UNIVERSITY OF CALIFORNIA, SAN DIEGO

Implicit Solvent Method Development and Application: Fast Molecular Surfaces, Constant pH and Accelerated Dynamics, and Rational Drug Design

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Bioinformatics

by

John Mongan

Committee in charge:

Professor J. Andrew McCammon, Co-chair
Professor Gary Huber, Co-chair
Professor David A. Case
Professor Elizabeth Komives
Professor Mauricio Montal
Professor Peter Wolynes

2006

The dissertation of John Mongan is approved, and it is acceptable in quality and form for publication on microfilm:

_____

_____

_____

_____

_____
Co-chair

_____
Co-chair

University of California, San Diego

2006

For Thuy

who always tells me
whether it makes sense

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Science is always the work of many people; this is particularly true in scientific education. I am deeply indebted to the many, many people who played a role in the work that follows. Firstly, I must thank my longest-serving teachers, my parents and family, who encouraged my scientific curiousity and gave me the best imaginable start. I am grateful to my teachers and professors at the Bayside, Martin Luther King and Tamalpais High Schools and Stanford University for the excellent scientific and academic foundation they provided me.

It has been a privilege to work with my advisors Andy McCammon and Dave Case, who provided me with guidance when I needed it and freedom when I could use it. I am likewise privileged to have had portions of my work guided by Seth Cohen and Alexey Onufriev. I thank my committee members, Gary Huber, Elizabeth Komives, Mauricio Montal and Peter Wolynes for their thought, time and suggestions that have benefitted the work presented here. This work involved a great deal of collaboration and discussion with coauthors and coworkers, with particularly notable contributions from Charlie Brooks, Justin Gullingsrud, Donald Hamelberg, Wonpil Im, Jana Khandogin, Jana A. Lewis, Jens Erik Nielsen, Alex Perryman, David T. Puerta, Julie Schames, Tony Shen, Carlos Simmerling, and Jessica M. J. Swanson. Science is an expensive pursuit; I am grateful to the people and organizations that have funded my contributions: the National Institute of Health, through the Medical Scientist Training Program; the Taft family of La Jolla; and Burroughs Wellcome, through La Jolla Interfaces in Science.

Finally, I am exceedingly grateful for the presence of Thuy in my life, as she has been a central source of support and advice, scientific and otherwise, throughout my time in graduate school.

Chapter 2 is a reprint in full of material that appeared in *Limitations of atom-centered dielectric functions in implicit solvent models.* Jessica M.J. Swanson, John Mongan and J. Andrew McCammon. Journal of Physical Chemistry B, **109**(31) 14769-14772, August 2005. I was the secondary researcher and author of this work.

Chapter 3 is a preprint in full of *Generalized Born with a simple, robust molecular volume correction.* John Mongan, Carlos Simmerling, J. Andrew McCammon, David

A. Case and Alexey Onufriev. Submitted to Proteins: Structure, Function and Bioinformatics. I was the primary researcher and author of this work.

Chapter 4 contains material that appeared in *Biomolecular simulations at constant pH.* John Mongan and David A. Case. Current Opinion in Structural Biology, **18**(2), 157-63, April 2005. I was the primary author of this work.

Chapter 5 is a reprint in full of material that appeared in *Constant pH Molecular Dynamics in Generalized Born Implicit Solvent.* John Mongan, David A Case and J. Andrew McCammon. Journal of Computational Chemistry, **25**(16), 2038-48, December 2004. I was the primary author and researcher for this work.

Chapter 6 is a reprint in full of material that appeared in *Accelerated Molecular Dynamics: A promising and efficient simulation method for biomolecules.* Donald Hamelberg, John Mongan and J. Andrew McCammon. Journal of Chemical Physics **120**(24), 11919-29, June 2004. I was the secondary author and researcher for this work.

Chapter 7 is a reprint in full of material that appeared in *Interactive Essential Dynamics.* John Mongan. Journal of Computer-Aided Molecular Design, **18**(6), 433-36, June 2004. I was the sole author and researcher for this work.

Chapter 8 is a reprint in full of material that appeared in *Potent, Selective Pyrone-Based Inhibitors of Stromelysin-1.* David T. Puerta, John Mongan, Ba L. Tran, J. Andrew McCammon, and Seth M. Cohen. Journal of the American Chemical Society, **127**(41), 14148-49, October 2005. I was the secondary author and conducted the computational portion of the research for this work. Co-authors of the article conducted the synthesis and assays or supervised and directed the work.

Chapter 9 is a preprint in full of *Evaluation and binding mode prediction of thiopyrone-based inhibitors of anthrax lethal factor.* Jana A. Lewis, John Mongan, J. Andrew McCammon and Seth M. Cohen. Submitted to Angewandte Chemie International. I was co-primary author and conducted the computational portion of the research for this work. Co-authors of the article conducted the assays or supervised and directed the work.

VITA

1999	B. S. in Chemistry with Honors and Distinction, Stanford University
2000-2008	Trainee, Medical Scientist Training Program, University of California San Diego
2003-2004	Fellow, Taft Family Physical Sciences Fellowship, University of California San Diego
2004-2006	Fellow, La Jolla Interfaces in Science, Burroughs Wellcome Fund
2006	Ph. D. in Bioinformatics, University of California San Diego

PUBLICATIONS

Programming Interviews Exposed. John Mongan and Noah Suojanen; 254 pages, Wiley, 2000.

Complete, Randomly Ordered Traversal of Cyclic Directed Graphs. US Patent #6,189,116, issued 2001. John T. Mongan and Dorothy M. Cribbs.

Network Distributed Automated Testing System. US Patent #6,304,982, issued 2001. John T. Mongan, Dorothy M. Cribbs and John R. DeAguiar.

Monte Carlo Automated Test Generator. US Patent #6,378,088, issued 2002. John T. Mongan.

Accelerated Molecular Dynamics: A promising and efficient simulation method for biomolecules. Donald Hamelberg, John Mongan and J. Andrew McCammon. Journal of Chemical Physics **120**(24), 11919-29, June 2004.

Interactive Essential Dynamics. John Mongan. Journal of Computer-Aided Molecular Design, **18**(6), 433-36, June 2004.

Constant pH Molecular Dynamics in Generalized Born Implicit Solvent. John Mongan, David A Case and J. Andrew McCammon. Journal of Computational Chemistry, **25**(16), 2038-48, December 2004.

Biomolecular simulations at constant pH. John Mongan and David A. Case. Current Opinion in Structural Biology, **18**(2), 157-63, April 2005.

Limitations of atom-centered dielectric functions in implicit solvent models. Jessica M.J. Swanson, John Mongan and J. Andrew McCammon. Journal of Physical Chemistry B, **109**(31) 14769-14772, August 2005.

Potent, Selective Pyrone-Based Inhibitors of Stromelysin-1. David T. Puerta, John Mongan, Ba L. Tran, J. Andrew McCammon, and Seth M. Cohen. Journal of the American Chemical Society, **127**(41), 14148-49, October 2005.

FIELDS OF STUDY

Major Field: Bioinformatics

Studies in biomolecular structure and function
Professor J. Andrew McCammon, University of California San Diego

Studies in biomolecular dynamics
Professor David A. Case, The Scripps Research Institute

Studies in implicit solvation
Professor Alexey Onufriev, Virginia Tech.

Studies in computer-aided drug design
Professor Seth M. Cohen, University of California San Diego

ABSTRACT OF THE DISSERTATION

Implicit Solvent Method Development and Application: Fast Molecular Surfaces, Constant pH and Accelerated Dynamics, and Rational Drug Design

by

John Mongan
Doctor of Philosophy in Bioinformatics

University of California San Diego, 2006

Professor J. Andrew McCammon, Co-chair
Professor Gary Huber, Co-chair

Implicit solvation provides a means of accelerating and improving the efficiency of computational biomolecular studies by eliminating explicit solvent degrees of freedom while still representing the effects of solvation. Evidence is provided supporting the importance of defining the implicit solvent-solute boundary such that solvent is excluded from spaces smaller than a water molecule. The pairwise analytical generalized Born (GB) model, a popular implicit solvent model, is extended to incorporate this property. Methods for conducting molecular dynamics simulations at a constant pH, rather than the traditional constant protonation state, are reviewed and a constant pH method employing a consistent GB-based Hamiltonian for conformational and protonation state sampling is developed. Even with the improved efficiency of implicit solvent, it is difficult to achieve sufficient sampling in molecular dynamics. This problem is addressed by accelerated molecular dynamics, a technique for accelerating sampling that requires no advance knowledge of the potential energy landscape is presented. Analysis of molecular dynamics data is aided by Interactive Essential Dynamics, a tool for visualization of principal component analysis results. Implicit solvent methods are applied to the computer-aided design of inhibitors for the zinc(II) proteases stromelysin-1 and anthrax lethal factor. Inhibitors with IC_{50} of 100 nM and 14 μM are reported for stromelysin-1 and lethal factor, respectively. Use of the GB model developed here allows for accurate

elucidation of the binding mode of the lethal factor inhibitor, while GB models that allow solvent in spaces smaller than a water molecule identify an incorrect binding mode.

Chapter 1

Introduction

We are creatures who originated in the sea, and in a sense have never left it: our bodies are well encapsulated facsimiles of our earliest environment. This is because the functions of biomolecules that give us life—protein folding, enzyme catalysis, phospholipid bilayer formation—occur only in aqueous solution. Because water plays such a crucial role in biomolecular structure and function, computational studies of these systems must accurately represent the effects of water if they are to yield useful results.

The most straightforward and physically rigorous method of simulating water is to create many individual water molecule representations surrounding the biomolecular system of interest. This approach is called *explicit solvation*, because the water molecules are explicitly represented as particles in the system. While conceptually simple, explicit solvation has a number of drawbacks, mostly centering on issues of computational efficiency. Due to long-range ordering in water, the layer of water used to solvate a molecule must be fairly thick, usually at least 8–10 Å. This substantially increases the number of particles that must be simulated; if the solute molecule of interest is small, there may be many times more solvent atoms than solute atoms in the system. Motions of biomolecular solutes occur slowly in explicit solvent. Water is an easily yielding medium on the human scale, but on molecular size and time scales, it is not. The speed of any movement or conformational change of a solute is limited by the time it takes for water molecules to move out of the way. Probably the most important limitation of explicit solvation is that at any point in a simulation, the arrangement of

water molecules represents only one configuration out of a large ensemble of configurations. In almost all cases, the computational result of interest is not a result for just a single solvent configuration, but a result averaged over the entire ensemble of solvent configurations. Therefore, explicit solvation must always be coupled with a solvent configuration sampling method such as molecular dynamics or Monte Carlo sampling. Collecting a sufficient sample to accurately estimate ensemble average properties can be very time consuming.

An alternative approach, *implicit solvation*, avoids representation of individual water molecules in favor of a solvent region that simulates the ensemble average effects of water. Implicit solvation also has its limitations, particularly when interactions of the solute molecule with individual solvent molecules are important. Nevertheless implicit solvent is sufficiently accurate for most applications, and it effectively addresses the efficiency problems of explicit solvation: no solvent particles need be added to the system, there are no water molecules to impede motions and since the model calculates averaged properties, there is no need to sample multiple solvent configurations. This dissertation explores the development and improvement of implicit solvent methods, and the applications made possible by the improved efficiency of implicit solvation.

Since an implicit solvent method simulates a solvent region having the properties of water, a key component of any implicit solvent model is a definition of the boundary between the solvent region and the solute. Chapter 2 investigates the errors that result from the use of a class of boundary definitions, often employed for their simplicity and computational efficiency, that allow the solvent region to extend into spaces smaller than a water molecule. To address these problems, chapter 3 introduces an extension to a commonly used implicit solvent model, the pairwise generalized Born (GB) model, that modifies the boundary definition to exclude solvent from places where a water molecule cannot fit.

One of the earliest applications of implicit solvation was to the calculation of protein pK_a values. Implicit solvent models are well suited to studies involving protonation changes as there is no need for solvent re-equilibration after protonation state transitions. Chapter 4 examines a number of methods for predicting pK_a values and protonation states in biomolecules, with a focus on constant pH molecular dynamics. Unlike earlier

protonation state methods, which use mostly or entirely static structures, constant pH molecular dynamics maintains the coupling between protonation state and conformation by sampling these properties in conjunction with each other. A constant pH method using a single GB-based Hamiltonian for simultaneous molecular dynamics sampling of conformation and Monte Carlo sampling of protonation state is presented in chapter 5.

Obtaining sufficient sampling is a major problem in studies of biomolecules, as there are a very large number of degrees of freedom and many biologically interesting processes occur only on timescales that are orders of magnitude longer than what can be feasibly simulated with traditional techniques. Use of implicit solvation in general, and efficient implementations such as GB in particular, alleviates this difficulty by accelerating sampling somewhat and making it more computationally efficient. However, this degree of acceleration may not be suffcient. Furthermore, the improved accuracy and physical realism achieved by many of the previously described advances comes at the expense of making the sampling problem more difficult. The boundary definition modifications discussed in chapters 2 and 3 provide more accurate results by reintroducing energetic barriers found in explicit solvation to GB implicit solvation, but the time required to traverse these barriers slows sampling. Likewise, constant pH molecular dynamics allows for greater physical realism by eliminating the constant protonation state assumption of traditional molecular dynamics, but in doing so adds additional protonation state degrees of freedom to the system, complicating sampling. Finally, in some cases existing implicit solvent methods may not provide sufficient accuracy to simulate the phenomena of interest, so it may be necessary to forgo the efficiencies of implicit solvent for explicit solvent. For all these reasons, it is useful to have means for accelerating sampling. *Accelerated molecular dynamics*, a method for enhancing sampling rates in molecular dynamics is presented in chapter 6. Acceleration is achieved by biasing the potential energy function to reduce the depth of energy minima, thus reducing the effective height of energy barriers. This method is distinguished by requiring essentially no knowledge of the potential energy surface or location of “interesting” conformations, as is necessary for many other potential biasing methods. Additionally, unlike temperature based enhanced sampling methods, it does not increase particle velocities, so there is no

need to reduce the length of the time step.

Molecular dynamics simulations rapidly generate large quantities of data, involving motions on a wide range of time and size scales. Much of this motion is relatively uninformative small scale thermal vibration, which can obscure the more interesting slow, large scale motions. One technique for separating these motions is principal component analysis (PCA) applied to trajectory data, a combination often called *essential dynamics*. PCA is used to explore conformation–protonation state coupling in chapter 5 and to measure rates of conformational sampling in chapter 6. The mathematics behind PCA are fairly simple, so the technique has been widely implemented, but tools for the visual analysis of PCA results have been lacking. Therefore, *Interactive Essential Dynamics*, a tool for visualization of these data that was developed to aid in the analyses conducted in chapters 5 and 6 is presented in chapter 7.

A particularly useful application of implicit solvation is in calculation of binding free energies for proteins and small molecule ligands, since this forms the basis of rational drug design. Chapters 8 and 9 discuss the computational study and design of inhibitors for two zinc(II) proteases, stromelysin-1 and anthrax lethal factor. Results in chapter 9 underscore the importance of continued improvement of the models discussed here: a widely used GB model identifies an incorrect pose of the inhibitor within the lethal factor enzyme as having the lowest energy, but the GB model developed in chapter 3 identifies the correct binding mode.

In summary, this dissertation will identify and address some of the outstanding issues in current methods of implicit solvation, and describe improvements in accuracy that can be realized with minimal effect on computational efficiency. Applications of implicit solvation to improvement of the physical realism and efficiency of molecular dynamics, through constant pH molecular dynamics and accelerated molecular dynamics will be presented. Finally, the advances in implicit solvation are highlighted by improved results in application to rational drug design.

Chapter 2

Limitations of atom-centered dielectric boundaries

ABSTRACT

Many recent advances in Poisson-Boltzmann and generalized Born implicit solvent models have used atom-centered polynomial or Gaussian functions to define the boundary separating low and high dielectric regions. In contrast to the Lee and Richards molecular surface, atom-centered surfaces result in inter-atomic crevices and buried pockets of high dielectric which are too small for a solvent molecule to occupy. This chapter shows that these interstitial high dielectric regions are of significant magnitude in globular proteins, that they artificially increase solvation energies, and that they distort the free energy surface of non-bonded interactions. These results suggest that implicit solvent dielectric functions must exclude interstitial high dielectric regions in order to yield physically meaningful results.

2.1 Introduction

Continuum solvent models have become an increasingly useful tool in the characterization of biomolecular systems. The most popular such methods employ either the Poisson-Boltzmann (PB) or generalized Born (GB) models, treating the solute as a set of

point charges in a low dielectric cavity and the surrounding solvent as a uniform high dielectric medium. The PB model is generally considered to be more accurate and is often used to benchmark GB models. GB has found more extensive application in dynamical simulations, however, because it is computationally efficient and more amenable to force calculations.

One of the main challenges in the use of PB for dynamics has been the determination of numerically stable and accurate forces. Most PB calculations have used a dielectric boundary based on the molecular surface (MS) as defined by Lee and Richards,[1] which results in forces that are unstable over time, lack analytical definition, converge poorly, and are sensitive to grid discretization.[2] Furthermore, an abrupt dielectric transition results in numerical instability regardless of the location of the boundary. Recent advances in PB methods have avoided these difficulties by using overlapping atom-centered Gaussian or polynomial functions to define the solute surface, resulting in analytically defined, differentiable dielectric functions with smooth transitions between low and high dielectric values.[2,3] These dielectric definitions increase force stability and computational efficiency. However, unless modifications such as those presented by Luo *et al.*[4] and Lee *et al.*[5] are employed, they result in inter-atomic crevices and buried pockets of high dielectric that are too small for a solvent molecule to occupy.[4–6] It was originally postulated that the consequences of these regions, henceforth called *interstitial high dielectrics*, would be minimal and that either a MS or an atom-centered surface definition should be physically and theoretically equivalent,[2,7] but more recent work has suggested that atom-centered surfaces are physically flawed.[5,6] Nevertheless, implicit solvent models based on unmodified atom-centered dielectric functions are becoming increasingly popular in the biophysical community.[2,3,7–10] This chapter reports results showing that atom-centered surfaces create interstitial high dielectric regions of significant magnitude in globular proteins, increase solvation energies, and distort the free energy surface of non-bonded interactions. Although similar results are expected for most atom-centered smoothed dielectric boundaries, the focus here is on the spline surfaces (SS) introduced by Im *et al.*[3] and implemented in the Adaptive Poisson-Boltzmann Solver (APBS),[11] the PBEQ module in CHARMM,[12] and the GBSW model.[8]

2.2 Methods

The protein surface and energy calculations were performed with APBS 0.3.2 using a grid resolution of 0.2 Å. To facilitate comparison between the different dielectric boundaries, the Nina *et al.* optimized radii have been chosen to define the van der Waals surface (vdWS), MS, and SS.[13,14] For results with different boundary conditions to be comparable, the radii must be rescaled for the MS and particularly for the SS; the recommended rescaling has been performed. To check that results were not specific to Nina *et al.* radii, they were verified with AMBER optimized radii,[15] parm22 radii,[12] and Bondi radii.[16] The latter two were augmented by the spline window width ($w = 0.3$Å) for the SS: a simple but reasonable scaling. APBS versions 0.3.2 and earlier have a flaw in the MS algorithm that overestimates MS volumes by 2-5% and underestimates solvation energies by 1-3%. Results shown here were calculated with a modified algorithm that corrects this problem. The energy calculations were performed with zero bulk ionic strength, a temperature of 300 K, a solvent dielectric of 80, a solute dielectric of 1 and charges from the CHARMM22 all atom force field for which the Nina *et al.* radii were optimized.

Implicit solvent potentials of mean force (PMFs) for hydrogen bond formation were calculated by combining solvation, Coulomb and van der Waals energies. PB solvation energies were calculated with APBS and the same parameters as used for protein solvation energies except for a finer grid resolution of 0.1 Å. GBMV and GBSW solvation energies, Coulombic energies and vdW energies were calculated with CHARMM 31a1. AMBER GB solvation energies were calculated using the igb=1 model in AMBER 8. Because AMBER GB models are not compatible with atoms having zero radius, the hydrogen radii were increased to 0.8 Å. This made the outer surfaces of the hydrogen atoms approximately coincident with the surfaces of the atoms to which they were bound. The explicit solvent PMFs were calculated by WHAM from results of umbrella sampling in TIP3P solvent. Umbrella sampling was carried out using the PMEMD module of sander, modified to apply harmonic restraints to only the y and z coordinates of the peptides. Due to the use of different force fields for implicit and explicit solvent measurements, no quantitative comparison should be made. However, explicit solvent

potentials calculated with the CHARMM force field have the same general shape.[17]

2.3 Results and Discussion

To probe the magnitude of interstitial high dielectric regions in globular proteins the solute volumes generated by vdWS, MS, and SS were compared. Figure 2.1 shows the MS and SS dielectric values on a plane intersecting a structure of Intestinal Fatty Acid Binding Protein (IFABP) taken from a molecular dynamics (MD) simulation. Although MD conformations might be expected to contain more interstitial high dielectrics than NMR or crystal structures, surprisingly similar plots were obtained for all the systems in table 2.1. The solute volumes, reported in table 2.1, show a consistent trend across all 6 structures: substantial interstitial high dielectrics with the SS definition. Within the MS volume the SS renders 65–262 $Å^3$ of interstitial high dielectric space with a value of 80 and an additional 502–1121 $Å^3$ with values over 20. The total interstitial high dielectric space ranges from 12% for the crystal structures to 15% for the NMR and MD structures. The quantitative effects of these regions on electrostatic solvation energies are shown in table 2.2; SS energies are overestimated by 11–21%, only slightly less than the overestimation by vdWS.

The volume and dielectric value of the interstitial spaces created by the SS can be decreased by using a larger spline smoothing window for SS or longer Gaussian tails for Gaussian surfaces, but interstitial high dielectrics can not be eliminated altogether.[2] Unfortunately, longer tails overestimate the size of solvent exposed atoms and create unphysical bulges around overlapping and adjacent atoms.[5] These expanded dielectric boundaries yield solvation energies and forces that are severely underestimated. For example when a spline window of 1.0 Å is applied to the systems in table 2.1, interstitial high dielectrics with $\varepsilon > 20$ are essentially eliminated, but the volume of lowered dielectric outside the MS is increased dramatically and the solvation energies are underestimated by 35% to 48%. This over or under-estimation of solvation energies by SS has also been reported by Lee *et al.*[6] who used hybrid explicit/implicit solvation energies to test various continuum surface definitions.

Table 2.1: Solute volumes for different surface definitions. Solute volumes ($Å^3$) for FKBP12 (1FKG), lysozyme (1AKI), protease (1KZK) and IFABP (1AEL). The SS is larger because almost one fourth of its volume has dielectric (ε) values less than 80 but greater than 1. Interstitial high dielectric volumes are measured by the volume within the MS that has high dielectric values in the SS. The total percentage of interstitial high dielectrics relative to the MS volume, shown in parenthesis, range from 11% to 14%, with the largest values for the MD and NMR structures.

protein	**vdWS**	**MS**	**SS**	**interstitial high dielectric**					
	$\varepsilon = 1$	$\varepsilon = 1$	$\varepsilon < 80$	$1 < \varepsilon < 20$		$20 < \varepsilon < 80$		$\varepsilon = 80$	
1FKG-xtal	14863.3	16087.5	18508.6	1345.8	(8.4%)	502.5	(3.1%)	65.3	(0.4%)
1AKI-xtal	17495.8	18954.9	21473.0	1519.4	(8.0%)	608.1	(3.2%)	100.5	(0.5%)
1KZK-xtal	27528.0	29937.1	33885.2	2512.8	(8.4%)	1006.9	(3.4%)	139.7	(0.5%)
1KZK-md	27617.3	30245.4	34081.5	2576.2	(8.5%)	1120.9	(3.7%)	202.3	(0.7%)
1AEL-nmr	17589.4	19361.9	21869.6	1532.7	(7.9%)	797.0	(4.1%)	196.8	(1.0%)
1AEL-md	19318.5	21585.4	24579.4	1930.0	(8.9%)	1031.1	(4.8%)	262.5	(1.2%)

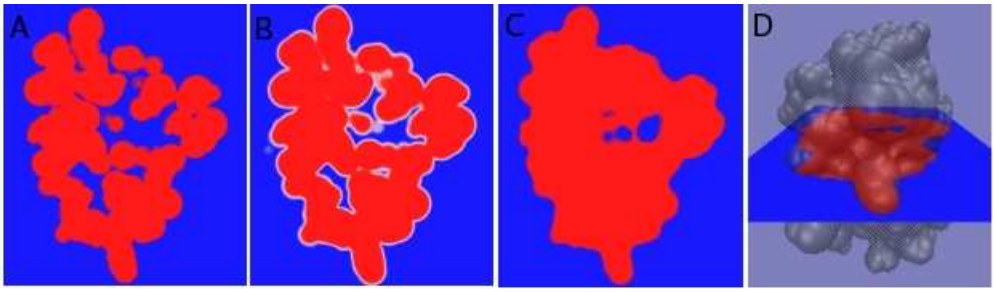

Figure 2.1: Dielectric maps of IFABP. Dielectric values on a plane intersecting IFABP for the vdWS (A), SS (B) and MS (C). Red regions have $\varepsilon = 1$, blue have $\varepsilon = 80$ and white regions have intermediate dielectric values. The location of the intersecting plane is shown in (D).

Table 2.2: Electrostatic solvation energies for different surface definitions. All energies are reported in kcal/mol. The SS yields energies much larger than the MS and similar to the vdWS. Percentages of vdWS and SS overestimation relative to the MS energies are given in parenthesis.

protein	MS	vdWS		SS	
1FKG-xtal	-1475.8	-1680.4	(13.9%)	-1638.3	(11.0%)
1AKI-xtal	-1724.3	-2034.7	(18.0%)	-1976.7	(14.6%)
1KZK-xtal	-2196.0	-2585.6	(17.7%)	-2499.9	(13.8%)
1KZK-md	-2162.5	-2549.1	(17.9%)	-2475.9	(14.5%)
1AEL-nmr	-2247.4	-2838.2	(26.3%)	-2727.0	(21.3%)
1AEL-md	-1979.3	-2368.9	(19.7%)	-2301.9	(16.3%)

While the ability to calculate atomic forces in a PB model is an important advance, such forces can have useful application only if the potential they are derived from accurately represents the physics of the system. In particular, solvation models employed in dynamical simulations must be capable of accurately calculating high energy as well as low energy conformations. Therefore, dynamics may constitute a more demanding test of a solvation model than calculating solvation energies of static structures, which tend to be dominated by low energy configurations of atoms.

Hydrogen bonds are of particular interest in simulations of biomolecules; since solvation effects make a large contribution to these interactions, the PMF for the separation of a hydrogen bond can be used as a test of the quality of a solvation model. The PMF of hydrogen bonding between the delta hydrogen and the epsilon nitrogen of two delta protonated histidines calculated with a variety of solvation methods is shown in figure 2.2. Both PB and GB results based on a MS dielectric boundary faithfully represent the important features of the explicit solvent PMF: a narrow minimum and a significant barrier to separation of the hydrogen bond. The energetic barrier in the MS PMFs comes about because the electrostatic energy rises rapidly as soon as the hydrogen bond participants are separated, but the solvation energy does not substantially increase in magnitude (become more negative) until the bond is sufficiently separated that the solvent probe will fit between the participants. The SS based implicit solvent models have good performance near the minimum and at long distances, but fail to capture the appropriate energetic barrier. At the separation where the MS PMF energy peaks, the SS has large interstitial high dielectrics, which result in a more negative solvation energy. This produces an artifactual minimum—or in less extreme cases, a shoulder—near a location where the PMF should have a maximum. AMBER (igb=1) GB results, based on the model of Hawkins, Cramer and Truhlar,[18] illustrate that the lack of a barrier to separation is a general feature of implicit solvent models that allow interstitial high dielectrics. In comparison to the SS results, the AMBER GB PMF is somewhat broader near the minimum, but is smoother and avoids the second minimum seen in the SS PMF.

