMD predictions does not depend heavily on the identity of the titrating side chain.

For residues which do not have precisely determined experimental p$K_{\rm a}$'s, but rather only bounds, the CpHMD method predictions are within the experimental bounds in 24 out of the 31 cases (77%). Considering the large number of these residues, and the fact that these residues are the only cases where burial in the hydrophobic core appears to anomalously stabilize the charged form relative to the solvent-exposed form of the residue, precisely determined experimental p$K_{\rm a}$'s would offer additional information about the accuracy and limitation of the CpHMD method. For the cases where our prediction is outside the experimentally determined range, the absolute difference between the predicted p$K_{\rm a}$ and experimental bound is on average 1.8 pK units, the largest of which (L37K and A132K) is about 3 pK units.

We plotted the predicted versus measured p$K_{\rm a}$ values (see upper panel of Figure 2.1). Most data points fall above the diagonal line which represents perfect prediction, indicating that there is a systematic underestimation of the p$K_{\rm a}$'s. The p$K_{\rm a}$'s are shifted to favor the neutral form for residues which have precisely measured values. This is expected since all mutation sites are not exposed to solvent. Consequently, the p$K_{\rm a}$'s of Asp and Glu residues are shifted higher than the model values of 4 and 4.4, respectively, while the p$K_{\rm a}$'s of Lys and Arg are shifted lower than the model values of 10.4 and 12.5, respectively. As a result of the opposite direction in p$K_{\rm a}$ shifts, the calculated and measured p$K_{\rm a}$ values tend to cluster around pH 7, which makes it difficult to analyze the correlation between calculation and experiment. Therefore, we plotted the predicted and measured p$K_{\rm a}$ shifts (see lower panel of Figure 2.1). The p$K_{\rm a}$ shift (ΔpK_a), which is the difference between the model p$K_{\rm a}$ and that measured or calculated in the protein environment, has been suggested to be a more informative measure of the accuracy of p$K_{\rm a}$ prediction methods[76], because it reflects the difference in the free energy of charging the residue in the protein environment and in solution. Linear regression of the predicted p$K_{\rm a}$ shifts versus measured values gives a correlation coefficient of 0.88 with a slope of 0.93, which reveals that

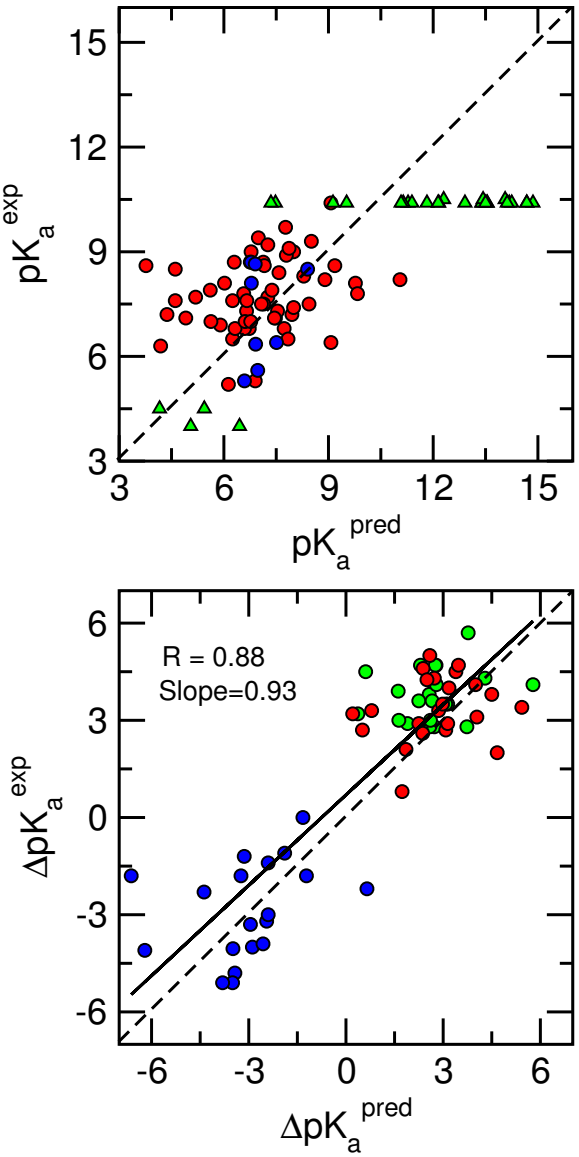

Figure 2.1:
Comparison of predicted and experimental pK_a values of SNase mutants. Comparison between the predicted and experimental pK_a values. Data are divided into three categories: blind predictions with precisely determined experimental pK_a's (red circles); those with experimentally determined upper or lower bounds (green triangles); and calculations with previously published experimental data (blue circles)(*upper*). Comparison between the predicted and experimental pK_a shifts (the difference between the model pK_a and that measured or calculated in the protein) for residues with precisely determined experimental pK_a's. Arg residues are not included because experimental measurements only gave upper or lower bounds of the pK_a values. Residue types are shown separately: Asp (red), Glu (green), and Lys(blue). Model pK_a's for Asp, Glu, Lys, and Arg are 4.0, 4.4, 10.4, 12.5, respectively[92](*lower*).
Linear regression is performed and shown as solid line. Correlation coefficient and slope are also given on the plot.

the prediction is in good agreement with experiment, although there is a systematic underestimation of the magnitude of p$K_{\rm a}$ shifts.

This systematic error is most likely attributable to underestimation of the desolvation penalty for atoms in the protein interior which is a well-known weakness of GB models using overlapping spheres to calculate the solvent-solute dielectric boundary. Underestimation of the desolvation energy is manifested in an underestimation of the p$K_{\rm a}$ shift. We will return to the discussion of this and other limitations of GB-based CpHMD later.

For Asp and Glu the CpHMD method underestimates the magnitude of the p$K_{\rm a}$ shift for 36 out of a total of 45 residues, while for Lys the p$K_{\rm a}$ shift is underestimated for 11 out of a total of 19 residues. Since the p$K_{\rm a}$ shifts of Asp and Glu residues are all positive, underestimation of the shift leads to underestimation of the absolute p$K_{\rm a}$'s as seen in the upper panel of Figure 2.1. Figure 2.2 shows the histogram of the absolute errors in the prediction. 14 or 22% of predictions have an error within 0.5 pK units, and 41 or 64% of predictions have an error within 1.5 pK units. The CpHMD method consistently gives p$K_{\rm a}$ values in good agreement with experiment. Only 2 predictions have an error above 3.5 pK units with the largest being 4.8 pK units. In our attempt to correlate the prediction errors with the characteristics of the titrating site, we found that the only common feature that can be linked to the magnitude of error is how deep the mutation site is buried in the protein. Therefore, we averaged the absolute prediction errors for all residues located at a particular mutation site, and compared the calculated average errors with the distances from the mutation site to the center of the protein. Mutation sites which have an average absolute error above 1.5 pK units are generally more deeply buried, while mutation sites with an average error below 1.5 pK units are found to be located closer to the protein surface, although none of the mutation sites are exposed to solvent. This finding is illustrated in Figure 2.3.

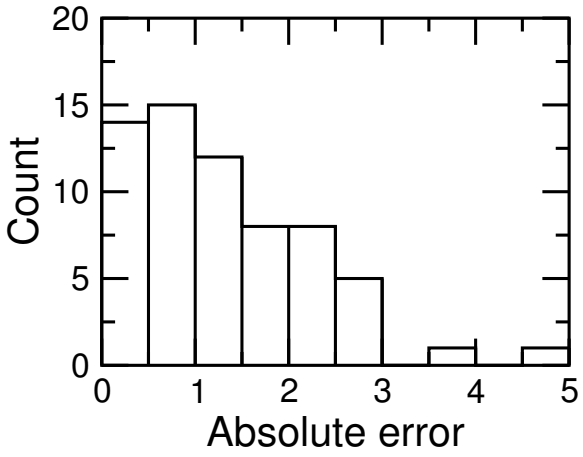

Figure 2.2: Histogram of SNase mutants pK_{a} prediction absolute error. Histogram of the absolute errors of predictions for SNase mutants with precisely determined experimental pK_{a} values.

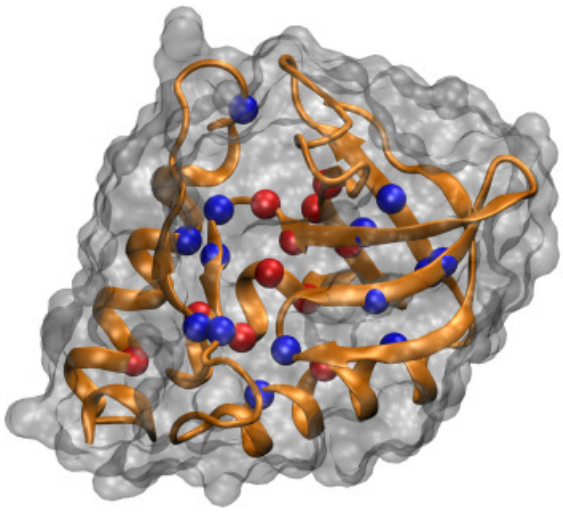

Figure 2.3: Location of mutation site in SNase three-dimensional structure. Location of mutation sites which have an average absolute error for all residues above 1.5 pK units (red) and below 1.5 pK units (blue). Mutation sites with larger errors are generally more deeply buried. Image was rendered using the VMD program [93].

The correlation between the prediction errors and depth of burial can be mainly attributed to the inaccuracy of the generalized Born (GB) implicit-solvent model that underlies the CpHMD simulations as will be discussed later in detail. In what follows we investigate the strengths and limitations of the CpHMD method for p$K_{\rm a}$ predictions. We discuss the physical origins of the inaccuracy in the CpHMD-based p$K_{\rm a}$ predictions.

2.4.2 Strengths of continuous constant-pH molecular dynamics for p$K_{\rm a}$ predictions

Prediction accuracy and conformational relaxation

A major strength of the CpHMD method for p$K_{\rm a}$ prediction is that it explicitly accounts for the effects of conformational dynamics[82]. Based on the prediction data, set we found that the CpHMD method performs equally well for rigid and flexible residues. Table 2.1 shows the root-mean squared fluctuations (RMSF) for residues

Table 2.1: Comparison of p$K_{\rm a}$ accuracy and conformational fluctuation of titratable residues

Mutant	RMSF	CpHMD	Expt.	Abs Error
V23D	0.78	6.7	6.8	0.1
L38D	1.00	6.6	6.8	0.2
A58D	2.26	6.3	6.8	0.5
A90D	1.19	7.1	7.5	0.4
A109D	1.55	7.1	7.5	0.4
N118D	3.36	6.6	7.0	0.4
T41E	2.54	6.3	6.5	0.2
T62E	0.82	7.3	7.7	0.4
A109E	1.55	7.4	7.9	0.5
A132E	1.29	6.8	7.0	0.2

RMSF (Å) refers to the root-mean-squared fluctuation averaged over all atoms in the residue. All root-mean-squared fluctuation (RMSF) values were calculated from TREX-CpHMD trajectories collected at pH 7.

with very accurately predicted p$K_{\rm a}$'s (errors less than 0.5 pK units). The RMSF

values range from below 1 Å (more rigid) to over 3 Å (more flexible), suggesting that the extent of local mobility does not affect the accuracy of pK_a calculations using the CpHMD method. It has been intensively discussed in the literature as to what is the optimal protein internal dielectric constant (ϵ_p) one should use in PB calculations[76,89,94,95]. Assignment of a protein internal dielectric constant is necessary to implicitly account for the dielectric relaxation due to intrinsic dynamics and charging-induced conformational relaxation[94]. By contrast, in CpHMD and other MD-based methods, the internal dielectric constant is set to one while dielectric relaxation of the protein is explicitly captured by the direct coupling between conformational dynamics and protonation equilibria.

We noticed that different mutations have different effects on the conformational flexibility of the protein. Since all mutants in this data set are the result of replacement of a hydrophobic side chain in a solvent-excluded site by an ionizable one, flexibility generally increases in the mutation site.

Convergence of TREX-CpHMD simulations

As noted in the previous work, based on 9 proteins of different fold and size[65], incorporation TREX to enhance conformational sampling[80,81] significantly accelerates the convergence of protonation-state sampling, allowing calculated pK_a values to converge within 1 ns per replica. To examine the convergence in the current prediction set, we evaluated the cumulative fraction of unprotonated state (S value) for all titration simulations. In the majority of the simulations, the cumulative S value plateaus after 600 ps, consistent with our previous estimate[65]. The only exceptions are in the simulations of mutants G20D, T41D, T41E, A90K, N100E, V104D, A109E, and N118E, where the S value continues to change after 1 ns. The time evolution of the S values reveals that prolonged simulations would increase the pK_a's for G20D, A90K, N100E, and A109E, and decrease those for T41D, T41E, and N118E, all of which

would bring the calculation in closer agreement with experiment. For example, Figure 2.4 shows the cumulative unprotonated fractions for the four T41 mutants. The unprotonated fractions for T41K and T41R stabilized within the first one-hundred exchange cycles while for T41D and T41E the values increase over the duration of the simulations.

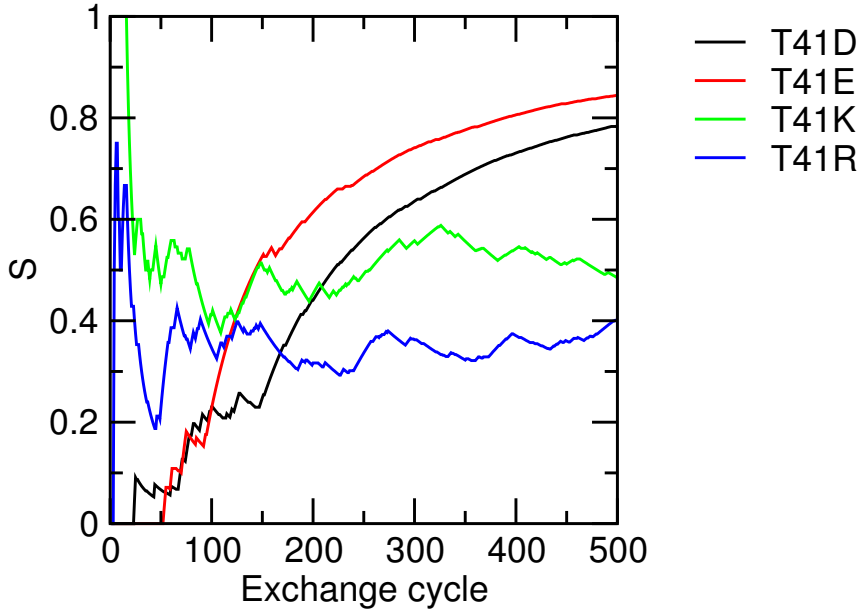

Figure 2.4: Time series of unprotonated fractions of T41 mutants. Cumulative unprotonated fractions of T41 mutants of Δ+PHS. S values were calculated from 298 K replica at pH 7 for T41D and T41E, pH 8.5 for T41K, and pH 14.0 for T41R.

Except for N118E, where the titrating residue is located in a loop, all other titrating residues in the mutants are located in a more rigid secondary structure element and deeply buried in the hydrophobic core of the protein, where local conformational rearrangement is slower as compared to residues that are closer to the surface.

2.4.3 Sources of prediction error

In recent years GB implicit-solvent models have become a powerful tool in theoretical studies of protein dynamics and folding[74]. However, a number of applications have revealed that GB simulations tend to overestimate stability of salt bridges and

protein compaction while underestimating hydrophobic interactions and protein mobility as compared to simulations conducted with explicit water models[83,96,97]. These and other deficiencies in GB models pose limitations on the accuracy of pK_a calculations using the GB-based CpHMD simulations. In an early work[66], it was found that the salt-bridge problem leads to overestimation of the absolute pK_a (down) shifts for solvent-exposed acidic residues. Subsequently, it was shown that this deficiency can be largely overcome[65] through fine tuning the GB input radii for more accurate description of solvent-mediated polar and charged interactions in proteins[98] as well as improvement of protonation-state sampling using the TREX for enhanced conformational sampling[80,81]. Here we investigate two other issues that need to be addressed in order to further enhance the accuracy of CpHMD-based pK_a predictions.

Desolvation energy of deeply buried residues

In the prediction data set, all titrating residues are mutations of hydrophobic internal sites, most of which are deeply buried in the protein. The pK_a shifts of buried residues are dominated by the desolvation energy, which is the reduction in the self-solvation energy, as compared to contributions from the cross term which describes the interactions with the neutral protein background and other titratable residues (see Eq. 1.11). The desolvation energy favors the neutral form, resulting in positive pK_a shifts for acidic residues such as Asp and Glu and negative pK_a shifts for basic residues such as Lys and Arg, as can be seen from the experimental data presented in Figure 2.1. As discussed earlier, our predictions systematically underestimate the absolute pK_a shifts, which suggests an underestimation of the desolvation energy. To test this hypothesis, we used the more accurate GBMV model to estimate the effective Born radii of the titrating atoms (carboxylate oxygen of Asp/Glu or amine nitrogen of Lys) for the structures extracted from the TREX-CpHMD simulations[78,99] and examined the correlation with the degree of burial of the titration site. GBMV offers

more accurate Born radii because it uses the solvent-excluded molecular surface in the integration of solute volume (see later discussions). Figure 2.5 shows the correlation between the calculated relative Born radius using the generalized Born with simple switching function (GBSW) and GBMV model and the depth of burial, measured as the distance, averaged over the trajectory, from the titration site, nitrogen for lysine and the average carboxylate oxygen position for Asp and Glu, to the center of mass of the protein. The relative Born radius is defined as the ratio between the effective

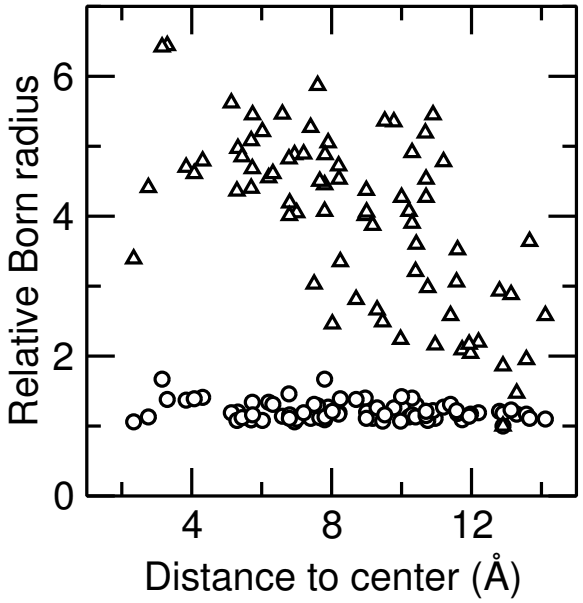

Figure 2.5: Comparision of degree of burial and relative Born radius calculated using two GB models. The degree of burial is measured as the distance from the titration site, nitrogen (Lys) or average carboxylate oxygen position (Asp and Glu) , to the center of mass of the protein. The relative Born radius is defined as the ratio between the effective Born radius of the titration atom and the same atom fully exposed to solvent. Data from the GBSW-based TREX-CpHMD simulations are shown as circles. GBMV calculations (triangles) were performed using the structures from the 298 K replica of the TREX-CpHMD simulations.

Born radius in the protein and the solvent-exposed value. Surprisingly, while the distance to the center covers a wide range, from 3 to 14 Å, the relative Born radius from the GBSW model varies only from 1.1 to 1.7 and does not show a significant correlation with the former. In sharp contrast, the relative Born radii from the GBMV

calculation span a wide range, from 1.5 to 6.4, and are clearly correlated with the depth of burial. We also notice that for atoms with a large distance to the protein center (close to the protein surface), the relative Born radii from both models are similar, while for atoms with a small distance to the protein center (deeply buried), the GBMV model gives relative radii that are 2 to 4 times the values from the GBSW model.

As the self solvation energy is inversely proportional to the effective Born radius, a larger radius means a smaller (absolute) self-solvation energy and larger absolute pK_{a} shift relative to the model value. Thus, the analysis based on Figure 2.5 demonstrates that the use of a more accurate implicit-solvent model such as GBMV would likely improve the prediction of pK_{a} shifts for deeply buried residues. We note that although adjustment of the GB input radii greatly reduces the errors in the calculation of solvent-mediated interactions[98], underestimation of the effective Born radii for buried atoms remains a major problem in GBSW and other GB models which use overlapping van der Waals spheres to represent the solute-solvent dielectric boundary[100]. In these models, the solvent-inaccessible crevices between van der Waals spheres are excluded in the calculation of solute volume leading to underestimation of effective Born radii. The GBMV model, which uses molecular surface, incorporates the solvent re-entrant regions and therefore offers more accurate calculation of effective Born radii.

Local conformational environment of residues close to surface

Although adjustment of the solute-solvent dielectric boundary can largely reduce errors in the description of solvent-mediated interactions, there is a limit to the improvement using the current forms of GB models, as demonstrated previously by others[96] and recently by us[101]. This aspect can be sensitively tested in the pK_{a} calculations of solvent-exposed residues or those close to the solute-solvent interface

using GB-based CpHMD simulations. We examine here the pK_a of A132K. The titrating residue Lys132 is close to the protein surface, and therefore, according to our previous analysis (Figure 2.5), should have a small error in the calculated desolvation energy. The X-ray crystal structure of the background protein shows one water within 4 Å and another water molecule within 7 Å from the mutation site. Given the fact that lysine is larger than alanine, Lys132 of A132K should be more solvated.

Thus, it seems puzzling that the calculated pK_a is around 7.4, 3 units below the experimentally estimated lower bound. To understand the large pK_a error, we examined the conformational environment of Lys132. Lys132 is near the end of the C-terminal helix. As compared to a control simulation with the TIP3P water model, which gives more accurate conformational sampling than GB models, Lys132 and its adjacent residues in the C-terminal helix show significantly increased mobility. The heavy-atom RMSD with respect to the starting structure increased to above 2.5 Å after 500 ps in the CpHMD simulation based on the GBSW model. In contrast, the RMSD was small and remained stable in the explicit-solvent simulation (data not shown). The pronounced mobility of the C-terminal helix in the GBSW based CpHMD simulation may be related to the less accurate representation of the interaction between Lys132 and Glu129 which is expected to stabilize the helix.

In the GB-based simulation, Lys132 and Glu129 form a salt-bridge interaction until Lys132 rotates away (via a switch in χ_3 angle) at 900 ps. In the explicit-solvent simulation, by contrast, the distance between Lys132 and Glu129 samples both salt-bridge like contact and solvent-separated interaction, and the switch in χ_3 angle does not occur.

We note that the distance between Lys132 and Glu129 has a significant impact on the protonation state of Lys132. Thus, we suggest that the large prediction error for the pK_a of Lys132 is due to the limited accuracy of the GB model in representation of the local conformational environment of the titratable side chain.

2.5 Conclusion

Prediction of pK_a shifts of internal residues offers a stringent test of the ability of pK_a prediction methods to describe energetic contributions from desolvation and Coulomb interactions with the environment as well as the effects due to dielectric relaxation. The blind prediction targets comprising buried residues of designed mutants of SNase and variants are particularly challenging because, unlike in naturally occurring proteins, desolvation of the titrating residue is not always compensated by stabilizing Coulomb interaction with nearby charged sites resulting in extremely large pK_a shifts. The data presented shows that TREX-CpHMD titration provides calculated pK_a values for buried residues that are typically within 1.5 pK units. For 87 residues that are blind predictions, the total RMSD from the experimental data is 1.69 pK units, with the majority (64%) of predictions having errors below 1.5 pK units. There are only two outliers giving errors of 4 and 5 pK units, demonstrating the consistency of the method. The performance revealed from the blind prediction data is consistent with the previous benchmark studies based on naturally occurring buried residues[65,82]. Our analysis of the prediction accuracy and conformational relaxation of the titration site supports the notion that dielectric heterogeneity and relaxation need to be explicitly taken into account in a physics-based method in order to quantitatively predict pK_a's of internal residues[76,89,94]. Because of simultaneous protonation and conformational sampling, the CpHMD method has the advantage of allowing the observation of conformational reorganization upon titration, thus providing a dynamic view of the causes of pK_a shifts in proteins, and the observation of mutation-induced conformational relaxation. Our data also demonstrates that, in most cases (see below), the convergence of protonation-state sampling in the TREX-CpHMD simulation is rapid and within 1 ns per replica, in agreement with our previous studies[65,82].

Benchmarking a large number of pK_a's has allowed identification of two ma-

jor limitations of the current CpHMD implementation. The GBSW implicit-solvent model overestimates the solvation energy of deeply buried sites, resulting in the systematic under-prediction of p$K_{\rm a}$ shifts for the most deeply buried residues. While most of large errors occur at the deeply buried sites, we also noticed large deviations from experimental p$K_{\rm a}$'s for surface residues such as in A132K. A close examination of the conformational sampling has revealed a second limitation of the CpHMD method, namely, the inaccurate representation of local conformational environment of the titrating site, which is another prerequisite for accurate p$K_{\rm a}$ calculation using microscopic approaches such as CpHMD. Analysis of p$K_{\rm a}$ convergence has revealed that the conformational rearrangement of deeply buried sites is slow and prolonged simulations (beyond 1 ns per replica) can improve the p$K_{\rm a}$ prediction. This observation suggests that further enhancement of conformational sampling may be necessary. On the other hand, slower conformational dynamics of hydrophobic cores may also be related to the crude approximation of the non-polar solvation energy in the current GB implementation. Recent studies of protein dynamics and folding have shown that GB models with a surface-area dependent term for non-polar solvation can not accurately model hydrophobic interactions[97,102], which is perhaps the major reason for over-compaction[83] and reduced diffusivity in proteins and other hydrophobic assemblies[101]. While progress is being made (see recent development by Levy and coworkers[103]) it remains to be seen whether these deficiencies can be overcome.

In conclusion, we have demonstrated that TREX-CpHMD titration is a reliable tool for prediction of protein p$K_{\rm a}$ values, but there is still room for improvement. To address the limitation due to the underlying solvent model, we have recently developed a hybrid approach where solvent is modeled explicitly in order to provide more accurate conformational sampling, while the free-energy of protonation/deprotonation is estimated using the GB model to provide efficiency and convergence[104]. Ongoing test results are encouraging. We find that by using the explicit-solvent model

for conformation sampling, the errors of pK_{a} prediction for the worst cases in the blind prediction data set are drastically reduced. We have also developed a two-dimensional replica-exchange scheme where a random walk in both temperature and pH space is possible. Ongoing tests suggest that this approach can reduce random errors, which are estimated to be about 0.4 pK units, by a factor of five. These improvements will further enhance the accuracy of CpHMD-based pK_{a} predictions and provide atomically-detailed insights into pH-dependent electrostatic phenomena that are difficult to obtain by experimental measurements.

Chapter 3

Improving protonation state sampling of continuous constant-pH molecular dynamics

One of the limitations of continuous constant-pH molecular dynamics is slow pK_a convergence. To address this issue, we outline two methods to accelerate the convergence of pK_a values.

3.1 Abstract

Comparison of pK_a values calculated from continuous constant-pH molecular dynamics (CpHMD) and experimental data is the most direct method for validating continuous constant-pH molecular dynamics (CpHMD); however, in order for such comparison to be useful, the calculated values must be converged. Two methods for enhancing protonation state sampling in CpHMD are proposed, implemented, and tested on single amino acids.

First, we test the Langevin algorithm in propagating titration coordinates. We find that the stochastic forces introduced in the Langevin approach enhance protonation state convergence for amino acids.

We then apply a pH-replica exchange (pHREX) sampling protocol and analyze convergence of the protonation state populations for amino acids. We use deterministic or Langevin titration with and without pHREX and find that pHREX enhances convergence regardless of which method is used for titration coordinate propagation.

The methods proposed accelerate protonation state sampling for amino acids, a result which will likely carry over to more complex systems such as proteins and make CpHMD simulations a more efficient and useful technique.

3.2 Introduction

In simulations of complex systems, such as proteins, where there are multiple minima separated by energy barriers, it is difficult to accumulate accurate conformational distributions. In CpHMD simulations, this multiple-minima problem is compounded because, in addition to the protein and solvent degrees of freedom, the variability of titrating groups adds additional protonation degrees of freedom. In CpHMD, we are left the challenge of accurately calculating probabilities of each residue being protonated and deprotonated at a certain pH value while simultaneously sampling all energetically accessible conformations.

The problem of accurately and reproducibly sampling distributions of protonation states in the context of CpHMD simulation was recognized early-on. In the initial report of CpHMD simulation of proteins, it was found that even with 1 ns of sampling, some protein pK_{a} values differed by more than 1 pK unit between separate trials[64]. Because pK_{a} values are logarithmic quantities, a deviation of 1 pK unit translates to an order of magnitude change in the relative concentration of protonated versus deprotonated states. The origin of this imprecision was attributed to difficulties in overcoming energy barriers associated with side chain packing and hydrogen-bonding patterns[64].

In an attempt to address the protonation-state sampling issue, which is a specific example of the difficulties encountered when attempting to derive quantities from molecular dynamics (MD) simulations by sampling, CpHMD was extended to allow side chains with equivalent (carboxylate side chains) or *quasi*-equivalent (histidine) proton binding sites to compete for proton uptake[66]. This procedure allowed more rapid interconversion between equivalent protonated forms of these groups and alleviated some of the problems associated with high energy barriers of rotation. Another improvement came with the application of the temperature-replica exchange (TREX) sampling protocol[80,81] to CpHMD which was shown to enhance barrier crossing and

provide more accurate sampling and pK_a calculation[65]. With TREX, the root-mean-squared deviation (RMSD) between 5 trials was reduced to 0.16 and 0.12 pK units for aspartic acid and histidine residues[65].

Although TREX was successful in reducing the deviation between independent trials, this method is computationally expensive. The number of required replicas scales as $O(f^{1/2})$ for a system with f degrees of freedom[105]. Thus, for large systems, and especially for simulations in explicit solvent, TREX quickly becomes prohibitively expensive.

To address the need to increase protonation state sampling, and provide more precise pK_a values without resorting to brute-force TREX, we implemented and tested two separate modifications to the CpHMD technique. The first is the implementation of a Langevin integrator for the propagation of titration coordinates. Langevin dynamics has been shown to accelerate interconversion between equatorial and axial conformations of N-acetylalanyl-N'-methylamide[106]; therefore, we expected that interconversion between protonated and unprotonated states will be enhanced similarly. The second modification is the implementation of a pH-based Hamiltonian replica exchange (pHREX) protocol which specifically targets titration degrees of freedom[104]. Hamiltonian replica exchange[107] has been shown to enhance free-energy calculation convergence in many contexts including the calculation of the association of small molecules to surfaces[108], the calculation of absolute hydration and binding free energies[109,110], and protein folding[111,112]. Hamiltonian replica exchange has also been used for loop modeling and protein structure refinement[113]. This method has the advantage over TREX in that only certain targeted degrees of freedom are excited. This reduces the number of replicas required for efficient exchange between replicas. We find that both modifications systematically improve the convergence of protonation state populations, a result which will be useful in simulations of complex systems such as proteins where obtaining converged protonation state distributions is espe-

cially challenging.

3.3 Methods

3.3.1 Langevin dynamics

In the Langevin dynamics approach to simulate molecular motion, Newton's equation

$$m\ddot{x} = F(t) \tag{3.1}$$

is replaced with Langevin's stochastic differential equation[114]

$$m\ddot{x} = F(t) - \zeta\dot{x} + R(t). \tag{3.2}$$

The influence of a heat bath is modeled as a random force *R(t)*, which is independent of the particle position or velocity, is Gaussian with a zero mean and variance of

$$\langle R(t)R(t')\rangle = 2m\gamma k_b T\delta(t-t') \tag{3.3}$$

where k_b is Boltzmann's constant, T is the absolute temperature, $\delta(t-t')$ is the Dirac delta function, and $\gamma = \zeta/m$.

3.3.2 pH-replica exchange

We outline the general replica exchange approach , but more detailed explanations can be found elsewhere[80,81,107]. In replica exchange, N non-interacting copies, or replicas, of the system are simulated at a ladder of conditions. These conditions may be different temperatures or different force-fields. At each condition, regular molecular dynamics is run and periodically interrupted and an exchange of conditions between a pair of (usually neighboring) replicas is attempted. In the exchange move, detailed balance (the requirement that probability of forward and backward are equivalent)

is imposed so that the process will converge to an equilibrium distribution. The Metropolis-criteria is applied and the exchange probability is defined as

$$P = \begin{cases} 1 & \text{if } \Delta \leq 0 \\ exp(-\Delta) & \text{otherwise,} \end{cases} \tag{3.4}$$

where Δ is the exchange energy. For the case of replicas being simulated at different pH values, the only portion of the total energy that changes is the pH-bias; therefore,

$$\Delta = \beta(U^{\mathrm{pH}}(\{\theta_i\};\mathrm{pH}') + U^{\mathrm{pH}}(\{\theta_i'\};\mathrm{pH}) - U^{\mathrm{pH}}(\{\theta_i\};\mathrm{pH}) - U^{\mathrm{pH}}(\{\theta_i'\};\mathrm{pH}')). \tag{3.5}$$

Here $\beta = (RT)^{-1}$, the first two terms are the pH-biasing potential energies (see Eq. 1.22) for the first and second replica after the exchange, and the last two terms are the corresponding energies before the exchange.

3.3.3 Analysis

To quantify the convergence of the unprotonated fractions, which are used to calculate $\mathrm{p}K_{\mathrm{a}}$ values, we make use of correlation-time analysis and block standard-error analysis.

Correlation-time analysis. The autocorrelation function describes the similarity, or correlation, of a quantity, f at time t, $f(t)$, with the quantity at a later time, $f(t')$. The value can be calculated for a time-ordered trajectory for all values of $t' - t = \Delta t \leq T$, where T is the total simulation length. The autocorrelation function is defined by

$$c_f(t') = \frac{\langle [f(t) - \bar{f}][f(t+t') - \bar{f}] \rangle}{\sigma_f^2} \tag{3.6}$$

where $\langle \dots \rangle$ indicates an average over all Δt intervals considered, $\bar{f}$ is the average of the quantity of interest and σ_f^2 is the variance of f. The autocorrelation function is at a maximum at $c_f(0) = 1$, and tends toward zero as Δt increases and $f(t')$ looses

memory of $f(t)$ [115].

The correlation time, τ_f, quantifies the amount of time required for $f(t')$ to loose all correlation with $f(t)$ and is defined as

$$\tau_f = \int_0^\infty c_f(t')dt' \tag{3.7}$$

where the integration is typically accomplished numerically or the autocorrelation function is fit to an assumed analytic function and a decay rate is extracted from the fitting procedure [115]. In this work the correlation time is calculated by numerically integrating the autocorrelation function.

Block standard-error analysis. In block standard error (BSE) analysis, we attempt to extract an estimate of the statistical error of an average value from a single simulation. This is accomplished by breaking the simulation into M blocks of length n snapshots. The average value is calculated for each block resulting in M values for the average of interest, $\bar{f}$. Next, the standard deviation among the averages is calculated, σ_n, and used to estimate the overall error using the expression

$$BSE(f, n) = \frac{\sigma_n}{\sqrt{M}}. \tag{3.8}$$

At long block-length, the BSE function plateaus and gives a reliable estimate of the error when the block-length becomes significantly greater than the correlation time of f [115,116].

3.3.4 Simulation details

All simulations were carried out using the CHARMM simulation package [39] and the all-atom CHARMM22/CMAP force-field for proteins [36]. The GBSW implicit-solvent model [79] was used with a refined set of atomic input radii [98,117] to define the molecular boundary for the GB calculation. The surface tension coefficient was

set to 0.005 kcal mol^{-1} Å^2. The SHAKE algorithm was applied to all bonds and angles involving hydrogen to allow a 2 fs time step, and conformational dynamics was propagated by the Langevin algorithm with a collision frequency of 5 ps^{-1} at 300 K. For titration simulations using Langevin propagation of titration coordinates, the collision frequency was set to 5 ps^{-1}. In simulations where titration dynamics was propagated by deterministic (non-Langevin) dynamics, a Nosé-Hoover chain was used for temperature control of titration degrees of freedom using the default settings. The Langevin integrator was implemented in the PHMD module of CHARMM, and pHREX was implemented in the MMTSB Tool Set[91]. All simulations were run for 2 ns. Single pH simulation were run at pH 4.0, 4.4, 6.5, and 10.4 for Asp, Glu, His, and Lys, respectively. For pHREX, three pH conditions were used. The pH conditions were 4.0($\pm$1.0), 4.4($\pm$1.0), 6.5($\pm$1.0), and 10.4($\pm$1.0) for Asp, Glu, His, and Lys, respectively. Exchanges were attempted every 500 steps or 1 ps. For each simulation, 10 trials were run with a unique set of initial velocities.

3.4 Results and Discussion

3.4.1 Langevin titration

We begin by examining the convergence of all model compounds using Langevin propagation of titration coordiantes and compare the convergence behavior to simulations using deterministic propagation. Arguably the most robust and straightforward approach for evaluating sampling quality and convergence is to evaluate the deviation between repeated trials[115]. The standard deviation of the unprotonated fractions, σ(S), between 10 trials is shown in Figure 3.1 over the course of the 2 ns simulation. For all four residues, at < 0.75 ns, Langevin titration has deviations lower than those from the deterministic integrator. This improvement is especially pronounced for aspartic acid and histidine for which Langevin titration reduced the deviations

by a factor of approximately two compared to the deterministic results. For all four residues, it is clear that Langevin titration significantly improves sampling for sub-nanosecond simulation times. The improvement for lysine however is marginal and the deviations at 2 ns are equivalent. For histidine and glutamic acid, which have two competing proton binding sites, the deviation is 0.05 from Langevin titration which is at the same level as the residue having a single-proton binding site, lysine, from deterministic titration.

