Exploring the Solvent Environment of Biomolecular Systems

by

Evan J. Arthur

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Chemistry)
in The University of Michigan
2015

Doctoral Committee:

Professor Charles L. Brooks III, Co-chair
Professor Kevin J. Kubarych, Co-chair
Professor Heather A. Carlson
Professor Eitan Geva
Professor David Sept

To my Parents

Acknowledgements

I graciously thank my two advisors Charles Brooks III and Kevin Kubarych for their shaping me into a better scientist. I hope to carry and pass on their curiosity and enthusiasm for discovery.

I also owe great thanks for current and previous lab members. In particular I thank John King, Carlos Baiz, Logan Ahlstrom, Michael Garrahan, and David Braun for mentorship and helping train me as a productive chemist and programmer.

I also thank my committee members Professor Charles L. Brooks III, Professor Kevin J. Kubarych, Professor Eitan Geva, Professor Heather A. Carlson, and Professor David Sept for their invaluable insight into the development and completion of the work presented in this thesis.

Last, but not least, I thank my family for their continued support and encouragement as I develop my passion and career as a chemist.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

1D: one dimensional

2D: two dimensional

2DIR: two-dimensional infrared

3D: three dimensional

3D-RISM: three-dimensional reference interaction site model

AMBER: Assisted Model Building with Energy Refinement

ASP: aspartic acid

AUE: average unsigned error

BphC: biphenyl dioxygenase

CD: circular dichroism

CHARMM: Chemistry at HARvard Macromolecular Mechanics

CPHMD: Constant pH Molecular Dynamics

CORM: carbon monoxide releasing molecule

CPU: central processing unit

CUDA: Compute Unified Device Architecture

DFG: difference frequency generation

FACTS: Fast Analytical Continuum Treatment of Solvation

FFCF: frequency–frequency correlation function

fs: femtosecond

FTIR: Fourier transform infrared

GB: generalized Born

GBMV: Generalized Born using Molecular Volume

GBSA/OBC: Generalized Born Surface Area from Onufriev, Bashford, and Case

GBSW: Generalized Born with a Simple sWitching function

GLU: glutamic acid

GPU: graphics processing unit

GROMACS: GROningen MAchine for Chemical Simulations

HB: hydrogen bond

HEWL: hen egg white lysozyme

HEWL-RC: hen egg white lysozyme ruthenium carbonyl complex

HH: Henderson-Hasselbalch

HIS: histidine

HuLys: human lysozyme

HuLys-RC: human lysozyme Ru-carbonyl complex

MCCE: multi-conformation continuum electrostatics

MD: molecular dynamics

MC: Monte Carlo

MEAD: macroscopic electrostatics with atomic detail

MM: molecular mechanics

μs: microsecond

NAMD: NAnoscale Molecular Dynamics program

ns: nanosecond

ODNP: Overhauser dynamic nuclear polarization

OpenCL: Open Computing Language

OpenMM: Open Molecular Mechanics

OKE: optical Kerr effect

ps: picosecond

PB: Poisson-Boltzmann

PME: particle-mesh Ewald

RESP: Restrained Electrostatic Potential

REX: replica exchange

RMSD: root-mean-square deviation

SNase: staphylococcal nuclease

TFE: 2,2,2-trifluoroethanol

Abstract

Solvent interactions at the protein-solvent interface facilitate many biological processes such as protein-protein recognition, protein-DNA binding, and a variety of enzymatic mechanisms. Consequently, developing a comprehensive understanding of solvation effects has been pursued for many decades, and promises benefits to many branches of biomolecular science. The following series of studies explore the maturation and improvement of several computational solvent models and analytical methods for studying protein-solvent interactions, and is divided into two principle sections. In the first half we create a detailed analysis of the protein-solvent interface, and explore dynamic mechanisms proteins employ for structural stability. The second half we follow the development and refactoring of accurate implicit solvent models to take advantage of modern parallel processing chips, and in doing so we enable new timescales for studying conformational equilibria and titration states.

Recent developments in 2DIR spectroscopy have enabled the study of site-specific hydration dynamics on protein and membrane surfaces. In the first three chapters we explore the development and significance of this new technology. The lifetime decay of signal amplitude and the spectral diffusion from metal carbonyl probe molecules report local water concentration and dynamics, respectively. By site-specifically bonding ruthenium dicarbonyl ($Ru(CO)_2$) probes to different protein domains on lysozymes, we find direct

experimental evidence for spatially-heterogeneous hydrophobicity. Additionally, we find that hydration dynamics are slowed down on protein surfaces relative to those of bulk water, and that protein clusters can cooperatively reduce water dynamics over 15 Å from a protein surface. In addition to supporting these experimental findings, the subsequent MD simulations of the lysozyme systems indicated that specific features of solvent interaction at the protein-solvent interface originate from a collective behavior of local amino acids. Not only was it confirmed that solvent interaction around histidines can be modulated by nearby residues, but that average solvation around identical alpha helices on two homologous lysozymes can be dissimilar. These findings show that even with an abundance of identical residues, homologous proteins do not necessarily share similar interactions with solvents. Our results provide an intuitive picture of the dynamic aspects of protein hydration, and illustrate how proteins control their local solvent environments to facilitate biological processes.

Next we explore protein solvation from a more coarse-grained perspective through benchmarking and improving the Generalized Born implicit solvent model with a Simple sWitching function (GBSW). Implicitly represented solvent speeds up molecular simulations by reducing the system size and eliminating the need to equilibrate solvent molecule conformations. Additionally, through adjustments to solvation free energy parameters, applying instantaneous changes in pH or salt concentration are relatively straightforward to implement. Constant pH Molecular Dynamics (CPHMD) is one such pH model that adjusts the partial charges of titrating atoms to simulate the effects of pH on a given solute. Such usefulness, however, has often been accompanied by poor scaling of algorithms for large system sizes, and poor utilization of modern parallel computing hardware. With the new availability of graphics chips containing thousands of processing cores, there is a great

opportunity in refactoring these aging implicit solvent models into efficiently parallel processes. During our benchmarking study of CPHMD it was found to predict pK_a values of residues to within 1 pK_a unit, but was also far too slow to be used for high-throughput applications. Through reconstructing and parallelizing the algorithms of GBSW and CPHMD on graphics processing units (GPUs) we offer an improvement over the original algorithm by about 1-2 and 2-3 orders of magnitude respectively, depending on the system size and nonbonded cutoff parameters. The algorithms also scale better with system size than the originals, which broadens their applicability in both high-throughput and large-system studies.

Chapter 1

Introduction

1.1 Proteins in Aqueous Solution

Many biological processes are mediated by the presence of water. Phenomena such as protein-ligand binding,[1-6] protein-protein recognition,[7,8] and ice crystal inhibition,[9] derive function both by dynamically constraining the movements of water molecules, and by harnessing water's electrostatic pressure.[10-12] Understanding this mediation promises to benefit pursuits of drug discovery, and further complete our modeling of biology. As we develop experimental techniques for analyzing aqueous solvents interacting with protein and investigate the dynamic mechanisms of those experiments through computer modeling, it is essential that we account for the presence of water with sufficient detail.[3,13-21]

The spectroscopic experiments explored in this thesis report hydrogen bond dynamics. In order to simulate systems relevant to these experimental results we need to simulate the solvent in explicit, all-atom detail. These high-resolution models take many forms, and are broadly categorized into polarizable, such as SWM4-DP[22] and SWM4-NDP[23]; and static charge models, such as SPC,[24] SPC/E,[25] TIP3P,[26] and TIP4P[26]. Interestingly, SPC/E, despite being a relatively crude 3-point static charge model, has shown excellent

correspondence with experimental properties of water such as diffusion coefficients and melting temperatures.[27-29] In addition to capturing bulk properties, SPC/E has been useful in studing fine details of water-solute interactions such as dynamical slowdown near hydrophobic surfaces, and rapid angular jumps associated with hydrogen bond reorganization.[17,30-32] Due in part to its accuracy and low computational cost, SPC/E was used as the explicit solvent of choice in our modeling solvent interaction with macrobiomolecules. As we will see in later chapters, SPC/E was used to observe the fine angstrom-scaled details of average water placement at the protein-solvent interface, as well as long reaching effects, such as nanometer-scaled dynamic water slowdown near protein interfaces.

1.2 Solvating Proteins in Water-Trifluoroethanol Cosolvent

Due to the large number of non-water components in the intracellular medium,[33,34] our studies of protein-solvent interfaces must address the effects of a cosolvent environment. Thus we chose a well-understood protein-cosolvent system, specifically lysozyme proteins in trifluoroethanol (TFE) mixtures,[35] from which we begin our investigations of the protein-solvent interface. Previous NMR and circular dichroism studies of HEWL provided for us concentrations of TFE that neither change the helical content nor the tertiary contacts of lysozyme.[36,37] In later chapters we will explore such systems using two-dimensional infrared spectroscopy (2DIR), and in doing so we gain an understanding on how solvation and dehydration can differ depending on the specific location on a protein. Additionally we explore generating compatible forcefield parameters for TFE, and we verify its ability to reproduce preferential solvation.[38] Much in the same way the SPC/E model reproduces many of water's experimental bulk properties despite only capturing water's

small size and strong dipole, our TFE model reproduces features of a water-TFE mixture by simply having a relative distribution of hydrophobic and hydrophilic atom groups. Although such solvent systems are unnatural, each result gives us insight into how the heterogeneous distribution of partial charges of protein surfaces in turn produces a heterogeneous pattern of preferential solvation.

1.3 Enabling pH-Coupled Conformational Dynamics

Proteins typically maintain their native structure and optimal functionality under a narrow range of pH.[39-41] Consequently, many biological systems tightly control local solvent pH to tune the effectiveness of enzymes, or to promote a useful protein conformation.[39,42,43] Mitochondrial ATP synthase utilizes a trans-membrane proton gradient to power its rotary catalysis mechanism,[44-46] and the departure from a normal pH range is known to be a driving force in forming the amyloid fibrils associated with Alzheimer's disease.[47,48] Additional examples of pH driven processes include the proton-activated gate mechanism of the KcsA potassium channel,[49] and the catalytic pathway of dihydrofolate reductase.[50] Finally, a notable survey by Aguilar et al. showed that about 60% of the protein-ligand complexes indicated that at least one titratable residue of the protein assumed a different protonation state between bound and unbound states.[51] Although important to many biological processes, pH-dependence in biomacromolecule simulations is greatly limited to short timescales, and is generally restricted to nanosecond-long timescales when proteins are simulated in full-atomic detail. This limitation effectively bars much observation of detailed, large-scale conformational dynamics and protein relaxation.

In the latter portion of this thesis we follow the process of parallelizing and improving of the constant pH molecular dynamics (CPHMD) titration method developed by

Lee et al.[52,53] This method is a form of continuous titration, and it enables the simultaneous transformation of multiple residues among protonation and tautomeric states. The result is a pH simulation method that can calculate pK_a values of protein structures to within 1 pK unit,[54] and when coupled to coarse-grained systems, can resolve the dominant folding pathway of the pH-sensitive HdeA homodimers.[55] The parallelization process begins with refactoring the Generalized Born implicit water model with a Simple sWitching function (GBSW) model[56] to function effectively on graphics processing units (GPUs) with thousands of parallel cores. Algorithmic improvements were made to enable better scaling with both system size and number of available parallel cores. The result was a version of GBSW that ran about 30 times faster and scaled better than its previous implementation. With the solvent model in place, we then added components of CPHMD inside the GBSW processes to gain an efficient and effective model of titration. We achieve speed increases of between 2 and 3 orders of magnitude over the original CPHMD original algorithm, and consequently, we enable microsecond-long simulations of biological processes to be computed in all-atom detail using relatively inexpensive GPUs. This new tool promises to bring detailed answers for many more questions regarding pH-coupled protein conformational change, as well as make CPHMD's pKa predictions fast enough and cost-effective enough to be appropriate for high-throughput applications.

1.4 Thesis Outline

In **Chapter 2** we introduce the methodology of probing local solvent environments using 2DIR, and we show that solvation and dehydration can differ depending on the specific location on a protein. Hen egg white lysozyme (HEWL) and human lysozyme (HuLys) offer homologous protein topologies, each with one solvent-exposed histidine.

Although the two proteins are 77% similar by amino acid sequence and are structurally different by only 0.54 Å root-means-square, the histidines are located on different domains of the protein. The H15 on HEWL is located on a turn adjacent to an alpha-helix, and the H78 on HuLys is located on a region without secondary structure. Local environments around these histidines were probed by covalently attaching a ruthenium-carbonyl vibrational chromophore. In initial studies the vibrational lifetime of the chromophore in H_2O and D_2O was used to measure not only the presence of water, but also the hindering of hydrogen bond reorientation dynamics in the nearby hydration water. It was found that different water dynamics correlate strongly with the local surface structure of the protein. The H15 probe location of HEWL is a low-curvature region solvated by orientationally constrained water, whereas the H78 site of HuLys is high-curvature and unstructured, and solvated by bulk-like water. To test the connection between constrained water and the thermodynamic driving force for dehydration by an amphiphilic co-solvent trifluoroethanol (TFE), lifetime measurements were made in a series of D_2O/TFE solutions. In pure D_2O, both sites were found to be hydrated based on their sub-5 ps vibrational lifetimes, which are consistent with water-assisted relaxation.[57] Upon addition of TFE, however, the sites displayed markedly distinct responses. The lifetime of the probe at the H15 site of HEWL exhibited an order-of-magnitude slowdown in a 10% (v/v) TFE solution consistent with local dehydration, whereas the H78 labeled site of HuLys showed no TFE-dependent vibrational lifetime changes at any of the experimental concentrations.

In **Chapter 3** we investigate the results of the previous chapter and explore the heterogeneous nature of preferential solvation of lysozyme by TFE-water mixtures. We use explicit solvent MD simulations to model human and hen egg white lysozymes mixed with water and different concentrations of trifluoroethanol. We then aligned each trajectory by

lowest protein backbone-atom root-mean-square deviation (RMSD) to one common structure. From these trajectories we then compute time-averaged three-dimensional histograms of the number density of solvent relative to each protein's structure. These values represent the spatial distribution of both the probability of finding a type of solvent atom and solvent density. Using these data, we mapped out trends of trifluoroethanol interacting with lysozyme surfaces and suggest a possible explanation for the observed phenomena in the spectroscopic experiments. Finally, we made a spatially dependent, solvent-centric comparison of homology between HEWL and humLys. We find that the 2DIR studies' reporting that the H78 site of HuLys is more hydrophilic than H15 site of HEWL is a reasonable conclusion. Additionally, we investigate how the homology of protein structure does not necessarily translate to similarities in solvent structure and composition, even when observing identical side chains.

In **Chapter 4** we use the spectroscopic tools and simulation framework from previous chapters to explore crowding effects near protein surfaces. Again we use 2DIR spectroscopy of ruthenium-carbonyl complexes bound to lysozyme proteins. By observing the vibrational relaxation of the probes we determine local hydration dynamics at the probe binding sites. We place the lysozyme-probe complex in aqueous solutions of PEG400 (8–9mer) ranging from 0 to 80% PEG400 by volume, and compare these results to previously reported experiments using glycerol.[58] Then we carry out a parallel experiment with the probe complex in varying concentrations of excess lysozyme ranging from 20 to 160 mg/mL, and we observe the effects of protein self-crowding. Interestingly, we find an abrupt dynamical transition of the protein and hydration dynamics induced by crowding, and the results suggest a dynamic hydration shell around the protein extending 15–20 Å, resulting in collective hydration for interprotein separations of 30–40 Å. To support and

elucidate a possible mechanism for these observations, MD simulations of protein-crowded were performed by arranging lysozyme molecules in varying distances from each other. The hydration structure and dynamics of the resultant trajectories were analyzed by averaging hydrogen bond counts and lifetimes in water near the hydration sites. Long-ranged slowdown of water dynamics similar to the experiments was observed. The consensus from both experiment and simulation, then, is the existence of two distinct dynamical regimes of biomacromolecules in solvent into "undercrowded" and "overcrowded" conditions.

Chapter 5 begins a series of studies where we explore and improve the implicit modeling of solvent. We begin with benchmarking the accuracy of the constant pH with molecular dynamics (CPHMD) model, and recapitulate the titratable residue pK_a values of staphylococcal nuclease variants. In previous work by García-Moreno *et al.*, the conformational role of aspartic and glutamic acids (GLU) in Δ+ PHS were studied in detail.[59] All such residues were titrated for pK_a calculations by measuring the pH dependence of the chemical shifts of Cγ or Cδ with two-dimensional HBHC(CBCG)CO experiments,[59] which led to a comprehensive quantification of the changes of internal energy within Δ + PHS in relation to introducing a hydrophilic residue into the hydrophobic core of the protein. The shielding effect of surrounding hydrophobic amino acids can reduce solvent interactions and consequently increases residue pK_a values by as much as 5 pK units. The measured perturbation in pK_a values for these systems provides an experimental basis for testing and assessing the accuracy of the CPHMD model. We find that for all variants of staphylococcal nuclease, including those variants that lack a fully-solved crystal structure, CPHMD correctly predicts perturbations in pK_a values to within 1 pK unit. Although accurate and applicable to a wide range of systems, CPHMD is too slow to be useful for large systems or high-throughput studies. For instance, converging titrating residues of the

nucleases to optimal protonation states required over a week of simulation time.

In **Chapter 6** we continue our studies of implicit solvent by refactoring the generalized Born with a simple switching function (GBSW) solvent model so it functions well on highly-parallel graphics chips. With the availability of graphics processing units (GPUs) carrying up to thousands of parallel processing cores and their newer ability to compute complex mathematical functions using C-like languages such as Open Computing Language (OpenCL) and Compute Unified Device Architecture (CUDA), a new frontier of GPU-powered ultra-parallel molecular dynamics software has come into being. Programs such as CHARMM,[60] AMBER,[61] OpenMM,[62] GROMACS,[63] and NAMD[64] all offer GPU-accelerated options for many types of simulations, all of which can replace the computational power of much larger computer networks with a single graphics card. Despite the fantastic improvements in molecular mechanics simulations afforded by GPUs, some algorithms remain challenging to parallelize. Notable among these are implicit solvent models, which either rely on recursive data processing or are inefficiently split into parallel functions. From the variety of implicit solvent methods for calculating solvation free energy, only those that use an uncoupled summation of Cramer-Truhlar-type atom-atom pairwise interactions,[65] such as GBSA/OBC,[66,67] have been implemented in GPU languages. Such implementations only required a retooled version of the neighboring atom interaction processes that were already developed for all-atom molecular mechanics.[62-64,66] This chapter we outline the implementation of a parallel, atom-coupled volumetric integration approach to calculating solvation free energy using the GBSW algorithm. Depending on the system size and nonbonded force cutoffs, the new GBSW algorithm offers speed increases of between one and two orders of magnitude over previous implementations while maintaining similar levels of accuracy. We also demonstrate that these speed enhancements now make

accessible folding studies of peptides and potentially small proteins by utilizing our GPU-accelerated GBSW model to fold the model system chignolin.

Chapter 7 extends the work performed in the GBSW solvent model to include the CPHMD model in its new highly-parallel processing platform. CPHMD models the influence of pH on a system by extending the Hamiltonian to include a continuous, pH-sensitive λ coordinate for each titrating residue. Each λ coordinate determines which protonation state a residue resides in. As such, each titrating residue has different sets of partial charges for each titration and tautomeric state, and a potential energy function that connects the λ coordinates to the local charge and pH environment. This model permits neighboring titrating residues to interact, and allows all residues to titrate simultaneously. By deconstructing the potential energy function and calculating most of it in parallel along with the GBSW solvent model, we see speed improvements in CPHMD of 2 and 3 orders of magnitude over its original form. With such speed improvements, the pH model is now appropriate for a much wider range of system sizes and trajectory lengths, and hopefully will enable the fine-tuning and wider acceptance of pH modeling in MD simulations.

The last chapter summarizes the results and draws general conclusions in the context of solvent modeling and analysis, and we discuss the new directions for adapting those models to a new frontier of highly-parallel processing hardware.

1.5 References

1. O. Rahaman, S. Melchionna, D. Laage, and F. Sterpone, "The Effect of Protein Composition on Hydration Dynamics," Phys. Chem. Chem. Phys. **15**(10), 3570-76, (2013).

2. B. Q. Q. Wei, W. A. Baase, L. H. Weaver, B. W. Matthews, and B. K. Shoichet, "A Model Binding Site for Testing Scoring Functions in Molecular Docking," J. Mol. Biol. **322**(2), 339-55, (2002).

3. K. W. Lexa, and H. A. Carlson, "Full Protein Flexibility Is Essential for Proper Hot-Spot Mapping," J. Am. Chem. Soc. **133**(2), 200-02, (2011).

4. I. Halperin, B. Y. Ma, H. Wolfson, and R. Nussinov, "Principles of Docking: An Overview of Search Algorithms and a Guide to Scoring Functions," Proteins: Struct., Funct., Genet. **47**(4), 409-43, (2002).

5. P. A. Sigala, J. M. M. Caaveiro, D. Ringe, G. A. Petsko, and D. Herschlag, "Hydrogen Bond Coupling in the Ketosteroid Isomerase Active Site," Biochemistry **48**(29), 6932-39, (2009).

6. P. A. Sigala, D. A. Kraut, J. M. M. Caaveiro, B. Pybus, E. A. Ruben, D. Ringe, G. A. Petsko, and D. Herschlag, "Testing Geometrical Discrimination within an Enzyme Active Site: Constrained Hydrogen Bonding in the Ketosterold Isomerase Oxyanion Hole," J. Am. Chem. Soc. **130**(41), 13696-708, (2008).

7. S. Jones, and J. M. Thornton, "Principles of Protein-Protein Interactions," Proc. Natl. Acad. Sci. U.S.A. **93**(1), 13-20, (1996).

8. G. A. Papoian, J. Ulander, and P. G. Wolynes, "Role of Water Mediated Interactions in Protein-Protein Recognition Landscapes," J. Am. Chem. Soc. **125**(30), 9170-78, (2003).

9. K. Meister, S. Ebbinghaus, Y. Xu, J. G. Duman, A. DeVries, M. Gruebele, D. M. Leitner, and M. Havenith, "Long-Range Protein-Water Dynamics in Hyperactive Insect Antifreeze Proteins," Proc. Natl. Acad. Sci. U.S.A. **110**(5), 1617-22, (2013).

10. E. E. Fenn, D. E. Moilanen, N. E. Levinger, and M. D. Fayer, "Water Dynamics and Interactions in Water-Polyether Binary Mixtures," J. Am. Chem. Soc. **131**(15), 5530-39, (2009).

11. U. Sreenivasan, and P. H. Axelsen, "Buried Water in Homologous Serine Proteases," Biochemistry **31**(51), 12785-91, (1992).

12. Y. Levy, and J. N. Onuchic. in *Annu. Rev. Biophys. Biomol. Struct.* Vol. 35 *Annual Review of Biophysics* 389-415 (2006).

13. A. Kovalenko, and F. Hirata, "Three-Dimensional Density Profiles of Water in Contact with a Solute of Arbitrary Shape: A Rism Approach," Chem. Phys. Lett. **290**(1-3), 237-44, (1998).

14. A. Fernandez, "Epistructural Tension Promotes Protein Associations," Phys. Rev. Letters **108**(18), (2012).

15. J. T. King, E. J. Arthur, C. L. Brooks III, and K. J. Kubarych, "Site-Specific Hydration Dynamics of Globular Proteins and the Role of Constrained Water in Solvent Exchange with Amphiphilic Cosolvents," J. Phys. Chem. B **116**(19), 5604-11, (2012).

16. T. Imai, K. Oda, A. Kovalenko, F. Hirata, and A. Kidera, "Ligand Mapping on Protein Surfaces by the 3d-Rism Theory: Toward Computational Fragment-Based Drug Design," J. Am. Chem. Soc. **131**(34), 12430-40, (2009).

17. A. J. Patel, P. Varilly, S. N. Jamadagni, H. Acharya, S. Garde, and D. Chandler, "Extended Surfaces Modulate Hydrophobic Interactions of Neighboring Solutes," Proc. Natl. Acad. Sci. U.S.A. **108**(43), 17678-83, (2011).

18. C. Ma, J. Tran, F. Gu, R. Ochoa, C. Li, D. Sept, K. Werbovetz, and N. Morrissette, "Dinitroaniline Activity in Toxoplasma Gondii Expressing Wild-Type or Mutant Alpha-Tubulin," Antimicrob. Agents Chemother. **54**(4), 1453-60, (2010).

19. S. Genheden, T. Luchko, S. Gusarov, A. Kovalenko, and U. Ryde, "An Mm/3d-Rism Approach for Ligand Binding Affinities," J. Phys. Chem. B **114**(25), 8505-16, (2010).

20. F. Sterpone, G. Stirnemann, and D. Laage, "Magnitude and Molecular Origin of Water Slowdown Next to a Protein," J. Am. Chem. Soc. **134**(9), 4116-19, (2012).

21. D. Chandler, "Interfaces and the Driving Force of Hydrophobic Assembly," Nature **437**(7059), 640-47, (2005).

22. G. Lamoureux, A. D. MacKerell, and B. t. Roux, "A Simple Polarizable Model of Water Based on Classical Drude Oscillators," J. Chem. Phys. **119**(10), 5185-97, (2003).

23. G. Lamoureux, E. Harder, I. V. Vorobyov, B. Roux, and A. D. MacKerell, "A Polarizable Model of Water for Molecular Dynamics Simulations of Biomolecules," Chem. Phys. Lett. **418**(1-3), 245-49, (2006).

24. H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, and J. Hermans. in *Intermolecular Forces* (ed B. Pullman) 331 (Reidel, 1981).

25. H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma, "The Missing Term in Effective Pair Potentials," J. Phys. Chem. **91**(24), 6269-71, (1987).

26. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of Simple Potential Functions for Simulating Liquid Water," J. Chem. Phys. **79**(2), 926-35, (1983).

27. D. Laage, G. Stirnemann, F. Sterpone, R. Rey, and J. T. Hynes, "Reorientation and Allied Dynamics in Water and Aqueous Solutions," Annu. Rev. Phys. Chem. **62**(1), 395-416, (2011).

28. D. E. Moilanen, E. E. Fenn, Y.-S. Lin, J. L. Skinner, B. Bagchi, and M. D. Fayer, "Water Inertial Reorientation: Hydrogen Bond Strength and the Angular Potential," Proc. Natl. Acad. Sci. **105**(14), 5295-300, (2008).

29. P. Mark, and L. Nilsson, "Structure and Dynamics of the Tip3p, Spc, and Spc/E Water Models at 298 K," The Journal of Physical Chemistry A **105**(43), 9954-60, (2001).

30. A. J. Patel, P. Varilly, and D. Chandler, "Fluctuations of Water near Extended Hydrophobic and Hydrophilic Surfaces," J. Phys. Chem. B **114**(4), 1632-37, (2010).

31. A. J. Patel, P. Varilly, S. N. Jamadagni, M. F. Hagan, D. Chandler, and S. Garde, "Sitting at the Edge: How Biomolecules Use Hydrophobicity to Tune Their Interactions and Function," J. Phys. Chem. B **116**(8), 2498-503, (2012).

32. F. Pizzitutti, M. Marchi, F. Sterpone, and P. J. Rossky, "How Protein Surfaces Induce Anomalous Dynamics of Hydration Water," J. Phys. Chem. B **111**(26), 7584-90, (2007).

33. P. A. Srere, "Protein Crystals as a Model for Mitochondrial Matrix Proteins," Trends Biochem. Sci. **6**(1), 4-7, (1981).

34. A. B. Fulton, "How Crowded Is the Cytoplasm," Cell **30**(2), 345-47, (1982).

35. M. Buck, H. Schwalbe, and C. M. Dobson, "Main-Chain Dynamics of a Partially Folded Protein:15n Nmr Relaxation Measurements of Hen Egg White Lysozyme Denatured in Trifluoroethanol," J. Mol. Biol. **257**(3), 669-83, (1996).

36. J. F. Povey, C. M. Smales, S. J. Hassard, and M. J. Howard, "Comparison of the Effects of 2,2,2-Trifluoroethanol on Peptide and Protein Structure and Function," J. Struct. Biol. **157**(2), 329-38, (2007).

37. M. Buck, S. E. Radford, and C. M. Dobson, "A Partially Folded State of Hen Egg-White Lysozyme in Trifluoroethanol - Structural Characterization and Implications for Protein Folding," Biochemistry **32**(2), 669-78, (1993).

38. E. J. Arthur, J. T. King, K. J. Kubarych, and C. L. Brooks III, "Heterogeneous Preferential Solvation of Water and Trifluoroethanol in Homologous Lysozymes," J. Phys. Chem. B **118**(28), 8118-27, (2014).

39. J. E. Nielsen, and J. A. McCammon, "Calculating Pka Values in Enzyme Active Sites," Protein Science : A Publication of the Protein Society **12**(9), 1894-901, (2003).

40. D. Sali, M. Bycroft, and A. R. Fersht, "Stabilization of Protein Structure by Interaction of [Alpha]-Helix Dipole with a Charged Side Chain," Nature **335**(6192), 740-43, (1988).

41. B. Cannon, D. Isom, A. Robinson, J. Seedorff, and B. Garcia-Moreno, "Molecular Determinants of Pka Values of Internal Asp Residues," Biophys. J., 403A-03A, (2007).

42. G. Rabbani, E. Ahmad, N. Zaidi, S. Fatima, and R. Khan, "Ph-Induced Molten Globule State of Rhizopus Niveus Lipase Is More Resistant against Thermal and Chemical Denaturation Than Its Native State," Cell Biochem Biophys **62**(3), 487-99, (2012).

43. G. R. Wagner, and R. M. Payne, "Widespread and Enzyme-Independent Nϵ-Acetylation and Nϵ-Succinylation of Proteins in the Chemical Conditions of the Mitochondrial Matrix," J. Biol. Chem. **288**(40), 29036-45, (2013).

44. L. A. Baker, I. N. Watt, M. J. Runswick, J. E. Walker, and J. L. Rubinstein, "Arrangement of Subunits in Intact Mammalian Mitochondrial Atp Synthase Determined by Cryo-Em," Proc. Natl. Acad. Sci. **109**(29), 11675-80, (2012).

45. B. D. Cain, and R. D. Simoni, "Impaired Proton Conductivity Resulting from Mutations in the a Subunit of F1f0 Atpase in Escherichia Coli," J. Biol. Chem. **261**(22), 10043-50, (1986).

46. V. K. Rastogi, and M. E. Girvin, "Structural Changes Linked to Proton Translocation by Subunit C of the Atp Synthase," Nature **402**(6759), 263-68, (1999).

47. A. B. Clippingdale, J. D. Wade, and C. J. Barrow, "The Amyloid-B Peptide and Its Role in Alzheimer's Disease," Journal of Peptide Science **7**(5), 227-49, (2001).

48. C. M. Dobson, "Protein Folding and Misfolding," Nature **426**(6968), 884-90, (2003).

49. L. G. Cuello, D. M. Cortes, V. Jogini, A. Somporpisut, and E. Perozo, "A Molecular Mechanism for Proton-Dependent Gating in Kcsa," FEBS letters **584**(6), 1126-32, (2010).

50. E. E. Howell, J. E. Villafranca, M. S. Warren, S. J. Oatley, and J. Kraut, "Functional-Role of Aspartic Acid-27 in Dihydrofolate-Reductase Revealed by Mutagenesis," Science **231**(4742), 1123-28, (1986).

51. B. Aguilar, R. Anandakrishnan, J. Z. Ruscio, and A. V. Onufriev, "Statistics and Physical Origins of Pk and Ionization State Changes Upon Protein-Ligand Binding," Biophys. J. **98**(5), 872-80, (2010).

52. X. J. Kong, and C. L. Brooks III, "Λ-Dynamics: A New Approach to Free Energy Calculations," J. Chem. Phys. **105**(6), 2414-23, (1996).

53. M. S. Lee, F. R. Salsbury, and C. L. Brooks III, "Novel Generalized Born Methods," J. Chem. Phys. **116**(24), 10606-14, (2002).

54. E. J. Arthur, J. D. Yesselman, and C. L. Brooks III, "Predicting Extreme Pka Shifts in Staphylococcal Nuclease Mutants with Constant Ph Molecular Dynamics," Proteins: Struct., Funct., Bioinf. **79**(12), 3276-86, (2011).

55. L. S. Ahlstrom, S. M. Law, A. Dickson, and C. L. Brooks III, "Multiscale Modeling of a Conditionally Disordered Ph-Sensing Chaperone," J. Mol. Biol. **427**(8), 1670-80, (2015).

56. W. Im, M. S. Lee, and C. L. Brooks III, "Generalized Born Model with a Simple Smoothing Function," J. Comput. Chem. **24**(14), 1691-702, (2003).

57. J. T. King, M. R. Ross, and K. J. Kubarych, "Water-Assisted Vibrational Relaxation of a Metal Carbonyl Complex Studied with Ultrafast 2DIR," J. Phys. Chem. B **116**(12), 3754-59, (2012).

58. J. T. King, and K. J. Kubarych, "Site-Specific Coupling of Hydration Water and Protein Flexibility Studied in Solution with Ultrafast 2DIR Spectroscopy," J. Am. Chem. Soc. **134**(45), 18705-12, (2012).

59. C. A. Castaneda, C. A. Fitch, A. Majumdar, V. Khangulov, J. L. Schlessman, and B. E. Garcia-Moreno, "Molecular Determinants of the Pk(a) Values of Asp and Glu Residues in Staphylococcal Nuclease," Proteins: Struct., Funct., Bioinf. **77**(3), 570-88, (2009).

60. B. R. Brooks, C. L. Brooks III, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus, "Charmm: The Biomolecular Simulation Program," J. Comput. Chem. **30**(10), 1545-614, (2009).

61. D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods, "The Amber Biomolecular Simulation Programs," J. Comput. Chem. **26**(16), 1668-88, (2005).

62. P. Eastman, M. S. Friedrichs, J. D. Chodera, R. J. Radmer, C. M. Bruns, J. P. Ku, K. A. Beauchamp, T. J. Lane, L.-P. Wang, D. Shukla, T. Tye, M. Houston, T. Stich, C. Klein, M. R. Shirts, and V. S. Pande, "Openmm 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation," J. Chem. Theory Comput. **9**(1), 461-69, (2013).

63. B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, "Gromacs 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation," J. Chem. Theory Comput. **4**(3), 435-47, (2008).

64. J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, and K. Schulten, "Scalable Molecular Dynamics with Namd," J. Comput. Chem. **26**(16), 1781-802, (2005).

65. G. D. Hawkins, C. J. Cramer, and D. G. Truhlar, "Parametrized Models of Aqueous Free Energies of Solvation Based on Pairwise Descreening of Solute Atomic Charges from a Dielectric Medium," J. Phys. Chem. **100**(51), 19824-39, (1996).

66. V. Tsui, and D. A. Case, "Theory and Applications of the Generalized Born Solvation Model in Macromolecular Simulations," Biopolymers **56**(4), 275-91, (2000).

67. A. Onufriev, D. Bashford, and D. A. Case, "Exploring Protein Native States and Large-Scale Conformational Changes with a Modified Generalized Born Model," Proteins: Struct., Funct., Bioinf. **55**(2), 383-94, (2004).

Chapter 2

Site-Specific Dynamics of Water and Trifluoroethanol on Lysozyme Proteins

The work presented in this chapter has been published in the following papers:

1. J. T. King, E. J. Arthur, C. L. Brooks, III, and K. J. Kubarych, "Site-Specific Hydration Dynamics of Globular Proteins and the Role of Constrained Water in Solvent," The Journal of Physical Chemistry B **119**(19), 5604-5611 (2012).

2. J. T. King, E. J. Arthur, D. G. Osborn, C. L. Brooks, III, and K. J. Kubarych, "Biomolecular hydration dynamics probed with 2DIR spectroscopy: From dilute solution to a macromolecular crowd," Chinese Chemical Letters **26**(4), 435-438 (2015).

2.1 Introduction

Biological processes, from DNA replication to enzyme catalysis, occur in the presence of water. Water's indispensible role in biology has motivated efforts to uncover the degree to which it actively participates in chemical events.[2] As the universal solvent of living organisms, water has a remarkable ability to accommodate both hydrophilic solutes through strong electrostatic interactions, as well as hydrophobic solutes through subtle modifications to the hydrogen bonding network.[1,3-18] The hydration of large solutes (> 1 nm), such as

membranes and proteins, requires significant rearrangements of the hydrogen bonding network leading to the sacrifice of hydrogen bonds. Hydration water—water directly solvating the large solute—is thus structurally and dynamically constrained, restricting the configuration space as well as limiting dynamical flexibility. These constraints endow interfacial water with properties that are different from the bulk liquid. Whether or not one adopts a picture of protein dynamics as being "slaved" to the solvent, it is nevertheless clear that the preponderance of free energy changes attributable to the solvent arise from the relatively thin hydration layer of water solvating the protein.

The interest in studying and characterizing the properties of interfacial water arises from the extensive role that the protein-water interface plays in influencing such processes as small ligand binding, protein-protein recognition, and protein-DNA interactions.[7] Studies of orientational and spectral dynamics of water near lipid bilayers,[8] within reverse micelles,[9,17] or in the presence of small solutes indeed support the picture that limiting the configuration space can impose constraints on water's dynamics. Additionally, molecular dynamics simulations have been used extensively to study dynamics that may be difficult to access experimentally, such as the immediate hydration environments of proteins. Experimental evidence of water confinement near protein surfaces has been found by studying solvation dynamics of site-specific fluorescent probes of protein surfaces via ultrafast fluorescence upconversion. Recently, the combined constraining influence of both protein and lipids has enabled NMR measurements of local water structure and its mobility using reverse micelle-encapsulated ubiquitin. THz absorption spectroscopy has also been demonstrated to be a powerful technique for studying the hydration environments of proteins. Though these experiments provide evidence for constrained water, it remains unclear precisely which aspects of water's motion are most strongly affected by the interface.

In this chapter, we present evidence from ultrafast two-dimensional infrared spectroscopy that the primary dynamical distinction of hydration water is the protein's suppression of large-angle orientational jumps. The unique dynamics of water have been used previously to sense the presence of water using 2DIR through its influence on both vibrational lifetimes[31] and spectral dynamics.[32-34] Since our probe is able to identify regions of hydration while simultaneously distinguishing between constrained hydration water and bulk-like solvation, we are able to determine directly from experiment that an amphiphilic cosolvent (trifluoroethanol) preferentially dehydrates the protein in the region where the protein constrains the water dynamics. Our data also show that the cosolvent associates directly with the protein by replacing water in the hydration shell, rather than indirectly by disrupting the hydration layer from a distance. This detailed picture of the heterogeneous dynamics of "biological water" should provide a microscopic basis for a more complete understanding of interactions between domains in large proteins, as well as between proteins in large-scale assemblies and pathological aggregates.

Here, we present experimental evidence for constrained biological water solvating model enzymes, hen egg white lysozyme (HEWL) and human lysozyme (HuLys), using a vibrational probe of structure and dynamics of the interfacial water solvating the near-native protein. By leveraging the isotope dependence of the probe's vibrational relaxation in water (i.e. H_2O and D_2O) we are able to observe the influence of qualitatively distinct protein surfaces on the associated hydration dynamics. Using a relatively strong IR probe based on a transition metal carbonyl adduct, we are able to record 2DIR spectra with protein concentrations at the 100-200 μM level, which precludes complications due to the spatial coupling of hydration shells of other protein molecules in solution. The observed slowdown in water's dynamics is the result of protein surface-induced constraints placed on a subset of

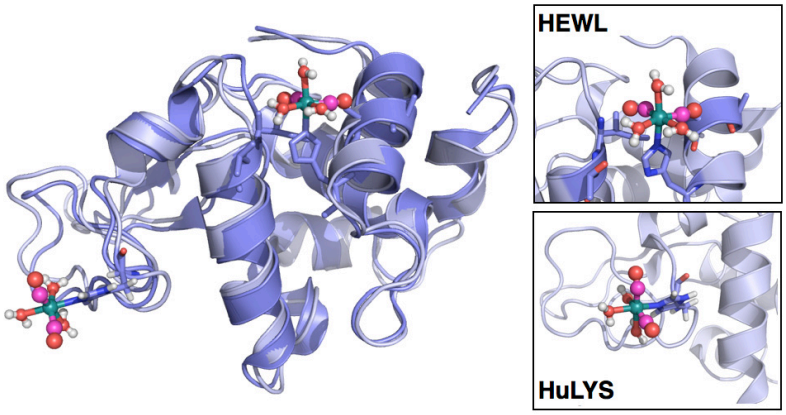

Figure 2.1 Crystal structures of HEWL-RC (PDB code 2XJW) overlaid with the crystal structure of native HuLys (PDB code 2ZIJ). The binding location of the metal carbonyl on the HEWL protein has been determined by X-ray crystallography. While no crystallographic data are available for the HuLys-RC complex, the binding location is proposed by comparison with the HEWL-RC complex.

water's fast dynamics, namely hydrogen bond switching events that occur through angular jumps, which has proven difficult to observe with other spectroscopic techniques.

We study the hydration environment of two homologous proteins, hen egg white lysozyme and human lysozyme. The crystal structure of the hen egg white lysozyme ruthenium carbonyl complex (HEWL-RC) shows a Ru-carbonyl complex bound to the lone His15 residue (Fig. 1).[34,36] While no crystal structure is available for the human lysozyme Ru-carbonyl complex (HuLys-RC), the structure is deduced by imposing the octahedral coordination found in HEWL-RC and by Fourier transform IR spectra which show identical carbonyl stretching frequencies for both HEWL-RC and HuLys-RC. HuLys has a single, solvent-exposed histidine residue (His78) which is the proposed binding location of the metal-carbonyl complex. There are several examples of metal carbonyl complexes binding to surface histidines. The linear and 2DIR spectra of HEWL-RC and HuLys-RC in D_2O are shown in Figure 2.3. The linear spectrum of HEWL-RC shows two small additional bands

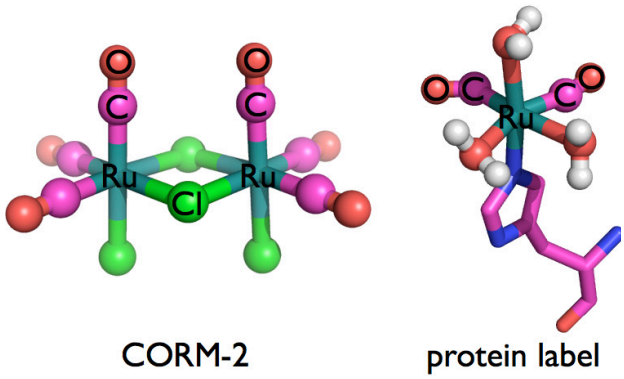

Figure 2.2 Structures of the vibrational chromophores used in this study. CORM-2 is used throughout the study as a model small molecule metal carbonyl. The key feature of the molecule that allows this comparison is the presence of multiple CO modes that are coupled, allowing for water-assisted vibrational relaxation.

corresponding to the low-population binding locations (Asp18 and Asp52) found in crystallography,[37] whereas the HuLys-RC shows only a single binding location. There is a slight (~1 cm^{-1}) shift in the vibrational frequencies of the two carbonyl modes, consistent with the metal center being coordinated to a histidine residue in both cases but having different local protein environments. In HEWL, His15 is in the highly structured α domain, whereas His78 of HuLys is located in the unstructured β domain (Figure 2.1). We note, however, that the unstructured domains of HEWL and HuLys are structurally similar.

The binding motif of the vibrational label to the proteins is shown in Figure 2.2. We also rely on comparisons with the small molecule dichloro-ruthenium(II) dimer (a so-called "carbon monoxide releasing molecule" often denoted CORM-2), which is the precursor to the labeling complex (referred to as CORM-3).[38] Because of the scarcity of water-soluble metal carbonyls, we rely on comparisons between the labeled proteins and the CORM-2 complex in both aqueous and organic solvents. While the molecules are clearly different, the comparison between these molecules is both robust and instructive. The crucial properties shared between these molecules are the presence of coupled CO chromophores as well as

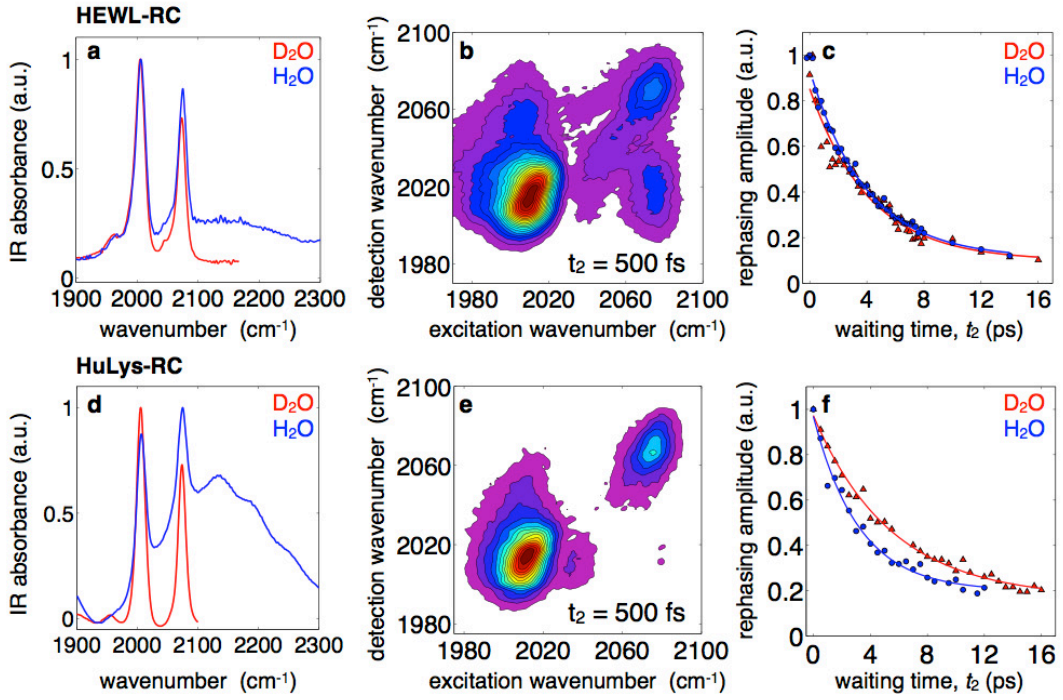

Figure 2.3 Linear FTIR spectra of HEWL-RC (a) and HuLys-RC (d) in D_2O and H_2O. The broad feature in the H_2O spectrum is the bend- libration combination band, centered at 2150 cm^{-1}. The 2DIR rephasing spectra for HEWL-RC (b) and HuLys-RC (e) in D_2O are shown for a waiting time of t2 = 500 fs. Monitoring the amplitude of the 2004 cm^{-1} peak as a function of waiting time, t2, provides the vibrational lifetime of the mode. For HEWL-RC there is no observable isotope effect in the vibrational relaxation between D_2O and H_2O (c), whereas HuLys-RC shows a very clear isotope effect (f). The lack of an isotope effect suggests solvation by slow constrained water, whereas hydration by bulk-like water leads to an observable isotope effect. These results demonstrate the heterogeneous nature of the water dynamics near a protein, where certain regions are hydrated by slow constrained water while other regions are hydrated by bulk-like water.

the presence H_2O of numerous low-frequency modes. The anharmonic coupling between the spectroscopic modes and the lower frequency modes of the molecule results in water-assisted relaxation in aqueous environments, a key aspect of the results and interpretations presented here.[39] The side chains that are in the immediate vicinity of the CO oscillators on HEWL-RC are isoleucine, phenylalanine, and alanine residues, which are nonpolar residues, as well as an arginine residue. The HuLys-RC probe is mostly exposed to the solvent, though it is neighbored by cysteine, leucine, and alanine. While the environment presented by the

protein is an important aspect of the dynamics felt by the vibrational probes, the observed lifetimes are dominated by the hydration water.