Discrepancies between MS and SS PMFs are most dramatic for interactions between sterically bulky groups, where the magnitude of the interstitial high dielectrics is largest, but are observed to a greater or lesser degree across a variety of hydrogen bond and salt

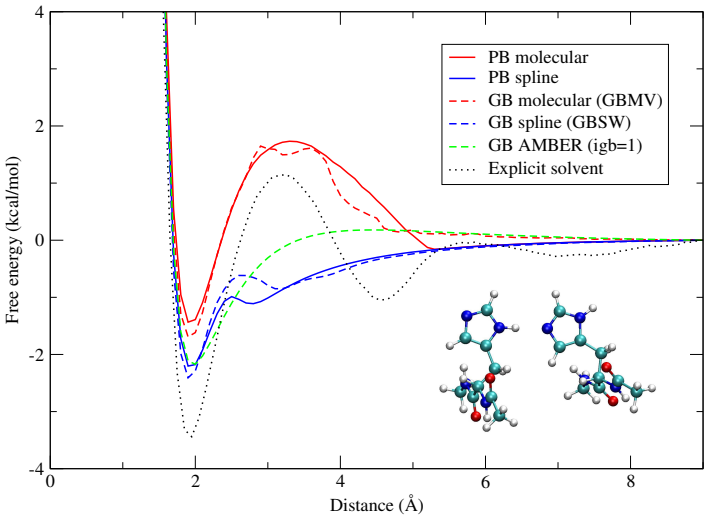

Figure 2.2: PMF for illustrated histidine-histidine hydrogen bond. Distances are measured between the hydrogen and nitrogen atoms participating in the bond. Spline-based dielectric boundaries fail to capture the free energy barrier to hydrogen bond separation because they allow interstitial high dielectrics near the hydrogen bond as it is separated.

bridge systems as seen in figures 2.2 through 2.6. For simplicity, apolar solvation contributions have been ignored in the implicit solvent PMFs presented here. A traditional surface area apolar term changes the depth of the minimum, but has no appreciable effect on the discrepancies between MS and SS PMFs.

2.4 Conclusion

The introduction of atom-centered dielectric functions has been a significant advance for PB force calculations. They can be analytically defined and easily smoothed allowing for numerical stability and increased efficiency. However, this chapter demonstrates that atom-centered surfaces produce large volumes of interstitial high dielectrics in globular proteins which artificially overestimate solvation energies and distort the free energy profile of non-bonded interactions such as hydrogen bonds and salt bridges. Dynamical simulations conducted using these dielectric boundaries will sample incorrect conformational ensembles. These findings suggest that although the optimal surface definition should be smooth and differentiable it should also exclude interstitial high

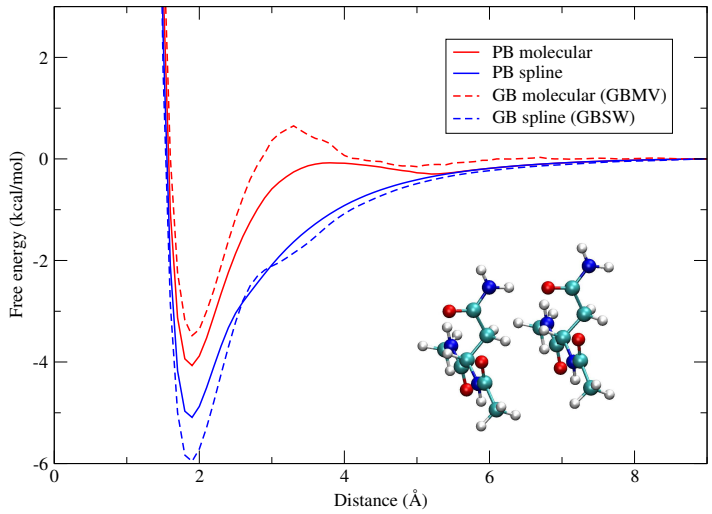

Figure 2.3: PMF for illustrated orientation of asparagine-asparagine hydrogen bond.

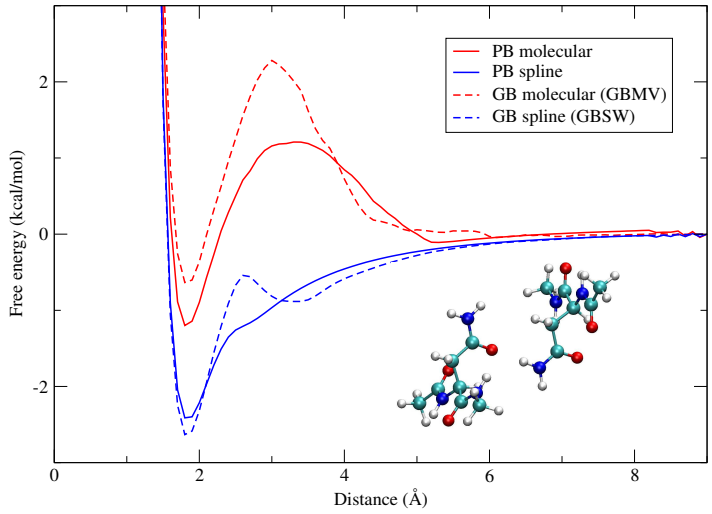

Figure 2.4: PMF for alternate orientation (illustrated) of asparagine-asparagine hydrogen bond.

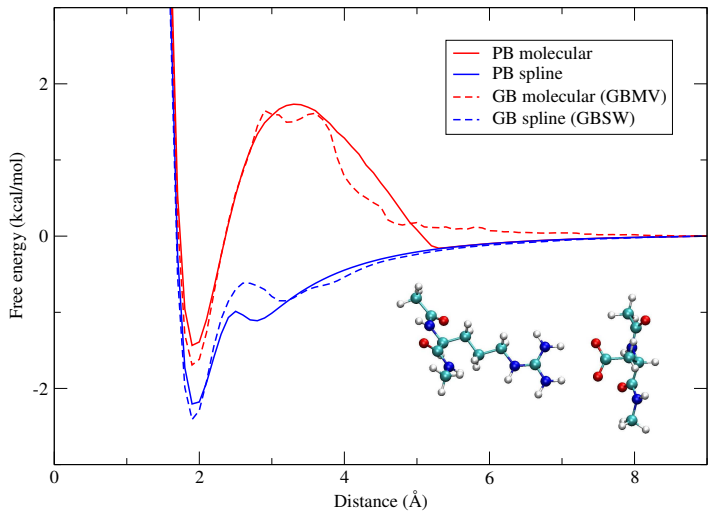

Figure 2.5: PMF for illustrated orientation of arginine-aspartate salt bridge.

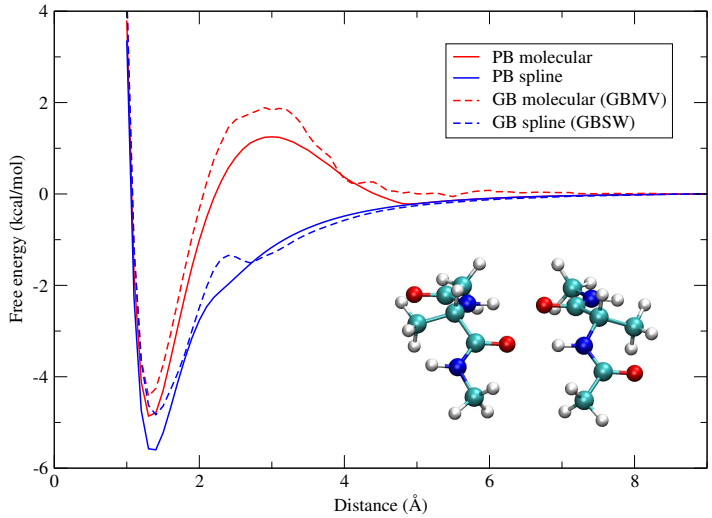

Figure 2.6: PMF for β-sheet hydrogen bonding model (alanine-alanine).

dielectrics as the MS does. Dielectric boundaries that address this issue, such as those proposed by Luo *et al*.,[4] and Lee *et al*.[5] will be critical for further improvement of PB and GB models.

This chapter is a reprint in full of material that appeared in *Limitations of atom-centered dielectric functions in implicit solvent models*. Jessica M.J. Swanson, John Mongan and J. Andrew McCammon. Journal of Physical Chemistry B, **109**(31) 14769-14772, August 2005. I was the secondary researcher and author of this work.

Chapter 3

Generalized Born with a simple, robust molecular volume correction

ABSTRACT

Generalized Born (GB) models provide a computationally efficient means of representing the electrostatic effects of solvent and are widely used, especially in molecular dynamics. A class of particularly fast GB models is based on integration over an interior volume approximated as a pairwise union of atom spheres—effectively, the interior is defined by a van der Waals rather than Lee-Richards molecular surface. The approximation is computationally effective, but if uncorrected, allows for non-physical interstitial high dielectric (water) regions between the atoms, leading to decreased accuracy. Here, an earlier pairwise GB model is extended by a simple analytic correction term that largely alleviates the problem by correctly describing the solvent-excluded volume of each pair of atoms. The correction term introduces a free energy barrier to the separation of non-bonded atoms. This free energy barrier is seen in explicit solvent and Lee-Richards molecular surface implicit solvent calculations, but has been absent from earlier pairwise GB models. The correction term yields hydrogen bond length distributions that are in better agreement with explicit solvent results. The robustness and simplicity of the correction preserves the efficiency of the pairwise GB models while making them a better approximation to reality.

3.1 Introduction

The effects of aqueous solvent are critical to the structure and function of biological macromolecules. Commonly, solvent is represented explicitly, by models of multiple water molecules, or implicitly, by a high dielectric region and additional apolar solvation terms. Although explicit solvent is a more physically rigorous representation, implicit solvent models have the advantage of dramatically reducing the degrees of freedom that must be sampled by eliminating those associated with the solvent. Additionally, implicit solvent models are often more computationally efficient than their explicit counterparts.

The solvation effects can be described by ΔG_{solv}: the free energy of transferring a given configuration of a molecule from vacuum to solvent. To facilitate calculation of ΔG_{solv}, it is typically decomposed into polar and nonpolar components: $\Delta G_{solv} = \Delta G_{pol} + \Delta G_{nonpol}$. Here, ΔG_{nonpol} is the free energy of introducing the solute molecule into solvent while electrostatic interactions between the solute and solvent are turned off, and ΔG_{pol} is the free energy change in the system resulting from turning these electrostatic interactions back on. In this work, the focus is on methods for calculating ΔG_{pol}.

Assuming that the solvent can be faithfully represented by a continuum dielectric region, the Poisson-Boltzmann (PB) equation is the most physically correct method of determining ΔG_{pol}, and has been widely used over the past decade.[19–25] Application of PB to molecular geometries requires numerical solution of second order partial differential equations, which is fairly computationally intensive and does not easily provide forces, although recent advances in PB methodology have improved the situation somewhat.[3,4,19,26] Alternatively, generalized Born (GB) models have become popular as a computationally efficient approximation to numerical solutions of the PB equation,[18,24,27–38] especially for use in dynamics.[39–49]

GB models evaluate polar solvation free energy as a sum of pairwise interaction terms between atomic charges. When the solute dielectric is 1 and the solvent dielectric is much greater than that of the solute,[50] the interactions can be accurately described by an analytical function first proposed by Still *et al.*,[28] that interpolates between the

Coulombic limit at long distances and the Born or Onsager limits at small distances,

$$\Delta G_{pol} \approx \Delta G_{GB} = -\frac{1}{2}\sum_{i,j} \frac{q_i q_j}{\sqrt{r_{ij}^2 + R_i R_j \exp\left(\frac{-r_{ij}^2}{4R_i R_j}\right)}} \left(1 - \frac{1}{\varepsilon_w}\right) \tag{3.1}$$

where r_{ij} is the distance between atoms i and j, q_i and q_j are partial charges and ε_w is the dielectric constant of water. The key parameters in this GB function are the effective radii of the interacting atoms, R_i and R_j, which represent each atom's degree of burial within the solute. More specifically, the effective radius of an atom is defined as the radius of a corresponding spherical ion having the same ΔG_{pol} as the self energy of this atom in the molecule. The self energy is the polar solvation free energy for the molecule with partial charges set to zero for all atoms except the atom of interest. The effective radius of an atom is larger than the intrinsic radius of its atom sphere because of the de-screening effects of surrounding atoms, reducing the extent to which the atom charge is screened by solvent. A computationally inefficient, but theoretically interesting method for determining effective radii is to derive them from self energies calculated using well converged numerical PB solutions. When these "perfect" effective radii are used, GB results are in close agreement with PB results,[51] which serve as a natural point of reference for assessing the accuracy of GB, since current GB models are an approximation to the more fundamental formalism of the PB equation. Although this form of GB is impractical for application, it suggests that in aqueous solution the GB function introduced by Still *et al.* is a minor source of error compared with the error introduced by (non-perfect) methods for estimating effective radii. Consequently, considerable effort has been spent on improving the way effective Born radii are computed.

In practice, effective radii for each atom are generally calculated by integration of an approximate electric field density due to the atom of interest over some definition of the molecule's volume,[5,18,23,29,36,47,52] although formulations based on surface integrals have also been proposed.[34,38] Here, the focus is on volume-based GB models which have traditionally used a Coulomb field integral,

$$I_i = \frac{1}{4\pi}\int_{\Omega_i} \mathbf{r}^{-4} d^3\mathbf{r} \tag{3.2}$$

where the origin is centered on atom i and Ω_i represents the volume inside the molecule but outside atom i. The effective radius is then calculated according to

$$R_i = \left(\rho_i^{-1} - I_i\right)^{-1} \tag{3.3}$$

where ρ_i is the intrinsic radius of atom i. Within the Coulomb field approximation (CFA) embodied by the integral in equation 3.2, it is assumed that the electric field generated by an atomic point charge is unaffected by the non-homogenous dielectric environment created by the solute, so that the field has the form described by Coulomb's law. The CFA is exact for a point charge at the center of a spherical solute, but it over estimates effective radii for molecular geometries[36] as well as for spherical regions when the charge is off center.[53] It has been suggested that some of the success of early GB models on small molecules may be attributed to fortuitous cancellation of errors in effective radius calculations between the over estimates of a CFA based integrand and the under estimates of a van der Waals (VDW) based region of integration.[47] Improved approximations based on empirical corrections to the CFA[5,8,36] or theoretical derivations originating with the Kirkwood formula[53,54] have significantly better agreement with effective radii calculated from PB self energies.

The integration in equation 3.2 can be performed numerically[5,8,28,34,36] or by an analytical pairwise approximation.[18,29,30,47,48,52] GB methods based on analytically approximated integrals are easily extended to calculate solvation forces and are generally faster than their numerically integrated counterparts,[55] so they have traditionally found greater application in dynamics.

Most pairwise approximations estimate the integral over a region formed by the union of atom spheres, which is equivalent to a VDW surface dielectric boundary. In calculating the effective radius for atom i, the contribution of every other atom $j \neq i$ to the integral is determined as a function of ρ_j and the distance between atoms i and j. Summation of these terms yields an overestimate of the total integral, due to overlap between descreening atoms. To correct for these overlaps, multiplicative scaling factors, S_x, are introduced to reduce the intrinsic radius of each descreening atom.

In contrast, PB calculations generally use a Lee-Richards molecular surface dielectric boundary, defined by rolling a solvent sphere over the surface of the molecule.[1]

Although there is no uniquely correct definition of the dielectric boundary, a van der Waals surface creates regions of interstitial high dielectrics that are smaller than a water molecule, while the Lee-Richards surface has the conceptually attractive advantage of excluding high dielectric from regions into which a water molecule is too large to fit. Differences between the Lee-Richards and VDW surface definitions are minimal for small molecules, where all atoms are well solvated, but become more substantial for macromolecules, where inclusion of interstitial high dielectrics in VDW-based models leads to overestimation of the solvation of interior atoms, relative to Lee-Richards results.[56] This may partially explain why early GB models that had good results for small molecules were less effective when applied to macromolecules.[41,47,52] Additionally, implicit solvent models that allow interstitial high dielectrics produce incorrect potentials of mean force between non-bonded atoms.[56] However, it is not practical to use the Lee-Richards surface directly in a GB model as it is fairly computationally intensive and can produce unstable or infinite forces for some molecular configurations.[19,26]

Attempts to reduce or eliminate the problems of interstitial high dielectrics in GB models have followed two paths. One approach, embodied by the GBMV2 model developed by Lee *et al.*,[5] has been to use numerical integration with adaptations for calculating forces in combination with an analytic surface definition that closely approximates the properties of the Lee-Richards surface. A CFA correction term is also employed in the integration. This GB model yields stable dynamics while providing excellent agreement with PB Lee-Richards surface results. However, both the analytic surface definition and the numerical integration are relatively slow, such that the fastest PB models approach the performance of GBMV2.[55] Furthermore, the reliance on numerical integration introduces artifacts, such as a lack of rotational invariance.

A different method (*OBC* GB), developed by Onufriev, Bashford and Case,[47] sought to extend the pairwise integration method (*HCT* GB) of Hawkins, Cramer, and Truhlar[18,29] to reduce the effect of interstitial high dielectrics. Based on the observation that effective radii for buried atoms are larger than for surface atoms, but much smaller than PB-derived "perfect" effective radii, this method modifies the radius calculation in

equation 3.3 by rescaling the integral from equation 3.2 according to

$$R_i = \left(\tilde{\rho}_i^{-1} - \rho_i^{-1}\tanh\left(\alpha I\tilde{\rho}_i - \beta(I\tilde{\rho}_i)^2 + \gamma(I\tilde{\rho}_i)^3\right)\right)^{-1} \tag{3.4}$$

where $\tilde{\rho}_i = \rho_i - 0.09$Å and α, β and γ are tunable parameters. When these parameters are set such that most radii are scaled up, the rescaled radii substantially improve agreement with PB solvation free energies, and the computational expense of the rescaling function is minimal so the efficiency of the Hawkins *et al*. model is retained. In addition, effective radii calculated with equation 3.4 are smoothly capped at about 30Å, avoiding problems with numerical instability and negative radii that can be encountered when using equation 3.3. However, by design, the rescaling function only affects atoms that are sufficiently buried that the interstitial high dielectrics can be accounted for in an averaged, geometry-independent manner. Uncompensated interstitial high dielectrics between more highly solvated surface atoms still affect solvation energies and potentials of mean force.

The work in this chapter attempts to combine the best aspects of both of these efforts in development of a GB model that adds a geometrically based molecular volume correction term accounting for interstitial high dielectrics to the pairwise approximated integration method. Since the correction term is, itself, a computationally efficient pairwise approximation, the performance and numerical benefits of analytical GB models are retained.

The shortcomings of the CFA are now well known, but rigorously derived non-CFA pairwise approximated GB models have only recently been described[54] and their stability and performance have not yet been extensively tested on biomolecules, so the model described here extends the Coulomb field-based *HCT* GB model.

3.2 Theory

An ideal volume correction term for a GB model based on VDW volume and the CFA would yield the integral of $\mathbf{r}^{-4}$ over the region inside the Lee-Richards molecular surface and outside the van der Waals surface. This region is designated the correction

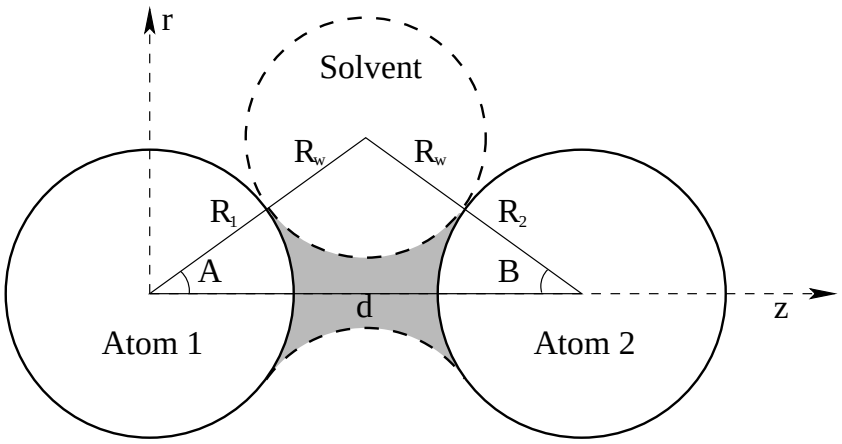

Figure 3.1: The neck region (shaded) is defined by the radius of atom 1, R_1, the radius of atom 2, R_2, the distance that separates them, d, and the radius of the solvent molecule, R_w. The coordinate system used for performing integration is also illustrated.

region.

$$\int_{LR} \mathbf{r}^{-4} d^3\mathbf{r} = \int_{VDW} \mathbf{r}^{-4} d^3\mathbf{r} + \int_{correction} r^{-4}\mathbf{r}^{-4} d^3\mathbf{r} \tag{3.5}$$

Since the *HCT* GB integration scheme calculates the value of the integral within the van der Waals surface, adding this correction term would yield an integral over the region within the molecular surface. In the general case, the correction region cannot be analytically defined. However, in the simple case of two closely spaced or overlapping atoms, the correction region forms an analytically definable "neck" region between the two atoms, as seen in figure 3.1. The general case of the correction region can be approximated by a union of these neck regions calculated pairwise between atoms. In the simplest form of this approximation, developed here, the integral for each atom includes corrections for only the neck regions in which the atom is directly involved. This simple form is a reasonably good approximation because the value of the integrand ($\mathbf{r}^{-4}$) is much higher in the nearby neck regions with which the atom is directly involved than in the distant portions of the correction region formed by interactions between other pairs of atoms.

Figure 3.1 illustrates how the geometry of the neck region is defined by four parameters: the radii of the two atoms, R_1 and R_2; the radius of the solvent molecule, R_w; and the distance between the two atoms, d. Derivations of the expressions for the CFA

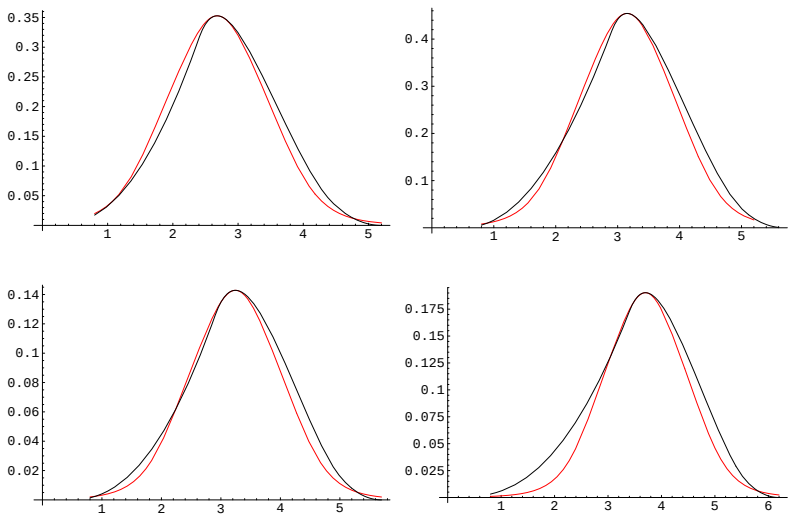

Figure 3.2: Values of numerical integration over the neck region (black) and analytical approximation (red) as a function of distance between atoms in angstroms. Left to right, top to bottom, radii (in angstroms) for atoms 1 and 2, respectively are 1.2 and 1.2; 1.2 and 1.7; 1.7 and 1.2; 1.7 and 1.7.

integrals over the neck region are given in Appendix I. Although the integrands in these expressions are fairly simple, the limits of integration are sufficiently complex to make analytical solution of the integrals impractical. The problem is simplified by considering that in the GB model, parameters R_1, R_2 and R_w have a relatively small set of discrete values (a single value, in the case of $R_w = 1.4$ Å), and so d is the only parameter with continuous values. With this view in mind, the function in four variables described by these integrals can be evaluated as a family of single variable functions of d, with each function determined by a particular set of values for R_1, R_2 and R_w. These functions of d can be plotted by solving the integrals numerically for a range of values of d, producing curves as shown in figure 3.2.

Numerical solution of these integrals is far too computationally costly for application in a GB model. Instead, they are replaced with an empirically determined analytic function shown in equation 3.6

$$neck\_integral(d) = \frac{m_0}{1 + (d - d_0)^2 + 0.3(d - d_0)^6} \tag{3.6}$$

This function is parameterized by the position (d_0) and value (m_0) of the maximum,

which are determined by numeric optimization (maximization) of the integral of $\mathbf{r}^{-4}$ over the neck region of figure 3.1. The values of d_0 and m_0 are dependent on R_1, R_2 and R_w, but since these variables have a small set of discrete values, tabulating all possible values of d_0 and m_0 is quite feasible (see Appendix II). As illustrated in figure 3.2, equation 3.6 is a very good approximation over the range of atomic radii typically encountered in biomolecules.

Applications of GB solvation models to dynamics require calculation of derivatives with respect to distance. Equation 3.6 is easily differentiated, yielding equation 3.7.

$$neck\_integral'(d) = -\frac{\left(2(d-d_0)+\frac{9}{5}(d-d_0)^5\right)m_0}{\left(1+(d-d_0)^2+\frac{3}{10}(d-d_0)^6\right)^2} \tag{3.7}$$

Ideally, neck integrals would be calculated only between atoms that are close enough to define a neck region ($d < R_1 + R_2 + 2R_w$): beyond this distance the neck integral and its first derivative with respect to d should be zero. However, the analytic approximation used here approaches zero asymptotically, and at $d = R_1 + R_2 + 2R_w$ its value is on the order of 10^{-3}. Truncating the function at this point would create a discontinuity which could lead to unstable dynamics. A variety of techniques could be employed to smooth this discontinuity; the simplest approach has been taken of continuing to calculate the neck correction for $d > R_1 + R_2 + 2R_w$ until d is large enough that the value of the function is sufficiently small that the error of truncating it is on the order of rounding error.

The neck correction described by the integrals in Appendix I and approximated by equation 3.6 is exact for a system of two atoms, but in the usual case of a molecule with more than two atoms, a strict summation of neck integrals calculated pairwise between atoms will tend to over estimate the integral over the correction region. Over estimation of the integral is due to overlap of neck regions with atoms not participating in the neck, as well as overlap with other neck regions, and must be corrected by scaling the contributions to the total integral.

The *GBn* model ("n" for neck) presented here takes a simple, two step approach to scaling. First, each neck integral value calculated in equation 3.6 is multiplied by

a scaling factor S_{neck} ($S_{neck} < 1$). Second, effective radii are calculated using equation 3.4 which provides descreening dependent scaling, as well as numerical stabilization for large effective radii. The two step scaling involves four parameters which must be optimized, S_{neck}, α, β and γ. Since the neck correction alone is expected to bring the integration volume closer to molecular volume, the optimal parameters of equation 3.4 are different from those used by the *OBC* model. The key difference between *GBn* and *OBC* GB can be best illustrated by a diatomic system such as that in figure 3.1: the *OBC* model will produce correct effective radii for only one value of atom-atom separation distance, while the *GBn* model should calculate accurate radii for this simple system across the entire range of interatomic distances.

Additionally, it is necessary to refit the intrinsic radius scaling factors, S_x. Although formally the S_x scaling factors merely correct overlaps, in practice they have been used as free parameters to optimize GB results for agreement with PB and experimental results.[18,41] As a result, the sets of S_x values used in the *HCT* and *OBC* GB models not only correct for atomic overlaps, but also correct for some of the effects of the CFA and interstitial high dielectrics (to the extent that this is possible on an averaged, geometry independent basis). Since the *GBn* model already accounts for interstitial high dielectrics with the neck term and has a different degree of CFA error due to the altered region of integration, it would clearly be inappropriate to use S_x sets that were fit for VDW regions of integration with the *GBn* model.

3.3 Results and Discussion

Parameters of the *GBn* model (S_{neck}, α, β, γ and the S_x parameters for atom types C, H, N and O) were optimized using the Nelder-Mead simplex algorithm.[57] The objective function that was minimized measured agreement between PB and GB solvation free energies over a training set consisting of structures from denaturation trajectories of apo myoglobin and protein L and structures representing potentials of mean force (PMF) for two hydrogen bonds and a salt bridge (see Methods for details of the objective function). The objective function has multiple local minima, so 100 minimizations were performed

Table 3.1: Optimized scaling parameters

Parameter	Value
α	1.095
β	1.908
γ	2.508
S_{neck}	0.362
S_H	1.091
S_C	0.484
S_N	0.700
S_O	1.066

starting from random initial points. Optimized parameter values producing the best overall performance are given in Table 3.1. Treatment of the S_x values as free parameters to optimize GB performance beyond their formal purpose of correcting overlap is made obvious by the values of S_O and S_H, which exceed 1. This represents a continuation of previous practice, although it may at first appear to be a divergence because previous sets of S_x values where all $S_x < 1$ may have been incorrectly interpreted as merely correcting overlaps.