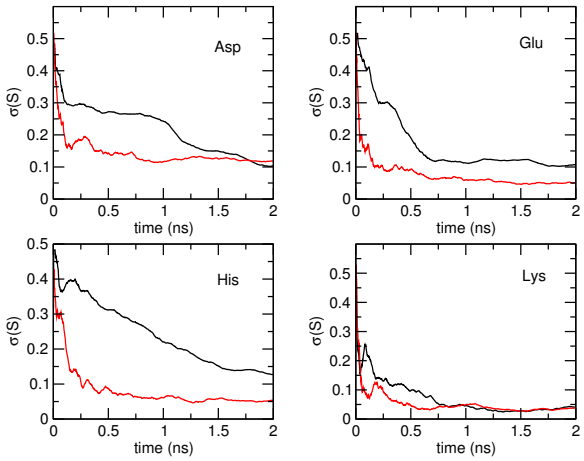

Figure 3.1: Standard deviations of unprotonated fractions using deterministic and Langevin titration. Standard deviations of 10 trials for each model compound as a function of simulation time for simulations using deterministic (black) and Langevin (red) titration coordinate propagation.

For single pH titration simulations which have a continuous trajectory, we can perform correlation-time analysis to examine the rate at which the titration coordinates sample the available protonation states. Autocorrelation functions for the four model compounds are shown in Figure 3.2. In agreement with the previous analysis of the standard deviations, there is little improvement in sampling for lysine using the Langevin titration approach, but for the two-proton binding residues, we see autocorrelation functions which decay much more quickly when titration if performed with Langevin titration. This indicates more rapid sampling of different protonation states, an observation that is made quantitative by calculating the correlation time

(see Eq. 3.7). The correlation times, calculated from the autocorrelation functions shown in Figure 3.2, are listed in Table 3.1. The correlation time for aspartic acid is reduce by a factor of 30, that for glutamic acid is reduced by a factor of 3, and the correlation time for histidine is reduced by an order of magnitude. Again, lysine shows little improvement as the correlation times are indistinguishable given the magnitude of the uncertainty.

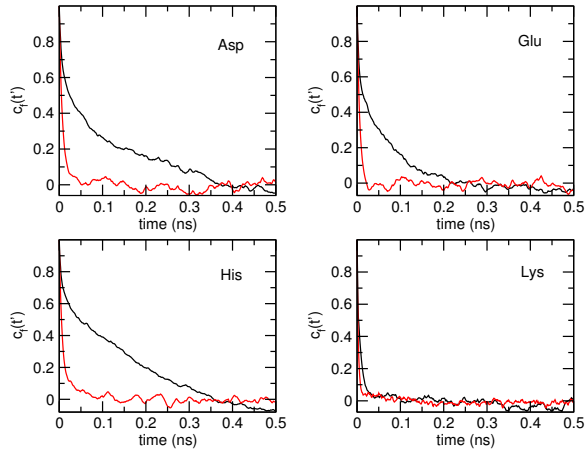

Figure 3.2: Autocorrelation functions of unprotonated fractions using deterministic and Langevin titration. Average autocorrelation function of 10 trials for each model compound as a function of simulation time for simulations using deterministic (black) and Langevin (red) titration coordinate propagation.

Table 3.1: Correlation times using deterministic and Langevin titration.

Residue	Correlation time (ps)	
	Deterministic	Langevin
Asp	87 ± 8	3 ± 2
Glu	43 ± 4	13 ± 2
His	110 ± 5	13 ± 2
Lys	0.2 ± 2	0.3 ± 1

Correlation times are defined by equation 3.7 and calculated by numerically integrating the autocorrelation functions of Figure 3.2. Errors given are the standard deviation between correlation times calculated for 10 separate trials.

3.4.2 pH-replica exchange with deterministic titration

We next examine convergence behavior when we apply pHREX. Figure 3.3 shows the standard deviation between the 10 trials over the course of the 2 ns simulation using deterministic titration with and without pH-exchange. Similar to the results from Langevin titration, for the residues that have two competing protonation sites pHREX reduces the deviation significantly for sub-nanosecond simulation time. Aspartic acid and glutamic acid benefit the most from pHREX sampling. At 2 ns the deviation between trials is reduced by approximately one-half for both of the acidic model compounds. The deviation resulting from pHREX simulation of lysine shows no improvement, and at the end of the 2 ns simulation the same is true for histidine.

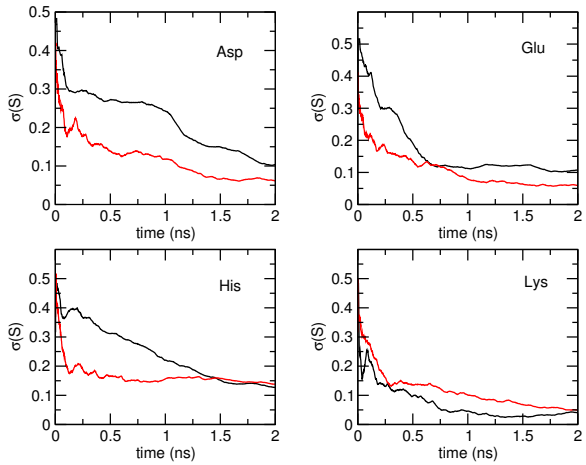

Figure 3.3: Standard deviations of unprotonated fractions from deterministic titration with and without pH-replica exchange. Standard deviation of 10 trials for each model compound as a function of simulation time for simulations using without pHREX (black) and with pHREX (red).

Since pHREX does not result in a continuous trajectories, application of correlation time analysis is not appropriate; however, we can make use of BSE analysis. The results from BSE analysis for pHREX simulations using deterministic titration are shown in Figure 3.4. With BSE, the error is estimated from the limiting value of the BSE curve. We first note that for lysine pHREX appears to offer no improvement. For the other residues which have two protonation sites, the error from BSE analysis

indicates that pH-exchange reduces the deviation by a factor of two.

Comparing the error estimates from the standard deviation between the trials (Figure 3.3) and the BSE analysis (Figure 3.4), we find that both methods give results that are encouragingly similar. This suggests that BSE is a reasonably reliable method for estimating error the statistical error without resorting to running multiple separate simulations.

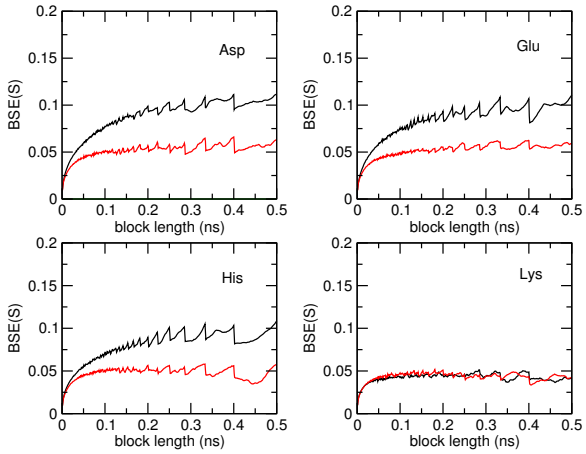

Figure 3.4: Block error analysis of unprotonated fractions from deterministic titration with and without pH-replica exchange. Average BSE of 10 trials for each model compound for simulations using deterministic propagation without pHREX (black) and with pHREX (red).

3.4.3 pH-replica exchange with Langevin titration

Finally, we examine the convergence behavior of model compounds when we combine Langevin titration with pHREX by employing the same analysis as before. As shown in Figure 3.5, the results using Langevin titration with and without pHREX are virtually identical for glutamic acid and lysine. Histidine has slightly greater deviation between trials with pHREX, while deviation for aspartic acid is significantly reduced when using Langevin titration with pHREX.

However, the results from BSE, shown in Figure 3.6, suggest that the combination of Langevin titration with pHREX slightly reduces the statistical uncertainty for all

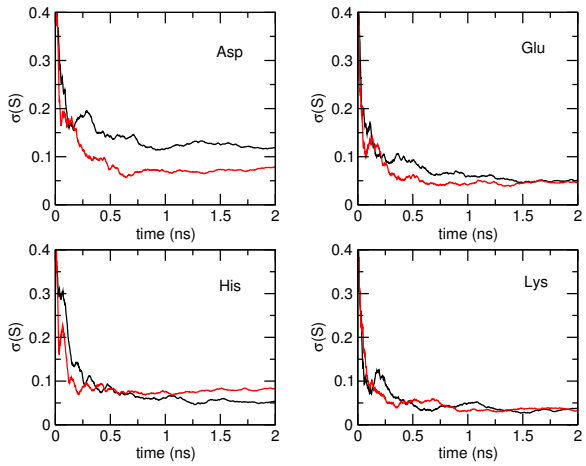

Figure 3.5: Standard deviations of unprotonated fractions using Langevin titration with and without pH-replica exchange. Standard deviation of 10 trials for each model compound as a function of simulation time for simulations using Langevin titration without pHREX (black) and with pHREX (red).

four model compounds.

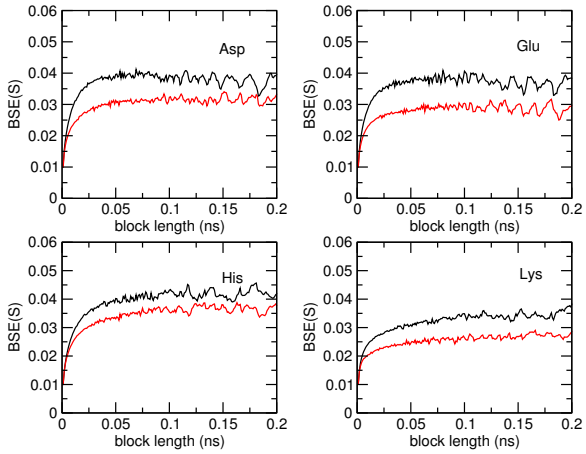

Figure 3.6: Block error analysis of unprotonated fractions using Langevin titration with and without pH-replica exchange. Average BSE of 10 trials for each model compound for simulations using Langevin titration without pHREX (black) and with pHREX (red).

3.5 Conclusion

In this work we have extended the CpHMD method[64,66,82] to allow Langevin propagation of titration coordinates and have implemented a variant of Hamiltonian

exchange specific to CpHMD simulations where the applied pH is used as the biasing coordinate. We have tested these techniques on four model compounds (Asp, Glu, His, and Lys) in implicit solvent.

We find that Langevin titration reduces the correlation time of the titration coordinates significantly for the three model compounds which have two competing proton-binding sites (Asp, Glu, and His), but for Lys, which has only a single titrating proton, we see no change in the correlation time. Langevin titration also reduces the standard deviation between replicate trials for residues with two proton binding sites. The application of pHREX reduces the deviation between trials for all four models compounds, but for aspartic acid and lysine the deviation between trials is similar at the end of the 2 ns simulation time. BSE analysis indicated reduced uncertainty for all residues except lysine when pHREX is applied. Combining Langevin titration with pHREX results in final standard deviations of around 0.05 calculated between the 10 trials for all four model compounds and slightly less when calculated using BSE. By combining Langevin titration with pHREX, the deviation between trials was reduced by approximately one-half when compared to the initial results using single-pH deterministic titration.

The modifications to CpHMD described, a Langevin integrator and pHREX, have been shown to improve the methods' ability to provide converged protonation-state populations of model compounds. Future work will be focused on using these methods for the calculation of pK_{a} values in proteins.

Chapter 4

Continuous constant-pH molecular dynamics in explicit solvent with pH-based replica exchange

In our tests of generalized Born based continuous constant pH molecular dynamics, one of the deficiencies of the method identified was a structural bias toward compact and rigid conformations that was inherited from the implicit-solvent model. In order to alleviate this problem, we developed a hybrid-solvent method that derives conformational states from the more accurate explicit-solvent representation while using the generalized Born model for protonation state energetics.

The following content was published in :

Journal of Chemical Theory and Computation

volume 7, pages 2617-2629, 2011

4.1 Abstract

A computational tool that offers accurate pK_{a} values and atomically detailed knowledge of protonation-coupled conformational dynamics is valuable for elucidating mechanisms of energy transduction processes in biology such as enzyme catalysis, electron transfer, as well as proton and drug transport. Towards this goal we present a new technique of embedding continuous constant-pH molecular dynamics within an explicit-solvent representation. In this technique we make use of the efficiency of the generalized Born (GB) implicit-solvent model for estimating the free energy of protein solvation, while propagating conformational dynamics using the more accurate explicit-solvent model. Also, we employ a pH-based replica exchange scheme to significantly enhance both protonation and conformational state sampling. Bench-

mark data of five proteins including HP36, NTL9, BBL, HEWL, and SNase yield an average absolute deviation of 0.53 and a root-mean-squared deviation of 0.74 from experimental data. This level of accuracy is obtained with simulation lengths of 1 ns per replica. Detailed analysis reveals that explicit-solvent sampling provides increased accuracy relative to the previous GB-based method by preserving the native structure, providing a more realistic description of conformational flexibility of the hydrophobic cluster, and correctly modeling solvent mediated ion-pair interactions. Thus, we anticipate that the new technique will emerge as a practical tool to capture ionization equilibria while enabling an intimate view of ionization-coupled conformational dynamics that is difficult to delineate with experimental techniques alone.

4.2 Introduction

Solution pH has a profound effect on the stability and function of proteins by changing the protonation states of titratable groups. Proteins can become denatured under extreme pH conditions. Enzymes are often catalytically active in a narrow pH range[22]. Protein-protein interactions[16] and protein-ligand binding[17] are also modulated by the protonation states of titratable groups. Accurate determination of active-site pK_{a} values informs about the catalytic mechanism of proteins[118]. Knowledge of the native- and denatured-state pK_{a} values can be used to quantify electrostatic effects on protein stability[119].

Although the importance of solution pH has long been recognized, molecular simulation techniques have traditionally neglected it. In a standard molecular dynamics (MD) simulation the protonation states of ionizable side chains are set at the beginning of the simulation based on the comparison of the desired pH condition and the solution or model compound pK_{a} values. This fixed protonation scheme can be a source of error in several instances. For example, if the pK_{a} values are near the pH of interest the protonated and deprotonated states should coexist, which obviously

is not reflected in simulation with fixed protonation states. Additionally, even when reasonable protonation states may be set for the initial conformation, conformational rearrangement may favor an entirely new set of protonation states.

In recent years, considerable effort has been made to develop methodologies that explicitly include pH as an external parameter in MD simulations, similar to temperature, allowing protonation states of ionizable groups to respond to changes in the chemical environment and external pH[52,54,56,60,63,64]. These constant-pH techniques differ in the way protonation states are updated. In the discrete methods, protonation states are periodically updated using Monte-Carlo sampling, while in the continuous approach titration coordinates are introduced and propagated simultaneously with the spatial coordinates (see a most recent review[82]). One of the most promising constant-pH techniques, termed continuous constant-pH molecular dynamics (CpHMD)[64,66] is based on the λ-dynamics approach to free-energy calculations[77], allowing ionizable groups to switch continuously between protonated and unprotonated forms. Protonation and deprotonation is accomplished in a manner similar to many free energy simulation techniques, where an alchemical coordinate, λ, is introduced. The novelty of the λ-dynamics approach lies in the fact that the alchemical coordinate is assigned to a fictitious λ-particle and the force on the particles is derived analytically. CpHMD has been shown to give accurate and robust predictions for protein pK_{a} values[82] and has opened a door to theoretical studies of pH-dependent protein dynamics and folding[67,68,120].

In the aforementioned CpHMD method, the generalized Born (GB) implicit-solvent model is used to calculate forces on both spatial and titration coordinates. The major advantage of using GB models in constant-pH methodologies is that convergence of pK_{a}'s can be achieved with a reasonable amount of sampling time, which has not been demonstrated feasible with explicit-solvent models (see more discussions later). Another benefit of using GB models within the CpHMD framework is that

forces on the titration coordinates can be computed analytically. However, as CpHMD and other GB-based constant-pH techniques are maturing into practical tools, problems inherited from the underlying GB models are becoming the limiting factor for further improvement of accuracy. Recent GB simulation studies have revealed several problems that seem difficult to overcome. Specifically, attractive electrostatic interactions are overestimated[98,121], and improvement through adjustment of GB input radii that define dielectric boundary[98,121] is limited[96]. Also, due to the lack of solvent granularity, GB simulations cannot reproduce the solvation peaks seen in the interaction free energy profiles from explicit-solvent simulations[98]. Furthermore, there have been noted problems with the stability of hydrophobic interactions[97,102], and overly compact and rigid unfolded states[83,122], which are likely due to the approximate nature of the non-polar solvation term based on solvent-accessible surface area (SA model). Finally, the inaccuracies of the GB/SA model in the representation of electrostatic and non-polar energetics result in a more favorable sampling of helical relative to extended states[83,123].

The limitations of GB models affect the accuracy and applicability of the CpHMD method in several ways. First, a small error in the electrostatic solvation energy calculated by the GB model alters the relative deprotonation free energy in reference to solution and therefore the pK_{a} shift. This type of "electrostatic" error is typically small for solvent-exposed residues, because the GB model, in particular GBSW used in this work, has been tuned to reproduce the explicit-solvent data of solvent-exposed polar or charged interactions[98]. However, the "electrostatic" error becomes significant for deeply buried residues[65,82] because the inaccuracy in the desolvation energies of deeply buried atoms remains an unsolved problem in GB models. Nonetheless, the electrostatic error is systematic[124] and a post correction may be introduced if necessary. The second type of GB-related error which affects the accuracy of $\Delta\Delta G^{\mathrm{deprot}}$ arises from the small distortion in the conformation or distribution of conforma-

tions. The impact of this "conformational" error on the protonation-state sampling is typically not systematic and the extent of the error is unpredictable. Finally, the dependence of conformational sampling on the GB model also hinders the application of the CpHMD method to poly-ionic systems such as DNA and RNA for which GB models are not well suited.

In light of the above considerations, we introduce here a method to extend the CpHMD framework to explicit-solvent simulations. In principle, forces on both spatial and titration coordinates can be derived from explicit-solvent sampling. However, the latter is not practical because a lengthy simulation time is required to accurately compute solvation-related forces based on explicit-solvent sampling. Consequently, we devise a method which takes advantage of the efficiency of the GB model to compute solvation forces on titration coordinates while propagating conformational dynamics via all-atom interactions in explicit solvent. Additionally, we implement a replica-exchange protocol based on the pH biasing energy to significantly accelerate the convergence of the simultaneous sampling of protonation and conformational space. Thus, by making use of the more accurate explicit-solvent sampling, the new method aims to improve the accuracy of CpHMD by reducing the aforementioned "conformational" error and to allow applications to many problems where implicit-solvent models are not feasible.

The rest of the chapter is organized as follows. First, we describe the explicit-solvent CpHMD method and the pH-based replica-exchange protocol. We then examine potential artifacts due to the use of both explicit- implicit-solvent schemes and the response of solvent molecules to titration. Next, we present and discuss results of model compound titrations and analyze the convergence behavior with the new sampling protocol. Finally, we benchmark the accuracy of the new method by calculating pK_{a} values of five proteins including HP36, NTL9, BBL, HEWL, and SNase. We compare the results with the GB-based CpHMD simulations and experiment. We

find that the explicit-solvent CpHMD offers slightly more accurate pK_{a} predictions but significantly deeper physical insights. Surprisingly, convergence of the explicit-solvent CpHMD titrations is achieved for all proteins with a simulation length of 1 ns per replica, suggesting that the new method will emerge as a powerful and practical tool for theoretical studies of electrostatic phenomena.

4.3 Methods

4.3.1 Continuous constant-pH molecular dynamics in explicit solvent

The key to CpHMD and other continuous titration methods is to simultaneously derive forces on the spatial and titration coordinates. While it is straightforward to compute forces on spatial coordinates in explicit-solvent simulations, there is inherent difficulty in the latter due to the need for very accurate estimate of the electrostatic desolvation free energy. In fact, attempts to directly calculate the free energy of charging titratable residues repeatedly during molecular dynamics by considering explicit interactions between solvent molecules and solute have encountered severe convergence problems in the context of both discrete[56] and continuous constant-pH MD methods[63,125]. Our own tests revealed that the variance in the instantaneous forces on the titration coordinates are up to an order of 100 kcal/mol per lambda unit, whereas the forces exerted from the pH biasing energy 1 pH unit away from the model compound pK_{a} is only 1.3 kcal/mol per lambda unit. Therefore, we decided to use a "mixed-solvent" scheme, where the GB model is used to derive forces on the titration coordinates, while the explicit-solvent model is used to propagate the spatial coordinates. To enable a direct coupling between solvent dynamics and proton titration of solute, we retain the λ-dependent scaling of van der Waals interactions involving titrating hydrogens and solvent molecules. An analogous "mixed-solvent" scheme has been developed by Baptista and coworkers and applied in the context of

the discrete constant-pH MD for protein titration studies[52]. One important difference is that their scheme does not include a direct (van der Waals) coupling between solvent dynamics and solute titration.

The caveat of the "mixed-solvent" scheme is that no formal Hamiltonian exists and potential artifacts may occur. Since the solvation-related force on titration coordinates is treated in a mean-field manner without explicitly accounting for the electrostatic interactions with nearby water molecules, inadequate or lagged response of solvent to the change in the charge state of the titrating site may occur. We expect this undesirable side effect to be minimal because of the aforementioned van der Waals coupling between solute protonation and solvent dynamics, and because in continuous evolution of titration coordinates, the energy change is small at each time step. Nevertheless, a preventive fix is to decrease the update frequency for λ coordinates (currently the same as spatial coordinates), thereby allowing relaxation of surrounding solvent molecules. Such a strategy has been demonstrated to be very effective in the discrete constant-pH molecular dynamics simulations using the "mixed-solvent" scheme[52]. Another source for potential artifacts in this and other "mixed-solvent" simulations is related to the fact that the total energy is no longer strictly conserved, which may result in a drift or pronounced fluctuation in temperature and energy of the system. We will examine these potential artifacts later in detail.

4.3.2 pH-replica exchange

It has been noted previously[60,64,66] that in constant-pH molecular dynamics the convergence of protonation-state sampling and resulting pK_{a} values is slow due to the tight coupling of conformational dynamics and protonation equilibria. To address this issue the temperature-replica exchange (TREX) protocol[80,81,126] was applied to enhance conformational sampling in the GB-based continuous[65] and discrete[61] constant-pH methods which has led to significant improvement in the convergence of

calculated pK_a values. A straightforward implementation of the TREX protocol in explicit-solvent simulations is however not effective because of the large number of replicas needed to account for the solvent degrees of freedom[127]. Recently, Simmerling and coworkers have proposed a mixed-solvent scheme to reduce the number of replicas[128], which may be incorporated into the explicit-solvent CpHMD presented in this work. One issue that was noted[128] and is currently being addressed[121], is the distorted conformational distribution due to inaccuracy of the underlying implicit-solvent model. To avoid this problem we decided to enhance the sampling of protonation space directly by making use of a replica exchange protocol based on the pH-biasing energy (Eq. 1.22). This protocol is a specific application of the reaction-coordinate replica-exchange method[107]. The reader is referred to Chapter 3 for a description of the pH-based replica exchange method.

4.3.3 Simulation details

Model compounds

As in the previous work[65,66], model compounds for Asp, Glu, His, and Lys side chains are single amino acids acetylated at N-terminus (ACE), and N-methylamidated at C-terminus (CT3). The model pK_a values (used in Eq. 6.4) were 4.0, 4.4, and 10.4 for Asp, Glu, and Lys, respectively[92]. The model pK_a of His was taken as 6.6 and 7.0 for the Nδ and Nϵ sites, respectively[129]. The model compound for the C-terminus attached to phenylalanine (CT-Phe) in HP36 was the acetylated C-terminal hexapeptide (KEKGLF) from HP36 with a measured pK_a of 3.2[130]. The parameters in the potential of mean force function U^{mod} were determined using thermodynamic integration (TI) in explicit solvent.[66]. Parametrization simulations at each combination of λ, and x for double-site titratable residues, were run for 1 ns. In the TI procedure, the protonation states of other titratable residues in the model peptide for CT-Phe were fixed because their pK_a's are at least 1 pH unit higher than the C-terminus.

Except for CT-Phe the ionic strength in the GB calculation was set to zero during the TI simulations following the previous protocol[66]. For CT-Phe the ionic strength was 150 mM in accord with experiment[130].

Proteins

Five proteins were studied in this work: the 45-residue binding domain of 2-oxoglutarate dehydrogenase multi-enzyme complex, BBL (PDB: 1W4H), the 36-residue subdomain of villin headpiece, HP36 (PDB: 1VII), the 56-residue N-terminal domain of ribosomal L9 protein, NTL9 (PDB: 1CQU), the 149-residue, of which 129 residues were resolved in the crystal structure, hyper-stable variant of staphyloccal nuclease Δ+PHS, SNase (PDB:3BDC), and the 129-residue hen egg white lysozyme, HEWL (PDB:2LZT). For all structures, the HBUILD facility of CHARMM[39] was used to add hydrogens. Unless otherwise specified, no explicit ions were added in the pHREX simulation because of the small simulation box and low ionic strengths used in experiment. See later discussions. The ionic strengths in the GB calculations were set to 200, 150, 100, 100, and 50 mM for BBL, HP36, NTL9, SNASE, and HEWL, respectively, consistent with the experimental conditions[89,130–133]. Unless otherwise noted, both N- and C-termini of proteins were left in the free, charged form. For SNase, the published crystal structure was missing residues 1-6 and 142-149. To avoid potential errors, the structure was acetylated at N-terminus and amidated at C-terminus. For NTL9, the C-terminus was amidated in accord with experiment[132].

Simulation protocol

We have implemented the explicit-solvent CpHMD method in a developmental version of CHARMM (c35b3)[39], and the pHREX sampling scheme in the MMTSB Tool Set[91]. All of the simulations described in this work were performed with the all-atom CHARMM22/CMAP force field[36] and the modified CHARMM version of the TIP3P water model[134] The solvation forces on the titration coordinates were calculated using the GBSW implicit-solvent model[79] with the refined[98] atomic input radii

of Nina et. al.[117]. The SHAKE algorithm was applied to all bonds and angles involving hydrogen to allow a 2-fs time step. Non-bonded electrostatic interactions were calculated using the particle-mesh Ewald summation with a charge correction to reduce pressure and energy artifacts for systems with a net charge[135]. In the GB calculation, all input parameters were identical to the previous work[65].

All simulations were performed under ambient pressure and temperature conditions using the Hoover thermostat[136] with Langevin piston pressure coupling algorithm[137]. Proteins and model compounds were built and then placed in a truncated octahedron water box of a size such that the distance between the solute and edges of the box was at least 14 Å. Water molecules within 2.6 Å of any heavy atom of the solute were deleted. Energy minimization was carried out in three stages. First, a harmonic restraint with a force constant of 50 kcal/(mol·Å) was applied to solute heavy atoms and the structure was energy minimized with 50 steps of the steepest descent (SD) and 200 steps of the adoptive basis Newton-Ralphson (ABNR) methods. Then the force constant was reduced to 25 kcal/(mol·Å) and the same minimization protocol was applied. Finally, the force constant was reduced to 10 kcal/(mol·Å) and the structure was energy minimized with 5 SD and 20 ABNR steps.

In the pHREX simulation of a model compound, three pH replicas, one at the reference pK_{a} and two at 1 pH unit above and below the reference value were used. Three independent pHREX simulations were conducted, where each simulation lasted 1.2 ns per replica and the first 200 ps was discarded in the pK_{a} calculation. For proteins, one pHREX simulation was performed. In the pHREX protocol, the pH spacing was 1 pH unit and the pH range extended at least 1 unit above and below the highest and lowest experimentally determined pK_{a} value for the protein. Specifically, for BBL the pH range is 2 to 9, for HP36, NTL9, and SNase it is 0 to 7, and for HEWL it is 0 to 9. Each pH replica was subjected to 4 ps of restrained equilibration without pH exchange, where a harmonic potential with the force constant of 10 kcal/(mol·Å) was

applied to all solute heavy atoms. Following equilibration, unrestrained simulation with the pHREX protocol was performed. The exchange in pH was attempted every 100 dynamic steps or 0.2 ps for model compound and 500 steps or 1 ps for protein simulations. The success rate for exchanges was at least 40%. Protein simulations lasted 2 ns and the first 0.25 ns was discarded in the analysis and pK_{a} calculation. Simulation of HP36 was run for 4 ns in order to observe pK_{a} behavior at longer simulation times.

Calculation of p$\boldsymbol{K}_{\mathrm{a}}$ values

To calculate the pK_{a} of a titratable site, we first recorded the population of protonated ($\lambda < 0.1$, N^{prot}) and unprotonated ($\lambda > 0.9$, N^{unprot}) states from simulations of different pH replicas. The resulting unprotonated fractions S at multiple pH values were then fitted to the following modified Hill equation, in accord with the commonly used model for fitting pH-dependent NMR chemical shifts[89],

$$S(\mathrm{pH}) = \frac{s_{\mathrm{A^-}} + s_{\mathrm{HA}} 10^{n(pK_a - \mathrm{pH})}}{1 + 10^{n(pK_a - \mathrm{pH})}}, \tag{4.1}$$

where n is the Hill coefficient, which represents the slope of the transition region of the titration curve[89], $s_{\mathrm{A_}}$ and s_{HA} are fitting parameters, which represent the extrapolated S values at extreme acidic and basic pH conditions for the observed titration event. Equation 4.1 becomes the Hill equation when protonation or deprotonation is complete in the simulated pH range, e.g., $s_{\mathrm{A^-}} = 1$ and $s_{\mathrm{HA}} = 0$, which was the case for nearly all residues. Occasionally, for acidic residues with significant negative pK_{a} shifts, s_{HA} deviated significantly from 0 as a result of incomplete protonation at the lowest pH condition. Finally, to account for the small systematic deviations of calculated pK_{a}'s of model compounds relative to the reference values, we made the following post-corrections, Asp (+0.2), Glu (+0.3), and His (-0.3) to the pK_{a} values of proteins.

4.4 Results and Discussion

4.4.1 Trajectory stability

Before applying explicit solvent CpHMD to titration simulations, it is important to examine potential artifacts due to caveats in the mixed scheme and the change in total net charge. As mentioned earlier, the proposed method does not conserve energy because the protonation states of titratable groups are changed using an implicit description of the electrostatic interactions with solvent, which may lead to drift or increased fluctuation in temperature and energy of the simulated system. Another source for potential artifacts is related to the fluctuating net charge of the system during proton titration. In the default implementation of Ewald summation a neutralizing plasma, which is a uniform distribution of a charge equal and opposite to the net charge, is added to the summation to avoid divergence in Coulomb energy for periodic systems[138]. This background plasma has been noted to introduce pressure artifacts for small net-charged systems, which could dramatically affect the dynamics of simulations at constant pressure[135]. Brooks and coworkers showed that the artifacts are drastically reduced by invoking a charge correction term[135]. We applied this correction term in all of our simulations.

To assess the extent of the spurious effects, we examined the temperature, pressure, and total potential energy of the system along the trajectory using two protocols. In the first protocol, a blocked lysine was subjected to CpHMD titration at pH 10.4. In the second protocol, a fixed-charge simulation was conducted using a neutral lysine with an otherwise identical simulation setup. As shown in Figure 4.1, the time series for temperature, pressure, and potential energy in the CpHMD titration of lysine (with 1:1 protonated and deprotonated states) is virtually indistinguishable from the conventional simulation of neutral lysine with fixed protonation state. The pressure fluctuations are quite large for both systems, but this is expected because of the small

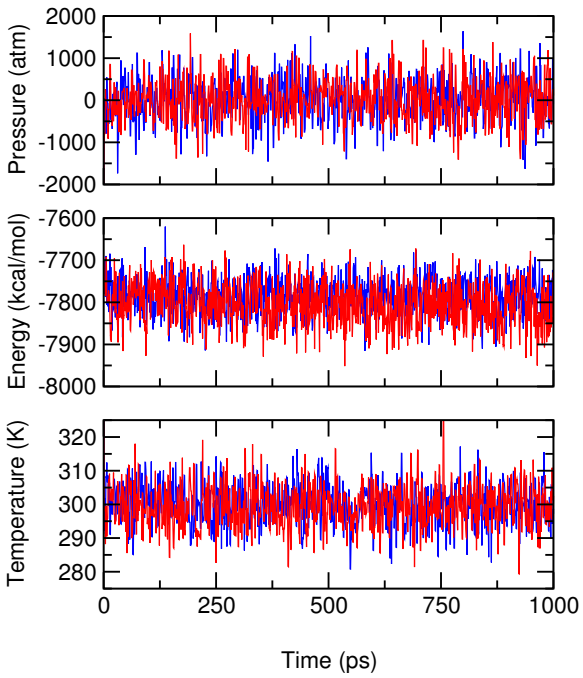

Figure 4.1: Comparison of pressure, energy, and temperature from fixed charged and continuous constant-pH molecular dynamics in explicit solvent simulations. Instantaneous pressure, potential energy, and temperature in the explicit-solvent CpHMD simulation of lysine at pH 10.4 (blue) and in the simulation of the neutral lysine with fixed protonation state (red).

size of the simulation box. Also, any energy leaking into or out of the system due to the non-conservative change in protonation state is not readily apparent as there is no visible drift in the total potential energy for this system. To further verify the stability of pressure, temperature, and potential energy, we performed CpHMD titrations for other model compounds and proteins. No systematic drift or increased fluctuation was observed in any of the three quantities at the simulation timescales (several nanoseconds) for either model compounds or proteins. Thus, we conclude that, with the net-charge correction and the Hoover thermostat, potential artifacts in pressure, temperature and potential energy are negligible.

4.4.2 Response of explicit solvent to titration

Although the van der Waals interactions between titratable hydrogen atoms and solvent molecules are explicitly described, the lack of explicit treatment of electrostatic interactions may have an undesirable effect such that water molecules cannot adjust quickly to a low energy position following a change in the titration coordinate. This could result in an unrealistic arrangement of solvent around the titrating site. To examine the response of explicit water molecules to solute titration, we calculated the radial distribution function (radial distribution function (RDF)) for the charged (protonated) and neutral (unprotonated) lysine from the (conventional) simulations (one for charged and one for neutral) and compared them with the RDF's from one CpHMD titration simulation. The latter simulation was conducted at a pH condition such that the charged and neutral populations are almost equal. As seen in Figure 4.2, the positions of maxima and minima in the RDF's of the charged and neutral forms of Lys are identical in the conventional simulations and CpHMD titration, which demonstrates that the water structure is qualitatively indistinguishable. To further investigate the reorientation of water molecules in response to titration, we took a closer look at the solute-solvent interactions that give rise to the peaks of the RDF's.

Interestingly and reassuringly, the relative orientation of lysine and the nearby water is identical in the conventional simulations and CpHMD titration. Figure 4.2 also shows the representative snapshots of the charged and neutral lysines interacting with an adjacent water molecule. When lysine is charged, it acts as a hydrogen-bond donor, interacting with the oxygen atom of water. When lysine is neutral, it acts as a hydrogen-bond acceptor, interacting with the hydrogen atom of water.

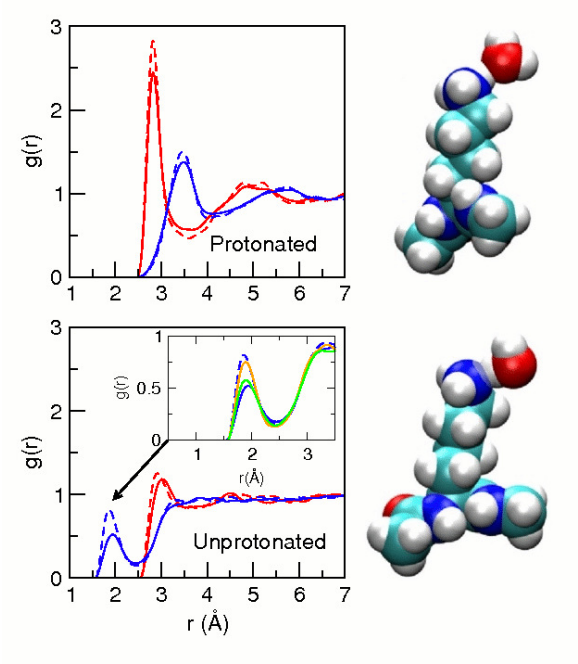

Figure 4.2: Response of explicit solvent molecules to explicit-solvent using continuous constant-pH molecular dynamics in explicit solvent. Radial distribution function for the titratable nitrogen atom of lysine to the hydrogen (blue) or oxygen (red) atom of water. Dashed lines are from the simulation with the fixed protonation state; solid lines are from the CpHMD titration with protonated (charged) and deprotonated (neutral) states coexisting. Snapshots of the interacting water and lysine are shown. The charged lysine donates a hydrogen bond to water (*upper*), while the neutral lysine accepts a hydrogen bond from water (*lower*). Simulations with fixed protonation states were run for 1 ns. The CpHMD titration time was 2 ns and the deprotonated fraction was about 0.5. The in-set gives RDFs when a very stringent cut-off ($\lambda >$ 0.99) is used to define the deprotonated state (green) and when λ values are updated every 10 MD steps in addition to the stringent cut-off (orange). Images were rendered using the VMD program [93].