Ultrafast 2DIR spectroscopy is used to study the hydration environments of HEWL-RC and HuLys-RC in pure water solvent (either H_2O or D_2O), as well as in solvent mixtures of D_2O and 2,2,2-trifluoroethanol (TFE) ranging from 0 to 20% TFE v/v. Because of the structural similarities of HEWL and HuLys (60% sequence homology, Cα rmsd = 1.1 Å), the two labeling locations, though occurring on different proteins, sample the heterogeneous protein structure as well as distinct solvation environments. The vibrational lifetime (T1) of the metal carbonyl probe is used as a reporter of the local solvation environment at the interfacial region of the protein. The lifetime is sensitive to the presence of water and has been shown to be an order of magnitude shorter in water (H_2O or D_2O)[38] than in either proteinaceous environments[40] or in polar organic solvents. Thus, the vibrational lifetime effectively acts as a water sensor positioned at the protein−water interface.

2.2 Water-Assisted Vibrational Relaxation

The sub-5- ps absolute vibrational lifetime of the CO modes reports on the presence of liquid water as the principal pathway for vibrational relaxation. We have previously shown that the vibrational lifetimes of metal-bound carbonyls are on the order of 50−100 ps.[38] Even in the highly polar solvent methanol, we find the vibrational lifetime of the small CORM-2 complex to be 42.25 ± 3 ps.[41] In water, however, the vibrational lifetime of CORM-2 is an order of magnitude smaller, an effect attributed to the high density of vibrational states in which to dissipate energy as well as the extremely rapid fluctuations of charge, both hallmarks of water solvation.[1] Water acts to facilitate the intramolecular

coupling of the solute vibrational degrees of freedom. The observed vibrational lifetimes of the protein-bound metal carbonyls also exhibit lifetimes on the order of 3−4 ps (Figure 2.3)[1,42-45] suggesting that the chromophores are sensitive to the interfacial water, which provides the dominant relaxation pathway. This conclusion, that the vibrational relaxation is sensitive mainly to the water hydrating the protein, was further verified using D_2O−TFE solvent exchange discussed in detail below, where we find that replacing the hydration water with an alcohol cosolvent results in a pronounced increase in the vibrational lifetime.

2.3 Constrained Water at the Protein Surface

The thermodynamic driving forces for hydrating small and large hydrophobic cavities differ according to the relative significance of enthalpic and entropic contributions. Small hydrophobes and small ions generally sustain water's local hydrogen bonding network through subtle rearrangements, so that free energy gradients arise from changes in entropy. Conversely, large hydrophobes disrupt hydrogen bonding, leading to driving forces dominated by enthalpic changes. Hence, one expects dynamical perturbations to reflect these distinct underlying free energy landscapes.

Figure 2.3 shows the Fourier transform IR (FTIR) spectrum of HEWL-RC in H_2O and D_2O. The vibrational probe has two IR-active CO modes located at 2004 and 2080 cm^{-1}. We focus on the low-frequency mode of both HEWL-RC and HuLys-RC for analysis. Using 2DIR spectroscopy, the vibrational lifetimes of the CO vibrational modes of HEWL-RC in H_2O and D_2O were extracted for the 2004 cm^{-1} mode and found to be 3.60 ± 0.18 and 3.73 ± 0.21 ps, respectively. This result is in stark contrast to what has previously been reported for water-assisted vibrational relaxation, where we observed pronounced isotope differences

between water and heavy water.[47]

The loss of the isotope effect can be explained in terms of the restraints that large, hydrophobic surfaces place on water's hydrogen bonding structure and dynamics. Comprising a subset of water's fast dynamics are hydrogen bonding switching events, which have been theoretically predicted[33] and experimentally supported[10] as occurring through abrupt angular jumps that involve large-scale motion of the hydrogen (or deuterium) atoms of the water.[32,33,48-50] Small molecules at low concentrations do not disrupt hydrogen bonding networks and, more importantly, do not significantly limit the configuration space available to hydrogen bond partners, allowing this subset of water's dynamics to occur unperturbed. Because the angular jump dynamics of water involve large displacements of the hydrogen atoms, these dynamics should also be particularly sensitive to isotope substitution. Hence, the solvent fluctuations that drive vibrational relaxation strongly reflect the dynamical differences between H_2O and D_2O. In fact, the water isotope effect on solvation dynamics had been successfully modeled from the perspective of Debye relaxation, which relates the macroscopic dielectric constant of water to the microscopic reorientation dynamics,[10] though the angular jump mechanism had not yet been identified.

The dynamical constraints exerted by extended hydrophobic surfaces on the surrounding water arise from the restrictions imposed on the hydrogen bonding network by the surface.[8] Extended surfaces limit both the configuration space available for hydrogen bonding as well as the associated dynamics, causing water to adopt geometries that are not favorable for hydrogen bond coordination while impeding switching events.[17,31,32] The structured region surrounding the HEWL His15 label is an excellent example of a natural

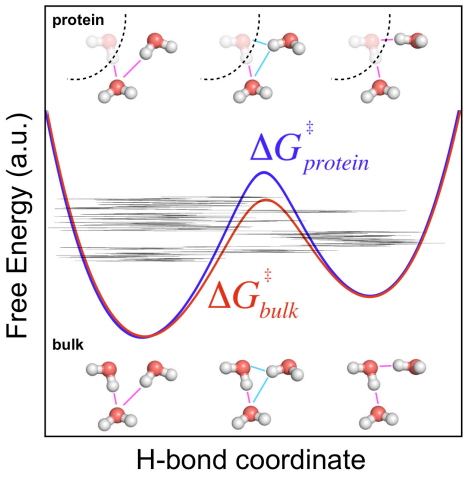

Figure 2.4 Cartoon depicting the free energy surface for hydrogen bond jumps. The transition state has been identified as a bifurcated hydrogen bond with both initial and final donors (shown with cyan hydrogen bonds). In regions of constrained hydration, the protein limits the availability of final donors, raising the free energy barrier by decreasing its entropy. Besides the relatively rare jumping events, the rapid intrawell fluctuations are able to induce enhanced anharmonic coupling, assisting vibrational relaxation for both water isotopes.

extended biological surface with low curvature, hence the surface slows down the water dynamics by limiting the available partners for fast hydrogen bond switching while inhibiting the required coordinated reorientation that accompanies large angular jumps. Because the switching events involve large displacements of the hydrogen atoms, these motions contribute significantly to the measured vibrational relaxation isotope effect, which is only observed when the hydrating water exhibits bulk-like dynamics. This interpretation of "hydrophobic slowing" of water's dynamics is consistent with what has been previously observed for small solutes at high concentrations, where neighboring solutes limit hydrogen bonding switching.

A cartoon depiction of hydrogen bond switching and its modification by the protein surface are shown in Figure 2.4. Since successful hydrogen bond switching events proceed through a bifurcated transition state where the switching hydrogen is fleetingly associated with both the initial and final partner O atoms, the free energy barrier is necessarily

influenced by the availability of such configurations. Relative to the bulk liquid, the protein interface deprives water molecules of potential partners, which reduces the availability of transition state candidates and lowers the entropy of the transition state. Nevertheless, hydrogen bond jumps are not the only source of environmental fluctuations leading to enhanced anharmonic coupling and the resulting carbonyl vibrational relaxation. The intrawell dynamics comprise the majority of these fluctuations (depicted by the stochastic trajectory in the cartoon), hence resulting in similarly rapid relaxation in both D_2O and H_2O.

The absolute value of the constrained H_2O/D_2O relaxation falls between the values of bulk-like H_2O and D_2O for both CORM-2 as well as HuLys-RC. While the dynamical nature of the solvent can be influential, it is only one component that determines the vibrational lifetime. The electric field generated by the solvent applies the force on the relaxing mode and is thus an important component of vibrational relaxation that we cannot probe directly. Because hydrophobic hydration is accompanied by dynamical and structural changes, the absolute lifetime observed for the constrained H_2O/D_2O will depend on any structural changes that occur in the hydration layer. Thus, the convergence of the H_2O and D_2O relaxation onto a single lifetime and the absolute value of the vibrational lifetime should be considered somewhat separately.

2.4 Heterogeneous Water Environments

The homologous structures of HEWL and HuLys allow us to investigate the water dynamics near two qualitatively distinct protein−water environments (Figure 2.1). We have previously discussed the lack of an observable isotope dependence of the vibrational relaxation of the HEWL-RC complex, where the probe is located on a structured, extended

protein surface. In the HuLys-RC complex, however, the probe is located in an unstructured and flexible region of the protein. In contrast to HEWL-RC, the isotope effect is clearly observed in HuLys-RC, where the relaxation time constants for the 2004 cm^{-1} mode are 3.12 $\pm$ 0.26 ps and 4.70 $\pm$ 0.38 ps in H_2O and D_2O, respectively. Despite being located at the protein surface, the measured solvation dynamics appears more consistent with small molecule hydration. Given that the vibrational probe is attached to a histidine residue in both proteins, the data indicate that some degree of collectivity at each site leads to the protein's heterogeneous influence on the hydrating water, as well as highlighting the role of surface topology on local hydrophobicity.[15]

The water dynamics surrounding the unstructured region of the HuLys-RC complex resembles what was previously observed for a small metal carbonyl, CORM-2, at low concentrations (~2 mM). This similarity suggests that the solvation of the unstructured region of the protein is similar to what is seen for a small molecule, namely, that the hydration environment is essentially bulk-like. The picture that emerges from these measurements is that a protein's ability to constrain hydration water dynamics is determined not only by the availability of solvent-exposed side chains capable of forming hydrogen bonds but also by the presence of a low-curvature surface topology. Though this view is consistent with the prevailing model of hydrophobic solvation,[20-24,53] our work shows clearly how a single, relatively compact globular protein can exhibit both extremes of hydration structure and dynamics.

The heterogeneous nature of the hydration dynamics of a protein raises interesting questions regarding the role, if any, of the dynamically constrained water in biological

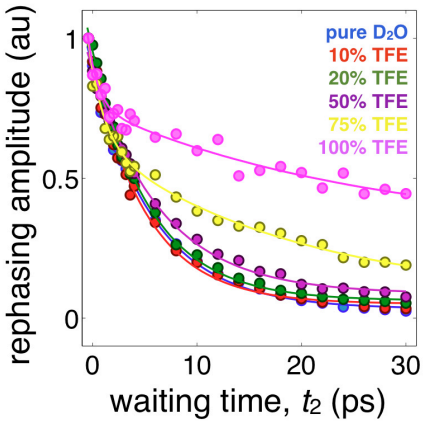

Figure 2.5 Vibrational relaxation of CORM-2 in D_2O/TFE mixtures, demonstrating the lifetime dependence on the cosolvent in the absence of preferential solvation.

processes. It has long been speculated that hydrophobic hydration—hydration environments that constrain water—leads to entropic driving forces for surface processes, all of which require protein dehydration as the initial step. A region of hydrophobic hydration can act as a "thermodynamic reservoir", where entropy is created by relaxing constraints on the hydrating water, in turn enabling greater participation in enthalpically favorable hydrogen bonding. We examine hydrophobic assembly below using an amphiphilic alcohol cosolvent.

2.5 D_2O-TFE Solvent Exchange

The tight interplay between protein dynamics and the hydration environment suggests that modulations can significantly impact a protein's dynamics, structure, and stability. The properties of a protein can be manipulated by adding small amounts of cosolvents, such as alcohols. Low concentrations of 2,2,2-trifluoroethanol (TFE), for example, can stabilize protein secondary structure through a mechanism that is generally attributed to preferential solvation of the protein by TFE, promoting intramolecular

hydrogen bonding within the protein by alleviating competition with external hydrogen bonding partners from hydrating water. At higher concentrations, however, lacking the driving force of hydrophobicity, the protein becomes unstable and partially denatures into an unfolded state characterized by a loosening of the helix packing even as the helices themselves remain stabilized.[41] Partial unfolding in lysozyme has been observed at TFE concentrations near 15% (v/v). The linear FTIR spectra of HEWL-RC in D_2O/TFE mixtures show no significant changes in either the amide region of the spectrum or the metal carbonyl stretch bands.[55]

To investigate the thermodynamic connection between constrained water and the driving force for surface processes such as preferential dehydration, we studied the influence of the amphiphilic cosolvent TFE on the vibrational lifetime of the protein-bound vibrational probe. We have shown that the dominant pathway of vibrational relaxation for the protein-bound probes is driven by the interfacial water dynamics. Therefore, dehydrating the protein surface surrounding the probe should result in measurable changes to its vibrational lifetime.

As a control experiment, we measured the vibrational lifetimes of CORM-2 in a series of D_2O/TFE mixtures (including pure D_2O, 10, 20, 50, and 75% TFE, and pure TFE). Figure 2.5 shows the vibrational relaxation of CORM-2 in the D_2O/TFE mixtures. At low concentrations, the vibrational lifetime remains dominated by water-assisted vibrational relaxation, only increasing from 4 to 6 ps over a range of 0−50% TFE. At higher concentrations, the relaxation becomes dominated by the TFE cosolvent, thereby increasing to 25 and 50 ps at 75% and pure TFE, respectively. It is clear that there is a nonlinear dependence of the vibrational lifetime on the solvent composition, likely due to the dominance of water-assisted vibrational relaxation as the most efficient relaxation pathway.

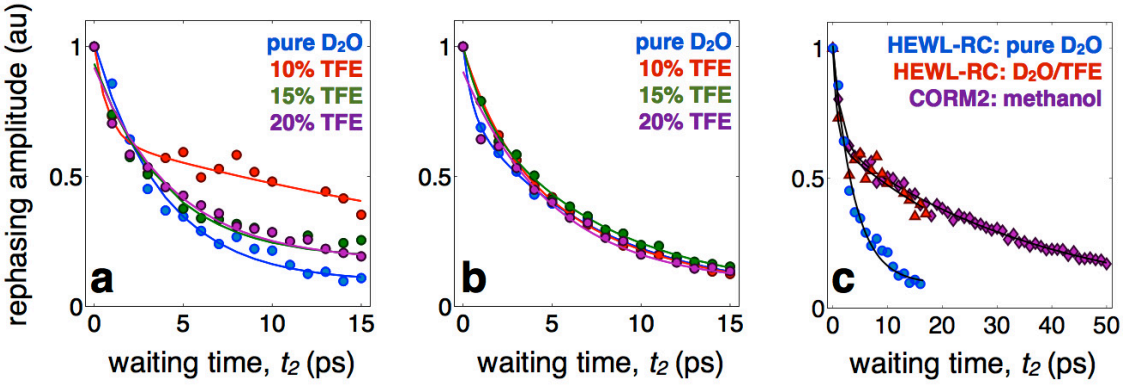

Figure 2.6 Vibrational relaxation for HEWL-RC (a) and HuLys-RC (b) in D2O/TFE mixtures ranging from 0% to 20% TFE v/v. The addition of small amounts of TFE results in a large increase in the vibrational lifetime of HEWL-RC, followed by a monotonic decrease upon further addition. The increase in lifetime at low concentrations is the result of preferential solvation, and the subsequent decrease in lifetime is the result of the onset of partial protein destabilization. In contrast, HuLys-RC shows no sensitivity to TFE, suggesting this region of the protein resists solvent exchange with TFE and remains hydrated. (c) A comparison of the cosolvent-dependent relaxation for HEWL-RC (circles) and CORM-2 (triangles) shows that at 10% TFE HEWL-RC indicates a local solvation environment with nearly no water, with a relaxation time scale similar to other metal carbonyls in alcohol environments.

Hence, significant changes in the vibrational lifetime are only observed when water is at a very low concentration. These data provide a baseline for vibrational lifetimes in D_2O/TFE mixtures in the absence of preferential solvation, which can be applied to the study of HEWL-RC and HuLys-RC in the presence of TFE.

Figure 2.6a shows the vibrational relaxation of HEWL-RC for four different TFE concentrations (0, 10, 15, 20% v/v). In pure D_2O, the vibrational lifetime is 3.73 ± 0.21 ps. Upon addition of 10% TFE the vibrational lifetime increases to 32.76 ± 1.15 ps,[56] suggesting that at low concentrations the alcohol dehydrates the protein near the vibrational probe in exchange for a preferred alcohol environment. Lacking water, the vibrational relaxation becomes significantly slower and resembles relaxation observed in CORM-2 in TFE environments (Figure 2.6c). In comparison to the T1 times for CORM-2 in D_2O/TFE

mixtures, the HEWL-RC surface surrounding the vibrational probe has a solvation composition that resembles a solution between 75% TFE and pure TFE, clearly showing there is a lack of water at the protein surface. The vibrational lifetime achieved through only the addition of 10% cosolvent provides clear evidence that the TFE is preferentially drawn to the protein at the hydrophobic region.

Further addition of TFE induces a decrease of the vibrational lifetime, resulting in relaxation times that reflect a homogeneous solution of water and TFE (15% TFE T1 = 3.99 ± 0.60, 20% TFE T1 = 4.41 ± 0.48 ps). This decrease in vibrational lifetime, which returns to characteristic time scales for water-assisted relaxation by 20% TFE, warrants additional discussion. This experimental observation, that preferential solvation at low TFE concentration is not sustained at higher TFE concentrations, suggests the emergence of structural instability of the protein at TFE concentrations above 10%. Previous reports using a combination of spectroscopic techniques have shown that TFE concentrations near 15% can promote significant structural changes, including some destabilization of protein tertiary structure.[54] Earlier work by Dobson using circular dichroism found TFE enhanced the overall helical content of the protein, but at the cost of destabilization of tertiary structure.[55,57-62] Our experimental results are consistent with helical portions of the protein being susceptible to dehydration and interactions with the hydrophobic portions of TFE. The decreased lifetime is consistent with the following scenario: As the constrained water is relieved and the protein alters its structure, the solvation environment becomes a mixture of D_2O/TFE as the collective influence of the extended hydrophobic surface is disrupted due to the loosened helix packing.

The mechanism by which small molecules denature proteins has been, and remains, an area of intense research. The present data suggest that there is a direct interaction

between the protein surface and TFE, leading to the formation of a dehydrated interface between the protein and the cosolvent. This cosolvent shell in turn can modify the limited water dynamics at the surface by supplying hydrogen bonding partners through the alcohol's hydroxyl group. This interpretation would be consistent with a mixed direct and indirect mechanism, where the cosolvent, directed to regions of constrained water, essentially coats the protein surface, promoting intraprotein hydrogen bonding and stabilizing secondary structure. Cosolvent association destabilizes the tertiary contacts between helices once the protein becomes so dehydrated that it loses the hydrophobic driving force to fold, resulting in partial denaturation. This picture of TFE-modulated lysozyme stability is consistent with thermodynamic measurements based on calorimetry and structural studies using NMR spectroscopy.

While we observe that the structured region of the HEWL-RC complex leads to constrained water that can drive solvent exchange, the unstructured β-region of HuLys is solvated by bulk-like water, suggesting that this region would not experience substantial solvent exchange. Figure 2.6b shows the vibrational relaxation of HuLys-RC in TFE/D_2O solution. Indeed, the vibrational relaxation of the label at this site shows no dependence on TFE, indicating that this location resists preferential dehydration by TFE and shows a solvation environment that might be expected for a simple mixture of D_2O/TFE (Figure 2.5). Comparing the experimental observations of HEWL-RC and HuLys-RC, it is clear that the interaction of TFE with the protein depends, to some degree, on the extended properties of the surface and not simply on individual amino acid residues since the vibrational probe is attached to a histidine residue in both cases.

The correlation between constrained water and solvent exchange demonstrates how the release of dynamically constrained water can drive hydrophobic association. It is known,

however, that for many association processes the entropic contribution is insufficient to account for the total change in free energy.[1] While the hydrophobic interaction between the protein surface and the hydration environment is indeed the driving force for such processes, its influence is not limited to entropic changes associated with liberating the water's constraints[65] since expelling hydration water affords enthalpic gains by restoring hydrogen bonding that is diminished near extended surfaces.[66] Moreover, since many macromolecular assembly processes are kinetically controlled, the time required to allow for the diffusive liberation of constrained water may be too long given that the approaching extended hydrophobic surfaces are both solvated by water with diffusivity that is lower than the bulk.

2.6 Conclusions

The results presented here provide a site-specific probe of heterogeneous hydration dynamics of large proteins in pure H_2O and D_2O and, more importantly, provide experimental evidence of the mechanism of the hydration slowdown. A key aspect of this work is the study of labeled proteins at micromolar concentrations, which allows an unobstructed observation of the influence of the protein surface on hydration water. The results indicate that a single lysozyme protein is capable of influencing its hydration environment. This result is to be contrasted to numerous other studies that rely on high concentrations of solute to observe slowed water dynamics, where the crowding of multiple solutes can cooperatively constrain the hydration water. Although crowding is a central aspect of in vivo chemical biology, it is essential to characterize a single protein's influence over its hydration environment in addition to the specific or nonspecific perturbations

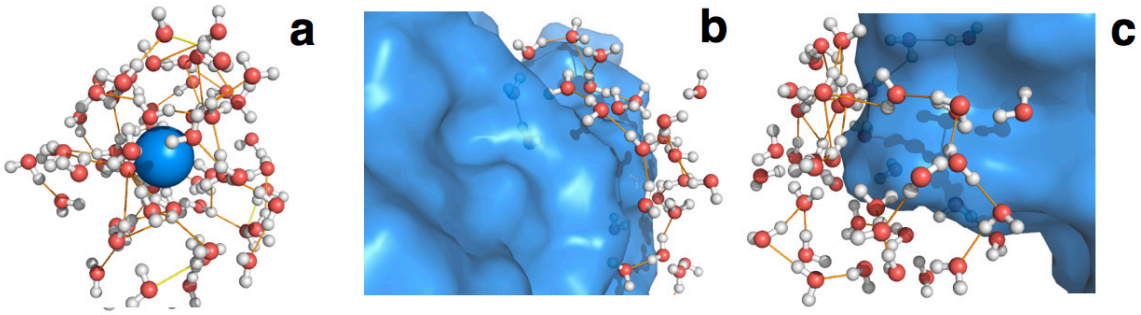

Figure 2.7 Cartoon demonstrating the effect of extended surfaces on hydrogen bonding switching events. Small molecules do not perturb the hydrogen bonding networks at small concentrations (a), while extended protein surfaces, like the surface found near the vibrational probe on HEWL-RC (b), can limit the hydrogen bonding network and the hydrogen bonding switching events. Loose, unstructured regions of proteins, like that surrounding the probe on HuLys-RC (c), act more like a collection of small molecules, where bulk-like dynamics can be preserved.

induced by crowders. Evidence for constrained water is found near the large, structured α domain of HEWL (Figure 7b), where the individual amino acid residues act in a cooperative manner to create an extended hydrophobic surface, depriving water of hydrogen bonding partners. In contrast, bulk-like water is found on the unstructured region of HuLys (Figure 7c), where the residues behave as independent solutes with a hydration environment resembling that of a small molecule. Around these individual residues water retains a bulk-like hydrogen bonding network, and the dynamics are not suppressed. It is important to note that it is precisely this unstructured and flexible region that acts as a flap over the substrate binding site.

In addition to the heterogeneous nature of hydration dynamics surrounding large globular proteins, this study also reveals the correlation between dynamically constrained water and the driving force for site-specific association at the protein surface. The free energy that is released upon dehydration of constrained water appears to be sufficient to drive the association of small molecules to the protein surface. Using a water sensing vibrational probe, we can distinguish between direct cosolvent–protein association and

indirect disruption of the hydration layer. On the basis of the marked changes in the vibrational lifetime, our data are consistent with direct displacement of water from the protein surface. This view is further supported by the subsequent cosolvent-induced destabilization caused by competing out the water to such an extent that the protein's tertiary structure loosens, evidenced by the infiltration of water in the absence of the structured, extended hydrophobic surface. Taken together, our data indicate that the spatially heterogeneous dynamics of hydration water is, to a significant degree, responsible for site-directed hydrophobic association, a perspective that should be helpful in rationalizing and perhaps in guiding the controlled disruption of deleterious protein−protein interactions.

Water's importance in biology cannot be overstated, but ample evidence shows that often only a small amount of water is truly necessary for function.[1,2] Water-sensitive vibrational probes on the surfaces of proteins will enable an experimental platform to systematically map interactions between proteins and other biomacromolecules including DNA and antibodies while simultaneously monitoring the role (or lack thereof) played by the thin layer of hydration water.

2.7 References

1. D. Chandler, "Interfaces and the Driving Force of Hydrophobic Assembly," Nature **437**(7059), 640-47, (2005).

2. P. Ball, "Water as an Active Constituent in Cell Biology," Chem. Rev. **108**(1), 74-108, (2008).

3. L. F. Scatena, M. G. Brown, and G. L. Richmond, "Water at Hydrophobic Surfaces: Weak Hydrogen Bonding and Strong Orientation Effects," Science **292**(5518), 908-12, (2001).

4. B. Bagchi, "Water Dynamics in the Hydration Layer around Proteins and Micelles," Chem. Rev. **105**(9), 3197-219, (2005).

5. L. R. Chieffo, J. T. Shattuck, E. Pinnick, J. J. Amsden, M. K. Hong, F. Wang, S. Erramilli, and L. D. Ziegler, "Nitrous Oxide Vibrational Energy Relaxation Is a Probe of Interfacial Water in Lipid Bilayers," J. Phys. Chem. B **112**(40), 12776-82, (2008).

6. E. E. Fenn, D. B. Wong, and M. D. Fayer, "Water Dynamics at Neutral and Ionic Interfaces," Proc. Natl. Acad. Sci. U.S.A. **106**(36), 15243-48, (2009).

7. M. D. Fayer, and N. E. Levinger. in *Annual Review of Analytical Chemistry, Vol 3* Vol. 3 *Annual Review of Analytical Chemistry* (eds E. S. Yeung, and R. N. Zare) 89-107 (2010).

8. A. A. Bakulin, C. Liang, T. L. C. Jansen, D. A. Wiersma, H. J. Bakker, and M. S. Pshenichnikov, "Hydrophobic Solvation: A 2DIR Spectroscopic Inquest," Accounts Chem. Res. **42**(9), 1229-38, (2009).

9. F. Pizzitutti, M. Marchi, F. Sterpone, and P. J. Rossky, "How Protein Surfaces Induce Anomalous Dynamics of Hydration Water," J. Phys. Chem. B **111**(26), 7584-90, (2007).

10. G. Stirnemann, P. J. Rossky, J. T. Hynes, and D. Laage, "Water Reorientation, Hydrogen-Bond Dynamics and 2DIR Spectroscopy Next to an Extended Hydrophobic Surface," Faraday Discuss. **146**, 263-81, (2010).

11. S. K. Pal, J. Peon, and A. H. Zewail, "Biological Water at the Protein Surface: Dynamical Solvation Probed Directly with Femtosecond Resolution," Proc. Natl. Acad. Sci. U.S.A. **99**(4), 1763-68, (2002).

12. W. Qiu, Y.-T. Kao, L. Zhang, Y. Yang, L. Wang, W. E. Stites, D. Zhong, and A. H. Zewail, "Protein Surface Hydration Mapped by Site-Specific Mutations," Proc. Natl. Acad. Sci. U.S.A. **103**(38), 13979-84, (2006).

13. L. Y. Zhang, L. J. Wang, Y. T. Kao, W. H. Qiu, Y. Yang, O. Okobiah, and D. P. Zhong, "Mapping Hydration Dynamics around a Protein Surface," Proc. Natl. Acad. Sci. U.S.A. **104**(47), 18461-66, (2007).

14. N. V. Nucci, M. S. Pometun, and A. J. Wand, "Site-Resolved Measurement of Water-Protein Interactions by Solution Nmr," Nat. Struct. Mol. Biol. **18**(2), 245-49, (2011).

15. N. V. Nucci, M. S. Pometun, and A. J. Wand, "Mapping the Hydration Dynamics of Ubiquitin," J. Am. Chem. Soc. **133**(32), 12326-29, (2011).

16. G. Stirnemann, J. T. Hynes, and D. Laage, "Water Hydrogen Bond Dynamics in Aqueous Solutions of Amphiphiles," J. Phys. Chem. B **114**(8), 3052-59, (2010).

17. Y. L. A. Rezus, and H. J. Bakker, "Observation of Immobilized Water Molecules around Hydrophobic Groups," Phys. Rev. Letters **99**(14), 4, (2007).

18. H. Frauenfelder, G. Chen, J. Berendzen, P. W. Fenimore, H. Jansson, B. H. McMahon, I. R. Stroe, J. Swenson, and R. D. Young, "A Unified Model of Protein Dynamics," Proc. Natl. Acad. Sci. U.S.A. **106**(13), 5129-34, (2009).

19. P. W. Fenimore, H. Frauenfelder, B. H. McMahon, and R. D. Young, "Bulk-Solvent and Hydration-Shell Fluctuations, Similar to Alpha- and Beta-Fluctuations in Glasses, Control Protein Motions and Functions," Proc. Natl. Acad. Sci. U.S.A. **101**(40), 14408-13, (2004).

20. I. Halperin, B. Y. Ma, H. Wolfson, and R. Nussinov, "Principles of Docking: An Overview of Search Algorithms and a Guide to Scoring Functions," Proteins: Struct., Funct., Genet. **47**(4), 409-43, (2002).

21. B. Q. Q. Wei, W. A. Baase, L. H. Weaver, B. W. Matthews, and B. K. Shoichet, "A Model Binding Site for Testing Scoring Functions in Molecular Docking," J. Mol. Biol. **322**(2), 339-55, (2002).

22. S. Jones, and J. M. Thornton, "Principles of Protein-Protein Interactions," Proc. Natl. Acad. Sci. U.S.A. **93**(1), 13-20, (1996).

23. G. A. Papoian, J. Ulander, and P. G. Wolynes, "Role of Water Mediated Interactions in Protein-Protein Recognition Landscapes," J. Am. Chem. Soc. **125**(30), 9170-78, (2003).

24. B. Jayaram, and T. Jain, "The Role of Water in Protein-DNA Recognition," Annu. Rev. Biophys. Biomol. Struct. **33**, 343-61, (2004).

25. C. L. Brooks, III, and M. Karplus, "Solvent Effects on Protein Motion and Protein Effects on Solvent Motion - Dynamics of the Active-Site Region of Lysozyme," J. Mol. Biol. **208**(1), 159-81, (1989).

26. V. Makarov, B. M. Pettitt, and M. Feig, "Solvation and Hydration of Proteins and Mucleic Acids: A Theoretical View of Simulation and Experiment," Accounts Chem. Res. **35**(6), 376-84, (2002).

27. M. Heyden, and M. Havenith, "Combining Thz Spectroscopy and Md Simulations to Study Protein-Hydration Coupling," Methods **52**(1), 74-83, (2010).

28. G. Niehues, M. Heyden, D. A. Schmidt, and M. Havenith, "Exploring Hydrophobicity by Thz Absorption Spectroscopy of Solvated Amino Acids," Faraday Discuss. **150**, 193-207, (2011).

29. C. T. Middleton, L. E. Buchanan, E. B. Dunkelberger, and M. T. Zanni, "Utilizing Lifetimes to Suppress Random Coil Features in 2DIR Spectra of Peptides," J. Phys. Chem. Lett. **2**(18), 2357-61, (2011).

30. Y. S. Kim, L. Liu, P. H. Axelsen, and R. M. Hochstrasser, "2DIR Provides Evidence for Mobile Water Molecules in Beta-Amyloid Fibrils," Proc. Natl. Acad. Sci. U.S.A. **106**(42), 17751-56, (2009).

31. D. Laage, and J. T. Hynes, "A Molecular Jump Mechanism of Water Reorientation," Science **311**(5762), 832-35, (2006).

32. D. Laage, G. Stirnemann, and J. T. Hynes, "Why Water Reorientation Slows without Iceberg Formation around Hydrophobic Solutes," J. Phys. Chem. B **113**(8), 2428-35, (2009).

33. F. Sterpone, G. Stirnemann, and D. Laage, "Magnitude and Molecular Origin of Water Slowdown Next to a Protein," J. Am. Chem. Soc. **134**(9), 4116-19, (2012).

34. T. Santos-Silva, A. Mukhopadhyay, J. D. Seixas, G. J. L. Bernardes, C. C. Romao, and M. J. Romao, "CORM-3 Reactivity toward Proteins: The Crystal Structure of a Ru(Ii) Dicarbonyl-Lysozyme Complex," J. Am. Chem. Soc. **133**(5), 1192-95, (2011).

35. A. M. Blanco-Rodriguez, M. Busby, C. Gradinaru, B. R. Crane, A. J. Di Bilio, P. Matousek, M. Towrie, B. S. Leigh, J. H. Richards, A. Vlcek, and H. B. Gray, "Excited-State Dynamics of Structurally Characterized Re-I(Co)(3)(Phen)(Hisx) (+) (X=83,109) Pseudomonas Aeruginosa Azurins in Aqueous Solution," J. Am. Chem. Soc. **128**(13), 4365-70, (2006).

36. S. L. Binkley, C. J. Ziegler, R. S. Herrick, and R. S. Rowlett, "Specific Derivatization of Lysozyme in Aqueous Solution with Re(CO)(3)(H_2O)(3)(+)," Chem. Commun. **46**(8), 1203-05, (2010).

37. R. Motterlini, and L. E. Otterbein, "The Therapeutic Potential of Carbon Monoxide," Nat. Rev. Drug Discovery **9**(9), 728-U24, (2010).

38. J. T. King, M. R. Ross, and K. J. Kubarych, "Water-Assisted Vibrational Relaxation of a Metal Carbonyl Complex Studied with Ultrafast 2DIR," J. Phys. Chem. B **116**(12), 3754-59, (2012).

39. C. Ventalon, J. M. Fraser, M. H. Vos, A. Alexandrou, J. L. Martin, and M. Joffre, "Coherent Vibrational Climbing in Carboxyhemoglobin," Proc. Natl. Acad. Sci. U.S.A. **101**(36), 13216-20, (2004).

40. J. T. King, J. M. Anna, and K. J. Kubarych, "Solvent-Hindered Intramolecular Vibrational Redistribution," Phys. Chem. Chem. Phys. **13**(13), 5579-83, (2011).

41. J. T. King, E. J. Arthur, C. L. Brooks, III, and K. J. Kubarych, "Site-Specific Hydration Dynamics of Globular Proteins and the Role of Constrained Water in

Solvent Exchange with Amphiphilic Cosolvents," J. Phys. Chem. B **116**(19), 5604-11, (2012).

42. F. H. Stillinger, "Structure in Aqueous Solutions of Nonpolar Solutes from the Standpoint of Scaled-Particle Theory," J. Solution Chem. **2**(2-3), 141-58, (1973).

43. K. Lum, D. Chandler, and J. D. Weeks, "Hydrophobicity at Small and Large Length Scales," J. Phys. Chem. B **103**(22), 4570-77, (1999).

44. J. D. Smith, R. J. Saykally, and P. L. Geissler, "The Effects of Dissolved Halide Anions on Hydrogen Bonding in Liquid Water," J. Am. Chem. Soc. **129**(45), 13847-56, (2007).

45. I. T. S. Li, and G. C. Walker, "Signature of Hydrophobic Hydration in a Single Polymer," Proc. Natl. Acad. Sci. U.S.A. **108**(40), 16527-32, (2011).

46. C. Y. Lee, J. A. McCammon, and P. J. Rossky, "The Structure of Liquid Water at an Extended Hydrophobic Surface," J. Chem. Phys. **80**(9), 4448-55, (1984).

47. M. Ji, M. Odelius, and K. J. Gaffney, "Large Angular Jump Mechanism Observed for Hydrogen Bond Exchange in Aqueous Perchlorate Solution," Science **328**(5981), 1003-05, (2010).

48. D. E. Moilanen, D. Wong, D. E. Rosenfeld, E. E. Fenn, and M. D. Fayer, "Ion-Water Hydrogen-Bond Switching Observed with 2DIR Vibrational Echo Chemical Exchange Spectroscopy," Proc. Natl. Acad. Sci. U.S.A. **106**(2), 375-80, (2009).

49. G. Stirnemann, F. Sterpone, and D. Laage, "Dynamics of Water in Concentrated Solutions of Amphiphiles: Key Roles of Local Structure and Aggregation," J. Phys. Chem. B **115**(12), 3254-62, (2011).

50. B. J. Schwartz, and P. J. Rossky, "The Isotope Effect in Solvation Dynamics and Nonadiabatic Relaxation: A Quantum Simulation Study of the Photoexcited Solvated Electron in D_2O," J. Chem. Phys. **105**(16), 6997-7010, (1996).

51. A. Nicholls, K. A. Sharp, and B. Honig, "Protein Folding and Association - Insights from the Interfacial and Thermodynamic Properties of Hydrocarbons," Proteins: Struct., Funct., Genet. **11**(4), 281-96, (1991).

52. J. D. Eaves, J. J. Loparo, C. J. Fecko, S. T. Roberts, A. Tokmakoff, and P. L. Geissler, "Hydrogen Bonds in Liquid Water Are Broken Only Fleetingly," Proc. Natl. Acad. Sci. U.S.A. **102**(37), 13019-22, (2005).

53. A. Cammers-Goodwin, T. J. Allen, S. L. Oslick, K. F. McClure, J. H. Lee, and D. S. Kemp, "Mechanism of Stabilization of Helical Conformations of Polypeptides by Water Containing Trifluoroethanol," J. Am. Chem. Soc. **118**(13), 3082-90, (1996).

54. D. Roccatano, G. Colombo, M. Fioroni, and A. E. Mark, "Mechanism by Which 2,2,2-Trifluoroethanol/Water Mixtures Stabilize Secondary-Structure Formation in Peptides: A Molecular Dynamics Study," Proc. Natl. Acad. Sci. U.S.A. **99**(19), 12179-84, (2002).

55. J. F. Povey, C. M. Smales, S. J. Hassard, and M. J. Howard, "Comparison of the Effects of 2,2,2-Trifluoroethanol on Peptide and Protein Structure and Function," J. Struct. Biol. **157**(2), 329-38, (2007).

56. M. Buck, S. E. Radford, and C. M. Dobson, "A Partially Folded State of Hen Egg-White Lysozyme in Trifluoroethanol - Structural Characterization and Implications for Protein Folding," Biochemistry **32**(2), 669-78, (1993).

57. J. A. Schellman, "Protein Stability in Mixed Solvents: A Balance of Contact Interaction and Excluded Volume," Biophys. J. **85**(1), 108-25, (2003).

58. K. A. Sharp, and J. M. Vanderkooi, "Water in the Half Shell: Structure of Water, Focusing on Angular Structure and Solvation," Accounts Chem. Res. **43**(2), 231-39, (2010).

59. C. Tanford, "Protein Denaturation. C. Theoretical Models for the Mechanism of Denaturation," Adv. Protein Chem. **24**, 1-95, (1970).

60. L. Hua, R. Zhou, D. Thirumalai, and B. J. Berne, "Urea Denaturation by Stronger Dispersion Interactions with Proteins Than Water Implies a 2-Stage Unfolding," Proc. Natl. Acad. Sci. U.S.A. **105**(44), 16928-33, (2008).

61. H. S. Frank, and F. Franks, "Structural Approach to Solvent Power of Water for Hydrocarbons - Urea as a Structure Breaker," J. Chem. Phys. **48**(10), 4746-&, (1968).

62. B. J. Bennion, and V. Daggett, "The Molecular Basis for the Chemical Denaturation of Proteins by Urea," Proc. Natl. Acad. Sci. U.S.A. **100**(9), 5142-47, (2003).

63. P. D. Ross, and S. Subramanian, "Thermodynamics of Protein Association Reactions - Forces Contributing to Stability," Biochemistry **20**(11), 3096-102, (1981).

64. C. Tanford, "Interfacial Free-Energy and the Hydrophobic Effect," Proc. Natl. Acad. Sci. U.S.A. **76**(9), 4175-76, (1979).

65. A. M. Klibanov, "Improving Enzymes by Using Them in Organic Solvents," Nature **409**(6817), 241-46, (2001).

66. P. A. Srere, "Protein Crystals as a Model for Mitochondrial Matrix Proteins," Trends Biochem. Sci. **6**(1), 4-7, (1981).

Chapter 3

Heterogeneous Preferential Solvation of Water-Trifluoroethanol Cosolvents on Homologous Lysozymes

The work presented in this chapter has been published in the following papers:

1. E. J. Arthur, J. T. King, K. J. Kubarych, and C. L. Brooks, III, "Heterogeneous Preferential Solvation of Water and Trifluoroethanol in Homologous Lysozymes," The Journal of Physical Chemistry B **118**(28), 8118-8127 (2014).

3.1 Introduction

The interiors of metabolizing cells have high concentrations of proteins, nucleic acids, and small molecules that can constitute more than 40% of the total cellular mass. In some cases, the density of non-water components in cells exceeds 400 g/L, which makes cytoplasmic crowding on the same order as that found in protein crystals. Contributions to crowding effects arise not only from biomacromolecules, but also a plethora of smaller osmolytes varying from sugars, such as sucrose and trehalose, to polymers, such as polysaccharides and ribonucleic acids. Previous studies have shown that such crowding effects from cytoplasmic osmolytes can significantly change the thermodynamic and kinetic

properties of not only nucleic acids and proteins, but also of water molecules. Furthermore, the complex interplay of chemicals in the cytoplasm remains difficult to characterize as simple cosolvent systems, such as octanol-water mixtures. After decades of research, the molecular mechanisms and biological significance of osmolytes interacting with biomacromolecules remain an active area of study.

Water molecules interacting with hydrophobic solutes have fewer available hydrogen bonding partners relative to the bulk, which can result in significantly constrained movements and diffusion rates. When water solvates large molecules (>1 nm), the physical constraints cause large changes to its network of hydrogen bonds. These can halve the average time between hydrogen bond jumps, and slow diffusion by more than an order of magnitude. Dynamically constrained solvent is not only a structural component to biology, but its altered chemistry is also exploited by processes such as protein-ligand binding, protein-protein recognition, ice crystal inhibition,[45] and protein-DNA interactions.[46,47] It is therefore a necessity to molecular biology, especially when studying within the context of cell-like environments, to deconvolve the influence on hydration environments near proteins due to various interactions, such as van der Waals, electrostatics, and protein topology.

Previous studies of proteins interacting with cosolvents have shown that changes in transfer free energy of solvent molecules near a protein's surface relative to the bulk, or so-called "epistructural interfacial tension", receives electrostatic contributions from the protein's interfacial topology. This notion has led to accurate docking predictions of small molecules on a protein's surface using implicit-water methods such as the three-dimensional reference interaction site model (3D-RISM). However, further evidence has shown that not all protein-ligand systems may be mapped accurately without a dynamic, explicit representation of water intermediating protein-ligand associations. These studies have led to

reassessments of such simplified models, even by coupling them to molecular-dynamics (MD) simulations to increase conformational sampling of both protein and solvent. Owing to the complexity of liquid solvent and the rapidly fluctuating nature of protein topology, it may be premature to suggest a theoretical model short of an all-atom MD simulation that predicts protein-solvent interactions accurately. It may also be an equally arduous task to modify a topology-based approach, such as 3D-RISM, to represent accurate hydrophobic protein-solvent interfaces and three-body interactions for any particular protein-ligand-water system. Thus for this study we turn to all-atom MD simulations as a means to investigate biomolecular interactions in mixed-solvent systems.

Previous work by King et al.[32] on systems of lysozyme and trifluoroethanol used the method of two-dimensional infrared spectroscopy (2DIR) to investigate how solvation and dehydration can differ depending on the specific location on a protein. Hen egg white lysozyme (HEWL) and human lysozyme (HuLys) offer homologous protein topologies, each with one solvent-exposed histidine. Although the two proteins are 77% similar by amino acid sequence and are structurally different by only 0.54 Å root-means-square, the histidines are located on different domains of the protein. The H15 on HEWL is located on a turn adjacent to an alpha-helix, and the H78 on HuLys is located on a region without secondary structure. Local environments around these histidines were probed by covalently attaching a ruthenium-carbonyl vibrational chromophore. In initial studies the vibrational lifetime of the chromophore in H_2O and D_2O was used to measure not only the presence of water, but also the hindering of hydrogen bond reorientation dynamics in the nearby hydration water. It was found that different water dynamics correlate strongly with the local surface structure of the protein. The H15 probe location of HEWL is a low-curvature region solvated by orientationally constrained water, whereas the H78 site of HuLys is high-curvature and

unstructured, and solvated by bulk-like water. To test the connection between constrained water and the thermodynamic driving force for dehydration by an amphiphilic co-solvent trifluoroethanol (TFE), lifetime measurements were made in a series of D_2O/TFE solutions. In pure D_2O, both sites were found to be hydrated based on their sub-5 ps vibrational lifetimes, which are consistent with water-assisted relaxation.[55-57] Upon addition of TFE, however, the sites displayed markedly distinct responses. The lifetime of the probe at the H15 site of HEWL exhibited an order-of-magnitude slowdown in a 10% (v/v) TFE solution consistent with local dehydration, whereas the H78 labeled site of HuLys showed no TFE-dependent vibrational lifetime changes at any of the experimental concentrations.

These data indicate that the local solvent compositions at the two sites are different. Previous NMR and circular dichroism studies of HEWL confirms that a 10% concentration of TFE does not change the helical content nor the tertiary contacts of lysozyme, which supports the conclusion that the change in vibrational lifetime is not due to a change in protein conformation. This result is consistent with prior observations that helical regions on proteins (such as the H15 on HEWL) are preferentially solvated by TFE more than unstructured regions (such as the H78 on HuLys).[61] Additionally, this result suggests that local solvation structure and dynamics can be modified by local protein topology.

The simulations of the present study are designed to investigate these results and explore the heterogeneity of preferential solvation of lysozyme by TFE-water mixtures. In the current study we used explicit solvent MD simulations to model human and hen egg white lysozymes mixed with water and different concentrations of the cosolvent trifluoroethanol. We then aligned each trajectory by lowest protein backbone-atom root mean square deviation (RMSD) to one common structure. We used these trajectories to compute time-averaged three-dimensional (3D) histograms of the number density of solvent

relative to each protein's structure. These values represent the spatial distribution of both probability of finding a type of solvent atom and solvent density. Using these data we mapped out trends of trifluoroethanol interacting with lysozyme surfaces and suggest a possible explanation for the observed phenomena in the spectroscopic experiments. Finally, we made a spatially-dependent, solvent-centric comparison of homology between HEWL and HuLys.