Since the primary purpose of adding the neck correction is to improve the accuracy with which effective radii are calculated, one simple assessment is to compare these effective radii with "perfect" radii derived from PB calculations, as previously described. It should be noted that while GB has excellent agreement with PB when perfect radii are used directly,[51] small improvements in this agreement do not always translate to improved performance of the model. In particular, sets of scaling parameters optimized for minimal deviation between GB effective and perfect radii had poor performance on the higher level tests of model quality described below. Nevertheless, radius comparisons are instructive as rough quality measures and in identifying sources of error that may not be readily apparent when molecular solvation free energies are compared. It is most useful to compare inverse radii, as this most faithfully represents the contribution of the effective radii to the energy in equation 3.1. As shown in figure 3.3, the accuracy of effective radii calculated by the *GBn* model is improved (R_i^{-1} RMSD 0.092 vs 0.128) when compared to the *OBC* GB[47] model. Although there is improvement in the

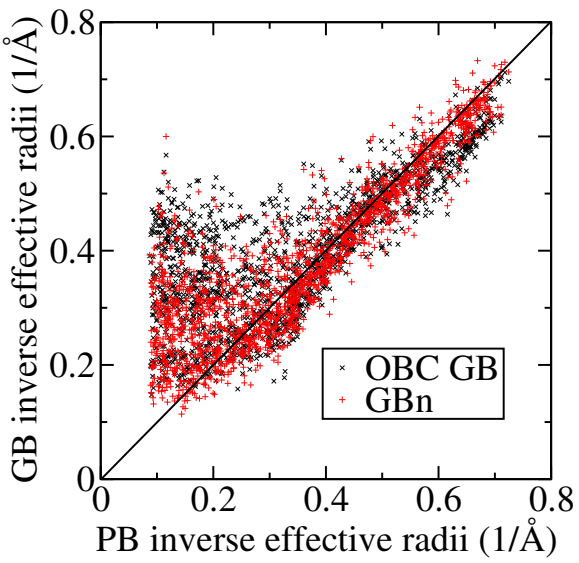

Figure 3.3: Scatter plot comparison of inverse effective radii calculated by the current GB neck model (red +) and earlier OBC GB model (black X) to inverse "perfect" PB radii for thioredoxin (PDB code 2TRX). Diagonal line indicates perfect agreement.

accuracy of large effective radii (left portion of the figure), these radii continue to have the largest errors. Errors seem to be largest for atoms near crevices that are slightly too small for a water molecule; presumably the pairwise appoximation is poorest here. The *OBC* GB model is selected as a reference for comparison because it is among the most recent and most accurate[55] pair-wise GB models that do not have a molecular volume correction beyond the "average" rescaling provided by equation 3.4.

A more direct test of GB model performance is comparisons of GB solvation free energies with those calculated by PB methods. Minimizing error across multiple conformations of the same system is of particular interest for GB methods that will be used in dynamics, as conformation-dependent errors will bias sampling. Figure 3.4 plots the difference between GB and PB solvation free energies for a series of conformations obtained from a thermal denaturation molecular dynamics trajectory. Error is reduced for the *GBn* model (standard deviation 6.4 kcal/mol) relative to the *OBC* GB model (standard deviation 7.2 kcal/mol). Solvation free energy errors are plotted as a function of the number of native tertiary contacts for the corresponding conformation to elucidate trends in error with respect to degree of denaturation. The GB model of Hawkins *et al.*

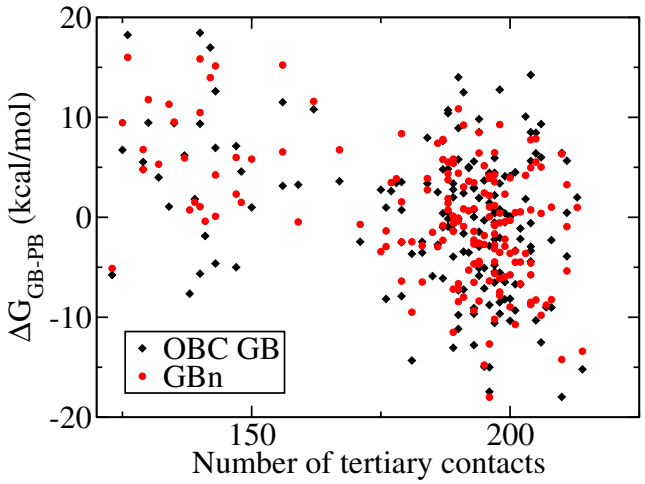

Figure 3.4: Relative deviation from PB solvation energy for *GBn* and *OBC* GB for a series of snapshots from a denaturation trajectory of protein A. *GBn* has a tighter clustering of points, indicating less random error than *OBC* GB (stdev 6.4 vs 7.2 kcal/mol), while maintaining a similar native state bias (trend of points across the plot). Average errors of -9.2 (*OBC* GB) and 68.9 (*GBn*) kcal/mol removed to facilitate comparison.

has significantly more negative errors for near-native conformations than for denatured conformations, but this native state bias is almost entirely corrected by the rescaling function in equation 3.4 employed by the *OBC* GB model.[47] As seen in figure 3.4, the *GBn* model has a very small native state bias, similar to *OBC* GB. Similar, slightly better results are obtained for conformations of protein L and apo-myoglobin; these results are not shown because they were used as part of the objective function in the optimization process and as such are likely to be less indicative of performance on other systems than the protein L results.

The improvements in effective radius and solvation free energy calculations described above represent useful but fairly incremental improvement over the existing *OBC* GB model. Indeed, the *OBC* GB model's performance is quite good on low free energy conformations, such as those found in crystal structures or sampled from molecular dynamics trajectories, making dramatic improvements on these structures unlikely. However, performance on higher free energy conformations is also important for common applications like dynamics and docking; here there is ample room for improvement on *OBC* GB. One common high free energy conformation is encountered in the free en-

ergy curve for separating a salt bridge or hydrogen bond, referred to here as a PMF to reflect the averaging of solvent degrees of freedom by the implicit solvent model. It has been shown that implicit solvent models that employ a molecular surface dielectric boundary have a free energy barrier to separation of the bond,[56] in qualitative agreement with explicit solvent results,[17] but models based on traditional pair-wise integration, even with average volume corrections such as *OBC* GB, fail to reproduce this behavior.

Since the *GBn* model attempts to approximate a molecular surface dielectric boundary it should be capable of reproducing the maximum in the PMF. As shown in figures 3.5 and 3.6 this result is seen in most cases, a distinct departure from implicit solvent models that allow interstitial high dielectrics.[56] In general, the *GBn* minima are less deep and the maxima are less high than the PB PMFs. This is probably a consequence of the CFA. The CFA underestimates the descreening contribution of nearby regions relative to more distant regions, because $\mathbf{r}^{-4}$ diminishes less rapidly than the higher order integrands of more accurate expressions.[36,53] Since the neck region is very close to the atom of interest, it seems likely that its effect is underestimated by the CFA, leading to a smaller difference between minimum and maximum. The shallow minima exhibited by the *GBn* model, most notable in the β-sheet model of figure 3.5, raise concerns that secondary structure may not be stable, possibly leading to denaturation. However, this has not been observed in molecular dynamics trajectories (see following), perhaps because the extent of destabilization is less in the protein environment than for these highly solvated model systems, or because the time scales of the simulations conducted here are not sufficient to observe these problems.

The primary purpose for the development of computationally efficient pairwise approximated GB is application in dynamcs; the *GBn* model, implemented in AMBER, was tested by conducting 10 ns molecular dynamics trajectories of ubiquitin and thioredoxin. As expected, the *GBn* model retains the computational efficiency of the *OBC* GB model running only 8-10% more slowly. Conformational stability of trajectories is commonly assessed by computing the RMSD of alpha carbons from their crystal coordinates; plots of the RMSD for thioredoxin and ubiquitin trajectories conducted using the *GBn* and *OBC* GB models are shown in figure 3.7. The *GBn* model maintains approximately the same high level of stability as *OBC* GB, with slightly higher RMSD in

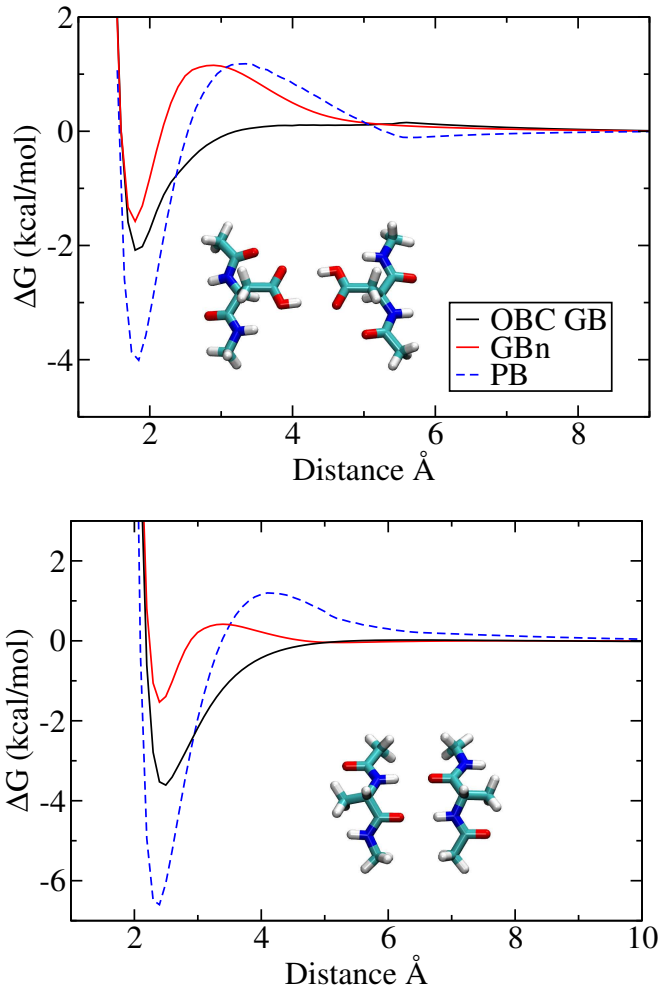

Figure 3.5: Potentials of mean force for hydrogen bonding systems not included in the objective function, calculated with three implicit solvent methods. Systems are two protonated aspartic acids and two alanines (β-sheet model). Potential includes electrostatic and van der Waals energies.

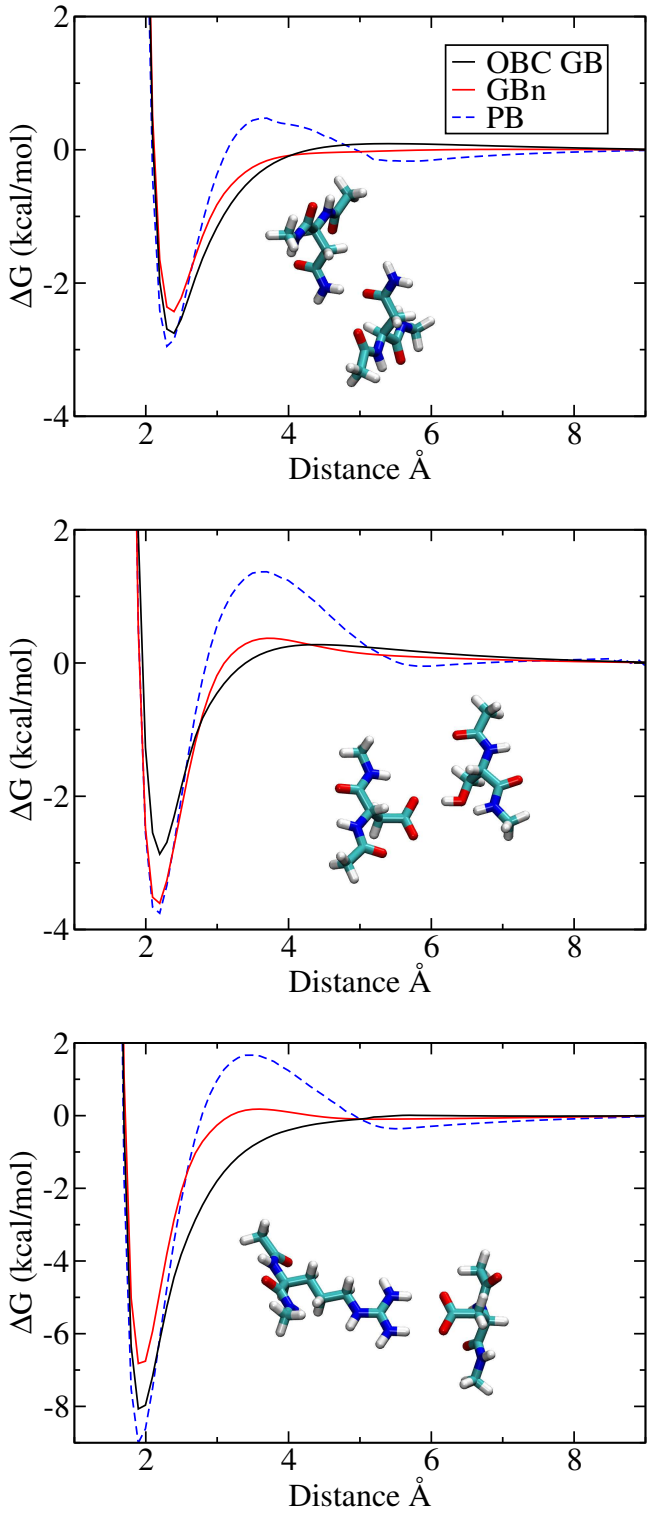

Figure 3.6: Potentials of mean force for hydrogen bonding and salt bridge systems included in the objective function, calculated with three implicit solvent methods. Systems are asparagine and asparagine; asparate and serine; arginine and aspartate. Potential includes electrostatic and van der Waals energies.

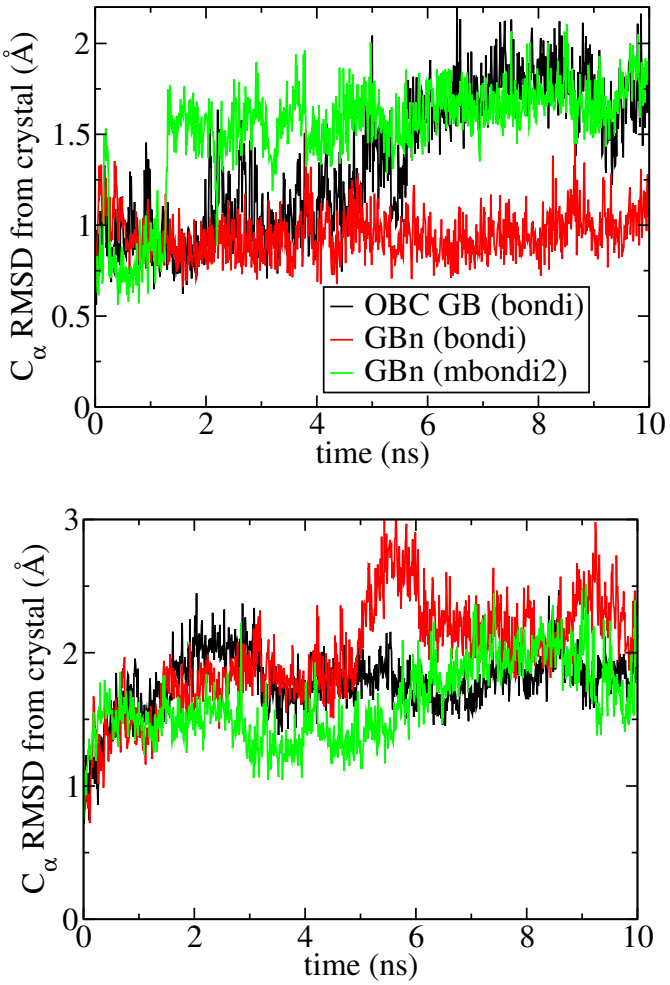

Figure 3.7: RMSD of α-carbons from crystal structure over the course of 10 ns of molecular dynamics of ubiquitin (left) and thioredoxin (right).

the thioredoxin trajectory and lower RMSD in the ubiquitin trajectory.

Performance of a GB model is affected by the set of atomic intrinsic radii used to define the dielectric boundary. Previous work has shown that for simulations conducted under the *HCT* or *OBC* GB models, structural stability is slightly increased and results are somewhat improved by increasing the intrinsic radius of hydrogens bound to nitrogen, H(N), from their Bondi radii[16] of 1.2 Å to 1.3 Å (forming the mbondi2 radius set).[41,47] As seen in figure 3.7, little benefit is realized by this change when using the *GBn* model.

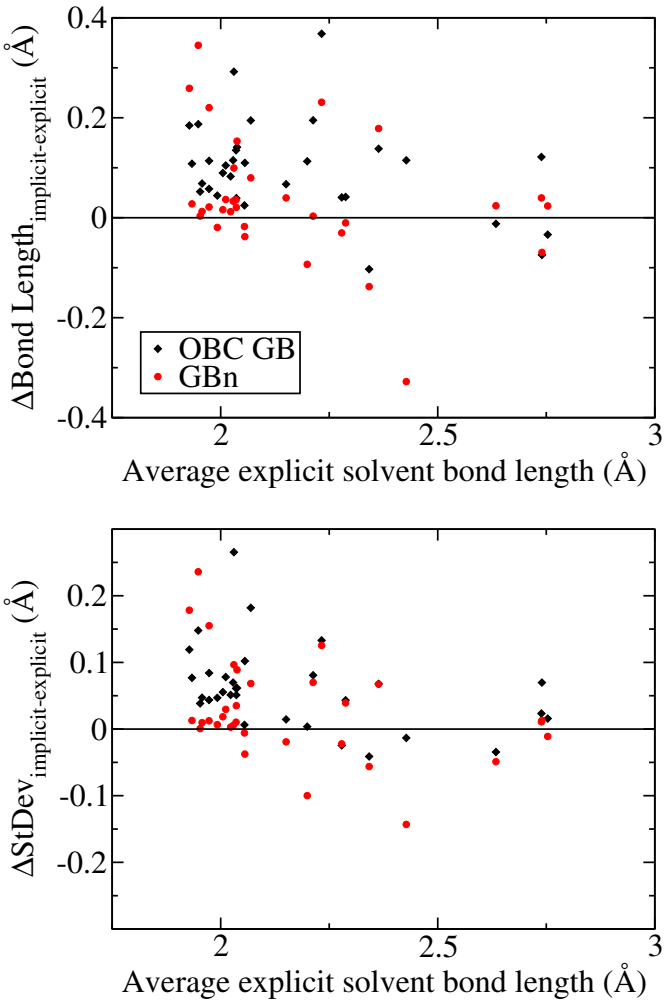

Figure 3.8: Ubiquitin backbone hydrogen bond length data collected over 10 ns of MD for TIP3P explicit solvent, *OBC* GB and *GBn*. Plots represent difference between implicit and explicit solvent bond length distribution mean (left) and standard deviation (right) as a function of mean explicit solvent bond length. Hydrogen bond lengths under the *GBn* model are generally closer to the zero line and thus in better agreement with explicit solvent results.

To examine whether the improved PMFs seen in figures 3.5 and 3.6 translate into improvements in the ensemble of macromlecular conformations sampled during MD, distributions of hydrogen bond lengths were compared between 10 ns ubiquitin trajectories conducted under *OBC* GB, *GBn* and TIP3P explicit solvation models. Figure 3.8 illustrates the differences in mean and standard deviation of hydrogen bond length for native backbone hydrogen bonds under the three solvation models. In nearly all cases, the *OBC* GB model yields hydrogen bonds with a higher mean length and standard deviation than in explicit solvent. As a consequence of the narrower potential wells seen in the PMFs, hydrogen bonds under the *GBn* model are generally shorter and their length distributions have lower standard deviations in better agreement with explicit solvent results than *OBC* GB. These differences are particularly noticeable for the shorter, more stable hydrogen bonds (left portions of the plots in figure 3.8), where length distributions are presumably mostly determined by the potential between bonding partners, while distributions for longer hydrogen bonds may be more affected by tertiary structural forces. The data in figure 3.8 suggest that the free energy barrier introduced by the neck correction affects not only dynamics and kinetic properties, but also average properties of the ensemble sampled by MD.

3.4 Methods

PB solvation energies and "perfect" radii were calculated using a modified version of APBS 0.3.2. The linearized PB model was employed along with the multiple Debye-Huckel boundary condition. Charge was discretized using the cubic B-spline method (spl2). Dielectric values were 1.0 for solute and 80.0 for solvent regions, except for "perfect" radii calculations, where solvent had dielectric 1000. A Lee-Richards type dielectric boundary (mol) was used. APBS versions 0.3.2 and earlier have a flawed molecular surface algorithm that overestimates solute volume; this flaw was fixed in the APBS version used here. All calculations were performed initially on a coarse grid and then on a smaller, finer grid using the coarse grid potential as boundary conditions. Grid spacings were 0.5/0.25 (coarse/fine) for protein solvation and perfect radii calculations

and 0.2/0.1 for PMF calculations.

GB effective radius, solvation energy and MD trajectories were calculated using a pre-release version of AMBER 9.[58] MD was carried out using the AMBER ff99 force field modified by the backbone torsional potentials in frcmod.mod_phipsi.1.[44] The timestep was 2 fs. Electrostatics were calculated using the neck GB model described here with a salt concentration of 0.1 M and no cut off. Non-polar solvation effects were represented using a surface area term of 0.005 kcal/mol·Å^2. Bonds involving hydrogen were constrained using SHAKE. Temperature was maintained at 300K using the Berendsen weak coupling method and a time constant of 2 ps for the thioredoxin trajectory and using Langevin dynamics with a Langevin constant of 1.0 for the ubiquitin trajectory. The crystal structures (2TRX and UBQ) were prepared for dynamics with 100 steps of steepest descent minimization during which all atoms were harmonically restrained with a weight of 1.0 kcal/mol·Å^2, followed by a 20 ps period of equilibration during which all atoms were harmonically restrained with a weight of 0.1 kcal/mol·Å^2.

The S_{neck}, α, β, γ and S_x parameters were optimized using the Nelder-Mead simplex algorithm[57] implemented by the SciPy library.[59] The objective function that was minimized measured agreement between PB and GB solvation free energies over a training set consisting of structures from denaturation trajectories of apo-myoglobin[60] and protein L[36] and structures representing varying degrees of separation of a salt bridge between aspartate and arginine and hydrogen bonds between two asparagine side chains and between serine and aspartate. The total value of the objective function was the sum of each system's contribution. For the structures from the denaturation trajectories, the difference between PB and GB solvation free energy was calculated for each structure and a linear regression was performed on these data points using the structure's time value (for apo myogloblin) or number of native tertiary contacts (for protein L) as the independent variable, yielding a regression line slope, *m*, and intercept, *b*. Additionally, the root mean square deviation (RMSD) between PB and GB solvation free energies for each structure was calculated. Each system's contribution to the objective function was defined as $RMSD - \left|\frac{b}{2}\right| + |m \cdot (\#\,of\,structures)|$. This term is designed to emphasize minimizing native state bias (represented by *m*) and random error while not overly penalizing systematic error for a particular system. Salt bridge and hydrogen bond systems

consisted of 80 configurations where the bonding partners were separated by 1 Å in the first configuration and are moved 0.1 Å further apart in each subsequent configuration (see figure 3.6 for picture of orientiations). PB and GB solvation free energies were calculated for each configuration, and the PB and GB solvation free energies were set to be equal at maximum separation by subtracting the energy calculated for maximal separation from that calculated for every other configuration. The objective function term for these systems was the RMSD of the adjusted errors multiplied by 10. The RMSD was increased by a factor of 10 to prevent the objective function from being dominated by the larger absolute errors of the larger protein systems. Since the objective function has multiple local minima, 100 minimizations were performed starting from random initial points. Initial points were chosen from the following intervals of a uniform random distribution: $S_{neck} \in [0.2, 0.5]$, $\alpha \in [0.5, 1.5]$, $\beta, \gamma \in [0.5, 3.0]$, and $S_{\{C,H,N,O\}} \in [0.6, 0.95]$.

3.5 Conclusion

The *GBn* model, presented here, extends current pairwise GB models with an intuitively attractive property: exclusion of high dielectric (representing water) from regions into which a water molecule is too large to fit. This extension is computationally efficient, slowing MD trajectories by only about 10%. Implementation of the neck correction is simple, requiring only a lookup table and (in the present implementation) approximately 30 lines of code. The neck GB model described here will be available in version 9 of the AMBER suite, and given its simplicity it should be straightforward to add the neck correction to any pairwise volumetric integration-based GB method. Although the correction is a pairwise approximation, it yields non-bonded PMFs with a free energy barrier to separation, a property unique to molecular surface-like dielectric boundaries. Additionally, non-bonded interactions have equilibrium properties in improved agreement with explicit solvent in protein molecular dynamics simulations conducted under the neck GB model.

The neck GB model is the fastest model that reproduces the essential characteristics of molecular surface dielectric boundaries, but it does not correlate as well with PB

results as the slower GBMV2 model of Lee *et al.*[5,56] One potential source of error is the fairly simplistic treatment of neck region overlaps in the current model. Some improvement might be realized by a higher order approach to overlaps, but the largest source of error appears to be the use of the CFA to define the integral used to calculate effective radii. Even with a perfect region of integration, errors due to the CFA are large, with effective radii overestimated by a factor of two in the worst case.[53] Despite the limitations imposed by the CFA, the current model serves as a proof of principle that a simple pairwise correction can produce an accurate approximation of molecular surface-like solvation properties. It is anticipated that a pairwise GB model based on the neck correction and a non-CFA integral, currently under development, will yield substantially improved accuracy.

3.6 Appendix: Neck region integrals

The neck region can be analytically defined using only basic trigonometry, but as the derivation is somewhat tedious, the details are provided here. As shown in figure 3.1, a triangle is formed by the centers of the atoms and the solvent molecule; the angles of the vertices centered at atoms 1 and 2 are defined as angle A and angle B, and their cosines can be expressed in terms of the four parameters d, R_1, R_2 and R_w, using the law of cosines, as shown in equation 3.8.

$$\cos A = \frac{d^2 + (R_1 + R_w)^2 - (R_2 + R_w)^2}{2d\,(R_1 + R_w)} \quad \cos B = \frac{d^2 - (R_1 + R_w)^2 + (R_2 + R_w)^2}{2d\,(R_2 + R_w)} \tag{3.8}$$

The system is cylindrically symmetric about an axis connecting the centers of the two atoms, so it is most naturally analyzed in cylindrical coordinates. The origin is placed at the center of atom 1 with the positive z axis extending toward the center of atom 2. There are three geometric cases for the neck region, illustrated in figure 3.9: (i) the atoms overlap and the neck region is ring shaped; (ii) the atoms are moderately separated forming a contiguous region; (iii) the atoms are widely separated such that the surface of the solvent molecule intersects the z axis, forming two discontinuous spike

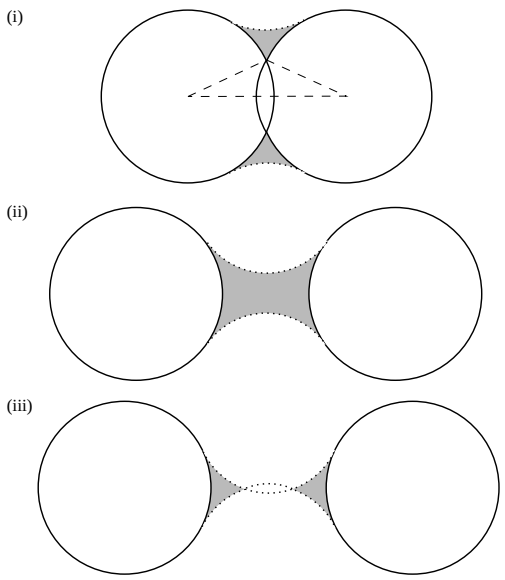

Figure 3.9: Three cases of neck regions (shaded) formed by atoms (solid circles) at varying separations. Dotted lines represent the surface of the solvent sphere. The leftmost vertex of the dashed triangle in (i) describes the angle A' referenced in equation 3.9. Although this figure shows two atoms with the same radius, neck regions may also be formed between atoms with unequal radii.

regions. When $d \geq R_1 + R_2 + 2R_w$ a solvent molecule can pass between the atoms and there is no neck region.