Despite the remarkable agreement in the positions of maxima and minima of the RDF's, the amplitude of the peaks from the CpHMD titration is reduced as

compared to those from conventional simulations. This reduction in the amplitude of RDF can be mainly attributed to the slight lag in water equilibration following a switch in protonation state, and to a lesser extent the cut-off chosen in our definition of protonated and deprotonated states. The in-set in Figure 4.2 shows that with a very stringent cut-off ($\lambda > 0.99$) there is small improvement in the amplitude of the RDF. If we use the stringent cut-off combined with the λ-update of every 10 MD steps the amplitude of the RDF is dramatically increased to nearly superimpose on the result from the simulation with fixed protonation state. If the frequency of switching protonation state is much slower, the RDF's would exactly match those calculated from the simulations at fixed charge. Baptista and coworkers showed that in the MD simulation, the reorganization time of water following the most dramatic protonation event from the fully neutral to doubly charged state of succinic acid is 1-3 ps[52]. Considering the average residence time at either protonation state in our simulation was on average about 1 ps and the transition between protonation states is continuous, water molecules have sufficient time to rotate to a favorable position following titration. Nevertheless, the data of lysine titration shows that the update frequency or time step for propagation of titration coordinates (currently set to be the same as the propagation of conformational dynamics) can be increased to ensure the full extent of water relaxation. A drawback is the slow down of protonation-state sampling.

4.4.3 Convergence and accuracy of model compound titrations

Before attempting to perform titration simulations of proteins, it is important to assess the required simulation time to reach converged values for the unprotonated fraction (S) of model compounds as well as the accuracy and precision of the calculated pK_a's. We first examine titration simulations conducted at a single pH value. Explicit-solvent CpHMD titration of a blocked lysine was performed at the pH equal

to the reference pK_a of 10.4. The S values stabilized at about 5 ns and there was little change over the remainder of the 10-ns simulation. We repeated the simulation twice with different randomly assigned velocities and observed a similar convergence time. Similar results were also found for the blocked Asp, Glu, and His which have two titration sites. The lengthy simulation time (5 ns) required for the convergence of pK_a values for single amino acids indicates the need for accelerated sampling. To directly enhance the protonation-state sampling, we applied the pH-based replica-exchange protocol with three replicas placed at pH values of 9.4, 10.4, and 11.4 in the lysine titration. The S values were converged within 1 ns for all model compounds, demonstrating significant acceleration over the single-pH simulation. We summarize these results in Table 4.1.

The uncertainty, or random error, in the calculated model compound pK_a's ranges from 0.02 to 0.11, which is similar to the range found in potentiometric and NMR titration experiments (see Table 4.1). To further assess convergence, we examine the reproducibility of S values and quality of fitting to the Henderson-Hasselbach equation. In Figure 4.3, results of three independent pHREX simulations (1 ns per replica) for Asp, Glu, His, and Lys are shown. The error in the S value ranges from 0.02 to 0.12, and the χ-square values of the fitting is virtually zero. Thus, the above data demonstrate that 1-ns pHREX titrations offer converged sampling for protonation equilibria.

Next we examine the accuracy of the calculated pK_a's of model compounds. As compared to the target reference values, the pK_a's of Asp, Glu and Lys are underestimated by 0.2–0.3 pH units while that of His is overestimated by 0.3 pH units (Table 4.1). There are two possible sources for the systematic deviations. The first possibility has to do with artifacts in simulations of net-charged periodic systems using Ewald potential. Even with the net-charge correction, Brooks and coworkers noted that the charged form may be slightly favored in the free energy simulation of a

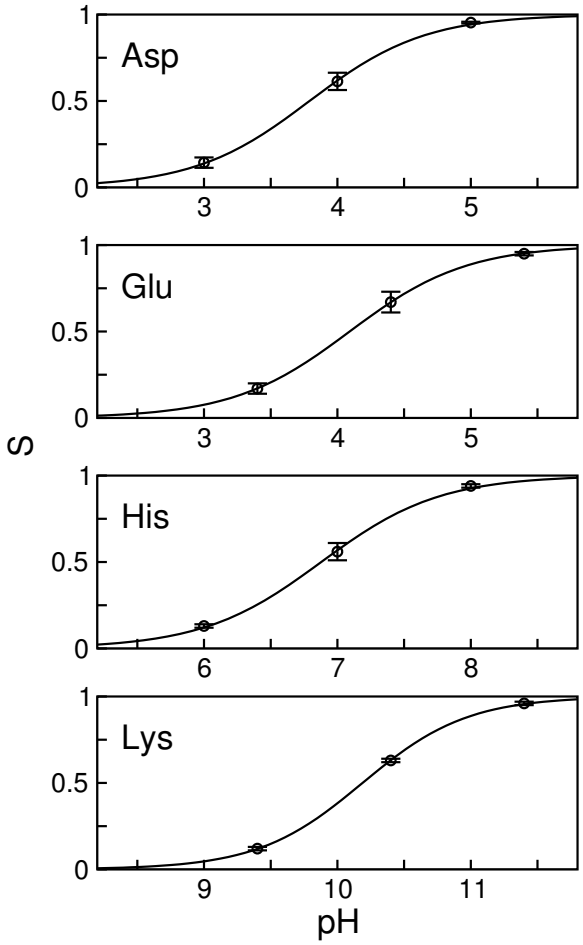

Figure 4.3: Titration curves for the blocked model compounds from continuous constant-pH molecular dynamics in explicit solvent simulations. Three independent pH-REX simulations were performed. Each REX simulation utilized three pH replicas with each replica running for 1 ns. The average unprotonated fractions S (calculated from the three runs and shown as circles) at three pH values were fit to the Henderson-Hasselbach equation and shown as lines. At each pH, an error bar indicates the range of the calculated S values, which is the largest at the pH closest to the pK_{a} value. These ranges are 0.10, 0.12, 0.10 and 0.02 for Asp, Glu, His, and Lys, respectively.

Table 4.1: Calculated and experimental pK_a values of model compounds

Residue	Calc[a]	Calc[b]	Ref[c]	Pace lab[d]
Asp	3.79±0.09	3.77±0.02	4.0	3.67±0.04
Glu	4.09±0.11	4.05±0.01	4.4	4.25±0.05
His	6.89±0.08	6.89±0.01	6.6/7.0	6.54±0.04
Lys	10.21±0.02	10.41±0.02	10.4	10.40±0.08
CT-Phe	3.38±0.06		3.2[e]	-

[a] Results using the standard simulation protocol where the λ value was updated every MD step and simulation length was 1.2 ns per pH replica. The average pK_a's obtained by fitting S data from three independent pHREX titrations are listed along with one half the difference between the highest and lowest calculated values. [b] Results from test simulations where the λ value was updated every 10 MD steps and the simulation length was 10 ns per pH replica. [c] Measured pK_a's based on the blocked single amino acids from Nozaki and Tanford[92]. These model pK_a's were used in the pH-biasing energy (Eq. 6.4). For His, the listed pK_a's are the microscopic values for δ and ϵ sites. The resulting macroscopic pK_a is 6.45[66]. Errors in the measurements are typically ±0.1–0.2[139]. [d] The most recent data from Pace lab based on potentiometric titrations of alanine pentapeptide Ac-AA-X-AA-NH_2 where X denotes the titrating residue[139]. [e] Measured pK_a of the C-terminal carboxylic acid in the C-terminal peptide of HP36 (sequence KEKGLF) based on the NMR titration data from Raleigh lab[130].

single ion and this deviation depends on the size of the simulation box[135]. Our tests however showed that increasing the box size did not affect the pK_a results for model compounds. We further ruled out the net-charge related artifact because the same systematic errors, e.g., underestimation of the pK_a's for Asp and Glu and overestimation of the pK_a for His were also observed in the GB-based CpHMD simulations[65].

The systematic errors in pK_a's indicate that the deprotonation free energy based on the potential of mean force function which is determined by the thermodynamic integration (TI) procedure does not exactly match that in the titration simulation. One possible reason for the discrepancy is the difference in water relaxation because in the TI simulation water has more time to relax at a specific λ value than in the titration simulation. To investigate this issue, we repeated the titrations with slower λ dynamics, updating λ value every 10 MD steps. Interestingly, the deviation for the pK_a of Lys is abolished but the deviation for Asp, Glu and His remains. Examination of the λ and x trajectories revealed that the two degenerate protonation states (doubly deprotonated in the case of Asp or Glu and doubly protonated in the case of His) occasionally experience prolonged residence time. In the absence of extensive analysis and consideration, we suggest that one route for correcting this bias is to make the barrier in the x (tautomeric) dimension a function of λ such that when λ approaches the degenerate protonation state interconversion becomes increasingly difficult. This is clearly a limitation that needs to be addressed in our future work. Nevertheless, since this bias is present in both model compound and protein titrations, the effect on the calculated pK_a shifts is negligible. To correct for the systematic deviations, we added post corrections for all the calculated pK_a values of proteins (see Simulation details).

4.4.4 Enhanced sampling of protonation and conformational states of proteins

We have demonstrated that the pHREX protocol significantly accelerates the pK_{a} convergence for model compounds. Now we show that the pHREX protocol significantly enhances sampling in both protonation and conformational space for proteins. Take the titration of HP36 as an example. Figure 4.4 displays the time series of the unprotonated fraction for Asp44 from one pHREX simulation and three single-pH simulations. In the single-pH simulations, Asp44 was trapped in the unprotonated form at pH 2.3 as a result of a persistent salt-bridge interaction with Arg15. In the pHREX simulation however, both protonated and unprotonated forms of Asp44 were sampled at pH 2 and pH 3, because the simulation was able to capture both formation and disruption of the salt bridge. Thus, by making use of the direct coupling between protonation events and conformational dynamics, the pHREX protocol allows the protein to overcome local energy barriers, while retaining the correct thermodynamic distribution. In this regard, pHREX has a similar effect as the TREX protocol, which significantly accelerates the sampling convergence of both protonation and conformational states in the GB-based CpHMD simulations[65].

4.4.5 Convergence and overall accuracy of protein titrations

In order for titration simulations to be practical, protonation-state sampling needs to converge within a reasonable amount of time. While we have shown that 1 ns of pHREX titration is sufficient for obtaining converged pK_{a}'s for model compounds, we also observed that 1-ns titration also yields converged pK_{a}'s for proteins, despite the fact that the degrees of freedom in a protein system may be orders of magnitude greater as compared to a model compound. This seemingly surprising observation is consistent with data from the GB-based CpHMD simulations[65,82], and can be attributed to the fact that pK_{a}'s are mainly determined by local environment. To

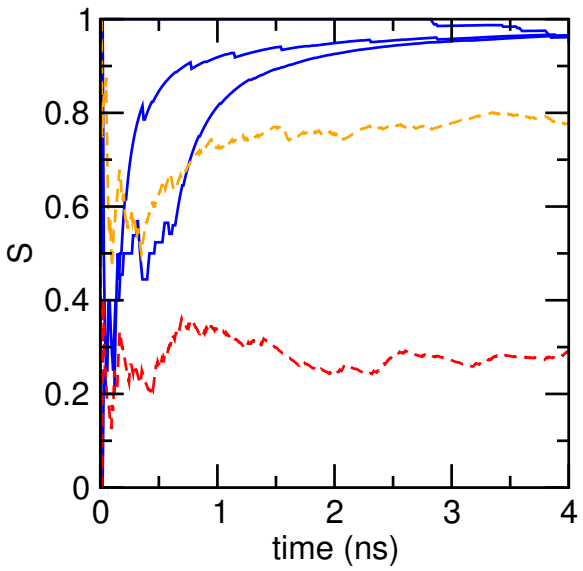

Figure 4.4: Enhancement of protonation-state and conformational sampling of protein pH-replica exchange. Cumulative unprotonated fraction of Asp44 of HP36. Data from the pHREX simulations are shown in red for replica at pH 2 and orange for replica at pH 3. Data from three independent single-pH simulations at pH 2.3 are shown in blue.

illustrate the rapid convergence in protein titrations, we monitor the times series of the S value and pK_{a} as well as the quality of fitting. In Figure 4.4 we can see that the S values for HP36 stabilize at 1 ns. The small fluctuation after 1 ns does not cause noticeable change in the pK_{a} value because of the logarithmic relationship between S and pK_{a}. Figure 4.5A shows that, after only a few hundred ps the calculated pK_{a}'s of the two histidines in BBL become stable and do not change in the remaining simulation time. This is encouraging given the fact that one of the histidines is buried and as such may require more sampling. Another indication of convergence is the quality of fitting to the HH equation. Figure 4.5B shows nearly perfect fits ($R^2 > 0.95$) for both residues based on the 1-ns titration data.

To assess the overall accuracy of the explicit-solvent CpHMD method, we performed titration on five test proteins, HP36, BBL, NTL9, SNase, and HEWL, and compared the calculated pK_{a}'s with experiment as well as the GB-based simulations, where the latter used the same pHREX protocol and salt as well as temperature con-

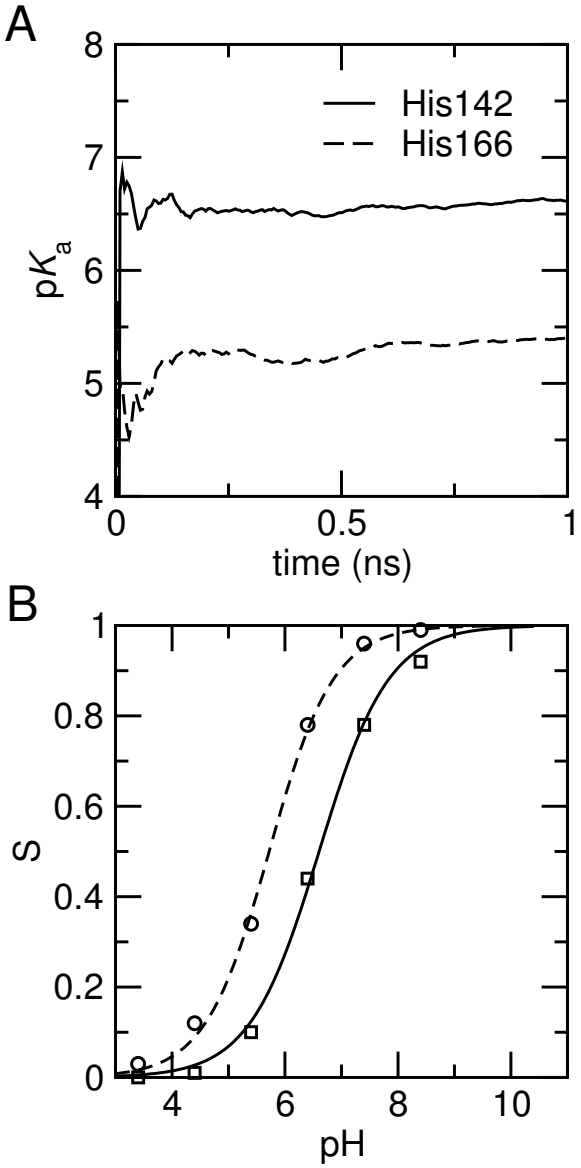

Figure 4.5: Convergence of p$K_{\rm a}$ values of BBL using continuous constant-pH molecular dynamics in explicit solvent. A. Time series of the calculated pK_a's for BBL from the explicit-solvent CpHMD simulations with pHREX protocol. The S values at pH 7 and 6 are used for His144 and His166, respectively. B. Titration data based on the 1-ns simulation and best fits to the modified HH equation (Eq. 4.1).

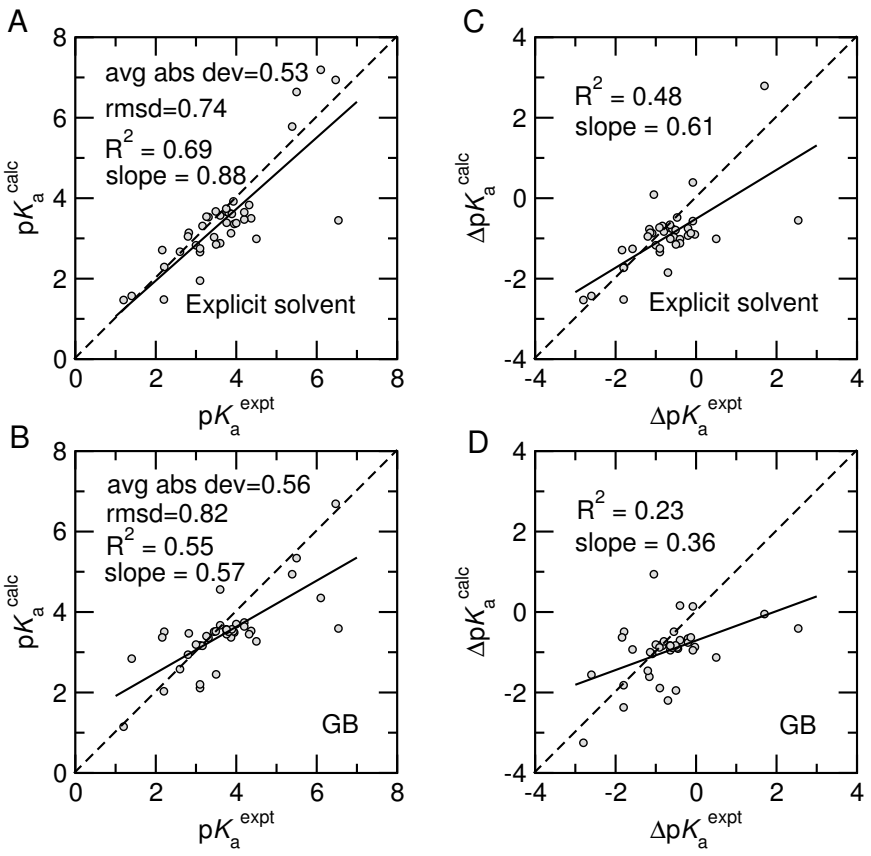

Figure 4.6: Comparison between calculated and experimental p$K_{\rm a}$ values and p$K_{\rm a}$ shifts relative to model values. Calculated p$K_{\rm a}$ values from the explicit-solvent and GB-based titrations are shown in A and B, respectively. Calculated p$K_{\rm a}$ shift from explicit-solvent and GB-based titrations are shown in C and D, respectively. Regression line (solid), slope, and R^2 value are shown on each plot as wells as y=x line (dashed) to facilitate visual comparison.

ditions. The results are presented in Table 4.2, 4.4, and 4.5 along with the estimates of statistical uncertainty, which was calculated as half of the difference between the pK_a's calculated from the first and last half of the 750 ps simulation. The total simulation length was 1 ns and the data from the first 250 ps was discarded. As a validation of convergence, the pK_a's calculated using 2-ns simulations are also listed. In reference to experimental data, the overall root-mean-squared deviation (RMSD) from the explicit-solvent titrations is 0.57, which is slightly lower than the RMSD from the GB-based titrations (0.68). As a more informative measure of calculation accuracy, linear regressions of the calculated versus measured pK_a shifts are shown in Figure 4.6 for the explicit-solvent and GB simulations. While the R^2 value and slope are 0.50 and 0.80 respectively from the explicit-solvent titrations, they are 0.24 and 0.45 from the GB titrations. Since the correlations are relatively low, we repeated the regression analysis by removing the data points with the four largest absolute pK_a shifts. The R^2 value from the explicit-solvent titrations dropped from 0.48 to 0.25, while R^2 from the GB simulations also dropped dramatically, from 0.23 to 0.06. Thus, the results show that the improvement due to explicit solvent is robust. Since the data set comprised of mainly acidic residues, the slopes being below 1 suggests that both simulations overestimate the negative pK_a shifts or underestimate the pK_a's. A close examination of the correlations reveals that the significantly improved agreement with experiment in the explicit-solvent titrations is due to reduction of relatively large errors for several groups. Thus, overall the explicit-solvent simulations offer increased accuracy for predicting protein pK_a's. The reasons in specific cases will be delineated next.

Small proteins BBL, HP36, and NTL9

We first examine the performance of the explicit-solvent CpHMD titrations for three small proteins with 36 to 56 residues and all-α as well as mixed α-β topologies. The results are listed in Table 4.2 along with the GB titration data. The convergence

of both explicit- and implicit-solvent titrations is excellent. The largest difference between the pK_a's calculated from the first and last half of the simulation is 0.3 units. Extending the explicit-solvent simulations to 2 ns leads to a pK_a change below 0.15 units and does not improve the agreement with experiment. Overall, the explicit-solvent data is similar to the GB data. The RMSD as well average absolute and maximum deviations from experiment in the explicit-solvent titration are 0.50, 0.44 and 0.87, respectively, similar to the GB titration. The deviations from experiment arise from the overestimation of the negative pK_a shifts of acidic residues in both explicit- and implicit-solvent titrations.

We examine two cases where the pK_a's from the explicit-solvent titration are at least 0.6 pH units different from the GB titration. In both cases, the explicit-solvent titration improves agreement with experiment. Asp23 is a residue where the explicit-solvent titration reduces the overestimation of the pK_a downshift of Asp23 from 0.9 to 0.3 units. This is because the salt-bridge interaction with the nearby amino terminus was over-stabilized in the GB simulation, a known problem in GB models[66].

His166 is the only buried residue in this data set. While being excluded from solvent, it also interacts with three nearby lysines. Thus, both desolvation and electrostatic repulsion destabilize the protonated or charged form of His166, leading to a downward pK_a shift relative to the model value. This is reflected in the experimental pK_a of 5.39, about 1.1 pH units lower than the model value. In the explicit-solvent titration the pK_a shift is underestimated by 0.41 pH units while it is overestimated in the GB titration by 0.55 pH units. Detailed analysis of the trajectories reveals the major cause of the difference to be structural. Figure 4.7A shows that in the explicit-solvent simulation, the conformations stayed close to the starting structure with the backbone RMSD centered at 2 Å. In the GB simulation, however, a conformational cluster developed that significantly deviates from the initial structure with the backbone RMSD centered at 4.9 Å. Figure 4.7B shows that while His166 is

Table 4.2: Calculated and experimental pK_{a} values of HP36, BBL, and NTL9

Residue	Explicit solvent[b]	GB	Expt[a]
BBL			
His142	6.94 ± 0.06 (6.83)	6.47 ± 0.03	6.47 ± 0.04
His166	5.78 ± 0.04 (5.90)	4.84 ± 0.19	5.39 ± 0.02
HP36			
Asp44	2.66 ± 0.09 (2.77)	3.17 ± 0.11	3.10 ± 0.01
Glu45	3.36 ± 0.31 (3.28)	3.49 ± 0.09	3.95 ± 0.01
Asp46	3.03 ± 0.09 (3.12)	3.51 ± 0.03	3.45 ± 0.12
Glu72	3.50 ± 0.21 (3.45)	3.53 ± 0.10	4.37 ± 0.03
CT-Phe	3.31 ± 0.20 (3.16)	3.16 ± 0.14	3.09 ± 0.01
			3.24 ± 0.12
NTL9			
Asp8	2.83 ± 0.07 (2.80)	3.19 ± 0.20	2.99 ± 0.05
Glu17	3.57 ± 0.14 (3.50)	3.67 ± 0.13	3.57 ± 0.05
Asp23	2.75 ± 0.16 (2.82)	2.11 ± 0.11	3.05 ± 0.04
Glu38	3.38 ± 0.30 (3.40)	3.70 ± 0.19	4.04 ± 0.05
Glu48	3.47 ± 0.17 (3.42)	3.74 ± 0.20	4.21 ± 0.08
Glu54	3.65 ± 0.22 (3.49)	3.64 ± 0.08	4.21 ± 0.08
Avg abs dev	0.44 (0.45)	0.36	
RMSD	0.50 (0.52)	0.47	
Max abs dev	0.87 (0.92)	0.99	

[a] pK_{a}'s determined by NMR titration for BBL[131], HP36[130], and NTL9[132]. [b] Values in parentheses were obtained from the 2-ns simulation.

slightly exposed to solvent in the explicit-solvent simulation it is fully enclosed in the GB simulation. Examination of the average distances to the nearby lysines reveals that the Coulomb interactions in both explicit-solvent and GB simulations are similar. Therefore, we suggest that the overestimation of the p$K_{\rm a}$ shift for His166 in the GB simulation is mainly due to the overestimation of desolvation penalty as a result of exaggerated cloistering of His166. Reduced mobility especially of buried sites has been also observed in other GB simulations[83].

Table 4.3: Effects of adding explicit ions on calculated p$K_{\rm a}$ values of NTL9

Residue	Calc[b]	Ions[c]	Expt[a]
Asp8	2.83 ± 0.07	2.91 ± 0.31	2.99 ± 0.05
Glu17	3.57 ± 0.14	3.38 ± 0.19	3.57 ± 0.05
Asp23	2.75 ± 0.16	2.98 ± 0.16	3.05 ± 0.04
Glu38	3.38 ± 0.30	3.48 ± 0.04	4.04 ± 0.05
Glu48	3.47 ± 0.17	3.42 ± 0.34	4.21 ± 0.08
Glu54	3.65 ± 0.22	3.52 ± 0.25	4.21 ± 0.08
Avg abs dev	0.41	0.40	
RMSD	0.48	0.49	
Max abs dev	0.73	0.78	

[a] pK_a's determined by NMR titration[132]. [b] Calculated p$K_{\rm a}$'s from explicit-solvent titrations without counter ions (as listed in Table 4.2). [c] Calculated p$K_{\rm a}$'s from simulations with an identical set up except for the addition of Cl^- ions such that the net charge of the protein at all pH conditions was minimized.

Although for these small proteins the explicit-solvent pK_a calculations are quite accurate, it is important to further discuss another issue concerning the explicit-solvent CpHMD method. Since the net charge is changing and may become large depending on the protonation state of the protein, we examined the effect of adding an approximate number of counter ions to minimize the net charge of the system in all pH conditions. Because of the large number of basic residues of NTL9 and the resulting net positive charge, NTL9 is an ideal test case to quantify the magnitude of the effect. As shown in Table 4.3 the calculated p$K_{\rm a}$ values in the simulations

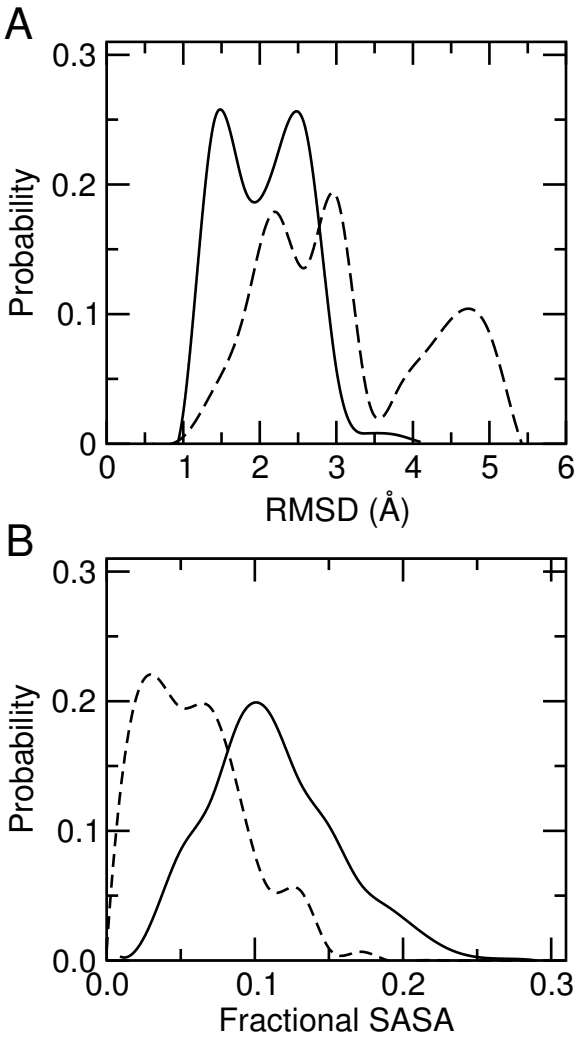

Figure 4.7: Comparison of BBL conformations from explicit- and implicit-solvent simulation. Structural comparison of BBL from explicit-solvent (solid) and GB (dashed) simulations at pH 5 A. Probability distributions of backbone RMSD. B. Ratio of the solvent accessible surface area (SASA) of His166 in BBL relative to the solvent-exposed value.

with neutralizing counter ions are virtually identical to those where no net-charge neutralizing ions were added. Thus, at least for the short simulation time required to obtain converged pK_a values, the data indicates that it is not necessary to include neutralizing ions.

SNase

The calculated pK_a's for a larger protein, a hyper-stable variant of the 149-residue SNase, are summarized in Table 4.4. SNase is a good test system because the structure-based continuum calculations gave very poor agreement with experiment presumably due to the lack of explicit treatment of protein flexibility [89]. Overall, the explicit-solvent titration offer a better agreement with experiment. The RMSD as well as the average absolute and maximum deviations in the explicit-solvent titration are 0.86, 0.46, and 3.09, respectively, while they are 0.96, 0.63, and 2.95 in the GB titration. Extending the explicit-solvent simulations to 2 ns give results that are very similar.

We first examine Asp95, for which the explicit-solvent titration was able to reduce the overestimation of pK_a from the GB-based titration from 1.21 to 0.55 units. The major reason for the improvement is related to the strength of the interaction with Lys70. In the crystal structure obtained at pH 8 the minimum distance between the charge centers on Asp95 and Lys70 is 4.7 Å, which suggests a salt-bridge interaction. Figure 4.8 shows the probability distribution of the minimum distance between the charge centers from the explicit-solvent and GB simulations. Although the average distance is identical at 6.1 Å, the difference lies in the distribution. The GB simulation sampled a uni-modal distribution centered around 7 Å. By contrast, the explicit-solvent simulation sampled two distinct populations, one centered at 2.8 Å, representing the conformations where Asp95 and Lys70 are closely associated, and another one centered at 7.1 Å, representing the conformations where the two side chains are rotated away from each other. The minimum region between the two populations

corresponds to the solvent-bridged conformations. The bimodal distribution seen in the explicit-solvent simulation is a direct result of including discrete solvent molecules and reflects a more realistic description of the ion pair interaction. However, the GB simulation neglects solvent granularity and models the ion-pair interaction in a mean-field manner, which results in a less tight salt-bridge pairing and an underestimation of the pK_a shift for Asp95.

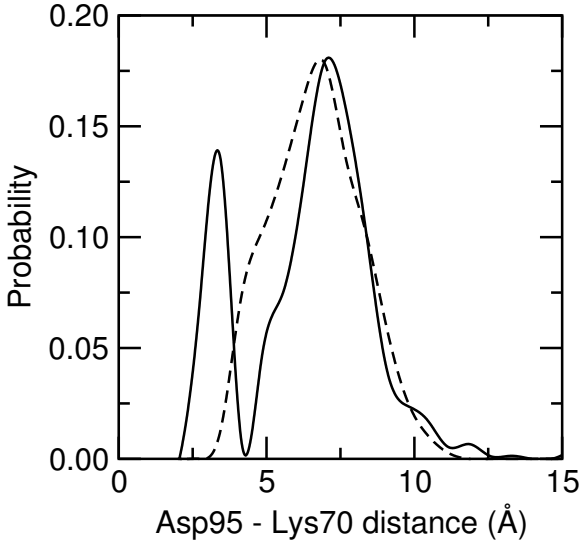

Figure 4.8: Comparison of Lys70-Asp95 of SNase salt-bridge distribution from explicit- and implicit-solvent simulation. Probability distribution of the minimum distance between the carboxylate oxygens of Asp95 and amino nitrogen of Lys70 of SNase from the explicit-solvent (solid) and GB (dashed) titrations at pH 3.

Another case where the inclusion of explicit solvent resulted in the more accurate pK_a calculation is for Asp77. The experimental measurement provides an upper bound of 2.2 for the pK_a. In the explicit-solvent simulation, the pK_a was calculated to be in the correct range, but in the GB simulation the pK_a shift was underestimated by at least 1 pH unit. Asp77 is within a hydrogen-bond distance of two backbone amide hydrogens of Asn119 and Thr120, which are located in a loop connecting a β-sheet motif to an α-helix (Figure 4.9, upper left snapshot). In Figure 4.9 we monitor the minimum distance between the carboxylate oxygens of Asp77 and the

Table 4.4: Calculated and experimental pK_{a} values of SNase

Residue	Explicit Solvent[c]	GB	Expt[a]
Glu10	3.14 ± 0.09 (3.33)	3.47 ± 0.01	2.82 ± 0.07
Asp19	2.29 ± 0.15 (2.49)	3.51 ± 0.02	2.21 ± 0.07[b]
			6.54 ± 0.06
Asp21	3.45 ± 0.28 (3.55)	3.59 ± 0.00	3.01 ± 0.01
			6.54 ± 0.02[b]
Asp40	3.13 ± 0.23 (3.35)	3.37 ± 0.09	3.87 ± 0.09
Glu43	3.83 ± 0.08 (3.76)	3.45 ± 0.00	4.32 ± 0.04
Glu52	3.92 ± 0.01 (3.88)	3.52 ± 0.02	3.93 ± 0.08
Glu57	3.67 ± 0.16 (3.64)	3.52 ± 0.01	3.49 ± 0.09
Glu67	3.66 ± 0.06 (3.67)	3.45 ± 0.06	3.76 ± 0.07
Glu73	3.53 ± 0.11 (3.54)	3.36 ± 0.13	3.31 ± 0.01
Glu75	3.54 ± 0.27 (3.58)	3.40 ± 0.06	3.26 ± 0.05
Asp77	< 0.0 (< 0.0)	3.14 ± 0.03	< 2.2
Asp83	2.54 ± 0.12 (2.84)	3.50 ± 0.04	< 2.2
Asp95	2.71 ± 0.57 (2.97)	3.37 ± 0.06	2.16± 0.07
Glu101	3.64 ± 0.11 (3.67)	3.51 ± 0.01	3.81± 0.10
Glu122	3.61 ± 0.03 (3.75)	3.57 ± 0.01	3.89± 0.09
Glu129	3.74 ± 0.11 (3.71)	3.57 ± 0.12	3.75± 0.09
Glu135	3.39 ± 0.20 (3.44)	3.56 ± 0.03	3.76± 0.08
Avg abs dev	0.46 (0.48)	0.63	
RMSD	0.86 (0.85)	0.96	
Max abs dev	3.09 (3.00)	2.95	

[a] pK_{a} determined by NMR titration[89]. [b] The major transition when the experimental data was fit to a two-pK_{a} model. [c] Values in parantheses were obtained from the 2-ns simulation.

backbone amide hydrogen of Asn119 or Thr120. In the explicit-solvent simulation the distance was stable, fluctuating around 2 Å during the entire trajectory, revealing that the backbone hydrogen bonding between Asp77 and Asn119/Thr120 was intact. However, in the GB simulation, this interaction was disrupted as a result of the high mobility of the aforementioned loop (see Figure 4.9, upper right snapshot). This analysis suggests that the underestimation of the pK_a shift for Asp77 in the GB simulation is due to the distortion of local structure.

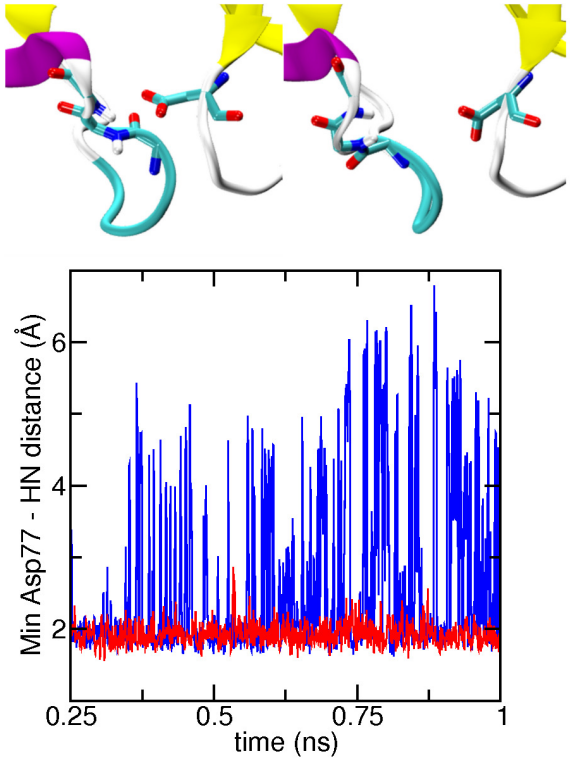

Figure 4.9: Comparison of Asp77 of SNase backbone hydrogen bond from explicit- and implicit solvent simulation. Comparison of the local environment of Asp77 of SNase from the explicit- and implicit-solvent titrations at pH 3. *Upper*. In the initial structure Asp77 forms backbone hydrogen bonds with Asn119 and Thr120 (left snapshot). These interactions were broken in the GB simulation (right snapshot). *Lower*. Time series of the minimum distance between the carboxylate oxygens of Asp77 and the backbone amide hydrogen of Thr119 or Asn120 from the explicit-solvent (red) and GB (blue) simulations at pH 3. Images were rendered using the VMD program[93].

The largest pK_{a} error from the explicit- and implicit-solvent titrations is for

Asp21, which interacts with Asp19 on the other end of the β-hairpin. NMR titration data showed two distinct transitions for the two residues[89]. The major transitions have the pK_{a} of 2.21, assigned to Asp19, and 6.54, assigned to Asp21[89].

The latter is the only upward shifted pK_{a} relative to the model value for SNase. Both the explicit- and implicit-solvent titrations were not able to reproduce the direction of the pK_a shift for Asp21 and underestimated the pK_{a} by about 3 pH units, although the explicit-solvent simulation was able to differentiate between the two pK_{a}'s. During the explicit-solvent simulation at pH 3, the average distance between the carboxylate oxygens of both residues was 3.7 Å. This close proximity was stabilized by a persistent hydrogen bond between the carboxylate oxygen of Asp19 and the backbone amide nitrogen of Asp21. However, the coupled titration behavior with two transitions was not observed when fitting the data for either Asp19 or Asp21. The only indication of coupling was a low Hill coefficient (0.56) for Asp19, which indicates anti-cooperativity, consistent with experiment[89]. We also examined the GB titration data. The interaction between Asp19 and Asp21 was very strong but both residues titrated with the same pK_{a} and the Hill coefficients were about 1. Thus, compared to the GB titration, the explicit-solvent simulation was able to provide, to some extent, the description of the coupled proton binding events for Asp19 and Asp21. However, the explicit-solvent simulation was not able to fully capture the negative cooperativity, which may be due to insufficient sampling.