3.2 Simulations

Two homologous lysozyme systems were simulated: hen egg white lysozyme (HEWL; PDB code 3IJU) and human lysozyme (HuLys; PDB code 2ZIJ). Eighteen replicas of both proteins were created, which consisted of three separate trajectories for each of six concentrations of TFE: 0%, 1%, 5%, 10%, 15%, and 20% by volume fraction (v/v). Water/TFE mixtures exhibit a nonideality of less than 10 mL per liter (less than 1%), so a ratio of molar fractions could be approximated by a ratio of volume fractions. Equation 1 shows how the precise number of TFE and water molecules could be calculated for a given cosolvent when assuming the solution behaves ideally.

$$\frac{V_m^{TFE}(\%TFE\,v/v)}{V_m^{H_2O}(\%H_2O\,v/v)} \approx \frac{x_{TFE}}{x_{H_2O}} = \frac{N_{TFE}}{N_{H_2O}} = \frac{G(r)_{TFE}}{G(r)_{H_2O}} \qquad \text{eq. 1}$$

V_m is molar volume, x is mole fraction, and N is the number of solvent molecules. The number of TFE and water molecules used in each simulation is listed in the original text for this chapter.

Hydrogen atoms were added to the proteins using the *pdb2gmx* utility in the GROningen MAchine for Chemical Simulations (GROMACS).[62] All replicas were solvated in SPC/E water[60] using the *genbox* utility in GROMACS with rectangular edges at least 20 Å from all protein atoms. Excess charge from the protein was neutralized by placing 8 chloride ions per lysozyme at random locations in the solvent using the *genion* utility in GROMACS. The TFE structure was energy-minimized using the Gaussian '03 software package.[63] An appropriate number of TFE molecules were added to each replica simultaneously with the chloride ions, using the *genion* utility from GROMACS. The locations of the TFE molecules were randomized for each replica to enhance the sampling of solvent configurations.

All 36 systems (2 proteins x 6 TFE concentrations x 3 independent trajectories) were simulated using the GROMACS macromolecular modeling package (version 4.5.5).[64] The *antechamber* program from the Antechamber package (version 1.25)[64] was coupled with Gaussian '03 to assign partial charges and to create an Amber-like forcefield for TFE. Partial charges were assigned using the Restrained Electrostatic Potential (RESP) method.[65] The remaining atoms of each replica were simulated using the AMBER99-all-atom force field.[66] Each replica was an isobaric-isothermal ensemble, and was maintained at 1 atm and 300 K using the Berendsen barostat and thermostat respectively.[57] A time coupling constant of 1 ps was used for both pressure and temperature, and the system compressibility was set to 4.5 x 10^{-5} bar. Electrostatic energies were determined using particle-mesh Ewald (PME) summations[58,67] with a Fourier-transform grid width of 1.2 Å, and real-space Coulomb and Lennard-Jones cutoffs of 9 Å. The magnitude of the PME-shifted potential at the cutoff was set to 10^{-5}, and the Leapfrog Verlet integrator was used with an integration time step of 1 fs. Each replica was energy minimized using a steepest-descent algorithm for 500 steps with a tolerance of 10 kJ mol^{-1} nm^{-1}, followed by an equilibration run for 50 ps, and finally a

production run of 20 nanoseconds (ns). Coordinates were saved every 1 picosecond (ps), which yielded a total of 60,000 structures for each protein at each concentration.

3.3 Volumetric Distribution Function of Solvents

Owing to the extremely low flexibility, high stability, and highly-conserved structure of the two lysozymes, all saved structures from all simulations represent fluctuations of one lysozyme system. The calculated circular dichroism (CD) shows an ellipticity of 10.1 ± 0.8 degrees, and the root-mean-square deviation (RMSD) of protein backbone atoms from the initial structure was 1.0 ± 0.1 Å. Additionally, no protein structure shows an RMSD of backbone atoms greater than 2 Å from any other structure, even between human and hen egg white lysozymes.[67,68] Although concentrations of TFE were simulated that would normally denature lysozyme, it may be that the mechanism of denaturing takes place on timescales longer than the 20 ns simulated in this study. These conditions permit the calculation of high-resolution three-dimensional (3D) solvent distribution functions centered on a relatively static protein structure.

First, periodic boundary conditions are used to align the protein at the center of each box. Then all saved structures from all simulation are aligned by least-squares fitting of protein backbone atoms to a single energy-minimized reference structure of HEWL. The reference structure is obtained from the first frame of one of the production runs of HEWL. Finally, time-averaged solvent distribution functions $G(r)$ are calculated for each trajectory using voxelized 3D histograms with a 1 Å^3 resolution using Equation 2, as performed in previous studies. The solvent distribution function $G(r_{xyz})$ is approximated by integrating the time-averaged solvent density ρ for a voxel of size $\Delta x \Delta y \Delta z$. The data is then normalized for

bulk density ρ_{bulk}, which resulted in a series of 36 maps of solvent distribution, each with a protein in the center.

$$G(r_{xyz}) = \left\langle G(r_{xyz},t)\right\rangle_t = \int_x^{x+\Delta x}\int_y^{y+\Delta y}\int_z^{z+\Delta z}\left(\frac{\left\langle \rho(x,y,z,t)\right\rangle_t}{\rho_{bulk}}\right)dx\,dy\,dz \qquad \text{eq. 2}$$

As an artifact of the least-squares fitting of saved structures, the corners of the periodic boundary boxes rotate during simulation. As such, *G(**r**)* data at the corners is not representative of bulk solvent in the solvent distribution functions, and was removed before analysis. This left a spherical volume of solvent density with a radius of 41 Å and an edge with bulk solvent density. This radius also maintains a minimum of 18 Å between all protein atoms and the edge of the spherical volume. Radial distribution functions of solvent from protein atoms indicate that no significant solvent clustering occurs much more than 9 Å from the surface of the protein.[67] Hence, interactions between the protein and the solvent, such as an enhanced solvent density, are not omitted from analyses by removing the corners. Furthermore, the edge of the data is representative of the time-averaged bulk density of water and TFE. The solvent densities at the edge of the spherical shape of the *G(**r**)* histograms are averaged to calculate the ρ_{bulk} of TFE and water.

All data within the *G(**r**)* histograms converged to consistent values in each voxel: the protein and water densities converged within the first 1-2 ns of simulation time, and TFE density within 5-14 ns. This indicates that a 20 ns simulation is sufficient to sample the atomic densities of 3D space sufficiently for further analysis.[58] TFE, relative to water, has a slower reorientation time and a slower diffusion time, thus *G(**r**)* functions of TFE require

more sampling to converge to one set of values, especially at lower concentrations. It may be, then, that shorter simulations would not sample a long enough trajectory to understand the average movements of TFE around lysozymes.

3.4 Local Percent of TFE by Volume

As noted in the simulation procedure, water/TFE mixtures are sufficiently ideal to translate a percent TFE by volume into a ratio of molecules to within 1% accuracy. Conversely, we can calculate the concentration of TFE v/v of a given volume from the number density of solvent molecules using Equation 1. Since the solvent distribution function $G(\mathbf{r})$ is the time-averaged number density of solvent molecules, we can relate it to equation 1 and find the percent TFE v/v of the solvent distribution function $G(r)$. By converting solvent atom counts per cubic angstrom into moles per cubic centimeter, we calculate of the percent TFE v/v for a single voxel. This calculation works wherever the volume in question contains solvent density from both water and TFE.

Every residue on human lysozyme shares a corresponding spherical volume with a residue on hen egg white lysozyme, except for the T43 which has no analogous residue on HEWL. These volumes can then be used to compare simulations with different solvent concentrations and different protein identities. These spheres of radius 7 Å, are centered at the average center of geometry of a residue's backbone atoms. The result is a total of 36 analogous spheres for every residue location, with each sphere consisting of 1437 voxels.

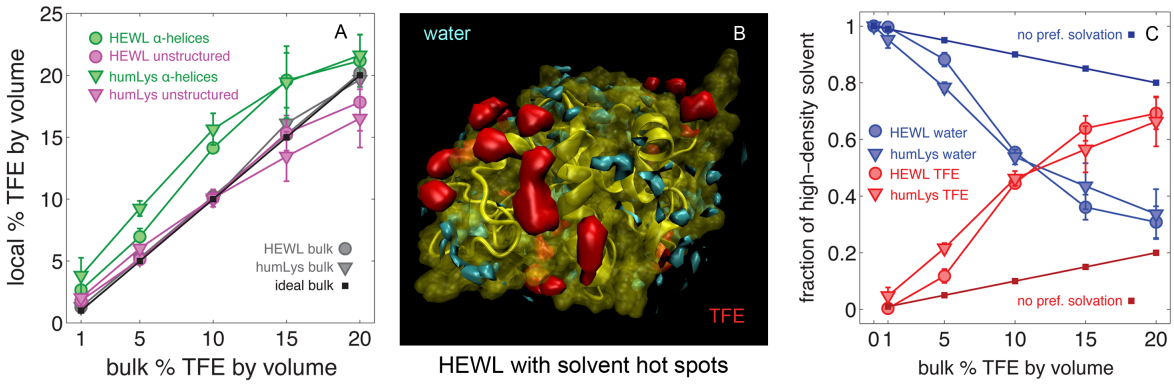

Figure 3.1 A) Percent TFE v/v calculated for the local environment of each surface-lying residue. Shown here are the average percentage of TFE for alpha-helices (green) and unstructured regions of the protein (magenta). Alpha-helices show a local increase in TFE relative to the bulk (grey/black), while unstructured regions show a relatively bulk-like concentration. The error bars are the standard deviation among the three parallel trajectories for each protein at each concentration. B) HEWL is shown as a visual cue for the general distribution and location of high density hot spots. TFE (red) and water (cyan) did not overlap in this data. C) The total volume of hot-spots for water and TFE exhibit a crossover near 10% TFE, beyond which the majority of hot spots are due to TFE. The error bars are the standard deviation among the three parallel trajectories for each protein at each concentration.

This selection encompasses 86% of volume with three times the bulk density of TFE. Since there is no clear method for rotating and realigning the grid of one residue to another, comparisons between non-analogous residues are not performed in this study.

The goal of this study is to analyze protein solvation, so buried residues are excluded from solvent analysis. A residue is considered buried if its average SASA is less than 17 $Å^2$, which led to 97 solvent exposed residues on HEWL, and 89 for HuLys.[58,69] Interestingly, human lysozyme on average had slightly less SASA than HEWL, and thus was slightly more spatially compact a protein.

As shown in Figure 3.1a, helical regions of the protein (green) show an enhanced concentration of TFE by up to 5.6% v/v relative to the bulk, while unstructured regions of the protein (magenta) show an enhanced concentration of water by up to 3.5% v/v relative to the bulk. By "helical" we mean both alpha-helical and 3/10-helices, and by "unstructured"

we mean turns, bends, and regions without secondary structure. This result is reasonable, since helices are both richer in solvent-exposed hydrophobic residues, and have been previously shown to be preferentially solvated by TFE. Unstructured regions, on the other hand, have more hydrophilic residues, and are preferentially solvated by water. A feature of high local concentrations of TFE (such as 15 and 20% by volume) is a greater standard deviation in solvent density data among parallel simulations. As mentioned in the previous section, this may be attributed to the longer time needed for TFE solvation data to converge.

What was not revealed in the data was a correlation between an individual residue's hydrophobicity and the local concentration of TFE. As discussed in later in the section "Insight into Site-Specific Dehydration near Lysozymes", a single residue's local concentration of TFE is most influenced by neighboring residue effects than its own hydrophobicity. Only when averaging over protein domains does a trend in TFE solvation become greater than the variance in the data.

Interestingly, the 50 residues with the highest local concentration of TFE from simulations of 15% bulk v/v TFE match more than 50% of the TFE-lysozyme crystal contacts in found in X-ray studies. These data indicate the forcefield choices reliably captures features of lysozyme in a water/amphiphilic cosolvent mixture.

3.5 Solvent Hot Spots

Hot spots contain a high number density of one solvent type. Within these regions of space, the probability density of a solvent is similar to that of the protein backbone atoms, which effectively makes them extensions of the protein's surface topology into the surrounding solvent. In terms of $G(\mathbf{r})$ data, hot spots are voxels that have an averaged local solvent density much higher than that of the bulk. For the simulations at 10% v/v TFE, the

*G(**r**)* functions show maxima over 2 and 12 times higher than the bulk density of water and TFE respectively. Isosurfaces enclosing these high-density regions on HEWL are shown in Figure 3.1b. No simulation shows hot spots extending further than 5 Å from protein atoms, indicating that stationary, high-density solvent clustering does not form in the bulk solvent during the simulations, and that large perturbations in solvent density do not extend beyond 5 Å from the surface of the protein. This also suggests protein-protein interactions from opposite sides of the lysozyme did not extend through the periodic boundaries of the solvent box.

The total volume of high-density solvent for all simulations averaged to 642 ± 91 $Å^3$, which indicates that a feature of the protein-solvent interface is a conserved volume of strongly-associated solute. With respect to the relative sizes of TFE and water molecules (126 $Å^3$ and 32 $Å^3$ respectively) this space corresponds to about 5 TFEs or 20 waters. What does change among different cosolvent concentrations is the identity of solvent dominating the hot spots. Interestingly, as the bulk concentration of TFE increased, the interfacial solvent environment shows a transition from being water-dominated to being TFE-dominated at the same concentrations that are known to denature lysozyme in experiments. High-density solvent is rich with water at low concentrations of TFE in the bulk, and in 15 and 20% TFE v/v in the bulk, the high-density solvent became dominated by TFE, as shown in Figure 3.1c. This observation is also reflected in the standard deviation of local concentration of TFE, as noted in the previous section. Although we see no evidence that the proteins denature during the simulations, this transition may lend insight into the mechanism that unfolds the protein. Lysozyme retains its native fold by maintaining a relatively consistent distribution of strongly-associated water. It is the removal of this water

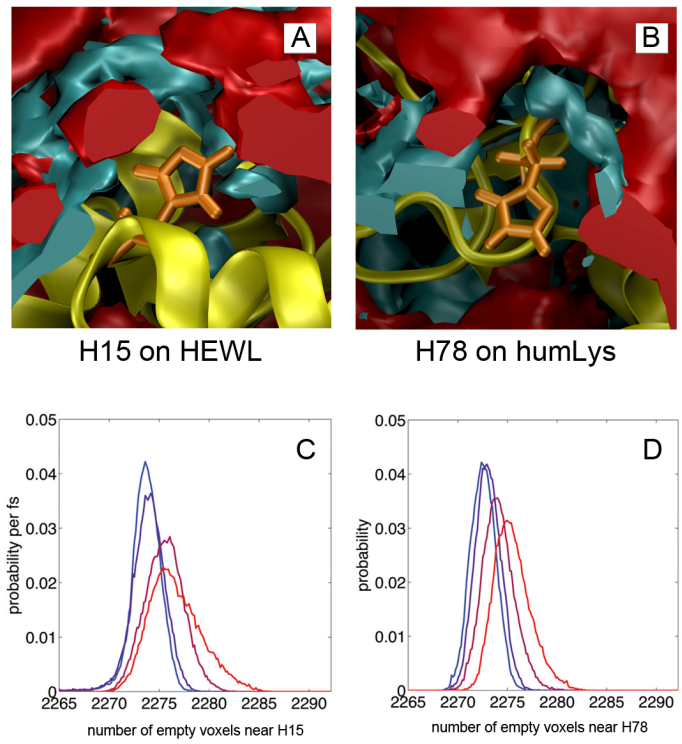

Figure 3.2 The local solvent structures near the histidines in simulations of 10% TFE v/v. A) HEWL (yellow) with histidine 15 (orange sticks) is surrounded by TFE (red) and water (blue) isosurfaces. B) Isosurfaces for histidine 78 on HuLys. Notice that both locations are surrounded by similar ratios of both solvent types. C) and D) show the probability of finding a number of empty voxels near the local environments around each histidine at various cosolvent concentrations. Notice that the distribution for HEWL's H15 site broadens out at much lower concentrations of TFE than HuLys' H78 site.

that leads to a non-native packing of the protein. Experiments have shown that lysozymes denaturing thermally also experience a disruption in their hydrogen bonding network before unfolding.[51,67,71]

3.6 Insight into Site-Specific Dehydration near Lysozymes

Studies of both model hydrophobic interfaces and biomolecules have provided insight into the nature of hydration water on the molecular scale. Patel et al. calculated the

probability density distributions of finding water near the solute-solvent interfaces of model systems, which included hydrophobic methyl groups, hydrophilic hydroxyl groups, melittin dimers, and biphenyl dioxygenase (BphC). Despite the chemical differences, it was found that the time-averaged number densities for water at the solvent interface are independent of the hydrophobicity of the surface itself. What differs markedly is the probability of finding a very small number of water molecules near each type of surface. That is, deviations from the average number density of water, corresponding to de-wetting, are much more likely in the vicinity of hydrophobic surfaces than hydrophilic ones.

Although the simulations of the current study cannot reach the level of precision in the work done by Patel et al., with some margin of error we can still infer the relative hydrophobicity of the two histidine sites. By counting the number of empty voxels around each histidine for each simulation we obtain the metric shown in Figure 3.2c and 2d, which shows a systematic increase in the number of waterless voxels with an increase in the concentration of TFE. The data clearly shows that beyond the variation of the data, when the histidines are exposed to higher concentrations of TFE one is more likely to find a vacuum-like environment around H15 of HEWL, and one is more likely to find a hydrated environment around H78. By the same logic from the studies of Patel et al., we therefore find H15 is becoming more hydrophobic with an increase in local TFE. This observation is likely influenced by the large difference in SASA between the residues: 55.1 and 175.1 $Å^2$ for H15 (HEWL) and H78 (HuLys) respectively. Figure 3.2a and b shows a visual reference of the relative surface area and solvent composition.

Although this particular TFE model is not properly tuned to exhibit a maximum number of evacuated voxels at the experimentally-analogous 10 % TFE by volume, it does support the hypothesis that TFE dehydrates the H15 location of the HEWL protein. These

data suggest that a direct mechanism of locally dehydrating the surface of lysozyme causes the change in signal amplitude from the protein label. As TFE removes neighboring water molecules, it also reduces the number of water molecules that can couple to the probe. The H78 on HuLys has more SASA, and consequently many opportunities for water to reach and couple with the probe.

3.7 Pearson Correlation Coefficients

Since all $G(\mathbf{r})$ functions are analogous 3D histograms, direct comparisons of the distribution of solvent density are made between pairs of simulations. Specifically, the local environments around each residue (detailed above as being 7 Å spherical volumes) are selected and analyzed by calculating Pearson correlation coefficients between sets of analogous voxels using Equation 4.

$$CorrCoef_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 (y_i - \bar{y})^2}} \qquad \text{eq. 4}$$

Here, the solvent densities of two local environments are compared by multiplying each normalized element x from one residue's $G(\mathbf{r})$ to its corresponding analogous element y from another residue's $G(\mathbf{r})$. This process converts the shapes of two solvent densities into a

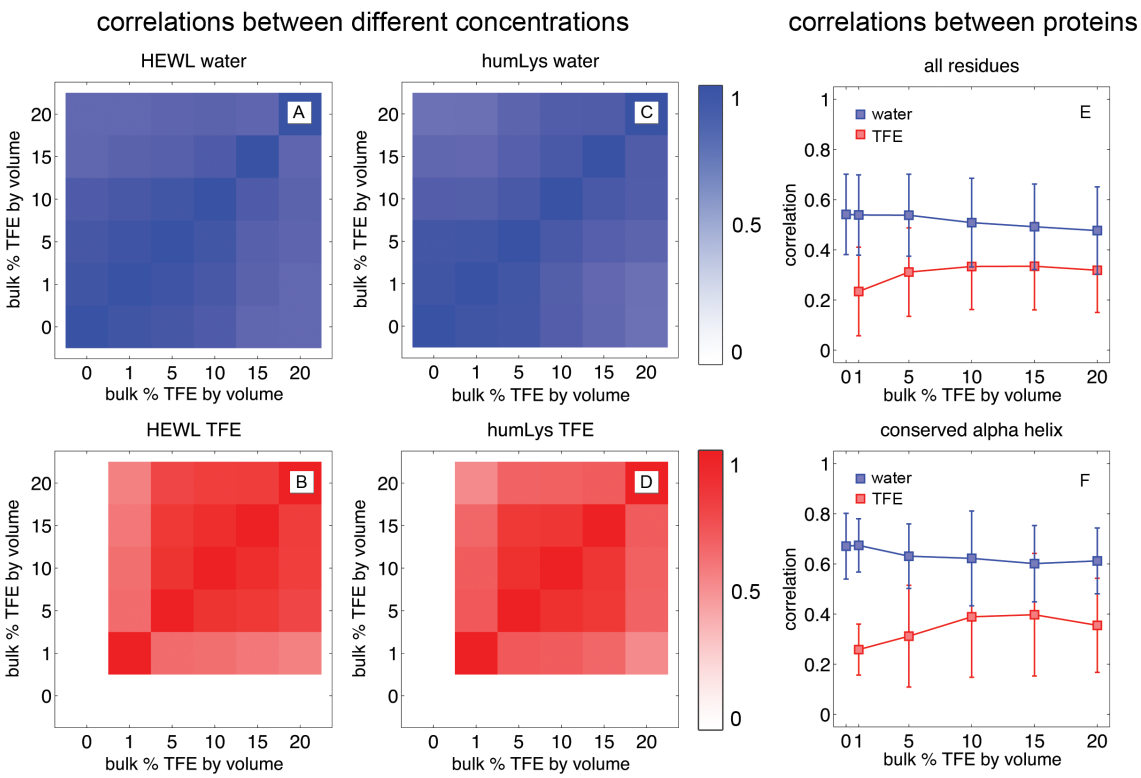

Figure 3.3 For all plots, only data from solvent-exposed amino acids are considered. Panels A through D show average correlation coefficients between amino acids of one protein (HEWL or HuLys) in solutions of different concentrations of TFE v/v. All correlations fall between 1 (on the diagonals) and 0.54 (at the corners) in these plots. Plots A and C are correlations of water densities at different concentrations, and plots B and D are correlations of TFE. Plot E is an average correlation of *G(r)* functions around each amino acid by comparing residues from HEWL to its homologue on HuLys. The error bars are the standard deviation of data among the correlations of amino acids. Plot F is the same analysis as seen in Plot E, except that only residues on the alpha helix that has 100 % conservation of residue identity. A stronger correlation is observed here, but due to neighboring effects of non-identical amino acids, the TFE distributions remained nonhomologous between the proteins.

value that indicates their relative similarity: 0 as non-correlative (no spatial overlap of data), 1 as a perfect correlation (a perfect spatial overlap of data), and -1 as a perfect anticorrelation. No attempt is made in this study to remedy the anti-aliasing artifacts of $G(\boldsymbol{r})$ data that occur when aligning non-analogous volumes. Thus no comparison between non-analogous locations are made (such as between two alanines on different protein domains).

Using Equation 4, we investigate three aspects of protein-solvent interactions: how

much simulation time is needed to converge on one solvent density distribution (comparing a residue site to itself at different times within the same simulation); what TFE and water interactions are conserved in cosolvent mixtures (comparing a site on one protein to itself in different cosolvent mixtures); and what TFE and water interactions are conserved between homologous proteins (comparing a site on HEWL to a homologous site on HuLys).

When calculating the convergence of solvent density around residues within a single simulation, we find that 20 ns of simulation provides sufficient sampling. Correlations of local $G(\mathbf{r})$ functions at each residue site are made between the instantaneous and time-averaged $G(\mathbf{r})$ functions. Due to the low flexibility and high stability of lysozyme systems as well as the high diffusion rate of the solvents, local $G(\mathbf{r})$ functions of water, the protein, and TFE converge within 2 ns, 2 ns, and 14 ns respectively, and had maximum correlations of 0.89, 0.75, and 0.62, respectively.[73] This indicates that each trajectory not only converges to a self-consistent atomic density, but also is well-correlated to the average of all densities. As such, the average $G(\mathbf{r})$ function of all 60 ns of simulation at each concentration is used as a representative atomic occupancy distribution of each protein in that corresponding environment.

Comparisons between identical amino acids at different concentrations of TFE revealed that for a single protein, there is a persistent configuration of solvent density (Figure 3.3a-d). Water and TFE have minimum correlations of 0.54 and 0.45 respectively, which indicates that even when placing a lysozyme in the extremes of 1% and 20% TFE, the local solvent density retains least a 45% overlap between any two simulations of that protein. When placed in solutions that showed better sampling for the cosolvent (such as in 5% and 10% TFE), the correlation coefficients between simulations rises even higher to 0.89 and 0.83 for water and TFE respectively. While comparing simulation data of one protein in

different concentrations of TFE, not only can representative information be gained from *G(r)* data between cosolvent concentrations for a protein, but also the lysozymes preferentially configure the solvent molecules on their surfaces regardless of the solvent composition. Remarkably, solvent molecules quickly find preferred configurations both when subjected to low sampling rates (such as TFE *G(r)* data in the 1% TFE simulations) and when experiencing lower diffusion of solvent molecules (such as in the 15 and 20% TFE simulations).

Next we explore what happens when imposing a hard cutoff, as defined in Equation 5. The This technique reduces the effects of noise on the correlation coefficient analyses, and presumably defines a more rigid shape to solvent configuration near the lysozymes.

$$G'(r) = \begin{cases} 0; & G(r) < (2 \times \rho_{bulk}) \\ 1; & G(r) \geq (2 \times \rho_{bulk}) \end{cases} \qquad \text{eq. 5}$$

*G(**r**)* functions then converted into rigid-boundary maps of high-density solvent where any location within *G(**r**)* with more than twice the bulk value of a solvent the ρ_{bulk} is 1, and every other space is 0. Correlation coefficients of *G'(**r**)* functions are decreased on average by 0.13 as compared to those reported in Figure 3.3a-d, indicating that the shapes of high-density solvent are also conserved between different bulk concentrations of TFE. This comparison also suggests there are thermodynamic minima on the protein for binding specific solvent components, and that these are maintained, at least in part, regardless of the bulk cosolvent composition.

When making comparisons between local solvent density around the two lysozymes, as shown in Figure 3.3e and f, TFE correlation coefficients are impacted much more than those of water. Hen egg white and human lysozymes are 77% similar and 60% identical

according a Smith-Waterman alignment.[74] When ignoring the shape of solvent density, the Pearson correlation coefficient between the local concentrations of TFE by volume around each residue is 0.55, as calculated with Equation 3. Presumably the proteins should have similar shapes of local solvent density, and comparably high correlation coefficients. Correlation coefficients between local volumes of the two proteins, shown in Figure 3.3e, average to 0.51 ± 0.03 for water and 0.26 ± 0.13 for TFE. These were 0.10 (water) and 0.22 (TFE) less than the lowest correlation values from Figures 3a-d.

Unexpectedly, while sequence alignment is a good predictor of similarity of *G(r)* data for water, it isn't for TFE. The two homologous lysozymes share similar shapes of *G(r)* data, which translates to correlations similar to identical amino acid sequence alignment. Even so, water's solvent density near HEWL is more similarly shaped between cosolvents of 0 and 20% TFE v/v than it is to HuLys with the same concentration of TFE. The correlations of averaged TFE density between the two proteins is even lower, indicating that although both lysozymes are similar in sequence, they have dissimilar interactions with water and TFE. Moreover, both proteins are more similar in their interaction with water than with TFE.

To ensure that noise in the *G(**r**)* functions were not falsely inflating the error of the analyses, parallel calculations were run with voxel volumes of 8 and 64 $Å^3$ (2 and 4 Å of voxel side lengths). The *G(**r**)* functions with reduced resolutions changed correlation coefficients by no more than 0.11, which indicates that the observations discussed above are resolution-independent.

In order to locate the sources of dissimilar solvent interactions, correlations are segregated by secondary structure type, residue identity, residue similarity, and hydrophobicity. Unfortunately, there are no apparent correlations of the shape *of G(r)* data between the two types of proteins beyond the variation of the data. Of particular interest is

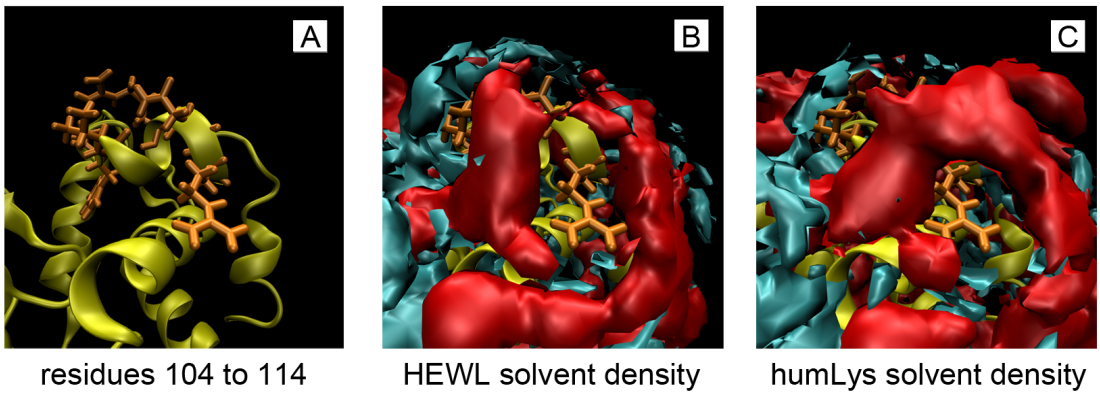

Figure 3.4 All three figures above show a reference lysozyme tertiary structure (yellow) and the residues of the alpha helix that are conserved between hen egg white and human lysozymes (orange). Since this study ignores buried residues, only the surface-lying residues 107, 108, 109, 112, 113, and 114 are shown as sticks. Panel A illustrates the configuration of side chains, and panels B and C overlay the protein with solvent density averaged from the three replicas at 10 % TFE. Even though this helix is completely conserved between the proteins, both in amino acid sequence and relative backbone RMSD, the averaged solvent densities of water (cyan) and TFE (red) are significantly different at this region. This difference illustrates that neighboring effects on solvent density from non-identical residues extend over many angstroms, and that a region with conserved amino acid sequence does not necessarily indicate a region with conserved solvent interactions.

the alpha helix from residues 105 to 114 on HEWL that is entirely conserved between the two lysozymes. An illustration of solvent density at the conserved alpha helix is shown in Figure 3.4, and their correlations are shown in Figure 3.3f. There is an average 0.11 and 0.03 increase in correlation for water and TFE respectively for this particular group, which still falls 0.17 short of the lowest correlations of that same group when comparing only one protein to itself in different cosolvents. Comparing homologous residues in the binding pockets of the proteins yields correlations that are lower than the average.

Evidentially, water and TFE interact with the specific details of protein surfaces differently. Water, being relatively small and having several axes of symmetry, resembles a more ideal solvent molecule than TFE. Its average interaction with the protein interface is conserved between HEWL and HuLys as much as their amino acid sequences. TFE, being

nine times larger, having fewer axes of symmetry, and having more internal degrees of freedom (such as dihedral angles), is more sensitive to influences from neighboring residues. Although the extent of these influences is unclear, they are long-ranged enough to disrupt the solvent density near the conserved alpha helix. In order to have similar solvent density at one homologous location between two proteins, it may require conserved topological features on the protein surfaces beyond the 7 Å radius used in the calculations of this study. Observing that TFE can influence water molecules as far as 8 Å away (twice the length of a TFE molecule) when in solution,[49] it is reasonable to expect that small differences in a protein's surface topology can have similarly long-reaching influences on solvent interactions.

Since the two proteins are highly conserved both in enzymatic mechanisms and physiological distribution among species,[1] the homology of solvent interactions may be unimportant to lysozyme chemical activity. Conversely, the similarity of averaged solvent interactions between two proteins may not indicate a structural homology. A well-equilibrated $G(\mathbf{r})$ function of solvent density may be a poor predictor for $G(\mathbf{r})$ functions of homologous systems, even with solvent molecules as small as TFE. When comparing a region of the protein with similar chemical function (and presumably similar charge distribution), such as the binding pocket, there always is a wide standard deviation of correlations between individual residues. For instance, W62 shows good correlations between the lysozymes in various cosolvents for both water and TFE, but a key catalytic residue D52 always shows a poor correlation. It may be that specific residues must maintain a certain number density of solvent interaction to maintain chemical properties (such as protein stability or catalytic reactivity). Other residues merely need to enforce electrostatic qualities in a reactive center. Even though TFE is not a target molecule for lysozyme

catalysis, this study suggests that targeted binding experiments with one lysozyme may not predict well results from similar experiments with another lysozyme.

3.8 Conclusions

Inspired by our experiments of mapping site-specific solvent interactions of lysozymes, we present here an analytical approach to using molecular dynamics for characterizing local interactions of lysozyme residues with a water-TFE cosolvent. This is a process of aligning all trajectories to one homologous structure, making a time-averaged 3D *G(r)* function of the data, and dividing *G(r)* into small volumes that encapsulate high-density solvent. As such we show a process for locating probable crystal contacts, observing preferential solvation trends, and comparing protein homology from the shape of averaged solvent density. These techniques are fully generalizable to proteins interacting with cosolvents of denaturants, small molecules, and salts.

We show that our trifluoroethanol forcefield mimics its basic chemical properties, such as preferentially solvating alpha helices more than unstructured regions of the protein and finding crystal contacts. Additionally we find that at concentrations above 10 percent TFE, water around the protein is displaced with TFE. This is consistent with a water displacement mechanism for TFE chemically denaturing lysozymes. Using our system setup we also found that it might be TFE displacing water hot spots on lysozyme that results in the protein denaturing. With regards to site-specific solvent dynamics, as with the ruthenium-dicarbonyl experiments on human and hen egg white lysozyme, displacing water on the surface of the protein can isolate regions of the protein from the bulk solvent and effectively shut off pathways of energy transfer from small molecule probes to the surrounding solvent.

Using 3D *G(r)* function we have a method for comparing the shape and overlap of averaged solvent density around proteins. We find that the two lysozymes conserve solvent hot spots despite being surrounded by different concentrations of TFE. We also find that homologous proteins may share similar interactions with one solvent, such as water, but not share similar interactions with another solvent, such as TFE. Larger solvent molecules with more degrees of freedom may have more pronounced effects from neighboring residues, and accordingly exhibit greater differences in average solvent interaction. Conversely, smaller solvents with several axes of symmetry, such as water, can have similar interactions with homologous proteins. What is very clear is that homologous proteins may be poor representations of one another when measuring solvent molecule interactions.

3.9 References

1. A. B. Fulton, "How Crowded Is the Cytoplasm," Cell **30**(2), 345-47, (1982).

2. P. A. Srere, "Protein Crystals as a Model for Mitochondrial Matrix Proteins," Trends Biochem. Sci. **6**(1), 4-7, (1981).

3. S. B. Zimmerman, and A. P. Minton, "Macromolecular Crowding - Biochemical, Biophysical, and Physiological Consequences," Annu. Rev. Biophys. Biomol. Struct. **22**, 27-65, (1993).

4. S. B. Zimmerman, and S. O. Trach, "Estimation of Macromolecule Concentrations and Excluded Volume Effects for the Cytoplasm of Escherichia-Coli," J. Mol. Biol. **222**(3), 599-620, (1991).

5. X. J. Zhang, J. A. Wozniak, and B. W. Matthews, "Protein Flexibility and Adaptability Seen in 25 Crystal Forms of T4 Lysozyme," J. Mol. Biol. **250**(4), 527-52, (1995).

6. K. B. Frederick, D. Sept, and E. M. De La Cruz, "Effects of Solution Crowding on Actin Polymerization Reveal the Energetic Basis for Nucleotide-Dependent Filament Stability," J. Mol. Biol. **378**(3), 540-50, (2008).

7. R. J. Ellis, and A. P. Minton, "Cell Biology - Join the Crowd," Nature **425**(6953), 27-28, (2003).

8. E. E. Fenn, D. E. Moilanen, N. E. Levinger, and M. D. Fayer, "Water Dynamics and Interactions in Water-Polyether Binary Mixtures," J. Am. Chem. Soc. **131**(15), 5530-39, (2009).

9. D. Chandler, "Interfaces and the Driving Force of Hydrophobic Assembly," Nature **437**(7059), 640-47, (2005).

10. I. Yu, K. Nakada, and M. Nagaoka, "Spatio-Temporal Characteristics of the Transfer Free Energy of Apomyoglobin into the Molecular Crowding Condition with Trimethylamine N-Oxide: A Study with Three Types of the Kirkwood-Buff Integral," J. Phys. Chem. B **116**(13), 4080-88, (2012).

11. I. Yu, Y. Jindo, and M. Nagaoka, "Microscopic Understanding of Preferential Exclusion of Compatible Solute Ectoine: Direct Interaction and Hydration Alteration," J. Phys. Chem. B **111**(34), 10231-38, (2007).

12. I. Yu, and M. Nagaoka, "Slowdown of Water Diffusion around Protein in Aqueous Solution with Ectoine," Chem. Phys. Lett. **388**(4-6), 316-21, (2004).

13. Q. Zou, B. J. Bennion, V. Daggett, and K. P. Murphy, "The Molecular Mechanism of Stabilization of Proteins by Tmao and Its Ability to Counteract the Effects of Urea," J. Am. Chem. Soc. **124**(7), 1192-202, (2002).

14. T. Arakawa, and S. N. Timasheff, "The Stabilization of Proteins by Osmolytes," Biophys. J. **47**(3), 411-14, (1985).

15. K. Gekko, and S. N. Timasheff, "Mechanism of Protein Stabilization by Glycerol - Preferential Hydration in Glycerol-Water Mixtures," Biochemistry **20**(16), 4667-76, (1981).

16. G. F. Xie, and S. N. Timasheff, "Mechanism of the Stabilization of Ribonuclease a by Sorbitol: Preferential Hydration Is Greater for the Denatured Than for the Native Protein," Protein Sci. **6**(1), 211-21, (1997).

17. N. A. Chebotareva, "Effect of Molecular Crowding on the Enzymes of Glycogenolysis," Biochemistry-Moscow **72**(13), 1478-90, (2007).

18. R. J. Ellis, "Macromolecular Crowding: Obvious but Underappreciated," Trends Biochem. Sci. **26**(10), 597-604, (2001).

19. J. T. King, E. J. Arthur, C. L. Brooks, III, and K. J. Kubarych, "Crowding Induced Collective Hydration of Biological Macromolecules over Extended Distances," J. Am. Chem. Soc. **136**(1), 188-94, (2014).

20. F. Alves, F. S. Oliveira, B. Schroeder, C. Matos, and I. M. Marrucho, "Synthesis, Characterization, and Liposome Partition of a Novel Tetracycline Derivative Using the Ionic Liquids Framework," J. Pharm. Sci. **102**(5), 1504-12, (2013).

21. C. Matos, B. de Castro, P. Gameiro, J. Lima, and S. Reis, "Zeta-Potential Measurements as a Tool to Quantify the Effect of Charged Drugs on the Surface Potential of Egg Phosphatidylcholine Liposomes," Langmuir **20**(2), 369-77, (2004).

22. P. Ball, "Water as an Active Constituent in Cell Biology," Chem. Rev. **108**(1), 74-108, (2008).

23. S. K. Pal, J. Peon, and A. H. Zewail, "Biological Water at the Protein Surface: Dynamical Solvation Probed Directly with Femtosecond Resolution," Proc. Natl. Acad. Sci. U.S.A. **99**(4), 1763-68, (2002).

24. G. Stirnemann, P. J. Rossky, J. T. Hynes, and D. Laage, "Water Reorientation, Hydrogen-Bond Dynamics and 2DIR Spectroscopy Next to an Extended Hydrophobic Surface," Faraday Discuss. **146**, 263-81, (2010).

25. F. Pizzitutti, M. Marchi, F. Sterpone, and P. J. Rossky, "How Protein Surfaces Induce Anomalous Dynamics of Hydration Water," J. Phys. Chem. B **111**(26), 7584-90, (2007).

26. A. A. Bakulin, C. Liang, T. L. C. Jansen, D. A. Wiersma, H. J. Bakker, and M. S. Pshenichnikov, "Hydrophobic Solvation: A 2DIR Spectroscopic Inquest," Accounts Chem. Res. **42**(9), 1229-38, (2009).

27. M. D. Fayer, and N. E. Levinger. in *Annual Review of Analytical Chemistry, Vol 3* Vol. 3 *Annual Review of Analytical Chemistry* (eds E. S. Yeung, and R. N. Zare) 89-107 (2010).

28. L. R. Chieffo, J. T. Shattuck, E. Pinnick, J. J. Amsden, M. K. Hong, F. Wang, S. Erramilli, and L. D. Ziegler, "Nitrous Oxide Vibrational Energy Relaxation Is a Probe of Interfacial Water in Lipid Bilayers," J. Phys. Chem. B **112**(40), 12776-82, (2008).

29. L. F. Scatena, M. G. Brown, and G. L. Richmond, "Water at Hydrophobic Surfaces: Weak Hydrogen Bonding and Strong Orientation Effects," Science **292**(5518), 908-12, (2001).

30. Y. Levy, and J. N. Onuchic. in *Annu. Rev. Biophys. Biomol. Struct.* Vol. 35 *Annual Review of Biophysics* 389-415 (2006).

31. U. Sreenivasan, and P. H. Axelsen, "Buried Water in Homologous Serine Proteases," Biochemistry **31**(51), 12785-91, (1992).

32. J. T. King, and K. J. Kubarych, "Site-Specific Coupling of Hydration Water and Protein Flexibility Studied in Solution with Ultrafast 2DIR Spectroscopy," J. Am. Chem. Soc. **134**(45), 18705-12, (2012).

33. O. Rahaman, S. Melchionna, D. Laage, and F. Sterpone, "The Effect of Protein Composition on Hydration Dynamics," Phys. Chem. Chem. Phys. **15**(10), 3570-76, (2013).

34. A. J. Patel, P. Varilly, and D. Chandler, "Fluctuations of Water near Extended Hydrophobic and Hydrophilic Surfaces," J. Phys. Chem. B **114**(4), 1632-37, (2010).

35. D. Prada-Gracia, R. Shevchuk, P. Hamm, and F. Rao, "Towards a Microscopic Description of the Free-Energy Landscape of Water," J. Chem. Phys. **137**(14), (2012).

36. B. Q. Q. Wei, W. A. Baase, L. H. Weaver, B. W. Matthews, and B. K. Shoichet, "A Model Binding Site for Testing Scoring Functions in Molecular Docking," J. Mol. Biol. **322**(2), 339-55, (2002).

37. K. W. Lexa, and H. A. Carlson, "Full Protein Flexibility Is Essential for Proper Hot-Spot Mapping," J. Am. Chem. Soc. **133**(2), 200-02, (2011).

38. I. Halperin, B. Y. Ma, H. Wolfson, and R. Nussinov, "Principles of Docking: An Overview of Search Algorithms and a Guide to Scoring Functions," Proteins: Struct., Funct., Genet. **47**(4), 409-43, (2002).

39. P. A. Sigala, D. A. Kraut, J. M. M. Caaveiro, B. Pybus, E. A. Ruben, D. Ringe, G. A. Petsko, and D. Herschlag, "Testing Geometrical Discrimination within an Enzyme Active Site: Constrained Hydrogen Bonding in the Ketosterold Isomerase Oxyanion Hole," J. Am. Chem. Soc. **130**(41), 13696-708, (2008).

40. P. A. Sigala, J. M. M. Caaveiro, D. Ringe, G. A. Petsko, and D. Herschlag, "Hydrogen Bond Coupling in the Ketosteroid Isomerase Active Site," Biochemistry **48**(29), 6932-39, (2009).

41. S. Jones, and J. M. Thornton, "Principles of Protein-Protein Interactions," Proc. Natl. Acad. Sci. U.S.A. **93**(1), 13-20, (1996).

42. G. A. Papoian, J. Ulander, and P. G. Wolynes, "Role of Water Mediated Interactions in Protein-Protein Recognition Landscapes," J. Am. Chem. Soc. **125**(30), 9170-78, (2003).

43. K. Meister, S. Ebbinghaus, Y. Xu, J. G. Duman, A. DeVries, M. Gruebele, D. M. Leitner, and M. Havenith, "Long-Range Protein-Water Dynamics in Hyperactive Insect Antifreeze Proteins," Proc. Natl. Acad. Sci. U.S.A. **110**(5), 1617-22, (2013).

44. B. Jayaram, and T. Jain, "The Role of Water in Protein-DNA Recognition," Annu. Rev. Biophys. Biomol. Struct. **33**, 343-61, (2004).

45. F. Sterpone, G. Stirnemann, J. T. Hynes, and D. Laage, "Water Hydrogen-Bond Dynamics around Amino Acids: The Key Role of Hydrophilic Hydrogen-Bond Acceptor Groups," J. Phys. Chem. B **114**(5), 2083-89, (2010).

46. C. N. Nguyen, and R. M. Stratt, "Preferential Solvation Dynamics in Liquids: How Geodesic Pathways through the Potential Energy Landscape Reveal Mechanistic Details About Solute Relaxation in Liquids," J. Chem. Phys. **133**(12), (2010).

47. A. Kovalenko, and F. Hirata, "Three-Dimensional Density Profiles of Water in Contact with a Solute of Arbitrary Shape: A Rism Approach," Chem. Phys. Lett. **290**(1-3), 237-44, (1998).

48. A. Fernandez, "Epistructural Tension Promotes Protein Associations," Phys. Rev. Letters **108**(18), (2012).

49. J. T. King, E. J. Arthur, C. L. Brooks, III, and K. J. Kubarych, "Site-Specific Hydration Dynamics of Globular Proteins and the Role of Constrained Water in Solvent Exchange with Amphiphilic Cosolvents," J. Phys. Chem. B **116**(19), 5604-11, (2012).

50. T. Imai, K. Oda, A. Kovalenko, F. Hirata, and A. Kidera, "Ligand Mapping on Protein Surfaces by the 3d-Rism Theory: Toward Computational Fragment-Based Drug Design," J. Am. Chem. Soc. **131**(34), 12430-40, (2009).

51. A. J. Patel, P. Varilly, S. N. Jamadagni, H. Acharya, S. Garde, and D. Chandler, "Extended Surfaces Modulate Hydrophobic Interactions of Neighboring Solutes," Proc. Natl. Acad. Sci. U.S.A. **108**(43), 17678-83, (2011).

52. C. Ma, J. Tran, F. Gu, R. Ochoa, C. Li, D. Sept, K. Werbovetz, and N. Morrissette, "Dinitroaniline Activity in Toxoplasma Gondii Expressing Wild-Type or Mutant Alpha-Tubulin," Antimicrob. Agents Chemother. **54**(4), 1453-60, (2010).

53. S. Genheden, T. Luchko, S. Gusarov, A. Kovalenko, and U. Ryde, "An Mm/3d-Rism Approach for Ligand Binding Affinities," J. Phys. Chem. B **114**(25), 8505-16, (2010).