For case (i), a second triangle can be formed between the centers of the two atoms and a point at which the surfaces of the atoms intersect. The angle formed by the vertex of this triangle that is located at the center of atom 1 is designated A' and its cosine is defined in equation 3.9.

$$\cos A' = \frac{d^2 + R_1^2 - R_2^2}{2dR_1} \tag{3.9}$$

Computation of a CFA term based on the neck region requires an expression for the integral of $\mathbf{r}^{-4}$ over the neck region. In the cylindrical coordinate system used here, $|\mathbf{r}| = \sqrt{r^2 + z^2}$ so when the volume element is included, the integrand becomes $r\left(r^2 + z^2\right)^{-2}$. Because of the cylindrical symmetry, the limits of integration over θ are always 0 to 2π. The upper limit of the integration over r is formed by the surface of the solvent molecule (dotted line in figures 3.1 and 3.9). As the expression defining the r coordinate of the

solvent surface as a function of z (that is, the perpendicular distance from the z axis to the dashed line in figure 3.1 for a given z) is somewhat complex, the notation is clarified by defining a function, solv, representing this expression:

$$\text{solv}(z, R_1, R_2, R_w, d) = (R_1 + R_w)\sqrt{1-\cos^2 A} - \sqrt{R_w^2 - (z - (R_1 + R_w)\cos A)^2} \tag{3.10}$$

The lower limit of integration for r is defined by the surface of atom 1, the z axis ($r = 0$), or the surface of atom 2, depending on the value of the z coordinate. In addition to defining the geometric extents of the neck region, the limits of integration over z are used to break the overall integral into pieces at the points where the r lower limit of integration changes. Thus in case (i) the integral has two contiguous pieces defined by three z limits: the coordinate at which the solvent molecule touches atom 1, the coordinate for the intersection of the two atoms and the coordinate where atom 2 touches the solvent molecule. Case (ii) has three contiguous pieces, with the extreme upper and lower limits defined by the locations that the solvent molecule touches the atoms, as in (i) and the two intermediate limits occurring where the lower r limit changes at the edges of atoms 1 and 2. Finally, case (iii) has two discontiguous spike regions, each of which is composed of two parts, where the z limits are the intersection of the atom and solvent molecule, the edge of the atom and the tip of the spike. The tips of the spikes are located at the two points where the solvent sphere intersects the z axis (see figure 3.9). The z coordinate of these intersections can be obtained by setting the function in equation 3.10 equal to zero and solving for z, yielding

$$\text{zinter}_{(-)}(R_1, R_2, R_w, d) = (R_1 + R_w)\cos A - \sqrt{R_1(R_1 + 2R_w)(-1 + \cos^2 A) + R_w^2 \cos^2 A} \tag{3.11}$$

$$\text{zinter}_{(+)}(R_1, R_2, R_w, d) = (R_1 + R_w)\cos A + \sqrt{R_1(R_1 + 2R_w)(-1 + \cos^2 A) + R_w^2 \cos^2 A} \tag{3.12}$$

Using the preceding definitions, the integrals of $\mathbf{r}^{-4}$ over the neck region for cases (i), (ii) and (iii) are presented in equations 3.13, 3.14 and 3.15.

$$\text{(i)}: \int_{neck\,region} \mathbf{r}^{-4} = \begin{array}{l} \int_{R_1\cos A}^{R_1\cos A'} \int_0^{2\pi} \int_{\sqrt{R_1^2 - z^2}}^{\text{solv}(z, R_1, R_2, R_w, d)} r\left(r^2 + z^2\right)^{-2} dr\, d\theta\, dz \\ + \int_{R_1\cos A'}^{d - R_2\cos B} \int_0^{2\pi} \int_{\sqrt{R_2^2 - (d-z)^2}}^{\text{solv}(z, R_1, R_2, R_w, d)} r\left(r^2 + z^2\right)^{-2} dr\, d\theta\, dz \end{array} \tag{3.13}$$

$$\begin{aligned}(\text{ii}):\int_{neck\,region}\mathbf{r}^{-4}= & \int_{R_1\cos A}^{R_1}\int_0^{2\pi}\int_{\sqrt{R_1^2-z^2}}^{\text{solv}(z,R_1,R_2,R_w,d)} r\left(r^2+z^2\right)^{-2}\,dr\,d\theta\,dz\\ & +\int_{R_1}^{d-R_2}\int_0^{2\pi}\int_0^{\text{solv}(z,R_1,R_2,R_w,d)} r\left(r^2+z^2\right)^{-2}\,dr\,d\theta\,dz\\ & +\int_{d-R_2}^{d-R_2\cos B}\int_0^{2\pi}\int_{\sqrt{R_2^2-(d-z)^2}}^{\text{solv}(z,R_1,R_2,R_w,d)} r\left(r^2+z^2\right)^{-2}\,dr\,d\theta\,dz\end{aligned} \tag{3.14}$$

$$\begin{aligned}(\text{iii}):\int_{neck\,region}\mathbf{r}^{-4}= & \int_{R_1\cos A}^{R_1}\int_0^{2\pi}\int_{\sqrt{R_1^2-z^2}}^{\text{solv}(z,R_1,R_2,R_w,d)} r\left(r^2+z^2\right)^{-2}\,dr\,d\theta\,dz\\ & +\int_{R_1}^{\text{zinter}_{(-)}(R_1,R_2,R_w,d)}\int_0^{2\pi}\int_0^{\text{solv}(z,R_1,R_2,R_w,d)} r\left(r^2+z^2\right)^{-2}\,dr\,d\theta\,dz\\ & +\int_{\text{zinter}_{(+)}(R_1,R_2,R_w,d)}^{d-R_2}\int_0^{2\pi}\int_0^{\text{solv}(z,R_1,R_2,R_w,d)} r\left(r^2+z^2\right)^{-2}\,dr\,d\theta\,dz\\ & +\int_{d-R_2}^{d-R_2\cos B}\int_0^{2\pi}\int_{\sqrt{R_2^2-(d-z)^2}}^{\text{solv}(z,R_1,R_2,R_w,d)} r\left(r^2+z^2\right)^{-2}\,dr\,d\theta\,dz\end{aligned} \tag{3.15}$$

3.7 Appendix: Coordinates of neck integral maxima

Table 3.2: Distance between atoms at which integral of $\mathbf{r}^{-4}$ over the neck region (defined in equations 3.13-3.15) has the maximum value, tabulated for a range of radii for atoms 1 and 2, assuming a solvent molecule (R_w) radius of 1.4 Å. These are the values used for d_0 in equations 3.6 and 3.7. Distances and atom radii in angstroms.

Atom 1 \ *Atom* 2	1.20	1.25	1.30	1.35	1.40	1.45	1.50
1.20	2.6797	2.7250	2.7719	2.8188	2.8656	2.9125	2.9609
1.25	2.7359	2.7813	2.8281	2.8750	2.9219	2.9688	3.0156
1.30	2.7922	2.8375	2.8844	2.9297	2.9766	3.0234	3.0719
1.35	2.8500	2.8953	2.9406	2.9859	3.0328	3.0797	3.1266
1.40	2.9062	2.9516	2.9969	3.0422	3.0891	3.1359	3.1828
1.45	2.9625	3.0078	3.0531	3.0984	3.1437	3.1906	3.2375
1.50	3.0188	3.0641	3.1078	3.1547	3.2000	3.2469	3.2922
1.55	3.0750	3.1203	3.1641	3.2094	3.2563	3.3016	3.3484
1.60	3.1313	3.1750	3.2203	3.2656	3.3109	3.3563	3.4031
1.65	3.1875	3.2313	3.2766	3.3203	3.3656	3.4125	3.4578
1.70	3.2437	3.2875	3.3313	3.3766	3.4219	3.4672	3.5125
1.75	3.3000	3.3422	3.3875	3.4312	3.4766	3.5219	3.5688
1.80	3.3547	3.3984	3.4422	3.4875	3.5313	3.5766	3.6234

Atom 1 \ *Atom* 2	1.55	1.60	1.65	1.70	1.75	1.80
1.20	3.0078	3.0562	3.1047	3.1531	3.2016	3.2500
1.25	3.0641	3.1109	3.1594	3.2078	3.2563	3.3047
1.30	3.1188	3.1672	3.2141	3.2625	3.3109	3.3594
1.35	3.1750	3.2219	3.2703	3.3172	3.3656	3.4141
1.40	3.2297	3.2766	3.3250	3.3719	3.4203	3.4688
1.45	3.2844	3.3313	3.3797	3.4266	3.4750	3.5234
1.50	3.3391	3.3875	3.4344	3.4813	3.5297	3.5781
1.55	3.3953	3.4422	3.4891	3.5359	3.5844	3.6313
1.60	3.4500	3.4969	3.5438	3.5906	3.6391	3.6859
1.65	3.5047	3.5516	3.5984	3.6453	3.6922	3.7406
1.70	3.5594	3.6063	3.6531	3.7000	3.7469	3.7953
1.75	3.6141	3.6609	3.7078	3.7547	3.8016	3.8484
1.80	3.6688	3.7156	3.7625	3.8094	3.8563	3.9031

Table 3.3: Maximum value of integral of $\mathbf{r}^{-4}$ over the neck region (defined in equations 3.13-3.15), tabulated for a range of radii for atoms 1 and 2, assuming a solvent molecule (R_w) radius of 1.4 Å. These are the values used for m_0 in equations 3.6 and 3.7. Distances and atom radii in angstroms.

Atom 1 \ *Atom* 2	1.20	1.25	1.30	1.35	1.40	1.45	1.50
1.20	0.35281	0.36412	0.37516	0.38594	0.39645	0.40670	0.41670
1.25	0.31853	0.32889	0.33902	0.34890	0.35855	0.36797	0.37717
1.30	0.28847	0.29798	0.30728	0.31637	0.32525	0.33392	0.34240
1.35	0.26199	0.27074	0.27930	0.28768	0.29587	0.30387	0.31170
1.40	0.23859	0.24666	0.25455	0.26228	0.26985	0.27725	0.28449
1.45	0.21783	0.22528	0.23258	0.23972	0.24673	0.25358	0.26029
1.50	0.19935	0.20624	0.21300	0.21962	0.22611	0.23247	0.23870
1.55	0.18285	0.18923	0.19550	0.20165	0.20767	0.21358	0.21938
1.60	0.16807	0.17400	0.17982	0.18553	0.19114	0.19664	0.20203
1.65	0.15480	0.16031	0.16573	0.17104	0.17626	0.18139	0.18642
1.70	0.14285	0.14798	0.15303	0.15798	0.16285	0.16764	0.17233
1.75	0.13207	0.13685	0.14155	0.14618	0.15073	0.15520	0.15959
1.80	0.12231	0.12677	0.13117	0.13549	0.13975	0.14393	0.14804

Atom 1 \ *Atom* 2	1.55	1.60	1.65	1.70	1.75	1.80
1.20	0.42646	0.43598	0.44527	0.45434	0.46319	0.47183
1.25	0.38615	0.39492	0.40348	0.41185	0.42001	0.42799
1.30	0.35069	0.35878	0.36669	0.37441	0.38196	0.38934
1.35	0.31936	0.32684	0.33416	0.34131	0.3483	0.35514
1.40	0.29158	0.29851	0.30529	0.31193	0.31842	0.32477
1.45	0.26686	0.27330	0.27959	0.28575	0.29179	0.29769
1.50	0.24480	0.25078	0.25664	0.26237	0.26799	0.27349
1.55	0.22505	0.23062	0.23607	0.24141	0.24665	0.25178
1.60	0.20732	0.21251	0.21759	0.22258	0.22747	0.23226
1.65	0.19135	0.19620	0.20095	0.20561	0.21018	0.21466
1.70	0.17694	0.18147	0.18591	0.19027	0.19455	0.19875
1.75	0.16390	0.16814	0.17230	0.17638	0.18039	0.18433
1.80	0.15208	0.15605	0.15995	0.16378	0.16754	0.17124

This chapter is a preprint in full of *Generalized Born with a simple, robust molecular volume correction.* John Mongan, Carlos Simmerling, J. Andrew McCammon, David A. Case and Alexey Onufriev. Submitted to Proteins: Structure, Function and Bioinformatics. I was the primary researcher and author of this work.

Chapter 4

Biomolecular simulations at constant pH

ABSTRACT

Like temperature and pressure, the solution pH is an important intensive thermodynamic variable that is commonly varied in experiments, and is used by cells to influence biochemical function. It is now becoming feasible to carry out practical molecular dynamics simulations that mimic the thermodynamics of such experiments, by allowing proton transfer between the system of interest and a hypothetical bath of protons at a given pH. These are demanding calculations, both because the energetics of charge changes upon protonation or deprotonation must be accurately modeled, and because such simulations must sample both molecular configurations and the large number of protonation states that are possible in a molecule with many titrating sites. Here, the history of these ideas and recent progress in meeting such challenges are discussed, looking at the design of algorithms and approximations that allow one to overcome some of their intrinsic difficulties.

4.1 Introduction

Solution acidity is an important thermodynamic variable that can affect biochemical function in a way that is often as profound as that of temperature or of the concentration of other allosteric effectors such as cofactors or phosphates. Many *in vitro* experiments mimic cellular compartments by regulating pH closely, commonly with buffering agents. The experimental study of titration behavior and the response of biomolecules to changes in pH has a long history, and there is a large amount known about the thermodynamics of proton binding.[61,62] Structural correlations are less well-developed, but are becoming of increasing interest as methods for monitoring site-specfic proton binding (particularly by NMR) become more routine.

There is also a long history of theoretical and computational approaches to study of behavior of proteins and nucleic acids as a function of solution acidity.[22,63–65] This is known to be a difficult problem, since almost all biomolecules have multiple sites that can bind or release protons, and these are coupled to one another in complex ways. In recent years, however, increases in computational power and new models for estimating the energetics of protonation/deprotonation events have led a number of investigators to seriously attempt simulations that allow the solution pH to be specified as an external variable in a manner that parallels the ways in which temperature or pressure are specified. This chapter outlines their history and current prospects.

4.2 Calculations of individual pK_a values in proteins

4.2.1 Thermodynamic integration and other free energy methods

In principle, the most rigorous way to estimate an individual pK_a value for a protein side-chain would involve a free energy simulation connecting the protonated and deprotonated forms of the molecule:

$$\mathrm{pK_a} = -\log_{10} \mathrm{K_a} = \frac{\Delta G}{k_B T \ln 10} \tag{4.1}$$

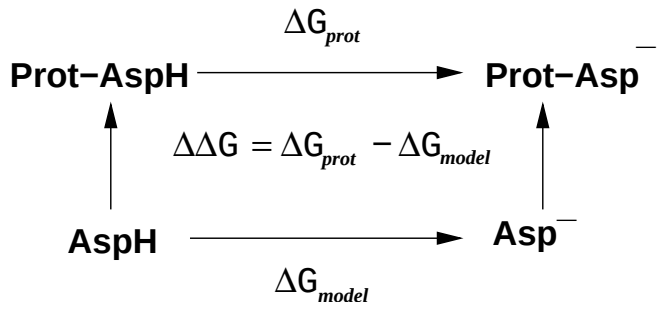

Figure 4.1: Thermodynamic cycle for calculation of protein sidechain pK_a values

In a molecular mechanics approach (where covalent bonds cannot be broken) this in practice would involve parallel, explicit solvent simulations on the protein of interest and on a model compound with the same functionality and with a known pK_a. The computed pK_a difference can then be added to the known model compound value to estimate the macromolecular result, as shown in figure 4.1. This model effectively assumes that energetic contributions outside the molecular mechanics model (such as the strength of the O—H chemical bond) are the same in the protein as in the model compound.

Given the importance of the problem, the large amount of experimental data that is available, and the relative maturity of free-energy methods in molecular simulations, it is surprising that this approach has not received more attention. The basic ideas are all present in some of the earliest free-energy simulations on proteins,[66] which anticipated many future developments, but which dealt with only a small piece of a protein and 80 water molecules. The computational cost of molecular dynamics free energy simulations is partly to blame for this neglect. Further, the use of cut-off schemes to truncate long-range electrostatic interactions, (which were in fairly common use up to a few years ago), gives generally unreliable results for perturbations that involve a net change in charge.[67] The adoption of periodic conditions and Ewald summations (or related methods) avoids this truncation. It has the disadvantage of introducing an artificial periodicity into the model,[68] and there are some subtle, and still controversial, questions arising about what boundary conditions to use in systems with a net charge. Nevertheless, simulations of the charging free energies of ions[67,69] have shown that periodic simulations can converge quickly (as a function of the size of the periodic unit

cell) to results fully consistent with models based on non-periodic approaches that treat long-range effects using a reaction field. Since such periodic simulations are now implemented in an efficient and parallel fashion in many popular simulation codes, the time seems right to re-evaluate what behavior should be expected for pK_a free enegy simulations.

Two examples of what to expect come from early studies of succinic acid[70] (which illustrates an ambitious method for handling multiple-site problems), and more recent work on thioredoxin and ribonuclease.[71] The protein results give reasonable agreement with experiment (for both explicit solvent and generalized Born simulations), but also show that protein response to a change in charge can take a long time (at least nanoseconds) to occur. Figure 4.2 shows the distribution of the energy difference between the protonated and deprotonated forms for simulations where the charge model interpolates between the neutral ($\lambda = 0$) and anionic ($\lambda = 1$) forms for a carboxylic acid. A system that follows linear response theory would show a Gaussian distribution.[72,73] The simulations obey this model closely at λ of 0.5 and 0.89, but near neutrality (at $\lambda = 0.11$) the energy gap fluctuations are no longer Gaussian: there are two populated subconformers (correpsonding to two populated sidechain orientations for ASP-26). This causes no difficulties for the free energy simulations, as long as all such conformational substates are adequately sampled, but illustrates a limitation for linear response models. As faster computers make free energy simulations more accessible, they may be expected to become the norm for careful studies.

4.2.2 Implicit solvent models using the Poisson-Boltzmann approach

The molecular dynamics free energy methods described above have a rigorous basis (within the limits of the force field being used), but require large amounts of computer time and are not readily adapted to handling the large numbers of titrating sites found in most proteins. The great majority of computational studies on protein titration behavior have thus used more simplified models, most commonly by treating the bulk aqueous environment as a continuum dielectric; the effects of bulk salt can also be treated in a continuum fashion, leading to the Poisson-Boltzmann equation for the electrostatic

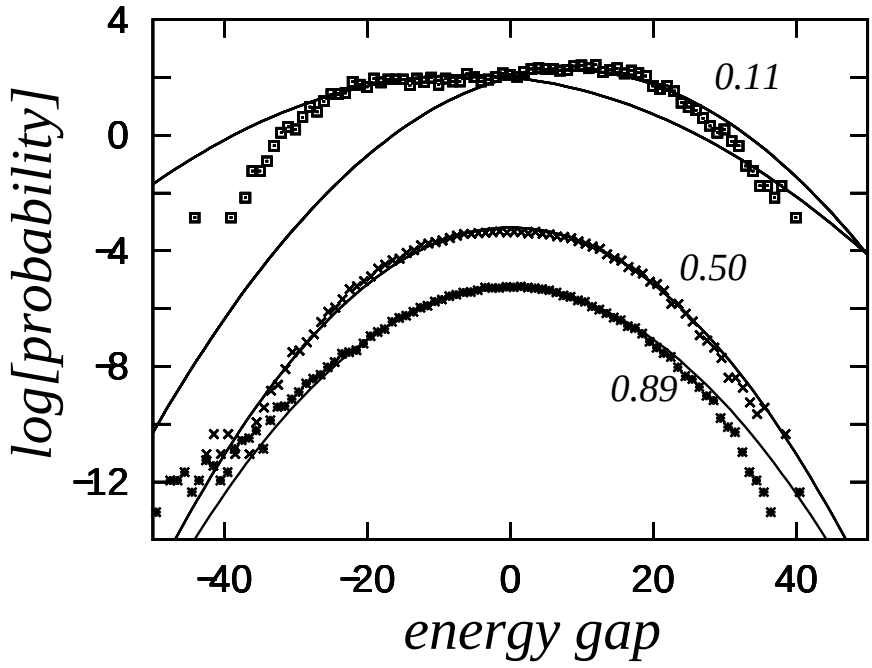

Figure 4.2: Probability profile for the enegy gap (the energy difference between the protonated and deprotonated forms, in kcal/mol) of ASP-26 in thioredoxin. Values of λ interpolate between the neutral form at $\lambda = 0$ and the ionized form at $\lambda = 1$. A Gaussian distribution would look like an inverted parabola in this semi-log plot. Data taken from Simonson *et al.*[71]

field.[25,64,74,75] The application of such a continuum model to the *entire* solvent region (including water molecules in direct contact with a solute) seems like a severe approximation, but actually gives quite a good account of solvation free energies and pK_a behavior in small, fairly rigid molecules.[76–80] This is undoubtedly due in part to a careful parameterization of the boundary between the solvent and solute, so that the average energetic consequences of even first-shell waters are incorporated into the continuum model. Models for this dividing surface can be made that appear to be transferable, and not overly dependent on the detailed chemical nature of the solute. (This effective boundary varies with temperature in a way that is not easy to model; for this reason, continuum models are much less successful in predicting quantities like solvation enthalpies and entropies.)

The extension of these ideas to proteins has a long history,[64,81] but the key problem

remains of how to describe the response of the protein itself to a change in protonation. The simplest model describes the protein interior itself as a continuous media with low dielectric constant.[82] This is a much more drastic physical assumption than treating water as a continuous dielectric medium: water molecules are very small and can reorient fairly easily, so that a representation as a continous dipole density seems appealing. Proteins, on the other hand, have a heterogeneous distribution of charges and dipoles, fluctuating about a native structure that may depend upon sidechain protonation states. In principle, a "protein dielectric constant" could provide a representation of the averaged response of its charges and dipoles to electrostatic fields, but there is no unique way to obtain the best value, and microscopic simulations usually only increase one's intuitive reservations about how such a simple model could possibly capture what must be a very complex set of interactions.[74,83]

There have been a variety of approaches to make more realistic models without going all the way to detailed simulations. The "protein-dipole Langevin dipole" (PDLD) model of Warshel and co-workers treats the protein as a set of particles, each bearing a charge and a polarizable dipole, and the nearby solvent as a lattice of orientable Langevin dipoles.[83–86] This procedure uses a reasonable model for flexibility in the protein and the nearby solvent (the distant, bulk solvent is still treated as a dielectric continuum), yet permits the efficient calculation of protonation thermodynamics. Rather than using polarizable dipoles for the protein, others have explicitly modeled certain types of motions, such as sidechain rotameric changes[22] or shifts among possible positions for the protons.[87–89] One can also carry out some averaging over protein configurations by applying any of these methods to snapshots from an explicit-solvent molecular dynamics simulation;[85,90,91] such combinations of MD simulation with periodic electrostatic analysis lead naturally to the *constant pH* approaches discussed below.

4.3 Constant pH simulations

The discussion so far has considered the thermodynamics of protonation of a single titrating site, where it makes sense to define a pK_a value as in Eq. 4.1, and the fraction

of sites that have a proton bound can be trivially calculated once K_a is known. Proteins, of course, have multiple titrating sites, and one must consider the relative energetics of different protonation possibilities, and then the statistics of thermal ensembles over the possible states. A rough "pK_a" can then be identified as the pH value at which the populations of the protonated and deprotonated forms are equal. In extreme cases, however, this may not be well-defined or unique, and it is really the properties of the entire titration curve that are of greatest interest.[61,65,92,93] Generally, proton binding is slightly anti-cooperative (because repulsive charge-charge interactions grow as more protons bind), but significant excursions from "normal" behavior can occur, either as the result of strong coupled interactions between nearby sites, or as a result of pH-dependent changes, such as global unfolding at low or high pH.

There are two main ways in which the multiple-site problem can be addressed. One approach makes a mean-field (or Tanford-Roxby) approximation, in which sites interact not with the particular protonation states of other sites, but with their (pH-dependent) average protonation; in effect, sidechains with a pK_a value near the solution pH will be represented with charge distributions that are intermediate between the protonated and deprotonated forms.[64,94] This is an appealing approximation, which forms the basis for the "continuous" constant pH models discussed below, but can break down when strongly coupled sites titrate in the same pH range.[95]

For a small number of titrating sites, it is feasible to avoid the mean-field approach, and to compute the complete partition function over all possible protonation states. For larger numbers of protonation states, one can use Monte Carlo (MC) sampling to estimate thermal averages.[64,96] This idea forms the basis for the "discrete protonation" constant pH models.

In general, protonation state sampling in all of these methods is driven by an equation describing the free energy of a protonation state transition

$$\Delta G = \Delta G_{MM} + \mathrm{pH}\, k_B T \ln 10 + \Delta G_{QM,\, proton\, solv} \tag{4.2}$$

where the first term describes the free energy change within the molecular mechanics force field, the second incorporates the pH dependence through the chemical potential of the proton bath and the third is term is a constant based in part on experimental data

that accounts for the proton solvation and quantum mechanical bond free energies. In general, ΔG_{MM} must be averaged over the solvent degrees of freedom—in many models this averaging is accomplished by use of implicit solvent.

4.3.1 Continuous protonation states with explicit solvent

The earliest published work on constant pH MD was conducted by Mertz and Pettitt.[97] They describe an explicit solvent, grand canonical MD method in which a reaction extent parameter, ξ, interpolates continuously between two Hamiltonians describing the protonated and deprotonated states. An equation of motion is derived for ξ. The adjustable parameter in this model is not the pH directly, but rather the difference in chemical potential between the reservoir representing the reactants (e.g. protonated solute and water) and the reservoir representing the products (e.g. deprotonated solute and hydronium). The pH of a simulation is calculated after the fact based on the average value of ξ and the known K_a for the system. Unique to this model is the explicit representation of free protons as hydronium ions (H_3O^+). This has the benefit of keeping the total charge constant between the two endpoint Hamiltonians. On the other hand, since each titratable group is coupled to a specific hydronium ion, changing the protonation state (changing the value of ξ) involves transferring charge from the titratable group to wherever the hydronium ion may be. This may not be a problem in simple systems, such as the application to acetic acid presented by Mertz and Pettitt, but in larger systems the non-physical dependence between protonation extent of a titratable group and electrostatic environment of the coupled hydronium ion may be expected to present a significant barrier to convergence.

A more recent development is the “acidostat” method of Börjesson and Hünenberger.[98] It involves continuous protonation states that are relaxed toward an equilibrium fractional protonation by weak coupling to a proton bath. This weak coupling is analogous to methods used for constant temperature and constant pressure MD. The method employs explicit solvation for MD as well as determining the equilibrium fractional protonation. Baptista has pointed out flaws in the theoretical basis of the acidostat method, and suggested that it involves an implicit mean field approximation.[99] Given this, he

suggests that the method is best evaluated empirically. Titration of a series of small amines using the acidostat method yields substantially correct fractional protonations at three different pH values, corresponding to predicted pK_a values within about 0.3 of experimental values. Fractional protonation is systematically higher than appropriate at low pH and lower than appropriate at high pH, indicating that titration curves produced under this method do not have the correct Henderson Hasselbalch shape. The authors attribute this to solvation differences between the model compounds and the titrated compounds, but this does not explain why the effect is still seen when the titrated compound is the same as the model compound. When the weak coupling time constant is set to be sufficiently long to prevent oscillation of fractional protonations (10 ps), it takes approximately 1 ns for methylamine to equilibrate to 10/11 protonation when starting from 1/11 protonation. Such slow equilibration may limit the ability of titratable groups in macromolecules to respond to changes in electrostatic environment, slowing convergence. In a second paper, Börjesson and Hünenberger explore titration of coupled sites under the acidostat method.[100] Titration of 1,4-diaminobutane from pH 8 to 12 produces a titration curve indicative of site-site correlation rather than one that would be expected from a mean field based method. These results were obtained by starting the simulations with the two titrating sites equilibrated to different fractional protonations based on apparent pK_a values from experiment. The appropriateness of equilibrating the two symmetry-related sites to different fractional protonations is questionable, given that their average fractional protonations should be the same. It is unclear whether or not results suggesting site-site correlations would have been obtained if both sites were started with the same fractional protonation. The method is also applied to polylysine, where helicity at a range of pH values shows good correlation with experimental observations.