HEWL

The last protein we consider is hen egg white lysozyme (HEWL), which has been used as a standard test system for many pK_{a} prediction methods[75,140]. Also, the most recent study of Nielsen and coworkers, where a consensus set of pK_{a}'s were derived from pH-dependent chemical shifts of different nuclei, makes HEWL the most vetted protein pK_{a} benchmark system available[133]. Table 4.5 lists the calculated pK_{a}'s from the explicit- and implicit-solvent titrations. Overall, the calculated pK_{a}'s

from the explicit-solvent titration are closer to experiment than the GB titration. The RMSD as well as the average absolute and maximum deviations in the explicit-solvent titration are 0.84, 0.70, and 1.50, respectively, while they are 0.93, 0.72, and 1.75, respectively, in the GB titration. Below we examine the cause for the significant differences between the explicit- and implicit-solvent titration data for residues Glu35 and Asp52.

Table 4.5: Calculated and experimental pK_a values of HEWL

Residue	Explicit Solvent[b]	GB	Expt[a]
Glu7	2.67 ± 0.01 (2.69)	2.58 ± 0.06	2.6 ± 0.2
His15	6.64 ± 0.10 (6.60)	5.34 ± 0.47	5.5 ± 0.2
Asp18	3.05 ± 0.13 (3.15)	2.94 ± 0.01	2.8 ± 0.3
Glu35	7.19 ± 0.15 (6.83)	4.35 ± 0.18	6.1 ± 0.4
Asp48	1.57 ± 0.48 (1.77)	2.84 ± 0.15	1.4 ± 0.2
Asp52	2.88 ± 0.08 (3.21)	4.56 ± 0.02	3.6 ± 0.3
Asp66	1.47 ± 0.60 (0.46)	1.15 ± 0.43	1.2 ± 0.2
Asp87	1.48 ± 0.41 (1.46)	2.03 ± 0.07	2.2 ± 0.1
Asp101	2.99 ± 0.09 (3.06)	3.27 ± 0.32	4.5 ± 0.1
Asp119	2.85 ± 0.05 (2.98)	2.45 ± 0.13	3.5 ± 0.3
CT-Leu	1.95 ± 0.37 (1.89)	2.20 ± 0.14	2.7 ± 0.2
Avg abs dev	0.70 (0.70)	0.72	
RMSD	0.84 (0.80)	0.93	
Max abs dev	1.50 (1.44)	1.75	

[a] Consensus pK_a's based on NMR titration using multiple nuclei[133]. [b] Values in parantheses were obtained from the 2-ns simulation.

The catalytic residues of HEWL are Glu35 and Asp52, which reside at the interface between two domains, and have the consensus pK_a's of 6.1 and 3.6, respectively. The experimental range of pK_a's calculated from chemical shifts of different nuclei were 6.0–6.8 for Glu35 and 3.4–4.0 for Asp52[133]. The pK_a's from the explicit-solvent simulation are 7.19 and 2.88, while those from the GB simulation are 4.35 and 4.56, respectively. Thus, considering the model values of 4.4 and 4.0 for Glu and Asp, the calculated pK_a shifts are in the correct direction in the explicit-solvent simulation but wrong in the GB simulation. Since the optimum pH for the activity of HEWL is

around 5[24], the pK_{a} calculation using the explicit-solvent CpHMD method is able to offer the correct protonation or charge states for the catalytic residues, which is not the case with the GB-based method. We note that the previous GB-based CpHMD simulations with the temperature-based replica-exchange protocol gave a correct direction of the pK_{a} shift for Glu35[65]. We examined the trajectory to delineate the cause for the significantly different pK_{a}'s. In the GB simulation, there is a significant rearrangement of the native structure. We plot the radius of gyration versus the heavy-atom RMSD using the explicit- and implicit-solvent simulation data (Figure 4.10). The conformations in explicit solvent have RMSD values, with respect to the crystal structure, ranging from 1.1 and 1.6 Å, and R_g values ranging from 14.1 to 14.4 Å. However, the conformations in the GB simulation have much larger RMSD (1.6–2.8 Å) and much smaller R_g (13.8–14.2 Å), which suggests a significant compaction and global deviation from the crystal structure. This global rearrangement of structure is propagated to the local conformational environment around the active-site residues, which can be seen from the differences in the solvent exposure of side chains.

At pH 6, Glu35 has an average solvent accessible surface area (SASA) of 18.9 Å^2 in the explicit-solvent simulation, which is similar to the value of 10 Å^2 based on the crystal structure but much smaller than the value of 38.9 Å^2 from the GB simulation. The significant increase in solvent exposure for Glu35 in the GB simulation leads to an overestimation of the self-solvation energy of Glu35, and thus an underestimation of the upward pK_{a} shift. For Asp52 the story is exactly reversed. The solvent exposure of Asp52 is underestimated in the GB as compared to the explicit-solvent simulation. At pH 4, the average SASA of Asp52 is 2.4 Å^2 in the GB simulation, whereas it is 25.4 Å^2 in the explicit-solvent simulation, which much closer to the initial value of 26.6 Å^2. Therefore, the self-solvation energy of Asp52 is underestimated in the GB simulation leading to a calculated pK_a value that is too high. Thus, HEWL is a clear case where

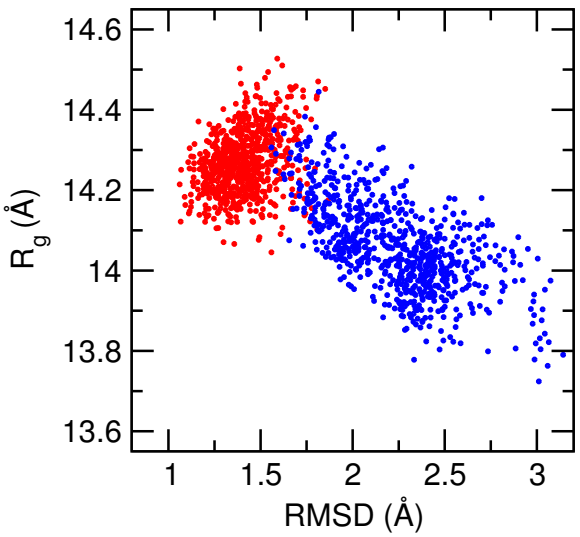

Figure 4.10: Comparison of conformational states of HEWL from explicit- and implicit-solvent simulations. Conformational states are described by root-mean-squared deviation (RMSD) and radius-of-gyration (R_g) of HEWL sampled in the explicit-solvent (red) and GB (blue) simulations at pH 6.

elimination of the "conformational" error introduced by GB can dramatically improve the accuracy of pK_a calculations for residues of biological significance.

4.5 Conclusion

Like temperature and pressure, solution pH is another important experimental condition that needs to be taken into account in molecular simulations in order to accurately capture physical reality. Motivated by the recent success of the GB implicit-solvent based CpHMD method in the accurate pK_a predictions and mechanistic studies of pH-dependent conformational dynamics of proteins, we have developed a robust approach to extend the CpHMD framework to explicit-solvent molecular dynamics simulations. In this approach, the explicit-solvent force field is used to drive conformational dynamics, while the GB model is used to efficiently estimate the role of solvent in modulating the cost of electrostatic free energy for protonation/deprotonation. The resulting explicit-solvent CpHMD method offers an increased accuracy and wider applicability as compared the GB-based CpHMD method, while retaining the efficiency

and robustness of the capability for proton titration. To overcome a critical hurdle related to the slow convergence of pK_{a} calculations, which has plagued CpHMD and other constant-pH methodologies, we have implemented a replica-exchange protocol based on the pH-biasing energy to directly accelerate protonation-state sampling. Remarkably, due to the tight coupling between titration and conformational degrees of freedom, this protocol also led to significant enhancement in conformational sampling, allowing pK_{a} to converge within 1 ns for small model compounds and large proteins. The random errors in the calculated pK_{a}'s for model compounds were about or below 0.1 pH units.

To benchmark the accuracy of the explicit-solvent based CpHMD method, we have calculated pK_{a}'s for five proteins and compared with results from the GB-based method and experiment. We found that the explicit-solvent titrations resulted in an average absolute error of 0.53 and RMSD of 0.74, on par with those from the GB-based titrations. However, by bringing the outliers closer to experimental values, the explicit-solvent method offers significantly improved correlation with experiment as compared to the GB-based method. Detailed analysis revealed that this improvement is due to more accurate conformational sampling in explicit solvent. For example, the explicit-solvent simulation preserved the structural integrity of the loop region, bringing the calculated pK_{a} of Asp77 from SNase closer to experiment. Compaction of HEWL in the implicit-solvent simulation caused distortion of the active site and large deviations in the calculated pK_{a} values for Glu35 and Asp52, while explicit-solvent simulation preserved the native conformation leading to a correct prediction of the protonation states at the optimum pH value for catalytic activity. Including solvent granularity enabled a more realistic description of ion-pair interactions, as was the case for Asp95 of SNase, where the explicit-solvent simulation gave a bimodal distribution representing both the close-range and solvent-separated interactions with Lys70, which resulted in a more accurate estimate of pK_a.

Finally, in the explicit-solvent simulation the hydrophobic cluster in BBL showed an increased mobility relative to the GB simulation, allowing His166 to be partially exposed to solvent, which resulted in a reduction in the pK_a shift due to desolvation penalty. The latter aspect is somewhat surprising, but is compatible with previous GB simulation studies revealing overly rigid hydrophobic assemblies[83,122]. It is also consistent with the experimental evidence[141] and previous simulation study[142] suggesting water penetration into the hydrophobic core of SNase. Although in the presented cases, the differences between the explicit-solvent and GB-based pK_a results are small (all within 0.5 pH units), our unpublished data shows that the explicit-solvent method offers improvement as high as 4 pH units for the worst prediction cases in the engineered mutants of SNase (Wallace and Shen, unpublished data).

While the results demonstrated in this work are encouraging, we note that several potential issues merit attention. First, a potential delay in the response of solvent reorganization to protonation/deprotonation may lead to unfavorable interactions or inaccuracy in the solvation energetics of the titration site. This problem can be effectively avoided by allowing a few additional dynamics steps between titration updates to allow relaxation of solvent around the titrating site, as has been demonstrated in the discrete constant-pH techniques[54]. Also, we identified a small bias towards the charged form in the titration of Asp, Glu and His residues due to the occasionally prolonged residence time of the two degenerate protonation states (doubly deprotonated in the case of Asp or Glu and doubly protonated in the case of His). Although the effect of this systematic error on the calculated pK_a shifts is likely minimal, it is clearly a limitation that needs to addressed in the future. Finally, the accuracy of pK_a calculations is still limited by the accuracy of the GB model to determine the deprotonation free energy. The largest deviation and the single outlier found in this work is Asp21 in SNase, where both explicit- and implicit-solvent simulations were not able to reproduce the direction of the positive pK_a shift, and underestimated the

pK_a by 3 pH units.

NMR data showed that the titration of Asp21 is coupled to that of Asp19, which has a negative pK_a shift. Although the explicit-solvent simulation was able to differentiate between the two pK_a's, it could not quantitatively reproduce the extent of the negative cooperativity in proton binding. One possible cause is that more exhaustive sampling may be required to fully capture coupled titration events. This issue deserves further investigation in our future studies. Another aspect that deserves further investigation is related to the effect due to ions. In the current work and previous GB-based CpHMD studies, an approximated Debye-Hückel model is applied in the GB electrostatic calculation to account for the bulk effect of salt screening, which may not be accurate for highly charged systems such as nucleic acids where local charge density can be very high. Finally, in order to apply the explicit-solvent CpHMD to studies of large-scale conformational changes, it may become necessary to combine with a method for global enhancement of conformational sampling such as the temperature-based replica-exchange scheme. Despite these remaining limitations, the current accuracy and precision of the explicit-solvent based CpHMD technique are encouraging, considering the fact that experimentally determined pK_a's can deviate by 0.5-1 pH units depending on the nuclei monitored[133]. Thus, we anticipate that explicit-solvent CpHMD simulations will emerge as a practical tool for gaining novel insights into protonation-related phenomena that are ubiquitous in biology and chemistry. Examples include the mechanism of proton channels, drug-efflux pumps, pH-dependent catalytic reactions of ribozymes, as well as titration behavior of mixed micelle systems.

Chapter 5

Unraveling a trap-and-trigger mechanism in the pH-sensitive self-assembly of spider silk proteins

We use hybrid-solvent CpHMD simulation to probe the pH-dependence of spider silk assembly, and find that there are twin electrostatic mechanisms that work in concert to control silk assembly. This atomically detailed mechanistic information may be useful for the design of novel silk based materials for engineering or biomedical applications.

The following content was published in :

Journal of Physical Chemistry Letters

volume 3, pages 658-662, 2012

5.1 Abstract

When the major ampullate spidroins (MaSp1) are called upon to form spider dragline silk, one of nature's most amazing materials, a small drop in pH must occur. Using a state-of-the-art simulation technique, constant-pH molecular dynamics, we discovered a few residues that respond to the pH signal in the dimerization of the N-terminal domain (NTD) of MaSp1 which is an integral step in the fiber assembly. At neutral pH the deprotonation of Glu79 and Glu119 leads to water penetration and structural changes at the monomer-monomer binding interface. At strongly acidic pH, the protonation of Asp39 and Asp40 weakens the electrostatic attraction between the monomers. Thus, we propose a "trap-and-trigger" mechanism whereby the intermolecular salt-bridges at physiologically relevant pH conditions always act as a stabilizing "trap" favoring dimerization. As pH is lowered to about 6,

Glu79 and Glu119 become protonated, triggering the dimerization and subsequent silk formation. We speculate that this type of mechanism is operative in many other pH-sensitive biological processes.

5.2 Introduction

Spider dragline silk is one of nature's most spectacular materials. The toughness of spider silk surpasses synthetic rubber and its strength is comparable to high-tensile steel[143]. These exceptional properties, combined with incompatibility, make silk and silk biomimetics attractive materials for future bio- and material engineering applications. Silk based products have been envisioned as drug delivery systems[144] and scaffolds for tissue engineering[145,146], adhesives, and microfluidic devices[147]. Dragline silk is made of proteins known as the major ampullate spidroins 1 and 2 (MaSp1 and MaSp2) which are stored in the major ampullate gland of spiders as a microemulsion[148,149]. The MaSp's are comprised of a repetitive domain containing polyalanine, glutamine- and glycine-rich motifs as well as the non-repetitive C-terminal and N-terminal domains[148,149]. During silk spinning the spidroins pass from the gland, through the exit duct, while being subjected to several chemical and physical forces leading to nano-composite fibers[148]. A major factor contributing to the transition from soluble proteins to silk fibers is acidification[13]. While solution pH drops from 7.2 in the gland to 6.3 within the first millimeter of the 20 mm exit duct[150], the pH at the distal end of the duct may be substantially lower[13,151]. Acidification has been found to increase the rate at which soluble silk proteins from *Euprosthenops australis* spiders[14] and *Bombyx mori* silkworms[152] aggregate suggesting this acid-bath treatment is a general method employed by nature for producing the delicate, yet durable, silk fibers.

Recently, several experiments have demonstrated that the NT domain of MaSp1 (NTD) is the pH sensing portion, which dimerizes, allowing spidroins to self-assemble

by a relay-like mechanism into silk fiber[14,149,153,154].

However, a detailed mechanism of the pH-dependent assembly process remains unclear. Askarieh *et al.* solved a crystal structure which showed that the NTD from *Euprosthenops australis* exists as a stable homodimer[14]. However, experiments based on electrospray ionization mass spectrometry with the same protein[154] and NMR with the protein from *Latrodectus*[155] indicated that NTD is mainly monomeric at neutral pH. The dimeric form is more stable at acidic pH and low concentrations of salt, consistent with the pull-down experiments using NTD's from both *Latrodectus* and *Nephila*[149]. Moreover, the interpretation of a pH-dependent red shift of tryptophan fluorescence differs as to whether it is the result of the conformational change in the dimer[14] or monomer[149].

The aforementioned inconsistency has prompted us to investigate the pH effect on the dimerization of NTD using a state-of-the-art simulation technique, continuous constant-pH molecular dynamics (CpHMD)[64,66] with conformational sampling in explicit solvent and pHREX[104], which allows us to determine the pK_{a} values of all ionizable side chains and probe the pH-dependent conformational dynamics in atomistic details. Our data indicates that the dimer becomes destabilized as the pH is increased above 6, as a result of the ionization of key residues Glu79 and Glu119 which leads to water penetration into the monomer-monomer interface.

5.3 Methods

5.3.1 Structure preparation

The initial structure of the dimeric NTD-MaSp1 was prepared by removing all hetero-atoms and adopting the first of the two side-chain orientations based on the crystal structure (PDB ID: 3LR2). Missing residues were built with the program Modeller[156]. Hydrogens were added with the HBUILD facility in the CHARMM

program[39] and the structure was placed in a truncated octahedral water box with dimensions 18 Å greater than the largest dimension of the protein. Sodium chloride was added such that the solvent ionic strength was 100 mM. Water molecules within 2.6 Å of any heavy atom were deleted and the system was energy minimized in several stages with progressively smaller harmonic restraint applied to heavy atoms of the protein.

5.3.2 Simulation details

The structures of the dimer and corresponding subunits as prepared above were used to initiate the pHREX titration simulations of the dimer and monomers, respectively. All simulations were conducted using the PHMD module[64,66,104] in the CHARMM program (version c35b3)[39]. The most recent extension[104] of the PHMD module which includes conformational sampling in explicit solvent and the pH-based replica-exchange (pHREX) protocol to accelerate barrier crossing and convergence was applied. The all-atom force fields, CHARMM22/CMAP[36] and TIP3P[134], were used to represent the protein and water atoms, respectively. Molecular dynamics simulations were run at ambient temperature and pressure using Hoover thermostat[136] and Langevin piston pressure coupling[137]. The SHAKE algorithm was applied to all bonds and angles involving hydrogen to allow a time-step of 2 fs. The electrostatic interactions were calculated using the particle-mesh Ewald summation. The van der Waals interactions were calculated with a switching function starting at 10Å and ending at 14 Å. The generalized-Born (GB) implicit model, GBSW[79], with a Debye-Hückel term for taking into account salt screening (ionic strength was set to 100 mM), was applied to calculate solvent-modulated electrostatic energies for the propagation of titration coordinates[104]. The implicit-solvent model GBSW[79] was employed with a refined set of atomic radii[98] to define the dielectric boundary. The GB calculation and update of titration coordinates were executed every 10 dynamic steps. In a

pHREX simulation, independent replicas are subject to molecular dynamics runs at ambient temperature and pressure but different pH conditions. An exchange between adjacent pH conditions was attempted every 500 dynamic steps. The pHREX protocol was enabled through a Perl package, MMTSB Tool Set[91], which provides an interface with the CHARMM program. A total of 14 and 17 replicas was used for the monomer and dimer simulations, respectively. Simulations were carried out for 5500 exchange attempts (5.5 ns) per replica, resulting in the cumulative simulation time of 77 ns for the monomers and 93.5 ns for the dimer. To optimize the exchange frequency between neighboring pH replicas, trial 1 ns pHREX simulations were carried out with pH conditions 0–10 with 1 pH unit intervals. The exchange success ratios were examined, and additional replicas were added at 0.5 pH unit intervals between replicas with exchange ratios below 20%. For monomer simulations, there were 14 replicas at pH values of 0.0, 1.0, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 6.0, 7.0, 8.0, 9.0, and 10.0. For the dimer simulation there were 17 replicas at pH values of 0.0, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 5.5, 6.0, 6.5, 7.0, 8.0, 9.0, and 10.0. The coordinates and titration states were recorded after each exchange attempt. In the data analysis, the first 500 exchange cycles (0.5 ns per replica) were discarded.

5.3.3 Calculation of pK_{a}'s and pH-dependent dimer stability

pK_{a} values were calculated by fitting the unprotonated fractions (S) at each pH to the Hill equation given by

$$S = \frac{1}{1 + 10^{n(pK_a - pH)}} \tag{5.1}$$

where n (the Hill coefficient) and the pK_{a} are fitting parameters.

Analytical integration of the Wyman-Tanford linkage equation leads to the following expression for the pH-dependent free energy of dimer dissociation in analogy

to the pH-dependent protein stability[83],

$$\begin{aligned}\Delta\Delta G(\mathrm{pH}) &= \Delta G(\mathrm{pH}) - \Delta G(\mathrm{pH}^{ref}) = \\ &= RT\sum_i \frac{1}{n_i^M} ln \frac{1+10^{n_i^M(p\mathrm{K}_i^M - \mathrm{pH})}}{1+10^{n_i^M(p\mathrm{K}_i^M - \mathrm{pH}^{ref})}} \\ &- RT\sum_i \frac{1}{n_i^D} ln \frac{1+10^{n_i^D(p\mathrm{K}_i^D - \mathrm{pH})}}{1+10^{n_i^D(p\mathrm{K}_i^D - \mathrm{pH}^{ref})}},\end{aligned} \tag{5.2}$$

where the summation runs over all residues. n_i^D, $p\mathrm{K}_i^D$ and n_i^M, $p\mathrm{K}_i^M$ are the Hill coefficients and pK_a values for the ith residue in the dimeric and monomeric forms, respectively. In this work we set $\mathrm{pH}^{ref} = 8.0$.

5.3.4 Poisson-Boltzmann calculations

Electrostatic potential maps were calculated using the Poisson-Boltzmann (PB) facility in the CHARMM program[39]. The atomic charges on the titratable residues at different pH conditions were set as the average charge calculated from the pHREX titration simulation at the respective pH condition. The PB calculations used a salt concentration of 100 mM, a 1Å ion exclusion (Stern) layer, an internal dielectric constant of 4 and an external dielectric constant of 80. Electrostatic potential maps were rendered using the program VMD[93].

5.3.5 Error estimates

To estimate the uncertainty of the calculated pK_a values (fitting parameters k), we applied the well-known Monte Carlo "bootstrap" method[157]. The method comprises three steps: (1) generate a large number (we used 100) of independent bootstrap samples $S^*(i), i = 1...N$, where S represents the unprotonated fraction; (2) calculate the quantity of interest, i. e. the fitting parameter $k^*(i)$, for N bootstrap samples; and (3) calculate the standard deviation of the $k^*(i)$ values. For step (2) we assume that the probability of selecting a particular S value in each set S_i^* is given by a Gaussian

distribution centered at S^{final}(pH) and having a standard deviation calculated by block standard-error[115,116] of the unprotonated fraction (see Eq. 3.8). The error associated with the resulting pH-dependent free energy of dimer dissociation was calculated by propagating the estimated error in calculated pK_{a} values.

5.4 Results and Discussion

Considering the experimental findings[14,149,154,155], we hypothesized that the origin of the pH-dependent dimerization may be explained by the intermolecular interactions between monomers, and that ionization of a few residues could shift the monomer-dimer equilibrium. To test this hypothesis, we carried out pH titration simulations using the pHREX-CpHMD method in explicit solvent for the dimer and monomer forms of NTD starting from the crystal structure of the dimer (PDB ID: 3LR2) and two monomer units, respectively. Molecular dynamics at pH conditions of 0 to 10 was performed for 5.5 ns while simultaneously titrating all acidic and histidine residues and periodically attempting exchanges between pH replicas. The cumulative simulation time was 77 ns and 93.5 ns for the dimer and monomers, respectively. These simulations allowed us to determine pK_{a} values and obtain details of the conformational dynamics at each pH condition. The convergence of the pK_{a} values is illustrated in Figure 5.1 which shows that the pK_{a} values are very stable after the first few-thousand exchange cycles.

Using the pK_{a} values of all titratable residues in the dimer and unbound monomers (Table 5.1) we calculated the total charge of the dimer and monomers (Figure 5.2a). Using the pK_{a} shifts upon dimerization we obtained the pH-dependent changes in the dimer stability (Figure 5.2b) by analytically integrating the Wyman-Tanford linkage equation[18]

$$\partial \Delta G / \partial \text{pH} = \ln(10) \text{RT} \Delta \text{Q}^{\text{diss}} \tag{5.3}$$

Table 5.1: Calculated pK_a values of the unbound (monomer) and bound (dimer)

Residue	pK_a unbound	pK_a bound	ΔpK_a
Monomer A			
His6	6.83 ± 0.07	6.88 ± 0.07	0.05 ± 0.10
Glu17	4.06 ± 0.02	4.10 ± 0.03	0.04 ± 0.04
Asp39	3.05 ± 0.05	1.31 ± 0.06	-1.74 ± 0.07
Asp40	4.11 ± 0.05	4.62 ± 0.04	0.51 ± 0.06
Glu79	4.42 ± 0.04	6.26 ± 0.05	1.84 ± 0.07
Glu84	4.40 ± 0.04	4.91 ± 0.06	0.51 ± 0.07
Glu85	3.92 ± 0.04	3.89 ± 0.04	-0.03 ± 0.05
Glu119	4.23 ± 0.05	6.12 ± 0.04	1.89 ± 0.05
Asp134	3.83 ± 0.05	3.55 ± 0.07	-0.28 ± 0.08
CT-Ala	3.35 ± 0.04	3.36 ± 0.04	0.01 ± 0.05
Monomer B			
His6	7.10 ± 0.04	7.73 ± 0.07	0.63 ± 0.08
Glu17	4.15 ± 0.05	4.15 ± 0.04	0.00 ± 0.06
Asp39	2.80 ± 0.04	2.03 ± 0.06	-0.77 ± 0.07
Asp40	4.19 ± 0.05	3.13 ± 0.10	-1.06 ± 0.11
Glu79	4.43 ± 0.05	6.73 ± 0.06	2.30 ± 0.08
Glu84	4.48 ± 0.07	4.70 ± 0.04	0.22 ± 0.08
Glu85	3.97 ± 0.04	3.88 ± 0.04	-0.09 ± 0.06
Glu119	4.32 ± 0.03	6.71 ± 0.05	2.39 ± 0.06
Asp134	4.22 ± 0.06	3.58 ± 0.04	-0.64 ± 0.07
CT-Ala	3.37 ± 0.03	3.73 ± 0.05	0.36 ± 0.06

pK_a values determined by fitting the simulated unprotonated fractions to the Hill equation. Error bars are the standard deviation of 100 bootstrap trial fittings. ΔpK_a = pK_a bound - pK_a unbound

Error bars are the standard deviation (σ) of 100 bootstrap trial fittings.

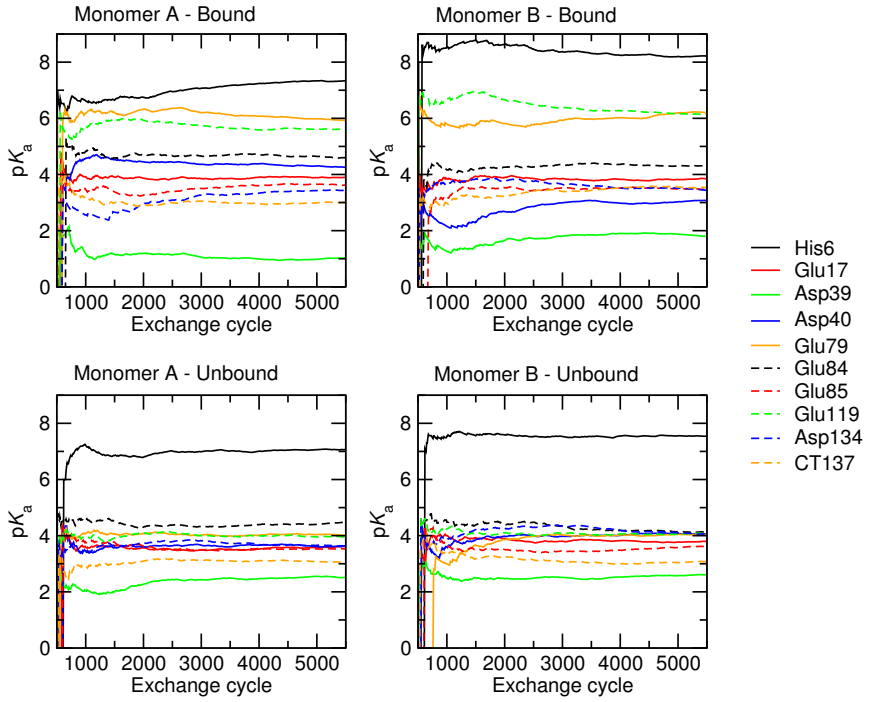

Figure 5.1: Time series of p$K_{\rm a}$ values of NTD monomer and dimer. Cumulative p$K_{\rm a}$ values estimated from the unprotonated fraction (S) at the pH value nearest the p$K_{\rm a}$ for all residues titrated.

where ΔG is the free energy of dimer dissociation (or dimer stability) and $\Delta Q^{\rm diss}$ is the change in the total charge as dimer dissociates. The integrated form of Eq. 5.3 is given by Eq. 5.2.

Our calculated isoelectric point (pI), 4.4 and 4.1 for the dimer and the two monomers, respectively (Figure 5.2a), are in good agreement with the value of 4.25 for the dimer determined by the measurements of electrophoretic mobility[14]. According to the calculated stability change (black curve in Figure 5.2b), the dimer is least stable at pH 8. As the solution pH is decreased to 6, the dimer is stabilized by about 3.5 kcal/mol. The stabilization continues as pH is lowered to near 4 where dimerization is favored by 11 kcal/mol as compared to pH 8. As the solution pH is reduced further the dimer is again destabilized.

The stability change of the dimer can be decomposed into contributions from each titratable residue assuming no coupling between their protonation equilibria. The

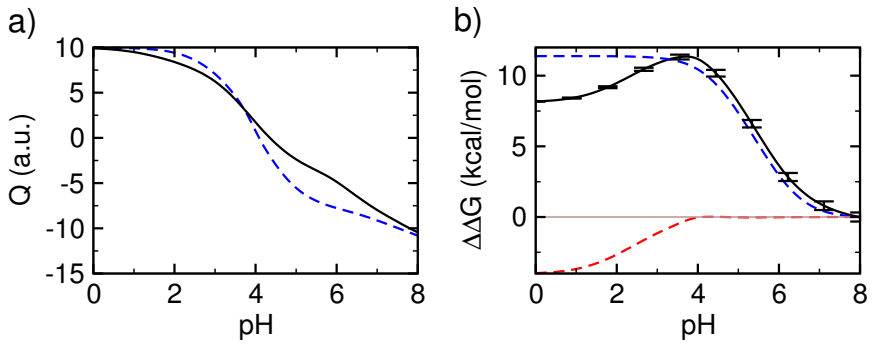

Figure 5.2: pH-dependent change in total charge of NTD monomer and dimer and stability of the NTD dimer. a) Total charge of the dimer (solid) and the two monomers (dashed) calculated at different pH conditions. b) Stability change (free energy of dimer dissociation), $\Delta\Delta G$) relative to the pH 8 condition calculated using all p$K_{\rm a}$ values (black), Glu79 and Glu119 (blue), and Asp39 and Asp40 (red). A horizontal line is drawn at $\Delta\Delta G$ of zero to guide the eye.

contributions from residues Glu79 and Glu119 dominate the stability change in the pH range of 4 to 8 (blue curve in Figure 5.2b) and they are also the major source for the change in the total charges upon dimerization in the same pH range (Figure 5.2a). This is because the p$K_{\rm a}$'s of Glu79 and Glu119 have the largest positive shifts upon dimerization. The p$K_{\rm a}$'s in the monomers are similar to the model value, around 4.4, but in the dimer they are shifted to above 6. Ionization of these residues significantly destabilizes the dimer. On the contrary, the p$K_{\rm a}$ values of Asp39 and Asp40 have the largest negative shifts upon dimerization. Therefore, they are together responsible for the stability change in the pH range 0-4 (red curve in Figure 5.2b). They are also major contributors to the change in the total charge upon dimerization.

We next examine in detail how ionization of Glu79 and Glu119 leads to the destabilization of the dimer at elevated pH. According to our calculation, as solution pH is increased from pH 4 to pH 8, the ionization of Glu79 and Glu119 destabilizes the dimer by more than 10 kcal/mol. When solution pH is 4 or 5, which is below the p$K_{\rm a}$ values of Glu79 and Glu119 (between 6.1 and 6.7), these residues are only weakly solvated. The distribution of the hydration number is centered around 2 (Figure 5.3).

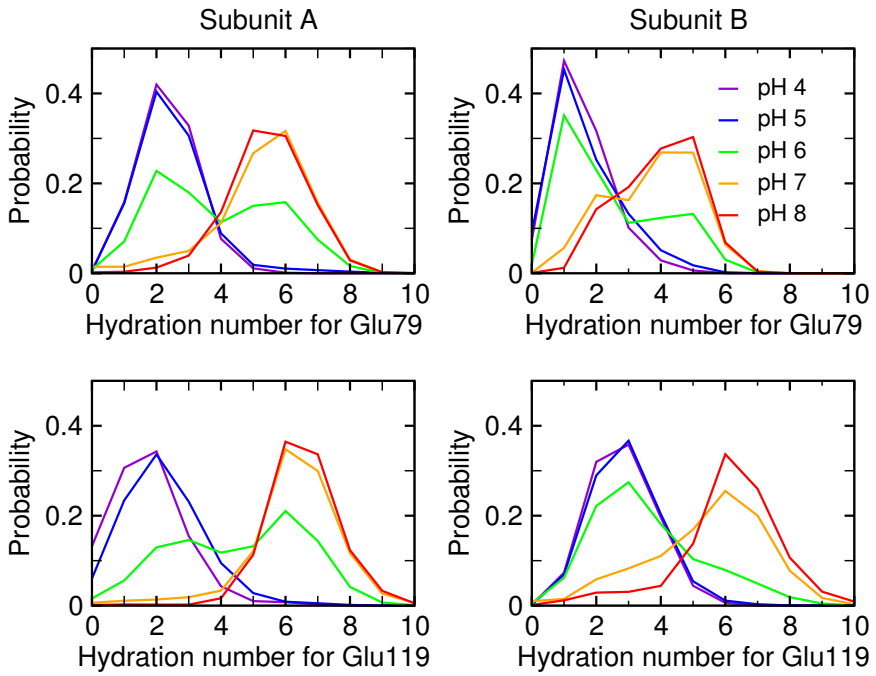

Figure 5.3: pH-dependent hydration of Glu79 and Glu119 in NTD dimer. Probability distributions of the hydration numbers for Glu79 and Glu119 in both monomer units at different pH conditions. Hydration number is defined as the number of water molecules within 3 Å of the side chain.

However, as pH is increased, the hydration number increases due to ionization of the side chains of Glu79 and Glu119. The distribution becomes bimodal at pH 6, while at pH 7 and 8, the maximum probability is shifted to around 6 or 7. The hydration of Glu79 and Glu119 at the pH above 6 is associated with the entrance of water molecules into the dimerization interface. The latter can be quantified by the change in the interfacial solvent accessible surface area (SASA) as a function of pH (Figure 5.4a). At pH 4 and 5, the interfacial SASA fluctuates around 1200-1300 Å^2 but the distribution shifts to larger values as pH is increased to 6. When pH is further increased to 7 and 8, the maximum probability for SASA is around 1500 Å^2. Thus, an increase of pH from 4 to 8 results in an increased solvent-exposure of the dimer interface by at least 200 Å^2. To further characterize the pH-induced water penetration, we calculated the radial distribution function (RDF) for water relative to the dimer center (Figure 5.4b). At pH 8, the RDF shows an increased density of water near the dimer center as compared to pH 4. The notable accumulation of water at around 10 Å and 12 Å corresponds to the positions of Glu79 and Glu119, respectively.

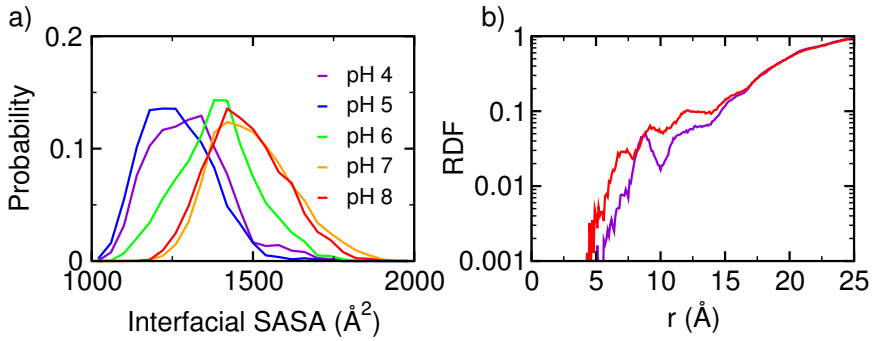

Figure 5.4: pH-dependent solvent exposure of NTD dimer interface. a) Probability distribution of the interfacial solvent-accessible surface area (SASA) at different pH conditions. b) Radial distribution function (RDF) of water to the dimer center (defined as the center of the Cα atoms in Phe73 and Ala74) at pH 4 (purple) and 8 (red).

As water molecules enter and the monomer-monomer interface opens up, confor-

mational rearrangement occurs. Although our simulation can not describe the full extent of the conformational change to due to the limited sampling time, it offer a glimpse at the initial events. At pH 6 and above, the contacts between residues from the opposite monomer subunits (A and B) are weakened. Most notably, the contact probabilities for Glu79(A)–Met71(B) and Met71(A)–Glu79(B), as well as for Glu119(A)–Met126 (B) and Met126(A)–Glu119(B) are significantly reduced (see Figure 5.5). The general trend of the weakened inter-molecular interactions at elevated pH conditions is another indication of the destabilization of the dimer. Moreover, we observe that upon ionization Glu119 rotates out of the binding interface into solution. The most probable distance between Glu119 and the dimer center is around 12 Å at pH 4, but at pH 8 the distribution becomes bimodal with a population centered at just greater than 15 Å (Figure 5.6).