54. F. Sterpone, G. Stirnemann, and D. Laage, "Magnitude and Molecular Origin of Water Slowdown Next to a Protein," J. Am. Chem. Soc. **134**(9), 4116-19, (2012).

55. J. T. King, M. R. Ross, and K. J. Kubarych, "Water-Assisted Vibrational Relaxation of a Metal Carbonyl Complex Studied with Ultrafast 2DIR," J. Phys. Chem. B **116**(12), 3754-59, (2012).

56. C. T. Middleton, L. E. Buchanan, E. B. Dunkelberger, and M. T. Zanni, "Utilizing Lifetimes to Suppress Random Coil Features in 2DIR Spectra of Peptides," J. Phys. Chem. Lett. **2**(18), 2357-61, (2011).

57. J. F. Povey, C. M. Smales, S. J. Hassard, and M. J. Howard, "Comparison of the Effects of 2,2,2-Trifluoroethanol on Peptide and Protein Structure and Function," J. Struct. Biol. **157**(2), 329-38, (2007).

58. M. Buck, S. E. Radford, and C. M. Dobson, "A Partially Folded State of Hen Egg-White Lysozyme in Trifluoroethanol - Structural Characterization and Implications for Protein Folding," Biochemistry **32**(2), 669-78, (1993).

59. M. D. Diaz, and S. Berger, "Preferential Solvation of a Tetrapeptide by Trifluoroethanol as Studied by Intermolecular Noe," Magn. Reson. Chem. **39**(7), 369-73, (2001).

60. B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, "Gromacs 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation," J. Chem. Theory Comput. **4**(3), 435-47, (2008).

61. H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma, "The Missing Term in Effective Pair Potentials," J. Phys. Chem. **91**(24), 6269-71, (1987).

62. M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam,

S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, A. Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, and J. A. Pople. (2003).

63. J. Wang, W. Wang, P. A. Kollman, and D. A. Case, "Automatic Atom Type and Bond Type Perception in Molecular Mechanical Calculations," J. Mol. Graphics Modell. **25**(2), 247-60, (2006).

64. J. M. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, "Development and Testing of a General Amber Force Field," J. Comput. Chem. **25**(9), 1157-74, (2004).

65. H. J. C. Berendsen, J. P. M. Postma, W. F. Vangunsteren, A. Dinola, and J. R. Haak, "Molecular-Dynamics with Coupling to an External Bath " J. Chem. Phys. **81**(8), 3684-90, (1984).

66. T. Darden, D. York, and L. Pedersen, "Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems," J. Chem. Phys. **98**(12), 10089-92, (1993).

67. E. J. Arthur, J. T. King, K. J. Kubarych, and C. L. Brooks, III, "Heterogeneous Preferential Solvation of Water and Trifluoroethanol in Homologous Lysozymes," J. Phys. Chem. B **118**(28), 8118-27, (2014).

68. W. Humphrey, A. Dalke, and K. Schulten, "Vmd: Visual Molecular Dynamics," J. Mol. Graphics Modell. **14**(1), 33-38, (1996).

69. M. S. Lehmann, S. A. Mason, and G. J. McIntyre, "Study of Ethanol Lysozyme Interactions Using Neutron-Diffraction," Biochemistry **24**(21), 5862-69, (1985).

70. A. Hedoux, R. Ionov, J. F. Willart, A. Lerbret, F. Affouard, Y. Guinet, M. Descamps, D. Prevost, L. Paccou, and F. Danede, "Evidence of a Two-Stage Thermal

Denaturation Process in Lysozyme: A Raman Scattering and Differential Scanning Calorimetry Investigation," J. Chem. Phys. **124**(1), (2006).

71. A. J. Patel, P. Varilly, S. N. Jamadagni, M. F. Hagan, D. Chandler, and S. Garde, "Sitting at the Edge: How Biomolecules Use Hydrophobicity to Tune Their Interactions and Function," J. Phys. Chem. B **116**(8), 2498-503, (2012).

72. T. F. Smith, and M. S. Waterman, "Identification of Common Molecular Subsequences," J. Mol. Biol. **147**(1), 195-97, (1981).

73. S. Jalili, and M. Akhavan, "Molecular Dynamics Simulation Study of Association in Trifluoroethanol/Water Mixtures," J. Comput. Chem. **31**(2), 286-94, (2010).

74. J. A. Nash, T. N. S. Ballard, T. E. Weaver, and H. T. Akinbi, "The Peptidoglycan-Degrading Property of Lysozyme Is Not Required for Bactericidal Activity in Vivo," J. Immunol. **177**(1), 519-26, (2006).

Chapter 4

The Effects of Crowding on Hydration Dynamics Near Lysozymes

The work presented in this chapter has been published in the following papers:

1. J. T. King, E. J. Arthur, C. L. Brooks, III, and K. J. Kubarych, "Crowding Induced Collective Hydration of Biological Macromolecules over Extended Distances," The Journal of the American Chemical Society **136**(1), 188-194 (2014).

4.1 Introduction

The hydrophobic effect is a powerful driving force crucial in biological systems,[2] playing a key role in protein folding[3-5] and membrane formation,[6] as well as directing surface association processes.[7,8] It has been predicted[9,10] and experimentally observed[11,12] that the energetic balance of hydrophobic hydration depends on the size of the hydrated molecule. For small solutes, the cost of hydration is largely entropic as the water enhances its local structure to minimize hydrogen bond losses, while the cost of hydrating larger molecules is largely borne by enthalpic contributions as the solute forces the disruption of water's hydrogen bonding network.[10] The corresponding dynamics of the surrounding water has been more difficult to access, though experiments and simulations are converging on a view

where small hydrophobes exert negligible influence over the dynamics of the surrounding water molecules when in dilute concentrations,[13-15] while large hydrophobic solutes can constrain and hinder the surrounding water by limiting the ability of hydrogen bond exchange.[14-16] The crossover occurs on the nanometer length scale, which is characteristic of proteins, lipids, and other biomolecules.

The perturbation of water by hydrophobic structures can have significant implications in cellular environments, where the structural and dynamic correlation lengths may extend well beyond the space available from interstitial water. Crowding effects are generally considered in terms of energetics focusing on protein stability and refolding kinetics,[17-24] where entropic forces arising from hard-core repulsions between macromolecules compete with enthalpic forces arising from weak attractions. Due to the challenging nature of experiments, dynamic aspects of crowding are more elusive, though progress in new methods of spectroscopy, including time-resolved fluorescence,[25] terahertz absorption,[26,27] NMR,[28,29] and 2DIR,[15] have allowed for the interfacial region of hydrated proteins to be studied directly. In particular, studies using THz absorption spectroscopy, coupled with molecular dynamics (MD) simulations, have found evidence of a dynamic hydration shell surrounding proteins ranging from 10 to 30 Å, depending on the protein.[26,27] As a striking example, antifreeze proteins were found to have a hydration environment that can extend upward of 30 Å.[27] Additionally, photon echo experiments of hemoglobin in erythrocytes[30] and optical Kerr effect (OKE) spectroscopy,[31,32] which measures the low-frequency Raman response, have been used to observe a general slowing of the system dynamics with increasing concentrations, though no dynamic transition was apparent from the data.

Within the context of crowding, there is a dichotomy between what can broadly be classified as "chemical" and "physical" effects. For instance, studies comparing monomeric

and polymeric sucrose (Ficoll 70) arrive at different conclusions. Pielak et al.[20] observe no difference in protein stability (chymotrypsin inhibitor 2), whereas Gruebele et al.[21] find pronounced differences in folding kinetics (phosphoglycerate). Our work focuses on dynamics using a similar comparison. If the differences in chemical interactions are minimal, is there a fundamental difference between macromolecular and small molecule crowding? In order to make progress, we have discovered that it is essential to perform experiments over a wide range of additive concentrations, as will be detailed below.

Questions remain regarding the relevant length and time scales associated with crowding. While ultrafast spectroscopic studies have uncovered the strong coupling between hydration water and protein flexibility, it is still unclear over what distances this coupling can persist, and whether the disruption of water upon crowding has a structural component or if it is a purely dynamic phenomenon. If there is a crowding dependence to the hydration structure, basic statistical mechanics tells us that there will be an energetic contribution due to the altered water–water and water–protein pair correlation functions. In the absence of a structural change, however, only dynamical measurements will be able to discern a detailed microscopic picture, as is the case, for example, with studies on the glass transition. In addition, measurements of diffusion in cellular environments show a general decrease in diffusion constants upon crowding,[24,33] but it is difficult to directly relate macroscopic diffusion constants to microscopic properties of the solvent, namely local solvent friction.

To address these issues, we use ultrafast two-dimensional infrared (2DIR) to study the picosecond dynamics of HEWL labeled with a transition metal carbonyl vibrational probe covalently attached to the surface exposed His15 residue (the labeled protein is referred to as HEWL-RC).[34] Metal carbonyls offer ideal vibrational probes for biological

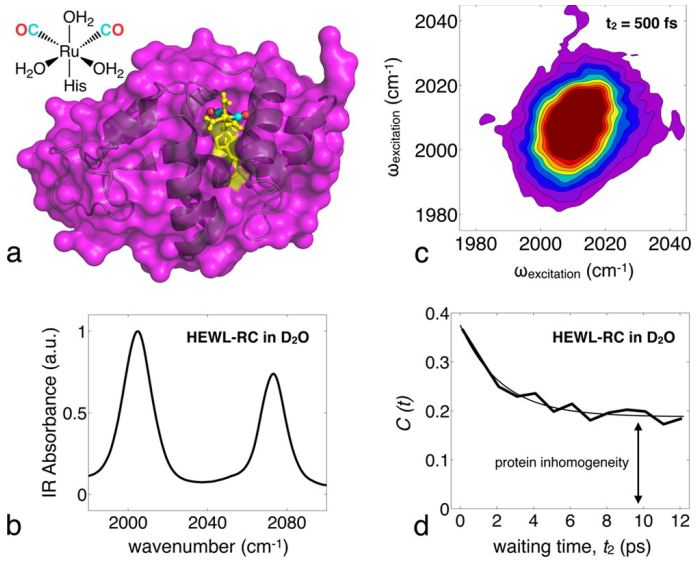

Figure 4.1 Crystal structure of HEWL-RC, linear and 2DIR spectra, example FFCF. (a) Structure of the metal–carbonyl vibrational probe and the crystal structure of the His 15 labeled HEWL carbonyl complex (probe site highlighted in yellow). (b) Linear FTIR spectrum and (c) 2DIR spectrum shown for the metal–carbonyl CO region. (d) Example of a typical frequency–frequency correlation function, showing an initial decay on the order of a few picoseconds corresponding to the hydration dynamics, followed by a static offset due to protein inhomogeneity that is not sampled within the experimental window.

molecules due to the inherent strength of the transition and the frequency of the vibrational modes, giving strong signal in a region of the IR spectrum that is free from the protein and water background.[14,15] Additionally, lysozymes are robust proteins that maintain structural integrity in crowded solutions.[35] The X-ray crystal structure of HEWL-RC is shown in Figure 4.1, as well as a linear FTIR spectrum of the C≡O modes of the vibrational probe. We study the dynamics of the system through the frequency–frequency correlation function (FFCF), a powerful observable unique to 2DIR that reports on the equilibrium structural fluctuations that modulate the transition frequency of a probe molecule. The surface location of the vibrational probe used here allows us to study both the hydration dynamics and the protein dynamics simultaneously. The FFCFs exhibit rapid initial picosecond decays due to motion of the hydration water, followed by a significant static offset arising from

fluctuations that are too slow to be fully sampled within the experimental window.[14] We attribute the static offset of the correlation function to slow protein fluctuations, though other work looking at similar correlation functions have suggested that the slow dynamics could arise from very slow exchange between surface water and bulk water.[25] While these contributions are difficult to distinguish experimentally, we believe that the spectral signatures between surface and bulk water are not as significant as the inhomogeneity arising from protein fluctuations. Since the region of the protein we probe experimentally is not located on the cleft region, but rather on an open, flat region of the protein, we do not expect idiosyncratically slow exchange of hydration water with the bulk. Simulations by Laage et al.[16] have used site-specific analysis around a protein surface and have found that the majority of the water molecules experience only a mild slowdown due to the protein surface, while a handful of water molecules located in cleft regions of the protein or in the interior, experience significant slowdown upward of 100 ps. Thus, distinct populations of hydration water can lead to mean residence times that are significantly longer than what the majority of the water experiences. Recent work on biomolecule hydration has highlighted the importance of considering metrics other than averages in describing interfacial water structure and thermodynamics.[36] In addition, slow translational motion of water from the surface to the bulk is more apparent through techniques such as NMR[28,29] and Overhauser dynamic nuclear polarization (ODNP),[37] whereas experiments that measure ultrafast correlation functions tend to be predominantly sensitive to local dynamics.

Though the vibrational relaxation of the probe precludes time resolving the protein motion, the magnitude of the static offset can be used as a proxy for the protein dynamics. Hence, a single probe's FFCF is sensitive to both the hydration and protein dynamics separately, offering a perspective that is generally not available from THz or OKE spectroscopy, where the two contributions are mixed. We measure the protein-hydration

dynamics of HEWL-RC in aqueous (D_2O) solutions of PEG400 (8–9mer) ranging from 0 to 80% PEG400 by volume, and compare these results to previously reported experiments using glycerol.[14] In addition, we carry out a parallel experiment of HEWL-RC in varying concentrations of excess lysozyme ranging from 20 to 160 mg/mL, which acts to self-crowd the labeled protein with a complex electrostatic surface, which contrasts starkly with that presented by the uncharged polymer crowder.

We present a comprehensive picture of the picosecond protein and hydration dynamics under crowding conditions. We find an abrupt dynamical transition of the protein and hydration dynamics induced by crowding, which is unique from the temperature dependent transition that is observed in hydrated proteins.[33,38] The results suggest a dynamic hydration shell around the protein extending 15–20 Å, resulting in collective hydration for interprotein separations of 30–40 Å. We also find that the collective water dynamics can be up to an order of magnitude slower than that for bulk water. In addition, we find that the presence of this transition seems to be due to the macromolecular nature of the crowding agent since it is absent in the case of solvation by glycerol/water solutions. The existence of two distinct regimes, each of which is largely dynamically decoupled from the fine details of the surrounding solvent fluctuations, suggests the partitioning of biomacromolecules into "undercrowded" and "overcrowded" conditions. Based on our measurements, many cellular environments can be classified as being "overcrowded."

4.2 Polymer Crowding

There is experimental evidence that PEG400 adopts a compact structure when in dilute aqueous solution.[39,40] For example, small angle neutron scattering results show that the

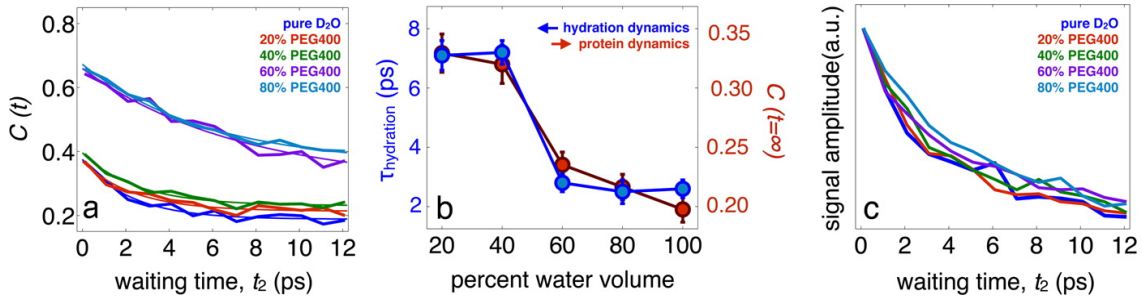

Figure 4.2 Interfacial water and protein dynamics of HEWL-RC in D_2O/PEG mixtures. (a) FFCFs for HEWL-RC in D_2O/PEG mixtures, ranging from pure D_2O to 80% PEG by volume. (b) Hydration time scale, obtained by the initial decay of the correlation function, and the protein dynamics, estimated by the static offset of the correlation function, plotted as a function of solvent composition. A strong coupling is clear from the data, with both the hydration and protein dynamics slowing down as glycerol is added to the system. There is also a sharp dynamic transition occurring at roughly 60% PEG. We suggest this transition results from the extended protein hydration environment overlapping with the PEG hydration environment. (c) The vibrational relaxation, estimated from the rephasing signal amplitude, lacks any PEG400 dependence suggesting that the protein remains fully hydrated in the region around the probe.

radius of gyration of PEG400 measured at 1% (v/v) in D_2O is 2 nm,[39,40] which is similar in size to a typical protein. The structure of PEG400 at high concentrations, however, remains unclear, though it has been proposed that the short polymer adopts an entangled structure. Nevertheless, the effect of PEG on protein and hydration dynamics should be largely due to the volume it excludes and the associated perturbation of its hydration environment, where the protein and hydration water reside largely in the pores of the entangled polymer solution.

The protein and hydration dynamics were studied in D_2O/PEG400 solvent mixtures of 0, 20, 40, 60, and 80% PEG400 v/v. Figure 4.2 shows the FFCFs for each solution and the experimental fits, consisting of a single exponential decay (due to hydration water) and a static offset (due to slow protein dynamics). In pure D_2O, the hydration dynamics occur with a 2.7 ps time constant, which is slower than that of bulk D_2O by a factor of 2. This observation has previously been reported[14] and is in quantitative agreement with MD simulations of Laage and co-workers that specifically investigated the influence of the

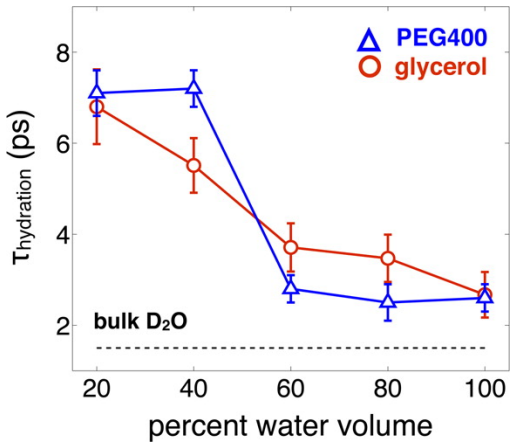

Figure 4.3 Comparison of interfacial water dynamics of HEWL-RC in solutions of glycerol and PEG400. While the magnitude of the hydration dynamics slowdown induced by each cosolvent is similar at high concentrations, the dynamic transition is observed only in the presence of the macromolecular crowding agent.

protein on extended hydrogen bond jumps of the hydration water.[16] At high PEG400 (80% v/v) concentration, the hydration dynamics slow by nearly a factor of 4, and the protein contribution increases by about 75% relative to pure D_2O. Surprisingly, a dynamic transition is observed around 50% D_2O where there is a significant, abrupt slowing of the protein-hydration dynamics. On either side of this transition, the protein and hydration dynamics are only weakly coupled to the polymer concentration, though the protein dynamics and the hydration dynamics stay strongly coupled to each other at all solvent compositions (evident from the correlation between $\tau_{hydration}$ and $C(t = \infty)$ in Figure 4.2). To ensure the protein is not dehydrated by PEG, at least in the local region of the probe molecule, we use the vibrational lifetime, which we have shown to be a unique observable capable of reporting on local hydration levels.[15] The lifetimes shown in Figure 4.4 exhibit decay times consistent with water-assisted relaxation at all PEG400 concentrations, ensuring that the local area of the probe remains fully hydrated.

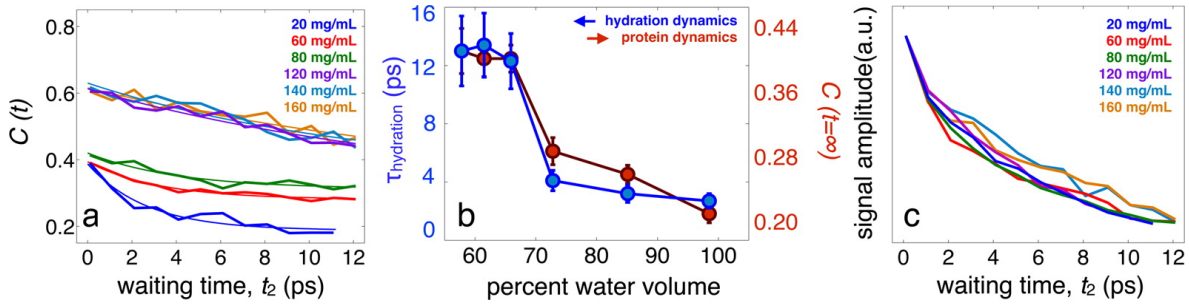

Figure 4.4 Interfacial water and protein dynamics of HEWL-RC in the presence of excess lysozyme. (a) FFCFs for HEWL-RC in self-crowding conditions, ranging from 20 to 160 mg/mL. (b) Hydration time scale, obtained by the initial decay of the correlation function, and the protein dynamics, estimated by the static offset of the correlation function, plotted as a function of solvent composition. A strong coupling is clear from the data, with both the hydration and protein dynamics slowing down as excess lysozyme is added to the system. Similar to the PEG400 crowding, a dynamical transition is observed at sufficient crowding, though this transition occurs at lower concentrations of HEWL because of the more significant constraining effect that HEWL has on surrounding waters. (c) Vibrational lifetimes estimated through the signal amplitude of the rephasing spectrum again show a consistently short lifetime, consistent with a lack of protein–protein interactions that would result in surface dehydration and increased lifetimes.[1-64]

These results are fundamentally different from previous observations made on HEWL-RC in D_2O/glycerol solutions.[14] In those experiments, we observed a gradual, uniform slowdown of the protein-hydration dynamics as a function of glycerol concentration, with no clear signs of a dynamical transition. Additionally, the slowdown in hydration dynamics was significantly more mild than what would be expected for the viscosity increase, demonstrating a weak coupling between interfacial water and the bulk solution. It is noteworthy that similar nonlinear scaling of interfacial water around liposomes has recently been observed using an NMR-based technique, Ovehauser dynamic nuclear polarization, which measures hydration water through the incorporation of a free-radical probe.[37] For our current and previous results, a comparison of the interfacial water dynamics is shown in Figure 4.3. The influence of either glycerol or PEG400 on the hydration dynamics has similarities and differences. While the magnitude of the slowdown induced by high concentrations of either cosolvent is similar, and thus the coupling between the

interfacial water and bulk solvent remains weak,[14,37] the presence of a dynamic transition is observed only with the macromolecular crowding agent.

4.3 Self-Crowding

The presence of surface charges, site-specific interactions, and an intricate surface topology makes proteins a more complex and biologically relevant crowding agent than a simple polymer. Additionally, proteins have well-defined structures that are often not significantly perturbed by concentration, which is not necessarily the case for PEG400. Here, we use unlabeled lysozyme to serve as the crowding agent to determine if the presence of a critical crowding level could exist in cell-like environments. In addition to providing a more realistic crowding agent, the well-defined shape and structure of lysozyme allows for the protein–protein distances to be estimated for a given concentration of protein.

A starting solution of HEWL-RC was prepared at 20 mg/mL, then excess lysozyme was added to concentrations up to 160 mg/mL. As before, we use the vibrational lifetime (Figure 4.4c) to ensure that no protein–protein contacts alter the hydration of the protein surface. Similar to the PEG400 data, the vibrational lifetimes exhibit negligible lysozyme concentration dependence, suggesting that the protein remains fully hydrated.

The FFCFs and fit parameters are shown in Figure 4.4. As with PEG400, there is a clear dynamic transition. The transition occurs at a higher water composition (~70%) than with the PEG400 crowding agent, which is attributed to the more significant constraining effect of HEWL on the surrounding waters. This view is supported by the fact that lysozyme is a highly charged (pI = 11) protein at neutral pH, and the dynamical constraints placed on the hydration water reduces the local dielectric, effectively extending the electrostatic

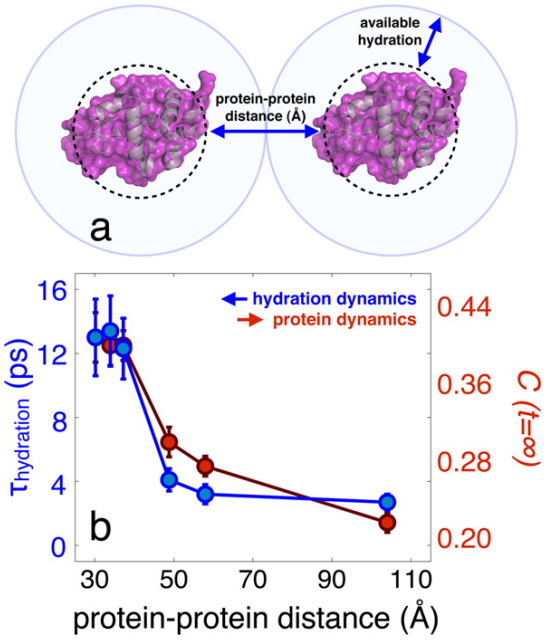

Figure 4.5 Hydration and protein dynamics of HEWL-RC in crowding conditions plotted as a function of protein–protein distance. (a) The protein–protein distance is defined as the average surface-to-surface distance between proteins using a spherical approximation, which can be estimated for each concentration. (b) Assuming a homogeneous mixture, the average surface-to-surface distance between proteins can be estimated, revealing that the transition occurs at a protein–protein distance of 30–40 Å.

footprint of the protein.[41] Assuming a homogeneous mixture[42] and a spherical approximation to the volume (computed using the van der Waals surface) to approximate the size of lysozyme, we estimated the typical protein–protein distances (surface-to-surface) at each crowding concentration. This distance is only an idealized estimate assuming homogeneous protein solution, and should be viewed as an estimated upper limit.[43] Plotting the protein-hydration dynamics in terms of our estimated protein–protein distance (Figure 4.5) reveals that this transition occurs at distances around 30–40 Å, suggesting a dynamical influence of the hydration water extending upward of 15–20 Å extending from each surface.

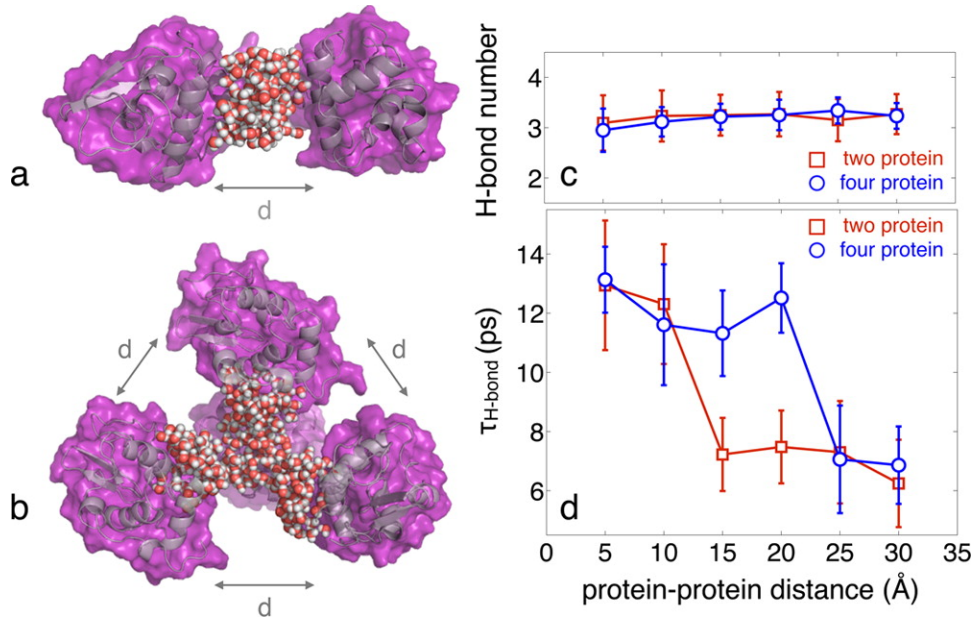

Figure 4.6 Example of the simulation analysis where (a) two proteins are separated by a set distance *d* and the bridging water is selected for analysis and (b) four proteins are arranged tetrahedrally, all of which are separated by the same variable distance. The water that was selected for analysis is shown. (c) Hydrogen bond number of the crowded water as a function of protein–protein distance. In each case, there is no clear transition in the average hydrogen bonds per water molecule, suggesting no significant change in structure. A slight downward trend is observed as the interprotein distance is reduced, though this is the result of a higher relative contribution from the interfacial water, which has fewer hydrogen bonds than bulk water. (d) Hydrogen bond correlation times of the crowded water as a function of protein–protein distance. The occurrence of a dynamic transition is found between 10 and 15 Å for two proteins and 20–25 Å for the four protein simulation. In each case, only a weak coupling is observed before and after the dynamic transition. The results not only demonstrate a percolation-like transition of water dynamics upon crowding, but also show that the distance of this transition is a function of the degree and geometry of crowding.

4.4 Molecular Dynamics Simulations

The experimental results motivated efforts to simulate hydration dynamics of proteins under crowding conditions. In one case, two proteins are separated by a variable distance, and in the other case four proteins are arranged tetrahedrally, with surfaces separated by a variable distance (Figure 4.6a,b). Each configuration (two- and four-protein geometries) was replicated in six individual simulations with average surface-to-surface distances ranging from 5 to 30 Å (see the methods section for more details).[43] The water

between the protein structures was selected for analysis of both the average hydrogen bond number (Figure 4.6c) as well as the hydrogen bond correlation time (Figure 4.6d). The hydrogen bond correlation time is reported as the 1/e time, alleviating complications of fitting a nonexponential relaxation. For the hydrogen bond number, there is a slight decrease at small interprotein distances reflecting the proportional increase in interfacial water, which exhibits reduced hydrogen bonding relative to bulk water. There is no observed threshold behavior in the extent of hydrogen bonding, suggesting a lack of significant structural changes of the water upon crowding. The dynamics of the water, however, show a very strong dependence on the crowding, including a dynamic transition that occurs at a critical protein–protein separation. In the two-protein simulation, this critical distance was found to be 10–15 Å, whereas in the four-protein simulation we observed this transition at 20–25 Å. This distance is consistent with what was observed experimentally (30–40 Å), though there is a clear dependence of the dynamic transition on the configuration and geometry of crowding. The decoupling of the dynamics from crowding above and below the dynamic transition observed experimentally is also evident in the simulations.

Surprisingly, the dynamic transition is accompanied by no significant net structural changes, as seen in the average hydrogen bonding number of the interfacial water remaining constant (Figure 4.6c). The lack of any clear structural signature accompanying the dynamical transition is similar to glassy[44] and jammed systems.[45] Our observations are the first examples of a purely dynamical transition induced by macromolecular crowding. The lack of a significant change in the degree of hydrogen bonding differs qualitatively from previous studies based largely on inelastic neutron scattering experiments. At hydration levels over an order of magnitude lower than what we consider here, there is clear evidence for a pronounced change in water structure.[46-48] Using comparisons with simulation, several workers have identified percolation transitions, where at a threshold hydration level, there is

a significant increase in the size of the largest hydrogen bonded cluster solvating the protein.[46] Such abrupt structural changes leading to hydrogen bonding networks that span large areas of the protein–water interface have been interpreted in terms of percolation theory. The new dynamical transition that we have identified here is distinct from these previous observations of hydrogen bond percolation.[33] Most importantly, our highest protein concentration is 160 mg/mL, which corresponds to a hydration level (h = mass of D_2O/mass of lysozyme) of h = 6. Neutron scattering experiments and accompanying simulations are carried out at hydration levels less than h = 1. These studies on hydrated protein powders represent an extreme case of crowding, while here the studies were performed on more dilute aqueous solutions. The dynamical transition that we observe occurs at comparably much larger values of protein hydration, highlighting the subtle nature of the collective hydration leading to a transition of the hydration water dynamics without significantly distorting its structure.

4.5 Discussion on Hydrogen Bond Networks

Water is capable of forming extensive hydrogen bonding networks that reorganize in a collective manner through an angular jump mechanism.[49,50] Furthermore, the barrier to hydrogen bond jumps is dominated by entropic contributions arising from the availability of hydrogen bonding accepting partners.[51] Hydrogen bond exchange dynamics can be stifled by limiting the configuration space available for accepting waters, and thus larger hydrophobic molecules are capable of hindering hydrogen bond dynamics while small hydrophobes have a negligible effect.[13] The collective nature of hydrogen bond motion can lead to spatially extended dynamic perturbations, inducing long-range coupling effects in crowded environments.[26] Extended collective motion of water over distances of 30–40 Å has been

observed not only in crowded protein solutions,[27] but also in water pools confined within reverse micelles.[52] In each case, the transition to collective water motion is found to be abrupt.

The measured retardation factors of crowded water are roughly 5 and 10 for PEG400 and lysozyme, respectively, relative to bulk D_2O. Given that the expected concentrations of macromolecules inside of cells is on the order of 300 mg/mL,[53] the experimental results suggest that the majority of water within cells is involved in slow, collective hydration, with only trace amounts of "bulklike" water present, despite 50–70% water content by volume. The long-range disruption of water dynamics around macromolecules is likely to be a general property of compact proteins, and particular proteins, such as antifreeze proteins,[27] may leverage the perturbation to carry out a function.

Since this study primarily investigates dynamics as sensed on the picosecond time scale, there is no straightforward link to biological function, which spans a vast range of time scales.[54] However, recent work has suggested fast fluctuations of proteins have significant implications on longer time scale dynamics, such as conformational sampling[55] and possibly enzyme activity.[54] Due to the strong coupling between the low-frequency fluctuations of proteins and the hydration water,[14,56-58] the observed jamming-like transition of the water is accompanied by a transition in the fast protein dynamics. In crowded environments, these low-frequency modes are significantly slowed from what they would be in solutions with excess water (Figures 2b and 4b). The collective hydration environment in crowded conditions effectively increases the viscosity felt by the protein, and thus, the protein undergoes pronounced slowing at a critical crowding concentration. Based on our estimated macromolecular crowder concentration threshold, it would appear that most, if not all, regions of the cell are "overcrowded."

4.6 Conclusions

We carried out two parallel experiments measuring the protein-hydration dynamics of HEWL-RC in both solutions of D_2O/PEG400 and solutions of excess lysozyme to act as crowding agents. From the experimental results we draw three conclusions. (1) Both PEG and protein crowders induce a dynamical transition, where the coupled protein-hydration dynamics exhibit a sharp slowdown above a critical degree of crowding indicative of an independent-to-collective hydration transition. It is observed that water in sufficiently crowded environments is roughly an order of magnitude slower than bulk water. (2) Using the results from self-crowding, we estimate that the distance between protein surfaces at which this transition occurs is 30–40 Å, which is a striking manifestation of the collective and coordinated behavior of strongly hydrogen bonding environments. (3) The macromolecular nature of the crowder is essential as demonstrated through comparisons between PEG400 crowding and previously reported glycerol/water solutions. While similar degrees of slowing are found at high concentrations of both, the presence of a dynamical transition is observed only in the PEG400 experiments. Simulation results confirm the experimental findings, while introducing an additional observation. In contrast to previous studies of protein hydration, where hydrogen bonding in the hydrating water is perturbed by the protein, our simulations indicate no significant changes in hydrogen bonding. Rather, the observed and simulated abrupt transition is purely dynamical in nature, and reflects the long-range influence of protein surface-induced constraints on water's orientational flexibility.

These results suggest that little to no "bulklike" water is present within cellular environments. Instead, biological macromolecules are hydrated by significantly constrained water that that in turn can strongly modulate the flexibility and dynamics of the biomolecules. Future work will be dedicated to studying the connection between the

picosecond dynamics of the hydration water, which we suggest to be the origin of dynamical crowding effects, on much longer processes, such as protein folding and catalytic activity. The partitioning of hydration dynamics into two apparent regimes suggests that large scale implicit solvent simulations of biomolecules may be able to produce realistic dynamics by adopting distance-dependent frictional damping. Based on our observations of distinct under- and overcrowded regimes, perhaps as few as two macromolecule-specific friction values are needed to capture the essential dynamical contrast between isolated and crowded macromolecules. Macromolecule-modified hydration dynamics has also been related to a change in the local dielectric constant,[41,59] a quantity which enters both the generalized Born model of solvation[60] as well as the accurate estimates of donor–acceptor distances in Förster resonant energy transfer experiments.[61] In both cases, the distance dependent solvation dynamics may produce qualitative deviations from conventional models based on a homogeneous dielectric continuum. With new methods such as site-specific 2DIR and other techniques, it is becoming clear that the complexity of biomolecule hydration can be addressed experimentally and linked directly to simulation, likely providing insight into the active nature of water in mediating biological processes.

4.7 Methods

Protein Labeling

Hen egg white lysozyme (HEWL) was purchased from Sigma Aldirich (bioultra, >98%). No further purification steps were taken. HEWL (approximately 2 mg/mL) was then combined in a 1:1 ratio with tricarbonylchloro(glycinato)ruthenium(II) in D_2O (Sigma) and stirred at room temperature for 1 h. The resulting labeled protein we refer to as HEWL-RC. The resulting product was purified in a desalting column (GE Healthcare, PD-10

Disposable Desalting Column), which removes unreacted tricarbonylchloro(glycinato)-ruthenium(II). The reaction was carried out on the morning of the experiments, and no HEWL-RC was stored to be used at a later date.[34]

2DIR Spectroscopy

Mid-IR pulses are generated through two home-built dual stage optical parametric amplifiers (OPAs) coupled with difference frequency generation (DFGs) which are pumped with a regeneratively amplified Ti:Sapphire laser. The mid-IR pulses are then split into fields E_1, E_2, E_3, and E_{LO} with respective wavevectors k_1, k_2, k_3, and k_{LO} (75 fs, 150 cm^{-1} bandwidth, 400 nJ/pulse), where the first three pulses are focused onto the sample in a box geometry to generate a third-order nonlinear signal, and the final pulse is used for heterodyne detection. We implement an upconversion detection technique that mixes a highly chirped pulse centered at 800 nm and fwhm = 160 ps with the mid-IR signal and local oscillator in a sum-frequency crystal (MgO doped $LiNbO_3$) to allow for detection in the visible with a silicon CCD camera. The detection frequency of the 2DIR spectrum is provided by the spectrometer. The excitation frequency is measured by scanning the time delay between the first two pulses and then Fourier transforming over the generated coherence period. A series of 2D spectra are then acquired as a function of waiting time between the excitation pulse pair and the detection pulse, which is stepped from 0 to 12 ps.

Molecular Dynamics Simulations

Simulations were designed using previously-shown methods.[43] Protein crowding was simulated by analyzing the interstitial water of two protein configurations: two proteins near each other, and four proteins in a packed tetrahedral configuration. Six replicas of each configuration were made by varying the average separation between protein surfaces into a

gradient of distances: 5, 10, 15, 20, 25, and 30 Å.

Hydrogen-bond (HB) autocorrelation functions and the average number of HB partners per water were calculated for interstitial water. Cutoffs for HB partners were defined as acceptor–donor distances of less than 3.5 Å (O–O distance), and acceptor–donor-hydrogen angles of less than 30° as outlined by Skinner et al.[62] The center of each protein was calculated as the mean position of all protein atoms. For each protein in each replica, spheres of water each with a radius of 10 Å were selected around the protein atom closest to the overall center. Hydrogen-bond autocorrelation functions were calculated at 15 ps intervals using the g_hbond utility from GROMACS. These functions were averaged to obtain the mean 1/e time. The average number of hydrogen bonds per water were calculated for each saved frame using in-house MATLAB code and the cutoff criteria detailed previously.

4.8 References

1. P. A. Sigala, J. M. M. Caaveiro, D. Ringe, G. A. Petsko, and D. Herschlag, "Hydrogen Bond Coupling in the Ketosteroid Isomerase Active Site," Biochemistry **48**(29), 6932-39, (2009).

2. P. Ball, "Water as an Active Constituent in Cell Biology," Chem. Rev. **108**(1), 74-108, (2008).

3. A. Nicholls, K. A. Sharp, and B. Honig, "Protein Folding and Association - Insights from the Interfacial and Thermodynamic Properties of Hydrocarbons," Proteins: Struct., Funct., Genet. **11**(4), 281-96, (1991).

4. P. Liu, X. H. Huang, R. H. Zhou, and B. J. Berne, "Observation of a Dewetting Transition in the Collapse of the Melittin Tetramer," Nature **437**(7055), 159-62, (2005).

5. M. S. Cheung, A. E. Garcia, and J. N. Onuchic, "Protein Folding Mediated by Solvation: Water Expulsion and Formation of the Hydrophobic Core Occur after the Structural Collapse," Proc. Natl. Acad. Sci. U.S.A. **99**(2), 685-90, (2002).

6. O. P. Hamill, and B. Martinac, "Molecular Basis of Mechanotransduction in Living Cells," Physiol. Rev. **81**(2), 685-740, (2001).

7. T. N. Bhat, G. A. Bentley, G. Boulot, M. I. Greene, D. Tello, W. Dallacqua, H. Souchon, F. P. Schwarz, R. A. Mariuzza, and R. J. Poljak, "Bound Water-Molecules and Conformational Stabilization Help Mediate an Antigen-Antibody Association," Proc. Natl. Acad. Sci. U.S.A. **91**(3), 1089-93, (1994).

8. C. J. Tsai, S. L. Lin, H. J. Wolfson, and R. Nussinov, "Studies of Protein-Protein Interfaces: A Statistical Analysis of the Hydrophobic Effect," Protein Sci. **6**(1), 53-64, (1997).

9. K. Lum, D. Chandler, and J. D. Weeks, "Hydrophobicity at Small and Large Length Scales," J. Phys. Chem. B **103**(22), 4570-77, (1999).

10. D. Chandler, "Interfaces and the Driving Force of Hydrophobic Assembly," Nature **437**(7059), 640-47, (2005).

11. I. T. S. Li, and G. C. Walker, "Signature of Hydrophobic Hydration in a Single Polymer," Proc. Natl. Acad. Sci. U.S.A. **108**(40), 16527-32, (2011).

12. J. G. Davis, K. P. Gierszal, P. Wang, and D. Ben-Amotz, "Water Structural Transformation at Molecular Hydrophobic Interfaces," Nature **491**(7425), 582-85, (2012).

13. D. Laage, G. Stirnemann, and J. T. Hynes, "Why Water Reorientation Slows without Iceberg Formation around Hydrophobic Solutes," J. Phys. Chem. B **113**(8), 2428-35, (2009).

14. J. T. King, and K. J. Kubarych, "Site-Specific Coupling of Hydration Water and Protein Flexibility Studied in Solution with Ultrafast 2DIR Spectroscopy," J. Am. Chem. Soc. **134**(45), 18705-12, (2012).

15. J. T. King, E. J. Arthur, C. L. Brooks, III, and K. J. Kubarych, "Site-Specific Hydration Dynamics of Globular Proteins and the Role of Constrained Water in Solvent Exchange with Amphiphilic Cosolvents," J. Phys. Chem. B **116**(19), 5604-11, (2012).

16. F. Sterpone, G. Stirnemann, and D. Laage, "Magnitude and Molecular Origin of Water Slowdown Next to a Protein," J. Am. Chem. Soc. **134**(9), 4116-19, (2012).

17. A. P. Minton, "Excluded Volume as a Determinant of Macromolecular Structure and Reactivity," Biopolymers **20**(10), 2093-120, (1981).

18. S. B. Zimmerman, and A. P. Minton, "Macromolecular Crowding - Biochemical, Biophysical, and Physiological Consequences," Annu. Rev. Biophys. Biomol. Struct. **22**, 27-65, (1993).

19. Y. Wang, M. Sarkar, A. E. Smith, A. S. Krois, and G. J. Pielak, "Macromolecular Crowding and Protein Stability," J. Am. Chem. Soc. **134**(40), 16614-18, (2012).

20. L. A. Benton, A. E. Smith, G. B. Young, and G. J. Pielak, "Unexpected Effects of Macromolecular Crowding on Protein Stability," Biochemistry **51**(49), 9773-75, (2012).

21. A. Dhar, A. Samiotakis, S. Ebbinghaus, L. Nienhaus, D. Homouz, M. Gruebele, and M. S. Cheung, "Structure, Function, and Folding of Phosphoglycerate Kinase Are Strongly Perturbed by Macromolecular Crowding," Proc. Natl. Acad. Sci. U.S.A. **107**(41), 17586-91, (2010).

22. M. S. Cheung, D. Klimov, and D. Thirumalai, "Molecular Crowding Enhances Native State Stability and Refolding Rates of Globular Proteins," Proc. Natl. Acad. Sci. U.S.A. **102**(13), 4753-58, (2005).

23. M. Feig, and Y. Sugita, "Variable Interactions between Protein Crowders and Biomolecular Solutes Are Important in Understanding Cellular Crowding," J. Phys. Chem. B **116**(1), 599-605, (2012).

24. S. R. McGuffee, and A. H. Elcock, "Diffusion, Crowding & Protein Stability in a Dynamic Molecular Model of the Bacterial Cytoplasm," PLoS Comput. Biol. **6**(3), (2010).

25. W. Qiu, Y.-T. Kao, L. Zhang, Y. Yang, L. Wang, W. E. Stites, D. Zhong, and A. H. Zewail, "Protein Surface Hydration Mapped by Site-Specific Mutations," Proc. Natl. Acad. Sci. U.S.A. **103**(38), 13979-84, (2006).

26. S. Ebbinghaus, S. J. Kim, M. Heyden, X. Yu, U. Heugen, M. Gruebele, D. M. Leitner, and M. Havenith, "An Extended Dynamical Hydration Shell around Proteins," Proc. Natl. Acad. Sci. U.S.A. **104**(52), 20749-52, (2007).

27. K. Meister, S. Ebbinghaus, Y. Xu, J. G. Duman, A. DeVries, M. Gruebele, D. M. Leitner, and M. Havenith, "Long-Range Protein-Water Dynamics in Hyperactive Insect Antifreeze Proteins," Proc. Natl. Acad. Sci. U.S.A. **110**(5), 1617-22, (2013).

28. N. V. Nucci, M. S. Pometun, and A. J. Wand, "Site-Resolved Measurement of Water-Protein Interactions by Solution Nmr," Nat. Struct. Mol. Biol. **18**(2), 245-49, (2011).

29. N. V. Nucci, M. S. Pometun, and A. J. Wand, "Mapping the Hydration Dynamics of Ubiquitin," J. Am. Chem. Soc. **133**(32), 12326-29, (2011).

30. B. L. McClain, I. J. Finkelstein, and M. D. Fayer, "Dynamics of Hemoglobin in Human Erythrocytes and in Solution: Influence of Viscosity Studied by Ultrafast Vibrational Echo Experiments," J. Am. Chem. Soc. **126**(48), 15702-10, (2004).

31. K. Mazur, I. A. Heisler, and S. R. Meech, "Water Dynamics at Protein Interfaces: Ultrafast Optical Kerr Effect Study," J. Phys. Chem. A **116**(11), 2678-85, (2012).

32. N. T. Hunt, L. Kattner, R. P. Shanks, and K. Wynne, "The Dynamics of Water-Protein Interaction Studied by Ultrafast Optical Kerr-Effect Spectroscopy," J. Am. Chem. Soc. **129**(11), 3168-72, (2007).

33. W. Doster, S. Cusack, and W. Petry, "Dynamical Transition of Myoglobin Revealed by Inelastic Neutron-Scattering," Nature **337**(6209), 754-56, (1989).