4.3.2 Continuous protonation states with implicit solvent

A potential of mean force (PMF) based method, titled “Implicit Titration,” has been presented by Baptista and co-workers.[101] In implicit titration, MD is conducted using a force field based on a PMF averaged over the protonation state distribution appropriate for the selected pH. This averaging effectively produces a continuous representation

of the protonation state at each titratable site. The distribution of protonation states is calculated with a continuum electrostatics (CE) method, and since the distribution is dependent on the configuration of the system, it is recomputed at intervals throughout the MD. The implementation of this method that is applied to bovine pancreatic trypsin inhibitor (BPTI) employs rather crude CE and MD electrostatics (Tanford-Kirkwood and distance-dependent dielectric), but these could easily be substituted with more rigorous methods, including explicit solvent MD. The implementation that is presented also uses a mean field approximation. Baptista et al. discuss how this approximation could be eliminated, but under the proposed method, the partial charge used for calculating the Coulombic interaction of an atom in a titratable group may depend on the atom with which it is interacting. The implications for dynamics and analysis of having multiple partial charges simultaneously assigned to the same atom are unclear and not explored in the paper. Further difficulties with the method involve the use of CE to compute the distribution of protonation states used for the PMF. It is computationally infeasible to run the CE calculations at each MD step, so most MD steps occur under a PMF that is appropriate for a recent configuration in the trajectory, but not exactly correct for the current one. Also, since CE is used to calculate the distribution of protonation states, this distribution is determined only by the configuration of the solute, not the solvent, which is not entirely correct if the MD is conducted in explicit solvent.

A recent work by Lee *et al.*[102] describes a novel approach for avoiding mean field approximations in a continuous protonation state model, drawing on the ideas of λ-dynamics. In this constant pH model, a potential is constructed along a λ coordinate interpolating between the protonated and deprotonated states. The potential for each titratable group is based on a model compound's experimentally known pK_a and previously calculated PMF along the λ coordinate. As in the Mertz and Pettitt model,[97] equations of motion are used to propagate a fictitious particle along the λ coordinate. Mean field approximations are avoided by introducing an energetic barrier centered at $\lambda = 1/2$, which forces protonation states away from intermediate, mixed protonation states and toward values representing full protonation or deprotonation. Timesteps where titratable groups have intermediate protonation states are excluded from predicted pK_a calculations. Barrier height is a tunable parameter, trading off between protonation state transition rate

and fraction of simulation time spent in intermediate protonation states. Titration of the aspartate model compound produces a titration curve that is well fit by the Henderson Hasselbalch equation, but the curve is shifted 0.3 pK_a units above the model compound pK_a of 4.0. The authors attribute this error to poor convergence. Given that the points on the titration curve were taken from 10 ns simulations, this suggests that even very simple molecules may converge quite slowly under this model. Another limitation of the current implementation is that since protonation coordinates are interpolated between the extreme protonation states, accurate representation of titratable groups having more than two protonation states requires propagating the protonation state fictitious particle in higher dimensional spaces. An extension of the original method shows significant improvement with three-state protonation models of carboxylic acids and histidines implemented with a two dimensional protonation coordinate space[103] However, continuation of this approach to three or more dimensions seems to involve prohibitive difficulties with sampling of the protonation space and calculation of the model potential function. Details of the protonation state fictitious particle trajectories have not been published, but it is likely that there is significant oscillation of protonation states within the energy wells representing protonation and deprotonation. Convergence problems are apparent in application to proteins, where 1 ns titrations starting from different initial velocities produced predicted pK_a values differing by more than 1 pK_a unit for many residues. Despite the convergence issues, the model produces predicted pK_a values that are in good agreement with experimental values. The average absolute error is 1.6 for hen egg white lysozyme, 1.2 for turkey ovomucoid, 0.9 for bovine pancreatic trypsin inhibitor (BPTI). The published implementation of this method employs generalized Born (GB) solvent, but in principle the method could be used with explicit solvent. Explicit solvent would be expected to further slow convergence, though, so use of the current model with explicit solvent is probably not computationally feasible.

4.3.3 Discrete protonation states with explicit solvent

The method put forth by Bürgi *et al.* is notable for being the only published approach that uses MC sampling of discrete protonation states without using CE to determine tran-

sition energies.[104] Instead, transition energies are calculated using thermodynamic integration (TI) in explicit solvent. Use of a single, consistent explicit-solvent electrostatics model involves fewer approximations than CE based models and might be expected to yield superior results. In practice, restrictions imposed by the computational demands of explicit solvation limit the effectiveness of the model. To increase the number of MC trials that can be performed, TI occurs over short periods of approximately 20 ps of dynamics. The meaning of the free energy difference calculated this way is unclear: the period of TI is neither short enough that it represents the protonation state transition free energy of any single conformation, nor long enough that it represents the transition free energy sampled over the entire ensemble of conformations. Since so much computer time is expended on the TI, the trajectory is assembled by concatenating the TI segments, using the "forward" direction (current protonation state to proposed state) for MC steps that are accepted and the "reverse" direction for those that are rejected. This has the undesirable effect of perturbing the dynamics of a titratable group every time it is involved in an MC trial, even if the step is rejected. Bürgi et al tested the method by applying it to HEWL and comparing predicted pK_a values derived from their simulations to experimental values. Results were in rough qualitative agreement with experiment (*i.e.* pK_a shifts with respect to model compounds were generally in the right direction) but quantitative accuracy and precision were poor for most residues. Even after 3 ns of simulation, predicted pK_a values seem to be far from convergence, most likely because the computational expense of using TI in MC trials limits the number of trials that can be performed.

4.3.4 Discrete protonation states with explicit and implicit solvent

Some of the shortcomings of the implicit titration model were addressed by Baptista and co-workers in a second model, "stochastic titration."[105] In this method, discrete protonation states are represented by alternate sets of partial charges. Short (0.2 - 5.0 ps) segments of MD are conducted in explicit solvent. Between MD segments, protonation states are altered using Monte Carlo sampling based on protonation state energies for the current conformation taken from Poisson Boltzmann (PB) calculations. After a

protonation state change, the solvent is equilibrated by running a few picoseconds of MD with the solute conformation fixed. One drawback to this method is that there is no single, consistent Hamiltonian for the whole system, due to the use of explicit solvent MD and PB electrostatics for protonation state sampling. The method was successfully applied to the titration of succinic acid; but it is difficult to predict performance in titration of a protein based on results from such a simple system. As the authors point out, optimum values of parameters may be system dependent. In particular, the finding that results are insensitive to the length of the MD segment may not be true for proteins, which have far more complicated energy landscapes than succinic acid.

4.3.5 Discrete protonation states with implicit solvent

A method very similar to stochastic titration has been explored in a series of papers by Antosiewicz and co-workers.[106–109] They also use MC based on PB energies to sample over discrete protonation states, with MC steps separated by short MD segments. In an initial application to ovomucoid third domain,[106] they used Langevin dynamics with a uniform dielectric constant of 15 for MD and mean-field protonation state probabilities at each titratable site for MC sampling. This first implementation uses a complicated series of heating, cooling, and minimization phases during MD, so it is more in the spirit of using MD as a sampling method to produce an ensemble of conformations than for producing continuous trajectories. Titration curves for a small peptide closely match experimental results, and RMS error for pK_a predictions in ovomucoid third domain is 0.8-1.1 (depending on dielectric parameters). A revised implementation using the analytical continuum electrostatics (ACE) method of CHARMM for dynamics and MC sampling that eliminates the mean-field approximation has been applied to a heptapeptide derived from ovomucoid third domain[107] and succinic acid.[108] The temperature changes and minimization of the earlier implementation are eliminated, but MD is restarted with randomized velocities after each protonation state change. This implementation achieves reasonable agreement with experimental data for the peptide and an acylated derivative and close agreement with both experimental data and the earlier results of Baptista *et al.*[105] in application to succinic acid. Like Baptista *et al.*'s stochastic titration, the An-

tosiewicz methods do not have a consistent whole-system Hamiltonian, since different implicit solvent models are used for dynamics and protonation. Due to the computational expense of PB calculations, MC steps are relatively infrequent (every 1 to 5 ps) in these PB based methods, but still much more frequent than in Bürgi *et al.*'s[104] TI based constant pH.

Greater efficiency can be obtained by using the generalized Born (GB) model[27] for both the dynamics and the protonation state sampling.[110] This approach, described in detail in the following chapter, permits much more frequent Monte Carlo protonation steps; trials are made every 10 fs, so that on average, any given sidechain would be tested for a protonation change about ten times per picosecond. These frequent changes of the protonation state do not appear to adversely affect the stability of the simulation. Unlike the PB based methods, which sample over all relevant protonation states at each MC step, the current implementation changes the state of at most one titratable group on each MC step. This limitation may slow convergence for tightly coupled titratable groups. Titration of the aspartate model compound by a series of 1 ns simulations produces a very close fit to the expected Henderson Hasselbalch curve. The method was applied to lysozyme, where 1 ns trajectories at a range of pH values are less well converged but nevertheless yield pK_a values with only 0.82 RMS error with respect to experimental data. Side-chain titrations predicted for lysozyme after 1 ns of simulation are almost independent of which crystal structure was used as a starting point, as one would expect for a fully dynamical method. This is in contrast to models that use a single static structure, where results can vary widely depending upon which crystal structure is selected.[111]

Although constant pH MD can be used to predict pK_avalues, and these predictions are the most straightforward means of method validation, a major goal of constant pH is improved realism for biomolecular simulations. Many interesting biomolecular processes are pH-dependent phenomena, and the relevant protonation states are not known in advance. Even in systems where the pH dependence is less obvious, it has been shown that ensemble of conformations sampled in fixed protonation state MD is biased towards the selected protonation state.[112] In contrast, constant pH MD allows for sampling over the biologically more meaningful ensemble of conformations at fixed pH.

This chapter contains material that appeared in *Biomolecular simulations at constant pH*. John Mongan and David A. Case. Current Opinion in Structural Biology, **18**(2), 157-63, April 2005. I was the primary author of this work.

Chapter 5

Constant pH molecular dynamics in generalized Born implicit solvent

ABSTRACT

A new method is proposed for constant pH molecular dynamics (MD), employing generalized Born (GB) electrostatics. Protonation states are modeled with different charge sets, and titrating residues sample a Boltzmann distribution of protonation states as the simulation progresses, using Monte Carlo sampling based on GB derived energies. The method is applied to four different crystal structures of hen egg-white lysozyme (HEWL). pK_a predictions derived from the simulations have root mean square (RMS) error of 0.82 relative to experimental values. Similarity of results between the four crystal structures shows the method to be independent of starting crystal structure; this is in contrast to most electrostatics-only models. A strong correlation between conformation and protonation state is noted and quantitatively analyzed, emphasizing the importance of sampling protonation states in conjunction with dynamics.

Reproduced from *Constant pH Molecular Dynamics in Generalized Born Implicit Solvent*. John Mongan, David A. Case and J. Andrew McCammon. Journal of Computational Chemistry, **25**(16), 2038-48, December 2004.

5.1 Introduction

Protein structure and function are strongly dependent on solvent pH. This dependence is due to changes in the predominant protonation state of titratable groups (chiefly side chains of certain amino acids and termini of peptide chains) as solvent pH changes. The protonation state of a titratable group is determined by the solvent pH, and the relative acidity of the group, measured by its pK_a. The instantaneous pK_a of a given group is influenced by its electrostatic environment, which is determined by the protein conformation and protonation state of other titratable groups. Protonation state, in turn, has a strong effect on protein conformation, due principally to the charge differences between different protonation states.

Due to the tight coupling between protein conformation and protonation state, the importance of solvent pH in molecular dynamics (MD) simulations of proteins has long been recognized. Traditionally, treatment of pH in MD has been limited to setting a constant protonation state for each titratable group. This approach has many drawbacks. First, assigning protonation states requires knowledge of pK_a values for the protein's titratable groups. Second, if any of these pK_a values are near the solvent pH there may be no single protonation state that adequately represents the ensemble of protonation states appropriate at that pH. Finally, since the assumed protonation states are constant, this approach decouples the dynamic dependence of pK_a and protonation state on conformation.

The solution pH is an important extrinsic thermodynamic variable, analogous to temperature or pressure, that is readily controlled experimentally and has considerable spatial and temporal variation in living organisms. It is natural to seek simulation methods that allow the user to directly specify the pH as an input variable. In the past decade, a number of models have been proposed for performing MD at constant pH with dynamic protonation states. These methods have been reviewed in detail in the preceding chapter.

This chapter introduces a model using generalized Born (GB) implicit solvation[35] that combines the best aspects of discrete protonation state constant pH models. The same GB electrostatics are used for calculating protonation state transition energies and

dynamics, so the potentials are consistent. Furthermore, calculation of transition energies using GB is fast and there is no need for solvent equilibration, so sampling is fast. This model is tested using simulations of hen egg-white lysozyme, examining convergence, stability, agreement with experimental pK_a values and correlation between conformation and protonation. Close agreement between predicted and experimental pK_a values suggest that this method accurately samples protonation states, providing a more physically realistic basis for studying dynamics of systems with titratable groups.

5.2 Theory and Methods

5.2.1 Algorithm

The proposed method employs GB solvated MD, with periodic Monte Carlo sampling of protonation states. Between Monte Carlo steps, the system evolves according to standard generalized Born solvated MD.[51,113] This sampling scheme and the justification for it are essentially the same as those described by Baptista *et al.*,[105] with the exception that there is no solvent equilibration step since the MD is conducted in implicit solvent.

At each Monte Carlo step, a titratable site and a new protonation state for that site are randomly chosen. A transition free energy for the protonation or deprotonation is calculated according to

$$\Delta G = k_{\mathrm{B}} T \left(\mathrm{pH} - \mathrm{pK}_{\mathrm{a,ref}}\right) \ln 10 + \Delta G_{\mathrm{elec}} - \Delta G_{\mathrm{elec,ref}} \tag{5.1}$$

where k_B is the Boltzmann constant, T is temperature, pH is the specified solvent pH, $pK_{a,ref}$ is the pK_a of the appropriate reference compound (see section 5.2.4 and table 5.1), ΔG_{elec} is the electrostatic component of the free energy calculated for the titratable group in the protein, and $\Delta G_{\mathrm{elec,ref}}$ is the electrostatic component of the transition free energy for the reference compound, a free dipeptide amino acid described in section 5.2.4. This equation is based on a division of the total transition free energy into electrostatic and non-electrostatic portions. The non-electrostatic transition free energy comprises all free energy contributions not accounted for in the GB electrostatics, in-

cluding the quantum mechanical bond free energy and proton solvation free energy. It is difficult to calculate the non-electrostatic transition free energy, but it can be assumed to have approximately the same value independent of electrostatic environment. Under this assumption, a reference compound with known pK_a can be introduced to cancel the non-electrostatic portion of the transition free energy, resulting in equation 5.1. The electrostatic portion of the transition free energy (ΔG_{elec}) is calculated by taking the difference between the potential calculated with the charges for the current protonation state and the potential calculated with the charges for the proposed state; since there is no need for solvent equilibration, this is done in a single step. Equation 5.1 can then be used to calculate the total transition free energy, as all other terms are known. This method of calculating transition free energies is similar to that employed by Bürgi *et al.*,[104] except that in this model only charges change between different protonation states, while they change van der Waals radii as well. Changing van der Waals radii may be added in a further refinement of this model, but good results are seen with changing only charges.

The total transition free energy, ΔG, is used as the basis for applying the Metropolis criterion to determine whether the transition will be accepted. If the transition is accepted, MD is continued with the titratable group in the new protonation state; if not, MD continues with no change to the protonation state.

Computationally, the time to evaluate a Monte Carlo step is less than that required for an MD step, so constant pH MD using this approach is only slightly slower than traditional constant protonation state GB MD.

Under this model, the total charge on the molecule is generally non-zero and changes when a titratable group changes protonation state. Since GB solvation does not employ periodic boundary conditions and the free energy associated with introducing and solvating a charge are included in the non-electrostatic portion of the transition free energy accounted for above, the changing total charge does not present a problem.

5.2.2 Molecular dynamics

MD was performed using a pre-release version of AMBER 8,[113] modified to implement the algorithm described above (these modifications have been included in the

released version of AMBER 8). The ff99 force field[114] was employed. The first GB model developed by Onufriev, Bashford and Case[47,51,52] (igb=2) was used for solvation. Salt concentration (Debye-Hückel based) was set at 0.1 M. The cutoff for non-bonded interactions and computation of effective Born radii was 30 Å. Solute temperature was weakly coupled to a Berendsen temperature bath at 300 K with a time constant of 2 ps. Lengths of bonds including hydrogen were constrained using SHAKE. The time step was 2 fs.

5.2.3 Protonation state models

Titratable group models were developed for the side chains of aspartate, glutamate, histidine, lysine and tyrosine. Protonation states for a given group differ only in partial charges. When a group's protonation state changes, charges on all of its sidechain atoms are changed to reflect the new state. Titratable hydrogens have zero charge in the deprotonated state. The titratable hydrogens in aspartate, glutamate and tyrosine have zero van der Waals radius in the AMBER force field, so when their charge is zero they have no non-bonded interactions with the system, although they retain defined positions. Van der Waals radii for titratable hydrogens on lysine and histidine were left unchanged at their ff99 value of 0.6 Å. This does not seem to substantially affect results, since pK_a predictions for amine (LYS) and carboxylic acid (ASP and GLU) residues were of comparable quality (see section 5.3.3).

Partial charges were taken from the protonated and deprotonated residue definitions in AMBER 99 force field. This force field does not define a deprotonated tyrosine; charges for the deprotonated tyrosine were calculated using Antechamber[115] with the RESP charge method based on HF/6-31G* calculations conducted with Gaussian 98.[116] Although the largest charge changes in these charge sets are concentrated near the titratable proton, every atom has some charge difference between protonated and deprotonated forms. If peptide backbone charges are changed when the protonation state changes, it is not possible to use a single reference free energy. Due to the 1-4 electrostatic interactions defined in the AMBER force field, backbone atoms have specific electrostatic interactions with side chain atoms of neighboring residues. If backbone

charges are allowed to change, then the free energy difference between protonated and deprotonated forms becomes sequence dependent. To avoid this problem, backbone charges were fixed at the values defined for the protonated state across all protonation states. A charge correction was added to the beta carbon of the deprotonated state such that the total charge difference between protonated and deprotonated states was 1.

In this charge-change only model of protonation states, a deprotonated group can gain a proton only at the location of a zero charge "ghost" proton. A titratable group may unrealistically favor the deprotonated state if its ghost proton rotates into an unfavorable position for protonation. This problem is especially severe for carboxylic acids, where the *syn* location for the proton is much more favorable than the *anti*. When a ghost proton moves into the *anti* position it is unlikely to protonate, and unlikely to move until it protonates, since no forces act on a ghost proton. This problem is addressed by building a carboxylic acid model with two protons on the oxygen, kept 180 degrees apart by an improper torsion. Since rotation of the carboxylic group to exchange the oxygen atoms is also slow, two protons are defined on each oxygen of the carboxylic groups. The protonation state charge sets are defined such that no more than one of the four protons has a non-zero charge at any time.

5.2.4 Reference compounds

Reference free energy differences for the titratable groups were calculated for single amino acids as dipeptide (blocked) molecules, having the sequence acetyl—amino acid—methyl amine. A titration of the dipeptide reference compound was performed with solution pH set to $pK_{a,ref}$. (Reference pK_a values are listed in table 5.1.) $\Delta G_{elec,ref}$ was adjusted based on the results of this titration to give equal populations in the protonated and deprotonated states for titrations of the reference compound having pH equal to pK_a. For the simple case of a titratable group with only two protonation states, $\Delta G_{elec,ref}$ should be equal to the free energy difference calculated between the states by thermodynamic integration (TI). Calculations of $\Delta G_{elec,ref}$ were checked by performing TI between the protonated and deprotonated states using the parameters described in section 2.2 to calculate ΔG_{TI} according to

Table 5.1: Reference pK_a values for titratable side chains. Reference pK_a values for aspartate, glutamate, lysine and tyrosine were taken from Bashford *et al*.[117] The reference pK_a values for histidine were taken from Kyte.[118]

Residue	$pK_{a,ref}$
Aspartate	4.0
Glutamate	4.4
Histidine-δ	6.5
Histidine-ε	7.1
Lysine	10.4
Tyrosine	9.6

$$\Delta G_{\mathrm{TI}} = \int_0^1 \left\langle \frac{\partial V}{\partial \lambda} \right\rangle_\lambda d\lambda \tag{5.2}$$

where V is the potential and λ is the coupling parameter between the charges for the protonated and deprotonated states. Eleven equally spaced values were used for λ. At each value of λ, the reference compound was equilibrated for 40 ps and sampled for 1.6 ns. The free energy difference was calculated in Mathematica[119] by numerical integration of a fourth degree polynomial fit to the $\frac{\partial V}{\partial \lambda}$ values. In all simple cases ΔG_{TI} matched $\Delta G_{\mathrm{elec,ref}}$ to within 0.05 kcal/mol. This consistency is of course expected, and is really just a check on the correctness of the implementation of the Monte Carlo algorithm. It has recently been demonstrated that ΔG_{TI} values calculated using the GB model adopted here are similar to those computed using explicit solvent models.[71]

Calculations for the carboxylic residues were complicated by having four protonated states defined (*syn* and *anti* on each of the oxygens). For these residues, ΔG_{TI} (which was calculated between the deprotonated state and one of the *syn* protonated states) differed from $\Delta G_{\mathrm{elec,ref}}$ by approximately $k_B T \ln 2$ due to the statistical effects of multiple protonated states; (only the two *syn* sites see appreciable populations.) In addition to balancing the relative populations of the protonated and deprotonated states, it is important that the relative proton affinities of the *syn* and *anti* states are correct. Based on relative populations of the *syn* and *anti* states in test titrations of the model compound, the free energy of the *syn* state was calculated to be 1.6-1.9 kcal/mol lower than the *anti* state. This is in close agreement with quantum mechanical calculations and experimen-

tal estimates,[120] so it was assumed that the force field accurately accounts for the free energy difference between the *syn* and *anti* states and no adjustment was made to the relative energies of these states.

5.2.5 Test system molecular models

Hen egg white lysozyme (HEWL) was selected as the test system because it is well studied[104,111,121] and has a number of residues with pK_a values that differ markedly from their reference values. Structures 1AKI, 1LSA, 3LZT and 4LYT from the PDB were selected as starting crystal structures. The structures were chosen to facilitate comparison with earlier work[104] and provide a diversity of crystal properties. The structures are from orthorhombic, tetragonal, triclinic and monoclinic space groups, respectively. 3LZT is at high resolution (0.92 Å), 4LYT is at low resolution (2.5 Å) and 1LSA has crystal contacts that have been problematic in earlier studies.

Each structure was prepared using WHAT IF[122] to optimize the hydrogen bond network[123] (by flipping side chains of HIS, ASN and GLN) and strip crystal waters. Hydrogens were added to the structures using the LEaP module of AMBER. They were then minimized with 100 steps of steepest descent followed by 100 steps of conjugate gradient using the sander module of AMBER and the MD parameters described in section 5.2.2.

Simulations starting from the 1AKI structure were performed at 0.5 pH increments from pH 2.0 to 4.0 with aspartates and glutamates titrating, from pH 4.5 to 6.5 with aspartates, glutamates and histidine titrating, and from pH 9.0 to 12.0 with tyrosines and lysines titrating. Simulations starting from 1LSA, 3LZT and 4LYT were performed at 1.0 pH increments from pH 2.0 to 7.0 with all aspartates, glutamates and histidines titrating and from 9.0 to 12.0 with all tyrosines and lysines titrating. There were 10 fs between Monte Carlo steps. Non-titrating residues were fixed at their most probable protonation states (protonated for basic residues and deprotonated for acidic residues). Protonation state models for terminal residues have not yet been created, so terminal residues are fixed at their most likely neutral pH protonation state in all simulations: protonated for the N-terminus and LYS-1 side chain and deprotonated for the C-terminus.

This is approximation is expected to have little effect on the titrating sidechains. The C-terminal residue is approximately 10 Å from the nearest acid-pH titrating group and the N-terminal residue is nearly 15 Å from the nearest basic-pH titrating group, so direct interactions are small. There may also be an indirect interaction in the high pH simulations due to perturbation of the conformations sampled because the N-terminus (experimental pK_a of 7.8–8.0[121]) is held in the protonated state. The C-terminus is sufficiently acidic (experimental pK_a of 2.63–2.87[121]) that the indirect interaction should be negligible.

5.2.6 pK_a prediction calculations

Constant pH simulations can be analyzed in a fashion entirely analogous to that used for experiments that give protonation information for individual side chains as a function of pH. As long as the protonation fraction is a monotonic function of pH, the pK_a of a side chain can be defined as the pH value for which the protonated and deprotonated populations are equal. The special case of an ideal titratable group having no interactions with other titratable groups has a sigmoidal titration curve, and behavior characterized by the Henderson-Hasselbalch (HH) equation

$$\mathrm{pK_a} = pH - \log_{10}\left(\frac{[A^-]}{[HA]}\right) \tag{5.3}$$

Following the reasoning of Baptista *et al.*,[105] the system is assumed to be ergodic, so the ratio of time that a titratable group spends in the protonated and deprotonated states can be used as a ratio of concentrations. This can be combined with the pH according to equation 5.3 to yield a prediction of the pK_a. When a titratable group has sufficiently weak interactions with other titratable groups, its behavior is well described by the HH equation, and pK_a values calculated from simulations at different pH values will differ only by random error. As interactions increase, the HH equation will not adequately describe the titration curve.

Titration data are often represented in a Hill plot, where $\log_{10}\left(\frac{[A^-]}{[HA]}\right)$ is plotted versus pH. A titration curve for a titratable group governed by the HH equation has the form of a straight line with a Hill coefficient (slope) of 1. Titrating groups with non-HH

behavior will have Hill coefficients that differ from 1. The Hill coefficient can be determined by linear regression of the titration data points on a Hill plot. Since the coefficient calculated by regression may differ from 1 due to random error or non-HH titration behavior, a *t*-test should be used to decide whether the Hill coefficient suggests statistically significant non-HH behavior. In such cases, further simulations can be conducted to plot a full titration curve.

5.3 Results and Discussion

A method for constant pH MD simulations should be computationally efficient and capable of reproducing accurate titration curves. Furthermore, when applied to macromolecules, an ideal method would yield pK_a predictions in close agreement with experimental values, converge rapidly to these predictions and maintain the stability of the trajectory. As previously mentioned, the proposed method is only slightly more computationally expensive than traditional GB MD. Here, the proposed method is evaluated based on how well it meets the remainder of these criteria, and non-HH behavior and conformation-protonation state correlations suggested by the results are investigated.

5.3.1 Convergence

Small systems, such as the reference compounds, converge to the relative protonation state populations predicted by the HH equation within a few nanoseconds of simulated time. For example, a titration curve for five 1 ns simulations of the aspartate model compound, shown in figure 5.1, closely matches the predicted titration curve.

Convergence in larger systems, such as HEWL, is much more difficult to achieve. As seen in figure 5.2, the predicted pK_a value for most residues stabilizes within a few hundred picoseconds. This stabilization does not represent convergence; the same residue may stabilize at a significantly different pK_a value if a different random seed is used in an otherwise identical simulation. Since the random error due to incomplete convergence may produce larger effects than those caused by a small change in pH, titration curves for titrating groups in proteins can be noisy. This is demonstrated by

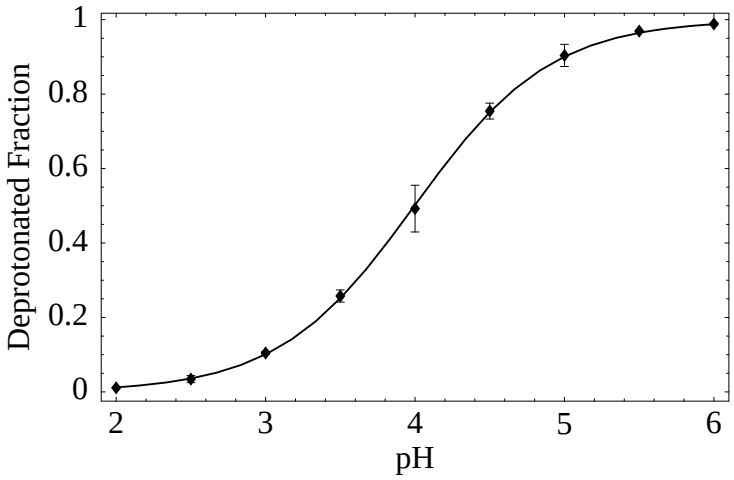

Figure 5.1: Deprotonated fraction for five 1 ns simulations with different initial velocities of aspartate model compound at pH values between 2 and 6. Data points represent average deprotonation over the five simulations, error bars illustrate standard deviations and the solid line is a best-fit curve.

comparing the protein titration curve seen in figure 5.3 to the model compound data of figure 5.1, noting that the protein data are for GLU-7, one of the better converged residues in HEWL. Despite the precision problems posed by these random errors, pK_a predictions are generally fairly accurate, and the impact of noise can be reduced by combining results from multiple simulations, as seen in section 5.3.3.