For Glu79 there is a very slight shift outward, which is a result of the expansion of the dimer structure (Figure 5.8). For Glu119, along with the overall movement of the dimer, there is a distinct change in the χ_2 angle at basic pH and a rotation of Helix 5 towards solution which allows Glu119 to become more solvated (see Figure 5.7).

At pH 4, the χ_2 angle of Glu119 samples $\pm 180^\circ$ and there is a minor population at 70°, but at pH 8 χ_2 is predominately 70° and the position at $\pm 180^\circ$ is not sampled. The related conformational rearrangement can be readily seen by comparing the snapshots taken from the simulation at pH 4 and pH 8 (Figure 5.8).

Our simulated titration data reveals two key residues responsible for the pH-dependent dimerization of NTD-MaSp1. At slightly acidic pH Glu79 and Glu119 are protonated, but at pH above 7 these residues become ionized. Burial of these charged residues in the hydrophobic environment of the dimer interface is unfavorable which results in a destabilization of the dimer by about 3.5 kcal/mol when pH is raised from 6 to 8 (Figure 5.2). Thus, according to our data, mutation of Glu79 to a neutral residue Gln should favor the dimerization at neutral or elevated pH, consistent with

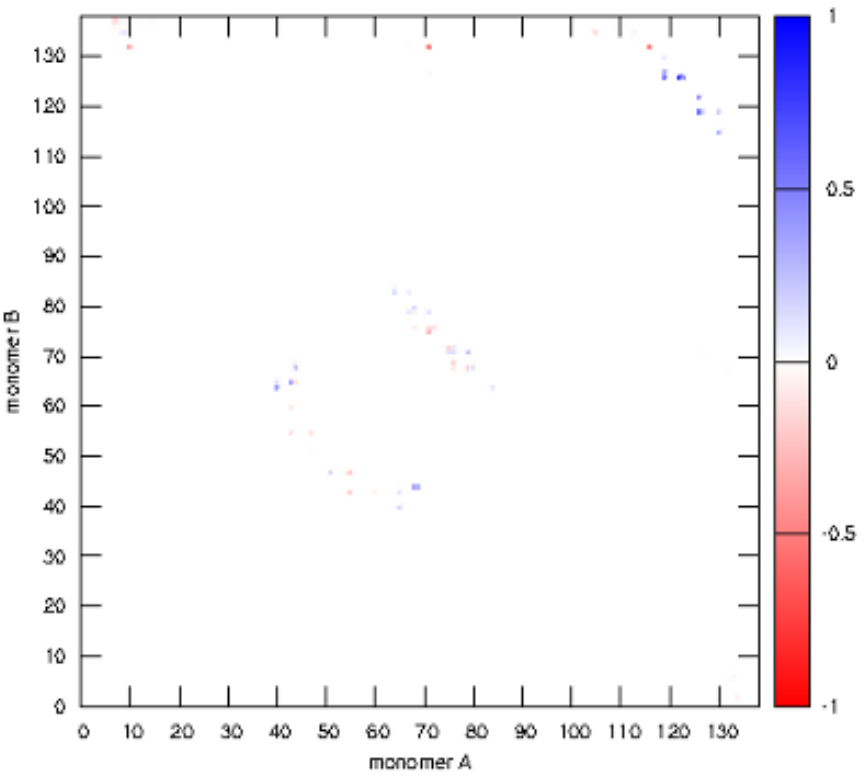

Figure 5.5: Difference (pH 4 less pH 8) contact probability map of monomer-monomer side chain interactions. Positive value indicates more probable contact at pH 4, while negative value indicates contact more probable contact at pH 8. Residues are considered to be in contact if the geometric centers of the side chain heavy atoms are within 7Å.

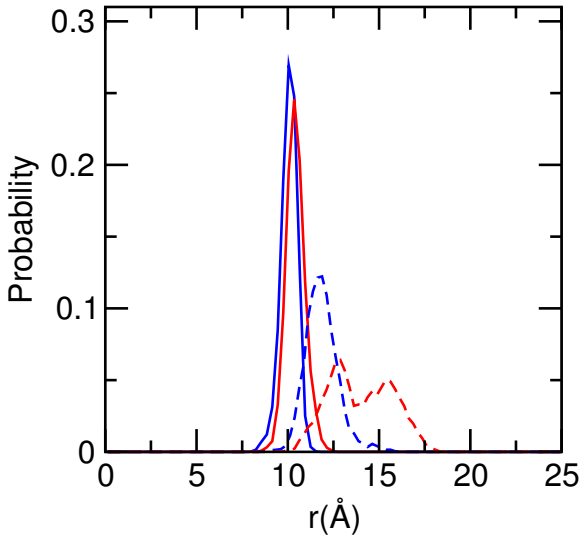

Figure 5.6: Probability distribution of Glu79 and Glu119 distance to dimer center. Probability distribution of the distance to the center-of-mass of the dimer from Glu79 (solid) and Glu119 (dashed) at pH 4 (blue) and pH 8 (red).

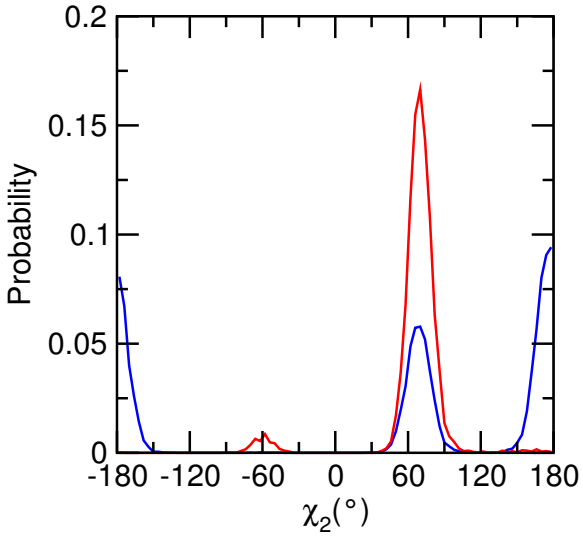

Figure 5.7: Probability distribution of Glu119 side chain orientation. Probability distribution of the χ_2 angle of Glu119 at pH 4 (blue) and pH 8 (red).

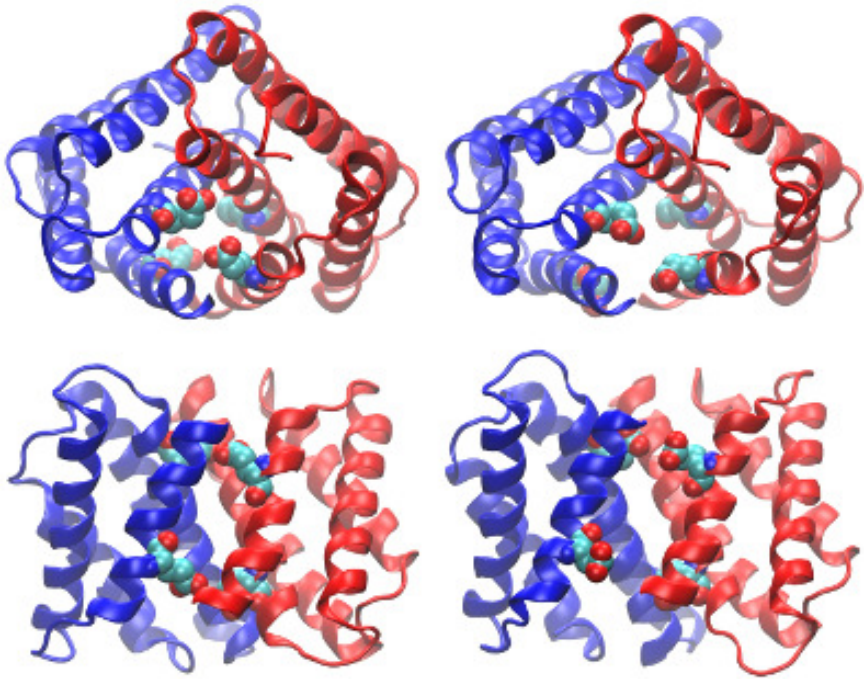

Figure 5.8: pH-dependent conformational rearrangement of NTD dimer. Top (*upper*) and side (*lower*) views of the snapshots taken from the simulation at pH 4 (*left*) and pH 8 (*right*). Glu79 and Glu119 are explicitly shown. Images were rendered using the VMD program[93].

the electrospray ionization mass spectrometry data which showed that mutant D79N is able to dimerize at pH 6.8 and 7 in contrast to the wild type[154].

Our dynamics data shows that ionization of the interfacial residues Glu79 and Glu119 at or above pH 7 causes the dimer interface to open up, which allows water molecules to enter, thereby weakening the intermolecular interactions that are responsible for holding the two monomer units together (Figure 5.3, 5.4 and 5.8). The pH-induced conformational change in the dimer is consistent with the deuterium exchange data which showed decreased deuteration level at pH 6 relative to pH 7[154]. The same set of experiments also showed that mutation D40N, E84Q or D40N/E84Q inhibits the dimer formation at low pH. However, these data do not necessarily imply that ionization of Asp40 or Glu84 promotes the dimer stability at low pH. This is because Asp39, Asp40 and Glu84 are clustered together in the crystal structure, and mutation of one side chain likely perturbs the electrostatic interactions of the other two. To rationalize these experimental data, additional simulation based on the mutant structures would be necessary, which is beyond the scope of the current paper.

Our simulation also reveals that the dimer interface is further stabilized by electrostatic interactions. Asp39 forms salt-bridges with Arg60 or Lys65 of the opposite subunit (see Figure 5.9). These favorable interactions are reflected in the negative pK_{a} shift of Asp39 upon dimerization. Asp40 also interacts with Lys65 of the other subunit, although the extent of the pK_{a} shift is less than that of Asp39. As pH is decreased below 4, Asp39 and Asp40 become protonated, the fraction of the tightly bound interactions with Arg60 and Lys65 is drastically reduced. Thus, ionization of Asp39 and Asp40 is responsible for the reduced dimer stability at pH below 4 (blue curve in Figure 5.2b).

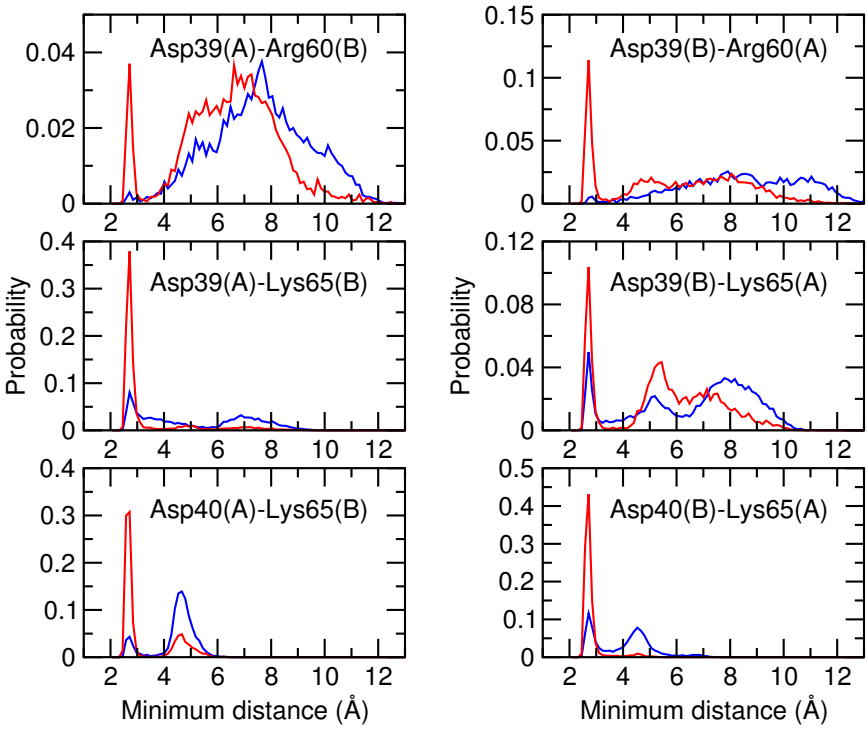

Figure 5.9: Probability distributions of NTD inter-monomeric salt-bridges. Probability distribution of the minimum distance between heavy atoms in Asp39-Arg60, Asp39-Lys65 and Asp40-Lys65 when the acidic residue is fully deprotonated (red) and fully protonated (blue). The distribution for the pH condition The subunit is indicated in the parenthesis.

5.5 Conclusion

In summary, we have identified a pH-modulated electrostatic system that controls the dimerization of NTD-MaSp1 (see Figure 5.10). In the pH range 4-6, the dimer is

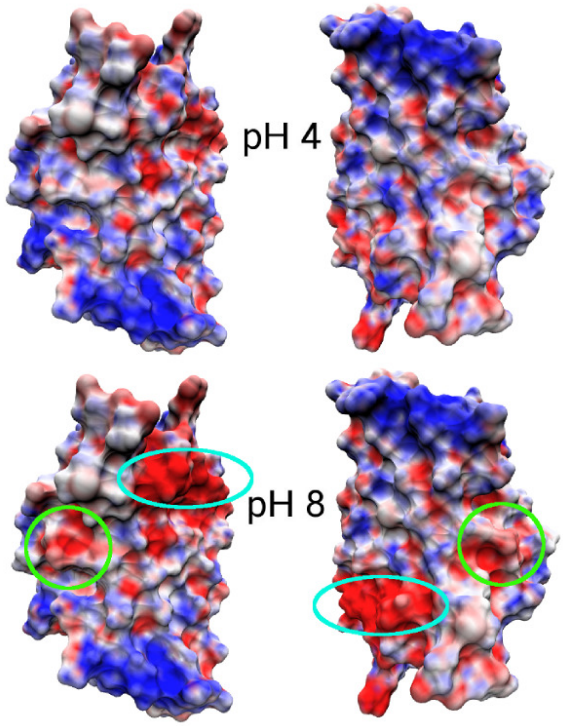

Figure 5.10: pH-dependent electrostatic potential of NTD dimer subunits. Electrostatic surface potential maps are calculated using the average charges derived from the unprotonated fractions of titratable residues at each pH. The subunits are separated and oriented to show the dimer interface. The locations of Glu79 (turquoise) and Glu119 (green) are circled. Images were rendered using the VMD program [93].

stabilized by the salt bridges (Asp39-Arg60, Asp39-Lys65, Asp40-Lys65) formed at the opposite poles of the subunits. Glu79 and Glu119 are protonated and buried in the interface. At pH above 6, these two residues become deprotonated and introduce a large excess negative potential resulting in a large desolvation penalty for the formation of the homodimer. Thus, we propose that a "trap-and-trigger" mechanism controls dimerization where the opposite poles at physiologically relevant pH conditions always act as a stabilizing "trap" favoring dimerization. However, the acidic residues at the hydrophobic interface must be protonated, and protonation of these

residues is the "trigger" that causes the NTD to dimerize and act as a pH-sensitive relay during silk formation.

Chapter 6

Explicit-solvent continuous constant-pH molecular dynamics with reaction field electrostatics and charge leveling

Spurred by the attempts of others to develop a fully explicit-solvent continuous constant-pH molecular dynamics (CpHMD) approach, we implemented techniques that allow the system net-charge to remain neutral as titration proceeds and give a proper treatment on long-range electrostatic interactions. Our data shows that explicit-solvent CpHMD can deliver pK_a values in good agreement with experiment when these techniques are applied. This indicates that explicit-solvent CpHMD can be used as a reliable tool for controlling solution pH in molecular dynamics simulations.

6.1 Abstract

The use of constant-pH molecular dynamics (pHMD) is becoming increasingly popular. Most pHMD approaches are dependent in part or whole upon an implicit-solvent model; however, tests of fully explicit-solvent pHMD on small peptides[69] and nucleotides[70] have recently emerged. Encouraged by these results, we perfect a method that combines the continuous constant-pH molecular dynamics framework with a reaction-field treatment of long-range electrostatics and a charge-leveling technique. We implemented these methods in the CHARMM[39,40] simulation program and tested them on a series of aliphatic dicarboxylic acids and proteins. We find agreement with experiment is poor if the net charge of the system is allowed to vary, but coupling ionization of titratable sites to the charging/neutralization of co-ions, which act as a charge reservoir, allows good agreement with experiment to be obtained. For protein residues which participate in strong electrostatic and hydrogen-bond interactions, our data indicates that insufficient sampling of the energetically available

conformations is a major source of difficulty in the calculation of protein pK_{a} values.

6.2 Introduction

Solution pH is an extremely important factor in chemical and biological processes. Proton uptake and release, driven by pH, can affect the energetics[158] and kinetics[159] of protein-protein complex formation. Conformational transitions of proteins between different folded states[160] or between native and denatured states can be modulated by pH[86,161,162]. Protein-ligand binding[18,163] and enzymatic activity of proteins[164,165] and ribozymes[166] can be affected by pH. Solution pH can also control morphological characteristics of large protein complexes such as aggregates of β-amyloid[167,168] and spider-silk proteins[14]. In solutions of fatty-acids, and other titratable amphiphilic molecules, changes in solution pH can trigger vesicle-micelle-bilayer transitions[169,170]. Due to the fundamental importance of pH, there have been several molecular dynamics techniques proposed where protonation states are allowed to respond to their local chemical environment and the specified pH. These methods, broadly referred to as constant-pH molecular dynamics (pHMD), have been reviewed elsewhere[82,140], but will briefly be discussed here.

The pHMD approaches fall into two categories: methods which use continuous protonation states and those that use discrete protonation states. In the discrete approach, fixed protonation state molecular dynamics (MD) is carried out and periodically interrupted, then protonation states are updated by Monte-carlo (MC) sampling. These methods differ in the solvent model used during the MD stage, either explicit[52,56,62] or implicit[59,60] and in how the change in energy is calculated after an update of the protonation states.

In the second continuous approach, additional titration degrees of freedom are added for every titratable site and are propagated alongside conformational dynamics. These methods are referred to as continuous constant-pH molecular dynamics

(CpHMD) since protonation states are controlled by continuous variables. The most widely applied, and arguably most successful, of these methods initially used a generalized Born (GB) implicit-solvent model for propagating both conformation and titration coordinates[64–66]. Later, in an attempt to circumvent conformational error introduced by GB while retaining the ability to efficiently evaluate the energetic effects of solvation, CpHMD was extended to allow conformational sampling to be carried out in explicit solvent, while the titration coordinates were propagated via GB energetics. This hybrid-solvent CpHMD approach, combined with pH-replica exchange (pHREX), offered more realistic conformational sampling, and thus delivered more accurate pK_{a} values for a series of proteins[104].

The more accurate conformational sampling provided by explicit solvent in hybrid-solvent CpHMD was critical for understanding the origin and direction of pK_{a} shifts of a probe fatty-acid molecule embedded in cationic, anionic, and neutral micelles in a comparison of implicit- and hybrid-solvent CpHMD[171]. Hybrid-solvent CpHMD has been used to investigate the microscopic origins of the pH-dependence of spider dragline silk assembly[172] and to calculate the thermodynamic coupling between a large-scale conformational transition and the ionization of an internal residue in a staphylococcal nuclease mutant[173]. Although hybrid-solvent CpHMD shows promise, there are some inherent limitations of the method. For instance, since conformation and titration sampling are controlled by different energy models, the energetics of conformation and protonation state are not strictly coupled. This decoupling may lead to inaccuracies when simulating pH-dependent conformational transitions[173]. The mixed-energy scheme of hybrid-CpHMD also prevents the method from being combined with the widely-used temperature-replica exchange (TREX)[104] sampling protocol. Additionally, the GB model used in hybrid-solvent CpHMD has been shown to require time-consuming parametrization of the molecular boundary to reproduce explicit-solvent energetics[98] and there is a limit to how closely GB can match explicit-

solvent results[101].

Recently there have been reports of CpHMD simulations of proteins[69] and nucleic acids[70] in explicit solvent where the protonation transitions are driven not by GB, as in the hybrid-solvent approach, but by forces originating from interactions with all atoms of the system. True explicit-solvent CpHMD is very attractive because the conformation and titration energetics are strictly coupled, and parametrization of GB is avoided altogether. However, in previous reports, although model compound titration was successful and tests on small derivatives of the model compounds were carried out, little[70] to no[69] comparison with experimental pK_{a} data was reported.

The comparison of calculated and experimental pK_{a} values is the most direct way to test how accurately CpHMD can model protonation and pH-dependent conformational equilibria of titratable systems. Without such comparisons, whether explicit-solvent continuous constant-pH molecular dynamics (ECpHMD) can be relied upon as a predictive tool for studying the effects of pH in complex systems, as have the GB[65,124] and hybrid-solvent approaches[104,171], remains an open question. In consideration of the emergence of ECpHMD and the lack of comparison to experimental pK_{a} values, we have extended the CpHMD module of CHARMM[39,40] to allow ECpHMD simulation in combination with a generalized reaction field (GRF) treatment of long-range electrostatics[41].

The GRF method is readily adaptable to the CpHMD framework because forces are calculated from strictly-pairwise interactions; unlike smooth Ewald methods where the electrostatic force is calculated in part from a convolution over the charge interpolation grid[174]. The GRF method has been shown to give results comparable to the more expensive Ewald methods in simulations of RNA[175], small peptides[176,177], protein crystals[178], highly-charged proteins[179], as well as the calculation of protein-folding kinetics[180]. Additionally, in a mixed Poisson-Boltzmann/explicit-solvent pHMD method, GRF electrostatics gave more accurate pK_{a} values than particle

mesh Ewald (PME)[54].

We have also implemented a method to circumvent a fundamental problem associated with finite-system simulations at constant pH; a non-neutral and pH-dependent net charge. To address this problem, we couple ionization of titratable sites to the simultaneous neutralization of co-ions in solution. This approach allows the net-charge of the system to remain constant and neutral as titration proceeds. Lastly, we modified the distributed-replica (REPDSTR) module of CHARMM to allow pHREX or TREX to be combined with CpHMD.

We first present the simulation methods used in our study and then test pHREX-CpHMD on a series of aliphatic dicarboxylic acids. We calculate the pK_{a} values of dicarboxylic acids where carboxyl groups are separated by intervening chains of methylene groups of length two to seven using CpHMD using implicit and explicit solvent. We compare pK_{a} values calculated using ECpHMD with and without the charge-leveling procedure.

We find that the difference between the first and second ionization pK_{a} values is severely overestimated (average absolute deviation of nearly 2 pK units) when ECpHMD is applied without neutralizing the net charge, but with charge leveling good agreement with experiment is obtained. The average absolute deviation of the first-ionization pK_{a} values and the difference between the first and second pK_{a}'s are both 0.18 pK units (excluding succinic acid) from ECpHMD simulation with charge leveling, while the average absolute deviation of the first pK_{a} values calculated using GB is 0.43. We examine pH-dependent conformational propensities and changes in solvent distributions as a function of pH for the longest dicarboxylic acid (azelaic acid) which has seven methylene groups separating the titration sites.

We then calculate pK_{a} values of protein side chains. We test the method on two small proteins (HP36 and BBL) and a moderately sized enzyme (HEWL) which are known to have pK_{a} values that deviate significantly from model compound values

due electrostatic and desolvation effects. For the most compact protein (HP36) we calculated p$K_{\rm a}$ values with and without charge leveling. We also compare the results obtained using charge leveling at zero ionic strength and at a salt concentration that matches experiment. We find that the calculated p$K_{\rm a}$ values deviate significantly (average absolute deviation $\geq$ 2.1 pK units) without charge leveling, but this deviation is reduced to 1.2 pK units when charge leveling is applied. The addition of salt to match experiment reduces the average absolute deviation further to 0.7 units. Results for BBL are in good agreement with experiment, having a root-mean-squared deviation (RMSD) of 0.3 pK units. For HEWL, the results are mixed. The p$K_{\rm a}$ values of surface residues are in good agreement with experiment; however, for several residues which have the ability to form salt bridges and hydrogen bonds, p$K_{\rm a}$ values deviate by over 1 pK unit as a result of inadequate conformational sampling.

This work represents, to the best of our knowledge, the first test of ECpHMD's ability to quantitatively predict p$K_{\rm a}$ values for complex molecules, while demonstrating the necessity to maintain charge neutrality and sample all relevant conformational states for the accurate calculation of p$K_{\rm a}$ values.

6.3 Methods

6.3.1 Explicit-solvent continuous constant-pH molecular dynamics with charge leveling

Although previously outlined in Chapter 1, we describe the CpHMD framework here for completeness. The CpHMD approach, based on the λ-dynamics technique[77], uses an extended Hamiltonian to simultaneously propagate spatial (real) and titration (virtual) coordinates. The total Hamiltonian of the system can be written as

$$\mathcal{H}(\{r_a\},\{\theta_i\}) = \sum_a \frac{m_a}{2}\dot{r}_a^2 + U^{\rm int}(\{r_a\}) + U^{\rm hybr}(\{r_a\},\{\theta_i\}) + \sum_i \frac{m_i}{2}\dot{\theta}_i^2 + U^*(\{\theta_i\}), \quad (6.1)$$

where $a = 1, N_{\text{atom}}$ is the index for atomic coordinates, and $i = 1, N_{\text{titr}}$ is the index for the continuous variables (θ_i) which are related to the titration coordinates (λ_i) by

$$\lambda_i = sin^2(\theta_i). \tag{6.2}$$

Boundaries are naturally imposed on titration coordinates through the sine function, where $\lambda_i = 0$ corresponds to the protonated state and $\lambda_i = 1$ corresponds to the unprotonated state. For residues with two competing titration sites, a second continuous variable is included to allow interconversion between tautomers.

In Eq. 6.1, the first term is the kinetic energy of the real system (atoms), U^{int} is the internal potential energy which is independent of titration, and U^{hybr} is the non-bonded energy involving titration sites which enables coupling between conformation and titration degrees of freedom. The last term (U^*) comprises a biasing potential (U^{barr}) to suppress intermediate values of λ (see Eq. 1.23), the model compound potential of mean force (PMF)(U^{mod}), and the term that imposes pH-dependence onto the protonation equilibria (U^{pH}). Together they are written as

$$U^*(\{\theta_i\}) = \sum_i (-U^{\text{mod}}(\theta_i) + U^{\text{pH}}(\theta_i) + U^{\text{barr}}(\theta_i)), \tag{6.3}$$

where the pH-dependent term is given by

$$U^{\text{pH}}(\lambda_i) = log(10) k_b T (pK_a^{\text{ref}} - \text{pH})\lambda_i, \tag{6.4}$$

and pK_a^{ref} is the experimentally determined pK_a of a reference molecule. For titratable groups with a single proton-binding site, the model compound PMF is fit with a harmonic potential given by

$$U^{\text{mod}}(\theta_i) = A(sin^2(\theta_i) - B)^2 \tag{6.5}$$

where A and B are fitting parameters. For residues with two tautomeric states, the PMF is both second-order in λ and x (an additional tautomeric degree of freedom) [66]. To determine the parameters A and B, we fit the derivative of Eq. 6.5 to the mean force calculate at several values of θ_i.

In order to maintain charge neutrality, we couple deprotonation of each acidic site to neutralization of a chloride ion and deprotonation of each basic site to ionization of a sodium ion. The net reaction under consideration is

$$\mathrm{AH} + \mathrm{Cl}^- \rightleftharpoons \mathrm{A[H]}^- + \mathrm{Cl}^0 \tag{6.6}$$

for each acidic site and

$$\mathrm{BH}^+ + \mathrm{Na}^0 \rightleftharpoons \mathrm{B[H]}^0 + \mathrm{Na}^+ \tag{6.7}$$

for each basic site. Protons on the product-side of the above chemical equilibria are in brackets to indicate that their electrostatic and vdW interactions have been turned off. In the CpHMD approach we do not allow these product-protons to move into solution, but instead account for the effects of proton concentration by including the pH-bias energy term (Eq. 6.4). In the proposed charge-leveling procedure, the PMF for each titration process is the sum of the individual PMFs for the deprotonation of a reference compound and the ionization or neutralization of a co-ion. Since on each side of the equilibria (Eq. 6.6 and 6.7) only one species carries a net-charge, interaction between the titration site and it's co-ion is minimized.

The hybrid-energy term (U^{hybr}) is written as a sum of the non-bonded energy terms; van der Waals, Coulombic, and generalized reaction field. Each has explicit dependence upon the titration degrees of freedom and are together written as

$$U^{\mathrm{hybr}}(\{r_a\},\{\theta_i\}) = U^{\mathrm{vdW}}(\{r_a\},\{\theta_i\}) + U^{\mathrm{Coul}}(\{r_a\},\{\theta_i\}) + U^{\mathrm{GRF}}(\{r_a\},\{\theta_i\}) \tag{6.8}$$

where the latter term (U^{GRF}) is given by

$$U^{\mathrm{GRF}} = -\frac{q_a q_b}{4\pi\epsilon_0\epsilon_{in}}\left(\frac{0.5C_{RF}r_{ab}^2}{R_c^3} + \frac{1-0.5C_{RF}}{R_c}\right). \tag{6.9}$$

Here, r_{ab} is the distance between two atoms, q_a and q_b are the respective (possibly λ dependent) instantaneous partial charges (see Eq. 1.19), and ϵ_{in} is the relative dielectric constant (typically set to one) within the the non-bonded cutoff radius R_c. C_{RF} includes ionic strength dependence, governs the magnitude of the reaction field term, and is given by

$$C_{RF} = \frac{(2\epsilon_{in} - 2\epsilon_{out})(1+\kappa R_c) - \epsilon_{out}(\kappa R_c)^2}{(\epsilon_{in} + 2\epsilon_{out})(1+\kappa R_c) + \epsilon_{out}(\kappa R_c)^2} \tag{6.10}$$

where ϵ_{out} is the relative dielectric constant of the surrounding medium (e.g. the value for water) and κ is the inverse Debye screening length[41] where $\kappa^2 = 8\pi q^2 I/ek_bT$ and I is the ionic strength.

With our charge-leveling procedure, the free-energy difference between the deprotonation of the reference compound in solution and that in the environment of interest includes contributions from the titrating site as well as the co-ion and can be written as

$$\Delta G_{env} - \Delta G_{sol} = (\Delta G_{env}^{titr} - \Delta G_{sol}^{titr}) + (\Delta G_{env}^{ion} - \Delta G_{sol}^{ion}). \tag{6.11}$$

Similar to the quantum mechanical effects (see discussion of CpHMD in Chapter 1), the co-ion will largely stay in solution and the charging/neutralization process will not be significantly affected by the presence of the titrating site ($\Delta G_{env}^{ion} \approx \Delta G_{sol}^{ion}$). Thus, the majority of the change in deprotonation free energy will be due to the titration reaction.

6.3.2 Data analysis

To test the predictive accuracy of our proposed ECpHMD method, we examine two sets of test systems. First, we calculate p$K_{\rm a}$ values for a series of dicarboxylic acids with varying lengths of methylene chains separating the carboxyl moieties. Considering the stepwise deprotonation equilibria of aliphatic dicarboxylic acids given by

$$\mathrm{HO_2C(CH_2)}_n\mathrm{CO_2H} \overset{K_1}{\rightleftharpoons} \mathrm{HO_2C(CH_2)}_n\mathrm{CO_2^-} + \mathrm{H^+} \overset{K_2}{\rightleftharpoons} {}^-\mathrm{O_2C(CH_2)}_n\mathrm{CO_2^-} + \mathrm{H^+}, \tag{6.12}$$

the total average protonation ($\langle P \rangle$) at each pH is given by

$$\langle P \rangle = \frac{10^{pK_2-pH} + 2 \times 10^{pK_1+pK_2-2pH}}{1 + 10^{pK_2-pH} + 10^{pK_1+pK_2-2pH}} \tag{6.13}$$

where pK_1 and pK_2 are the first and second deprotonation p$K_{\rm a}$'s[59,181]. Secondly, we calculate the p$K_{\rm a}$ values of acidic residues and histidines of three proteins. The p$K_{\rm a}$ values of proteins are calculated by fitting the unprotonated fraction (S) at each pH to the Hill equation given by

$$S = \frac{1}{1 + 10^{n(pK_a-pH)}} \tag{6.14}$$

where n (the Hill coefficient) and the p$K_{\rm a}$ are fitting parameters.

To quantify the correlations between two protonation equilibria and between protonation state and conformation, we make use of cross-correlation analysis. The normalized cross-correlation function between two time series (x and y) shifted by an offset (Δr) is defined by

$$R_{xy}(\Delta r) = \sum_i \frac{[x(i) - \mu_x][y(i + \Delta r) - \mu_y]}{\sigma_x \sigma_y}, \tag{6.15}$$

where μ is the population mean and σ is the standard deviation. Normalized cross-

correlation values range from -1 (completely anti-correlated) to +1 (completely correlated).

6.3.3 Simulation details

All simulations were carried out using CHARMM[39,40]. Modifications were made to the PHMD module to allow the application of GRF and the charge-leveling procedure. All simulations used the CHARMM22/CMAP force field[36]. The SHAKE algorithm was applied to all bonds and angles involving hydrogen to allow a 2 fs time step. The titration coordinates were propagated using the Langevin algorithm with a collision frequency of 5 ps^{-1}. All simulations were carried out at 300 K. The mass of the fictitious lambda particles was set to 10 atomic mass units for dicarboxylic acids and amino acids, but 20 atomic mass units for proteins to match slower conformational dynamics.

We used the GBSW model[79] with the atomic input radii of Nina et. al.[117] and a surface tension coefficient of 0.005 kcal mol^{-1} $Å^2$ for all implicit-solvent CpHMD simulations of dicarboxylic acids. In the GB simulations, conformational dynamics was propagated via the Langevin algorithm with a collision frequency of 5 ps^{-1} and non-bonded interactions were truncated at a cut-off radius of 20 Å using a switching function.

Our ECpHMD simulations used the modified CHARMM TIP3P water model[134], an updated sodium vdW radius[182], and a modified sodium chloride vdW interaction distance to reduce sodium chloride contact-ion pair formation in concentrated solutions[183]. Non-bonded interactions were truncated at 14 Å, beyond which electrostatic effects were treated by GRF. All explicit-solvent simulations were carried out with periodic-boundary conditions at ambient temperature and pressure using the the Hoover thermostat[136] and Langevin piston pressure-coupling algorithm[137] as in previous hybrid-solvent CpHMD simulations[104]. In the GRF term, ϵ_{in} was

set to 1.0 and ϵ_{out} was set to 80.0. The ionic strength in C^{RF} was set to zero in dicarboxylic acid simulations as the experimental data was extrapolated to zero ionic strength. In ECpHMD simulations of proteins, the ionic strength was set to 50 mM for HEWL[133], 200 mM for BBL[131,184], and 100 mM for for HP36[130] to match the experiments. A cubic water-box with 30 Å edge-length was used for all ECpHMD simulations of dicarboxylic acids, while for proteins, edge-lengths were 14 Å greater than the longest dimension of the proteins.

We calculated pK_{a} values of three proteins: the 45-residue binding domain of 2-oxoglutarate dehydrogenase multi-enzyme complex, BBL (PDB: 1W4H), the 36-residue subdomain of villin headpiece, HP36 (PDB: 1VII), and the 129-residue hen egg white lysozyme, HEWL (PDB: 2LZT). Protein preparation was carried out as described previously[104], except that C-terminal residues were amidated due to uncertainty in the reference C-terminus pK_{a} value.

In ECpHMD simulations (where specified), we added a chloride co-ion for each acidic site and a sodium co-ion for each titrating basic site (i.e. histidine) in order to maintain charge-neutrality. We then added additional sodium or chloride to neutralize the net charge. Finally, additional sodium chloride was added to reach the experimental ionic strength. Ions were placed randomly within the simulation box and the position of each ion was relaxed with 100 MC moves using a constant dielectric model with $\epsilon = 80.0$. All water molecules within 2.6 Å of an ion or a solute heavy atom were then deleted. Water molecules and ions were subjected to energy minimization with restrained solute prior to starting ECpHMD.

Simulations using pHREX[104,185] and TREX[80,81] were carried out using the modified REPDSTR module in CHARMM. Exchanges were attempted every 500 MD steps (1 ps) for all replica-exchange simulations (pHREX or TREX). The pHREX simulations of dicarboxylic acids were carried out for 5 ns per replica with pH conditions ranging from 3.0 to 7.0 with a pH interval of 0.5 pH units, except for succinic acid

where the highest pH had to be increased to 10.0. The pHREX simulations of amino acids are described later (see discussion of amino acid reference PMF calculation). The pHREX simulations of proteins used the same pH conditions as used in previous work[104]. Simulations were carried out for 10 ns per replica for the small proteins (HP36 and BBL) and 7 ns per replica for the larger enzyme (HEWL).

The *temperature generator for REMD-simulations* website (http://folding.bmc.uu.se/remd/)[186] was used to estimate the temperature conditions for a target exchange ratio of 0.3 given the size of the protein and the number of solvent molecules for TREX simulations of proteins. A TREX simulation of HEWL was carried out at pH 0 with 30 replicas ranging from 300–330 K for 1.5 ns per replica. Separate TREX simulations were conducted for HP36 at pH 0, 2, 4 and 6. TREX simulations of HP36 used 16 replicas with temperatures ranging from 300–338 K. Each TREX simulation of HP36 was run for 2 ns per replica.

6.3.4 Deriving reference compound potential of mean force

Dicarboxylic acids and co-ions

For both implicit- and explicit-solvent simulations of dicarboxylic acids, the reference deprotonation event was the first ionization of azelaic acid ($n = 7$) and the reference pK_{a} was 4.55[187]. Thermodynamics integration (TI) was carried out at θ values of 0.2, 0.4, 0.6, 0.7854, 1.0, 1.2, and 1.4 for 500 ps in explicit-solvent and 100 ps in implicit-solvent. For PMF calculations of co-ions in explicit solvent, the same θ values were used and TI simulations were run for 500 ps. The cumulative-average force indicated that the TI simulations were converged. The PMFs of azelaic acid deprotonation and of sodium neutralization and chloride ionization in explicit-solvent are shown in Figure 6.1.