34. T. Santos-Silva, A. Mukhopadhyay, J. D. Seixas, G. J. L. Bernardes, C. C. Romao, and M. J. Romao, "CORM-3 Reactivity toward Proteins: The Crystal Structure of a Ru(Ii) Dicarbonyl-Lysozyme Complex," J. Am. Chem. Soc. **133**(5), 1192-95, (2011).

35. L. Huang, R. Jin, J. Li, K. Luo, T. Huang, D. Wu, W. Wang, R. Chen, and G. Xiao, "Macromolecular Crowding Converts the Human Recombinant Prpc to the Soluble Neurotoxic Beta-Oligomers," FASEB J. **24**(9), 3536-43, (2010).

36. A. J. Patel, P. Varilly, S. N. Jamadagni, M. F. Hagan, D. Chandler, and S. Garde, "Sitting at the Edge: How Biomolecules Use Hydrophobicity to Tune Their Interactions and Function," J. Phys. Chem. B **116**(8), 2498-503, (2012).

37. J. M. Franck, J. A. Scott, and S. Han, "Nonlinear Scaling of Surface Water Diffusion with Bulk Water Viscosity of Crowded Solutions," J. Am. Chem. Soc. **135**(11), 4175-78, (2013).

38. M. Tarek, and D. J. Tobias, "Role of Protein-Water Hydrogen Bond Dynamics in the Protein Dynamical Transition," Phys. Rev. Letters **88**(13), (2002).

39. K. L. Linegar, A. E. Adeniran, A. F. Kostko, and M. A. Anisimov, "Hydrodynamic Radius of Polyethylene Glycol in Solution Obtained by Dynamic Light Scattering," Colloid J. **72**(2), 279-81, (2010).

40. G. Lancz, M. V. Avdeev, V. I. Petrenko, V. M. Garamus, M. Koneracka, and P. Kopcansky, "Sans Study of Poly(Ethylene Glycol) Solutions in D_2O," Acta Phys. Pol., A **118**(5), 980-82, (2010).

41. F. Despa, A. Fernandez, and R. S. Berry, "Dielectric Modulation of Biological Water," Phys. Rev. Letters **93**(22), (2004).

42. A. Shukla, E. Mylonas, E. Di Cola, S. Finet, P. Timmins, T. Narayanan, and D. I. Svergun, "Absence of Equilibrium Cluster Phase in Concentrated Lysozyme Solutions," Proc. Natl. Acad. Sci. U.S.A. **105**(13), 5075-80, (2008).

43. J. T. King, E. J. Arthur, C. L. Brooks, III, and K. J. Kubarych, "Crowding Induced Collective Hydration of Biological Macromolecules over Extended Distances," J. Am. Chem. Soc. **136**(1), 188-94, (2014).

44. C. A. Angell, "Formation of Glasses from Liquids and Biopolymers," Science **267**(5206), 1924-35, (1995).

45. V. Trappe, V. Prasad, L. Cipelletti, P. N. Segre, and D. A. Weitz, "Jamming Phase Diagram for Attractive Particles," Nature **411**(6839), 772-75, (2001).

46. A. Oleinikova, I. Brovchenko, N. Smolin, A. Krukau, A. Geiger, and R. Winter, "Percolation Transition of Hydration Water: From Planar Hydrophilic Surfaces to Proteins," Phys. Rev. Letters **95**(24), (2005).

47. N. Smolin, A. Oleinikova, I. Brovchenko, A. Geiger, and R. Winter, "Properties of Spanning Water Networks at Protein Surfaces," J. Phys. Chem. B **109**(21), 10995-1005, (2005).

48. H. Nakagawa, and M. Kataoka, "Percolation of Hydration Water as a Control of Protein Dynamics," J. Phys. Soc. Jpn. **79**(8), (2010).

49. J. D. Eaves, J. J. Loparo, C. J. Fecko, S. T. Roberts, A. Tokmakoff, and P. L. Geissler, "Hydrogen Bonds in Liquid Water Are Broken Only Fleetingly," Proc. Natl. Acad. Sci. U.S.A. **102**(37), 13019-22, (2005).

50. D. Laage, and J. T. Hynes, "A Molecular Jump Mechanism of Water Reorientation," Science **311**(5762), 832-35, (2006).

51. D. Laage, and J. T. Hynes, "Do More Strongly Hydrogen-Bonded Water Molecules Reorient More Slowly ?," Chem. Phys. Lett. **433**(1-3), 80-85, (2006).

52. D. E. Moilanen, E. E. Fenn, D. Wong, and M. D. Fayer, "Water Dynamics in Large and Small Reverse Micelles: From Two Ensembles to Collective Behavior," J. Chem. Phys. **131**(1), (2009).

53. S. B. Zimmerman, and S. O. Trach, "Estimation of Macromolecule Concentrations and Excluded Volume Effects for the Cytoplasm of Escherichia-Coli," J. Mol. Biol. **222**(3), 599-620, (1991).

54. K. A. Henzler-Wildman, M. Lei, V. Thai, S. J. Kerns, M. Karplus, and D. Kern, "A Hierarchy of Timescales in Protein Dynamics Is Linked to Enzyme Catalysis," Nature **450**(7171), 913-U27, (2007).

55. H. Vashisth, and C. L. Brooks, III, "Conformational Sampling of Maltose-Transporter Components in Cartesian Collective Variables Is Governed by the Low-Frequency Normal Modes," J. Phys. Chem. Lett. **3**(22), 3379-84, (2012).

56. V. Lubchenko, P. G. Wolynes, and H. Frauenfelder, "Mosaic Energy Landscapes of Liquids and the Control of Protein Conformational Dynamics by Glass-Forming Solvents," J. Phys. Chem. B **109**(15), 7488-99, (2005).

57. H. Frauenfelder, G. Chen, J. Berendzen, P. W. Fenimore, H. Jansson, B. H. McMahon, I. R. Stroe, J. Swenson, and R. D. Young, "A Unified Model of Protein Dynamics," Proc. Natl. Acad. Sci. U.S.A. **106**(13), 5129-34, (2009).

58. N. Q. Vinh, S. J. Allen, and K. W. Plaxco, "Dielectric Spectroscopy of Proteins as a Quantitative Experimental Test of Computational Models of Their Low-Frequency Harmonic Motions," J. Am. Chem. Soc. **133**(23), 8942-47, (2011).

59. R. Harada, Y. Sugita, and M. Feig, "Protein Crowding Affects Hydration Structure and Dynamics," J. Am. Chem. Soc. **134**(10), 4842-49, (2012).

60. D. Bashford, and D. A. Case, "Generalized Born Models of Macromolecular Solvation Effects," Annu. Rev. Phys. Chem. **51**(1), 129-52, (2000).

61. D. Beljonne, C. Curutchet, G. D. Scholes, and R. J. Silbey, "Beyond Forster Resonance Energy Transfer in Biological and Nanoscale Systems," J. Phys. Chem. B **113**(19), 6583-99, (2009).

62. R. Kumar, J. R. Schmidt, and J. L. Skinner, "Hydrogen Bonding Definitions and Dynamics in Liquid Water," J. Chem. Phys. **126**(20), (2007).

Chapter 5

Predicting pK_a Shifts Using Constant pH Molecular Dynamics

The work presented in this chapter has been published in the following papers:

1. E. J. Arthur, J. D. Yesselman, and C. L. Brooks, III, "Predicting extreme pK_a shifts in staphylococcal nuclease mutants with constant pH molecular dynamics," Proteins: Structure, Function, and Bioinformatics **79**(12), 3276-3286 (2011).

5.1 Introduction

Stability and function of many proteins and nucleic acids are dependent on the charge of titratable residues. Changes in the protonation state of these residues have the potential to trigger significant configurational variation. Some examples include the proton-gradient in mitochondria, which enables the rotary motion of ATP synthetase for virtually all known metabolizing life forms.[1,2] In addition, the catalytic mechanisms of numerous enzymes are driven by locally perturbed protonation equilibria at the active site.[3] Furthermore, amyloidogenic protein aggregation into oligomers is a pH driven process, demonstrating the role of ionization states in protein function.[4,5] To study these biological mechanisms, it is crucial to understand how they are dependent on the ionization states of

their amino acid residues.

Understanding these phenomena requires a system that describes the complex coupling between structure, chemical composition, and proton affinities as a function of proton concentration (pH). Residue-specific pK_a values provide a framework from which to begin to provide quantitative relationships among the above noted properties. However, the pK_a of a particular site and its tendency to ionize or accept a proton is highly responsive to the surrounding solvent environment as well as to charge–dipole and charge–charge interactions.[6-8] These in turn alter the specific tendency for that residue to change its ionization state, i.e., its pK_a. For extreme cases, such as aspartic acid (ASP)-96 in bacteriorhodopsin, the measured perturbation is at least 8.0 pK units greater than that of the isolated amino acid in pure water.[3] This creates a need for measuring the relative amino acid pK_a perturbations in a folded protein. Determining these experimentally, however, is nontrivial, although possible through a range of techniques.[9]

Experimentally investigating pK_a values involves titrating a species over a wide range of pH.[9] Most biologically functional proteins, however, are natively folded only within a very narrow pH range. Outside of these native conditions they often adopt non-native, denatured, or unfolded conformations. Since the pK_a values of an ionizable residue are highly dependent on its interactions with solvent and surrounding protein tertiary structures, titrating a protein to pH values outside of this range may not provide pK_a values relevant to its natively folded configuration.[7] To aid in both the calculation and interpretation of such experiments, theoretical tools have been developed to make pK_a predictions based on knowledge of the native protein structure. For many proteins, a reliable method of experimentally determining residue-specific pK_a values is either too cost prohibitive, or infeasible. Before such experimental methods become viable, computational tools are the

only means available for studying their pK_a values.[7,8]

The theoretical framework and computational methods to predict pK_a shifts in large molecules can be divided into three basic approaches: finite difference Poisson-Boltzmann based continuum electrostatics methods, empirical methods, and molecular dynamics (MD) coupled with explicit free energy estimates using explicit solvent or implicit solvent (generalized Born continuum electrostatics) methods. Empirical methods, such as PROPKA,[10,11] are based upon empirical algorithms that relate structural metrics to pK_a perturbation. Provided with sufficient relevant experimental data and an accurate structure of a protein, this method has been shown to yield predictions within 1 pK_a unit root-mean-squared deviation (RMSD) from experimental observation. This level of agreement with experimental pK_a values shows that the corresponding link between structural metrics and pK_a shifts is an important tool in understanding the electrostatic environments of proteins. Empirical methods, however, cannot be used to determine pK_a values without both extensive experimental data and a high-resolution protein structure.[11] Poisson-Boltzmann equation based methods, such as multi-conformation continuum electrostatics (MCCE)[12,13] and macroscopic electrostatics with atomic detail (MEAD),[14,15] calculate the macroscopic electrostatic effects of ion–ion and ion–dipolar interactions, such as between a titrating site and polar solvent molecules given the dielectric response of the protein interior. Provided with a high-resolution crystal structure, they offer predictions within 1 pK_a unit RMSD for residues with relatively high solvent exposure. Since the accuracy of this method is directly related to solvent interactions, it often leads to inaccurate predictions when the target titrating residue has little macroscopic solvent interaction, or if the target site's pK_a is significantly altered by conformation.[16] To explore poorly understood protein systems, relatively more brute-force methods using MD with simulated titration may be necessary.

MD simulations can derive information from virtually any protein system as long as atomic interactions can be parameterized into a consistent force field and explicit coordinates can be defined.[7,8] This provides the potential for MD based methods to estimate residue pK_a values of lower resolution or even partially solved structures. Calculating pK_a then relies upon parameterizing the solvent model. The effective Born radii of individual residues may be calculated from the shape of the protein's solvent exposure, and from that information ionization energies may be calculated. In comparative tests, MD based methods consistently provide more accurate pK_a estimates over a wider variety of protein residues and environments than other computational methods.[7,8]

There are two dominant approaches available for the inclusion of titrating sites in MD-based pK_a calculation methods: discrete and continuous. Discrete methods titrate residues using Monte Carlo (MC) sampling, which allow protons to be added and deleted from amino acids.[17]However, recurring instantaneous switches of protonation states by adding or deleting the protons result in discontinuities of energy and force calculations. In addition, only one proton addition or deletion move is made during a MC step, which contributes to slower convergence for systems with many ionizable groups.[17]Nevertheless, discrete protonation state methods coupled with MD have proven to be useful in exploring pK_a values of proteins.[18]

Continuous methods by definition allow a gradual change in the "titration" coordinates during the MD simulation. This permits continuous energy and force calculations, yields greater sampling rates, and enables the titration of multiple sites simultaneously. The accuracy and efficiency of continuous dynamical methods make them as a useful methodology for studying many proteins.[7,8,19]

In this article, we utilize a recently developed continuous method called constant pH

molecular dynamics (CPHMD).[6,20] It is a component of the CHARMM simulation and modeling package[21] and employs a variant of the λ dynamics methodology in CHARMM[22,23] and the generalized Born with simple switching (GBSW) implicit solvent model to mimic the effects of the solvent environment[24,25] with continuous atomic trajectories.[26] The dynamics of the titration coordinates for ionizable residues is characterized by as many as two continuous coordinates for each ionizable amino acid in the form (λ, *x*). The variable λ corresponds to the protonation state of the residue and *x* controls of the interconversion between tautomeric states.[6]For single site titrations, such as the atom NZ in lysine, *x* is unnecessary since there are no tautomerers. Residues with multiple protonation sites such as ASP are defined with three states, (λ= 1) for the deprotonated, (λ = 0, *x* = 1) for the OD1-protonated state, and (λ = 0, *x* = 0) for the OD2-protonated state. By simulating protonation in this manner, pK_a predictions are made with both rapid convergence and accurate predictions to within 1.0 pK units.[6,7,20]

CPHMD has been successfully employed in the prediction of the pK_a values of amino acids both in small peptides and in proteins. Recently Khandogin and coworkers demonstrated CPHMD's accuracy on turkey ovomucoid third domain and bovine pancreatic ribonuclease A, by predicting experimental pK_a values within 0.6 to 1.0 pK units, respectively.[6] Although their simulations verified CPHMD's ability to provide accurate pK_a estimates of ionizable side chains, almost all protein residues included in this study had relatively small pK_a perturbations of several pK units or less. Considering the earlier example of ASP-96 in bacteriorhodopsin, a perturbation of several units represents a narrow range of possible pK_a values for protein residues. In pursuit of computational methods to address these highly perturbed electrostatic environments, the methods must be able to calculate the pK_a of titrating amino acids regardless of the size of the perturbation. Therefore, it is

Residue	Experimental pKa[27]	Predicted pKa	Unsigned Error (CPHMD)	Unsigned Error (Null Model)
Asp-19	2.21	3.76	1.55	1.65
Asp-21	6.54	5.43	1.11	2.68
Asp-40	3.87	2.03	1.84	0.01
Asp-77	<2.2	0.79	1.41	1.66
Asp-83	<2.2	3.83	1.63	1.66
Asp-95	2.16	3.44	1.28	1.70
Glu-10	2.82	3.32	0.50	1.25
Glu-43	4.32	3.76	0.56	0.25
Glu-52	3.93	4.90	0.97	0.14
Glu-57	3.49	4.42	0.93	0.58
Glu-67	3.76	3.62	0.14	0.31
Glu-73	3.31	2.41	0.90	0.76
Glu-75	3.26	4.89	1.63	0.81
Glu-101	3.81	3.51	0.30	0.26
Glu-122	3.89	4.69	0.80	0.18
Glu-129	3.75	4.43	0.68	0.32
Glu-135	3.76	4.44	0.68	0.31

Table 5.1 Observed versus calculated pK_a values in Δ+PHS. pK_a values for residues beyond 141 were not reported here, because their coordinates are not solved in most of the crystal structures used during this study. This includes the 3BDC structure used to calculate the data for this table.

necessary to test CPHMD in predicting highly perturbed pK_a values for biologically relevant systems. Staphylococcal nuclease (SNase) represents an ideal example of such a system, because it has both decades of folding and structural research and a variety of hyperstable mutants, including many with highly perturbed pK_a values.[28-31]

SNase is a relatively small protein consisting of a single polypeptide chain of 149 amino acids with no disulfide bonds. Its simple structure, prevalence in nature, and lack of chaperon-assisted folding to achieve its native fold have made it a model system for studying protein folding, point mutations, and the role of amino acids in protein function. Using site-directed mutagenesis, the various roles of residues in SNase's stability and folding pathway have been discovered, leading to a thorough understanding of the protein.[27,31-33] Putting theory into practice, this information was used to develop a hyperstable variant of SNase, known as Δ + PHS. This variant has five point mutations (G50F, V51N, P117G, H124L, and S128A) and a truncation (residues 44–49).[31] It is extraordinary in its ability to remain in its native conformation both over a broad range of pH and temperature, and when subjected to additional point mutations.[28,31] This resilience enables all its ionizing residues to be titrated experimentally, even with the introduction of hydrophilic residues into the protein's hydrophobic core.[28,31]

In previous work by Garcia-Moreno *et al.*, the conformational role of aspartic and glutamic acids (GLU) in Δ+ PHS were studied in detail.[31] All such residues were titrated for pK_a calculations by measuring the pH dependence of the chemical shifts of Cγ or Cδ with two-dimensional HBHC(CBCG)CO experiments.[31] These results are summarized in Table 5.1 under "experimental pK_a." In addition, 27 point-mutation variants of Δ + PHS (two ASPs and 25 GLUs) were successfully created. Each variant was titrated to measure the pK_a at the mutation site by analyzing the pH correlation with changes in Gibbs free energy of

PDB	Mutation	Experimental pK_a[28]	Predicted pK_a	Unsigned Error (CPHMD)	Unsigned Error (Null Model)	RMSD (Å)
3H6M	V104E	9.4	7.58	1.8	5.33	1.3491
1TR5	I92E	9.0	6.78	2.2	4.93	1.3903
1TQO	I92E	9.0	7.27	1.7	4.93	1.4413
3EVQ	L25E	7.5	8.36	0.9	3.43	1.2621
3ERO	I72E	7.3	6.78	0.5	3.23	1.1948
3D4D	Y91E	7.1	5.49	1.6	3.03	1.3142

Table 5.2 Observed versus calculated pK_a values for buried charge mutants of Δ+PHS with crystallographically determined structures. RMSD are in Angstroms.

unfolding ($\Delta\Delta G°_{H2O}$) with GdnHCl as a denaturant. These results are given in Tables 5.2 and 5.3 under "experimental pK_a."[28] These experiments provide a comprehensive quantification of the changes of internal energy within Δ + PHS in relation to introducing a hydrophilic residue into the hydrophobic core of the protein. The shielding effect of the surrounding hydrophobic amino acids greatly reduces solvent interactions with the glutamic and ASP mutations, and consequently increases their pK_a values by as much as 5 pK units. The measured perturbation in pK_a values for these systems provides an experimental basis for testing and comparing the accuracy of CPHMD simulations in the calculation of highly perturbed pK_a values of these acidic side chains.

The calculations we present below provide a significant test of the robustness of CPHMD predictions of pK_a. We consider four sets of calculations for GLU and ASP

Mutation	Experimental pK_a[28]	Predicted pK_a	Unsigned Error (CPHMD)	Unsigned Error (Null Model)	RMSD (Å)
L125E	9.1	6.83	2.3	5.03	1.2716
L103E	8.9	7.35	1.6	4.83	1.3857
L36E	8.7	7.10	1.6	4.63	1.2542
V66E	8.5	6.39	2.1	4.43	1.2862
V99E	8.4	7.19	1.2	4.33	1.2762
V39E	8.2	4.55	3.7	4.13	1.4638
A109E	7.9	4.41	3.5	3.83	1.3753
V74E	7.8	8.40	0.6	3.73	1.2469
A58E	7.7	5.20	2.5	3.63	1.3959
T62E	7.7	6.93	0.8	3.63	1.2945
N100E	7.6	5.76	1.8	3.53	1.3650
F34E	7.3	7.26	0.0	3.23	1.1966
V23E	7.1	6.95	0.1	3.03	1.2704
A132E	7.0	6.50	0.5	2.93	1.3416
L38E	6.8	6.33	0.5	2.73	1.1974
T41E	6.8	6.52	0.3	2.73	1.3341
A90E	6.4	6.74	0.3	2.33	1.4063
L37E	5.2	6.15	1.0	1.13	1.1912
G20E	4.5	5.46	1.0	0.43	1.3533
N118E	4.5	2.50	2.0	0.43	1.2462

Table 5.3 Observed versus calculated pK_a values for buried charge mutants of Δ+PHS with crystallographically determined structures. RMSD are in Angstroms.

residues in Δ + PHS: (1) predicting the pK_a values for each GLU and ASP in the Δ + PHS structure, (2) the value of each point mutation for proteins with solved crystal structures, (3) those of each point mutation without crystallographically determined structures, and (4) calculating the pK_a values of specific residues in systems similar to Δ + PHS. The first set of calculations confirms that our computational methods can accurately predict the pK_a values for this protein. The second and third studies explore the accuracy of pK_a calculations for proteins of less understood systems. The last set of calculations investigates the use of similar crystal structures to study a target system. Mutants without solved structures were built in CHARMM by mutating the Δ + PHS structure. The computational results are compared with NMR titrations to establish the overall quality and capability of CPHMD pK_a predictions over a range of perturbed pK_a systems. It should be noted that this protocol was not a blind study. The calculations within this article were carried out over the course of 2 years, which both preceded and followed the release of the measured pK_a values of SNase and Δ + PHS. This study represents an ongoing effort to assess the accuracy of the replica exchange (REX)-CPHMD process during its development.

5.2 The REX-CPHMD Method

REX, or parallel tempering, is a method of increasing barrier crossing rates by simulating an ensemble of proteins distributed through temperature space.[34] During a REX simulation a single protein structure is replicated and simulated in parallel over an exponentially spaced temperature range. After a defined time (replica cycle), the replicas are allowed to exchange atomic configurations with adjacent temperature windows based on the Metropolis criterion.[34] This technique has shown success in modeling protein folding and

peptide dynamics[34] and has been incorporated into numerous simulation environments.[7,35,36] As it concerns this study, it was used to enhance sampling of the protein conformational space around the vicinity of the native fold as well as the conformations of the tautomeric states of the titrating amino acids during CPHMD.

CPHMD is a methodology developed by Brooks and coworkers that assigns titration coordinates to ionizable hydrogen atoms, (λ, x), which are propagated simultaneously with atomic coordinates.[6,20] These coordinates control a smooth turning on or off of van der Waals and electrostatic interactions of hydrogen atoms in these groups, which enables a direct coupling between conformation and protonation states.[20]

In the REX-CPHMD protocol, λ and x coordinates are recorded at the end of each replica cycle for all titrating residues as defined in Equation 1.

$$N^{unprot} = \sum N \ (\lambda > 0.9;\ x < 0.1 \vee x > 0.9)$$
$$N^{prot} = \sum N \ (\lambda < 0.1;\ x < 0.1 \vee x > 0.9) \qquad \text{eq. 1}$$

As such, x defines the dominant tautomer during the cycle ($x < 0.1$; $x > 0.9$) and λ indicates whether that tautomer is protonated ($\lambda < 0.1$) or deprotonated ($\lambda > 0.9$). The non-physical regions of λ and x space that are not representative of protonated or deprotonated configurations enable a continuous transition between protonation states. Barriers are added to the energy functions for these coordinates to diminish the time spent in such states.[6,20] After completing all REX cycles, analysis was performed using the CPHMD tools within the MMTSB Tool Set (rexanalysis.pl) to collect all titration coordinates into the values N^{prot} and N^{unprot}.[35] With enough REXs, the population of states converges to the probability of state (S) as defined in Equation 2.

$$S^{unprot} = \frac{\rho^{unprot}}{\rho^{unprot} + \rho^{prot}} \approx \frac{N^{unprot}}{N^{unprot} + N^{prot}} \qquad \text{eq. 2}$$

S^{unprot} is the probability of a residue being unprotonated. ρ^{unprot} and ρ^{prot} are the probabilities associated with the unprotonated and protonated states. S^{unprot} is related to pK_a in the Henderson-Hasselbalch (HH) equation given in Equation 3.

$$S^{unprot} = \frac{1}{1 + 10^{n(pK_a - pH)}} \qquad \text{eq. 3}$$

In this equation, the Hill coefficient (n) and the pK_a can be fit given a set of S and pH values. In this study, 10 to 15 (pH, S) points per titrating residue were found to give the optimal trade-off between accuracy and computational time. For residues titrating multiple protonation sites, such as aspartic and glutamic acids, pK_a values for each site are calculated separately. These pK_a values are combined into a total pK_a via Equation 4.

$$pK_a = \log_{10}(10^{pK1} + 10^{pK2}) \qquad \text{eq. 4}$$

Here we arrive at an experimentally observable quantity. Now we review the system design and setup for comparing to other work.

5.3 Simulation Setup

Simulations

All REX-CPHMD simulations were run using the aarex.pl tool as part of the MMTSB Tool Set,[35] which performs REX simulations using the PHMD[6,20] and GBSW[24] modules within the CHARMM program environment.[21] Simulations were performed using the CHARMM22 all-atom force field for proteins[37] with CMAP[38,39] and optimized GB input radii.[38] This protocol was intended to follow closely to that performed by Khandogin and Brooks, and thus unprotonated fractions (S) of residues were calculated for pH values between pH = 2 and pH = 9 in all cases.[7] For residues with highly perturbed pK_a values, this range was extended by several pH units.

During each simulation, the protein was replicated in 8–16 temperature windows spanning from 298 K to 400 K. This range of temperatures was chosen so that the exchange ratio was approximately 35–45%.[7] All replicas were run simultaneously through exchange cycles: each cycle consisted of 500 dynamic steps (a total of 1 ps) followed by an exchange attempt. During an exchange attempt, adjacent temperature windows were allowed to exchange replica structures based on the Metropolis criterion.[33] The total sampling time of each protein was 4 ns. Debye-Hückel screening[24] of charge–charge interaction was used to represent the 150 mM salt concentration in the solvent.[7] All simulations were included a Nosé-Hoover thermostat to maintain the desired temperature for each window.[26] For the GB calculations, a smoothing length of 0.6Å at the dielectric boundary with 24 radial integration points up to 20Å and 38 angular integration points were used. The nonpolar solvation energy was computed using the surface tension coefficient of 0.03 kcal mol^{-1} Å^{-2}.[40] The SHAKE algorithm allows a 2 fs time step when applied to hydrogen bonds, and a 22Å

distance cutoff was applied to truncate the non-bonded in non-bonded energy evaluations.

Structures of Δ + PHS were processed according to their availability, which led to a division into two groups for this study: those with solved protein structures, and those without. All solved structures, including Δ + PHS and many of its mutants, are listed in 5.2 as their corresponding PDB codes. These structures were downloaded from the Protein Databank www.pdb.org.[41] For those without solved structures, the Δ + PHS structure was computationally mutated as explained in the following section.

Each PDB file was processed to remove all non amino acid residues and to convert the PDB file into a CHARMM supported format with convpdb.pl from the MMTSB toolset.[35] During this step, the ligand thymidine-3′,5′-diphosphate was removed to make the crystal structures match those used during the NMR analyses performed by Isom *et al.*[28] Structures were minimized for 500 steps with steepest descents and harmonic restraints (10 × mass) on heavy atoms. All titrating residues were patched appropriately so that CPHMD could recognize them correctly. The GLU and ASP patches represent doubly protonated residues with the hydrogen atoms bound to the ionizing oxygen.

Modeling Salt Effects

As has been shown in earlier calculations, the accurate recapitulation of experimentally measured pK_a values depends on modeling both the aspects of the solvent environment and the influence of ionic strength correctly.[7] To model solvent in our REX-CPHMD calculations, we use the optimized GBSW model[38] together with the simple Debye-Hückel correction introduced into GB models by Case *et al.*[42,43]

Simulating Residue Point Mutations

For Δ + PHS mutants without a PDB structure, coordinates were generated computationally from the Δ + PHS PDB structure (3BDC) using mutate.pl from the MMTSB toolset.[35] This protocol eliminates an amino acid at a user-specified location, and replaces it with the desired mutation. The structures were minimized using steepest descents for 500 steps with harmonic restraints (10 × mass) on all heavy atoms using minCHARMM.pl. Several mutants had significant atom clashes after running mutate.pl. These structures underwent 100 steps of steepest descents all-atom minimization using minCHARMM.pl to resolve the structural conflicts, followed the 500 step energy minimization with harmonic restraints on heavy atoms.

As a measure of confidence in the method, the average structure was calculated from each simulation trajectory and then compared with its original PDB of Δ + PHS by a backbone-based RMSD analysis of structural alignment. These values are given in Tables 5.2 and 5.3. The low values suggest that the mutations are accommodated without requiring significant reorganization of the protein.

5.4 Modeling Staphylococcal Nuclease's Ionizable Residues

Δ + PHS

The pK_a values of all 17 carboxylic acids in Δ + PHS were determined from 3BDC, as shown in Table 5.1. There is a reasonable agreement between the observed and calculated pK_a values, with an average unsigned error (AUE) of 0.99 pK units. Fifty-nine percent (59%) of the residues had an error of <1 pK unit. This suggests that our protocol is able to determine pK_a values of diprotic residues for this protein, even if they are in a greatly

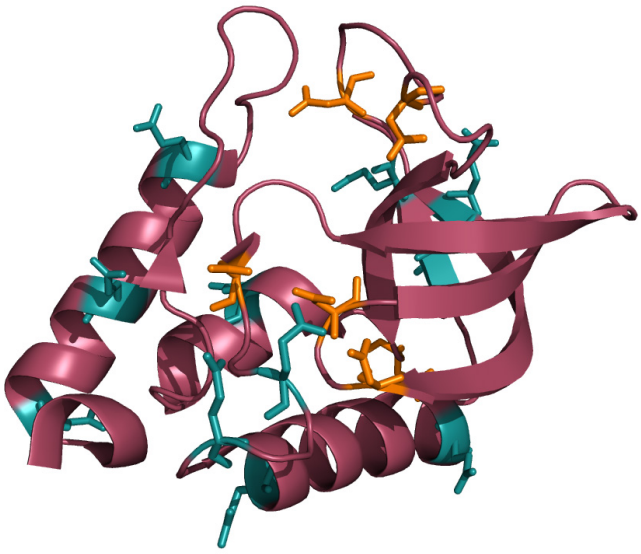

Figure 5.1 Locations of ionizable residues in Δ + PHS as displayed on the PDB crystal structure 3BDC. The Δ + PHS variant of staphylococcal nuclease is shown here with all ionizing residues highlighted. Glutamic acid is cyan, and aspartic acid is orange.

perturbed state. These findings are consistent with previous studies using CPHMD in that an AUE of 1 pK unit or less was achieved for proteins containing ionizable side chains in the core.[7]

Figure 5.1 shows that the titrated residues in our calculations sample a variety of solvent-exposed environments. GLU residues at α-helical locations (57, 67, 101, 122, 129, 135) showed an average error of 0.5 pK_a units, those in β-sheets (10, 73, 75) showed an error of 1.0 unit, and those in flexible side-chains (43, 52) showed an error of 0.8 units. ASP residues (19, 21, 40, 77, 83, 95) were all on flexible side-chains, and showed an AUE of 1.5 units.

Of the titrating residues, seven had errors in calculated pK_a values that were >1 pK unit from experimental values, six of these were ASP. Surprisingly, four of these six residues (Asp 19, Asp 21, Asp 40, and Asp 95) were in unstructured regions relatively far from the

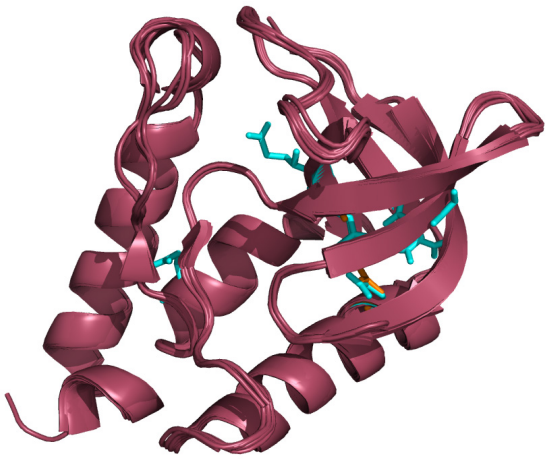

Figure 5.2 Apparent tertiary structure similarity between various solved crystal structures used in this study. Δ + PHS staphylococcal nuclease, its 6 solved PDB structures, and three structural homologues are all shown overlaid with one another. The mutated residues are shown in red. All mutants had an RMSD of <0.35Å, indicating that even with the introduction of hydrophilic residues into the protein's interior, the structure of Δ + PHS is not significantly distorted.

center of the protein. Previous research suggested that when a titrating residue has a large surface area exposed to solvent its ionization state is well defined by the GBSW model, resulting in a better pK_a prediction.[7] This phenomenon will be explored further in the discussion section.

Δ + PHS mutants with known structure

Of the 27 Δ + PHS mutants studied in this work, five (5) had solved coordinates. Figure 5.2 illustrates their similarity by overlapping their secondary structural representations. The RMSD between any two proteins was less than 0.35Å. Their six (6) corresponding PDB codes and calculated pK_a values are shown in Table 5.2. We list only the pK_a values of residues reported by the NMR titration experiments. There is good agreement between the observed and calculated ionization equilibria, with an AUE of 1.5 pK units.

The stability of proteins was monitored during the simulation by the RMSD between the initial and average structures of each simulation. The RMSD of all simulations averaged to 1.3Å (specific values are shown in Table 5.3). This indicates that the conformational changes and fluctuations that occurred during the simulations are relatively small, even when the proteins were subjugated to a wide range of pH conditions. This also indicates that such fluctuations are greater than the structural differences between different mutants.

Δ + PHS compared with I92E mutants

The mutant GLU pK_a values for two I92E structures were predicted (1TR5 and 1TQO), which provides some insight into the sensitivity of CPHMD to conformational differences in the starting structures of the proteins. The two structures had an RMSD of 0.85Å from each other, and an RMSD of 1.10Å when compared with Δ + PHS. This suggests that in the case of Δ + PHS, conformational rearrangements near the point of mutation are comparable with differences in multiple ground state configurations. These rearrangements can be explained as the energy cost of allowing Glu 92 access to solvent.

When comparing the ionization of all titrating residues between Δ + PHS and its I92E mutants, most aspartic and GLU residues titrated to values <1 pK unit from each other, as seen in Table 5.4. This falls within 1 pK_a unit of error, as seen in previous research.[7] Residues outside of this margin include all residues on flexible regions of the protein, such as all ASP residues. These residues sample a wide range of fluctuations in the environment, which may require a longer time to converge to a correct pK_a estimate. There was a consistent trend that corresponding residues yielded similar pK_a predictions, which suggests that the conformational changes induced by point mutations do not destroy the overall

Residue	Experimental pKa[27]	Predicted pKa (3BDC)	Predicted pKa (1TR5)	Predicted pKa (1TQO)	Averaged Deviation from Experimental
Asp-19	2.21	3.76	1.55	3.58	0.445
Asp-21	6.54	5.43	5.63	5.64	1.01
Asp-40	3.87	2.03	2.54	2.48	1.585
Asp-77	<2.2	0.79	1.25	0.52	1.18
Asp-83	<2.2	3.83	4.3	3.63	1.865
Asp-95	2.16	3.44	3.42	3.83	1.27
Asp-143	3.86	--	--	3.74	--
Glu-10	2.82	3.32	4.55	3.88	1.115
Glu-43	4.32	3.76	3.4	3.60	0.74
Glu-52	3.93	4.9	4.81	5.09	0.925
Glu-57	3.49	4.42	4.5	4.59	0.97
Glu-67	3.76	3.62	3.87	3.76	0.015
Glu-73	3.31	2.41	3.21	3.20	0.5
Glu-75	3.26	4.89	4.69	4.53	1.53
Glu-92	--	--	6.78	7.27	--
Glu-101	3.81	3.51	3.38	3.29	0.365
Glu-122	3.89	4.69	5.16	4.94	1.035
Glu-129	3.75	4.43	4.14	4.24	0.535
Glu-135	3.76	4.44	4.54	4.56	0.73
Glu-142	4.49	--	--	4.41	--

Table 5.4 Comparison of Δ + PHS pK_a values (all titrating residues) to its I92E mutant residues

accuracy of the calculation for other ionizing residues. This opens the possibility that when predicting pK_a values, a solved structure may not be necessary; if an approximation of the secondary and tertiary structures can be found, pK_a values might still be predicted using REX-CPHMD. The remaining calculations in this study are designed to explore this possibility.

Δ + PHS mutants with modeled structure

Eighteen (18) of the reported pK_a values from previous analyses did not have a corresponding solved structure in the PDB. Assuming that the solved structure of Δ + PHS is an adequate approximation of the system, models for these proteins were created by computationally mutating the Δ + PHS PDB file 3BDC. For these mutants, the results from our pK_a calculations appear in Table 5.3. Changes in the amino acid sequence of Δ + PHS, and our modeling of them, could affect the quality of the calculated pK_a values. However, these changes are apparently small enough to allow accurate predictions of the pK_a values for the mutated proteins to within an AUE of 1.4 pK_a units. This indicates that even in the absence of a crystallographically determined starting structure, the CPHMD methodology can yield accurate predictions of pK_a shifts with an AUE similar to those calculated from solved crystal structures. A caveat here, is that this technique requires a near-match of crystal structure to model the chemistry of the target system.

Calculation of a single residue

During this study, all residues were titrated simultaneously for every structure. This ensured that all cooperative protonation interactions between nearby titrating residues were considered. When the pK_a of only a single titrating residue is desired, however, it may be

PDB	Mutation	Experimental pKa	Predicted pKa	Unsigned Error (CPHMD)	Unsigned Error (Null Model)	RMSD (Å)
1U9R	V66E[28]	8.9	8.15	0.8	4.83	1.1162
2OXP	V66D[44]	8.8	7.50	1.3	4.73	1.0719
2OEO	I92D[45]	7.5	7.47	0.0	3.43	1.4566

Table 5.5 Calculated and experimental pK_a values of Δ + PHS mutants modeled from nonexact matches of amino acid sequences.

more efficient to titrate only the target residue. This was tested by calculating the pK_a value of the GLU residue of the I92E (1TR5) mutant by allowing only the mutant residue to titrate. The calculation produced a value of 6.4 pK units, compared with 6.8 pK units when all ionizable residues were allowed to titrate. Since titrating residues don't significantly alter the ionization equilibria of distant parts of the system, these results suggests that the differences in accuracy by simulating the titration of one residue may be small enough to allow accurate pK_a prediction. The caveat for performing only a single-site titration during a REX-CPHMD simulation is that it ignores any cooperative protonation chemistry and the subsequent dynamics influenced by it. This simplification can greatly reduce the computational cost of modeling pK_a changes in large systems with many titrating residues by reducing time to reach convergence.

Calculation from similar PDB structures

In many cases, atomic coordinates are not available for a particular protein from crystallographic or NMR studies. This portion of the study investigates the accuracy of pK_a predictions when using a PDB with a similar tertiary structure to the target one to determine pK_a values. Three mutants of Δ + PHS were matched with three PDB files that had nearly identical conformations to Δ + PHS: 1U9R, 2OXP, and 2OEO. These pairings, including their experimental pK_a values, appear in 5.5. To illustrate their similarity with Δ + PHS, all of these structures appear in Figure 5.2 overlaid with the other structures homologous to Δ + PHS.

The results from the pK_a calculations were surprisingly accurate, especially considering that 2OEO (similar to Δ + PHS I92E), provided the most accurate result despite lacking five ionizable lysines from the Δ + PHS/I92D structure used in the experimental calculations. Since these residues only titrate at dissimilar pH values than GLU, it is unlikely these changes to the sequence had substantial effects on the target ASP-92 mutation. These results suggest that REX-CPHMD can provide accurate pK_a calculations from a similar structure even in the absence of an exact match of amino acid sequences. These also suggest that approximating the tertiary conformation of a protein may be sufficient to predict its pK_a values accurately.

V39E and A109E mutants

The two simulations that yielded the poorest outcome for calculated pK_a values, V39E and A109E, were examined for structural exceptions that may have caused their unusually high deviation. In both cases, the mutant residue was on an unstructured region of

the protein, and both residues flipped their orientations outward in the averaged structures from their respective simulations. The conformational change then exposed the GLU residues to more solvent than had they remained in the interior of the protein, thereby lowering their calculated pK_a values. This change is evident in both structures' having relatively large RMSD values between the average structure and the initial structure. This conformational change may be due to the understabilization of local salt bridges that would otherwise pull the residues into the interior of the protein or have arisen from model preparation and equilibration protocols. The averaged structure of the V39E mutant appears to have a stable GLU39–LYS110 salt bridge that exposes the V39E mutation to more solvent (leading to a reduced pK_a). During the calculation, however, the GLU39–ARG35 salt bridge may be the dominant orientation of the mutant site, which would draw the GLU into the interior of the protein (leading to an elevated pK_a). The A109E mutant showed an average structure with a solvent-exposed LYS108–GLU109 salt bridge. This bridge may have been overstabilized relative to the ARG105–GLU109 salt bridge that would draw the mutant residue into the core of the protein. These residues could be exceptions to the current update of the GBSW force field.[38]

Comparison with similar work

During the course of this study, a publication with many similar results to this article was published by Wallace *et al.*[46] Although they calculated pK_a values both in CHARMM and using an identical GBSW force field, their calculations yielded a somewhat lower AUE of 1.1 pK units. This difference appears to have arisen from the linear fitting of the HH equation to single pH points. This technique involves calculating and averaging pK_a values from several (or one as in their case) points where S^{unprot} is nearly 0.5, and assuming that the Hill

coefficient (n) is equal to unity. To test this, a single S^{unprot} fraction from this study was used

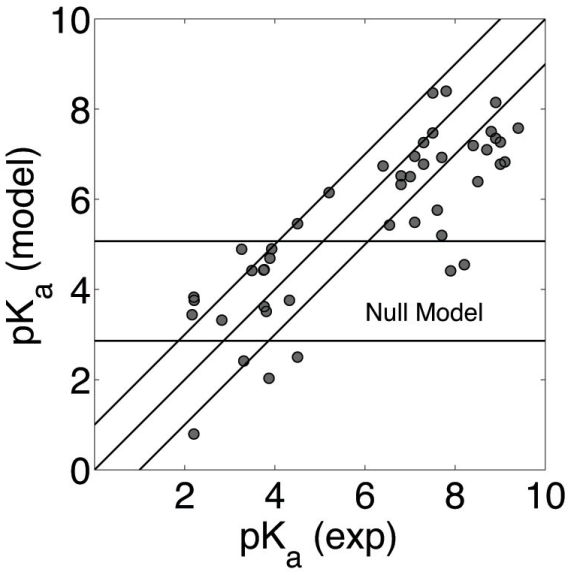

Figure 5.3 Calculated versus experimental pK_a. All pK_a values that had a corresponding experimental pK_a value are presented in this graph. This includes all values from Tables 1-3, and 5. A perfect prediction would presumably place all points along a 45° incline from the origin. The ideal range of ±1 pK unit error from this diagonal has been highlighted. The null model region is the horizontal range of ±1 pK unit error from unperturbed ASP and GLU pK_a values of 3.86 and 4.07, respectively. As shown, CPHMD excels in discovering and mapping large perturbations in pK_a.

to calculate each pK_a value available. The results gave an AUE identical to that from the Wallace *et al.* article (1.1 pK units), and an average unsigned difference from the HH fit of less than 0.3 pK units per residue. When the pH values were chosen closest to this study's calculated pK_a values, the calculations yielded an identical AUE as the HH-equation curve fitting method (1.3 pK units), and an average unsigned difference from the HH fit of less than 0.3 pK units per residue. This indicates that more accurate pK_a values may be calculated with fewer points than fitting a complete HH equation curve, when the appropriate single pH value has been determined. The caveats of this method are that it may require manually choosing the data points used to solve the linear fit, and it is clearly not applicable when multiple sites are of interest.

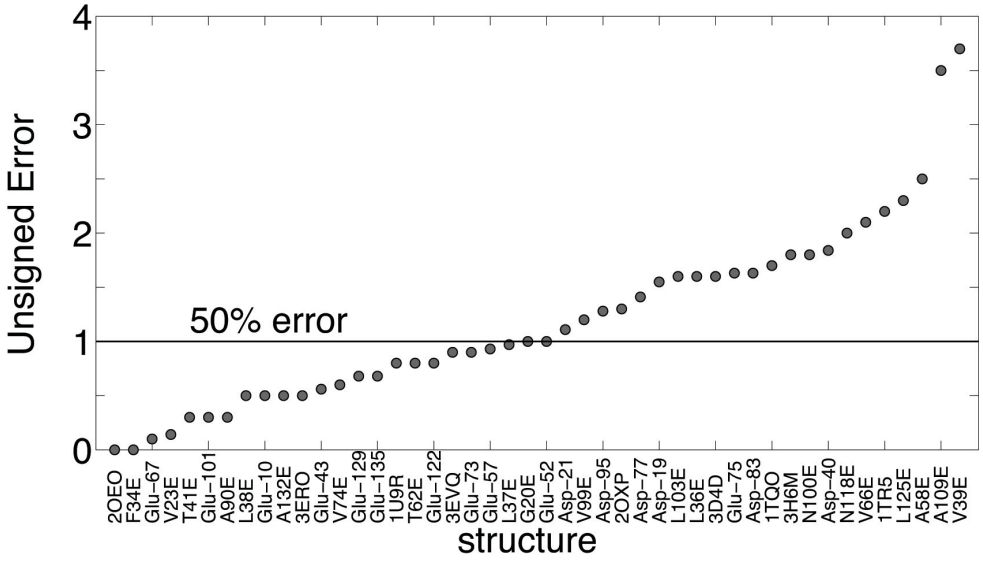

Figure 5.4 pK_a values of GLU and ASP residues in 29 internal positions in staphylococcal nuclease. This is a list of mutations in order of increasing unsigned difference of experimental determination of apparent pK_a value, and its calculated value using CPHMD. Approximately half (48%) of the calculated values had a difference of <1 pK unit.

5.5 Discussion

Making use of the REX enhanced sampling protocol and the improved parameterization of the GBSW implicit solvent model, we determined the pK_a shifts of a large number of SNase buried charge mutants. Our study provides accurate calculations of the ionization properties of buried charge groups in proteins, and supports our REX-CPHMD method as a useful tool for studying pK_a shifts.[47] In addition, the titrating groups in the mutants of this study have among the most-perturbed carboxylic acid pK_a values observed.[28] Being able to predict such titration shifts accurately suggests that CPHMD simulations and the GBSW implicit water model provide a robust methodology for exploring electrostatic environments of protein interiors.