The major limiting factor on convergence appears to be conformational sampling. As shown in section 5.3.5, the instantaneous pK_a is strongly dependent on conformation, so if two simulations sample conformation space differently, it should be expected that they would have differing protonation state populations. Sufficiently complete conformational sampling is achievable for small systems, but is currently computationally infeasible for systems the size of HEWL.

Accepting that complete conformational sampling is out of reach for HEWL, titrations of 1 ns were performed to allow sufficient time for predicted pK_a stabilization, if not convergence. For the simulation shown in figure 5.2, each of the titrating residues was evaluated for a protonation state transition an average of 11,000 times, of which between 160 and 840 transitions were accepted.

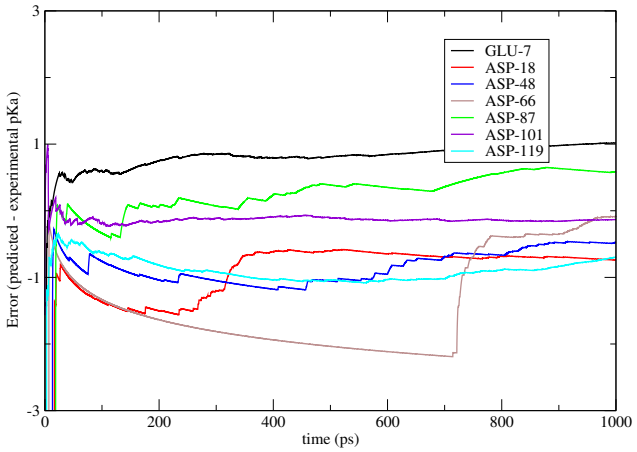

Figure 5.2: Time evolution of predicted pK_a for acidic residues in simulation of HEWL at pH 3.0, starting from structure 1AKI. Each point represents the predicted pK_a calculated from all protonation data collected up to that time in the simulation. Residues ASP-52 and GLU-35 do not converge due to H-bonding issues and large offset (see text), respectively, and are not shown on this plot.

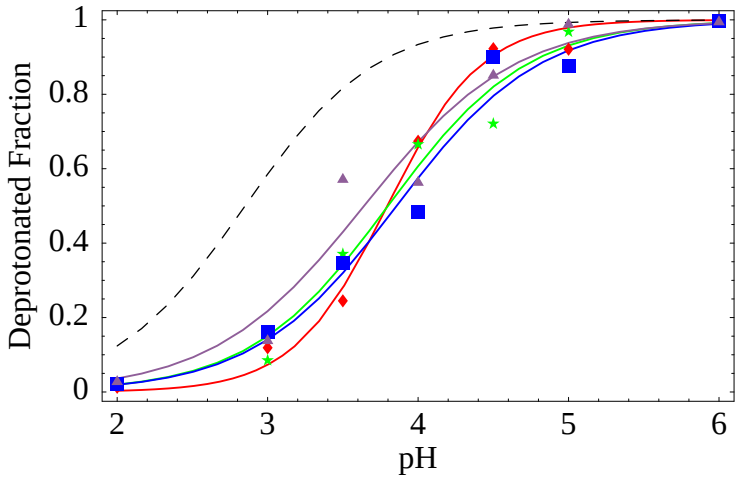

Figure 5.3: Deprotonated fraction for GLU-7 at pH values between 2.0 and 6.0. Each data point represents the average deprotonation of four 1 ns simulations starting from one of four crystal strucutres (1AKI, 1LSA, 3LZT and 4LYT). Solid lines are best-fit curves; the dashed line shows the expected titration curve based on experiment.

5.3.2 Simulation stability

Since this method involves instantaneous changes in protonation state, which result in non-physical discontinuities in energy and force, system stability across protonation state changes was examined. When the protonation state changes, there is a discontinuous change in total energy reported by AMBER equal to ΔG_{elec}. Most of this energy change represents transfer of energy between the energy modes governed by the force field and those outside the scope of the force field (e.g. quantum mechanical energy of the bond and solvation free energy of the proton). The remainder of the change represents Boltzmann sampling of the energy levels of different protonation states. Even when temperature regulation is removed, there is no trend to the changes in total energy. Average kinetic energy does follow the fluctuations in total energy, so temperature fluctuations are increased somewhat by this method. However, even in the worst case of a very small system (the reference compound) with no temperature regulation, root mean square temperature fluctuations were only about twice as high for a simulation where protonation state changed rapidly as they were for a simulation where protonation state was fixed. In the more common case of a larger system with temperature regulation and slower protonation state changes, temperature fluctuations are fairly small—on the order of 5 K for the HEWL titrations described here.

Conformational stability, as well as energetic stability, is of interest in biomolecule simulations. Experimentally, HEWL is a stable protein across a wide range of solvent pH values, and this should be reflected in the trajectories of the titrations. Figure 5.4 compares alpha carbon root mean square deviation (RMSD) versus crystal structure for titrations starting from 1AKI at a range of pH values to a non-titrating trajectory. Most importantly, this plot shows that after an initial relaxation period, RMSD stabilizes for each trajectory. One simulation was continued to 3 ns to confirm stability: there were no significant excursions beyond an RMSD of 2.5 Å. In general, RMSD for titrating trajectories increased more rapidly and stabilized at higher values than for the non-titrating trajectory. Traditional MD trajectories are known to be biased toward conformations that are compatible with their fixed protonation state;[112] it seems reasonable that allowing protonation states to change would reduce this bias and allow greater conformational

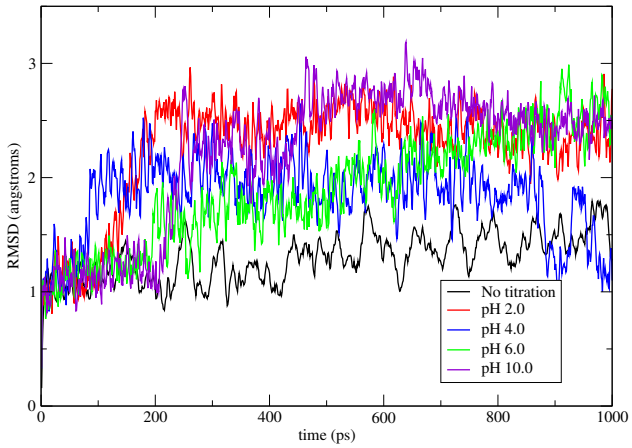

Figure 5.4: α-carbon RMSD from crystal coordinates for residues 4-125 of 1AKI structure of HEWL at five different solvent pH values. Plots are representative of behavior of other simulations. RMSD for constant protonation state (no titration) MD under similar conditions is shown for comparison.

sampling, producing a higher RMSD. One might also expect that simulations conducted at pH values close to the pH of the crystal structure would tend to sample conformations closer to the crystal structure. Indeed, figure 5.4 shows that the pH 4.0 trajectory stabilizes at the lowest RMSD relative to the 1AKI crystal structure, which was solved at pH 4.5.

5.3.3 pK_a predictions

Although the primary aim of this work is to improve the physical realism of dynamics, it is difficult to validate the quality of simulated dynamics as a function of pH, since HEWL appears to have no major structural changes in the pH range considered here. HEWL was selected because it has been well studied, so it provides a good test system for determining whether protonation states are accurately sampled, and accurate sampling is a prerequisite for simulating pH-dependent conformational changes. Therefore, pK_a values, which can be calculated from the simulations using equation 5.3, are compared to experimentally measured values as the primary quality measure used to validate the proposed method.

In comparing predicted pK_a values to experimental measurements, it is useful to

have a method for combining predictions from simulations conducted at different solvent pH values into a single composite pK_a prediction, taking account of their relative accuracies. This is commonly achieved by plotting the data at each pH on a Hill plot and performing linear regression. In regression, each data point is weighted according to the inverse of its variance (more properly, the variance of the distribution from which the data point is drawn). It is expected that the variance of a data point will be dependent on the absolute difference between the pH at which it was taken and the predicted pK_a. This follows from equation 5.3, which becomes increasingly sensitive to small changes in the number of time steps spent in each protonation state as the quotient of these numbers becomes very large or very small. It is computationally infeasible to run sufficient simulations to determine a separate variance model for each titratable group, so data for all titratable groups were pooled to determine a global variance model for the method. Figure 5.5 shows a scatter plot of absolute difference between the pH and predicted pK_a (offset) versus pK_a prediction error. The running (windowed) variance line on this plot shows that variance is roughly uniform when the offset is less than 2.0 pH units, and increases rapidly outside of this range. Since there are insufficient data to empirically determine a variance at each offset, a simplified variance model is drawn from these data: uniform variance for offsets less than 2.0 pH units and very high (effectively infinite) variance for larger offsets. This leads to uniform weights for the small offset data points and zero weight for those with large offset. Since the non-HH behavior in this system is small and in general does not affect the predicted pK_a, the number of free parameters in the fit is reduced by restricting the slope of the fitted line to 1. The composite pK_a calculated by the fitting process described here can be determined by the mathematically equivalent operation of averaging all pK_a predictions with an offset less than 2.0 pH units.

Tables 5.2 and 5.3 show pK_a values predicted from 1 ns simulations starting from the 1AKI structure of HEWL. In general, the composite predictions show close correspondence to experimental data, and variation between simulations at different solvent pH values is small. A few problematic cases are worthy of mention. ASP-52 has a hydrogen bond to ASN-46 in all four crystal structures. While this bond is maintained, ASP-52 is prevented from protonating. In most simulations, this hydrogen bond is sta-

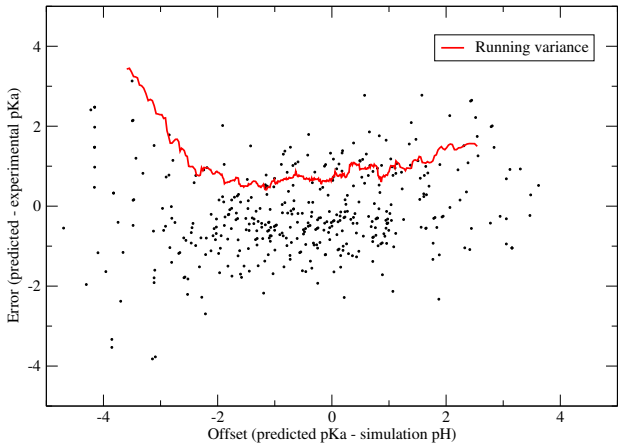

Figure 5.5: Scatter plot of difference between predicted pK_a and solvent pH (offset) versus difference between predicted and experimental pK_a (prediction error). Points represent all predictions with offsets between -5 and 5 made. Variance is calculated with a window size of 40 data points.

ble throughout all or nearly all of the simulation, leading to a very low predicted pK_a. The simulations yielding the best results for ASP-52 (pH 2.0, 2.5 and 3.5) were those in which ASP-52 and ASN-46 were dissociated for much of the simulation. A similar effect of hydrogen bonding leading to erroneously low pK_a predictions was seen with GLU-35, but to a lesser extent. Results for the three tyrosine residues were markedly poorer than results for the other residue types. This may be due to slower conformational sampling due to steric hindrance of the large aromatic ring, and ignoring the effect of the polarizability of the aromatic ring.

The overall quality of the pK_a value predictions can be measured by the root mean square (RMS) error of predicted pK_a values relative to experimental values, which is 0.86 for simulations starting from the 1AKI structure, as shown in table 5.4. RMS error for null model predictions, where each residue's pK_a is predicted to be equal to the reference value given in table 5.1, are also shown in table 5.4. The current method gives predictions that are an overall improvement on the null model, and superior for each type of titrating residue, except lysine. Prediction results for lysine are actually more accurate than for any other residue (RMS error 0.64), but due to the very small shifts of these residues' pK_a values from reference values, the null model RMS error is very low.

Table 5.2: pK_a predictions for acidic residues of HEWL, calculated from 1 ns simulations starting from 1AKI at specified pH. Composite pK_a is the average of predictions with absolute offset less than 2.0 (see text for discussion). Where experimental pK_a values[121] were given as a range, the midpoint of the range is used; where only an upper bound was given, the upper bound is used.

	pH 2.0	pH 2.5	pH 3.0	pH 3.5	pH 4.0	pH 4.5	pH 5.0	pH 5.5	pH 6.0	pH 6.5	Comp.	Exp.
ASP-18	0.97	2.02	1.92	2.05	1.98	2.12	2.20	1.93	-∞	-∞	1.83	2.66
ASP-48	2.84	0.75	2.03	1.29	1.94	0.92	2.17	1.96	-∞	-∞	1.87	2.50
ASP-52	2.15	2.04	-0.14	2.05	-∞	2.38	-∞	2.54	-∞	-∞	2.08	3.68
ASP-66	2.57	1.51	1.92	0.40	-∞	-∞	-∞	2.82	-∞	-∞	2.00	2.00
ASP-87	3.14	2.17	2.66	2.36	3.12	0.84	-∞	-∞	-∞	-∞	2.69	2.07
ASP-101	3.56	3.76	3.96	3.90	3.80	3.67	4.07	3.85	3.69	4.09	3.80	4.09
ASP-119	3.70	2.29	2.50	2.40	2.36	2.46	2.49	2.69	3.15	-∞	2.60	3.20
GLU-7	3.89	3.88	3.87	3.99	4.18	3.42	3.93	3.82	2.68	3.39	3.85	2.85
GLU-35	3.87	5.97	6.48	5.27	5.65	5.60	6.78	5.31	5.19	5.36	5.32	6.20
HIS-15	-	-	-	-	6.26	5.87	6.69	6.69	6.71	6.26	6.45	5.71

Table 5.3: pK_a predictions for basic residues of HEWL, calculated from 1 ns simulations starting from 1AKI at specified pH. Composite pK_a is the average of predictions with absolute offset less than 2.0 (see text for discussion). Where experimental pK_a values were given as a range, the midpoint of the range is used; where only an upper bound was given, the upper bound is used.

	pH 9.0	pH 9.5	pH 10.0	pH 10.5	pH 11.0	pH 11.5	pH 12.0	Comp.	Exp.
TYR-20	11.21	10.73	9.91	11.86	11.02	10.82	10.84	10.86	10.3
TYR-23	11.54	12.02	12.43	11.45	11.38	11.08	11.27	11.30	9.8
TYR-53	10.88	10.38	11.08	11.41	10.48	11.18	10.89	10.90	12.1
LYS-13	9.77	9.12	9.54	10.08	9.38	9.38	9.34	9.58	10.5
LYS-33	9.47	9.93	9.44	10.04	9.36	9.55	9.62	9.63	10.6
LYS-96	9.86	9.70	9.96	10.22	10.13	8.98	9.96	9.97	10.8
LYS-97	9.55	9.87	10.08	10.15	10.13	10.17	10.21	10.02	10.3
LYS-116	9.97	10.14	10.27	10.14	10.33	10.11	10.20	10.17	10.4

Table 5.4: RMS errors of predicted pK_a values from experimental values. All structures refers to composite pK_a predictions using data from all simulations on 1AKI, 1LSA, 3LZT and 4LYT. Null model RMS errors are provided for comparison; in the null model, all residues are predicted to have the reference pK_a values given in table 5.1.

	All structures	1AKI	1LSA	3LZT	4LYT	Null model
All residues	0.82	0.86	0.77	0.88	0.95	1.19
Aspartates	0.69	0.80	0.72	0.86	0.78	1.34
Glutamates	0.88	0.94	0.97	1.01	0.54	1.68
Histidine	0.21	0.74	0.11	0.10	0.01	0.69
Tyrosines	1.29	0.88	1.10	1.23	1.69	1.50
Lysines	0.64	0.83	0.57	0.51	0.78	0.21

The results reported here are a significant improvement on the explicit solvent TI-based constant pH MD pK_a predictions reported for 1AKI HEWL by Bürgi *et al.*,[104] which have RMS error of 2.8-3.8 and seem to be far from convergence in 3 ns titrations. They are also more accurate than Lee *et al.*'s continuous protonation state results for HEWL, which had RMS error of 1.31 for non-terminal residues.[102] The constant pH method employing Poisson-Boltzmann protonation state sampling described by Walczak and Atosiewicz had RMS error of 0.81-1.12 (depending on parameters) in application to ovomucoid third domain.[106] The ovomucoid third domain may represent an easier prediction problem than HEWL, as it has fewer residues with pK_a values that are substantially shifted relative to reference pK_a values; these strongly shifted residues have the greatest errors in the Walczak and Atosiewicz method. It is difficult to compare the quality of the proposed GB constant pH MD method to that of Baptista *et al.*,[105] as they have only reported on application of their method to succinic acid, and not to proteins.

The Poisson-Boltzmann-based pK_a prediction methods employed by Baptista *et al.* and Walczak and Atosiewicz for protonation state sampling have a long history[124] and continue to be an active area of research.[112,121,125] Non-PB-based electrostatics methods have also found success.[126,127] When the best of these electrostatics-only methods are applied to crystal structures they provide somewhat more accurate predictions of pK_a values (RMS error of 0.5-0.7) in less computer time than the proposed GB constant

Table 5.5: Composite pK_a predictions for simulations starting from the 1AKI, 1LSA, 3LZT and 4LYT crystal structures. No value is shown for ASP-66 in the 3LZT simulations because none of the predictions had offsets with magnitude less than 2.

	1AKI	1LSA	3LZT	4LYT	Exp.
ASP-18	1.83	1.69	2.38	2.55	2.66
ASP-48	1.87	2.48	2.04	2.38	2.5
ASP-52	2.08	2.68	1.75	2.65	3.68
ASP-66	2.00	1.18	-	2.19	2.0
ASP-87	2.69	2.66	2.32	3.34	2.07
ASP-101	3.80	3.74	3.76	3.96	4.09
ASP-119	2.60	2.45	2.17	1.98	3.2
GLU-7	3.85	3.72	3.89	3.58	2.85
GLU-35	5.32	5.14	5.23	5.97	6.2
HIS-15	6.45	5.82	5.61	5.70	5.71
TYR-20	10.86	10.82	10.98	11.62	10.3
TYR-23	11.30	11.42	11.39	12.21	9.8
TYR-53	10.90	11.25	10.84	11.10	12.1
LYS-13	9.58	9.87	9.97	9.56	10.5
LYS-33	9.63	9.66	9.94	9.47	10.6
LYS-96	9.97	10.33	10.15	10.04	10.8
LYS-97	10.02	10.04	9.94	9.86	10.3
LYS-116	10.17	10.12	10.19	10.12	10.4

pH MD method. Although these methods are fairly accurate, they can be very sensitive to details of the crystal structure because all atomic positions are fixed, and they often produce widely varying pK_a value predictions for different crystal structures of the same protein.[111] Models (PB and non-PB-based) that allow for some conformational rearrangement have much less dependence on crystal structure,[85,87,88] and a dynamics-based method should be immune to these effects. This was tested by running simulations starting from three additional crystal structures (PDB identifiers 1LSA, 3LZT and 4LYT). These structures were chosen for maximum diversity of crystal characteristics, as described in section 5.2.5. As seen in table 5.5 and summarized in table 5.4, pK_a value predictions were highly consistent across the four structures, with a total variation in RMS error of only 0.18 pH units. This stands in contrast to a recent electrostatics study of these structures, which yielded RMS errors of 1.01, 1.44, 1.15 and 2.03 for 1AKI, 1LSA, 3LZT and 4LYT, respectively.[111]

5.3.4 Non-Henderson-Hasselbalch behavior

Titrating residues were tested for non-HH behavior (titration curves that do not match the sigmoidal shape defined by equation 5.3) using the procedure described in section 5.2.6; these results are shown in table 5.6. First, it should be noted that, as in most proteins, the magnitude of non-HH behavior is small—in all significant cases, it is less than 0.35 deviation in pK_a prediction for every 1 unit change in solution pH. Furthermore, any error will tend to be canceled by the opposing effects of predictions made from simulations with pH above and below the residue's pK_a, so ignoring non-HH behavior in the lysozyme pK_a calculations above is a reasonable approximation. The results in table 5.6 also justify the use of only single-site MC moves for this system—none of the interactions are strong enough for any titratable group to block protonation state changes in a nearby group. Nevertheless, some interaction between titrating residues leading to non-HH behavior is expected for lysozyme, and it is reassuring that the proposed method reproduces these effects.

The bold lines in table 5.6 indicate which residues have statistically significant non-HH behavior. LYS-96 and LYS-97 interact with each other due to their obvious proximity in both primary and tertiary structure. LYS-116 projects toward TYR-23 (titrating N to O distance 7.7 Å in 1AKI); the non-HH effect on TYR-23 is presumably lost in noise due to the poor tyrosine results. ASP-87 most likely interacts with HIS-15, which reaches 90% confidence for non-HH behavior. ASP-101 does not appear to have specific interactions with any single titrating group. However, weak interactions have statistical significance for ASP-101 because it is one of the best converged residues, and as such has little noise. The analysis for GLU-35 is dominated by three data points with very negative errors, representing simulations in which GLU-35 was significantly H-bonded. Eliminating these outliers increases the p-value from 0.017 to 0.7.

5.3.5 Conformation-protonation correlation

A major motivation for the implementation of this method is the idea that protonation state and conformation are strongly coupled, such that they cannot be adequately studied in isolation (*e.g.* electrostatics-based pK_a predictions and traditional molecular

Table 5.6: Hill coefficients for titration data determined by linear regression. P-value is the significance level at which the Hill coefficient differs from one. Residues with p-values less than 0.05 are indicated in boldface.

Residue	Hill coefficient	P-value
GLU-7	1.006	0.897
LYS-13	0.966	0.755
HIS-15	0.816	0.100
ASP-18	0.763	0.149
TYR-20	0.740	0.131
TYR-23	0.955	0.864
LYS-33	0.939	0.446
GLU-35	**0.670**	**0.017**
ASP-48	0.995	0.984
ASP-52	0.838	0.528
TYR-53	0.908	0.393
ASP-66	1.003	0.993
ASP-87	**0.751**	**0.039**
LYS-96	**0.796**	**<0.001**
LYS-97	**0.865**	**0.004**
ASP-101	**0.885**	**0.030**
LYS-116	**0.908**	**0.002**
ASP-119	1.076	0.575

dynamics). Results of these simulations support this idea: the protonation and conformation of ASP-18 in the pH 2.5 simulation starting from 1AKI is examined as an example.

Essential dynamics (ED)—principal component analysis (PCA) of trajectory data—is a useful technique for separating functionally important, slow, large scale motions from local fluctuations,[128] discussed in detail in chapter 7. Projections of a particular snapshot from the trajectory onto the most significant principal components (the eigenvectors having the largest eigenvalues) can be used as a dimensionally reduced representation of a molecular conformation. Best results in correlating conformation to protonation for a particular residue are obtained when the atoms included in the PCA are limited to the residue and its immediate neighbors. The results presented here are based on PCA[129] of ASP-18 and all atoms within 7.5 Å of ASP-18 in the 1AKI crystal structure. In figure 5.6, position of the data points represents conformation (projection onto the two eigenvectors with the largest eigenvalues) while shade represents degree of protonation (darker is more protonated). A strong qualitative association between conformation and protonation is apparent: the conformational cluster in the upper left is almost entirely deprotonated, while the cluster in the lower right is predominantly protonated.

Plots such as figure 5.6 illustrate correlation between conformation and protonation, but deriving quantitative data directly from such a plot necessitates partitioning into conformational clusters which, if done by hand, is subjective and effectively limited to two or three dimensions. Clustering algorithms provide an objective means for identifying clusters and can operate in high-dimensional spaces that are not readily visualized. A *k*-means clustering algorithm[130] (Euclidean distance) was employed to conformationally cluster one thousand 1 ps snapshots, represented by their projections onto the 10 largest eigenvectors, and pK_a was calculated separately for each cluster. There is no obvious choice for the value of *k* (the number of clusters): values that are too low may force distinct conformations with different protonation characteristics into a single cluster, while values that are too high may divide what should be a single cluster into two clusters. It seems prudent to try a range of values, increasing *k* until it is clear that further increases will not identify clusters with unique protonation properties. For instance, in table 5.7

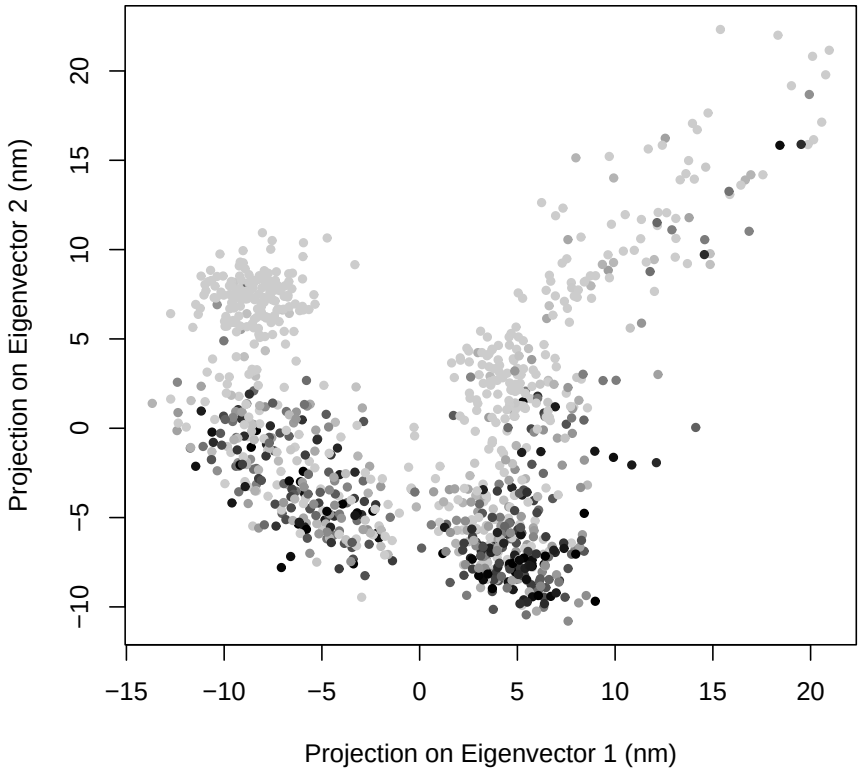

Figure 5.6: Coupling of conformation and protonation are illustrated in this plot, where spot location represents conformation and shade represents protonation. Specifically, principal component analysis was performed on a 1 ns trajectory at pH 2.5 beginning from 1AKI. Only atoms within 7.5 Å of ASP-18 were included in the analysis. This plot shows the projections of 1 ps snapshots from the trajectory onto the first two (largest eigenvalues) principal components. Shading represents fraction of time spent protonated in the 1 ps window surrounding the snapshot: black represents fully protonated, lightest gray represents fully deprotonated.

Table 5.7: pK_a values calculated for conformational clusters. Clusters were generated using the k-means algorithm with Euclidean distances. Data points to be clustered were projections of trajectory snapshots onto the first 10 principal components (see text). pK_a values were calculated for each cluster from the combined protonation state data for each snapshot assigned to the cluster.

	k=2	k=3	k=4	k=5	k=6	k=7
1	1.95	1.48	0.85	0.91	0.75	0.55
2	2.07	1.95	1.49	1.53	1.39	1.39
3		2.15	2.15	1.64	1.61	1.61
4			2.24	2.26	1.66	1.66
5				2.35	2.24	2.16
6					2.34	2.27
7						2.34

it is clear that k of 2 or 3 is too small to separate distinct protonation properties. $k = 4$ identifies a very acidic cluster with pK_a of less than 1.0, a cluster with pK_a near 1.5 and two more clusters with pK_a greater than 2.0. Increasing k beyond 4 serves only to non-productively subdivide these clusters.