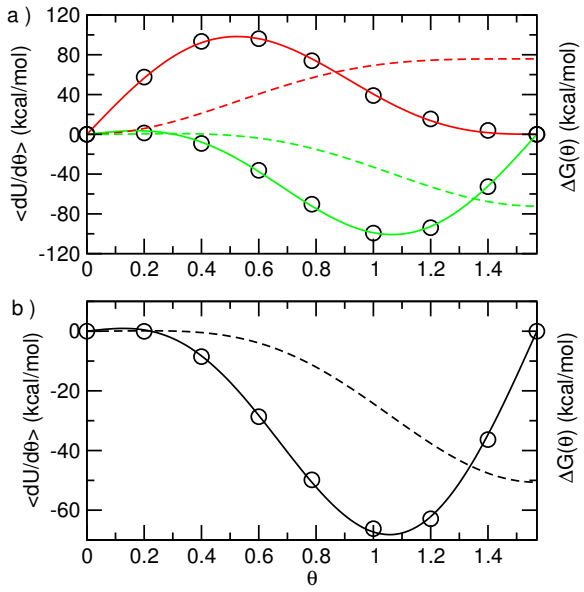

Figure 6.1: Potentials of mean force for co-ions and azelaic acid from explicit-solvent continuous constant-pH molecular dynamics simulations. Average force at each θ value is shown as circles and the fitting-function are shown as solid lines. The PMFs (dashed lines) in each plot were obtained by integrating the fitting functions. Average-force values are denoted by the y-axis label on the left. Values of the PMF are denoted by the y-axis label on the right: a) Chloride neutralization (red) and sodium ionization (green) and b) First deprotonation reaction of azelaic acid.

Amino acids

The molecules used for the calculation of the reference PMFs for protein side chains were amino acids acetylated at the N-terminus and N-methylamidated at the C-terminus, as used previously[64–66,104]. Unlike previous work, model compound parametrization was carried out using an iterative approach. The model compound PMFs of amino acids in implicit solvent can be fit with the derivative of Eq. 6.5; however, as noted by others[69], we observed deviation from this functional form using explicit solvent. Figure 6.2 shows the TI data for lysine deprotonation from our ECpHMD simulations with the best-fit U^{mod} derivative. As shown in the figure, there is significant deviation (up to 5 kcal/mol) between the average force and the best-fit curve. Since the required quantity is the free-energy difference, the actual path from the protonated to unprotonated state is not critical and only the net change is important. To derive the reference compound PMFs for amino-acid side chains us-

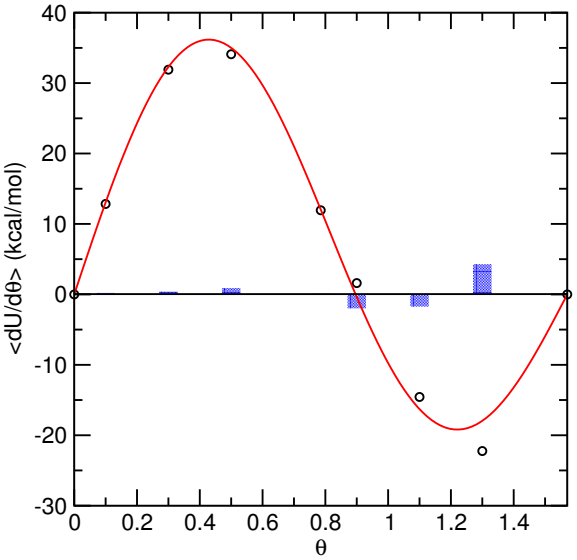

Figure 6.2: Average forces for Lys deprotonation from explicit-solvent continuous constant-pH simulations. Average force at each θ value is shown as a circles best-fit U^{mod} (Eq. 6.5) derivative is shown as solid line and residuals of the fit are shown as bars.

ing ECpHMD, we began by using PMF parameters from our previous hybrid-solvent

work[104] and calculated the pK_a's using pHREX-ECpHMD. The pHREX-ECpHMD simulations were carried out using three pH conditions: Asp (4 ± 1), Glu (4.4 ± 1), His (6.5 ± 1), and Lys (10.4 ± 1). Five separate trials were run and that data was combined to calculate the pK_a values. By relating the deviation in the calculated pK_a value with the error in the deprotonation free-energy, after several rounds we arrived at a set of parameters for the deprotonation reactions in explicit-solvent that exactly (within the uncertainty) cancels out ΔG_{sol}^{sim} (see Eq. 1.27) and reproduces the experimental model compound pK_a values, as shown in Table 6.1. Titration curves for the model compounds are shown in Figure 6.3.

Table 6.1: Calculated and experimental pK_a values of amino acids from explicit-solvent continuous constant-pH molecular dynamics simulations.

Residue	Calc pK_{a}	Calc Hill	Ref pK_{a}
Asp	4.03 ± 0.07	0.99 ± 0.08	4.0
Glu	4.43 ± 0.10	0.96 ± 0.07	4.4
His	6.54 ± 0.13 (6.76†/7.05‡)	1.03 ± 0.16	6.45* (6.6†/7.0‡)
Lys	10.41 ± 0.05	1.14 ± 0.23	10.40

The pK_{a} values were calculated by combining data from five separate trials. The length of each simulation was 4 ns per replica. Average values and errors were calculated from 100 trial bootstrap fittings[157]. Reference pK_{a}'s are taken from Nozaki and Tanford[92] except for Nδ- and Nϵ-sites of His[129]. †Nδ site pK_{a} values. ‡Nϵ site pK_{a} values. *Macroscopic pK_{a} for His[66].

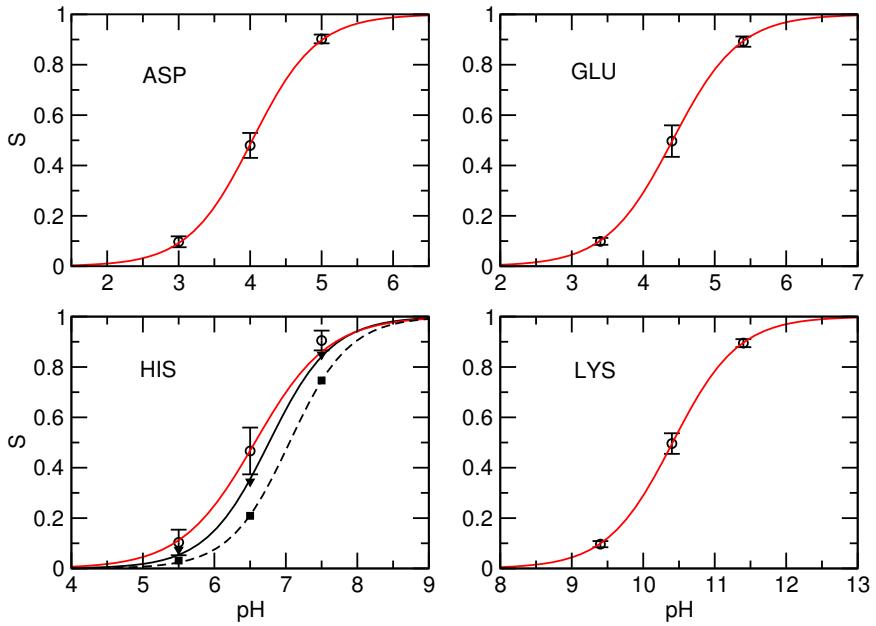

Figure 6.3: Amino acid titration curves from pH-replica exchange explicit-solvent continuous constant-pH simulations. Average S values are shown as circles and error bars indicate the standard deviation between 5 trials. Solid lines are the best-fit Hill equation (see Eq. 6.14). Data and best fit curves for δ- (triangles and solid black line) and ϵ-sites (squares and dashed black line) are shown for histidine.

6.4 Results and Discussion

6.4.1 Dicarboxylic acids

6.4.1.1 Accuracy of Calculated pK_a values

We begin by calculating first and second pK_a values for the aliphatic dicarboxylic acid series described previously (see chemical equation 6.12). This is a classic problem used to test our understanding of the electrostatic influence on acidic ionization that was first investigated using analytical theory nearly a century ago[188,189]. This series of molecules has some attractive qualities as a test case for computer simulation. There are two interacting titration sites so the accurate calculation of both pK_a values is non-trivial. The small size of the molecules means simulations can be carried out quickly and adequate sampling of all relevant conformations should not be an issue. Also, the experimental data is extrapolated to zero ionic strength[187] so the added complexity in the deprotonation energetics due to salt-screening is not present.

Since we compare the pK_a values from GB and explicit-solvent CpHMD, we show the model PMF parameters and deprotonation free-energy results from the two solvent models in Table 6.2. It is worth noting that the deprotonation free-energy from ECpHMD is 25% greater than the value obtained using GB-based CpHMD.

The first and second pK_a values, as well as the pK_a shifts, from experiment and simulation are shown in Table 6.3. We calculated the pK_a values using GB and ECpHMD. The explicit-solvent simulations were conducted both with (E+CL) and

Table 6.2: Potential of mean force parameters and deprotonation free energy for azelaic acid

Solvent	Parameters		ΔG^{deprot}(kcal/mol)
	A	B	
GB	-62.39	0.18	-39.93
Explicit	-56.16	0.05	-50.54

Table 6.3: Experimental and calculated pK_a's and pK_a shifts of dicarboxylic acids

Acid		Expt[a]	GB	E-CL	E+CL
	n		pK_1		
succinic	2	4.19	8.08 (0.07)	7.0 (0.04)	7.2 (0.2)
glutaric	3	4.34	3.57 (0.02)	3.7 (0.1)	4.0 (0.3)
adipic	4	4.42	4.04 (0.03)	4.1 (0.1)	4.7 (0.2)
pimelic	5	4.48	4.17 (0.04)	4.2 (0.1)	4.5 (0.4)
suberic	6	4.52	4.19 (0.04)	4.3 (0.1)	4.6 (0.2)
azelaic	7	4.55	4.20 (0.03)	4.2 (0.1)	4.4 (0.2)
			pK_2		
succinic	2	5.48	9.05 (0.05)	9.3 (0.1)	7.7 (0.1)
glutaric	3	5.42	5.05 (0.05)	6.6 (0.1)	5.0 (0.4)
adipic	4	5.41	5.42 (0.04)	7.1 (0.1)	5.4 (0.2)
pimelic	5	5.42	5.25 (0.06)	7.0 (0.1)	5.3 (0.4)
suberic	6	5.40	5.25 (0.05)	7.1 (0.2)	5.2 (0.2)
azelaic	7	5.41	5.18 (0.06)	7.1 (0.1)	5.3 (0.3)
			ΔpK_a		
succinic	2	1.29	0.97 (0.09)	2.3 (0.1)	0.5 (0.2)
glutaric	3	1.08	1.48 (0.05)	2.9 (0.1)	1.0 (0.5)
adipic	4	0.99	1.38 (0.05)	2.9 (0.1)	0.7 (0.3)
pimelic	5	0.94	1.08 (0.07)	2.8 (0.1)	0.8 (0.6)
suberic	6	0.88	1.06 (0.06)	2.8 (0.2)	0.6 (0.3)
azelaic	7	0.86	0.98 (0.07)	2.9 (0.1)	0.8 (0.4)

[a] [187] n refers to the number of methylene groups in the chemical structure $HO_2C(CH_2)_nCO_2H$. Values in parentheses are the uncertainty calculated as the standard deviation of pK_a values calculated from 1 ns windows.

without charge-leveling co-ions (E-CL). The GB and E-CL results both have random errors significantly smaller than those seen in E+CL, as shown in Table 6.3. We suspected that the relatively large pK_a value uncertainty from E+CL is a result of slow diffusion of ions. To test this, we analyzed the correlation between the sodium-ion distribution and pimelic acid pK_a values from 1 ns windows. We investigated this correlation for pimelic acid because it has the greatest pK_a value uncertainty. We combined E+CL trajectories from all pH conditions and calculated the sodium to carboxylate-oxygen radial distribution functions (RDF) from 1 ns simulation windows (see Figure 6.4).

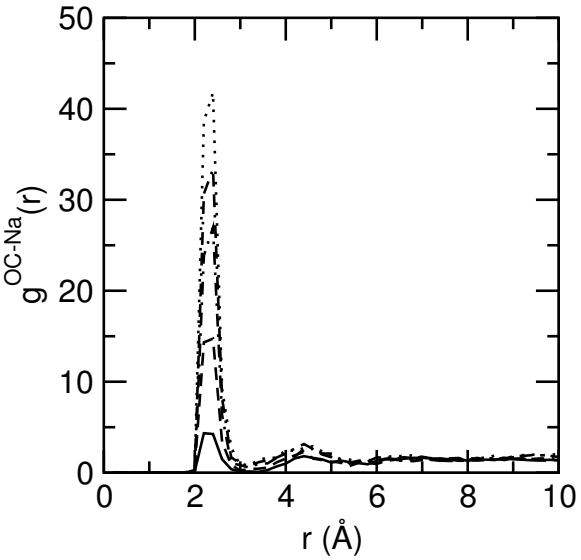

Figure 6.4: Pimelic acid carboxyl-oxygen to sodium radial distribution functions. RDF of pimelic carboxyl-oxygen to sodium from 1 ns windows. Each trace is the RDF from a separate 1 ns window.

We then performed linear-regression of the relative intensity of the RDF at a separation distance of 2.5 Å versus the first and second pK_a values. The sodium to carboxylate-oxygen RDF at 2.5 Å corresponds to the contact-ion pair distance. Regression of the first pK_a gives a slope of -0.97 and an R^2 value of 0.74. Similar results are obtained when we performed regression of the RDF intensity versus the second pK_a (slope = -0.87 and R^2 = 0.67). This data demonstrates that over a 1

ns simulation window, changes in the population of carboxylate-sodium contact-ion pairs correlates with the observed pK_a values. The pK_a values decrease as more sodium and carboxyl-oxygen contact-ion pairs are observed.

Even thought the GB and E-CL results are more precise, the first pK_a values deviate from the experimental value by more than the calculated error, while the results from E+CL are in agreement (within the calculated error) with nearly all experimental first pK_a values, except for succinic acid. Comparing the accuracy of the first pK_a values, excluding succinic acid, from the three simulations, the average absolute deviation from E+CL is 0.18, the average absolute deviation from E-CL is 0.38, while that from GB is the poorest of the three at 0.43. The deviation between the calculated and experimental first pK_a's of succinic acid is 3-4 units for all three simulations indicating that the model compound PMF of succinic acid deviates significantly from that of azelaic acid.

E-CL fails miserably for the second pK_a's, having an average absolute deviation of 1.6 pK units. This indicates that the electrostatic repulsion between the first and second carboxyl groups, when ionized, is grossly overestimated when ECpHMD is conducted without neutralizing the net charge. On the other hand, when including co-ions which compete for charge with the titration sites such that the net charge of the systems is neutral, the average absolute deviation is reduced by nearly an order of magnitude to 0.18 pK units. This data indicates that in order to obtain quantitative agreement between calculation and experiment, one cannot simply ignore fluctuations of the net charge.

Our method of adding discrete co-ions that serve as charge reserviors is not the only possible approach to enforce charge neutrality. For example, instead of adding additional co-ions, one could imagine a scheme where the excess charge is distributed to the solvent molecules such that each of them would carry a fractional net-charge. This approach would be very cumbersome and difficult to implement, because the

addition and removal of charge from every water molecule would have to be coupled simultaneously to the titration of *all* titratable sites. Our approach is straightforward and provides calculated p$K_{\rm a}$ values for the dicarboxylic acids that are in good agreement with experiment.

The next quantity to be compared is the difference between the first and second p$K_{\rm a}$ values. The trend follows as before since this quantity is simply the difference between the first and second pK_a's. The average absolute deviation is 1.9 from E-CL, 0.24 from GB, and 0.19 from E+CL. The data in Table 6.3 is shown graphically in Figure 6.5 to facilitate comparison. There is a splitting of the first and second p$K_{\rm a}$

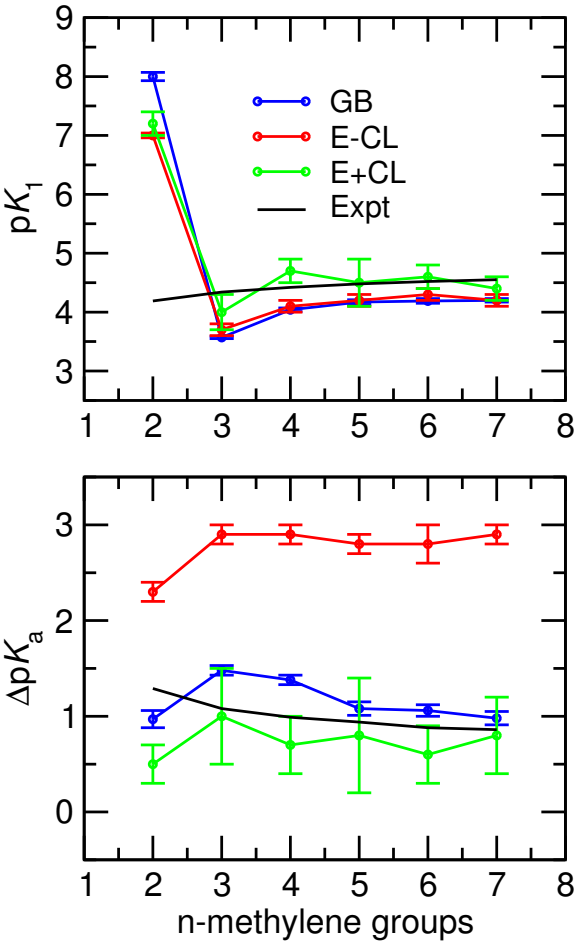

Figure 6.5: First p$K_{\rm a}$'s and p$K_{\rm a}$ shifts of dicarboxylic acids. The first p$K_{\rm a}$'s and p$K_{\rm a}$ shifts (pK_2-pK_1) for dicarboxylic acids with n intervening methylene groups. Error bars indicate the standard deviation of the p$K_{\rm a}$ values calculated from 1 ns windows.

values. There are two sources of the splitting. The first source arises from statistical factors due to the number of equivalent species on each side of the chemical equilibria. Looking at Eq 6.12, it is immediately apparent that $K_1 = 4K_2$ in the absence of charge-charge interactions between the carboxyl groups. Thus, the statistical factor is responsible for 0.6 units of $\Delta \mathrm{p}K_\mathrm{a}$. The second source of the $\mathrm{p}K_\mathrm{a}$ splitting, which is more difficult to accurately account for, is electrostatic in nature. Deprotonation of the first carboxyl group imposes an electrostatic penalty for deprotonation of the second, and causes the $\Delta \mathrm{p}K_\mathrm{a}$'s to be greater than 0.6 units. Similarly, it is reasonable to expect that the instantaneous protonation states of the two equivalent carboxyl groups should be anti-correlated with one another. If one carboxyl group is ionized, this should favor protonation of the other given that we are at an intermediate pH that allows both groups to be protonated. We calculated the normalized cross-correlation function with an offset ranging from -100 to 100 exchange cycles (see Eq. 6.15) for glutaric acid at pH 4 to examine the extent of correlation, or coupling, between the two carboxyl groups protonation states. We present the data from the three different simulations (GB, E-CL, and E+CL) for glutaric acid (see Figure 6.6), because the $\mathrm{p}K_\mathrm{a}$ shift of glutaric acid was calculated with good accuracy and the correlation between protonation states was the greatest of the dicarboxylic acids, excluding succinic acid. The cross-correlation at zero delay time correlates with the accuracy of the $\mathrm{p}K_\mathrm{a}$ shift. There is a strong anti-correlation between the two titratable groups protonation states and the calculated $\Delta \mathrm{p}K_\mathrm{a}$ is severely overestimated from E-CL. The cross-correlation is reduced and the accuracy of $\Delta \mathrm{p}K_\mathrm{a}$ improves with GB. From E+CL, the cross-correlation is virtually nonexistent while $\Delta \mathrm{p}K_\mathrm{a}$ is the most accurate. Thus, it appears that overestimation of the electrostatic repulsion between ionized carboxyl groups from GB and E-CL causes exaggerated correlation between the carboxyl group's protonation states and the calculated $\Delta \mathrm{p}K_\mathrm{a}$ to be overestimated.

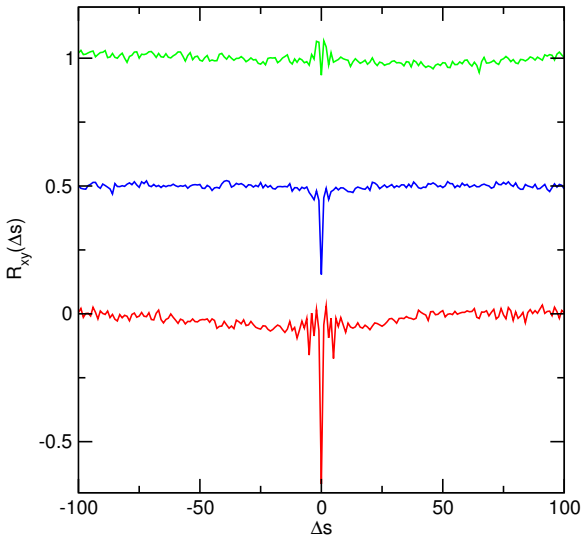

Figure 6.6: Correlation between the protonation states of glutaric acid. Cross-correlation between protonation states of glutaric acid at pH 4 from GB (blue), explicit solvent (red), and explicit solvent with charge leveling (green). Baseline of cross-correlation functions from explicit-solvent with charge leveling and GB have been shifted for clarity.

pH-dependent protonation, conformation, and solvent distribution of azelaic acid

With reasonable pK_a values obtained for the dicarboxylic acids (excluding succinic acid), we move on to analyze the pH-dependent protonated fractions, protonation state populations, and the exchange-efficiency of pHREX-ECpHMD. These data are shown for azelaic acid in Figure 6.7. The upper panel of Figure 6.7 shows the average number of bound protons at each pH, as well as the fit to Eq. 6.13. The data are fit well by the model as demonstrated by a fitting-correlation of >0.999. The middle panel of Figure 6.7 shows the fraction of each protonated form. At pH 3, azelaic acid is almost completely doubly protonated, at intermediate pH the largest population is singly protonated, and at pH values above the second p$K_a$the molecule is almost completely in the fully unprotonated form. The lower panel of Figure 6.7 shows that the exchange ratio (number of successful exchanges divided by the number attempted) between replicas is at a minimum where the doubly protonated, singly protonated,

and fully unprotonated states coexist. This indicates that titration events reduce the number of successful exchanges between neighboring replicas.

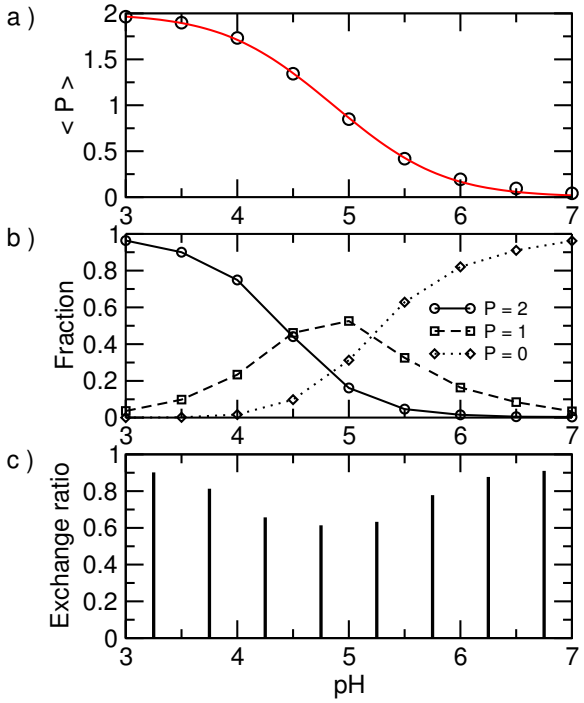

Figure 6.7: Populations and exchange rates of azelaic acid. a) Average protonation state ($\langle P \rangle$) at each pH. Data is shown as circles and best-fit (Eq. 6.13) curve is shown as a solid line. b) Fraction of population having two ($P = 2$), one ($P = 1$), and zero ($P = 0$) bound protons at each pH. c) Exchange ratio between neighboring replicas.

Looking at the conformations of azelaic acid at different pH values, we see that when azelaic acid is doubly protonated it preferentially occupies conformations with a shorter end-to-end distance when fully unprotonated. The distribution of end-to-end distance at pH 7 has greater intensity at 9 and 10 Å and reduced intensity at distances below 8.5 Å, relative to the distribution at pH 3 (Figure 6.8a). This is expected since there is electrostatic repulsion between carboxylate groups. Considering the solvent organization at different pH values, at more basic pH there is an increase in the carboxyl-oxygen and water-oxygen RDF at 3 Å which corresponds to the distance at which hydrogen-bonding between the carboxyl groups and water occurs (Figure 6.8b). The decrease in intensity of this peak at basic pH indicates that hydrogen bonding

when water is the donor is more stable than when the carboxyl group is the hydrogen-donor. There is an increase in the sodium carboxyl-oxygen RDF at a separation distance near 2.5 Å at pH 7 that corresponds to the carboxyl-oxygen sodium contact-ion pair (Figure 6.8c). There is also an increase in the water-hydrogen carboxyl-oxygen RDF at a separation distance near 2 Å at pH 7 that corresponds to water donating a hydrogen bond to solvate the carboxyl group (Figure 6.8d).

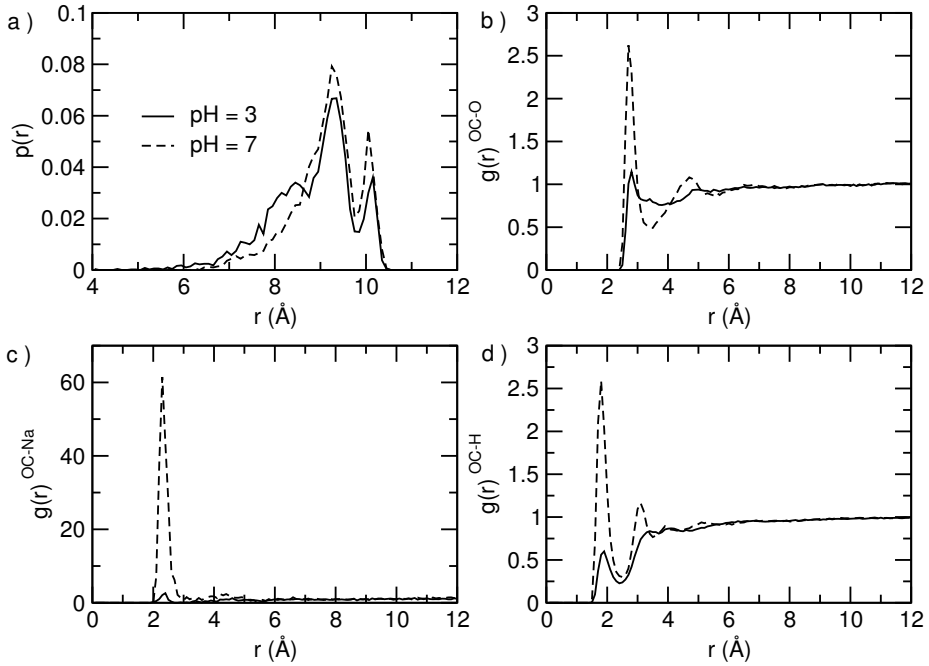

Figure 6.8: Conformation and solvent distributions of azelaic acid at different pH conditions. a) Probability distribution of distances between carboxyl-carbons. b) RDF between carboxyl-oxygen and water-oxygen. c) RDF between carboxyl-oxygen and sodium. d) RDF between carboxyl-oxygen and water-hydrogen.

6.4.2 Proteins

The effect of charge leveling on pK_a calculation accuracy

With encouraging results for dicarboxylic acids, we tested ECpHMD on a small protein (HP36) which has several acidic residues with pK_a's downshifted relative to standard values. We conducted pHREX-ECpHMD using three setups: without charge-leveling co-ions (E-CL), with charge-leveling co-ions (E+CL), and with

charge-leveling and additional salt (E+CL+Salt) to match experiment[130]. These simulations allow us to independently probe the effect of an overall change in the net charge at different pH conditions and the effects of salt-induced electrostatic screening. The results are shown in Table 6.4.

The calculated p$K_{\rm a}$ order from E-CL is in agreement with the experimental values;

Table 6.4: Experimental and calculated pK_a values of HP36

Residue	Expt[a]	Calc.		
		E-CL	E+CL	E+CL+Salt
Asp44	3.10 (0.01)	< 0.0	0.75	1.86
Glu45	3.95 (0.01)	2.16	3.82	4.41
Asp46	3.45 (0.12)	1.57	3.92	3.64
Glu72	4.37 (0.03)	3.01	4.26	4.77
	RMSD	≥ 2.13	1.20	0.70

[a] [130]. In the E+CL+Salt simulation, the ionic strength was 150 mM as in experiment. p$K_{\rm a}$ values were calculated using 2 ns simulations.

however, the magnitudes of the p$K_{\rm a}$ shifts are severely overestimated. Considering that, without charge leveling, the net charge becomes increasingly positive as the pH is reduced and successive acidic residues are protonated (first Glu72, then Glu45, Asp46, and finally Asp44). The overestimation of p$K_{\rm a}$ shifts is in line with our conjecture, and our results for dicarboxylic acids, that it is important to neutralize the net-charge of the system to accurately model bulk-deprotonation equilibria in MD simulations. Further supporting this argument, the RMSD of calculated p$K_{\rm a}$ values is reduced to 1.20 pK units when charge leveling is applied. In ES+CL, the p$K_{\rm a}$ shift of Asp44 is still overestimated as a result of an apparent overestimation of the Asp44-Arg55 interaction strength (see later discussions). Addition of sodium chloride to match experiment brings the calculated p$K_{\rm a}$ of Asp44 more in line with the experimental value, from 0.75 at $I = 0$, to 1.86 at $I = 150$ mM. Although these simulations are relatively short, they demonstrate that reasonable protein p$K_{\rm a}$ values can be obtained using fully explicit-solvent CpHMD. Our data indicate that net charge

neutralization is necessary in order to obtain good agreement with experiment. Also, careful and deliberate system setup to more closely match experimental conditions gives more favorable results.

Accuracy of calculated pK_a values

Although calculated pK_a values for dicarboxylic acids and HP36 are encouraging, we sought to test how the pK_a's of HP36 vary over longer simulation time. We also calculated pK_a values of proteins with pK_a shifts resulting from burial in hydrophobic environments to further test the reliability of pK_a calculation using pHREX-ECpHMD with charge leveling. Experimental and calculated pK_a values for HP36, BBL, and HEWL from multi-nanosecond simulations are given in Table 6.5. The RMSD of calculated pK_a values using the entire the simulation (where pK_a values were calculable) ranges from 0.3 for BBL to 1.1 for HEWL, while the value is 0.8 for HP36. The average absolute deviation is below one pK unit for all three proteins. Overall, this level of accuracy is on par with pK_a calculation accuracy from hybrid-solvent CpHMD[104].

We plot the calculated versus experimental pK_a values in Figure 6.9. The slope of the regression line is 1.1 and the R^2 value is 0.83. Although our previous study using hybrid-solvent CpHMD included results for two additional proteins, our regression data indicates that a slightly better correlation with experiment is obtained using ECpHMD. As depicted in Figure 6.9, 3 out of 22 residues have an absolute error of $\geq$1 pK unit. The maximum pK_a errors are 1.6 for Asp44 of HP36 and Asp52 of HEWL and 1.2 for Glu35 of HEWL. Possible sources of such error will be discussed further in what follows. It is worth noting that the largest deviation between calculated and experimental pK_a values of BBL is only 0.6 units.

Table 6.5: Calculated and experimental pK_a values of HP36, BBL, and HEWL

Protein	Residue	Expt[a]	Calc[b]	
			8-10 ns	0-10 ns
HP36	Asp44	3.10 (0.01)	1.2	1.5 (0.3)
	Glu45	3.95 (0.01)	3.9	4.0 (0.3)
	Asp46	3.45 (0.12)	3.8	3.3 (0.5)
	Glu72	4.37 (0.03)	4.3	4.2 (0.3)
		Avg. abs. dev.	0.6	0.5
		RMSD	1.0	0.8
			8-10 ns	0-10 ns
BBL	Asp129	3.88 (0.02)	3.5	3.3 (0.6)
	Glu141	4.46 (0.04)	4.1	4.2 (0.3)
	His142	6.47 (0.04)	6.2	6.2 (0.4)
	Asp145	3.65 (0.04)	4.0	3.9 (0.7)
	Glu161	3.72 (0.05)	4.0	4.0 (0.3)
	Asp162	3.18 (0.04)	3.0	3.0 (0.3)
	Glu164	4.50 (0.03)	4.8	4.8 (0.3)
	His166	5.39 (0.02)	4.8	4.8 (0.3)
		Avg. abs. dev.	0.3	0.3
		RMSD	0.3	0.3
			8-10 ns	0-10 ns
HEWL	Glu7	2.6 (0.2)	3.3	3.6 (0.4)
	His15	5.5 (0.2)	5.6	5.5 (0.4)
	Asp18	2.8 (0.3)	3.2	2.7 (0.7)
	Glu35	6.1 (0.4)	7.3	7.3 (0.5)
	Asp48†	1.4 (0.2)	0.4	0.08 (0.5)
	Asp52	3.6 (0.3)	5.5	5.2 (0.4)
	Asp66†	1.2 (0.2)	0.5	0.4 (0.2)
	Asp87†	2.2 (0.1)	2.2	1.7 (0.6)
	Asp101	4.5 (0.1)	4.7	4.4 (0.4)
	Asp119	3.5 (0.3)	3.8	3.4 (0.6)
		Avg. abs. dev.	0.7	0.7
		RMSD	0.9	0.9

[a] pK_a values determined by NMR titration for HP36[130], BBL[131,184], and HEWL[133]. [b] Uncertainty of the calculated values, in parentheses, are the standard deviations of the 1ns windows. † Only includes values from 1 ns windows for which calculation of pK_a values was possible (see Figure 6.12).

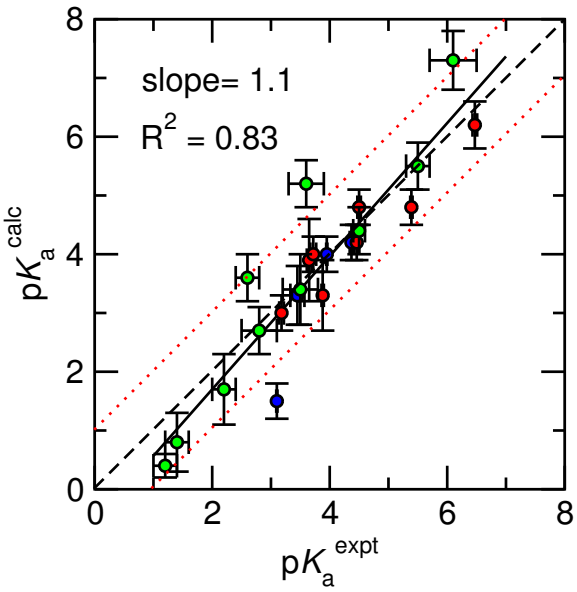

Figure 6.9: Experimental versus calculated pK_a values of proteins. Data is shown for HP36 (red), BBL (blue), and HEWL (green). Regression line (solid), y=x line (dashed black), and lines showing 1 pK unit deviation from experiment (dotted red) are shown as well as regression slope and R^2 value.

Variation of pK_a values over time and exchange efficiency

We calculated pK_a values from 1 ns windows to examine pK_a stability over the course of the simulations. These "pK_a trajectories" are depicted for HP36 in Figure 6.10, for BBL in Figure 6.11, and for HEWL in Figure 6.12. There are fluctuations of 1 pK unit around the mean values, but little drift over the course of the simulations for most of the residues. Exceptions to this are seen for Asp48 and Asp87 of HEWL where the pK_a values were initially incalculable, then after several nanoseconds the pK_a's began to steadily increase.

The exchange ratio is at a minimum at pH values where a large number of titratable groups pK_a values reside for all three proteins. Thus, it is clear that a linear spacing of replicas in pH space is suboptimal. This decrease in the number of successful exchanges is analogous to TREX results for protein folding. There is a bottle-neck in the exchange efficiency at the folding temperature, because there is an energy gap between the folded and unfolded states and they co-exist at the folding temperature.

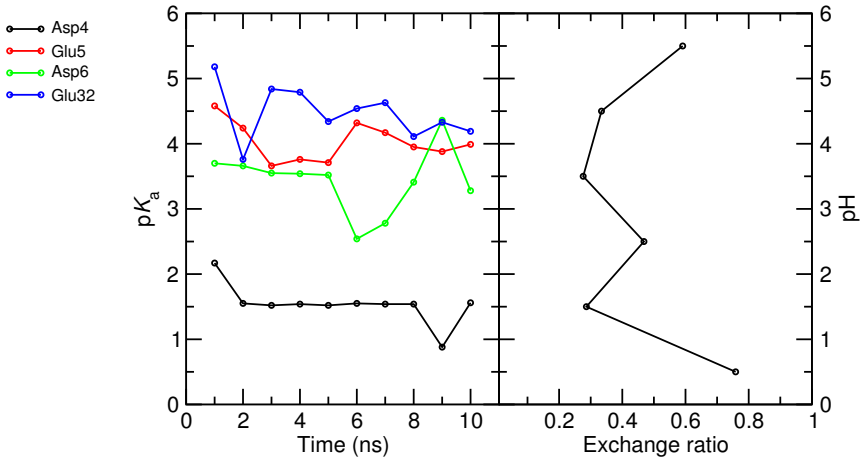

Figure 6.10: Time series of p$K_{\rm a}$ values and exchange ratio for HP36. (*Left*) p$K_{\rm a}$ values of each residue calculate from 1 ns windows taken from the 10 ns pHREX simulation. (*Right*) Exchange ratios between neighboring replicas calculated from the entire 10 ns pHREX simulation.