When taking the perspective of a null model, where all GLU and ASP are assumed to have fixed pK_a values of 4.07 and 3.86, respectively,[48] the AUE of predicting pK_a values is

similar to CPHMD when observing amino acids with a small perturbation. Results in Table 5.1 show that the null model had an AUE of 0.85 pK units, while CPHMD had an AUE of 0.99 units. The null model fails when large perturbations are being observed. The low-pK_a bias for ASP residues in Δ + PHS, for instance, was consistently modeled better with CPHMD by several tenths of a pK unit. As Figure 5.3 illustrates, when the perturbation of the amino acid is more than one unit, CPHMD calculations are significantly better. When considering all pK_a predictions within this experiment, the analogous result from the null model prediction has a mean AUE of 3.54 pK units, as compared with the AUE of 1.31 units with CPHMD. A relative confidence level of CPHMD is shown in Figure 5.4 by listing the complete comparative statistics of this study. All calculated residues that had corresponding experimental data are listed by order of increasing error. 48% had an error below 1 pK unit. This margin contains 58% of Δ + PHS residues, 44% predictions from PDB files, and 50% of predictions from modeled structures.

We note that although pK_a is defined by protein structure, no strong correlations were found between the error of the pK_a prediction and large-scale structural phenomena within the scope of this study. These include conformational changes caused by the relaxation of the protein, changes in residue volume from the mutation, and proximity to the bound ligand thymidine-3′,5′-diphosphate present in the PDB structure . The R^2 values of these trends were 0.29, 0.002, and 0.001, respectively. This indicates that the methodology may not be significantly improved by accommodating such conformational trends or exceptions. This provides insight into the robustness of CPHMD: our method repeatedly yields accurate predictions of pK_a values almost irrespective to such phenomena.

The one trend consistent enough throughout this study was the under-prediction of pK_a values, as seen in Figure 5.3. When calculating residue pK_a values of Δ +PHS mutants,

23 of 29 values were underpredicted. This suggests that CPHMD may systematically overstabilize the ionized form of the residues studied, and indicates avenues of refinement in the updated GBSW-specific force field created in previous work.[28] To refine the protocol significantly, adjustments may need to be made to the force field and titrating residue patches to increase the perceived perturbation of residue pK_a values.

While refinements should be made to improve the accuracy of the CPHMD protocol, this study provides a modest benchmark of its capability to predict highly perturbed pK_a values of buried charge residues in proteins. This promises to aid the evaluation and characterization of ionization in protein interiors, which could give valuable insight into the mechanism of pH-based biological activity.

5.6 References

1. V. K. Rastogi, and M. E. Girvin, "Structural Changes Linked to Proton Translocation by Subunit C of the Atp Synthase," Nature **402**(6759), 263-68, (1999).

2. V. Ovchinnikov, B. L. Trout, and M. Karplus, "Mechanical Coupling in Myosin V: A Simulation Study," J. Mol. Biol. **395**(4), 815-33, (2010).

3. T. K. Harris, and G. J. Turner, "Structural Basis of Perturbed Pk(a) Values of Catalytic Groups in Enzyme Active Sites," IUBMB Life **53**(2), 85-98, (2002).

4. J. W. Kelly, "Alternative Conformations of Amyloidogenic Proteins Govern Their Behavior," Curr. Opin. Struct. Biol. **6**(1), 11-17, (1996).

5. J. Khandogin, and C. L. Brooks, III, "Linking Folding with Aggregation in Alzheimer's Beta-Amyloid Peptides," Proc. Natl. Acad. Sci. U.S.A. **104**(43), 16880-85, (2007).

6. J. Khandogin, and C. L. Brooks, III, "Constant Ph Molecular Dynamics with Proton Tautomerism," Biophys. J. **89**(1), 141-57, (2005).

7. J. Khandogin, and C. L. Brooks, III, "Toward the Accurate First-Principles Prediction of Ionization Equilibria in Proteins," Biochemistry **45**(31), 9363-73, (2006).

8. J. A. Wallace, and J. K. Shen. in *Methods in Enzymology* Vol. 466 (eds L. J. Michael, K. A. Gary, and M. H. Jo) 455-75 (Academic Press, 2009).

9. R. L. Thurlkill, G. R. Grimsley, J. M. Scholtz, and C. N. Pace, "Pk Values of the Ionizable Groups of Proteins," Protein Sci. **15**(5), 1214-18, (2006).

10. D. C. Bas, D. M. Rogers, and J. H. Jensen, "Very Fast Prediction and Rationalization of Pk(a) Values for Protein-Ligand Complexes," Proteins: Struct., Funct., Bioinf. **73**(3), 765-83, (2008).

11. H. Li, A. D. Robertson, and J. H. Jensen, "Very Fast Empirical Prediction and Rationalization of Protein Pka Values," Proteins: Struct., Funct., Bioinf. **61**(4), 704-21, (2005).

12. E. G. Alexov, and M. R. Gunner, "Incorporating Protein Conformational Flexibility into the Calculation of Ph-Dependent Protein Properties," Biophys. J. **72**(5), 2075-93, (1997).

13. R. E. Georgescu, E. G. Alexov, and M. R. Gunner, "Combining Conformational Flexibility and Continuum Electrostatics for Calculating Pk(a)S in Proteins," Biophys. J. **83**(4), 1731-48, (2002).

14. D. Bashford, and K. Gerwert, "Electrostatic Calculations of the Pka Values of Ionizable Groups in Bacteriorhodopsin," J. Mol. Biol. **224**(2), 473-86, (1992).

15. D. Bashford. in *Scientific Computing in Object-Oriented Parallel Environments* Vol. 1343 *Lecture Notes in Computer Science* (eds Y. Ishikawa, R. Oldehoeft, J. Reynders, and M. Tholburn) 233-40 (Springer Berlin / Heidelberg, 1997).

16. D. Bashford, "Macroscopic Electrostatic Models for Protonation States in Proteins," Front. Biosci. **9**, 1082-99, (2004).

17. A. M. Baptista, V. H. Teixeira, and C. M. Soares, "Constant-Ph Molecular Dynamics Using Stochastic Titration," J. Chem. Phys. **117**(9), 4184-200, (2002).

18. J. T. Mongan, D. A. Case, and J. A. McCammon, "Constant Ph Molecular Dynamics in Generalized Born Implicit Solvent," Abstr. Pap. Am. Chem. Soc. **229**(Part 1), U768, (2005).

19. S. Kannan, and M. Zacharias, "Enhanced Sampling of Peptide and Protein Conformations Using Replica Exchange Simulations with a Peptide Backbone Biasing-Potential," Proteins: Struct., Funct., Bioinf. **66**(3), 697-706, (2007).

20. M. S. Lee, F. R. Salsbury, and C. L. Brooks, III, "Constant-Ph Molecular Dynamics Using Continuous Titration Coordinates," Proteins: Struct., Funct., Bioinf. **56**(4), 738-52, (2004).

21. B. R. Brooks, C. L. Brooks, III, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus, "Charmm: The Biomolecular Simulation Program," J. Comput. Chem. **30**(10), 1545-614, (2009).

22. J. L. Knight, and C. L. Brooks, III, " Λ -Dynamics Free Energy Simulation Methods," J. Comput. Chem. **30**(11), 1692-700, (2009).

23. X. J. Kong, and C. L. Brooks, III, " Λ -Dynamics: A New Approach to Free Energy Calculations," J. Chem. Phys. **105**(6), 2414-23, (1996).

24. W. Im, M. S. Lee, and C. L. Brooks, III, "Generalized Born Model with a Simple Smoothing Function," J. Comput. Chem. **24**(14), 1691-702, (2003).

25. J. Chen, C. L. Brooks, III, and J. Khandogin, "Recent Advances in Implicit Solvent-Based Methods for Biomolecular Simulations," Curr. Opin. Struct. Biol. **18**(2), 140-48, (2008).

26. S. Nose, "A Unified Formulation of the Constant Temperature Molecular-Dynamics Methods," J. Chem. Phys. **81**(1), 511-19, (1984).

27. B. Cannon, D. Isom, A. Robinson, J. Seedorff, and B. Garcia-Moreno, "Molecular Determinants of the Pka Values of the Internal Asp Residues," Biophys. J., 403A-03A, (2007).

28. D. G. Isom, C. A. Castañeda, B. R. Cannon, P. D. Velu, and E. B. García-Moreno, "Charges in the Hydrophobic Interior of Proteins," Proc. Natl. Acad. Sci. **107**(37), 16096-100, (2010).

29. Y. Arata, R. Khalifah, and O. Jardetzky, "Nmr Relaxation Studies of Unfolding and Refolding of Staphylococcal Nuclease at Low Ph," Ann. N.Y. Acad.Sci. **222**(DEC31), 230-39, (1973).

30. A. Erickson, and R. H. Deibel, "Production and Heat-Stability of Staphylococcal Nuclease," J. Appl. Microbiol. **25**(3), 332-36, (1973).

31. C. A. Castaneda, C. A. Fitch, A. Majumdar, V. Khangulov, J. L. Schlessman, and B. E. Garcia-Moreno, "Molecular Determinants of the Pk(a) Values of Asp and Glu Residues in Staphylococcal Nuclease," Proteins: Struct., Funct., Bioinf. **77**(3), 570-88, (2009).

32. K. Baran, C. Fitch, J. Schlessman, and B. Garcia-Moreno, "Molecular Determinants of Pka Values of Ionizable Residues Involved in Clusters and Networks: Contributions by Short-Range Interactions and by Local Conformational Fluctuations," Biophys. J. **88**(1), 38A-38A, (2005).

33. D. A. Karp, A. G. Gittis, M. R. Stahley, C. A. Fitch, W. E. Stites, and B. E. Garcia-Moreno, "High Apparent Dielectric Constant inside a Protein Reflects Structural Reorganization Coupled to the Ionization of an Internal Asp," Biophys. J. **92**(6), 2041-53, (2007).

34. Y. Sugita, and Y. Okamoto, "Replica-Exchange Molecular Dynamics Method for Protein Folding," Chem. Phys. Lett. **314**(1-2), 141-51, (1999).

35. M. Feig, J. Karanicolas, and C. L. Brooks, III, "Mmtsb Tool Set: Enhanced Sampling and Multiscale Modeling Methods for Applications in Structural Biology," J. Mol. Graph. Model. **22**(5), 377-95, (2004).

36. B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, "Gromacs 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation," J. Chem. Theory Comput. **4**(3), 435-47, (2008).

37. A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus, "All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins," J. Phys. Chem. B **102**(18), 3586-616, (1998).

38. J. H. Chen, W. P. Im, and C. L. Brooks, III, "Balancing Solvation and Intramolecular Interactions: Toward a Consistent Generalized Born Force Field," J. Am. Chem. Soc. **128**(11), 3728-36, (2006).

39. M. Feig, A. D. MacKerell, and C. L. Brooks, III, "Force Field Influence on the Observation of Pi-Helical Protein Structures in Molecular Dynamics Simulations," J. Phys. Chem. B **107**(12), 2831-36, (2003).

40. D. Sitkoff, K. A. Sharp, and B. Honig, "Accurate Calculation of Hydration Free-Energies Using Macroscopic Solvent Models," J. Phys. Chem. **98**(7), 1978-88, (1994).

41. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," Nucleic Acids Res **28**(1), 235-42, (2000).

42. V. Tsui, and D. A. Case, "Theory and Applications of the Generalized Born Solvation Model in Macromolecular Simulations," Biopolymers **56**(4), 275-91, (2000).

43. D. Bashford, and D. A. Case, "Generalized Born Models of Macromolecular Solvation Effects," Annu. Rev. Phys. Chem. **51**(1), 129-52, (2000).

44. M. S. Chimenti, C. A. Castañeda, A. Majumdar, and B. GarcÌa-Moreno E, "Structural Origins of High Apparent Dielectric Constants Experienced by Ionizable Groups in the Hydrophobic Core of a Protein," J. Mol. Biol. **405**(2), 361-77, (2011).

45. B. Garcia-Moreno, J. J. Dwyer, A. G. Gittis, E. E. Lattman, D. S. Spencer, and W. E. Stites, "Experimental Measurement of the Effective Dielectric in the Hydrophobic Core of a Protein," Biophys. Chem. **64**(1-3), 211-24, (1997).

46. J. A. Wallace, Y. Wang, C. Shi, K. J. Pastoor, B.-L. Nguyen, K. Xia, and J. K. Shen, "Toward Accurate Prediction of Pka Values for Internal Protein Residues: The Importance of Conformational Relaxation and Desolvation Energy," Proteins: Struct., Funct., Bioinf. **79**(12), 3364-73, (2011).

47. C. A. Fitch, S. T. Whitten, V. J. Hilser, and E. B. Garcia-Moreno, "Molecular Mechanisms of Ph-Driven Conformational Transitions of Proteins: Insights from Continuum Electrostatics Calculations of Acid Unfolding," Proteins: Struct., Funct., Bioinf. **63**(1), 113-26, (2006).

48. T. E. Creighton, *Proteins: Structure and Molecular Properties* (W. H. Freeman and Company, New York, 1993)

Chapter 6

Implementation of the GBSW Water Model on Modern Graphics Processors

The work presented in this chapter has been published in the following paper:

1. E. J. Arthur, and C. L. Brooks, III, "Parallelization and Improvements of the Generalized Born Model with a Simple sWitching Function for Modern Graphics Processors," in progress.

6.1 Introduction

An accurate representation of solvent in molecular dynamics simulations plays a vital role in recapitulating molecular conformation and energetics. This is especially true for studying biological macromolecules such as nucleic acids and proteins, where the solvent environment can be a driving force of observed phenomena.[1-5] Traditionally in biomolecular simulations, the solvent (generally water) is represented by atomically-detailed molecules and counterions that surround a solute molecule. While such explicitly-represented solvent models are often considered the most detailed approach to molecular simulations, they can be cost-prohibitive when used for long timescales and large systems.[6] In order to reduce boundary-condition artifacts and to better describe experiments, a given system may comprise of as much as 95% water-related atoms.[7,8] The computational load of accounting

for non-bonded pairwise interactions and the need to equilibrate configurations of water and counterions can make many systems prohibitively expensive to simulate.[6]

For purposes of exploring conformational equilibria of a large solute molecule, implicit solvent models can be used to mimic solvent effects without requiring the computational load of simulating a large bulk of solvent.[9-13] Although implicit solvent omits atomic-level interactions between the solvent and solute, such as hydrogen bonding, such setups offer straightforward methods of calculating solvation free energy, salt effects, and continuous changes to pH.[11,13-17] Additionally, continuum solvent obviates the need to maintain structural equilibria of water and counterions, so conformational changes of the solute often occur on shorter timescales. For instance Tsui and Case have shown that A-form DNA converges into a more optimal B-form conformation within 20 ps as compared to 500 ps when using explicit solvent.[7,12] Such enhanced dynamics have been useful in exploring protein folding mechanisms and protein-protein interactions.[4,18]

Many successful implicit solvent models are based on the assumption that a protein's interior is a uniform, low dielectric region of space filled with partially-charged atoms, and that this protein is surrounded by a featureless high-dielectric solvent.[19,20] The exact solution of this approximation is given by the numerical solution of the finite-difference Poisson-Boltzmann (PB) equation. Although PB implicit solvation grants simulation speed gains by reducing the system size, its poor scalability has been a principle bottleneck in exploring the dynamics of large biological systems.[21-23]

In the pursuit of finding a more efficient method of solvating bio-macromolecules, the generalized Born (GB) implicit solvent model has been developed as a computationally cheaper approximation of PB solvent.[13,19,20] This method of calculating a system's electrostatic free energy relies upon the solute atom's locations, atomic partial charges, and

the effective distance between an atom and the solvent-solute dielectric boundary, or Born radius. The most accurate GB formula for calculating the electrostatic free energy of solvation (ΔG^{elec}) was first proposed by Still et al., and follows the form[20]

$$\Delta G^{elec} = -\frac{1}{2}\sum_a \sum_b \tau \frac{q_a q_b}{f_{ab}^{GB}} \qquad \text{(eq. 1)}$$

where

$$f_{ab}^{GB} = \left[r_{ab}^2 + R_a^{Born} R_b^{Born} \exp(-r_{ab}^2 / (4 R_a^{Born} R_b^{Born})) \right]^{1/2} \qquad \text{(eq. 2)}$$

Here R_a^{Born} represents the Born radius of atom a, r_{ab} is the distance between atoms a and b, and q is the partial charge of the atoms. τ is the conversion factor that scales the Born energy by the difference in dielectric values at the dielectric boundary.

$$\tau = 1/\varepsilon_p - 1/\varepsilon_s \qquad \text{(eq. 3)}$$

Here ε_p and ε_s are the dielectric values of inside the solute molecule (such as a protein) and solvent respectively. Should a low concentration of salt be present in the simulation, the electrostatic energy can be modified by a Debye-Huckel screening parameter κ as follows,[17]

$$\tau = 1/\varepsilon_p - \exp(-\kappa f_{GB})/\varepsilon_s \qquad \text{(eq. 4)}$$

The accuracy and speed of GB implicit solvent models depend heavily on the method used for calculating the Born radius, and those various methods are what distinguish each model. Some popular models include using an empirically-driven spatial symmetry function of atom placement such as in the Fast Analytical Continuum Treatment of Solvation (FACTS);[24] atom-atom pairwise potentials as in Generalized Born Surface Area from Onufriev, Bashford, and Case (GBSA/OBC);[11,25] and atomic volume exclusion such as in the Generalized Born with a Simple sWitching function (GBSW)[13] and Generalized Born using Molecular Volume (GBMV).[26] Atomic volume exclusion algorithms make few assumptions regarding the shape of molecules and the placement of atoms. As we will develop later in this study, these algorithms integrate energy contributions from groups of neighboring atoms, which is effective at capturing atomic overlap and buriedness. As such, they often excel at reproducing solvation free energies, but usually at a higher computational cost and lower scalability relative to other models.[27] In this study we will look at improving the speed and scalability of the accurate GBSW model.

GBSW has over a decade of research and parameterization. Aside from gaining a well-characterized set of atomic and fitting parameters, its functionality has been extended to include pH, implicit membranes, and coarse-graining.[4,13,14,28-32] Unfortunately GBSW scales poorly with system size, and systems larger than 1,000 atoms running on one central processing unit (CPU) core proceed at speeds of less than 1 nanosecond (ns) per day. Several methods of improving its speed include using more processing cores, improving the algorithm, or improving the hardware. With additional CPU cores modest speed improvements can be seen, and systems of up to 10,000 atoms can be simulated for single ns/day. When using additional cores, few speed increases are seen above about 20 cores. Additionally, systems with more than 8 cores today are expensive, and cost-limiting to many

research groups. Thus we focus on algorithmic improvements to allow GBSW to utilize more cores, and on hardware improvements to take advantage of newer and more affordable parallel processing hardware.

With the availability of graphics processing units (GPUs) carrying up to thousands of parallel processing cores and their newer ability to compute complex mathematical functions using C-like languages such as Open Computing Language (OpenCL) and Compute Unified Device Architecture (CUDA), a new frontier of GPU-powered ultra-parallel molecular dynamics software has come into being. Programs such as CHARMM,[6] AMBER,[33] OpenMM,[34] GROMACS,[35] and NAMD[36] all offer GPU-accelerated options for many types of simulations, all of which can replace the computational power of much larger computer networks with a single graphics card. Despite the fantastic improvements in molecular mechanics simulations afforded by GPUs, some algorithms remain challenging to parallelize. Notable among these are implicit solvent models, which either rely on recursive data processing or are inefficiently split into parallel functions. From the variety of implicit solvent methods for calculating solvation free energy, only those that use an uncoupled summation of Cramer-Truhlar-type atom-atom pairwise interactions,[37] such as GBSA/OBC,[11,25] have been implemented in GPU languages. Such implementations only required a retooled version of the neighboring atom interaction processes that were already developed for all-atom molecular mechanics.[11,34-36] This study represents the first implementation of a parallel, atom-coupled volumetric integration approach to calculating solvation free energy using the GBSW algorithm.

Due to OpenMM's achievements and effectiveness in harnessing GPUs, the CHARMM-OpenMM interface was developed to combine the capabilities of the two software packages. As such, the robust algorithms and range of methods supported in

CHARMM are used to design new simulation methods, and these methods are run using OpenMM's efficient processes that have been developed and optimized for modern GPU architectures[6,34] Additionally, the GBSA/OBC model already in place in OpenMM offers a GB framework that forms a basis for our new GBSW code. In this study we outline a highly-parallelized version of the Generalized Born implicit solvent model with a Simple sWitching function within the CHARMM-OpenMM interface.[14] First we present some of the underlying theory of how GBSW calculates the solvation free energy and Born radii. Then we delve into the implementation of the algorithm in its original Fortran90 format, and how functions were refactored for a parallel CUDA implementation in the OpenMM software package. Please refer to the original text of this chapter to explain why particular numerical cutoffs were chosen, and to describe the hardware setup used for benchmarking. Finally we review the speed improvements achieved by the new algorithm, its ability to fold chignolin a linear peptide chain, and future directions for developing the model.

6.2 Effective Born Radii

Like many other GB solvent models, GBSW uses the atomic self-contribution of the Still equation (eq. 1) to calculate the Born radius from the electrostatic free energy. The self-term for atom a reduces eq. 1 to

$$\Delta G_a^{elec} = \frac{-\tau}{2}\frac{q_a^2}{R_a^{Born}} \qquad \text{(eq. 5)}$$

The self-energy is then approximated in two energy terms, the Coulomb field approximation term $\Delta G_a^{elec,0}$, and an empirical correction term $\Delta G^{elec,1}$, in the following relationship:[26]

$$\Delta G_a^{elec} \approx \alpha_0 \Delta G_a^{elec,0} + \alpha_1 \Delta G_a^{elec,1} \qquad \text{(eq. 6)}$$

where α_0 and α_1 are empirical fitting coefficients. By default, these coefficients are -0.1801 and 1.81745 respectively.[13] The first interaction term is derived from the Coulomb-field approximation for electric displacement, and it calculates the work function for removing the partial charge of an atom a a distance from a dielectric boundary. This term is evaluated to

$$\Delta G_a^{elec,0} = \frac{-\tau q_a^2}{8\pi} \int_{solvent} \frac{1}{\left(r_{a,\mathbf{r}}\right)^4} dV \qquad \text{(eq. 7)}$$

Here the integral is evaluated over all solvent volume V, and $r_{a,\mathbf{r}}$ is the radial distance between the point in space $\mathbf{r}$ and atom a. The Coulomb-field approximation from equation 7 systematically underestimates the electrostatic solvation free energy as calculated by exact Poisson-Boltzmann methods, and consequently overestimates atomic Born radii.[26] Lee et al. demonstrated that this underestimation could be greatly reduced by adding the Born energy correction term $\Delta G^{elec,1}$.[26,38] The term is computed as follows.[13]

$$\Delta G_a^{elec,1} = \frac{-\tau q_a^2}{2} \left[\frac{1}{4\pi} \int_{solvent} \frac{1}{\left(r_{a,\mathbf{r}}\right)^7} dV \right]^{1/4} \qquad \text{(eq. 8)}$$

Finally we solve for the Born radius using eq. 5 – 8 and we arrive at

$$\left(R_a^{Born}\right)^{-1} = \alpha_0 \left[\frac{1}{4\pi} \int_{solvent} \frac{1}{\left(r_{a,\mathbf{r}}\right)^4} dV \right] + \alpha_1 \left[\frac{1}{4\pi} \int_{solvent} \frac{1}{\left(r_{a,\mathbf{r}}\right)^7} dV \right]^{1/4} \qquad \text{(eq. 9)}$$

Thus we have the basic construction for evaluating Born radii. Next we explain the details of the switching function that define the dielectric boundary in GBSW.

6.3 Switching Function

In GBSW, the solvent volume is defined as the region of space excluded by the van der Waals spheres of the solute molecule's atoms. Rather than integrate over all space to infinity, we integrate the volume of the solute molecule and calculate the volume inclusive function $V_{solute}(\mathbf{r}-\mathbf{r}_a)$ where $\mathbf{r}-\mathbf{r}_a$ is a location in Cartesian space centered on atom a. Then $1-V_{solute}(\mathbf{r}-\mathbf{r}_a)$ becomes the solvent exclusion function, and we get the following Born radius equation.

$$\left(R_a^{Born}\right)^{-1} \approx \alpha_0 \left[\frac{1}{4\pi} \int d\mathbf{r} \frac{1-V_{solute}(\mathbf{r}-\mathbf{r}_a)}{\left(r_{a,\mathbf{r}-\mathbf{r}_a}\right)^4} \right] + \alpha_1 \left[\frac{1}{4\pi} \int d\mathbf{r} \frac{1-V_{solute}(\mathbf{r}-\mathbf{r}_a)}{\left(r_{a,\mathbf{r}-\mathbf{r}_a}\right)^7} \right]^{1/4} \qquad \text{(eq. 10)}$$

Here, $r_{a,\mathbf{r}-\mathbf{r}_a}$ is the radial distance between the point in space $\mathbf{r}-\mathbf{r}_a$ and atom a. Being a function of all atoms of the solute, $V_{solute}(\mathbf{r}-\mathbf{r}_a)$ can be expressed as a product of each atom's volume as follows.

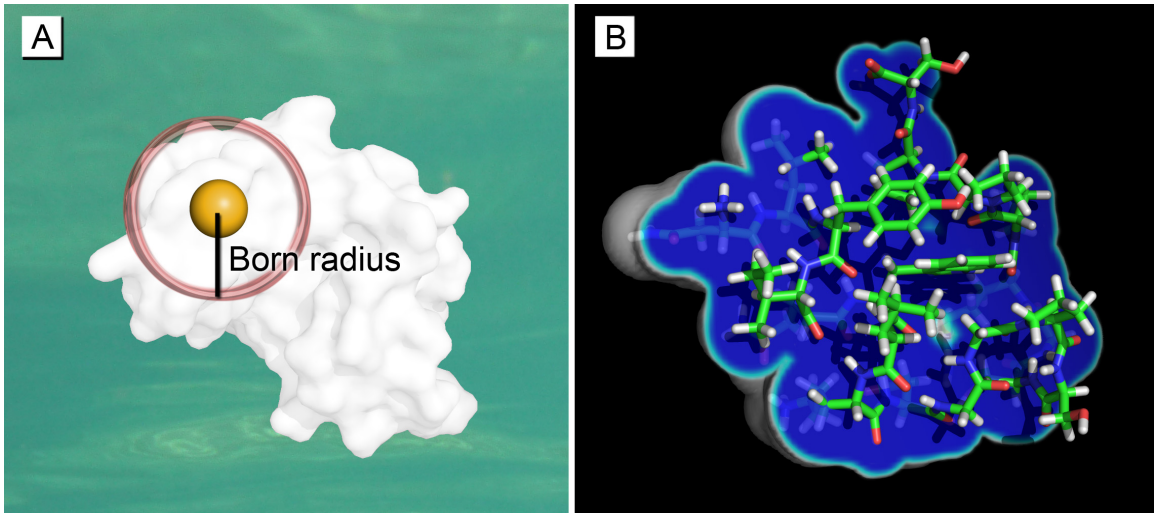

Figure 6.1 A) The Born radius (black line) of an atom (yellow ball) is shown with respect to a Trp-Cage miniprotein. Notice it is the approximate distance to the solvent. B) The same protein is shown with a partially-removed isosurface of atom density, which ranges from "solute" (dark blue) to "solvent" (white). The switching function exists in between (cyan), which makes the solute-solvent transition continuous.

$$V_{solute}(\mathbf{r}-\mathbf{r}_a) = \prod_{b}^{solute\ atoms} v_b(r_{b,\mathbf{r}-\mathbf{r}_a}) \qquad \text{(eq. 11)}$$

Here $r_{b,\mathbf{r}-\mathbf{r}_a}$ is the distance between atom b and the point in space $\mathbf{r}-\mathbf{r}_a$, and v_b is the atomic volume-inclusive function. This function determines whether a given point in space is inside (0), or outside (1) an atom. Since discontinuities in the dielectric boundary can cause numerical instability in calculations of solvation forces, we employ a simple switching function in v_b from which GBSW gains its name. The switching function blurs the hard boundary with a cubic function, and continuously links the interior and exterior of an atom in the following relationship.

$$v_a(r)=\begin{cases}0 & r \le R_a^{atom}-s_w \\ \frac{1}{2}+\frac{3}{4s_w}\left(r-R_a^{atom}\right)-\frac{3}{4{s_w}^3}\left(r-R_a^{atom}\right)^3 & R_a^{atom}-s_w < r < R_a^{atom}+s_w \\ 1 & r \ge R_a^{atom}-s_w \end{cases} \quad \text{(eq. 12a)}$$

The term v_a is a function of distance r from the center of atom a. R_a^{atom} is the atomic radius of atom a that defines a dielectric boundary that is consistent with PB calculations, and s_w is the switching length that determines the thickness of the switching function. The default switching length in GBSW is 0.3 Å.[13] One of the most notable benefits of the switching length is it fills small voids with atomic density, which in turn corrects for an underestimation on the Born radius when integrating over small crevices. Figure 6.1a illustrates the Born radius relative to an atom inside a molecule, and figure 6.1b shows a cross-section of the switching function.

Interestingly, when the atomic radii are optimized to recapitulate the exact Born energy from PB calculations, they differ from the van der Waals radius used for Lennard-Jones potentials. Chen et al. have produced the latest such modifications to atomic radii for amino acid side chains, which include larger radii for methyl carbons and zero radii for hydrogens.[29,39]

In equation 12a we find a method for including the low-dielectric environment of an implicit membrane. Much like how the solvent volume exclusion functions (eq. 11 and 12a) simulate the low dielectric of a protein's interior by removing atom-sized volumes from the solvent, the low dielectric environment inside of a membrane is simulated by removing a slab of solvent volume in the following manner:[13,32]

$$v_{mem}(\mathbf{r}^z)=\begin{cases}0 & |\mathbf{r}^z|\le R^{mem}-s_w \\ \frac{1}{2}+\frac{3}{4s_w}\left(|\mathbf{r}^z|-R^{mem}\right)-\frac{3}{4{s_w}^3}\left(|\mathbf{r}^z|-R^{mem}\right)^3 & R^{mem}-s_w<|\mathbf{r}^z|<R^{mem}+s_w \\ 1 & |\mathbf{r}^z|\ge R^{mem}+s_w\end{cases} \quad \text{(eq. 12b)}$$

Here we see that if the absolute value of the z-coordinate of a point in space $\mathbf{r}^z$ is such that it is less than the membrane's half-thickness, R^{mem}, then the membrane's switching function applies to that point in space. This option enables a membrane-like low dielectric to interact with the solute molecule, which scales an atom's Born radius by both its buriedness in a solute molecule, and by its buriedness in a membrane. This setup has been useful in predicting structures of transmembrane domains of G protein-coupled receptors.[40]

6.4 Numerical Quadrature

Im et al.[13] optimized a spherical quadrature method for calculating the Born radii as a means to sample the atomic density surrounding each atom, and rapidly calculate the volume exclusion functions $1-V_{solute}(\mathbf{r}-\mathbf{r}_a)$. The setup involves placing points of integration (quadrature points) around each atom, determining the v_a values for each atom near those quadrature points, calculating the value of $1-V_{solute}(\mathbf{r}-\mathbf{r}_a)$, and scaling the result of each point by a corresponding volumetric weight. Retooling equation 9 with quadrature points results in

$$\left(R_a^{Born}\right)^{-1}\approx\alpha_0\left[\sum_{quad}w_{quad}\frac{1-V_{solute}(\mathbf{r}_{quad}+\mathbf{r}_a)}{\left(r_{a,\,\mathbf{r}_{quad}+\mathbf{r}_a}\right)^2}\right]+\alpha_1\left[\sum_{quad}w_{quad}\frac{1-V_{solute}(\mathbf{r}_{quad}+\mathbf{r}_a)}{\left(r_{a,\,\mathbf{r}_{quad}+\mathbf{r}_a}\right)^5}\right]^{1/4} \quad \text{(eq. 13)}$$

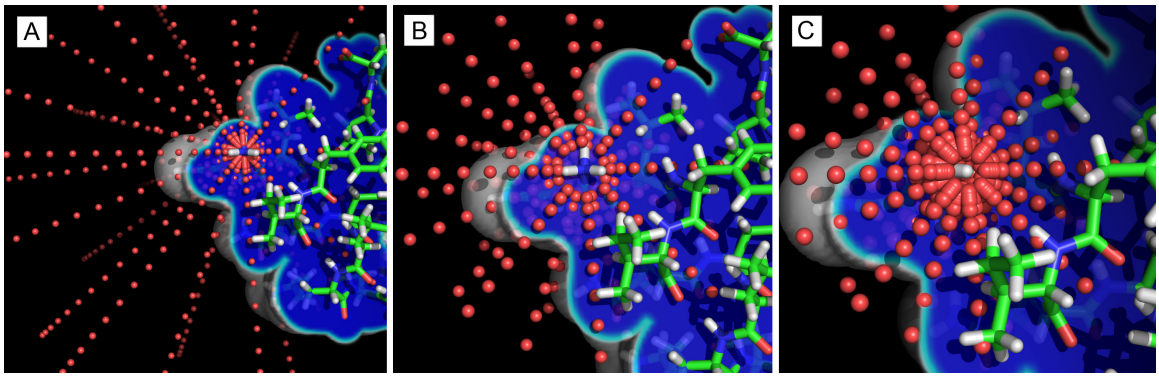

Figure 6.2 A) The original 1200 quadrature point cloud (red dots) around the R1 N atom of Trp-Cage miniprotein. Each point samples its local atom density $V_{solute}(\mathbf{r})$ as being 1 (dark blue; inside an atom), 0 (outside an atom), or in between (cyan), and offers a contribution to the atom's Born radius. B) The modified 350 quadrature point cloud which retains the quadrature points that most contribute to the Born energy and Born force. Points near the atom's center are assumed to have $V_{solute}(\mathbf{r}) = 1$, and points far from the atom are assumed to have $V_{solute}(\mathbf{r}) = 0$. The remaining points recapitulate greater than 99.5% of the original forces vectors and atom-wise energy. C) The modified 500 quadrature point cloud for hydrogen atoms, here around the R1 HT2 atom of Trp-Cage. Similarly to the nitrogen atom's quadrature scheme, points very close and very far from the atom's center are unnecessary to calculate explicitly. Since hydrogen has a smaller atomic radius than nitrogen, fewer quadrature points near the atom's center could be omitted.

Here w_{quad} is the integration weight for a quadrature point, and $\mathbf{r}_{quad} + \mathbf{r}_a$ is the placement of that quadrature point in space as projected from atom a. Meanwhile, $r_{a,\mathbf{r}_{quad}+\mathbf{r}_a}$ is the distance between atom a and the quadrature point $\mathbf{r}_{quad} + \mathbf{r}_a$. The default quadrature point cloud is comprised of 1200 points that are determined by the combination of a 50-point Lebedev quadrature[41] and two second-order Gaussian-Legendre quadratures.[42] The angular Lebedev quadrature distributes 50 points on the surface of a unit sphere, and each radius of both Gaussian-Legendre quadratures scales a Lebedev quadrature to sample a volume. The 24 radii are determined by a 5-radius Gaussian-Legendre quadrature from 0.5 Å to 1 Å, and another 19-radius quadrature from 1 Å to 20 Å. The result is 1200 points that sample a spherical volume around each atom, and 500 points within a 1 Å radius around each atom. In addition to offering a high sampling rate near each atomic center, the quadrature setup

also provides an integration process that is straightforward to parallelize: the integration value at each point can be performed independently from the others. Figure 6.2 illustrates the quadrature point cloud around each atom. In addition to offering a high sampling rate near each atomic center, the quadrature setup also provides an integration process that is straightforward to parallelize: the integration value at each point can be performed independently from the others. Figure 6.2 illustrates the quadrature point cloud around each atom. Notice that the $(4\pi)^{-1}$ from eq. 12 is included in the weights of the Levadev quadrature.

6.5 Nonpolar Energy

The nonpolar contribution ΔG^{np} represents the energy used to cavitate a solvent around a solute. Although physically relevant, this component of the energy is not included in the GBSW model by default. Nevertheless, we discuss its implementation.

Schaefer and coworkers along with modifications from Jay Ponder calculated the nonpolar energy of solvation with the following relationship,[34,43]

$$\Delta G^{np} = \sum_a \Delta G_a^{np} = 4\pi\gamma \sum_a \left(R_a^{atom} + R^{probe}\right)^2 \left(\frac{R_a^{atom}}{R_a^{Born}}\right)^6 \qquad \text{(eq. 14)}$$

Here ΔG^{np} is a sum of nonpolar contributions from each atom a, each of which is derived from a relationship among the atomic radii R_a^{atom}, Born radii R_a^{Born}, the phenomenological constant γ, and the "probe radius" R^{probe}, which corresponds to the radius of a water-molecule-sized sphere. The original fraction had an exponent of 1 but unpublished work

from Ponder found that a higher-order exponent of 6 better captured the solute molecule's solvent-accessible surface area. This equation establishes that the smaller the Born radius of an atom, the closer it is to the solvent, and thus it has a larger contribution to the surface area. Conversely an atom with a large Born radius is far from the solvent-accessible surface area, and thus gives a smaller surface area contribution. The higher exponent greatly enhances this relationship, and better removes the contributions to the surface area from buried atoms with larger Born radii. Again we note that the nonpolar contribution to solvation free energy is small, and ignored by default in GBSW. Should the nonpolar energy and forces be enabled during a simulation, we adopted the formalism already used in OpenMM to calculate it.[34]

6.6 Calculating the Forces

Calculating the forces of implicit solvation becomes complicated because the effective Born radius of an atom is a function of all solute atoms in the system. Thus the force on any atom also depends on the placement of every other atom in the system. When we deconvolve the force with respect to atom-atom distances and Born radii, we arrive at,

$$\frac{\partial \Delta G^{elec}}{\partial \mathbf{r}_a} = \frac{\partial \Delta G^{elec}}{\partial r_{ab}} \frac{\partial r_{ab}}{\partial \mathbf{r}_a} + \sum_b \frac{\partial \Delta G^{elec}}{\partial R_b^{Born}} \frac{\partial R_b^{Born}}{\partial \mathbf{r}_a} \qquad \text{(eq. 15)}$$

Here we find two terms. The first force component is centralized on atom a, and is a Coulomb-like interaction between atom pairs. When expressed in greater detail, it becomes

$$\frac{\partial \Delta G^{elec}}{\partial r_{ab}} \frac{\partial r_{ab}}{\partial \mathbf{r}_a} = \frac{\tau}{4} \sum_{ab} \frac{q_a q_b \left[4 - \exp\left(-D_{ab}\right) \right]}{\left(f_{ab}^{GB} \right)^3} \left(\mathbf{r}_b - \mathbf{r}_a \right) \qquad \text{(eq. 16a)}$$

where

$$D_{ab} = \frac{r_{ab}^2}{R_a^{Born} R_b^{Born}} \qquad \text{(eq. 17)}$$

For systems with a low salt concentration, and a non-zero Debye-Huckel screening parameter we instead arrive at

$$\frac{\partial \Delta G^{elec}}{\partial r_{ab}} \frac{\partial r_{ab}}{\partial \mathbf{r}_a} = \frac{1}{4} \sum_{ab} \frac{q_a q_b \left[4 - \exp\left(-D_{ab}\right) \right]}{\left(f_{ab}^{GB} \right)^3} \left(\mathbf{r}_b - \mathbf{r}_a \right) \left(\frac{\tau}{f_{bc}^{GB}} - \kappa \frac{e^{-\kappa f_{bc}^{GB}}}{\varepsilon_s} \right) \qquad \text{(eq. 16b)}$$

Here the $\mathbf{r}_b - \mathbf{r}_a$ gives direction to the force vector. Meanwhile, the atomic charges, distances, and Born radii scale the force. We note that when a and b are equal, the force of this component is zero. The second component arises from the electric displacement of solute atoms in a continuous dielectric, and effectively is the interaction between an atom and the molecular surface. When interpreted in the context of GBSW's quadrature, we see that it emerges as an atom-quadrature point interaction. The force is scaled first by deriving the Still equation (eq 3) with respect to the change in neighboring atoms' Born radii:

$$\frac{\partial \Delta G^{elec}}{\partial R_b^{Born}} = \frac{\tau}{2} \sum_{bc} \frac{q_b q_c \exp\left(-D_{bc}\right)}{\left(f_{bc}^{GB} \right)^3} \left(R_c^{Born} + \frac{r_{bc}^2}{4 R_b^{Born}} \right) \qquad \text{(eq. 18a)}$$

For systems with a non-zero Debye-Huckel screening parameter we alternatively arrive at

$$\frac{\partial \Delta G^{elec}}{\partial R_b^{Born}} = \frac{1}{2}\sum_{bc} \frac{q_b q_c \exp\left(-D_{bc}\right)}{\left(f_{bc}^{GB}\right)^2}\left(R_c^{Born} + \frac{r_{bc}^2}{4R_b^{Born}}\right)\left(\frac{\tau}{f_{bc}^{GB}} - \kappa\frac{e^{-\kappa f_{bc}^{GB}}}{\varepsilon_s}\right) \qquad \text{(eq. 18b)}$$

The remaining component of the force computes the atom-molecular surface interaction as atom-quadrature point $\partial R_b^{Born} / \partial \mathbf{r}_a$ contributions. This becomes

$$\frac{\partial R_b^{Born}}{\partial \mathbf{r}_a} = \left(R_b^{Born}\right)^2 \sum_{quad} w_{quad}\left[\frac{\alpha_0}{\left(r_{b,\mathbf{r}_{quad}}\right)^2} - \frac{\alpha_1}{4\left(r_{b,\mathbf{r}_{quad}}\right)^5}\left(\frac{\tau q_b^2}{2(\Delta G_b^{elec,1})}\right)^3\right]\frac{-\partial V_{solute}(\mathbf{r}_{quad}+\mathbf{r}_b)}{\partial \mathbf{r}_a} \qquad \text{(eq. 19)}$$

where

$$\frac{\partial V_{solute}(\mathbf{r}_{quad}+\mathbf{r}_b)}{\partial \mathbf{r}_a} = \sum_{b\neq a}\frac{V_{solute}(\mathbf{r}_{quad}+\mathbf{r}_b)}{v(r_{b,\mathbf{r}_{quad}+\mathbf{r}_b})}\left(\frac{3}{4s_w} - \frac{3}{4s_w{}^3}(r_{b,\mathbf{r}_{quad}+\mathbf{r}_b} - R_a^{PB})^2\right)\frac{\mathbf{r}_{quad}+\mathbf{r}_b-\mathbf{r}_a}{\left|\mathbf{r}_{quad}+\mathbf{r}_b-\mathbf{r}_a\right|} \qquad \text{(eq. 20)}$$

$$+ \sum_{a=b,a\neq c}\frac{V_{solute}(\mathbf{r}_{quad}+\mathbf{r}_c)}{v(r_{a,\mathbf{r}_{quad}+\mathbf{r}_c})}\left(\frac{3}{4s_w} - \frac{3}{4s_w{}^3}(r_{a,\mathbf{r}_{quad}+\mathbf{r}_c} - R_c^{PB})^2\right)\frac{\mathbf{r}_a-\mathbf{r}_{quad}+\mathbf{r}_c}{\left|\mathbf{r}_a-\mathbf{r}_{quad}+\mathbf{r}_c\right|}$$

The derivative of the volume exclusion function V_{solute} is split between one part where quadrature points from an atom a mediate an interaction with atoms b, and the inverse where atom a is interacting with a quadrature point of another atom c. Notice that for quadrature points where $V_{solute}(\mathbf{r}_{quad}) = \{0,1\}$ there will always be at least one switching function such that $\partial v(r)/\partial \mathbf{r}_a = 0$. Thus eq. 20 only receives contributions from quadrature points residing at the dielectric boundary where there is a nonzero slope of the volume exclusion function. We have arrived at the equations GBSW uses to generate both the solvation free energy, and its derivative force on all solute atoms. The most computationally-

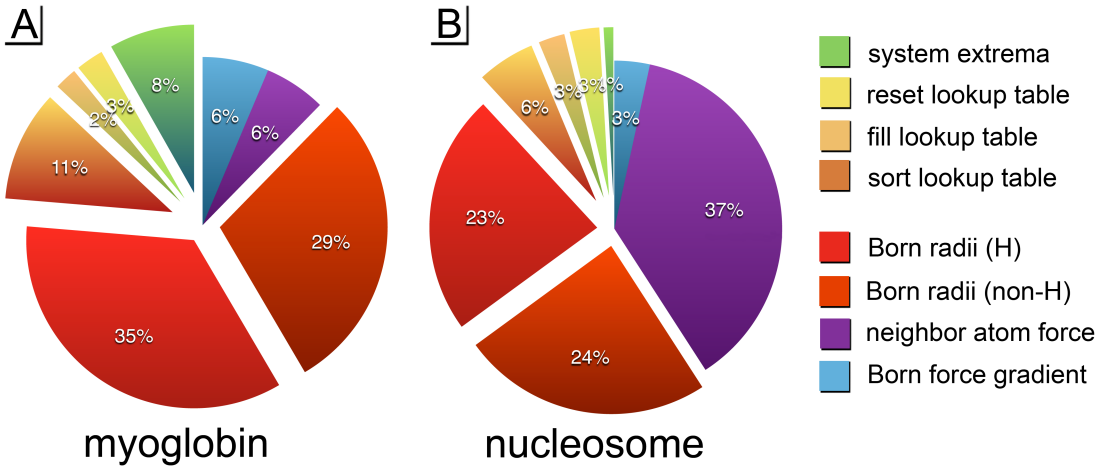

Figure 6.3 These are the approximate distributions of CPU time spent on two systems: A) myoglobin with 2459 atoms, and B) the eukaryotic nucleosome with 22481 atoms. Notice that the neighbor lookup table is the only kernel that doesn't scale approximately with O(N) complexity. In larger systems the neighbor-atom force becomes the most expensive part of the forcefield calculation.

demanding portion of this algorithm is calculating the quadrature point contributions. The demand arises both from the great number of quadrature points in the system, and their potential interaction with any atom in the solute. Fortunately these challenges were met well by the nature of GPU architecture. Now we explore the framework used to perform this calculation both efficiently and in parallel.

6.7 Function Design and Parallelization

The GPU languages CUDA and OpenCL are designed for massively-parallel processes, or kernels, that execute efficiently when a problem can be divided into many smaller parallel pieces. Due to their architecture, processing time is both related to the speed of each processing core as well as the number of processing cores. Graphics chips of today can have up to many tens of multiprocessors, each consisting of up to 64 discrete processing

cores. Thus kernels at minimum must be split into thousands of parallel tasks to take full advantage of the parallel architecture of today's GPUs.

The overall structure of a kernel running on a GPU is divided into blocks and threads, and data is stored in a hierarchical memory structure of increasing speed and decreasing capacity: global, shared, and local memory. Blocks are parallel processes that can ideally run in any order, and only communicate to each other on global memory. These processes are analogous to the multi-processor tasks in C++, Fortran, and other multi-processor CPU languages. Unlike the CPU, however, each of these parallel tasks can be further subdivided into groups of related threads on a GPU. On graphics cards, each thread has its own local memory, and threads can communicate through a high-speed shared memory as they process a calculation. Additionally, threads can initiate, stop, and synchronize with other threads, allowing for a precise level of control over both data management and speed. For instance, this study covers a kernel used for quadrature integration that designates one block for each atom, and one thread for each quadrature point. Midway through the calculation, all threads in a block share Born radii calculation results and intermediate values to compute Born radius gradients.