Mapping the clusters from principal component space back to atomic coordinates using Interactive Essential Dynamics, described in chapter 7, provides a means to identify the physical basis for the protonation behaviors exhibited by the different clusters. The centroid of each cluster is taken as the representative of the conformations in the cluster. The process of projecting a snapshot onto the principal components is reversed to generate atomic coordinates in Cartesian space from the centroid coordinates in principal component space.

Representations of the $k = 4$ cluster centroids in atomic coordinates are illustrated in figure 5.7. These images show that in this trajectory, LYS-13 adopts three distinct conformations, with varying distances from ASP-18 leading to a difference in pK_a of 1.4 between cluster 1 and cluster 4. The dramatic differences in the protonation states sampled in these conformational clusters demonstrates the coupling of protonation state and conformation and emphasizes the need to use techniques that maintain this coupling by analyzing protonation state in conjunction with dynamics.

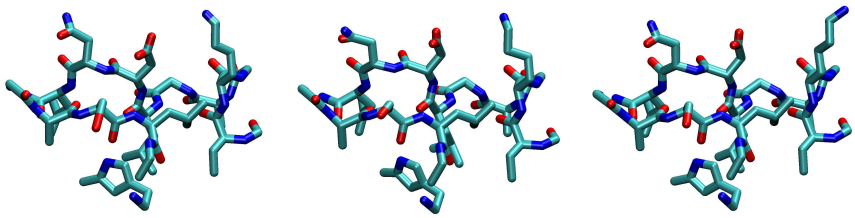

Figure 5.7: Rendered images[131] representing atomic coordinates corresponding to centroids of clusters 1, 2 and 4 (left to right, having pK_a values 0.85, 1.49, 2.24) from k=4 clustering in table 5.7. ASP-18 is top center, ASN-19 is in the upper left and LYS-13 is in the upper right. Note that bond lengths and angles are somewhat distorted due to the averaging effects of taking the centroid.

5.3.6 Summary

The method described here, combining GB MD with Monte Carlo sampling of discrete protonation states, provides a computationally efficient means for performing constant pH MD. As evidenced by close agreement between predicted and experimental pK_a values, this method accurately samples protonation states while producing a conformationally and energetically stable trajectory. Convergence is rapid for small molecules, but much slower for larger biomolecules. Slow convergence is due to slow conformational sampling and, in systems that have more strongly interacting titratable groups than HEWL, barriers to moving in protonation-state space.

The analysis of correlation between conformation and protonation state in these results illustrates the strong coupling between these aspects of molecular configuration; the ability to sample protonation states concurrently with conformation is an important step in improving the physical realism of MD simulations. The accurate constant pH MD achieved by this method will facilitate the study of pH and protonation state dependent dynamics that have been inaccessible with traditional MD.

This chapter is a reprint in full of material that appeared in *Constant pH Molecular Dynamics in Generalized Born Implicit Solvent*. John Mongan, David A Case and J. Andrew McCammon. Journal of Computational Chemistry, **25**(16), 2038-48, December 2004. I was the primary author and researcher for this work.

5.4 Appendix: Partial charges of titratable groups

5.4.1 Aspartate charges

	deprotonated	O2 *syn*	O2 *anti*	O1 *syn*	O1 *anti*
N	-0.4157	-0.4157	-0.4157	-0.4157	-0.4157
H	0.2719	0.2719	0.2719	0.2719	0.2719
CA	0.0341	0.0341	0.0341	0.0341	0.0341
HA	0.0864	0.0864	0.0864	0.0864	0.0864
CB	-0.1786	-0.0316	-0.0316	-0.0316	-0.0316
2HB	-0.0122	0.0488	0.0488	0.0488	0.0488
3HB	-0.0122	0.0488	0.0488	0.0488	0.0488
CG	0.7994	0.6462	0.6462	0.6462	0.6462
OD1	-0.8014	-0.5554	-0.5554	-0.6376	-0.6376
OD2	-0.8014	-0.6376	-0.6376	-0.5554	-0.5554
1HD2	0.0000	0.4747	0.0000	0.0000	0.0000
C	0.5973	0.5973	0.5973	0.5973	0.5973
O	-0.5679	-0.5679	-0.5679	-0.5679	-0.5679
2HD2	0.0000	0.0000	0.4747	0.0000	0.0000
1HD1	0.0000	0.0000	0.0000	0.4747	0.0000
2HD1	0.0000	0.0000	0.0000	0.0000	0.4747

5.4.2 Glutamate charges

	deprotonated	O2 *syn*	O2 *anti*	O1 *syn*	O1 *anti*
N	-0.4157	-0.4157	-0.4157	-0.4157	-0.4157
H	0.2719	0.2719	0.2719	0.2719	0.2719
CA	0.0145	0.0145	0.0145	0.0145	0.0145
HA	0.0779	0.0779	0.0779	0.0779	0.0779
CB	-0.0398	-0.0071	-0.0071	-0.0071	-0.0071
HB2	-0.0173	0.0256	0.0256	0.0256	0.0256
HB3	-0.0173	0.0256	0.0256	0.0256	0.0256
CG	0.0136	-0.0174	-0.0174	-0.0174	-0.0174
HG2	-0.0425	0.0430	0.0430	0.0430	0.0430
HG3	-0.0425	0.0430	0.0430	0.0430	0.0430
CD	0.8054	0.6801	0.6801	0.6801	0.6801
OE1	-0.8188	-0.5838	-0.5838	-0.6511	-0.6511
OE2	-0.8188	-0.6511	-0.6511	-0.5838	-0.5838
2HE2	0.0000	0.4641	0.0000	0.0000	0.0000
C	0.5973	0.5973	0.5973	0.5973	0.5973
O	-0.5679	-0.5679	-0.5679	-0.5679	-0.5679
2HE2	0.0000	0.0000	0. 4641	0.0000	0.0000
1HE1	0.0000	0.0000	0.0000	0. 4641	0.0000
2HE1	0.0000	0.0000	0.0000	0.0000	0. 4641

5.4.3 Histidine charges

	δ, ε protonated	δ protonated	ε protonated
N	-0.1506	-0.1506	-0.1506
H	0.1749	0.1749	0.1749
CA	-0.1394	-0.1394	-0.1394
HA	0.1036	0.1036	0.1036
CB	-0.1057	-0.2276	-0.2710
HB2	0.1021	0.0863	0.0546
HB3	0.1021	0.0863	0.0546
CG	0.0511	-0.0015	0.2784
ND1	0.0021	-0.2058	-0.4233
HD1	0.2584	0.3183	0.0000
CE1	-0.0333	0.1473	0.0260
HE1	0.2189	0.1222	0.1268
NE2	-0.1410	-0.6015	-0.0980
HE2	0.3534	0.0000	0.2669
CD2	-0.1438	0.0437	-0.2976
HD2	0.2135	0.1102	0.1604
C	0.6756	0.6756	0.6756
O	-0.5421	-0.5421	-0.5421

5.4.4 Tyrosine charges

	protonated	deprotonated
N	-0.4157	-0.4157
H	0.2719	0.2719
CA	-0.0014	-0.0014
HA	0.0876	0.0876
CB	-0.0152	-0.0858
HB2	0.0295	0.0190
HB3	0.0295	0.0190
CG	-0.0011	-0.2130
CD1	-0.1906	-0.1030
HD1	0.1699	0.1320
CE1	-0.2341	-0.4980
HE1	0.1656	0.1320
CZ	0.3226	0.7770
OH	-0.5579	-0.8140
HH	0.3992	0.0000
CE2	-0.2341	-0.4980
HE2	0.1656	0.1320
CD2	-0.1906	-0.1030
HD2	0.1699	0.1320
C	0.5973	0.5973
O	-0.5679	-0.5679

5.4.5 Lysine charges

	protonated	deprotonated
N	-0.3479	-0.3479
H	0.2747	0.2747
CA	-0.2400	-0.2400
HA	0.1426	0.1426
CB	-0.0094	-0.1096
HB2	0.0362	0.0340
HB3	0.0362	0.0340
CG	0.0187	0.0661
HG2	0.0103	0.0104
HG3	0.0103	0.0104
CD	-0.0479	-0.0377
HD2	0.0621	0.0115
HD3	0.0621	0.0115
CE	-0.0143	0.3260
HE2	0.1135	-0.0336
HE3	0.1135	-0.0336
NZ	-0.3854	-1.0358
HZ1	0.3400	0.0000
HZ2	0.3400	0.3860
HZ3	0.3400	0.3860
C	0.7341	0.7341
O	-0.5894	-0.5894

Chapter 6

Accelerated Molecular Dynamics

ABSTRACT

Many interesting dynamic properties of biological molecules cannot be simulated directly using molecular dynamics due to the limitations of the computationally feasible timescale. Such properties involve transitions over high free energy barriers, which are rare events in a simulation. To address this problem, a robust biasing function is proposed that can be used to increase the frequency of transitions over free energy barriers. The potential energy landscape is altered by adding a bias potential to the true potential such that the escape rates from potential wells are enhanced, which accelerates sampling in molecular dynamics simulations. The biased potential echoes the shape of the underlying true potential, focusing sampling on the minima. This approach, which can be extended to biomolecules, samples the conformational space more efficiently than conventional molecular dynamics simulations, and converges to the correct canonical distribution.

6.1 Introduction

Molecular dynamics simulation is one of the most widely used techniques in computational chemistry due, in part, to its simplicity and ability to accurately sample the conformational space of a molecular system. By integrating Newton's equations of motion, this technique evaluates the time-dependent behavior and evolution of a molecular

system as it samples conformational space. Therefore, with an accurate representation of the system's potential energy landscape, the conformational space can be easily sampled, and thermodynamic and kinetic properties can be calculated while studying a host of other structural and dynamic phenomena. As a result, molecular dynamics simulations have provided thorough information on local motions and conformational changes of proteins[132] and DNA[133–138]

However, for most biological systems of interest, available computational resources limit the simulation time to the nanosecond timescale, so conventional molecular dynamics cannot be used to adequately explore portions of the energy landscape separated by high barriers from the initial minimum. Furthermore, for most biological molecules, the energy landscape has multiple minima or potential energy wells separated by high free energy barriers, and during a molecular dynamics simulation the system is trapped in one or another local minimum for long periods of simulation time. Consequently, thermodynamic properties of interest for large biological systems cannot be directly calculated from simulations because of the non-ergodic nature of conventional molecular dynamics for systems with high free energy barriers.[139]

The dynamic evolution of biological molecules and many other molecular systems occurs through series of rare events as the systems move from sampling one potential energy basin to another.[140,141] Therefore, in order to perform realistic simulations of molecular systems, one has to be able to simulate series of rare transitions between potential energy minima. There have been a number of approaches introduced that are aimed at addressing this problem. These methods include conformational flooding,[142] replica exchange,[143] umbrella sampling,[144] self-guided MD,[145] and others reviewed by Berne and Straub.[146] Umbrella sampling, one of the more widely used approaches, involves construction of a compensating function, known as the umbrella, which is added to the true potential energy function in order to bias the sampling toward a particular set of conformations. However, construction of the umbrella requires prior knowledge of the conformations of interest.

This chapter introduces a molecular dynamics approach based on earlier work by Voter[140,141] that simulates infrequent events of molecular systems without any advance knowledge of the location of either the potential energy wells or barriers. Voter[140,141]

recently proposed a hyperdynamics method to speed up molecular dynamics simulations by reducing the amount of computational time systems spend in potential energy minima between crossing potential barriers. The scheme modifies the potential energy surface, $V(r)$, by adding a bias potential, $\Delta V(r)$, to the true potential such that the potential surfaces near the minima are raised and those near the barrier or saddle point are left unaffected. Statistics sampled on the biased potential are then corrected to remove the effect of the bias. In Voter's implementation of the bias potential, the Hessian matrix is diagonalized at each step, so that the transition state regions can be identified. This limits its use to small systems because of the computational cost of assembling and diagonalizing the Hessian. Alternatively, a prescription for a simple bias potential was proposed by Steiner *et al.*[147] and later used by Rahman and Tully[148] in which the bias potential is chosen such that the modified potentials is constant if the unmodified potential falls below a theshold. This simple definition of the bias potential does not require the diagonalization of the Hessian matrix at each step, and hence makes it possible for it to be applied to larger systems. This chapter presents a simple, robust way of altering the potential energy landscape that preserves the underlying shape of the potential energy surface, and allows for the simulation to be extended to larger molecular systems, like proteins. It is shown that this approach accurately and efficiently explores conformational space with improved sampling and converges to the correct canonical probability distribution.

6.2 Theory

The general idea behind the accelerated molecular dynamics scheme is depicted in figure 6.1. A continuous non-negative bias boost potential function $\Delta V(r)$ is defined such that when the true potential $V(r)$ is below a certain chosen value E, the boost energy, the simulation is performed on the modified potential $V^*(r) = V(r) + \Delta V(r)$, represented using dashed lines, and when $V(r)$ is greater than E, the simulation is performed on the true potential $V^*(r) = V(r)$. This leads to an enhanced escape rate from local minima when the simulation is performed on $V^*(r)$. The modified potential $V^*(r)$

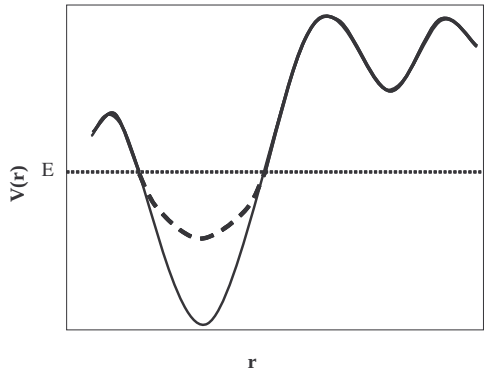

Figure 6.1: Schematic representation of the unmodified potential (solid line), the biased potential (broken line), and the threshold energy, E.

is related to the true potential, bias potential, and boost energy by

$$V^*(r) = \begin{cases} V(r), V(r) \geq E \\ V(r) + \Delta V(r), V(r) < E \end{cases} \tag{6.1}$$

During conventional molecular dynamics simulations of biological molecules on the unmodified potential surface, the systems extensively sample conformations around a local minimum without adequately sampling conformations elsewhere on the potential energy surface. Therefore, the primary goal of this work is to develop a method for large biological systems that is capable of accelerating the state to state evolution of a system relative to normal molecular dynamics. The bias potential increases the escape rate of the system from potential basins, and the subsequent state to state evolution of the system on the modified potential occurs at an accelerated rate with a non-linear time scale of Δt^*, where

$$\Delta t_i^* = \Delta t e^{\beta \Delta V[r(t_i)]} \tag{6.2}$$

This allows advancement of the clock at each step depending on the strength of $\Delta V(r)$, where Δt is the actual time step of the simulation on the unmodified potential. Hence, the total estimated simulation time becomes a statistical property and is given by equations 6.3 and 6.4.

$$t^* = \sum_i^N \Delta t_i^* = \Delta t \sum_i^N e^{\beta \Delta V[r(t_i)]} \tag{6.3}$$

$$t^* = t\langle e^{\beta\Delta V[r(t_i)]}\rangle \tag{6.4}$$

where N is the total number of molecular dynamics steps carried out during the whole simulation, and $\langle e^{\beta\Delta V[r(t_i)]}\rangle$ is termed the boost factor. The boost factor is a measure of the extent to which the simulation has been accelerated. At each step, the time step, Δt^*, is non-linearly dependent on the value of the bias potential, $\Delta V(r)$. It follows from equation 6.2 that $\Delta t^* = \Delta t$ when the system is on the true potential, $V(r)$, that is when $\Delta V(r) = 0$. If the choice of the boost energy E is very high, then the boost factor will be very large, leading to noisy statistics because the wells will not be sampled sufficiently. However, correct statistics will be obtained after many transitions and adequate sampling of the potential energy wells. Note that the times described here represent relative times spent sampling different portions of the conformational ensemble; in general, kinetic properties measured when the simulation is performed on $V^*(r)$ will not be correct.

It is important that this method yields correct canonical averages of an observable $A(r)$, so that thermodynamics and other equilibrium properties can be accurately determined from accelerated MD simulations. The equilibrium ensemble average value of any observable $A(r)$ on the normal potential $V(r)$ is given by

$$\langle A\rangle = \frac{\int drA(r)e^{-\beta V(r)}}{\int dre^{-\beta V(r)}} \tag{6.5}$$

Similarly, the ensemble average value of any observable $A(r)$ taken on the modified potential can be written as

$$\langle A^*\rangle = \frac{\int drA(r)e^{-\beta V^*(r)}}{\int dre^{-\beta V^*(r)}} \tag{6.6}$$

substituting for $V^*(r)$, yields

$$\langle A^*\rangle = \frac{\int drA(r)e^{-\beta V(r)-\beta\Delta V(r)}}{\int dre^{-\beta V(r)-\beta\Delta V(r)}} \tag{6.7}$$

Re-weighting the phase space of the modified potential by multiplying each configuration by the strength of the bias at each position, results in equations 6.8 and 6.9, the corrected ensemble average, which is equivalent to the equilibrium observable of $A(r)$ on the normal potential.

$$\langle A^C\rangle = \frac{\int drA(r)e^{-\beta V(r)-\beta\Delta V(r)}e^{\beta\Delta V(r)}}{\int dre^{-\beta V(r)-\beta\Delta V(r)}e^{\beta\Delta V(r)}} \tag{6.8}$$

$$\langle A^C \rangle = \frac{\int dr A(r) e^{-\beta V(r)}}{\int dr e^{-\beta V(r)}} = \langle A \rangle \tag{6.9}$$

Therefore, it can be seen that the accelerated molecular dynamics simulation method converges to the canonical distribution, and the corrected canonical ensemble average of the system is obtained by simply re-weighting each point in the configuration phase space on the modified potential by the strength of the Boltzman factor of the bias energy, $e^{\beta \Delta V[r(t_i)]}$, at that particular point. When the system is on the normal potential, the bias is zero.

Various approaches on how to define the bias potential, $\Delta V(r)$, have been studied.[144, 145, 149, 150] The bias or boost potential, $\Delta V(r)$, was defined as $E - V(r)$ by Rahman and Tully,[148] such that the modified potential becomes $V^*(r) = E$: a flat modified potential surface that they termed "puddles" covering energy wells. This implementation is very simple and computationally inexpensive, because the force while on the "puddle" potential is zero. However, there are some problems associated with this choice: the derivatives of the potential, $dV(r)/dr$, are discontinuous at points where the unmodified potential, $V(r)$, merges with the modified potential, $V^*(r)$, that is where $V(r) = E$. Therefore, Rahman and Tully devised a special integration technique used to traverse points where $V(r) = E$, which subtantially reduces the gain in computational efficiency. Also, at high values of the boost energy, E, the flat modified potential energy surface is raised above most transition state regions. The majority of the potential energy surface becomes flat, and the system experiences a random walk. Under these conditions, the system converges very slowly.

In contrast to the flat "puddles" employed by Rahman and Tully to fill energy minima, the modification of the potential energy surface proposed here is more akin to snow drifts. These "snow drifts" smooth the landscape by filling minima, but maintain the underlying shape of the unmodified potential energy surface and merge smoothly with the original potential at the threshold "boost energy" value E. Therefore, $\Delta V(r)$ is chosen such that the first derivative of $V^*(r)$ has no discontinuity, and the modified potential reproduces the shape of the minima even at high values of E. The choice of $\Delta V(r)$ is

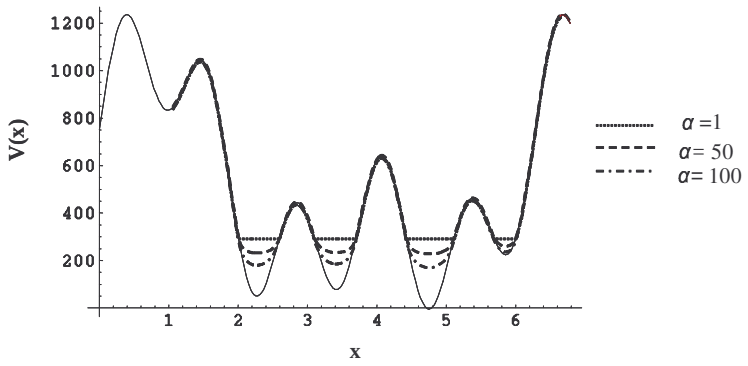

Figure 6.2: Schematic representation of a hypothetical potential energy function and several bias potentials plotted at various values of α and a relatively low value of the threshold boost energy, E.

given by

$$\Delta V(r) = \frac{(E - V(r))^2}{\alpha + (E - V(r))} \tag{6.10}$$

where α is a tuning parameter that determines how deep the modified potential energy basin will be. When α is zero, the modified potential is flat, $V^*(r) = E$, and equivalent to that adopted by Rahman and Tully.

Selection of E and α is important to the method and determines how aggressively the molecular dynamics will be accelerated. Therefore, appropriate choices of E and α are suggested by examining the effect of the bias potential on a hypothetical one-dimensional energy function at various values of E and α as shown in figures 6.2 and 6.3. One prescription for choosing E is that it should be greater than the local minima of $V(r)$, $V_{\min}$, near the starting structure. If E is less than $V_{\min}$, then the simulation will always be performed on the true potential which is simply a conventional MD simulation. Furthermore, since large molecules tend to have multiple minima very close together, calculating an average potential energy, $\langle V[r(t_i)]\rangle$, on the true potential over a short period of time starting with the initial structure, and using that as the minimum, $V_{\min}$, is an effective strategy. At low values of E as represented by the potential energy profiles presented in figure 6.2, the modified potential falls below most of the transition state regions and the relative probability of escape remains the same for the modified and unmodified potentials. Therefore the choice of α is not that critical to the overall potential

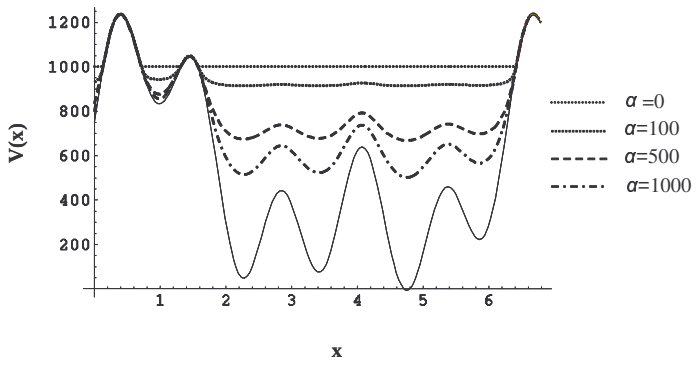

Figure 6.3: Schematic representation of a hypothetical potential energy function and several bias potentials plotted at various values of α and a relatively high value of the threshold boost energy, E.

energy landscape as long as α is not so small that the modified potential becomes flat (figure 6.2; $\alpha = 1$). Flat modified potential surfaces cause the calculated force to be discontinuous at points where $V^*(r) = E$.

On the other hand, when E is chosen to be high (figure 6.3), the value of α becomes important because at low values of α, as in the case when α = 0, the modified potential becomes isoenergetic in most places, and the molecular system experiences a random walk. Also, as discussed earlier, the first derivative of the modified potential becomes discontinuous at points where the modified potential is equal to E. Therefore, in order to maintain the basic shape of the potential energy surface at high values of E, and preserve the same potential energy wells that are present on the unmodified potential surface, α has to be set to a much higher value than zero. As shown by the plot in figure 6.3, α seems to produce the desired effect when it is set to a value close to $E - V_{\min}$ (figure 6.3; α = 1000). Therefore, E should be chosen to be greater than $V_{\min}$, with the magnitude depending on how aggressively one wants to sample the conformational space. A choice of $\alpha = E - V_{\min}$ will allow the modified potential energy surface to echo the shape of the potential wells and merge smoothly with the original potential.

6.3 Methods and Applications

Molecular dynamics simulations were carried out using the Cornell et al.[151] all-atom force field as shown in its simple form in equation 6.11:

$$V(r) = \sum_{bonds} K_r(r - r_{eq})^2 + \sum_{angles} K_\theta(\theta - \theta_{eq})^2 + \sum_{dihedrals} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}}\right] + \sum_{i<j} \frac{q_i q_j}{\varepsilon R_{ij}} \tag{6.11}$$

where the summations represent the harmonic bond and angle energy terms, the dihedral torsions, and nonbonded van der Waals and electrostatics interactions terms respectively. Since conformational changes in proteins involve changes in torsions to a much greater extent than any other degrees of freedom, this accelerated MD method has been applied to the sum of the dihedral torsions and the 1–4 non-bonded interactions as shown in equation 6.12, instead of the whole potential.

$$V_D(r) = \sum_{dihedrals} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}}\right]_{1-4} + \sum_{i<j} \left[\frac{q_i q_j}{\varepsilon R_{ij}}\right]_{1-4} \tag{6.12}$$

Therefore, equations 6.1 and 6.10 become:

$$V^*(r) = \begin{cases} V_o(r) + V_D(r), V_D(r) \geq E_D \\ V_o(r) + V_D(r) + \Delta V_D(r), V_D(r) < E_D \end{cases} \tag{6.13}$$

$$\Delta V_D(r) = \frac{(E_D - V_D(r))^2}{\alpha + (E_D - V_D(r))} \tag{6.14}$$

where $V_o(r)$ is the sum of the interaction potential without that of the dihedral torsions and the non-bonded 1–4 interactions, and E_D is the threshold energy or boost energy that is analogous to E.

Because this study was carried out using implicit solvation, the collisions of the molecular system with solvent are approximated by using the Langevin dynamics equation

$$m\frac{d^2x}{dt^2} = F(t) - \gamma m\frac{dx}{dt} + R(t) \tag{6.15}$$

where γ is the collision frequency, γm is the frictional coefficient, and R(t) is a random Gaussian force with zero mean. A value of 2.0 ps^{-1} was used for the collision frequency

as suggested by Loncharich et al.[152] The electrostatic interaction was treated using the generalized Born[35,51] implementation (igb = 2) in AMBER 7,[153] and the apolar solvation term was also included in the potential function with the surface tension parameter set to the default value of 0.005 kcal/mol Å^2. The SHAKE algorithm[154] was applied to all bonds involving hydrogen atoms, and an integration time step of 2.0 fs was used for the integration of the Langevin equation. All calculations were carried out using a version of the sander module in the AMBER 7 suite of programs that was modified to perform the accelerated molecular dynamics simulation. Several accelerated molecular dynamics simulations were carried out using various values of E_D. During the accelerated molecular dynamics simulations the configuration correction term was calculated at each step.

6.4 Results and Discussion

6.4.1 Correct Canonical probability distribution

Presently, conventional molecular dynamics simulations of biomolecules are generally unable to sufficiently sample the conformational space. Therefore, the proposed accelerated molecular dynamics approach extends the time scale of the simulation without prior knowledge of the potential energy surface. In order for this approach to be effective, the first question that needs to be addressed is whether this approach reproduces the canonical probability distribution after re-weighting the statistics of the accelerated molecular dynamics simulations. A normal MD simulation (with no bias potential) and several accelerated MD simulations using blocked alanine dipeptide (figure 6.4) were performed. This system was chosen because it is small, so complete conformational sampling is computationally feasible, and it represents the essential attributes of the backbone conformations seen in proteins.

One normal MD simulation and three accelerated MD simulations were carried out on the blocked alanine dipeptide. $\langle V_D(r) \rangle$ was calculated to be approximately 60 kcal/mol, and the values of the boost energy, E_D, and the tuning parameter, α, for the three accelerated simulations were set at 80, 90, and 100, and 20, 30, and 40 kcal/mol

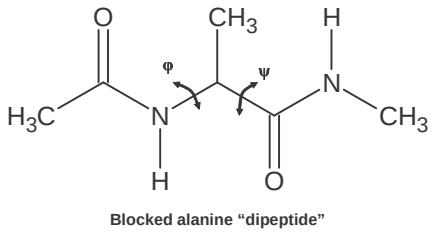

Figure 6.4: Alanine dipeptide

respectively. The simulations were carried out at 800K, so that the simulation could achieve convergence and to ensure sufficient transitions over high potential energy barriers. Each simulation was carried out over 5 x 10^7 steps of MD simulation (equivalent to 0.1 μs of normal MD simulation), and snapshots were collected for analysis every 50 steps.