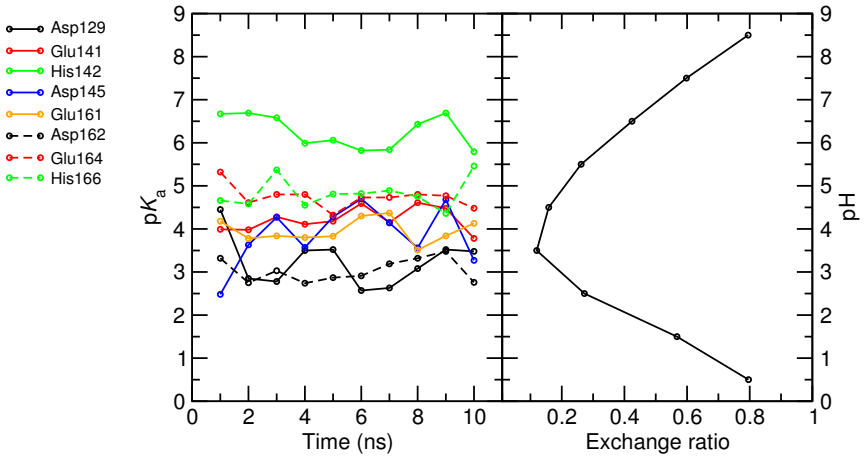

Figure 6.11: Time series of p$K_{\rm a}$ values and exchange ratio for BBL. (*Left*) p$K_{\rm a}$ values of each residue calculate from 1 ns windows taken from the 10 ns pHREX simulation. (*Right*) Exchange ratios between neighboring replicas calculated from the entire 10 ns pHREX simulation.

To overcome this limitation, a feedback loop can be applied to adjust the temperature of each replica and optimize the flow through the temperature conditions[190]. Such a scheme could be applied to pHREX as well. On the other hand, since the majority of titratable residues are on the surface of proteins and have pK_a values near the standard values, it may be sufficient to start pHREX simulations with additional replicas clustered around the model compound pK_a values. Further investigation to determine optimal pHREX parameters such as exchange-frequency and pH distribution is an important area of research that should be pursued in the future; however, it is beyond the scope of the present work.

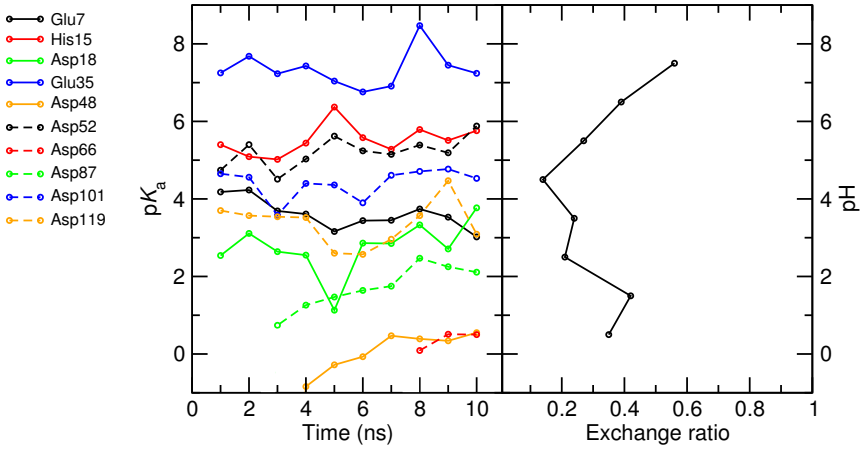

Figure 6.12: Time series of pK_a values and exchange ratio for HEWL. (*Left*) pK_a values of each residue calculate from 1 ns windows taken from the 10 ns pHREX simulation. (*Right*) Exchange ratios between neighboring replicas calculated from the entire 10 ns pHREX simulation.

pH-dependent protein conformations

In the final section, we look at the ways pH affects conformational states of the proteins. This analysis allows us to rationalize the observed pK_a shifts, identify possible sources of error, propose routes to correct these errors, and offers a window to the pH-dependent properties of the proteins.

BBL

We begin by the analyzing pH-dependent properties of BBL. The calculated pK_a values for this protein are the most accurate. The protein conformation shows little pH dependence; however, there is an increase in the solvent accessible surface area (SASA) for His166 and a decrease in the distance between His166 and nearby lysine amino groups as His166 is protonated (see Figure 6.13). The side chain is pulled out of the hydrophobic pocket and into solution due to a more favorable solvation energy as His166 becomes ionized. Also, as pH decreases, electrostatic repulsion between His166 and nearby lysine residues (Lys25, Lys40, and Lys44) favors a greater separation distance. Comparing the solvent exposure calculated here with our previous hybrid-solvent results, we see that the maximum SASA intensity from both simulations is centered around a fractional SASA of about 0.1. The pK_a shift from ECpHMD is slightly overestimated, while that from hybrid-solvent CpHMD is slightly underestimated. This is in line with our data from the benchmark GB study that GB tends to underestimate the desolvation energy for residues in a hydrophobic environment (see Chapter 2).

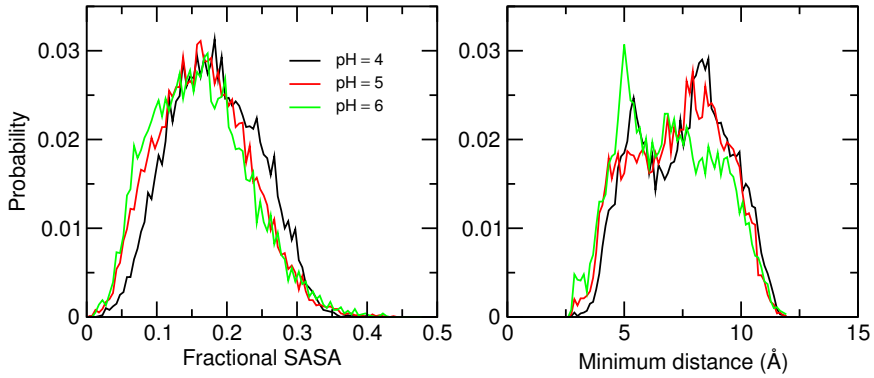

Figure 6.13: pH-dependent environment of His166 of BBL. (*Left*) Probability distribution of fractional solvent-accessible surface area (solvent accessible surface area divided by that of an isolated His residue). (*Right*) Minimum distances to the amine nitrogens of Lys25, Lys40, and Lys44 at different pH values.

HEWL

Two residues of HEWL show noticeable pK_a drift over the course of the simulation (see Figure 6.12). Asp48 was deprotonated even at pH 0 for the first 3 ns, but then

at 4 ns there was a conformational change that allowed the pK_{a} to be calculated. The pK_{a} of Asp48 continued to drift upward for the remainder of the simulation, indicating that further sampling may improve the aggreement with experiment.

In the X-ray structure, the Cγ of Asp48 is only 3.7 Å from the hydroxyl-oxygen of Ser50 and 3.3 Å from an amino-nitrogen of Arg61. We calculated the cross-correlation function between the protonation state of Asp48 and the distance from Asp48 side-chain-oxygen to both the hydroxyl-hydrogen of Ser50 and amino-hydrogens of Arg61. The cross-correlation values, at pH 0 and zero offset, between the protonation states of Asp48 and the distance from Asp48 to Ser50 and Arg61 hydrogen-bond donating hydrogens are -0.22 and -0.1, respectively. Thus, the protonation state of Asp48 is anti-correlated with the distance from Asp48 to both Arg61 and Ser50. Ser50 and Arg61 tend to act as hydrogen-bond donors when Asp48 is ionized, but the hydroxyl group of Ser50 rotates away from Asp48 and the side-chain of Arg61 moves into solution when Asp48 is protonated. Figure 6.14 shows the time series of the distance between the carboxyl-oxygen of Asp48 to the hydroxyl-hydrogen of Ser50 as well as the protonation state of Asp48 at pH 0. Although not shown, the trend in the time series of the Asp48-Arg61 distance is very similar. From this data, it is clear that hydrogen bonding between these residues and Asp48 stabilizes the ionized form of Asp48 and contributes to the downward pK_{a} shift. Figure 6.15 shows representative snapshots taken from pH 0 when Asp48 is ionized and neutral to illustrate the observed conformational change as Asp48 is protonated.

A similar effect is seen for Asp87. For the first 2 ns, Asp87 is mainly deprotonated and hydrogen bonded with Thr89. Asp87 is also initially near His15. These hydrogen-bond and electrostatic interactions favor the ionized form of Asp87, but eventually, at low pH, Asp87 becomes protonated, allowing it to move away from His15 and break the hydrogen bond with Thr89. The time series of the Asp87-Thr89 distance and protonation state of Asp87 at pH 1 are shown in Figure 6.16 and representative

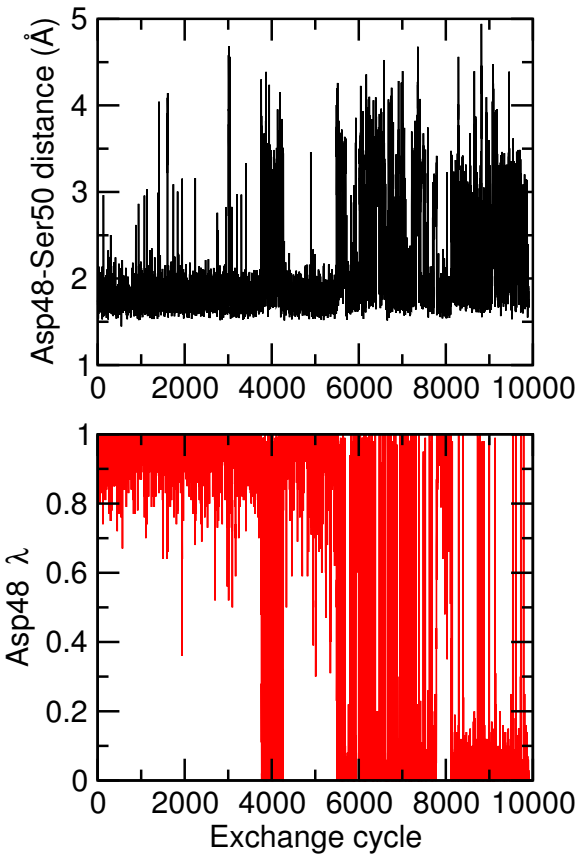

Figure 6.14: Correlation between Asp48 protonation state and Ser50 hydrogen bond. (*Upper*) Time series of minimum distance from Asp48 carboxyl-oxygen to Ser50 hydroxyl-hydrogen at pH 0. (*Lower*) Time series of Asp48 protonation state at pH 0.

snapshots when Asp87 is ionized and neutral are shown in Figure 6.17.

Asp66 of HEWL only became protonated at the lowest pH after 7 ns of simulation due to interactions with multiple residues (side chains of Arg61, Tyr53, Thr51, Ser60, Thr69, and the backbone NH group of Thr69 and Arg68) that stabilize the ionized state. Taken together, this data indicates that for side chains with strong interactions stabilizing specific side chain conformation/protonation states, pHREX suffers from inadequate sampling. This is in agreement with other's observations regarding the efficiency of pHREX; the method is very efficient for sampling protonation states at a given conformation, but the increase in conformational sampling is less dramatic[185]. After several nanoseconds, we observed transitions from the initial conformations that strongly favor deprotonation of these residues to conformations that facilitate proto-

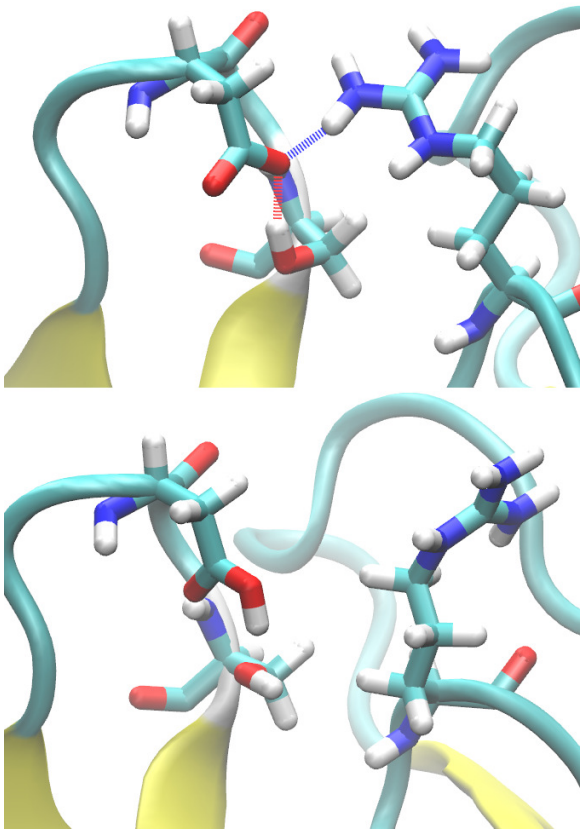

Figure 6.15: pH-dependent orientation of Asp48, Ser50, and Arg61 of HEWL. Representative snapshots of conformations of Asp48, Ser50, and Arg61 taken at pH 0. (*Upper*) Ser50 and Arg61 readily act as hydrogen bond donor forming hydrogen bond with Asp48, as is seen in the X-ray structure. (*Lower*) Protonation of Asp48 is correlated to rotation of Arg61 into solution, and rotation of Ser50 hydroxyl group. Images were rendered using the VMD program [93].

nation of Asp48 and Asp87, but substantially longer simulations would be required to accurately calculate conformational distributions and pK_a values. We suspected that the computationally more expensive TREX protocol could be used to accelerate transitions among these ionized and neutral conformational states as has been observed using GB-based CpHMD [65]. Short test-simulations using TREX at pH 0 confirmed this. Conformations in both the initial hydrogen-bonded and hydrogen-bond disrupted states for Asp48, Asp66, and Asp87 were observed within the first few exchange cycles.

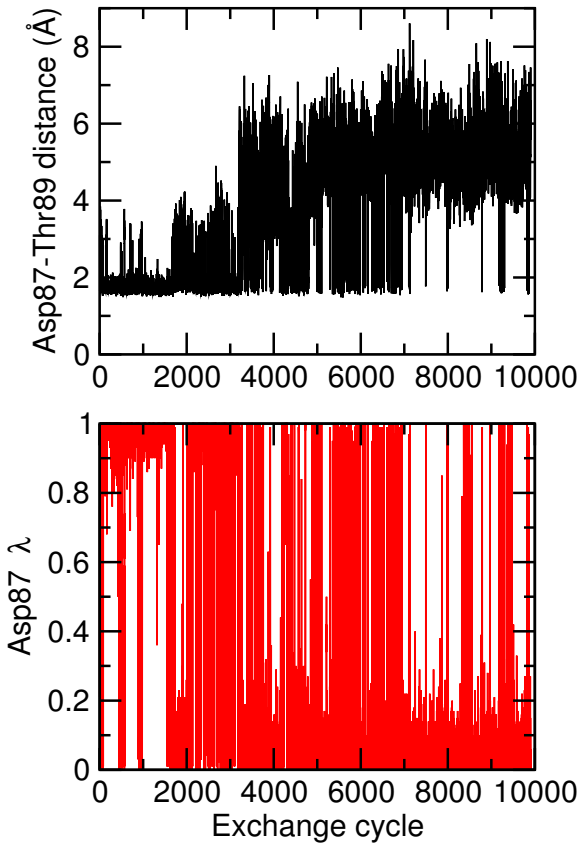

Figure 6.16: Correlation between Asp87 protonation state and Thr89 hydrogen bonding. (*Upper*) Time series of minimum distance from Asp87 carboxyl-oxygen to Thr89 hydroxyl-hydrogen at pH 1. (*Lower*) Time series of Asp87 protonation state (0 - protonated ; 1 - deprotonated) at pH 1.

The last residues of HEWL that we consider in detail are in the active site of the enzyme. The correct protonation states for Glu35 and Asp52 are critical for enzymatic activity. Optimal enzymatic activity occurs near pH 5 [24] where Glu35 is protonated and Asp52 is unprotonated. Our calculated pK_{a} values of 7.2 for Glu35 and 5.1 for Asp52 are both 1 pK unit greater than the experimental values; however, at pH 5 Glu35 is predicted to be protonated and Asp52 is predicted to be partially unprotonated, in reasonable agreement with experiment. Cross-correlation analysis of Glu35 and Asp52 protonation indicates no correlation between the titration equilibria. This is somewhat puzzling considering that the NMR titration curves (which follow the chemical shifts of the amide ^{15}N-atoms of Glu35 and Asp52) exhibit two distinct

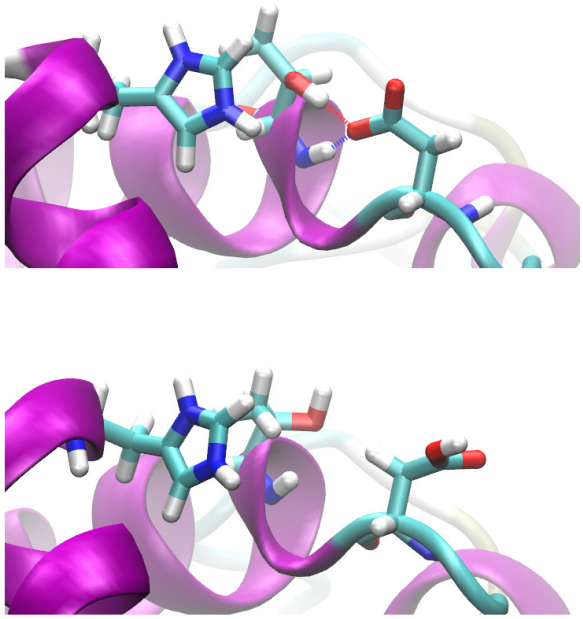

Figure 6.17: pH-dependent orientation of Asp87, Thr89, and His15 of HEWL. Representative snapshots of conformations of Asp87, Thr89, and His15 taken at pH 1. (*Upper*) Thr89 readily acts as hydrogen-bond donor and forms a hydrogen bond with Asp87, as is seen in the X-ray structure. (*Lower*) Protonation of Asp87 is correlated to rotation of Asp87 away from His15 and Thr89. This disrupts the Asp87-Thr89 hydrogen bond and increases the distance between Asp87 and His15. Images were rendered using the VMD program [93].

titration events [133]. To reconcile the experimental data with our results, we suggest that since the NMR chemical shift is much more sensitive to the local electrostatic environment than is the protonation equilibria, the complex titration curves exhibited by NMR spectroscopy do not necessarily indicate complex titration equilibria. There are many examples of non-standard titration curves in HEWL that are a result of "ghost" titrations. So-called "ghost" titrations occur when the chemical shift of an atom in one residue reports on titration of a spatially distant residue, while titration of the two is completely independent [133]. Our data suggests that this is the case for the titrations of Glu35 and Asp52.

We are then left to rationalize the sources of the pK_a shifts for Glu35 and Asp52. The solvent accessibility of these residues is stable over the course of the simulation,

both are partially buried, but Glu35 more-so than Asp52. Figure 6.18 shows the SASA for both Glu35 and Asp52 at pH 6 over the course of the simulation, as well as the probability distribution. The distribution is centered around 25 Å^2 for Asp52 and around 15 Å^2 for Glu35, in agreement with our previous hybrid-solvent CpHMD data[104]. We suggest that desolvation of these two residues is solely responsible for the positive pK_a shifts. Since Glu35 is more buried than Asp52, there is a greater desolvation penalty, and a greater upward pK_a shift. A slight overestimation of the desolvation penalty resulting from one of many possible sources (e.g. force field error, electrostatic treatment, inadequate sampling, etc.) could then explain the slight inaccuracy, which is comparable for the two residues.

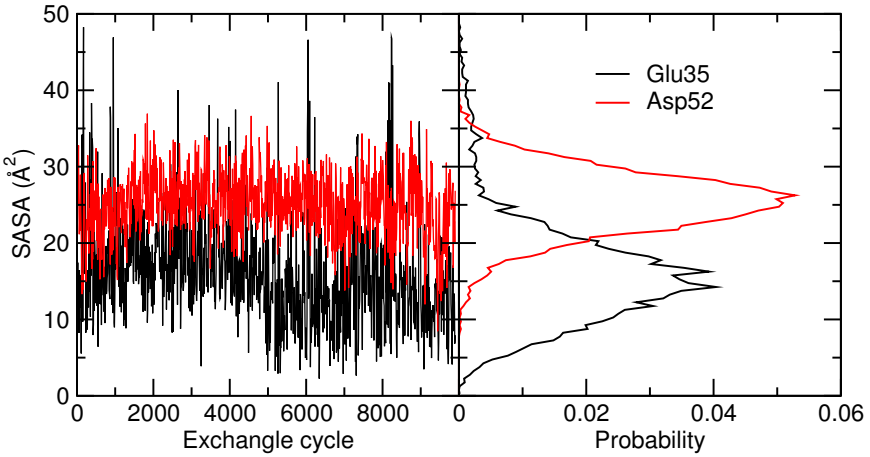

Figure 6.18: Solvent accessible surface area of Glu35 and Asp52 of HEWL. (*Left*) Time series of SASA of Glu35 and Asp52 and (*right*) probability distribution of SASA for these residues at pH 6.

HP36

HP36, the smallest protein we studied, has the most interesting story to tell concerning pH-dependent conformational states, coupled titration, and pK_a calculation error. To familiarize the reader with HP36, we show the NMR model of the protein in Figure 6.19. HP36 is a 36 residue mini-protein that folds on the microsecond timescale[191] and is composed of three α-helices. We denote these helices as I to III from the N- to C-termini. There are three acidic side chains (Asp44, Glu45, and

Asp46) on helix I. There is another acidic side chain (Glu72) on helix III and several basic residues. There is a small hydrophobic core formed by three phenylalanine residues of helix I and II.

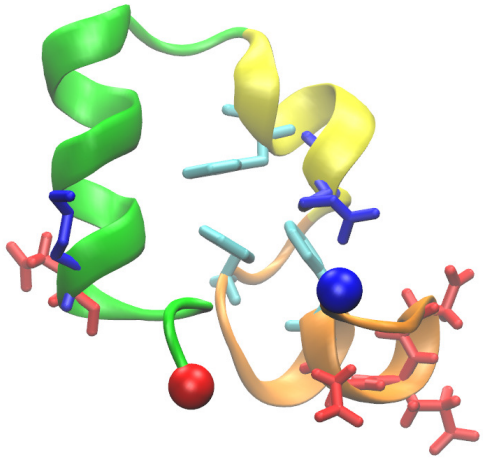

Figure 6.19: NMR structure of HP36. NMR structure of HP36, in cartoon representation, showing acidic residues (red) as well as Arg15 and Lys31 (blue). The Cα atoms of N-(blue) and C-terminal(red) residues are shown as spheres for reference. The three helices are depicted in different colors and the phenylalanine hydrophobic core is in cyan. Image was rendered using the VMD program[93].

The residue of HP36 with the largest experimental pK_{a} shift is Asp44. In the NMR model, the Cγ of Asp44 is 7.9 Å from Cζ of Arg55; however, in our simulations slight repositioning of the helices and side chain rearrangement quickly (within a few nanoseconds) reduces this value so that they form a salt bridge. The experimental pK_{a} of Asp44 is downshifted by one unit relative to the standard value, but in our simulations the down shift is overestimated by 1.5 units. There is a clear pH-dependence of the salt-bridge stability. Below the calculated pK_{a} of Asp44 the salt bridge is broken, but above it the salt bridge is stable as illustrated in Figure 6.20. Overestimation of the pK_{a} shift suggests two possible sources of error: either the force field (or treatment of electrostatics via GRF) overestimates the strength of the salt bridge or length of the simulation (much shorter than the lifetime of the salt bridge) causes apparent over-stabilization. If the latter is the case, this error could be corrected by

running longer simulations or utilizing more effective sampling techniques. Regarding the electrostatics treatment, comparison of salt-bridge strength using GRF and PME indicates that the strength of a salt bridge is slightly weaker using GRF[192]. This suggests that overestimation of the salt-bridge strength is not an artifact of GRF electrostatics. Whether this is a result of force field bias or inadequate sampling is a more difficult issue. Simulations run at constant protonation state have indicated that the salt-bridge lifetime in MD simulation may be on the order of hundreds of nanoseconds[193]. Although, pHREX has been shown to improve sampling when combined with hybrid-solvent CpHMD[104], to accurately sample both salt-bridge formed and broken states in ECpHMD, it appears that substantially longer simulations than conducted here may be required. Another method to accelerate sampling is to run TREX simulations at each pH value. When we conducted TREX for 2 ns per replica at pH 0, 2, 4 and 6, the calculated pK_{a} value of Asp44 increased by 0.5 pK units to 2.0. This suggests that TREX provides accelerates interconversion between conformational states and this increase in conformational rearrangement can result in more accurate pK_{a} values. However, the length of these test simulations were admittedly not long enough to confirm that the use of TREX will lead to better pK_{a} prediction overall.

In the titration of HP36, we observed correlated protonation events for Glu45 and Asp46. In the NMR structure, the Cγ of Asp46 is 7.7 Å from Cδ of Glu45. The normalized cross-correlation between the two residues protonation states at a delay time of zero is -0.2 at pH 4 and -0.1 at pH 3. The experimental (NMR derived) titration data also suggests coupled protonation equilibria. The experimental Hill value is 0.9 for Glu45 and 1.1 for Asp46[130]. Our calculated Hill value for Asp46 is 1.4±0.5 indicating cooperative proton binding, as was found in experiment. We also observed very weak correlation between Asp44 and Glu45 titration, although further comparison to experiment is not appropriate due to the inaccuracy in the calculated

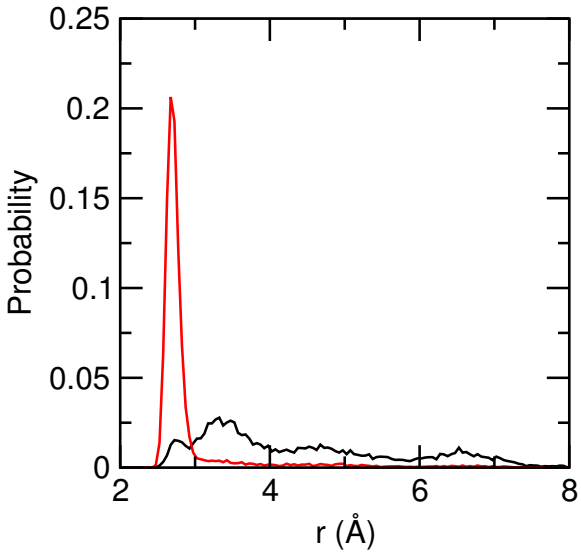

Figure 6.20: Probability distributions of Asp45-Arg55 distances in HP36. Probability distributions of the minimum distance from Asp45 carboxyl-oxygen to Arg55 side chain nitrogen at pH 0 (black) and pH 2 (red).

pK_{a} value of Asp44.

Lastly, we observed partial unraveling of helix III of HP36 at elevated pH. We calculated the PMF (shown in Figure 6.21) at pH 2 and 4 as a function of segment A (helix I and II) RMSD and segment B (helix II and III) RMSD to quantify the extent of the structural distortion. There is only one attractive basin centered at segment A RMSD of 2.2 and segment B RMSD of 2.5 at pH 2, while at pH 4 an additional basin is observed at segment B RMSD of 5, indicating a conformational change of helix III.

We further analyzed the helicity per residue and the overall helical content of the conformations extracted from the two attractive basins observed at pH 4 (see Figure 6.22). We see a reduction in the overall helical content in the non-native basin. The helicity per residue indicates that the end of helix III looses secondary structure beginning at residue 30.

Figure 6.23 shows the probability distribution of the distances between the carboxyl-oxygens of Glu32 and amine-nitrogen of Lys31 for conformations extracted from the

two basins in Figure 6.21. Looking more closely at the conformations extracted from the two basins in Figure 6.21, in the non-native basin a partial unraveling of helix III at the C-terminal allows Glu72 and Lys71 to come very close to one another and form a salt bridge. This salt bridge is not observed in conformations which retain native secondary structure. This indicates that a non-native salt bridge formed between Glu32 and Lys31, which is only possible in the partially unfolded state, stabilizes these non-native conformations.

Chemical denaturation (urea) data at pH 5.0 shows that the native state of the mutant K71M is slightly more stable than the wild-type protein[130]. Our data can be used to rationalize this experimental finding. At a pH above the pK_{a} of Glu72, when Glu72 is ionized, this residue participates in a non-native favorable interaction with Lys71 when HP36 is partially unfolded. Removing the basic residue involved in this non-native interaction stabilizes the folded state. Although the purpose of our simulations is not to exam the unfolded state, but rather to calculate native-state pK_{a} values, it is encouraging that from only 10 ns of sampling we begin to gather information regarding initial stages of the unfolding process.

In agreement with experimental stability measurements of the wild-type protein[130], our data indicates that the folded state becomes less stable at elevated pH and suggests that helix III is the least stable of the three helices. Experimental evidence indicates that there is significant structure in the unfolded state of HP36[195]. A truncated peptide containing only residues of helix I and II displays characteristics similar to the unfolded state[196] of the full length protein suggesting that the unfolded state of HP36 represents a loss of structure in helix III. Further, a triplet-triplet energy transfer study[197] corroborated the idea that the hydrophobic core formed by phenylalanines of helix I and II remains intact in the unfolded state, while helix III unravels. Several computational studies of HP36 folding have been undertaken, but the results are mixed and it is difficult to discern whether these observations

are governed by kinetics or thermodynamics. Simulations using an implicit-solvent model suggested helix III forms first[198,199] while explicit-solvent simulations of the truncated individual helices indicated that, in isolation, helix I is the most stable[200].

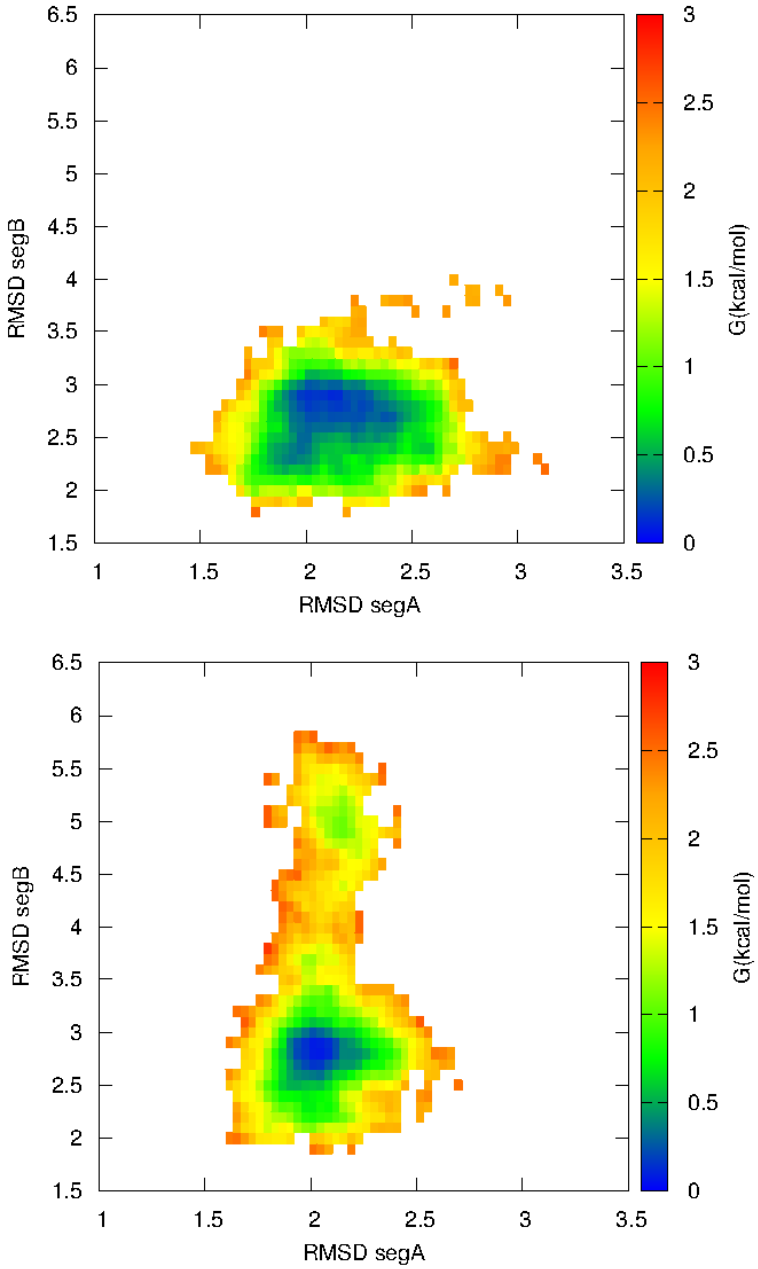

Figure 6.21: Free-energy surfaces of HP36. Two-dimensional free-energy surfaces of HP36 at pH 2 (*Upper*) and pH 4 (*Lower*). The first axis is RMSD of segment A (helix I and II) while the second axis is RMSD of segment B (helix II and III). The first nanosecond of the simulation was discarded for this and all subsequent analysis.

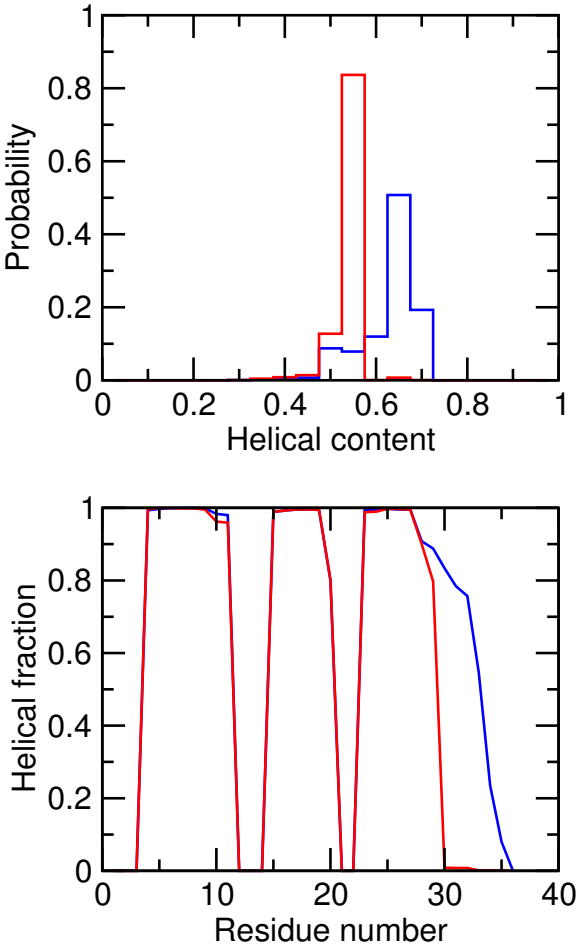

Figure 6.22: Secondary structure of HP36. (*Upper*) Probability density of the helical content (the fraction of residues that are part of an α-helix) and (*Lower*) the helical fraction (the fraction of time a specific residue is part of an α-helix) from structures extracted from native (blue) and non-native basins (red) in Figure 6.21. Secondary structure was calculated using the STRIDE program[194].

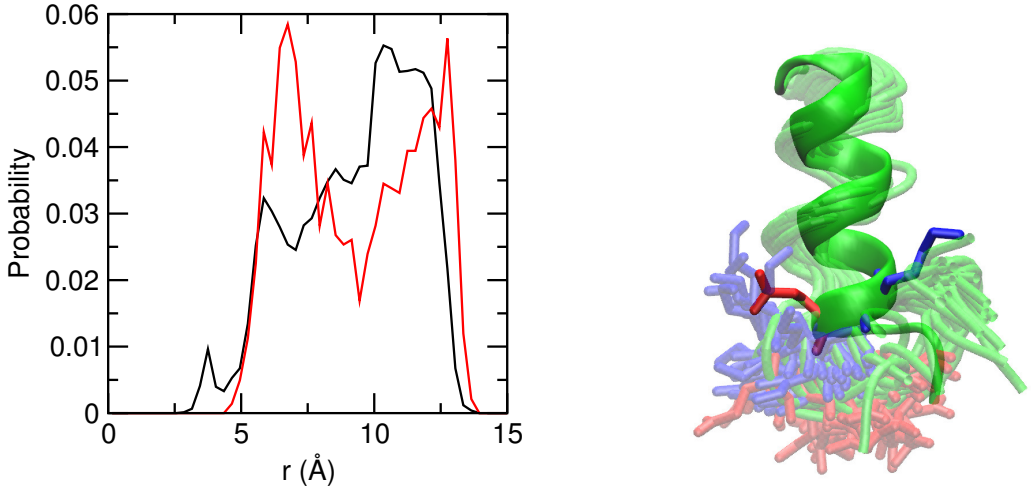

Figure 6.23: Distributions of HP36 Lys71 to Glu72 distances from native and non-native basins. (*Left*) Probability distribution of HP36 Lys71 amine-nitrogen to Glu72 carboxyl-oxygen minimum distance for conformations at pH 4 with segment B RMSD > 4 (black) and segment B RMSD < 4 (red). (*Right*) NMR model of helix *III* as cartoon showing Lys71 (blue) and Glu72 (red) with every $50th$ snapshot having segment B RMSD > 4 oriented and superimposed onto the native helix and rendered as transparent image.