Although a fast implementation of GBSW can be built directly from CUDA, a myriad of complexities arise if the code is to be robust on all computing systems. Depending on the year a GPU was manufactured, available memory, number of processors, and thread management capabilities are different. The OpenMM software toolkit developed by Eastman et al.[34] addresses this complexity through a rapid update cycle, and an execution step that effectively redesigns kernels to suit the available hardware. Its accomplishments were notable enough to the CHARMM community that a CHARMM-OpenMM interface was developed so that CHARMM could take advantage of GPUs in an efficient manner. Simulations can

now be designed using CHARMM's robust algorithms for processing macromolecules, and through the software interface, CHARMM controls OpenMM's kernels to run the dynamics of a simulation. The GBSW algorithm was developed as a stand-alone solvent model within OpenMM, and as part of the CHARMM-OpenMM interface. This way the GBSW kernels are slightly tailored to different GPU hardware under the guidance of OpenMM, and effectively utilize a wide range of available hardware.

The GBSW calculation is broken up into a total of 8 kernels, 4 of which organize an atom-lookup table, and the other 4 calculate the Born energies and forces on each atom. The lookup table is a multidimensional array that facilitates the rapid calculation of the Born radii, and is the most memory-intensive part of the calculation. Meanwhile calculating the Born energy and forces are the most computationally-intensive. In the next few sections we discuss the details of those kernels, and overcoming the challenges associated with them. The approximate time spent per kernel is shown in Figure 6.3.

6.8 Baseline of Error

As we explore the various alterations and assumptions made in this iteration of the GBSW algorithm, we must ensure that it accurately recapitulates the original algorithm. Since the quadrature points are fixed along Cartesian coordinates, they do not rotate with the system's atoms. Consequently there is an inherent rotational variance in the energy and force vectors produced by the integration algorithms, as shown in Figure 6.4. Rotational variance was explored using a 4,107-atom system generated using the small ribosome subunit proteins S1, S2, and S3 from the PDB 4V88. The standard deviation in forces and angles caused by rotational variance provides a benchmark of accuracy for the CUDA-GBSW algorithm: 0.32

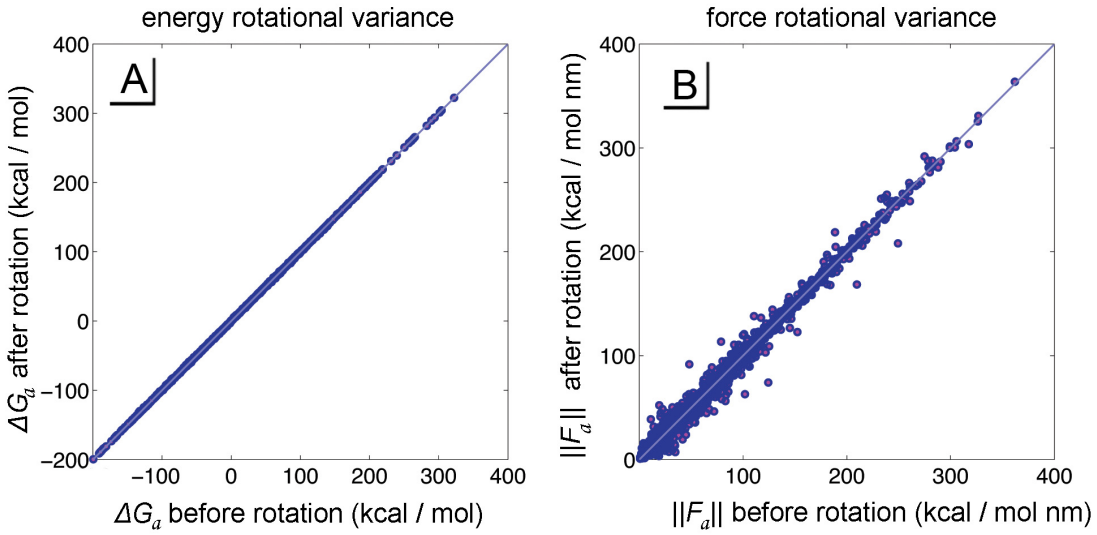

Figure 6.4 The rotational variance of the original GBSW forcefield was explored by randomly rotating a 4,107-atom system and observing the resulting changes in forces and energies of each atom. Shown here are the variations for one rotation in A) energy magnitudes and B) force magnitudes of individual atoms (small light blue dots). These data provided a minimum baseline of accuracy for GBSW as we altered the algorithm and made it suitable for parallel processing.

kcal/mol in magnitude of energy, 4.28 kcal/mol nm in magnitude of force, and 17.23 degrees in the angle of force.

6.9 Atom Lookup Table

An important assistant to calculating the Born radii is an atom-lookup table that returns the resident atoms at a given point in space. We found the most efficient memory structure is a voxelized representation of 3-dimensional Cartesian coordinates, where each XYZ grid coordinate contains an array of atoms residing at that gridpoint. An atom a is identified as residing inside all voxels that meet the distance criteria

$$r_{atom,voxel} \le R_a^{atom} + s_w + L_{voxel}\left(\sqrt{3}/2\right) \quad \text{(eq. 21)}$$

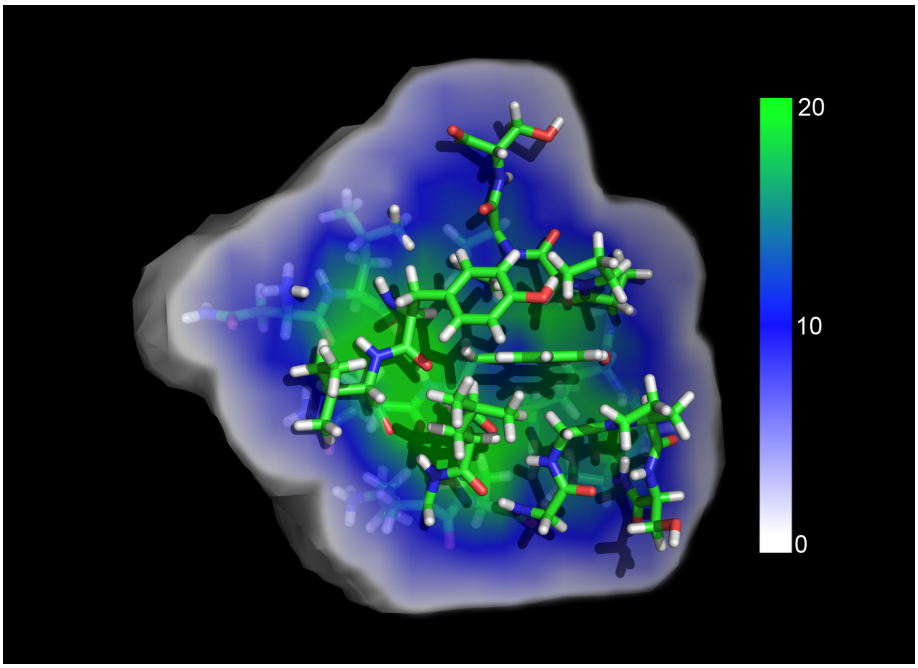

Figure 6.5 The lookup table is a multidimensional array that tracks which atoms exist in what part of space by using a 3D grid. Each grid voxel is a cube with a side length of 1.5 Å. Shown here is a cutaway representation of the number of atoms at each gridpoint, ranging from white (0 atoms) to green (20+ atoms). The condition that determines whether or not an atom resides in a voxel is shown in eq. 20.

where the atom's distance to a voxel $r_{atom,voxel}$ is small enough that part of the atom (or its switching function) may reside inside a cubic voxel of side length L_{voxel}. In the Fortran90 implementation of GBSW, the lookup table was a 3D rectangular array with index locations, another 3D rectangular array containing the length of the atom list at a gridpoint, and a large 1-dimensional (1D) array containing the concatenated atom lookup lists at each gridpoint. The two 3D grids provided the location of the local resident atom array within the large 1D array. Because the arrays were resized and reshaped with every timestep, this method gave a small footprint in computer memory.

Unfortunately, the dynamic rearrangement of data within a large 1D array is not efficiently parallelizable, and dynamically allocating memory is not possible inside of a GPU

process. To solve these problems we allocate a large 4-dimensional array that holds a small 1D array for every location in 3D space. By converting a point in space to a voxel in the first 3 dimensions, the lookup array returns the small 1D array containing the number of resident atoms at that location, and the atom indices of those atoms. An isosurface representation of the 3D portion of the lookup table is shown in Figure 6.5. The caveat of this method of data storage is that it requires enough memory to contain all reasonable configurations of the system before any kernels are executed (dimensions X * dimensions Y * dimensions Z * number of voxels per gridpoint). Consequently, this lookup array is the single largest memory requirement for most, if not all, GPU simulations using the GBSW implicit water model. Thankfully, though, graphics memory is relatively cheap and seems to be more plentiful with each new generation of GPU's. We find the parallelized 4D array requires about 11 times the memory footprint of the previous algorithm, but now each voxel can be processed in parallel.

A cubic voxel length L_{voxel} of 1.5 Å is used for the lookup table, which allows for a rapid filling of the lookup table, a smaller 1D atom list in each voxel, and ultimately a smaller memory footprint of the array. Decreasing this length increases memory requirements exponentially, but decreases the time spent calculating the energies and forces. This value may change as memory on GPU chips becomes more abundant. Meanwhile, the length of the 1D atom list at each voxel was set to 25 atoms. This length was found to contain a sufficient number of atom indices for an accurate GBSW calculation as shown in Figure 6.5. Although some regions of space may contain more than 25 atoms, there is a diminishing influence on Born forces and Born energy when additional atoms are accounted for. We note that setting 23 as the maximum number of atoms still provides accurate calculations of Born forces and energies.

Adding a buffer or an extension to the radial parameter $L_{voxel}(\sqrt{3}/2)$ can reduce the need to update the lookup table with every timestep. This change, however, not only increases the burden on memory allocation, but also increases the amount of time needed for the expensive Born radii calculation. Ultimately the GBSW algorithm is fastest when each voxel stores the shortest atom list possible. As a side note, the nearest-neighbor atom lookup tables used for Lennard-Jones and electrostatic force calculations can also be used to calculate a correct set of Born radii. This form of calculating GBSW requires that all quadrature points around an atom a need to check their distance from all neighboring atoms of atom a. This arrangement would make the Born radius calculation prohibitively inefficient, and was only used to check the accuracy of the calculations.

The lookup table kernels run in O(N) time, where for 2 kernels N is the number of atoms, and for the other 2 N is the number of voxels in the system. Because there are volumetric components to this calculation, conformation can change the speed of creating the atom lookup table. Despite its complexity, this CUDA algorithm is approximately 90 times faster than its previous CPU iteration, and it owes most of its speed increase to the fact that it now runs in parallel. A full description of the lookup table kernels follows in Table 6.1.

kernel	no. blocks	no. threads per block	description
System extrema	1	64	Calculate the maximum and minimum dimensions of the system, define the size and shape of the lookup table. When periodic boundaries are used, these calculations only need to be performed once and this kernel is ignored.
Reset lookup table	dimX * dimY * dimZ	1	Each block attends one voxel, and the single thread in each block sets the number of atoms per voxel to 0.
Fill lookup table	number of atoms	256	Each block attends one atom, and each thread attends one voxel near that atom. If the atom resides in this voxel (see eq. 28), then this atom's index is recorded.
Sort lookup table	dimX * dimY * dimZ	25	Each block sorts one voxel of space. Each thread sorts one atom in the 1D lookup list at the block's voxel. Atoms are sorted by distance to the voxel's center $r_{atom,\,voxel}$ using a parallel bubble-sort. Although $O(N^2)$ operations take place during the sort, the parallel architecture allows it to run in O(N) time. Since the sorted array finds quadrature points inside atoms faster than unsorted arrays, it speeds up the Born radius calculation by about 40 % over using an unsorted lookup table.

Table 6.1 Detailed description of the 4 lookup table kernels.

6.10 Born Radii, Forces, and Energy Calculation

The Born radii calculation consists of calculating the volume exclusion function V_{solute} for each quadrature point in parallel, and then combining those values to integrate the Born radii using equations 11 through 13. When is portion of the work is combined with a well-designed and sorted lookup table, parallel graphics processing offers a speedup of more than an order of magnitude over the single-core iteration of the algorithm.

Although the Born radii calculation is the most parallelizable part of the GBSW algorithm, it still remains the most computationally expensive kernel as shown in Figure 6.3. The end goal of this study is to speed up GBSW as much as possible while preparing it for a future of parallel processors, so we set out to determine precisely how many of the quadrature points need to be calculated. The contribution to the Born energy and forces diminishes with r^{-2} and r^{-5} as shown in eq. 13. Additionally, the default coefficients α_0 and α_1 indicate that the r^{-5} term provides the greatest contribution to the calculation. Noting these aspects, we can explore reducing the maximum for quadrature radius $r_{a,\mathbf{r}_{quad}+\mathbf{r}_a}$ by assuming that $V_{solute}(r)=0$ for various integration radii. Additionally, since many quadrature points are guaranteed to reside within atomic radii R_a^{atom}, quadrature points closest to the atom centers can be pre-integrated by assuming that $V_{solute}(r)=1$.

We conclude that an accurate calculation of Born energies and forces only requires 500 quadrature points for hydrogens and 350 points for heavier, non-hydrogen atoms. Not only do we reduce the total number of points needed for the algorithm, we also significantly reduce the number of threads required per block. When implemented on the GPU, this new integration setup is roughly 30 times faster than the previous single-core calculations of Born

radii. For a visual reference, Figure 6.2b and 6.2c show these integration points around atoms.

After calculating the Born radii, neighboring-atom facilities already exist in OpenMM that efficiently calculate $(\partial \Delta G^{elec} / \partial r)(\partial r / \partial \mathbf{r})$ and $(\partial \Delta G^{elec} / \partial R^{Born})$ from eq. 16 and eq. 18, respectively, in a single kernel. Additionally, this same kernel calculates the free energy of solvation ΔG_a^{elec} for each atom, which greatly speeds the majority of the force calculation. These facilities vary greatly in efficiency depending on input parameters and system configuration, but in general marginalize the time requirements for neighboring-atom interactions for systems smaller than 20,000 atoms, as shown in Figure 6.3.

The final component of GBSW is calculating the Born radius gradient $\partial R^{Born} / \partial \mathbf{r}$ from eq. 19. The original algorithm for GBSW was optimized for a minimum memory footprint, and consequently favored recalculating values over saving them to memory. As such the Born radius gradient was a stand-alone function that required a similar time as calculating the Born radii. In this iteration of GBSW we calculate the Born radius gradient at the same time as the Born radii, and then we save the resulting $\partial R_b^{Born} / \partial \mathbf{r}_a$ values in a vector array. As with the lookup table, this array was capped with a maximum number of gradient contributions per atom. Figure SI3 shows that a maximum of 196 quadrature interactions per atom does not significantly alter the force calculation, and so a generous cap of 256 interactions was used. The result is a rapid, parallel calculation of both the Born radii and their gradients from which we gain yet more speed improvement over the original algorithm.

kernel	no. blocks	no. threads per block	description
Calculate Born radii for hydrogens	number of atoms	500	Each block attends one hydrogen atom, each thread attends one quadrature point. The atom density V_{solute} at each quadrature point is calculated. Then those densities are integrated to generate the Born radius R^{Born} and the derivative $\partial R^{Born} / \partial \mathbf{r}$ from eq. 14 and eq. 23 respectively.
Calculate Born radii for heavy atoms	number of atoms	350	Each block attends one non-hydrogen atom, each thread attends one quadrature point. This kernel functions the same as the "*Calculate Born radii for hydrogens*" kernel and calculates R^{Born}, and the derivative $\partial R^{Born} / \partial \mathbf{r}$ for heavier atoms. The differences lie in the integration offsets and number of threads used to accommodate the larger radii of heavier atoms.
Calculate neighbor-atom force	number of neighbor-atom tiles	256	Each block attends one atom-atom comparison tile of the OpenMM neighbor-atom list, and each thread attends one atom-atom pair in that tile. The atom-atom interactions of all atoms, their Born radii, and charges are combined to calculate $(\partial \Delta G^{elec} / \partial r)(\partial r / \partial \mathbf{r})$ and $(\partial \Delta G^{elec} / \partial R^{Born})$ from eq. 20 and eq. 22 respectively. Additionally, the GBSW free energy of solvation ΔG^{elec} is calculated for the system.
Born force gradient	number of atoms	256	Each block attends an atom, and each thread attends a contribution of $\partial R_b^{Born} / \partial \mathbf{r}_a$. The final value of $(\partial \Delta G^{elec} / \partial R^{Born})(\partial R^{Born} / \partial \mathbf{r})$ from eq. 14 is calculated and added to the total force on each atom. If the option for calculating the nonpolar contribution to the solvation free energy is requested, it is calculated in this kernel following eq. 13.

Table 6.2 Detailed description of the 4 Born energy kernels.

With the exception of the neighbor-atom force kernel which runs in O(N) to O(N^2) depending on the system configuration and input parameters, the GBSW Born radii and Born force calculations all operate on O(N) time where N is the number of atoms. As we will discuss in the next section, GBSW shows promise in scaling well with system size. For sufficiently large systems, GBSW emerges as one of the fastest solvent methods available at the time of this study. A full description of the Born radii kernels follows in Table 6.2.

6.11 Accuracy and Speed Gains Exhibited by CUDA-GBSW

We have outlined the basic setup of a new parallel CUDA-GBSW algorithm that shows great speed increases over the original Fortran90 GBSW. Although the new algorithm shows the same size-dependent scaling as the original, the new algorithm maintains useful speeds of nanoseconds per day even when used to solvate systems greater than 100,000 atoms. During this study we subdivided the GBSW algorithm into many thousands of parallel tasks, all of which would benefit well with the addition of more processing cores in future graphics chips. This new solvation method is expected to gain speed benefits until each quadrature point thread has its own core. In a smaller system of 1000 atoms with 500 hydrogens, 250,000 parallel threads are used to calculate the Born radii of the hydrogen atoms. Such a system presumably would receive no speed improvements, only when more than a quarter-million cores exist on a single GPU.

We benchmarked the CUDA-GBSW algorithm and observed its ability to recapitulate the original GBSW algorithm in CHARMM as shown in Figure 6.6. For each point in Figure 6.6e a subselection of the small ribosomal subunit (PDB code 4V88) was used for the benchmark. Since these subselections were not necessarily as dense or compact

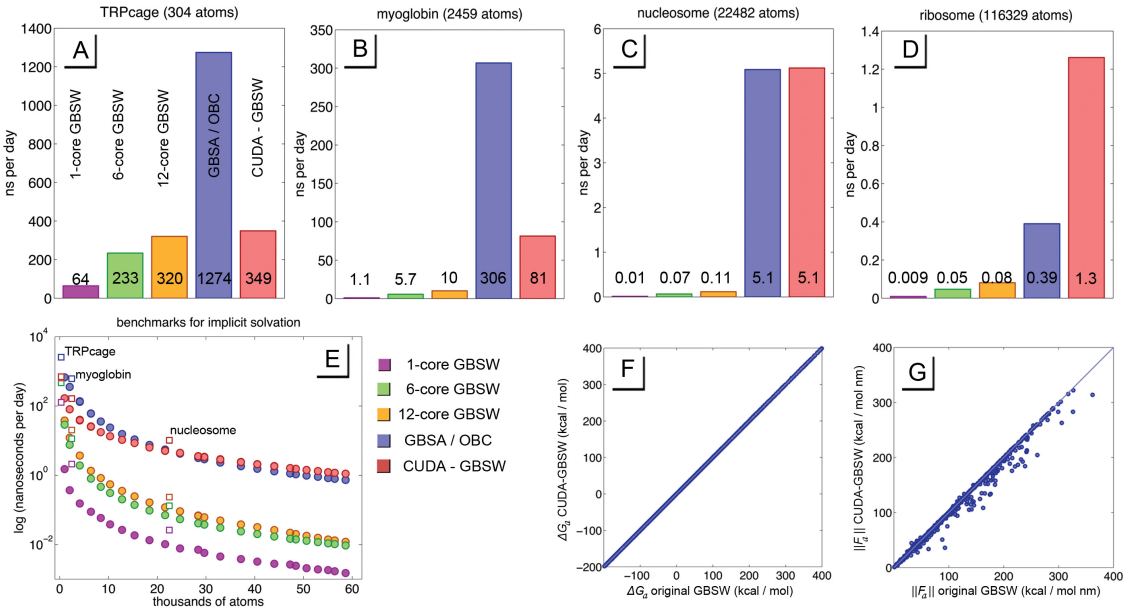

Figure 6.6 Above in plots A-D are benchmarks for specific systems in nanoseconds per day, which include benchmarks for the original CHARMM algorithm with 1, 6, and 12 cores (purple, green and yellow respectively), a benchmark for the GBSA / OBC forcefield running in OpenMM (blue), and the CUDA-GBSW forcefield discussed in this study (red). Plot E shows a logarithmic benchmark for various system sizes, all of which are comprised of one or more proteins from the small eukaryotic ribosomal subunit. The result is a smooth curve highlighting what system sizes receive what speed gains for various systems. The square icons represent benchmarks for the specific systems in plots A-D. These systems were TRPcage, myoglobin, nucleosome, and the small eukaryotic ribosomal subunit, with PDB codes 1L2Y, 1BVC, 1AOI, and 4V88 respectively. Plots F and G show the accuracy of CUDA-GBSW in recapitulating the forces and energies of the original GBSW algorithm in CHARMM.

a system as a folded protein, there was an added cost of processing a large, empty atom lookup grid. With this in mind, the log plot benchmarks slightly underestimate the CUDA-GBSW performance for more compact systems containing the same number of atoms.

We find that for smaller systems such as TRPcage (304 atoms), the CUDA-GBSW algorithm was only slightly faster than the 12-core multiprocessing GBSW, and less than one-third the speed of the CUDA-GBSA/OBC solvation method. However, CUDA-GBSW solvation scales mostly through O(N) scaling. It is an expensive calculation for each member N, but for large enough systems the better scaling compensates its complex algorithm. We

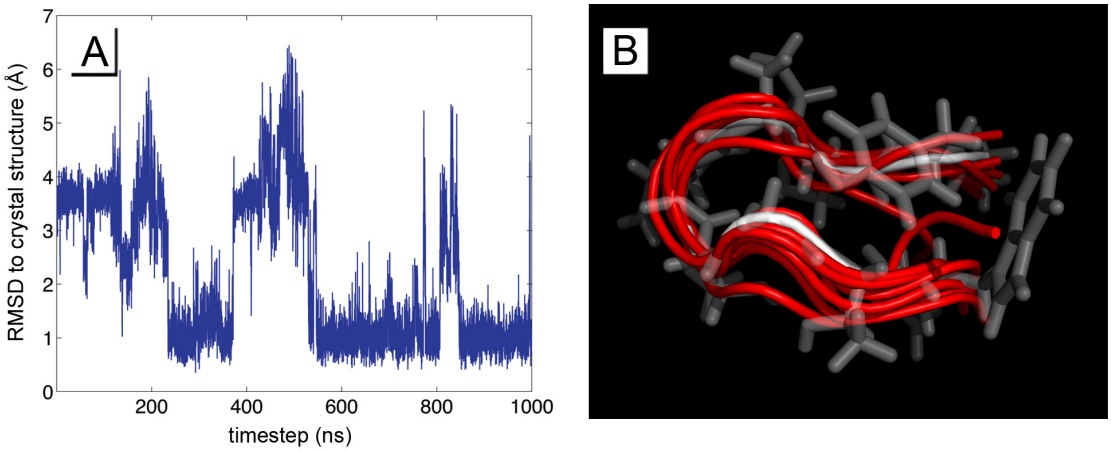

Figure 6.7 Chignolin was simulated in 8 replicas for 1 microsecond, and each trajectory was analyzed by RMSD to the PDB crystal structure 1UAO by backbone carbon atoms, and through an unbiased k-means clustering algorithm. A) shows a typical RMDS trajectory of comparing chignolin to the crystal structure, and B) overlays the dominant configurations from the k-means clustering (red) with the structure from 1UAO (white and transparent).

find that for systems greater than 22,000 atoms, CUDA-GBSW emerges as a more efficient implicit solvation method than GBSA/OBC, and over an order of magnitude faster than a multiprocessing Fortran90 GBSW solvation. When simulating large systems such as the small ribosomal subunit (116,329 atoms), CUDA-GBSW solvation is over 3 times faster than GBSA/OBC running on the same GPU.

We also reflect that CUDA-GBSW, despite its assumptions and simplifications, reproduces the original GBSW algorithm with less error than its inherent rotational variation. The differences between the two forcefields are 0.16 kcal/mol in magnitude of energy, 3.62 kcal/mol nm in magnitude of force, and 2.11 degrees in the angle of force. These amount to less than one percent difference between the two versions of GBSW. Thus we conclude that CUDA-GBSW accurately recapitulates the original algorithm of GBSW from CHARMM, and represents a good first iteration of the algorithm in modern parallel graphics processing languages.

6.12 Folding Chignolin

Chignolin is a 10-residue peptide consisting of 137 atoms, and when solvated with CUDA-GBSW runs at 438 ns/day in our computer setup. We simulated chignolin starting from in a linear, unfolded state in 8 replicas, each for 1 microsecond. The simulations were run using the CHARMM22 forcefield[44,45] using the Leapfrog Verlet integrator with an integration time step of 2 fs. These were NT simulations in an unbounded volume at a temperature of 298K using a Langevin heat bath. Atomic radii were optimized through work by Chen et al.[29] These simulations tested both the numerical stability of CUDA-GBSW during long simulations, and whether the algorithm and force field could find a reasonable structure for the native peptide. We analyzed the trajectories using unbiased k-means clustering to find the dominant conformation, and compared the trajectories to the crystal structure PDB 1UAO through backbone-atom root mean squared deviation (RMSD).

We found that of the 8 replicas, all trajectories explored configurations that were within 0.5 Å RMSD from the crystal structure. Additionally, the k-means clustering reported that 6 out of 8 trajectories were dominated by a structure within 1.5 Å RMSD of the crystal structure. Two simulations reported structures within 0.7 Å RMSD of PDB 1UAO. Figure 6.7 illustrates an RMSD trajectory, and a backbone-atom overlay of the k-means clustering results.

This exploration was designed only with testing the numerical stability of CUDA-GBSW in mind, and was not optimized for efficiency or accuracy. Nevertheless, it shows that CUDA-GBSW solvation is comparably efficient to other GPU-based GB models in folding chignolin,[46] and that GBSW running on GPUs remains appropriate for folding small

proteins when starting from a linear chain.[47-49] Additionally, these data indicate that the algorithm remains appropriate for exploring the conformational equilibria of small proteins.[47-49]

6.13 Future Directions

One of the greatest sources of accuracy and error of the GBSW algorithm lies in the placement of quadrature points around each atom. Through the better placement of each point, one may reduce the calculation time of the algorithm or enhance spatial sampling and reduce rotational variance. Each variation on quadrature point placement, though, carries the risk of requiring a full recalibration of the atomic radii and phenomenological constants. For instance, a variation of the Gaussian-Legendre quadrature was explored by using a single radial quadrature rather than two. This implementation, however, reduced sampling near the atomic centers and poorly recapitulated the PB free energies of solvation for each atom.

One possibility of increasing speed is by restricting the Lebedev quadrature only to integrate points away from neighboring atoms. Hydrogen atoms, for example, often lie inside the atomic radii of heavier atoms, and don't require a complete quadrature point cloud. Another option is to scan the solute molecule before a simulation to determine an optimal quadrature point setup for each atom. Such an option could begin the radial Gaussian-Legendre integration at an atom's switching function ($R_a^{atom} - s_w$), rather than the arbitrary distance of 0.5 Å from an atomic center as currently implemented in GBSW.

Although many more variations and improvements upon GBSW remain to be explored, what has been established is a parallel version that will improve greatly with each new generation of graphics chips for many years to come. Finally, we note that another

highly accurate, volumetric integration-based generalized Born model, GBMV,[26] has a similar algorithmic construction to the GBSW model we studied here. By applying similar GPU-based approaches, such as the CUDA-GBSW lookup table kernels, to the GBMV model, there is potential for giving the algorithm significant speed improvements. This remains a topic for future explorations.

6.14 References

1. S. M. Vaiana, M. Manno, A. Emanuele, M. B. Palma-Vittorelli, and M. U. Palma, "The Role of Solvent in Protein Folding and in Aggregation," J. Biol. Phys. **27**(2-3), 133-45, (2001).

2. V. Martorana, D. Bulone, P. L. San Biagio, M. B. Palma-Vittorelli, and M. U. Palma, "Collective Properties of Hydration: Long Range and Specificity of Hydrophobic Interactions," Biophys. J. **73**(1), 31-37, (1997).

3. E. J. Arthur, J. T. King, K. J. Kubarych, and C. L. Brooks, III, "Heterogeneous Preferential Solvation of Water and Trifluoroethanol in Homologous Lysozymes," J. Phys. Chem. B **118**(28), 8118-27, (2014).

4. L. S. Ahlstrom, S. M. Law, A. Dickson, and C. L. Brooks, III, "Multiscale Modeling of a Conditionally Disordered Ph-Sensing Chaperone," J. Mol. Biol. **427**(8), 1670-80, (2015).

5. B. H. Morrow, P. H. Koenig, and J. K. Shen, "Atomistic Simulations of Ph-Dependent Self-Assembly of Micelle and Bilayer from Fatty Acids," J. Chem. Phys. **137**(19), 194902, (2012).

6. B. R. Brooks, C. L. Brooks, III, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus, "Charmm: The Biomolecular Simulation Program," J. Comput. Chem. **30**(10), 1545-614, (2009).

7. I. T. E. Cheatham, and P. A. Kollma, "Observation of Thea-DNA Tob-DNA Transition During Unrestrained Molecular Dynamics in Aqueous Solution," J. Mol. Biol. **259**(3), 434-44, (1996).

8. J. T. King, E. J. Arthur, C. L. Brooks, III, and K. J. Kubarych, "Crowding Induced Collective Hydration of Biological Macromolecules over Extended Distances," J. Am. Chem. Soc. **136**(1), 188-94, (2014).

9. B. N. Dominy, and C. L. Brooks, III, "Development of a Generalized Born Model Parametrization for Proteins and Nucleic Acids," J. Phys. Chem. B **103**(18), 3765-73, (1999).

10. D. Bashford, and D. A. Case, "Generalized Born Models of Macromolecular Solvation Effects," Annu. Rev. Phys. Chem. **51**(1), 129-52, (2000).

11. V. Tsui, and D. A. Case, "Theory and Applications of the Generalized Born Solvation Model in Macromolecular Simulations," Biopolymers **56**(4), 275-91, (2000).

12. V. Tsui, and D. A. Case, "Molecular Dynamics Simulations of Nucleic Acids with a Generalized Born Solvation Model," J. Am. Chem. Soc. **122**(11), 2489-98, (2000).

13. W. Im, M. S. Lee, and C. L. Brooks, III, "Generalized Born Model with a Simple Smoothing Function," J. Comput. Chem. **24**(14), 1691-702, (2003).

14. J. Khandogin, and C. L. Brooks, III, "Constant Ph Molecular Dynamics with Proton Tautomerism," Biophys. J. **89**(1), 141-57, (2005).

15. J. Khandogin, and C. L. Brooks, III, "Toward the Accurate First-Principles Prediction of Ionization Equilibria in Proteins," Biochemistry **45**(31), 9363-73, (2006).

16. M. S. Lee, F. R. Salsbury, and C. L. Brooks, III, "Constant-Ph Molecular Dynamics Using Continuous Titration Coordinates," Proteins: Struct., Funct., Bioinf. **56**(4), 738-52, (2004).

17. J. Srinivasan, M. W. Trevathan, P. Beroza, and D. A. Case, "Application of a Pairwise Generalized Born Model to Proteins and Nucleic Acids: Inclusion of Salt Effects," Theor. Chem. Acc. **101**(6), 426-34, (1999).

18. W. Im, J. Chen, and C. L. Brooks, III. in *Adv. Protein Chem.* Vol. Vol. 72 (eds L. B. Robert, and B. David) 173-98 (Elsevier Academic Press, 2006).

19. R. Constanciel, and R. Contreras, "Self Consistent Field Theory of Solvent Effects Representation by Continuum Models: Introduction of Desolvation Contribution," Theoret. Chim. Acta **65**(1), 1-11, (1984).

20. W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson, "Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics," J. Am. Chem. Soc. **112**(16), 6127-29, (1990).

21. J. Warwicker, and H. C. Watson, "Calculation of the Electric Potential in the Active Site Cleft Due to Å-Helix Dipoles," J. Mol. Biol. **157**(4), 671-79, (1982).

22. I. Klapper, R. Hagstrom, R. Fine, K. Sharp, and B. Honig, "Focusing of Electric Fields in the Active Site of Cu-Zn Superoxide Dismutase: Effects of Ionic Strength and Amino-Acid Modification," Proteins: Struct., Funct., Bioinf. **1**(1), 47-59, (1986).

23. A. Nicholls, and B. Honig, "A Rapid Finite Difference Algorithm, Utilizing Successive over-Relaxation to Solve the Poisson–Boltzmann Equation," J. Comput. Chem. **12**(4), 435-45, (1991).

24. U. Haberthür, and A. Caflisch, "Facts: Fast Analytical Continuum Treatment of Solvation," J. Comput. Chem. **29**(5), 701-15, (2008).

25. A. Onufriev, D. Bashford, and D. A. Case, "Exploring Protein Native States and Large-Scale Conformational Changes with a Modified Generalized Born Model," Proteins: Struct., Funct., Bioinf. **55**(2), 383-94, (2004).

26. M. S. Lee, F. R. Salsbury, and C. L. Brooks, III, "Novel Generalized Born Methods," J. Chem. Phys. **116**(24), 10606-14, (2002).

27. J. L. Knight, and C. L. Brooks, III, "Surveying Implicit Solvent Models for Estimating Small Molecule Absolute Hydration Free Energies," J. Comput. Chem. **32**(13), 2909-23, (2011).

28. J. Chen, "Effective Approximation of Molecular Volume Using Atom-Centered Dielectric Functions in Generalized Born Models," J. Chem. Theory Comput. **6**(9), 2790-803, (2010).

29. J. H. Chen, W. P. Im, and C. L. Brooks, III, "Balancing Solvation and Intramolecular Interactions: Toward a Consistent Generalized Born Force Field," J. Am. Chem. Soc. **128**(11), 3728-36, (2006).

30. X. Zhu, P. Koenig, M. Hoffmann, A. Yethiraj, and Q. Cui, "Establishing Effective Simulation Protocols for β- and α/β-Peptides. Iii. Molecular Mechanical Model for Acyclic β-Amino Acids," J. Comput. Chem. **31**(10), 2063-77, (2010).

31. J. L. Knight, J. D. Yesselman, and C. L. Brooks, III, "Assessing the Quality of Absolute Hydration Free Energies among Charmm-Compatible Ligand Parameterization Schemes," J. Comput. Chem. **34**(11), 893-903, (2013).

32. W. Im, M. Feig, and C. L. Brooks, III, "An Implicit Membrane Generalized Born Theory for the Study of Structure, Stability, and Interactions of Membrane Proteins," Biophys. J. **85**(5), 2900-18, (2003).

33. D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods, "The Amber Biomolecular Simulation Programs," J. Comput. Chem. **26**(16), 1668-88, (2005).

34. P. Eastman, M. S. Friedrichs, J. D. Chodera, R. J. Radmer, C. M. Bruns, J. P. Ku, K. A. Beauchamp, T. J. Lane, L.-P. Wang, D. Shukla, T. Tye, M. Houston, T. Stich, C. Klein, M. R. Shirts, and V. S. Pande, "Openmm 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation," J. Chem. Theory Comput. **9**(1), 461-69, (2013).

35. B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, "Gromacs 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation," J. Chem. Theory Comput. **4**(3), 435-47, (2008).

36. J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, and K. Schulten, "Scalable Molecular Dynamics with Namd," J. Comput. Chem. **26**(16), 1781-802, (2005).

37. G. D. Hawkins, C. J. Cramer, and D. G. Truhlar, "Parametrized Models of Aqueous Free Energies of Solvation Based on Pairwise Descreening of Solute Atomic Charges from a Dielectric Medium," J. Phys. Chem. **100**(51), 19824-39, (1996).

38. M. S. Lee, M. Feig, F. R. Salsbury, and C. L. Brooks, III, "New Analytic Approximation to the Standard Molecular Volume Definition and Its Application to Generalized Born Calculations," J. Comput. Chem. **24**(11), 1348-56, (2003).

39. M. Nina, D. Beglov, and B. Roux, "Atomic Radii for Continuum Electrostatics Calculations Based on Molecular Dynamics Free Energy Simulations," J. Phys. Chem. B **101**(26), 5239-48, (1997).

40. M. Michino, J. Chen, R. C. Stevens, and C. L. Brooks, III, "Foldgpcr: Structure Prediction Protocol for the Transmembrane Domain of G Protein-Coupled Receptors from Class A," Proteins: Struct., Funct., Bioinf. **78**(10), 2189-201, (2010).

41. V. I. Lebedev, and D. N. Laikov, "A Quadrature Formula for the Sphere of the 131st Algebraic Order of Accuracy," Doklady Mathematics **59**(3), 477-81, (1999).

42. M. Abramowitz, and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables* (Dover, New York, 1965), "Chapter 25.4, Integration".

43. M. Schaefer, C. Bartels, and M. Karplus, "Solution Conformations and Thermodynamics of Structured Peptides: Molecular Dynamics Simulation with an Implicit Solvation Model," J. Mol. Biol. **284**(3), 835-48, (1998).

44. A. D. MacKerell, M. Feig, and C. L. Brooks, III, "Extending the Treatment of Backbone Energetics in Protein Force Fields: Limitations of Gas-Phase Quantum Mechanics in Reproducing Protein Conformational Distributions in Molecular Dynamics Simulations," J. Comput. Chem. **25**(11), 1400-15, (2004).

45. A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus, "All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins," J. Phys. Chem. B **102**(18), 3586-616, (1998).

46. H. Nguyen, J. Maier, H. Huang, V. Perrone, and C. Simmerling, "Folding Simulations for Proteins with Diverse Topologies Are Accessible in Days with a Physics-Based Force Field and Implicit Solvent," J. Am. Chem. Soc. **136**(40), 13959-62, (2014).

47. J. Chen, and C. L. Brooks, III, "Can Molecular Dynamics Simulations Provide High-Resolution Refinement of Protein Structure?," Proteins: Struct., Funct., Bioinf. **67**(4), 922-30, (2007).

48. J. Chen, and C. L. Brooks, III, "Implicit Modeling of Nonpolar Solvation for Simulating Protein Folding and Conformational Transitions," Phys. Chem. Chem. Phys. **10**(4), 471-81, (2008).

49. W. Im, J. Chen, and C. L. Brooks, III. in *Adv. Protein Chem.* Vol. Volume 72 173-98 (Academic Press, 2005).

Chapter 7

Refactoring the Constant pH Molecular Dynamics Method for Modern Graphics Processors

The work presented in this chapter has been published in the following paper:

1. E. J. Arthur, and C. L. Brooks, III, "Efficient Implementation of the Constant pH with Molecular Dynamics Method on Modern Graphics Processors," in progress.

7.1 Introduction

Proteins typically maintain their native structure and optimal functionality under a narrow range of pH.[1-3] Consequently, many biological systems tightly control local solvent pH to tune the effectiveness of enzymes, or to promote a useful protein conformation.[1,4,5] Mitochondrial ATP synthase utilizes a trans-membrane proton gradient to power its rotary catalysis mechanism,[6-8] and the departure from a normal pH range is known to be a driving force in forming the amyloid fibrils associated with Alzheimer's disease.[9,10] Additional examples of pH driven processes include the proton-activated gate mechanism of the KcsA potassium channel,[11] and the catalytic pathway of dihydrofolate reductase.[12] Finally, a notable survey by Aguilar et al. showed that about 60% of the protein-ligand complexes indicated

that at least one titratable residue of the protein assumed a different protonation state between bound and unbound states.[13] Although important to many biological processes, pH-dependence in biomacromolecule simulations remains a nonstandard tool that awaits both wider acceptance, and finer tuning of its models.

Typical molecular dynamics (MD) simulations fix all amino acid protonation states to those of isolated residues in a neutral pH environment. While this pH-insensitive approach is sufficient to fold some proteins and observe their conformational equilibria,[14] it arguably fails to capture phenomena dependent on local ionization effects of side-chains or perturbations to a residue's pK_a.[15,16] This failure is particularly problematic for histidine residues, in that they have two hydrogens that titrate with near-neutral pK_a values. This ionizability indicates that at biologically-relevant pH environments histidine's protonation state and tautomeric configuration are often unclear.[17] In recent decades a series of models of varying complexity and accuracy promise to bring accurate pH responsiveness to MD simulations.

Protonation-state modeling of amino acids in MD simulations is based on setting up a pH-sensitive extended Hamiltonian that modifies the forcefield parameters and structure of a given molecule. This began by discretely-titrating protons, and progressing a simulation using instantaneous switches between protonated and unprotonated states. Mertz and Pettitt used an open system Hamiltonian to model the titration of acetic acid,[18] and Sham et al. applied a linear response approximation through the protein-dipoles Langevin-dipoles model to calculate lysozyme residue pK_a values.[19] Additional work has been done where Monte Carlo (MC) sampling guides the protonation state of an otherwise classical MD simulation. Baptista et al. used explicitly-represented solvent molecules with an implicit solvent Poisson-Boltzmann (PB) function to determine protonation states.[20,21] Meanwhile, Mongan et al. utilized generalized Born (GB) implicit solvation both for the solute, and to add a solvation

free energy component protonation function.[22] While all these discrete models can predict pK_a values for individual amino acids to within one pK unit, they are computationally expensive. Whether the expense stems from the need to relax numerical instabilities caused by instantaneous protonation / deprotonation events, or from the MC algorithms' ability to titrate only one hydrogen at a time, such methods may require an unreasonable amount of time to study large systems with many titratable groups.

One possible solution to the inefficiencies inherent with discrete titration methods is to use continuous titration of hydrogen atoms. Lee et al. developed one such method called constant pH molecular dynamics (CPHMD), which uses λ-dynamics coupled to transitions between protonation states.[23,24] This method uses the Generalized Born implicit water model with a Simple sWitching function (GBSW) model,[25] or the related Generalized Born with Molecular Volume (GBMV) model,[24] to efficiently couple the protonation state to the solvation free energy of the molecule. The following year, Khandogin and Brooks introduced proton tautomerism capabilities to the method, which allows multi-site titrating residues, such as histidines, to be modeled accurately.[26] Since the method is continuous, there are no instantaneous protonation/deprotonation events, and multiple residues can titrate simultaneously. Additionally, such continuous titration methods allow for the efficient coupling of protonation states among neighboring residues. The result is a pH simulation method that can calculate pK_a values of protein structures to within 1 pK unit,[16] and can resolve the dominant folding pathway of the pH-sensitive HdeA homodimers.[15]

CPHMD's efficiency, however, is bound by the rate-limited component of the calculation: the GBSW solvent model. As such, when running on a single-core central processing unit (CPU), CPHMD achieves 1 nanosecond (ns) of simulation time per day when simulating a solute system of about 1,000 atoms. Since typical uses of CPHMD, such

as predicting pK_a shifts of protein residues, may require many nanoseconds of simulation time,[16] even smaller proteins, such as lysozymes, may require about a week to converge on useful results. Larger systems, such as asymmetric viral capsid subunits with tens of thousands of atoms, may require unreasonably long simulation times if captured in full atomic detail. Fortunately, the GBSW solvent model has recently been rewritten to function on new, parallel graphics processing unit (GPU) hardware, and is now between 1 and 2 orders of magnitude faster than its CPU counterpart.[16] By incorporating the CPHMD model into the GPU-GBSW algorithm, there holds the promise of speeding up pH simulations substantially.

This study represents an increment in the ongoing adaptation of efficient and useful algorithms onto parallel-processing GPUs. Such chipsets can contain thousands of processing cores, and are able to process C-like languages such as Open Computing Language (OpenCL) and Compute Unified Device Architecture (CUDA). This combination of features has opened up a new frontier of parallel processing where expensive computer clusters can be replaced with single, affordable graphics cards. Simulation packages such as CHARMM,[27] AMBER,[28] OpenMM,[29] GROMACS,[30] and NAMD[31] all offer GPU-accelerated options for many types of studies, and most of those options receive speed increases of greater than an order of magnitude over their CPU counterparts.

Due to OpenMM's effectiveness in harnessing the capabilities of GPUs with a wide variety of hardware, a CHARMM-OpenMM interface was developed to combine the strengths of both simulation packages.[27,29] CHARMM's robust algorithms can be used to design and parameterize a simulation, and OpenMM's efficient programming can be used to propagate dynamics.[27,29] Now with the recent incorporation of the GBSW solvent model into the CHARMM-OpenMM interface, many of CHARMM's algorithms parameterized for use

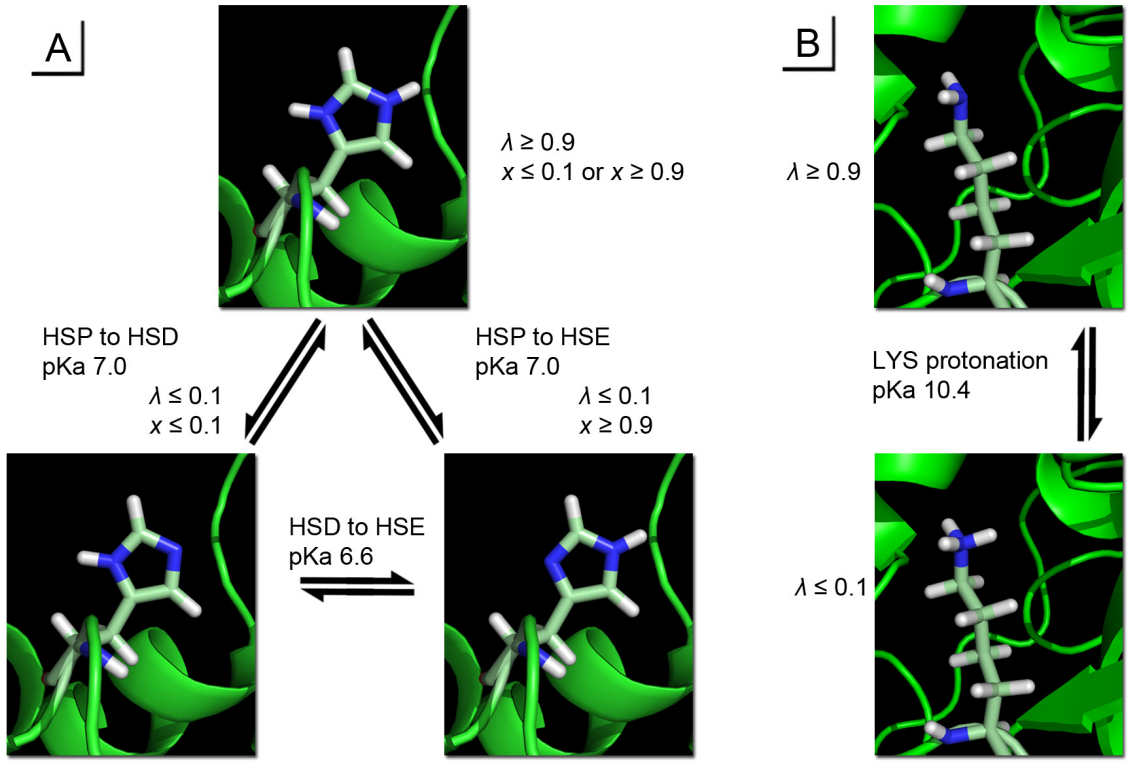

Figure 7.1 Shown are cartoons of the protonated and unprotonated states of A) histidine and B) lysine. Also noted are the reference pK_a values of each transition, as well as the λ values at each state.

with GBSW, such as CPHMD, can be adapted for parallel processing on GPUs as well. In this study we take advantage of the recent incorporation of GBSW onto GPUs, and discuss the adaptation of CPHMD onto this new parallel architecture. First we explain the underlying theory behind λ-dynamics, and how the λ coordinate is propagated. Then we delve into how it was originally implemented for CHARMM, and examine fitting CPHMD into the GBSW algorithm. Here we discuss the algorithmic improvements, and show how much of the forces on λ are calculated alongside the free energy of solvation. Finally we review the speed improvements achieved by the new algorithm, and future directions for pH simulations.