6.5 Conclusion

We propose a variant of ECpHMD that makes use of the GRF treatment of long-range electrostatics and a charge-leveling procedure that keeps the net charge neutral by coupling titration to the charging and neutralization of co-ions which act as a charge reservoir. We tested our method on two series of molecules: aliphatic dicarboxylic acids and proteins. Our results from the dicarboxylic acid series and the small protein HP36 indicate that charge neutralization is necessary in order to obtain accurate pK_{a} values. The method we propose yields results in good agreement with experiment for cases where conformational sampling is adequate such as dicarboxylic acids and the protein BBL. We find that the inclusion of additional ions to match experiment can improve pK_{a} calculation accuracy for proteins.

Concerning overall accuracy, pHREX-ECpHMD performs on-par with hybrid-solvent CpHMD. However, in certain cases where there are larger energy barriers that separate conformations favored by different protonation states, as is the case of salt bridges and hydrogen bonding, pHREX-ECpHMD suffer from inadequate sampling. Given this limitation, ECpHMD can be combined with the TREX protocol providing a straightforward route to increase conformational sampling and pK_{a} accuracy.

This work represents, to the best of our knowledge, the most rigorous test of the accuracy of ECpHMD and paves the way for further development and application of the method.

Bibliography

(1) Rivas, G.; Ferrone, F.; Herzfeld, J. *EMBO Rep.* **2004**, *5*, 23–227.

(2) van den Berg, B.; Ellis, R. J.; Dobson, C. M. *EMBO J.* **1999**, *18*, 6927–6933.

(3) van den Berg, B.; Wain, R.; Dobson, C. M.; Ellis, R. J. *EMBO J.* **2000**, *19*, 3870–3875.

(4) Homouz, D.; Perham, M.; Samiotakis, A.; Cheung, M. S.; Wittung-Stafshede, P. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 11754–11759.

(5) Sasahara, K.; McPhie, P.; Minton, A. P. *J. Mol. Biol.* **2003**, *326*, 1227–1237.

(6) Minton, A. P. *Biophys. J.* **2005**, *88*, 971–985.

(7) Attri, P.; Venkatesu, P.; Lee, M.-J. *J. Phys. Chem. B* **2010**, *114*, 1471–1478.

(8) Stigter, D.; Alonso, D. O.; Dill, K. A. *Proc. Natl. Acad. Sci. USA* **1991**, *88*, 4176–4180.

(9) Garcia-Moreno, B. *J. Biol.* **2009**, *8*, 98.

(10) Matlin, K. S.; Reggio, H.; Helenius, A.; Simons, K. *J. Cell Biol.* **1981**, *91*, 601–613.

(11) Yoshimura, A.; Kuroda, K.; Kawasaki, K.; Yamashina, S.; Maeda, T.; ichi Ohnishi, S. *J. Virol.* **1982**, *43*, 284–293.

(12) Dawson, J. E.; Šečkutė, J.; De, S.; Schueler, S. A.; Oswald, A. B.; Nicholson, L. K. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 8543–8548.

(13) Vollrath, F.; Knight, D. P.; Hu, X. W. *Proc. R. Soc. Lond. B* **1998**, *265*, 817–820.

(14) Askerieh, G.; Hedhammar, M.; Nordling, K.; Saenz, A.; Casals, C.; Rising, A.; Johansson, J.; Knight, S. D. *Nature* **2010**, *465*, 236–238.

(15) Bierzynski, A.; Kim, P. S.; Baldwin, R. L. *Proc. Natl. Acad. Sci. USA* **1982**, *79*, 2470–2474.

(16) Sheinerman, F. B.; Norel, R.; Honig, B. *Curr. Opin. Struct. Biol.* **2000**, *10*, 153–159.

(17) Warshel, A. *Acc. Chem. Res.* **1981**, *14*, 284–290.

(18) Wyman, Jr., J. *Adv. Protein Chem.* **1964**, *19*, 223–286.

(19) Di Cera, E.; Gill, S. J.; Wyman, J. *Proc. Natl. Acad. Sci. USA* **1988**, *85*, 5077–5081.

(20) Mason, A. C.; Jensen, J. H. *Proteins* **2008**, *71*, 81–91.

(21) Tanford, C. *Adv. Protein Chem.* **1970**, *24*, 1–95.

(22) Warshel, A. *Biochemistry* **1981**, *20*, 3167–3177.

(23) Fersht, A. R. *J. Mol. Biol.* **1972**, *64*, 497–509.

(24) Bartik, K.; Redfield, C.; Dobson, C. M. *Biophys. J.* **1994**, *66*, 1180–1184.

(25) Chivers, P. T.; Prehoda, K. E.; Volkman, B. F.; Kim, B.-M.; Markley, J. L.; Raines, R. T. *Biochemistry* **1997**, *36*, 14985–14991.

(26) Dao-pin, S.; Sauer, U.; Nicholson, H.; Matthews, B. W. *Biochemistry* **1991**, *30*, 7142–7153.

(27) Pérez-Cañadillas, J. M.; Campos-Olivas, R.; Lacadena, J.; del Pozo, A. M.; Gavilanes, J. G.; Santoro, J.; Rico, M.; Bruix, M. *Biochemistry* **1998**, *37*, 15865–15876.

(28) Lide, D. R., Ed. *CRC handbook of chemistry and physics*, 92nd ed.; CRC Press LLC: Boca Raton,FL, 2012.

(29) McCammon, J. A.; Gelin, B. R.; Karplus, M. *Nature* **1977**, *267*, 585–590.

(30) Ponder, J. W.; Case, D. A. *Adv. Protein Chem.* **2003**, *66*, 27–85.

(31) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

(32) Jorgensen, W. L.; McDonald, N. A. *J. Mol. Struct. (Theochem)* **1998**, *424*, 145–155.

(33) Halgren, T. A. *J. Comput. Chem.* **1996**, *17*, 520–552.

(34) Oostenbrink, C.; Villa, A.; Mark, A. E.; van Gunsteren, W. F. *J. Comput. Chem.* **2004**, *25*, 1656–1676.

(35) MacKerell Jr., A. D. et al. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

(36) Mackerell, Jr., A. D.; Feig, M.; Brooks, III, C. L. *J. Comput. Chem.* **2004**, *25*, 1400–1415.

(37) Buck, M.; Bouguet-Bonnet, S.; Pastor, R. W.; MacKerell Jr., A. D. *Biophys. J.* **2006**, *90*, L36–L38.

(38) Jones, J. E. *Proc. R. Soc. Lond. A* **1924**, *106*, 463–447.

(39) Brooks, B. R. et al. *J. Comput. Chem.* **2009**, *30*, 1545–1614.

(40) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.

(41) Tironi, I. G.; Sperb, R.; Smith, P. E.; van Gunsteren, W. F. *J. Chem. Phys.* **1995**, *102*, 5451–5459.

(42) Spoel, D. V. D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005**, *29*, 1701–1718.

(43) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.

(44) Schaefer, M.; Karplus, M. *J. Phys. Chem.* **1996**, *100*, 1578–1600.

(45) Karplus, M.; Petsko, G. A. *Nature* **1990**, *347*, 631–639.

(46) Michel, J.; Foloppe, N.; Essex, J. W. *Mol. Inf.* **2010**, *26*, 570–578.

(47) Paschek, D.; Hempel, S.; García, A. E. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 17754–17759.

(48) Best, R. B.; Mittal, J. *Proteins* **2011**, *79*, 1318–1328.

(49) Voelz, V. A.; Bowman, G. R.; Beauchamp, K.; Pande, V. S. *J. Am. Chem. Soc.* **2010**, *132*, 1526–1528.

(50) Mertz, J. E.; Pettitt, B. M. *Int. J. Supercomput. Appl. High Perform. Comput.* **1994**, *8*, 47–53.

(51) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H. *J. Chem. Phys.* **1953**, *21*, 1087–1092.

(52) Baptista, A. M.; Teixeira, V. H.; Soares, C. M. *J. Chem. Phys.* **2002**, *117*, 4184–4200.

(53) Machuqueiro, M.; Baptista, A. M. *J. Phys. Chem. B* **2006**, *110*, 2927–2933.

(54) Machuqueiro, M.; Baptista, A. M. *Proteins* **2008**, *72*, 289–298.

(55) Machuqueiro, M.; Baptista, A. M. *Proteins* **2011**, *79*, 3437–3447.

(56) Bürgi, R.; Kollman, P. A.; van Gunsteren, W. F. *Proteins* **2002**, *47*, 469–480.

(57) Beveridge, D. L.; DiCapua, F. M. *Annu. Rev. Biophys. Biophys. Chem.* **1989**, *18*, 431–492.

(58) Schaefer, M.; Sommer, M.; Karplus, M. *J. Phys. Chem. B* **1997**, *101*, 1663–1683.

(59) Dlugosz, M.; Antosiewicz, J. M. *Chem. Phys.* **2004**, *302*, 161–170.

(60) Mongan, J.; Case, D. A.; McCammon, J. A. *J. Comput. Chem.* **2004**, *25*, 2038–2048.

(61) Meng, Y.; Roitberg, A. E. *J. Chem. Theory Comput.* **2010**, *6*, 1401–1412.

(62) Stern, H. A. *J. Chem. Phys.* **2007**, *126*, 164112.

(63) Börjesson, U.; Hünenberger, P. H. *J. Chem. Phys.* **2001**, *114*, 9706–9719.

(64) Lee, M. S.; Salsbury, Jr., F. R.; Brooks III, C. L. *Proteins* **2004**, *56*, 738–752.

(65) Khandogin, J.; Brooks III, C. L. *Biochemistry* **2006**, *45*, 9363–9373.

(66) Khandogin, J.; Brooks III, C. L. *Biophys. J.* **2005**, *89*, 141–157.

(67) Khandogin, J.; Chen, J.; Brooks III, C. L. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 18546–18550.

(68) Khandogin, J.; Brooks III, C. L. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 16880–16885.

(69) Donnini, S.; Tegeler, F.; Groenhof, G.; Grubmüller, H. *J. Chem. Theory Comput.* **2011**, *7*, 1962–1978.

(70) Goh, G. B.; Knight, J. L.; III, C. L. B. *J. Chem. Theory Comput.* **2012**, *8*, 36–46.

(71) Bashford, D.; Case, D. A. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129–152.

(72) Girvin, M. E.; Rastogi, V. K.; Abildgaard, F.; Markley, J. L.; Fillingame, R. H. *Biochemistry* **1998**, *37*, 8817–8824.

(73) Pielak, R. M.; Chou, J. J. *Protein Cell* **2010**, *1*, 246–258.

(74) Chen, J.; Brooks III, C. L.; Khandogin, J. *Curr. Opin. Struct. Biol.* **2008**, *18*, 140–148.

(75) Khandogin, J. In *Multi-scale quantum models for biocatalysis*; York, D. M., Lee, T.-S., Eds.; Springer: New York, 2009; Chapter 10. Modeling protonation equilibria in biological macromolecules, pp 261–284.

(76) Schutz, C. N.; Warshel, A. *Proteins* **2001**, *44*, 400–417.

(77) Kong, X.; Brooks, III, C. L. *J. Chem. Phys.* **1996**, *105*, 2414–2423.

(78) Lee, M. S.; Feig, M.; Salsbury, Jr., F. R.; Brooks, III, C. L. *J. Comput. Chem.* **2003**, *24*, 1348–1356.

(79) Im, W.; Lee, M. S.; Brooks III, C. L. *J. Comput. Chem.* **2003**, *24*, 1691–1702.

(80) Hansmann, U. H. *Chem. Phys. Lett.* **1997**, *281*, 140–150.

(81) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141–151.

(82) Wallace, J. A.; Shen, J. K. *Methods Enzymol.* **2009**, *466*, 455–475.

(83) Shen, J. K. *Biophys. J.* **2010**, *99*, 924–932.

(84) Bashford, D.; Karplus, M. *Biochemistry* **1990**, *29*, 10219–10225.

(85) Isom, D. G.; Cannon, B. R.; Castañeda, C. A.; Robinson, A.; García-Moreno E., B. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 17784–17788.

(86) Isom, D. G.; Castañeda, C. A.; Cannon, B. R.; Velu, P. D.; García-Moreno E., B. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 16096–16100.

(87) Hynes, T. R.; Fox, R. O. *Proteins* **1991**, *10*, 92–105.

(88) Chen, J.; Lu, Z.; Sakon, J.; Stites, W. E. *J. Mol. Biol.* **2000**, *303*, 125–130.

(89) Castañeda, C. A.; Fitch, C. A.; Majumdar, A.; Khangulov, V.; Schlessman, J. L.; García-Moreno E., B. *Proteins* **2009**, *77*, 570–588.

(90) Schaftenaar, G.; Noordik, J. H. *J. Comput.-Aided Mol. Design* **2000**, *14*, 123–134.

(91) Feig, M.; Karanicolas, J.; Brooks III, C. L. *J. Mol. Graph. Model.* **2004**, *22*, 377–395.

(92) Nozaki, Y.; Tanford, C. *Methods Enzymol.* **1967**, *11*, 715–734.

(93) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Grapics* **1996**, *14*, 33–38.

(94) Sham, Y. Y.; Muegge, I.; Warshel, A. *Biophys. J.* **1998**, *74*, 1744–1753.

(95) Georgescu, R. E.; Alexov, E. G.; Gunner, M. R. *Biophys. J.* **2002**, *83*, 1731–1748.

(96) Geney, R.; Layten, M.; Gomperts, R.; Hornak, V.; Simmerling, C. *J. Chem. Theory Comput.* **2006**, *2*, 115–127.

(97) Wallace, J. A.; Shen, J. K. *Biochemistry* **2010**, *49*, 5290–5298.

(98) Chen, J.; Im, W.; Brooks III, C. L. *J. Am. Chem. Soc.* **2006**, *128*, 3728–3736.

(99) Lee, M. S.; Salsbury, Jr., F. R.; Brooks, III, C. L. *J. Chem. Phys.* **2002**, *116*, 10606–10614.

(100) Onufriev, A.; Case, D. A.; Bashford, D. *J. Comput. Chem.* **2002**, *23*, 1297–1304.

(101) Wang, Y.; Wallace, J. A.; Koenig, P. H.; Shen, J. K. *J. Comput. Chem.* **2011**, *32*, 2348–2358.

(102) Chen, J.; Brooks III, C. L. *Phys. Chem. Chem. Phys.* **2008**, *10*, 471–481.

(103) Gallicchio, E.; Paris, K.; Levy, R. M. *J. Chem. Theory Comput.* **2009**, *5*, 2544–2564.

(104) Wallace, J. A.; Shen, J. K. *J. Chem. Theory Comput.* **2011**, *7*, 2617–2629.

(105) Fukunishi, H.; Watanabe, O.; Takada, S. *J. Chem. Phys.* **2002**, *116*, 9058–9067.

(106) Loncharich, R. J.; Brooks, B. R.; Pastor, R. W. *Biopolymers* **1992**, *32*, 523–535.

(107) Okamoto, Y. *J. Mol. Graph. Model.* **2004**, *22*, 425–439.

(108) Mu, Y. *J. Chem. Phys.* **2009**, *130*, 164107.

(109) Jiang, W.; Hodoscek, M.; Roux, B. *J. Chem. Theory Comput.* **2009**, *5*, 2583–2588.

(110) Jiang, W.; Roux, B. *J. Chem. Theory Comput.* **2010**, *6*, 2559–2565.

(111) Mu, Y.; Xu, W. *J. Chem. Phys.* **2007**, *127*, 084119.

(112) Kannan, S.; Zacharias, M. *Proteins* **2007**, *66*, 697–706.

(113) Kannan, S.; Zacharias, M. *Proteins* **2010**, *78*, 2809–2819.

(114) Chandrasekhar, S. *Rev. Mod. Phys.* **1943**, *15*, 1–89.

(115) Grossfield, A.; Zuckerman, D. M. *Annu. Report Comput. Chem.* **2009**, *5*, 23–48.

(116) Flyvbjerg, H.; Petersen, H. G. *J. Chem. Phys.* **1989**, *91*, 461–466.

(117) Nina, M.; Beglov, D.; Roux, B. *J. Phys. Chem. B* **1997**, *101*, 5239–5248.

(118) Nielsen, J. E.; Mccammon, J. A. *Protein Sci.* **2003**, *12*, 1894–1901.

(119) Shen, J. K. *J. Am. Chem. Soc.* **2010**, *132*, 7258–7259.

(120) Khandogin, J.; Raleigh, D. P.; Brooks III, C. L. *J. Am. Chem. Soc.* **2007**, *129*, 3056–3057.

(121) Okur, A.; Wickstrom, L.; Simmerling, C. *J. Chem. Theory. Comput.* **2008**, *4*, 488–498.

(122) Voelz, V. A.; Singh, V. R.; Wedemeyer, W. J.; Lapidus, L. J.; Pande, V. S. *J. Am. Chem. Soc.* **2010**, *132*, 4702–4709.

(123) Roe, D. R.; Okur, A.; Wickstrom, L.; Hornak, V.; Simmerling, C. *J. Phys. Chem. B* **2007**, *111*, 1846–1857.

(124) Wallace, J. A.; Wang, Y.; Shi, C.; Pastoor, K. J.; Nguyen, B.-L.; Xia, K.; Shen, J. K. *Proteins* **2011**, *79*, 3364–3373.

(125) Börjesson, U.; Hünenberger, P. H. *J. Phys. Chem. B* **2004**, *108*, 13551–13559.

(126) Nadler, W.; Hansmann, U. H. E. *Phys. Rev. E* **2007**, *75*, 026109.

(127) Nadler, W.; Hansmann, U. H. E. *J. Phys. Chem. B* **2008**, *112*, 10386–10387.

(128) Okur, A.; Wickstrom, L.; Layten, M.; Geney, R.; Song, K.; Hornak, V.; Simmerling, C. *J. Chem. Theory Comput.* **2006**, *2*, 420–433.

(129) Bashford, D.; Case, D. A.; Dalvit, C.; Tennant, L.; Wright, P. E. *Biochemistry* **1993**, *32*, 8045–8056.

(130) Bi, Y. Studies of the folding and stability of the villin headpiece subdomain. Ph.D. thesis, Stony Brook University, 2008.

(131) Arbely, E.; Rutherford, T. J.; Sharpe, T. D.; Ferguson, N.; Fersht, A. R. *J. Mol. Biol.* **2009**, *387*, 986–992.

(132) Kuhlman, B.; Luisi, D. L.; Young, P.; Raleigh, D. P. *Biochemistry* **1999**, *38*, 4896–4903.

(133) Webb, H.; Tynan-Connolly, B. M.; Lee, G. M.; Farrell, D.; O'Meara, F.; Søndergaard, C. R.; Teilum, K.; Hewage, C.; McIntosh, L. P.; Nielsen, J. E. *Proteins* **2011**, *79*, 685–702.

(134) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.

(135) Bogusz, S.; Cheatham III, T. E.; Brooks, B. R. *J. Chem. Phys.* **1998**, *108*, 7070–7084.

(136) Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695–1697.

(137) Feller, S. E.; Zhang, Y.; Pastor, R. W.; Brooks, B. R. *J. Chem. Phys.* **1995**, *103*, 4613–4621.

(138) Hummer, G.; Pratt, L. R.; García, A. E. *J. Phys. Chem.* **1996**, *100*, 1206–1215.

(139) Thurlkill, R. L.; Grimsley, G. R.; Scholtz, J. M.; Pace, C. N. *Protein Sci.* **2006**, *15*, 1214–1218.

(140) Mongan, J.; Case, D. A. *Curr. Opin. Struct. Biol.* **2005**, *15*, 157–163.

(141) Denisov, V. P.; Schlessman, J. L.; García-Moreno E., B.; Halle, B. *Biophys. J.* **2004**, *87*, 3982–3994.

(142) Damjanović, A.; García-Moreno E., B.; Lattman, E. E.; García, A. E. *Proteins* **2005**, *60*, 433–449.

(143) Gosline, J. M.; Guerette, P. A.; Ortlepp, C. S.; Savage, K. N. *J. Exp. Biol.* **1999**, *202*, 3295–3303.

(144) Spiess, K.; Lammel, A.; Scheibel, T. *Macromol. Biosci.* **2010**, *10*, 998–1007.

(145) Altman, G. H.; Diaz, F.; Jakuba, C.; Calabro, T.; Horan, R. L.; Chen, J.; Lu, H.; Richmond, J.; Kaplan, D. L. *Biomaterials* **2003**, *24*, 401–416.

(146) Zhou, S.; Peng, H.; Yu, X.; Zheng, X.; Cui, W.; Zhang, Z.; Li, X.; Wang, J.; Weng, J.; Jia, W.; Li, F. *J. Phys. Chem. B* **2008**, *112*, 11209–11216.

(147) Omenetto, F. G.; Kaplan, D. L. *Science* **2010**, *329*, 528–531.

(148) Heim, M.; Keerl, D.; Scheibel, T. *Angew. Chem. Int. Ed.* **2009**, *48*, 3584–3596.

(149) Gaines, W. A.; Sehorn, M. G.; Marcotte Jr., W. R. *J. Biol. Chem.* **2010**, *285*, 40745–40753.

(150) Dicko, C.; Vollrath, F.; Kenney, J. M. *Biomacromolecules* **2004**, *5*, 704–710.

(151) Dicko, C.; Kenney, J. M.; Knight, D.; Vollrath, F. *Biochemistry* **2004**, *43*, 14080–14087.

(152) Matsumoto, A.; Chen, J.; Collette, A. L.; Kim, U.-J.; Altman, G. H.; Cebe, P.; Kaplan, D. L. *J. Phys. Chem. B* **2006**, *110*, 21630–21638.

(153) Hedhammar, M.; Rising, A.; Grip, S.; Martinez, A. S.; Nordling, K.; Casals, C.; Stark, M.; Johansson, J. *Biochemisty* **2008**, *47*, 3407–3417.

(154) Landreh, M.; Askarieh, G.; Nordling, K.; Hedhammar, M.; Rising, A.; Casals, C.; Astorga-Wells, J.; Alvelius, G.; Knight, S. D.; Johansson, J.; Jörnvall, H.; Bergman, T. *J. Mol. Biol.* **2010**, *404*, 328–336.

(155) Hagn, F.; Thamm, C.; Scheibel, T.; Kessler, H. *Angew. Chem. Int. Ed.* **2011**, *45*, 3795–3800.

(156) Šali, A.; Blundell, T. L. *J. Mol. Biol.* **1993**, *234*, 779–815.

(157) Efron, B.; Tibshirani, R. *Stat. Sci.* **1986**, *1*, 54–75.

(158) Bose, K.; Clark, A. C. *Protein Sci.* **2005**, *14*, 24–36.

(159) Xavier, K. A.; Willson, R. C. *Biophys. J.* **1998**, *74*, 2036–2045.

(160) Otosu, T.; Nishimoto, E.; Yamashita, S. *J. Biochem.* **2010**, *147*, 191–200.

(161) Teles, R. C.; de A. Calderon, L.; Medrano, F. J.; Barbosa, J. A.; Guimarães, B. G.; Santoro, M. M.; ; de Freitas, S. M. *Biophys. J.* **2005**, *88*, 3509–3517.

(162) Cho, J.-H.; Raleigh, D. P. *J. Mol. Biol.* **2005**, *353*, 174–185.

(163) Murtaugh, M. L.; Fanning, S. W.; Sharma, T. M.; Terry, A. M.; Horn, J. R. *Protein Sci.* **2011**, *20*, 1619–1631.

(164) Shim, J. H.; Benkovic, S. J. *Biochemistry* **1999**, *38*, 10024–10031.

(165) Schiaretti, F.; Bettati, S.; Viappiani, C.; Mozzarelli, A. *J. Biol. Chem.* **2004**, *279*, 29572–29582.

(166) Cottrell, J. W.; Scott, L. G.; Fedor, M. J. *J. Biol. Chem.* **2011**, *286*, 17658–17664.

(167) Abe, H.; Kawasaki, K.; Nakanishi, H. *J. Biochem.* **2002**, *132*, 863–874.

(168) Mehta, A. K.; Lu, K.; Childers, W. S.; Liang, Y.; Dublin, S. N.; Dong, J.; Snyder, J. P.; Pingali, S. V.; Thiyagarajan, P.; Lynn, D. G. *J. Am. Chem. Soc.* **2008**, *130*, 9829–9835.

(169) Micali, N.; Villari, V.; Consoli, G. M. L.; Cunsolo, F.; Geraci, C. *Phys. Rev. E* **2006**, *73*, 051904.

(170) Vlachy, N.; Merle, C.; Touraud, D.; Schmidt, J.; Talmon, Y.; Heilmann, J.; Kunz, W. *Langmuir* **2008**, *24*, 9983–9988.

(171) Morrow, B. H.; Wang, Y.; Wallace, J. A.; Koenig, P. H.; Shen, J. K. *J. Phys. Chem. B* **2011**, *115*, 14980–14990.

(172) Wallace, J. A.; Shen, J. K. *J. Phys. Chem. Lett.* **2012**, *3*, 658–662.

(173) Shi, C.; Wallace, J. A.; Shen, J. K. *Biophys. J.* **2012**, *102*, 1590–1597.

(174) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Hsing, L.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

(175) Nina, M.; Simonson, T. *J. Phys. Chem. B* **2002**, *106*, 3696–3705.

(176) Baumketner, A.; Shea, J.-E. *J. Phys. Chem. B* **2005**, *109*, 21322–21328.

(177) Monticelli, L.; Simões, C.; Belvisi, L.; Colombo, G. *J. Phys. Condens. Matter* **2006**, *18*, S329–S345.

(178) Walser, R.; Hünenberber, P. H.; van Gunsteren, W. F. *Proteins* **2001**, *43*, 509–519.

(179) Gargallo, R.; Hünenberger, P. H.; Avilés, F. X.; Oliva, B. *Protein Sci.* **2003**, *12*, 2161–2172.

(180) Robertson, A.; Luttman, E.; Pande, V. S. *J. Comput. Chem.* **2008**, *29*, 694–700.

(181) Ullmann, G. M. *J. Phys. Chem. B* **2003**, *107*, 1263–1271.

(182) Noskov, S. Y.; Roux, B. *J. Mol. Biol.* **2008**, *377*, 804–818.

(183) Luo, Y.; Roux, B. *J. Phys. Chem. Lett.* **2010**, *1*, 183–189.

(184) Arbely, E.; Rutherford, T. J.; Neuweiler, H.; Sharpe, T. D.; Ferguson, N.; Fersht, A. R. *J. Mol. Biol.* **2010**, *403*, 313–327.

(185) Itoh, S. G.; Damjanović, A.; Brooks, B. R. *Proteins* **2011**, *79*, 3420–3436.

(186) Patriksson, A.; van der Spoel, D. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2073–2077.

(187) Gane, R.; Ingold, C. K. *J. Chem. Soc.* **1931**, 2153–2169.

(188) Kirkwood, J. G.; Westheimer, F. H. *J. Chem. Phys.* **1938**, *6*, 506–512.

(189) Westheimer, F. H.; Kirkwood, J. G. *J. Chem. Phys.* **1938**, *6*, 513–517.

(190) Trebst, S.; Troyer, M.; Hansmann, U. H. E. *J. Chem. Phys.* **2006**, *124*, 174903.

(191) Wang, M.; Tang, Y.; Sato, S.; Vugmeyster, L.; McKnight, C. J.; Raleigh, D. P. *J. Am. Chem. Soc.* **2003**, *125*, 6032–6033.

(192) Rozanska, X.; Chipot, C. *J. Chem. Phys.* **2000**, *112*, 9691–9694.

(193) Gruia, A. D.; Fischer, S.; Smith, J. C. *Chem. Phys. Lett.* **2004**, *385*, 337–340.

(194) Frishman, D.; Argos, P. *Proteins* **1995**, *23*, 566–579.

(195) Tang, Y.; Rigotti, D. J.; Fairman, R.; Raleigh, D. P. *Biochemistry* **2006**, *43*, 3264–3272.

(196) Meng, W.; Shan, B.; Tang, Y.; Raleigh, D. P. *Protein Sci.* **2009**, *18*, 1692–1701.

(197) Reiner, A.; Henklein, P.; Kiefhaber, T. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 4955–4960.

(198) Zagrovic, B.; Snow, C. D.; Khaliq, S.; Shirts, M. R.; Pande, V. S. *J. Mol. Biol.* **2002**, *323*, 153–164.

(199) Lei, H.; Su, Y.; Jin, L.; Duan, Y. *Biophys. J.* **2010**, *99*, 3374–3384.

(200) Wickstrom, L.; Okur, A.; Song, K.; Hornak, V.; Raleigh, D. P.; Simmerling, C. L. *J. Mol. Biol.* **2006**, *360*, 1094–1107.

(201) Stites, W. E.; Gittis, A. G.; Lattman, E. E.; Shortle, D. *J. Mol. Biol.* **1991**, *221*, 7–14.

(202) Garcia-Moreno E., B.; Dwyer, J. J.; Gittis, A. G.; Lattman, E. E.; Spencer, D. S.; Stites, W. E. *Biophys. Chem.* **1997**, *64*, 211–224.

(203) Fitch, C. A.; Karp, D. A.; Lee, K. K.; Stites, W. E.; Lattman, E. E.; E., B. G.-M. *Biophys. J.* **2002**, *82*, 3289–3304.

(204) Karp, D. A.; Gittis, A. G.; Stahley, M. R.; Fitch, C. A.; Stites, W. E.; García-Moreno E., B. *Biophys. J.* **2007**, *92*, 2041–2053.

(205) Dwyer, J. J.; Gittis, A. G.; Karp, D. A.; Lattman, E. E.; Spencer, D. S.; Stites, W. E.; García-Moreno E., B. *Biophys. J.* **2000**, *79*, 1610–1620.

(206) Nguyen, D. M.; Reynald, R. L.; Gittis, A. G.; Lattman, E. E. *J. Mol. Biol.* **2004**, *341*, 565–574.

Appendix: Summary of SNase mutant pK_a predictions

pK_a calculations with published experimental data

Name (PDB ID)	Background	Res. Type	Res. ID	Calc.	Exp.	Error	Ref.
V66K(2snm)	wild type	K	66	7.5	6.4	1.1	I
V66K	PHS	K	66	6.9	6.4	0.6	II
V66K	Δ+PHS	K	66	7.0	5.6	1.4	III
V66D(2oxp)	PHS	D	66	6.8	8.7	-1.9	IV
V66E(1u9r)	PHS	E	66	8.4	8.5	-0.1	VII
				8.4	(8.7)	-0.3	V
I92E(1tqo)	Δ+PHS	E	92	6.9	9.0	-2.1	VII
				6.9	(8.7)	-1.8	VI
I92K(1tt2)	Δ+PHS	K	92	6.6	5.3	1.3	VII
				6.6	(5.6)	1.0	VI
I92D(2oeo)	Δ+PHS	D	92	6.8	8.1	-1.3	V

Previously published data are listed in parentheses.

pK_a predictions

Name (PDB ID)	Background	Res Type	Res ID	Calc	Expt	Error[a]	Ref
G20R	Δ+PHS	R	20	12.9	>10.4	COR	VII
G20E	Δ+PHS	E	20	4.2	<4.5	COR	
G20K	Δ+PHS	K	20	11.3	>10.4	COR	
G20D	Δ+PHS	D	20	2.3	<4.0	COR	
V23R	Δ+PHS	R	23	13.6	>10.4	COR	
V23E	Δ+PHS	E	23	7.5	7.1	0.4	
V23D	Δ+PHS	D	23	6.7	6.8	-0.1	
V23K	Δ+PHS	K	23	7.1	7.3	-0.2	
L25K(3erq)	Δ+PHS	K	25	4.2	6.3	-2.1	
L25E(3evq)	Δ+PHS	E	25	8.4	7.5	0.9	
L25D	Δ+PHS	D	25	7.7	6.8	0.9	
L25R	Δ+PHS	R	25	14.9	>10.4	COR	
F34R	Δ+PHS	R	34	11.8	>10.4	COR	
F34D	Δ+PHS	D	34	6.6	7.8	-1.2	
F34K	Δ+PHS	K	34	7.5	7.1	0.4	
F34E	Δ+PHS	E	34	7.5	7.3	0.2	
L36E	Δ+PHS	E	36	7.1	8.7	-1.6	
L36K(3eji)	Δ+PHS	K	36	8.0	7.2	0.8	
L36R	Δ+PHS	R	36	11.4	>10.4	COR	
L36D	Δ+PHS	D	36	5.6	7.9	-2.3	
L37E	Δ+PHS	E	37	6.1	5.2	0.9	
L37K	Δ+PHS	K	37	7.5	>10.4	>2.92	
L37D	Δ+PHS	D	37	5.1	<4.0	>1.05	
L37R	Δ+PHS	R	37	11.1	>10.4	COR	
L38D	Δ+PHS	D	38	6.6	6.8	-0.2	
L38R	Δ+PHS	R	38	13.5	>10.4	COR	
V39R	Δ+PHS	R	39	14.7	>10.4	COR	
V39E	Δ+PHS	E	39	8.9	8.2	0.7	
V39D	Δ+PHS	D	39	9.8	8.1	1.7	
V39K	Δ+PHS	K	39	8.0	9.0	-1.0	
T41R	Δ+PHS	R	41	14.2	>10.4	COR	
T41E	Δ+PHS	E	41	6.3	6.5	-0.3	
T41K	Δ+PHS	K	41	8.5	9.3	-0.8	
T41D	Δ+PHS	D	41	6.5	<4.0	>2.45	
A58E	Δ+PHS	E	58	5.2	7.7	-2.5	
A58R	Δ+PHS	R	58	13.4	>10.4	COR	
A58K	Δ+PHS	K	58	9.1	>10.4	>1.33	
A58D	Δ+PHS	D	58	6.3	6.8	-0.5	
T62R	Δ+PHS	R	62	14.7	>10.4	COR	

pK_a predictions (continued)

Name (PDB ID)	Background	Res Type	Res ID	Calc	Expt	Error[a]	Ref
T62K(3dmu)	PHS	K	62	6.0	8.1	-2.1	VII
T62D	Δ+PHS	D	62	6.3	8.7	-2.4	
T62E	Δ+PHS	E	62	7.3	7.7	-0.4	
I72K(2rbm)	Δ+PHS	K	72	9.2	8.6	0.6	
I72E(3ero)	Δ+PHS	E	72	6.7	7.3	-0.6	
I72D	Δ+PHS	D	72	6.3	7.6	-1.4	
V74D	Δ+PHS	D	74	8.3	8.3	0.0	
V74K	Δ+PHS	K	74	8.0	7.4	0.6	
V74R	Δ+PHS	R	74	13.4	>10.4	COR	
V74E	Δ+PHS	E	74	9.8	7.8	2.0	
A90D	Δ+PHS	D	90	7.1	7.5	-0.4	
A90K	Δ+PHS	K	90	7.2	8.6	-1.4	
A90E	Δ+PHS	E	90	9.1	6.4	2.7	
Y91D	Δ+PHS	D	91	4.4	7.2	-2.8	
Y91R	Δ+PHS	R	91	12.2	>10.4	COR	
Y91K	Δ+PHS	K	91	6.9	5.3	1.6	
Y91E(3d4d)	Δ+PHS	E	91	4.9	7.1	-2.2	
V99D	Δ+PHS	D	99	4.6	8.5	-3.9	
V99E	Δ+PHS	E	99	7.6	8.4	-0.8	
V99R	Δ+PHS	R	99	13.7	>10.4	COR	
V99K	Δ+PHS	K	99	7.8	6.5	1.3	
N100R	Δ+PHS	R	100	12.1	>10.4	COR	
N100K	Δ+PHS	K	100	3.8	8.6	-4.8	
N100E	Δ+PHS	E	100	4.6	7.6	-3.0	
N100D	Δ+PHS	D	100	5.9	6.9	-1.0	
L103E	Δ+PHS	E	103	7.8	8.9	-1.1	
L103R	Δ+PHS	R	103	9.5	>10.4	>0.88	
L103D	Δ+PHS	D	103	6.8	8.7	-1.9	
L103K(3e5s)	Δ+PHS	K	103	11.1	8.2	2.9	
V104D	Δ+PHS	D	104	7.8	9.7	-1.9	
V104R	Δ+PHS	R	104	13.5	>10.4	COR	
V104E	Δ+PHS	E	104	7.0	9.4	-2.4	
V104K(3c1f)	Δ+PHS	K	104	9.6	7.7	1.9	
A109K	Δ+PHS	K	109	7.3	9.2	-1.9	
A109R(3d4w)	Δ+PHS	R	109	14.1	>10.5	COR	
A109E	Δ+PHS	E	109	7.4	7.9	-0.5	
A109D	Δ+PHS	D	109	7.1	7.5	-0.4	

pK_a predictions (continued)

Name (PDB ID)	Background	Res Type	Res ID	Calc	Expt	Error[a]	Ref
N118E	Δ+PHS	E	118	5.4	<4.5	>0.94	VII
N118D	Δ+PHS	D	118	6.6	7.0	-0.4	
N118R	Δ+PHS	R	118	11.1	>10.4	COR	
N118K	Δ+PHS	K	118	9.1	>10.4	>1.27	
L125D	Δ+PHS	D	125	6.7	7.6	-0.9	
L125R	Δ+PHS	R	125	14.3	>10.4	COR	
L125E	Δ+PHS	E	125	7.9	9.1	-1.2	
A132D	Δ+PHS	D	132	5.6	7.0	-1.4	
A132K	Δ+PHS	K	132	7.4	>10.4	>3.05	
A132R	Δ+PHS	R	132	14.1	>10.4	COR	
A132E	Δ+PHS	E	132	6.8	7.0	-0.2	

[a] The uncertainty of experimental pK_a measurements using thermodynamic stabilities is 0.2–0.5 pK units[86]. COR means the calculated pK_a is within the experimentally determined bound.

References

I[201]; II[202]; III[203]; IV[204]; V[205]; VII[206];

VII. Data from "pKacoop" (http://amylase.ucd.ie/pKacoop/).