7.2 The underlying energy function for single-site titration

For clarity in following discussions, we present the underlying theory of CPHMD. We start by setting up the framework for a single residue with one titrating hydrogen. The rudimentary picture of titration events is an equilibrium association/disassociation reaction of a model compound $A_{(aq)}$ in aqueous solution from a titrating hydrogen.

$$AH_{(aq)} \leftrightarrow A^{-}_{(aq)} + H^{+}_{(aq)} \qquad \text{(eq. 1)}$$

Here, the protonation free energy is defined by

$$\Delta G^{\exp}\left(model\right) = -k_B T \ ln\ 10\left(pK_a^{\exp} - pH\right) \qquad \text{(eq. 2)}$$

k_B is Boltzmann's constant, and T is the temperature. We can approximate the above equations through classical simulations by interpreting the protonation interaction as a change in free energies:

$$\Delta G^{\exp}\left(protein\right) - \Delta G^{pH}\left(model\right) = \Delta G^{classical}\left(protein\right) - \Delta G^{model}\left(model\right) \qquad \text{(eq. 3)}$$

This relationship then leads to the estimate of experimental free energy of protonation for a single titrating site:

$$\Delta G^{\exp}\left(protein\right) = \Delta G^{classical}\left(protein\right) - \Delta G^{classical}\left(model\right) + \Delta G^{pH}\left(model\right) \qquad \text{(eq. 4)}$$

From this perspective, we infer that titratable groups have an intrinsic free energy of protonation that is perturbed by the protein environment mainly through nonbonded interactions. We model this perturbation by extending the system's Hamiltonian with a non-geometric dimension of λ. As mentioned in the introduction, the CPHMD model uses a series of λ coordinates are applied to a system where each λ value tracks the progress of protonation-deprotonation events. For a particular residue i, these coordinates are generated from

$$\lambda_i = \sin^2(\theta_i) \quad \text{(eq. 5)}$$

where i is the residue being titrated. In this form the θ variable is bound to all real numbers, and λ is bound to the continuous range $0 \le \lambda_i \le 1$. The sine-squared function then favors λ values near the boundary protonated (1) and unprotonated (0) states. Because λ is only physically relevant as it nears these boundary states, we impose cutoffs on interpreting λ. In CPHMD an unprotonated state is $\lambda_i \le 0.1$, a protonated state is $\lambda_i \ge 0.9$, and a mixed state is $0.1 < \lambda_i < 0.9$. Figure 7.1 illustrates the protonation states and their corresponding λ values. Potentials and their derivative forces on λ are then interpreted as potentials and forces on θ.

The potential energy that governs protonation states contains five λ-dependent components. We start with the pH dependence of the deprotonation free energy as follows from ΔG^{pH}. This potential is experimentally-verifiable, and connects λ to the pK_a of a

residue:

$$U^{pH}(\lambda_i) = \lambda_i (pK_a(i) - pH)(k_B T \ ln\ 10) \qquad \text{(eq. 6)}$$

Here $pK_a(i)$ is the pK_a of titrating group i. Next we have the potential of mean force (PMF) along the λ coordinate from ΔG^{model}. This term corresponds to the negative of free energy needed to deprotonate a model residue:

$$U^{model}(\lambda_i) = A_i(\lambda_i - B_i)^2 \qquad \text{(eq. 7)}$$

Equation 7 is a quadratic fit to the thermodynamic work potential of deprotonating a model compound, and it splits the protonation state into two low-energy wells that represent the protonation states. Then a barrier potential is added that disfavors mixed states of λ:

$$U^{barrier}(\lambda_i) = 4\beta_i(\lambda_i - 1/2)^2 \qquad \text{(eq. 8)}$$

The barrier scaling parameter β_i is an empirical coefficient designed to tune the propensity for a λ value to remain in either protonated or unprotonated states, and in the current iteration of CPHMD assumes a value of 2.5 or 1.75 kcal/mol. Finally, we arrive at the two charge-dependent potentials: the Coulombic and generalized Born. The classical Coulombic potential is

$$U^{elec}(\lambda_i) = \sum_{a,i}\sum_{b} K^{elec}\frac{q_{a,i}(\lambda_i)\ q_b}{r_{ab}} \qquad \text{(eq. 9)}$$

Here K^{elec} is Coulomb's constant, q_a and q_b are the partial charges of atoms a and b respectively, and r_{ab} is the distance between those atoms. Note that this potential for residue i includes the interactions between all atoms a in residue i to all other atoms in the system. Meanwhile, $q_a(\lambda_i)$ is a λ-dependent charge of atoms a, which follows the form

$$q_{a,i}(\lambda_i) = \lambda_i q_{a,i}^{unprot} + (1-\lambda_i) q_{a,i}^{prot} \qquad \text{eq. 10}$$

where charges on titrating atom a can be in protonated ($q_{a,i}^{prot}$) and unprotonated ($q_{a,i}^{unprot}$) states. Here we note that in an effective charge model of pH, titrating residues are allowed to interact. As such, any atom b from a titrating residue j interacting with residue i has its own $q_{b,j}^{prot}$ and $q_{b,j}^{unprot}$. Thus the partial charge q_b follows one of two possibilities:

$$q_b = \begin{cases} q_b & non-titrating \\ \lambda_j q_{b,j}^{unprot} + (1-\lambda_j) q_{a,j}^{prot} & titrating \end{cases} \qquad \text{eq. 11}$$

That is if atom b lies in a non-titrating residue, that atom's partial charge is simply from the standard partial charge from that residue's forcefield. If atom b lies in a titrating residue j

and its charge is affected by the protonation state of j, then its partial charge is derived from the same λ-dependent relationship from Equation 10. Since atoms near a titrating site can have their partial charges affected by titration states, many more than the titrating hydrogens can possess a λ-dependent charge state. We also note that at times $j = i$. The final λ-dependent potential is that from the GB solvent model as expressed in the Still equation:[32]

$$U^{GB}(\lambda_i) = \sum_{a,i} \sum_{b} \tau \frac{q_{a,i}(\lambda_i)\, q_b}{f_{ab}^{GB}} \qquad \text{eq. 12}$$

where

$$f_{ab}^{GB} = \left[r_{ab}^2 + R_a^{Born} R_b^{Born} \exp\left(-r_{ab}^2 / \left(4 R_a^{Born} R_b^{Born} \right) \right) \right]^{1/2} \qquad \text{eq. 13}$$

Here, $q_a(\lambda_i)$ and q_b follow the same form as in eq. 10 and 11 respectively, r_{ab} is the distance between atoms a and b, τ is the factor that scales the Born energy by the difference in dielectric values at the dielectric boundary, and the values R_a^{Born} and R_b^{Born} represent the Born radii of atoms a and b respectively. The Born radii are the effective distance between an atom and the solute-solvent dielectric boundary, and they are calculated through volumetric integration following the GBSW implicit solvent model.[25]

If we pull together the complete potential for a titrating residue i from equations 6 through 13, then we arrive at the form

$$U_i^{total}(\lambda_i) = U_i^{pH}(\lambda_i) + U_i^{model}(\lambda_i) + U_i^{barrier}(\lambda_i) + U_i^{elec}(\lambda_i) + U_i^{GB}(\lambda_i) + U_i^{VDW} + U_i^{internal} \quad \text{eq. 14}$$

The so-called "internal energy" term ($U^{internal}$) corresponds to the bond, angle, and torsional energy terms of a classical energy forcefield. In this model, the titration state is dynamically independent of this potential. Although several models of CPHMD include a λ-dependent van der Waals term (U^{VDW}),[26,33,34] during this study it was found that at most it contributes to less than 0.05 kcal/mol of a given residue's force on λ, while it nearly doubles the calculation time of CPHMD. With the observation that there is an average rotational variance of 4.8 kcal/mol Å in the force on λ due to the GBSW solvent model's integration algorithm, the force contribution from the potential U^{VDW} was considered negligible. Additionally, the default random force of Langevin dynamics has a standard deviation of 14.5 kcal/mol at 298K, which further marginalizes vdW forces on λ. Thus in the interest of speeding up the original algorithm, vdW force calculations were ignored in this implementation of CPHMD.

Although we now have the proper setup for addressing residues with a single titration site, such as in lysine, we need to address how CPHMD handles tautomerization in residues, such as in aspartic acid and histidine.

7.3 Proton Tautomerism

Similar to how one λ variable is used to track the progress of titration states of a residue, Khandogin and Brooks incorporated tautomeric behavior into CPHMD by providing residues with a second λ variable, called x, to track the progress of tautomeric states.[26] This arrangement is illustrated in Figure 7.1a with histidine. Just as in λ dynamics for titration states, tautomeric states are linearly interpolated between using the x variable. What results are the potentials become bivariate to both λ and x, and each tautomeric residue has four charge states: tautomer A in protonated and unprotonated states, and tautomer B in protonated and unprotonated states. What we shall see later is that residues can have equivalent states in this setup. Histidine's protonated state, for example, is a residue saturated with protons. As such tautomers A and B of the protonated state are equivalent. We now review the influence of including two λ parameters for a tautomeric titrating residue.

The pH dependent potential becomes

$$U^{pH}(\lambda_i, x_i) = \lambda_i\left[x_i\left(pK_a^A(i) - pH\right) + (1 - x_i)\left(pK_a^B(i) - pH\right)\right](k_B T\ ln\ 10) \qquad \text{eq. 15}$$

where the pK_a values of tautomers A and B are pK_a^A and pK_a^B respectively. While these pK_a values for aspartic acid and glutamic acid are equivalent, in residues with asymmetric titrating sites such as histidine they are not. The PMF for protonation becomes a bivariate polynomial from Equation 7, which expands into the general form

$$U^{model}(\lambda_i, x_i) = a_0\lambda_i^2 x_i^2 + a_1\lambda_i^2 x_i + a_2\lambda_i x_i^2 + a_3\lambda_i x_i + a_4\lambda_i^2 + a_5 x_i^2 + a_6\lambda_i + a_7 x_i + a_8 \quad \text{eq. 16}$$

The barrier potential is simply a summation of terms that disfavor the mixed states of both λ and x, and follows the form

$$U^{barrier}(\lambda_i, x_i) = 4\beta_i^{\lambda}(\lambda_i - 1/2)^2 + 4\beta_i^{x}(x_i - 1/2)^2 \quad \text{eq. 17}$$

Note that there are two barrier scaling parameters β_i^{λ} and β_i^{x} for λ and x. Although different biases for tautomeric and protonation transitions are possible in this equation, in the discussed CPHMD model they are identical for all titrating residues.

The charge-dependent potentials in Equations 9 and 12 are only modified in that charges for atoms can now be dependent on the new x coordinate. The Coulombic and generalized Born potentials then follow the forms

$$U^{elec}(\lambda_i, x_i) = \sum_{a,i}\sum_{b} K^{elec}\frac{q_{a,i}(\lambda_i, x_i)\, q_b}{r_{ab}} \quad \text{eq. 18}$$

and

$$U^{GB}(\lambda_i, x_i) = \sum_{a,i}\sum_{b} \tau\frac{q_{a,i}(\lambda_i, x_i)\, q_b}{f_{ab}^{GB}} \quad \text{eq. 19}$$

respectively. The bivariate charge $q_{a,i}(\lambda_i, x_i)$ then follows the form

$$q_{a,i}(\lambda_i, x_i) = \lambda_i\left[x_i\, q_{a,i}^{A,\,unprot} + (1-x_i) q_{a,i}^{B,\,unprot}\right] + (1-\lambda_i)\left[x_i\, q_{a,i}^{A,\,prot} + (1-x_i) q_{a,i}^{B,\,prot}\right] \quad \text{eq. 20}$$

Where charges on titrating atom a are derived from the protonated and unprotonated variants of both A and B tautomers, $q_{a,i}^{A,\,prot}$, $q_{a,i}^{A,\,unprot}$, $q_{a,i}^{B,\,prot}$, and $q_{a,i}^{B,\,unprot}$. Similarly, the charge on atom b emerges as

$$q_b = \begin{cases} q_b & non-titrating \\ \lambda_j\left[x_j\, q_{b,j}^{A,\,unprot} + (1-x_j) q_{b,j}^{B,\,unprot}\right] + (1-\lambda_j)\left[x_j\, q_{b,j}^{A,\,prot} + (1-x_j) q_{b,j}^{B,\,prot}\right] & titrating \end{cases} \quad \text{eq. 21}$$

We now arrive at a general-purpose setup for evaluating the underlying potential for continuous transitions among various charge states of a particular residue. Deriving the forces with respect to λ and x, while important, serves little purpose for illuminating the topics explored in the remainder of this study. With the framework above, we now can discuss the construction of the original algorithm, and the changes made to refactor it for efficient parallel processing on GPUs.

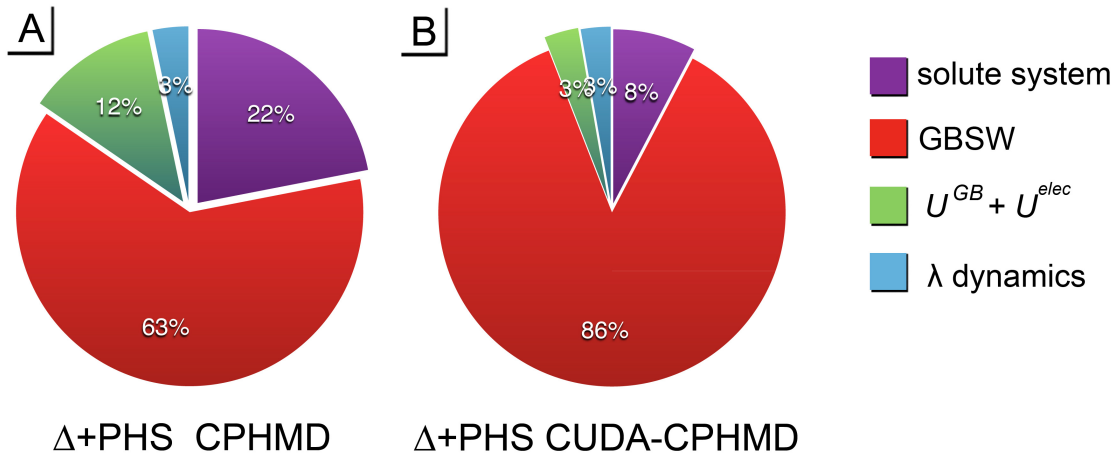

Figure 7.2 Shown are the approximate distributions of CPU time spent on running simulations components of Δ+PHS staphylococcal nuclease molecule. This protein contains 2132 atoms and 37 titrating residues. A) run on using the original algorithm using a single processing core in CHARMM. B) run using the newly refactored CUDA-CPHMD algorithm.

7.4 Refactoring CPHMD

The original CPHMD model was built with mathematical precision and function portability in mind. It is a stand-alone function that can be applied to both implicit and explicit solvent systems, and except for atom coordinate and Born radii updates, it receives no input from other functions during a simulation. During the course of a timestep, each titrating coordinate λ_i is scanned to identify the residue type (such as whether the residue has one or two titrating hydrogens), and then an appropriate functional is applied to calculate its pH (eq. 6 and 15), model (eq. 7 and 16), and barrier (eq. 8 and 18) potentials. Next, neighboring atom-atom interactions are scanned for whether one or both atoms reside in titrating groups. If a titrating atom-atom pair is found, then contributions to the electrostatic (eq. 9 and 18) and GB (eq. 2 and 19) potentials are integrated. Neighboring atom-atom pairs are then scanned again to calculate the VDW potential (ignored in this new

iteration of CPHMD). Finally, the force on θ is calculated, and λ via θ is advanced a timestep using Langevin dynamics.[35] In this setup there are several opportunities presented to us for improving the algorithm both in the efficiency of its execution in parallel, and by weaving portions of the calculation into existent functions elsewhere in the simulation.

We first note that the majority of clock cycles used for calculating λ dynamics are spent on neighboring atom-atom interactions when accumulating the electrostatic and GB potentials. While the calculations required for each atom pair are computationally cheap, the large number of interatomic interactions in a protein containing thousands of atoms can make this multitude of cheap calculations altogether expensive. As show in Figure 7.2a, about 12% of a 2000-atom simulation is spent only on this calculation.

Both CPHMD and the GBSW solvent model require calculating the Still equation (eq. 12 and 13) to address part of the neighboring atom potential, so a significant speed improvement can be made by placing all of CPHMD's atom-atom processes inside the neighboring atom process of the GBSW solvent model. This way, as GBSW produces the solute molecule's electrostatic solvation free energy and its derivative force on atoms, CPHMD processes neighboring atom potentials on λ simultaneously. Thus the large number of redundant atom-atom distance calculations can be reduced significantly during a simulation. This setup gains additional speedup through GBSW by using OpenMM's efficient parallel possessing of neighboring-atom interactions. As shown in 7.2, by combining the CPHMD and GBSW algorithms we see that pH modeling with CPHMD accounts for a much smaller fraction of the overall simulation time.

Due to the nature of parallel processing, bottlenecks are often created from the longest portions of non-parallel code. While a single-core process can be sped up dramatically by creating a case-by-case set of calculations, navigating through the additional

overhead to make the situation-specific decision can slow parallel processes down. Regarding the equations described earlier, a titrating residue with one tautomer requires fewer calculations than a titrating residue with two. As we place each residue's force calculations in parallel processes, however, the speed of the code is improved by regarding all titrating residues as possessing two tautomeric states. In this new implementation of CPHMD, single-titration residues, such as lysine, are given a meaningless x coordinates. Lysine then uses the longer barrier potential from eq. 16, where the x-coupled coefficients a_0, a_1, a_2, a_3, and a_5 are set to a value of 0.0. Without the overhead for residue identification, the longest calculation required, that is calculating the force on θ for a residue with two tautomeric states, is shortened. What results is a speed improvement when calculating all components of the total potential on λ coordinates. As shown in Figure 7.2b, using the parallel CUDA-CPHMD algorithm for a small system impacts the processing time by approximately 6%, as opposed to 15% for the original algorithm.

7.5 Benchmarking CUDA-CPHMD

We finally reach an efficient setup where using the CPHMD model results in little slowdown of the overall simulation time. We chose several systems to benchmark the new algorithm, and explore the speed benefits it offers. We chose the naja atra snake cardiotoxin (PDB: 1CVO),[36] the Δ+PHS hyperstable variant of staphylococcal nuclease (PDB: 3BDC),[37] and the asymmetric subunit of the bacteriophage HK97 head capsule (PDB: 2FT1).[38] This trio provided a range of system sizes and residue configurations. To add additional statistics, the 7 proteins of the HK97 head capsule were assembled into 6 additional subsystems, all of which appear in Figure 7.3 to show for a range of system sizes the speed dependence on

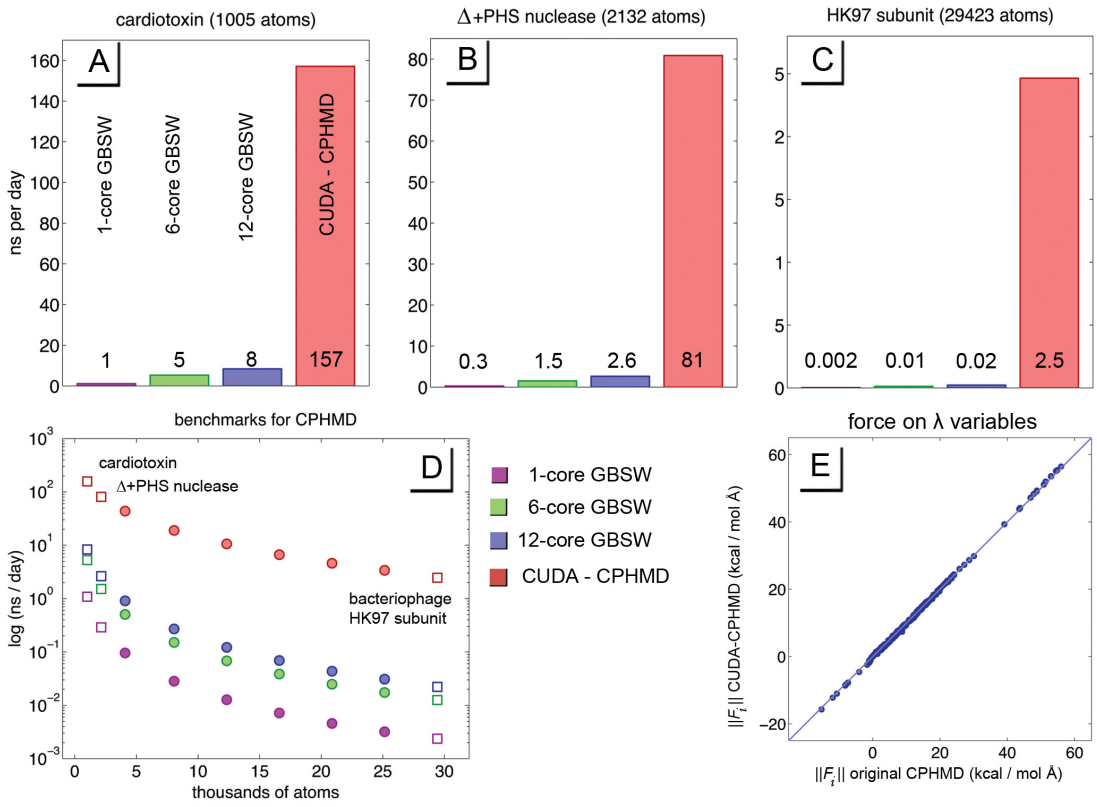

Figure 7.3 Shown are the benchmarks for the new CUDA-CPHMD algorithm. The individual systems tested were A) the naja atra snake cardiotoxin (PDB: 1CVO); B) the Δ+PHS hyperstable variant of staphylococcal nuclease (PDB: 3BDC); and C) the asymmetric subunit of the bacteriophage HK97 head capsule (PDB: 2FT1). As shown, the new algorithm is substantially faster than the original CPU algorithm by up to 3 orders of magnitude. In D) the same benchmarks from earlier are shown (squares) alongside subsystems from the 7 proteins of the bacteriophage subunit (circles). Notice that the CUDA algorithm scales more linearly with system size than its CPU-based counterpart. E) compares the force on λ as calculated on all 595 λ coordinates from both CPHMD algorithms. There is less than a 0.23 (kcal/mol Å) AUE between the two algorithms.

system size. All simulations were using the CHARMM22 forcefield[39,40] using the Langevin integrator with a timestep of 2 femtoseconds. These were NT simulations at 298K in unbounded volumes using the CUDA-GBSW solvent model, and CUDA-CPHMD to model titration states and advance λ coordinates. Atomic radii for the GBSW solvent model were provided through work by Chen et al.[41] We found speed improvements of between 1 and 3 orders of magnitude in the CUDA-CPHMD algorithm over its CPU counterpart.

As we combine the improved efficiency and parallel execution of both GBSW and CPHMD (shown in Figure 7.3a to 7.3d), substantial speed gains are found in this new version of pH modeling over its predecessor. For smaller 1,000-atom systems, we see a speed improvement of over 20-fold when comparing a 12-threaded CPHMD simulation to the new CUDA-CPHMD, and an improvement of over 150-fold when compared to the single-core algorithm (shown in Figure 7.3a). For larger 29,000 atom systems, we see speed improvement of over 1,000-fold (shown in Figure 7.3c). Since the neighboring-atom component doesn't scale linearly with system size, larger systems experience a greater calculation penalty than smaller ones. Fortunately, simple changes such as using nonbonded cutoffs can mitigate such problems. For instance, a nonbonded cutoff of 14 Å sped up the large viral capsid simulation to 6.7 ns/day (a 270% speed increase).

7.6 Accuracy of the CUDA-CPHMD algorithm

Speed gains in implementing CPHMD are an important goal both for increasing the algorithm's applicability to a wider range of system sizes, and for its ability to converge on useful results more rapidly. It's accuracy, however, must not be compromised as we reconfigure the execution of the algorithm. In Figure 7.3e we show that there is little difference between the original CPHMD and CUDA-CPHMD algorithms when calculating the force on λ. We maintain an average unsigned error (AUE) of less than 0.23 kcal / mol Å in this force, which is much less than the AUE of 4.8 kcal/mol Å caused by the rotational variance from the GBSW solvent model. We also note that 99.9% of the AUE between the two CPHMD methods is from the slight differences in Born radii calculated from the

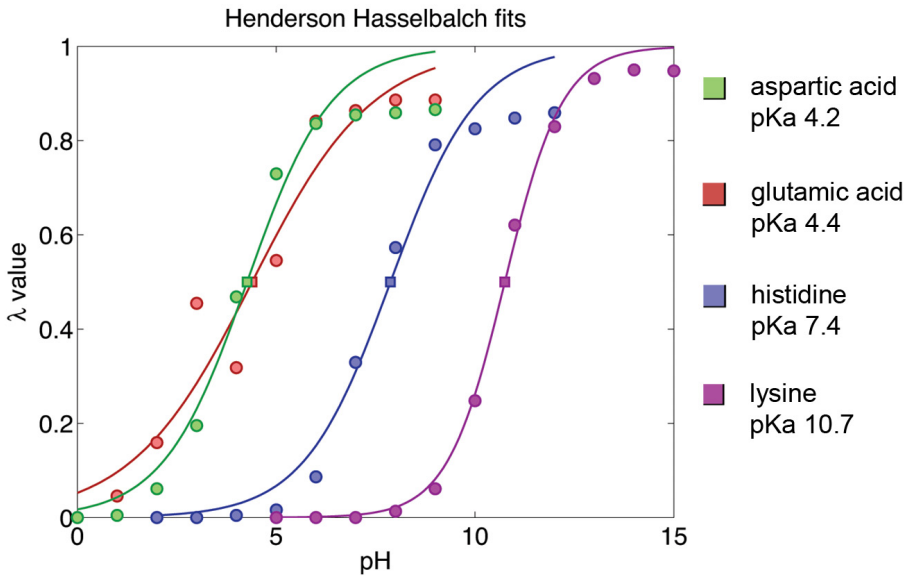

Figure 7.4 Above are the pK_a calculations for 4 single residues: aspartic acid, glutamic acid, histidine, and lysine. The protonation state (dots) were calculated from a fraction of λ values in pure unprotonated and protonated states. The point of inflection (boxes) of Henderson HasselBalch equation fits (lines) indicates the calculated pK_a values. Even without optimizing for efficiency, convergence of data, or simulation parameters, we find the calculated pKa values match those from the forcefield to within 0.5 pK units.

original and CUDA implementations of GBSW. Thus, we conclude that CUDA-CPHMD accurately reproduces the original algorithm's force on λ.

While CUDA-CPHMD may be able to produce the force on λ coordinates, we ran an additional test to see whether or not residue protonation states are also reproduced. Due to each residue's pH-dependent biasing potential, a single residue alone in solution presumably should find an optimal protonation state depending on the environmental pH. At pH environments below a residue's pK_a the residue should favor a protonated state $(\lambda_i \le 0.1)$, and conversely a residue exposed to a pH above its pK_a should favor an unprotonated state $(\lambda_i \ge 0.9)$. By calculating the fraction of protonated to unprotonated states of residues at various pH values and fitting the results to the Henderson-Hallelbalch

equation of states, we expect the point of inflection to reproduce the pK_a of that residue.

We ran simulations of aspartic acid, glutamic acid, histidine, and lysine to calculate their protonation states, as shown in Figure 7.4. These residues were simulated using the same setup from the benchmarking section as NT simulations in an unbound volume, and CUDA-CPHMD was used both to model titration states and advance λ coordinates. The backbone atom ends were capped with the ACE and CT2 hydrogens. Each dot in Figure 7.4 represents 200 ps of simulation time, and the residues ran at an average speed of 470 ns/day.

We find that without optimizing the simulations for speed, accuracy, or convergence of protonation states, that the pK_a values could be captured to within 0.5 pK units. Interestingly, all states reported a small, systematic overestimation of the pK_a, and the exact source of this discrepancy remains unclear. The CUDA-GBSW solvent model overestimates solvation energy by an average of approximately 0.16 kcal/mol. However, this overestimation of energy should bias deprotonation events to occur slightly more often, and thus lower the calculated pK_a. What is clear from these data, though, is that like its predecessor, the CUDA-CPHMD algorithm models the pH dependence of titration well.

7.7 Discussion and Future Directions

In this study we present a significantly faster version of the CPHMD algorithm adapted for parallel processing in the CHARMM-OpenMM interface. While algorithmically the new CUDA-CPHMD algorithm represents little change over its earlier version, the speed improvements are so great that previously-unreasonable simulations are now straightforward to perform. For instance, what may have been a year-long simulation of the HK97 head capsule can now be performed in about 160 minutes. With this newfound speed

is an opportunity to fine-tune the CPHMD titration model for a variety of protein systems, and to explore the impact of pH environments on side-chain dynamics both at the microsecond timescale and with all-atom detail.

Similarly to GBSW, the CPHMD model carries with it over a decade of research and parameterization. One model of particular interest is pH replica exchange (REX),[42] which has been shown to predict pKa values of protein structures within single nanoseconds of simulation time. Coupled with the improved speed of CPHMD, adapting REX would enable a useful and rapid method for characterizing the chemical environment of protein interiors.

7.8 References

1. J. E. Nielsen, and J. A. McCammon, "Calculating Pka Values in Enzyme Active Sites," Protein Science : A Publication of the Protein Society **12**(9), 1894-901, (2003).

2. D. Sali, M. Bycroft, and A. R. Fersht, "Stabilization of Protein Structure by Interaction of [Alpha]-Helix Dipole with a Charged Side Chain," Nature **335**(6192), 740-43, (1988).

3. B. Cannon, D. Isom, A. Robinson, J. Seedorff, and B. Garcia-Moreno, "Molecular Determinants of Pka Values of Internal Asp Residues," Biophys. J., 403A-03A, (2007).

4. G. Rabbani, E. Ahmad, N. Zaidi, S. Fatima, and R. Khan, "Ph-Induced Molten Globule State of Rhizopus Niveus Lipase Is More Resistant against Thermal and Chemical Denaturation Than Its Native State," Cell Biochem Biophys **62**(3), 487-99, (2012).

5. G. R. Wagner, and R. M. Payne, "Widespread and Enzyme-Independent Nϵ-Acetylation and Nϵ-Succinylation of Proteins in the Chemical Conditions of the Mitochondrial Matrix," J. Biol. Chem. **288**(40), 29036-45, (2013).

6. L. A. Baker, I. N. Watt, M. J. Runswick, J. E. Walker, and J. L. Rubinstein, "Arrangement of Subunits in Intact Mammalian Mitochondrial Atp Synthase Determined by Cryo-Em," Proc. Natl. Acad. Sci. **109**(29), 11675-80, (2012).

7. B. D. Cain, and R. D. Simoni, "Impaired Proton Conductivity Resulting from Mutations in the a Subunit of F1f0 Atpase in Escherichia Coli," J. Biol. Chem. **261**(22), 10043-50, (1986).

8. V. K. Rastogi, and M. E. Girvin, "Structural Changes Linked to Proton Translocation by Subunit C of the Atp Synthase," Nature **402**(6759), 263-68, (1999).

9. A. B. Clippingdale, J. D. Wade, and C. J. Barrow, "The Amyloid-Β Peptide and Its Role in Alzheimer's Disease," Journal of Peptide Science **7**(5), 227-49, (2001).

10. C. M. Dobson, "Protein Folding and Misfolding," Nature **426**(6968), 884-90, (2003).

11. L. G. Cuello, D. M. Cortes, V. Jogini, A. Somporpisut, and E. Perozo, "A Molecular Mechanism for Proton-Dependent Gating in Kcsa," FEBS letters **584**(6), 1126-32, (2010).

12. E. E. Howell, J. E. Villafranca, M. S. Warren, S. J. Oatley, and J. Kraut, "Functional-Role of Aspartic Acid-27 in Dihydrofolate-Reductase Revealed by Mutagenesis," Science **231**(4742), 1123-28, (1986).

13. B. Aguilar, R. Anandakrishnan, J. Z. Ruscio, and A. V. Onufriev, "Statistics and Physical Origins of Pk and Ionization State Changes Upon Protein-Ligand Binding," Biophys. J. **98**(5), 872-80, (2010).

14. K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, "How Fast-Folding Proteins Fold," Science **334**(6055), 517-20, (2011).

15. L. S. Ahlstrom, S. M. Law, A. Dickson, and C. L. Brooks, III, "Multiscale Modeling of a Conditionally Disordered Ph-Sensing Chaperone," J. Mol. Biol. **427**(8), 1670-80, (2015).

16. E. J. Arthur, J. D. Yesselman, and C. L. Brooks, III, "Predicting Extreme Pka Shifts in Staphylococcal Nuclease Mutants with Constant Ph Molecular Dynamics," Proteins: Struct., Funct., Bioinf. **79**(12), 3276-86, (2011).

17. D. Bashford, D. A. Case, C. Dalvit, L. Tennant, and P. E. Wright, "Electrostatic Calculations of Side-Chain Pka Values in Myoglobin and Comparison with Nmr Data for Histidines," Biochemistry **32**(31), 8045-56, (1993).

18. J. E. Mertz, and B. M. Pettitt, "Molecular-Dynamics at a Constant Ph," International Journal of Supercomputer Applications and High Performance Computing **8**(1), 47-53, (1994).

19. Y. Y. Sham, Z. T. Chu, and A. Warshel, "Consistent Calculations of Pk(a)'S of Ionizable Residues in Proteins: Semi-Microscopic and Microscopic Approaches," J. Phys. Chem. B **101**(22), 4458-72, (1997).

20. A. M. Baptista, P. J. Martel, and S. B. Petersen, "Simulation of Protein Conformational Freedom as a Function of Ph: Constant-Ph Molecular Dynamics Using Implicit Titration," Proteins: Struct., Funct., Genet. **27**(4), 523-44, (1997).

21. A. M. Baptista, V. H. Teixeira, and C. M. Soares, "Constant-Ph Molecular Dynamics Using Stochastic Titration," J. Chem. Phys. **117**(9), 4184-200, (2002).

22. J. T. Mongan, D. A. Case, and J. A. McCammon, "Constant Ph Molecular Dynamics in Generalized Born Implicit Solvent," Abstr. Pap. Am. Chem. Soc. **229**(Part 1), U768, (2005).

23. X. J. Kong, and C. L. Brooks, III, "Λ-Dynamics: A New Approach to Free Energy Calculations," J. Chem. Phys. **105**(6), 2414-23, (1996).

24. M. S. Lee, F. R. Salsbury, and C. L. Brooks, III, "Novel Generalized Born Methods," J. Chem. Phys. **116**(24), 10606-14, (2002).

25. W. Im, M. S. Lee, and C. L. Brooks, III, "Generalized Born Model with a Simple Smoothing Function," J. Comput. Chem. **24**(14), 1691-702, (2003).

26. J. Khandogin, and C. L. Brooks, III, "Constant Ph Molecular Dynamics with Proton Tautomerism," Biophys. J. **89**(1), 141-57, (2005).

27. B. R. Brooks, C. L. Brooks, III, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus, "Charmm: The Biomolecular Simulation Program," J. Comput. Chem. **30**(10), 1545-614, (2009).

28. D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods, "The Amber Biomolecular Simulation Programs," J. Comput. Chem. **26**(16), 1668-88, (2005).

29. P. Eastman, M. S. Friedrichs, J. D. Chodera, R. J. Radmer, C. M. Bruns, J. P. Ku, K. A. Beauchamp, T. J. Lane, L.-P. Wang, D. Shukla, T. Tye, M. Houston, T. Stich, C. Klein, M. R. Shirts, and V. S. Pande, "Openmm 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation," J. Chem. Theory Comput. **9**(1), 461-69, (2013).

30. B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, "Gromacs 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation," J. Chem. Theory Comput. **4**(3), 435-47, (2008).

31. J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, and K. Schulten, "Scalable Molecular Dynamics with Namd," J. Comput. Chem. **26**(16), 1781-802, (2005).

32. W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson, "Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics," J. Am. Chem. Soc. **112**(16), 6127-29, (1990).

33. S. Donnini, F. Tegeler, G. Groenhof, and H. Grubmüller, "Constant Ph Molecular Dynamics in Explicit Solvent with Λ-Dynamics," J. Chem. Theory Comput. **7**(6), 1962-78, (2011).

34. M. S. Lee, F. R. Salsbury, and C. L. Brooks, III, "Constant-Ph Molecular Dynamics Using Continuous Titration Coordinates," Proteins: Struct., Funct., Bioinf. **56**(4), 738-52, (2004).

35. G. E. Uhlenbeck, and L. S. Ornstein, "On the Theory of the Brownian Motion," Phys Rev **36**(5), 0823-41, (1930).

36. A. K. Singhal, K. Y. Chien, W. G. Wu, and G. S. Rule, "Solution Structure of Cardiotoxin V from Naja Naja Atra," Biochemistry **32**(31), 8036-44, (1993).

37. C. A. Castañeda, C. A. Fitch, A. Majumdar, V. Khangulov, J. L. Schlessman, and B. E. García-Moreno, "Molecular Determinants of the Pka Values of Asp and Glu Residues in Staphylococcal Nuclease," Proteins: Struct., Funct., Bioinf. **77**(3), 570-88, (2009).

38. L. Gan, J. A. Speir, J. F. Conway, G. Lander, N. Cheng, B. A. Firek, R. W. Hendrix, R. L. Duda, L. Liljas, and J. E. Johnson, "Capsid Conformational Sampling in Hk97 Maturation Visualized by X-Ray Crystallography and Cryo-Em," Structure **14**(11), 1655-65, (2006).

39. A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus, "All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins," J. Phys. Chem. B **102**(18), 3586-616, (1998).

40. A. D. MacKerell, M. Feig, and C. L. Brooks, III, "Extending the Treatment of Backbone Energetics in Protein Force Fields: Limitations of Gas-Phase Quantum Mechanics in Reproducing Protein Conformational Distributions in Molecular Dynamics Simulations," J. Comput. Chem. **25**(11), 1400-15, (2004).

41. J. H. Chen, W. P. Im, and C. L. Brooks, III, "Balancing Solvation and Intramolecular Interactions: Toward a Consistent Generalized Born Force Field," J. Am. Chem. Soc. **128**(11), 3728-36, (2006).

42. D. Sabri Dashti, Y. Meng, and A. E. Roitberg, "Ph-Replica Exchange Molecular Dynamics in Proteins Using a Discrete Protonation Method," J. Phys. Chem. B **116**(30), 8805-11, (2012).

Chapter 8

Discussion and Final Remarks

8.1 Final Thoughts on the Protein-Solvent Interface

From the work presented in the first half of this thesis (**Chapters 2-4**) we develop several observations pertaining to protein-water interactions. Most prominently we see that hydration dynamics are slowed down near protein surfaces on both nanometer and angstrom distances. Since proteins are not isotropic solutes, these dynamics are not homogeneously distributed. On the angstrom scale, we see a consistent volume of space surrounding protein surfaces where water molecules reside for long periods of time, and this volume shows little perturbance in the presence of small concentrations of salts or a trifluoroethanol cosolvent (below 10% v/v for lysozyme). We also found that when lysozyme was placed in solution with sufficient cosolvent to denature it in experiments, simulations showed that the trifluoroethanol replaced more than half of the water hot spots. Whether we interpret water at these hot spots as a structural component of lysozyme proteins, or as locations where the first hydration layer directly stabilizes the protein, we find evidence that water hot spots are

necessary for lysozyme's maintaining a native conformation.

Interestingly, predicting the location of such water hot spots shows little correlation with the relative hydrophobicity of individual residues, and seems to be a result of cooperative interactions from groups of residues. While the extent of this cooperativity is yet unresolved, it seems to be longer-reaching than individual water molecules. We find evidence of this by comparing water distributions on the two homologous lysozymes we studied; even the conserved alpha helix (residues 105-109) showed substantial variation in average water interactions. Regardless of its origin of solvent hot spots, we characterize water molecules in these hot spots through experiment and simulation to have low exchange rates with bulk water, slower orientational dynamics, and longer average lifetimes of hydrogen bonds.

As we observe longer nanometer-ranged changes in water dynamics, like other studies on the subject,[1-3] we find that there is an interstitial layer of solvent with slower dynamics between the protein surface and bulk water. This local hydration layer has faster diffusion rates and shorter hydrogen bond lifetimes than the water hot spots on protein surfaces, but significantly slower dynamics than bulk water. Interestingly, we find that the cutoff between local and bulk hydration dynamics is rather sharp. Our experiments report this cutoff to reside at distances of 30-40 Å from surfaces of lysozymes, and our simulations report the distance to be either at 10-15 Å or 20-25 Å from lysozymes surfaces depending on the relative number of proteins interacting with each other. Keeping in mind the high concentration of non-water components in the intracellular medium,[4,5] even using the lowest cutoff estimate suggests that there is little to no bulk-like water present within cellular environments. Instead, biological macromolecules are hydrated by significantly constrained water that in turn can strongly modulate the flexibility and dynamics of the biomolecules.

We hope that this new set of findings can add both to the significance of crowding

in biological systems, and to the effects of using differently-sized crowding agents. The work laid out, then, is determining which crowding molecules would be appropriate when, and characterizing the effects of such molecules for simulations. Whether the effects are present at atomically-represented cosolvent molecules, or a distance-dependent frictional dampening parameter, simulating such crowding effects may greatly enhance the accuracy of simulating water-mediated biological processes.

8.2 Expanding the Scope of Modeling Titration

In the second half of this thesis (**Chapters 5-7**) the most prominent result is simply that a long future of parallel processing has been brought to a series of accurate solvation methods. In creating the CUDA-GBSW algorithm, we mark the first implementation of accurate, volume-exclusion implicit solvation on graphics processing units. With that implementation we enable a relatively straightforward method of bringing models dependent on accurate calculations of solvation free energy to a fast parallel computing environment. Among those models is CPHMD, which allows for efficient simulations of pH-dependent titration states. The crowning success of this achievement, though, is not only are the algorithms faster, but they that were designed to receive speed gains from future advances in GPU technology until GPU chips have with millions of processing cores. Already, previously-unreasonable simulations are now trivial to perform, and a new set of time scales are accessible to interested scientists. For instance, the time requirements for simulating the Δ+PHS protein in microsecond-long trajectories used to take years, and now it can be performed in days. These timescales have enabled unprecedented capacity to test and improve the solvent models.

Although the algorithms of CUDA-GBSW and CUDA-CPHMD are robust and stable, improvements are recommended. The integration methods used in GBSW remain as the most computationally expensive portion of a simulation. A better placement of the integration points offers significant benefits to both speed and accuracy, but carries the risk of requiring a new set of atomic radii and optimal parameters. Fortunately, testing the accuracy or long-term stability of a simulation is far more accessible with the faster GPU version of the algorithm. The more important development, though, is in validating and improving pH dependence in a simulation. Although biology regularly utilizes altered titration states of amino acids both to regulate structure and function of proteins,[6-12] modeling pH remains a non-standard model. Aside from observing pH-dependent effects on conformational equilibria, what awaits the future is a deeper understanding in the significance of buried titrating residues.

8.3 References

1. A. Kuffel, and J. Zielkiewicz, "Why the Solvation Water around Proteins Is More Dense Than Bulk Water," J. Phys. Chem. B **116**(40), 12113-24, (2012).

2. J. T. King, E. J. Arthur, C. L. Brooks III, and K. J. Kubarych, "Site-Specific Hydration Dynamics of Globular Proteins and the Role of Constrained Water in Solvent Exchange with Amphiphilic Cosolvents," J. Phys. Chem. B **116**(19), 5604-11, (2012).

3. F. Sterpone, G. Stirnemann, and D. Laage, "Magnitude and Molecular Origin of Water Slowdown Next to a Protein," J. Am. Chem. Soc. **134**(9), 4116-19, (2012).

4. P. A. Srere, "Protein Crystals as a Model for Mitochondrial Matrix Proteins," Trends Biochem. Sci. **6**(1), 4-7, (1981).

5. A. B. Fulton, "How Crowded Is the Cytoplasm," Cell **30**(2), 345-47, (1982).

6. L. A. Baker, I. N. Watt, M. J. Runswick, J. E. Walker, and J. L. Rubinstein, "Arrangement of Subunits in Intact Mammalian Mitochondrial Atp Synthase Determined by Cryo-Em," Proc. Natl. Acad. Sci. **109**(29), 11675-80, (2012).

7. B. D. Cain, and R. D. Simoni, "Impaired Proton Conductivity Resulting from Mutations in the a Subunit of F1f0 Atpase in Escherichia Coli," J. Biol. Chem. **261**(22), 10043-50, (1986).

8. V. K. Rastogi, and M. E. Girvin, "Structural Changes Linked to Proton Translocation by Subunit C of the Atp Synthase," Nature **402**(6759), 263-68, (1999).

9. A. B. Clippingdale, J. D. Wade, and C. J. Barrow, "The Amyloid-B Peptide and Its Role in Alzheimer's Disease," J. Pept. Sci. **7**(5), 227-49, (2001).

10. C. M. Dobson, "Protein Folding and Misfolding," Nature **426**(6968), 884-90, (2003).

11. L. G. Cuello, D. M. Cortes, V. Jogini, A. Somporpisut, and E. Perozo, "A Molecular Mechanism for Proton-Dependent Gating in Kcsa," FEBS Lett. **584**(6), 1126-32, (2010).

12. E. E. Howell, J. E. Villafranca, M. S. Warren, S. J. Oatley, and J. Kraut, "Functional-Role of Aspartic Acid-27 in Dihydrofolate-Reductase Revealed by Mutagenesis," Science **231**(4742), 1123-28, (1986).