The conformational free energy plots shown as a function of backbone torsional angles are depicted in figure 6.5. The plot of the normal MD simulation represented in figure 6.5a shows that the alpha-helical region with the broadest energy minimum, where ϕ is less than 0 and ψ is between 0 and -60, is strongly populated when compared to other regions. Figures 6.5b-d show the free energy plot as a function of the backbone torsional angles of the three accelerated molecular dynamics simulation as E_D is increased from 80 to 100 kcal/mol. Also, it can be seen that the plots of the accelerated MD simulations shown are quite similar to that of the normal MD simulation in figure 6.5a. Therefore, it can be inferred that this approach leads to the proper calculation of the canonical distribution after re-weighting of the conformational space, and that the potential energy wells are accurately sampled for the accelerated MD simulations.

As previously discussed, the choice of α is very important in preserving the underlying shape of the potential energy surface at high values of E_D, so that the potential energy wells can be adequately sampled, as can be seen for a relatively high value of E_D in figure 6.5d. In order to thoroughly investigate the effect of α on the shape of the underlying potential energy landscape and the implication on the statistics generated, two additional sets of four accelerated MD simulations were carried out with varying values of α while keeping E_D constant at a relatively high value of 100 kcal/mol and low value

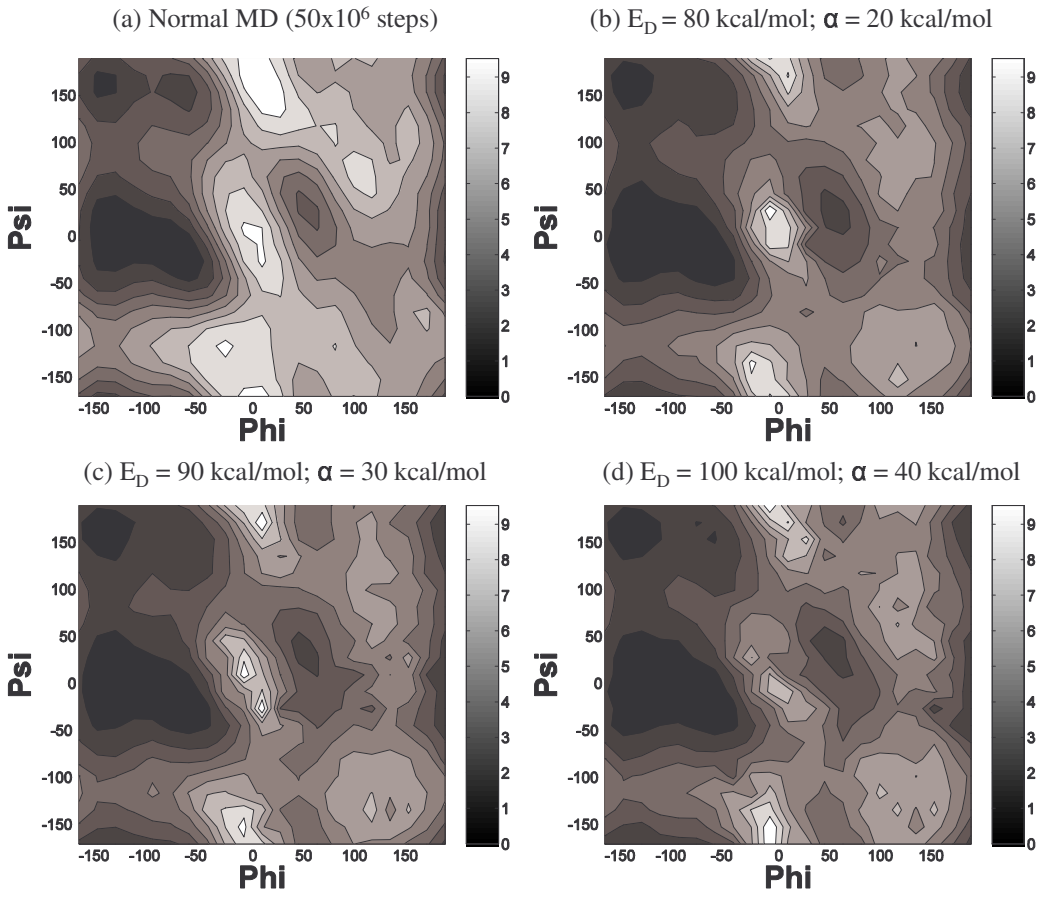

Figure 6.5: Alanine dipeptide backbone torsional free energy surface (kcal/mol) derived from histogram of snapshots from normal MD simulation and several accelerated MD simulations at various values of E_D.

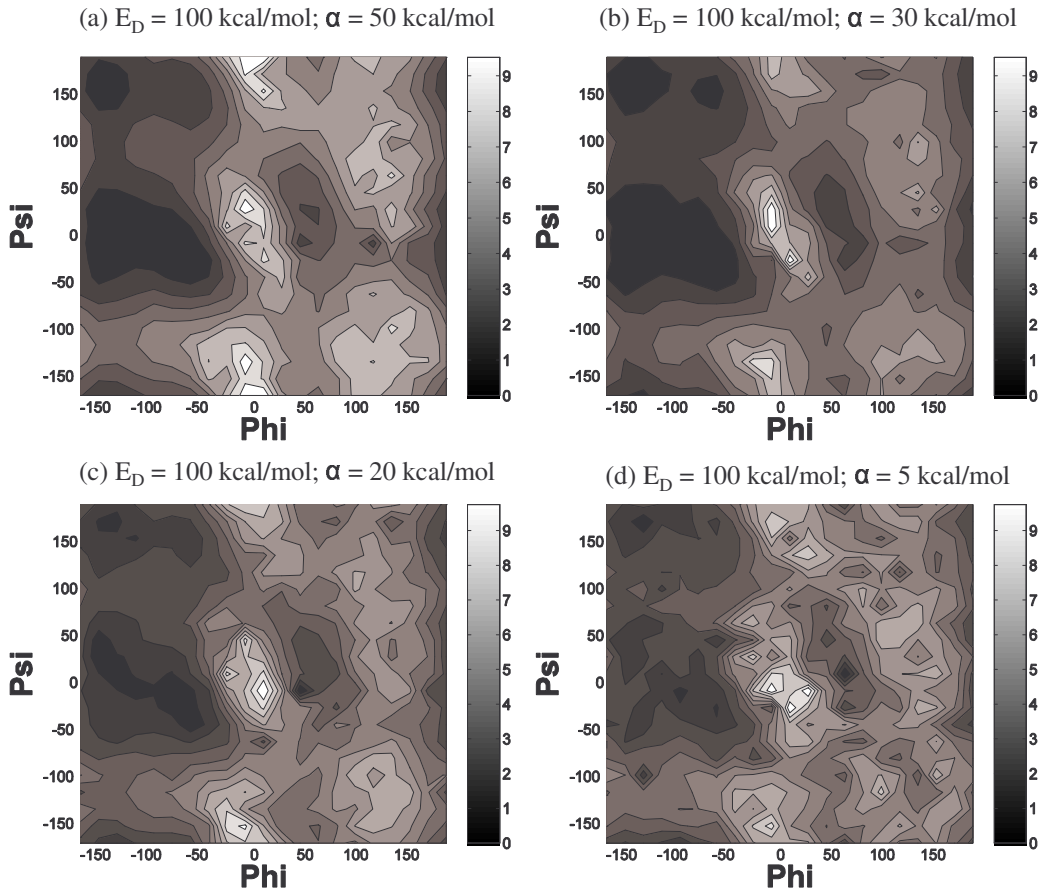

Figure 6.6: Alanine dipeptide backbone torsional free energy surface (kcal/mol) derived from histogram of snapshots from several accelerated MD simulations at with high E_D(100 kcal/mol) and a range of α values: (a) 50, (b) 30, (c) 20, and (d) 5 kcal/mol.

of 80 kcal/mol. At a relatively high value of E_D and high values of α, figure 6.6a, the plot is very similar to that of the normal MD simulation and represents the correct distribution of blocked dipeptide alanine. However, as α is decreased, the statistics generated become noisy and give rise to a spotty density plot (figure 6.6d). The noisy statistics at high E_D and low α are due to the modified potential energy landscape becoming flat and isoenergetic (figure 6.3). On this flat, isoenergetic potential surface, alanine dipeptide experiences a random walk and does not sufficiently sample the potential energy minima. The potential energy minima are not well defined on the flat modified potential surface, and the statistics are dominated by a few heavily weighted points.

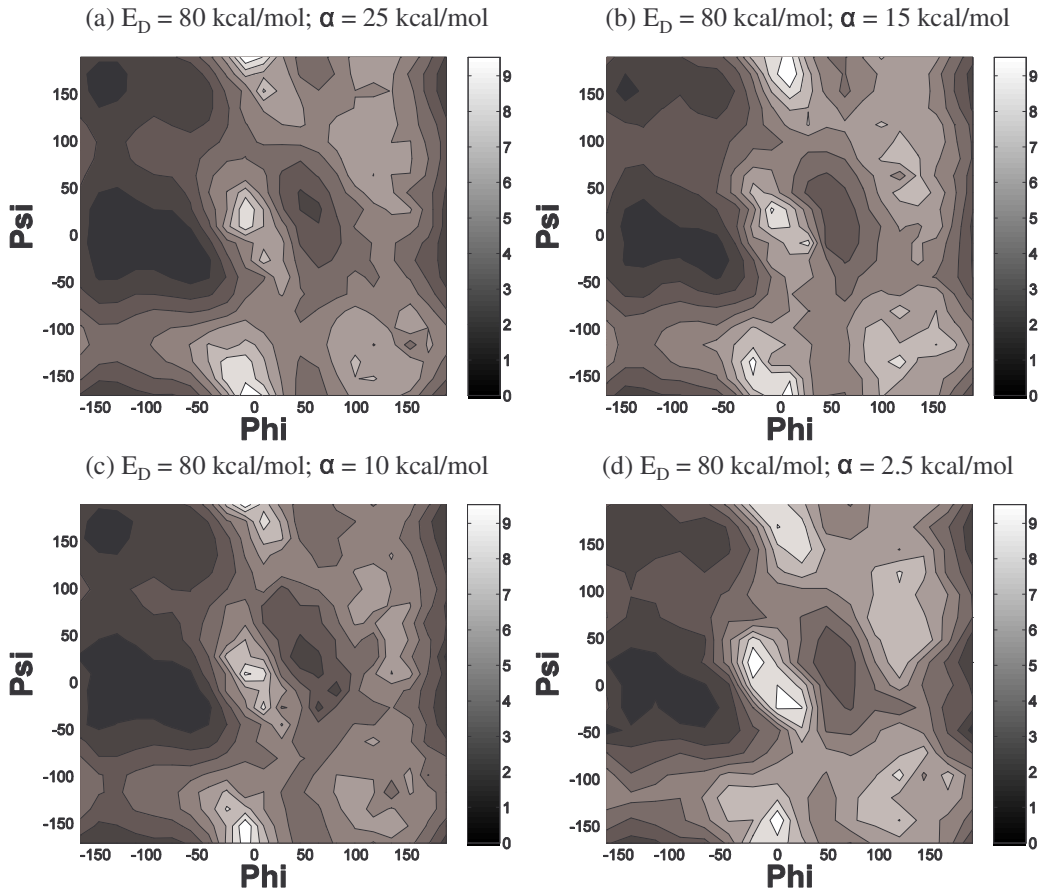

Figure 6.7: Alanine dipeptide backbone torsional free energy surface (kcal/mol) derived from histogram of snapshots from several accelerated MD simulations at with low E_D(80 kcal/mol) and a range of α values: (a) 25, (b) 15, (c) 10, and (d) 2.5 kcal/mol.

On the other hand, when E_D is relatively low as shown in figure 6.2, the re-weighted probability distributions for several values of α (high and low) converge to the correct canonical distribution (figure 6.7) and are quite similar to that of the normal MD simulation (figure 6.5). Low values of E_D position the modified surface below most of the transition state regions (figure 6.2), thus maintaining the accurate sampling of each well with the correct probability irrespective of the choice of α. Therefore, at relatively low values of E_D (figure 6.7) the potential energy wells are accurately sampled and the re-weighted free energy plots are similar to that of the unmodified potential.

6.4.2 Enhanced Sampling

This chapter has introduced an approach to extend the time scale of MD simulations and has shown that it converges to the correct canonical probability distribution. Another question that needs to be answered is whether this technique accelerates the sampling of the conformational space. This question has been addressed by carrying out a normal MD simulation and several accelerated molecular dynamics simulations at various values of E_D on hepta-alanine starting with the helical structure, since that is the predominant configuration for polyalanine. These simulations were conducted at 300K and then repeated at 400K. Each simulation was carried out over 5×10^6 steps of MD simulation (equivalent to 10 ns of normal MD simulation), with snapshots taken for analysis every 100 steps. The relative conformational free energy plots of the backbone angles of the third (ALA 3) and fourth (ALA 4) residues for the simulations are plotted in figures 6.8 and 6.9 respectively. During the normal MD, the peptide stayed close to the starting structure and sampled mainly the alpha-helical conformations. As E_D is increased, other potential energy wells are sampled that are not observed in the normal MD. It can be seen that after re-weighting, the alpha-helical conformation, which is known to be the predominant species in solution, tends to be sampled heavily. The conformational space of the peptide is extensively explored using the accelerated MD method, and more configurations are sampled as the value of the boost energy, E_D, is increased. Therefore, the higher the boost energy E_D is set, the more aggressive the sampling becomes. Similar effects are observed at 400 K. This is reflected in the extent of acceleration of the molecular dynamics, as estimated by the boost factor calculated for each simulation using equations 6.3 and 6.4.

Another interesting and informative way of looking at the simulation results is by performing a multivariate analysis such as principal components analysis (PCA), as described in chapter 7, on the concatenated snapshots of the normal and accelerated MD simulations. The positional covariance matrix of the Cartesian atomic coordinates of the backbone atoms of the three central alanine residues was calculated and then diagonalized to provide a set of eigenvectors representing different modes of conformational change and their corresponding eigenvalues. Each eigenvalue indicates the relative con-

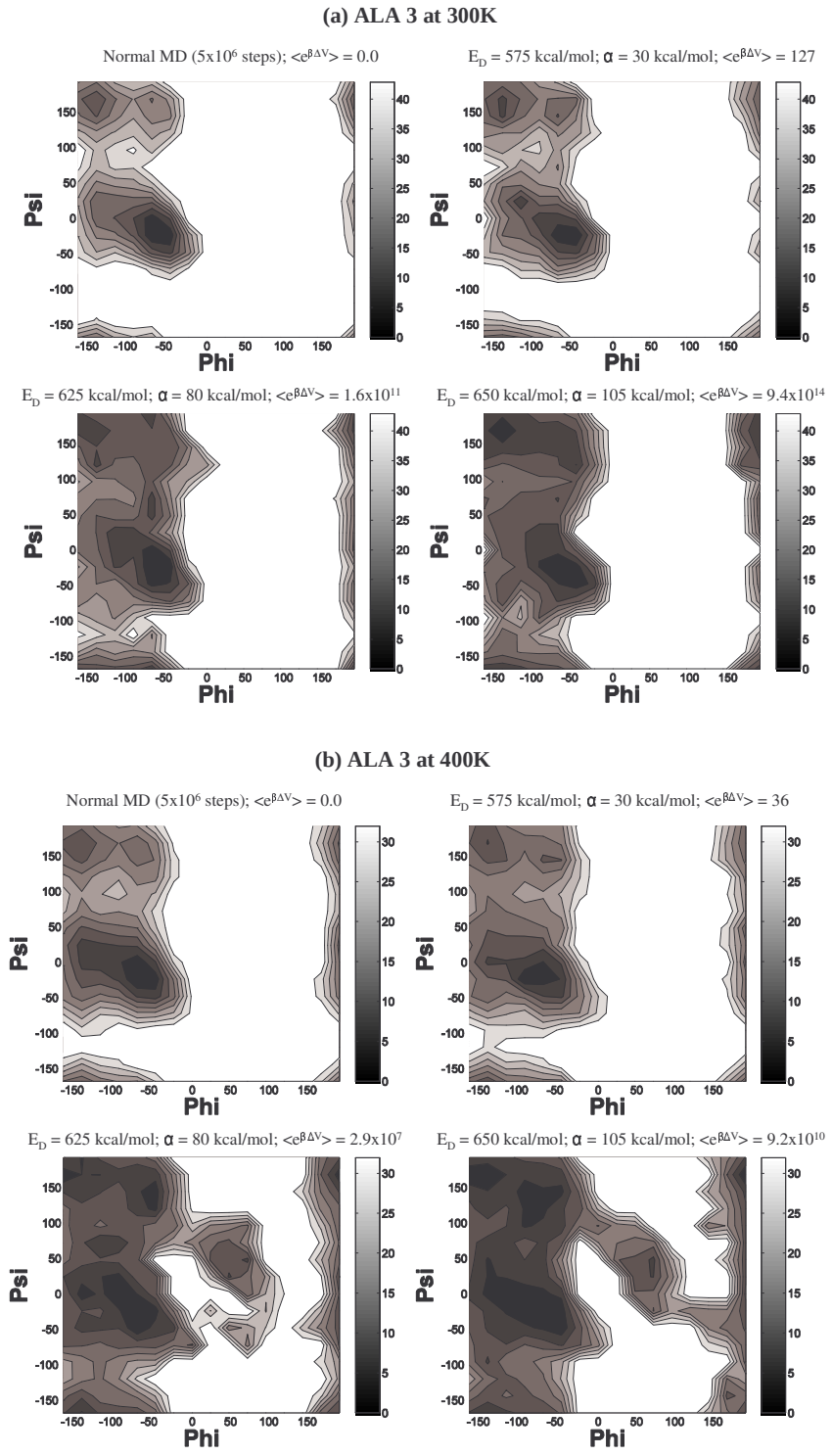

Figure 6.8: Backbone torsional free energy surface (kcal/mol) for ALA-3 of hepta-alanine derived from histogram of snapshots from normal and accelerated MD simulations at (a) 300K and (b) 400K.

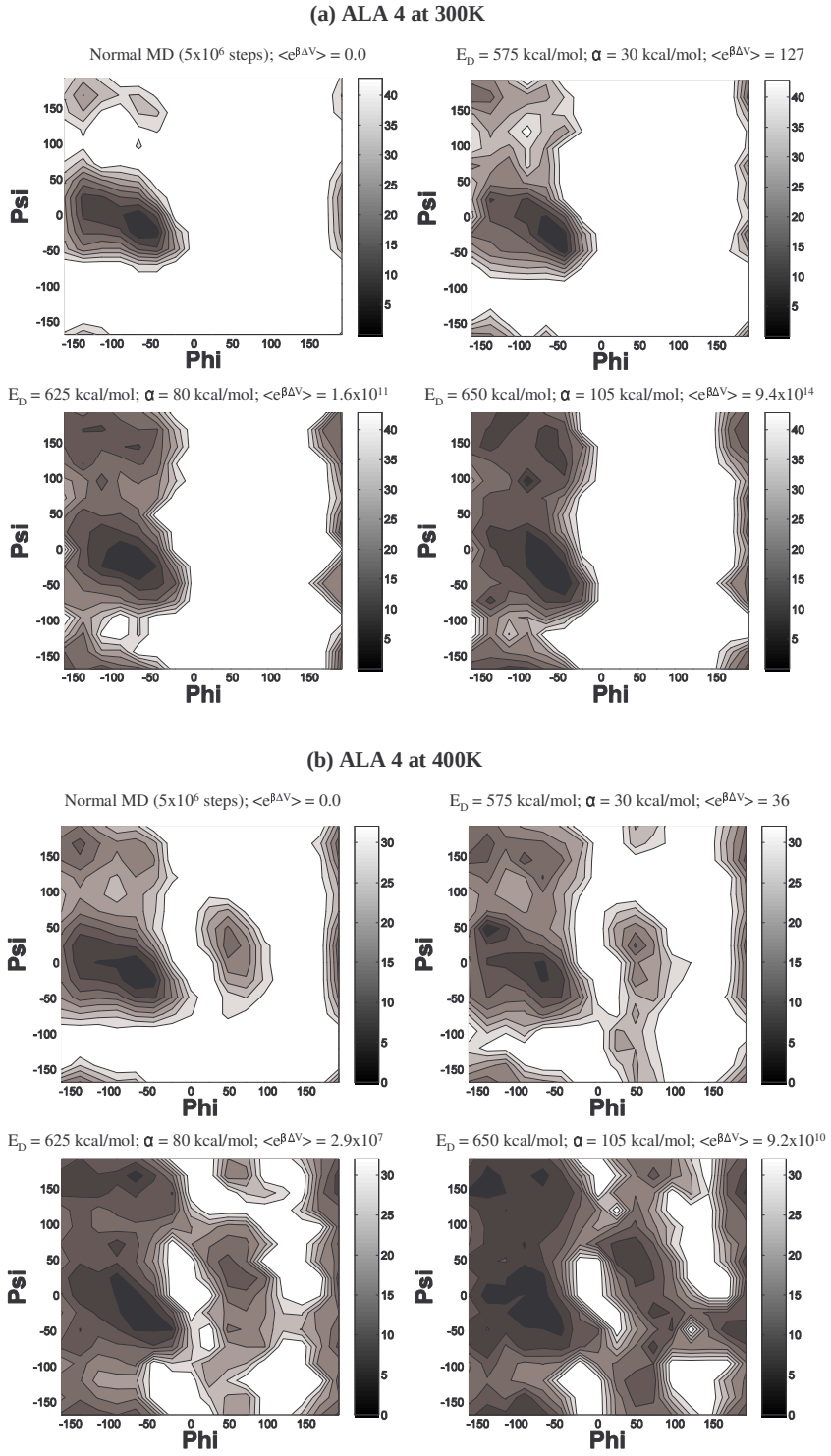

Figure 6.9: Backbone torsional free energy surface (kcal/mol) for ALA-4 of hepta-alanine derived from histogram of snapshots from normal and accelerated MD simulations at (a) 300K and (b) 400K.

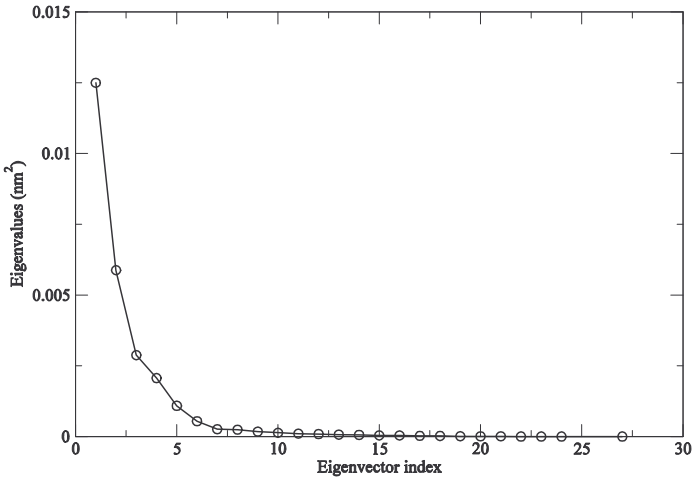

Figure 6.10: Eigenspectrum for principal component analysis of central three residues of hepta-alanine from accelerated MD trajectory with $E_D = 650$ kcal/mol.

tribution of the corresponding mode to the overall dynamics. The ranking of the eigenvalues is shown in figure 6.10, and it can be seen that the first two modes are the most dominant and contribute over 70% to the overall motion. Therefore, projections of the trajectory on the first and second modes for the normal and accelerated MD simulations at 300 K were analyzed and plotted in figure 6.11 as unweighted scatter plots and reweighted density plots. The multidimensional phase space reduced to 2D sampled by the peptide during the normal MD simulation is smaller than that sampled during the accelerated MD simulations as is evident from figure 6.11. Conformations from the normal MD simulation fall into two clusters, while the accelerated MD simulations sample these two clusters as well as two others not seen in the normal MD simulation.

For thorough insight into the sampling of the accelerated MD simulations and the conformations that are present in each cluster of the plots of the two dominant modes in figure 6.11, the plot of the accelerated MD when E_D = 650 kcal/mol was quantitatively separated into four clusters using a simple k-means clustering algorithm, as shown in figure 6.12. Furthermore, the phi and psi backbone angles of the third and fourth residues for the conformations in each cluster were plotted in figure 6.13. The cluster shown in black represents helical conformations where the backbone angles of the third and fourth residues are in the alpha helical region as shown in figure 6.13. The normal MD

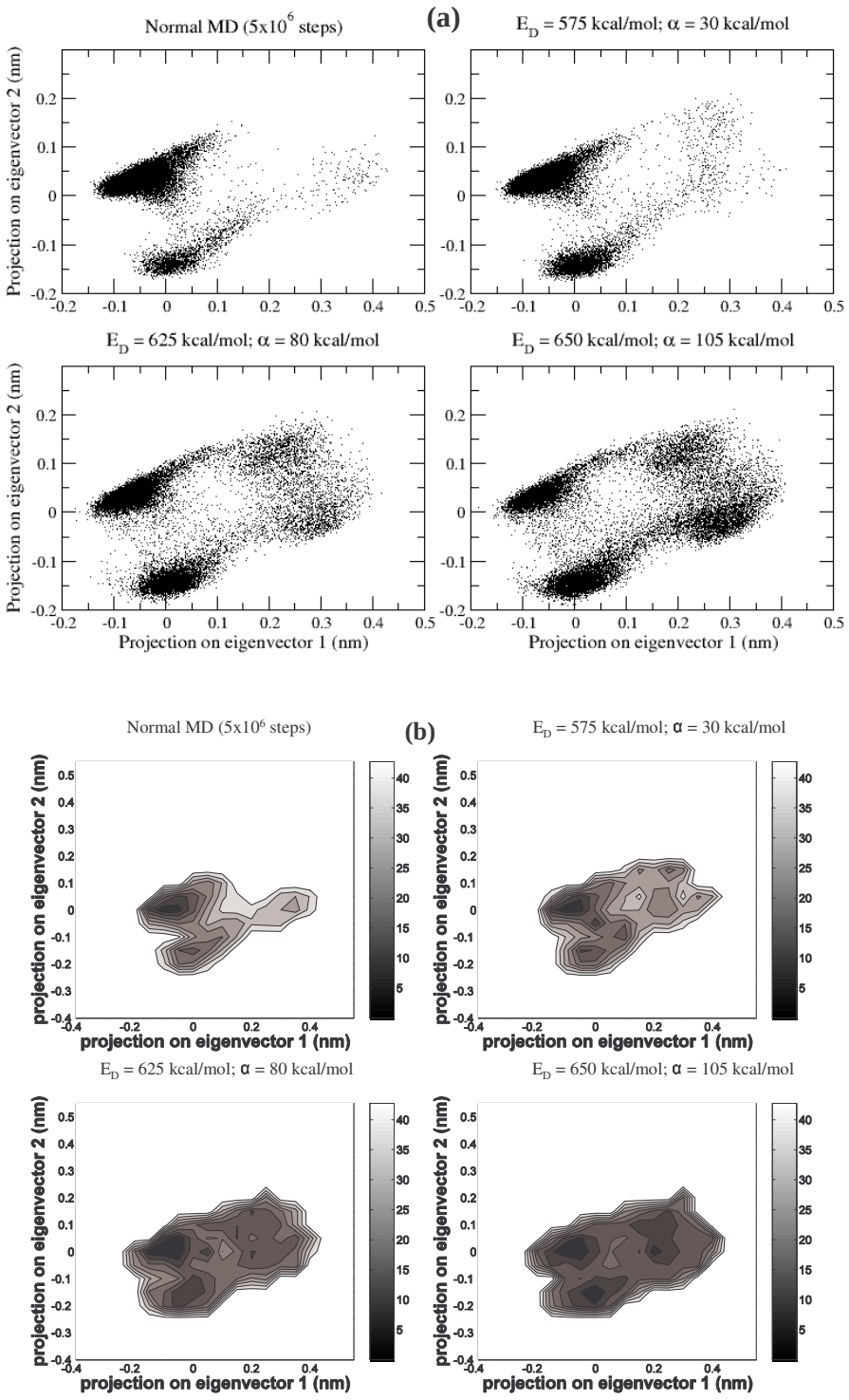

Figure 6.11: Projection of snapshots from normal and accelerated MD of hepta-alanine onto first and second principal components. Principal components were calculated based on the combined trajectories. Projections presented as (a) unweighted scatter plots and (b) Boltzmann reweighted density plots.

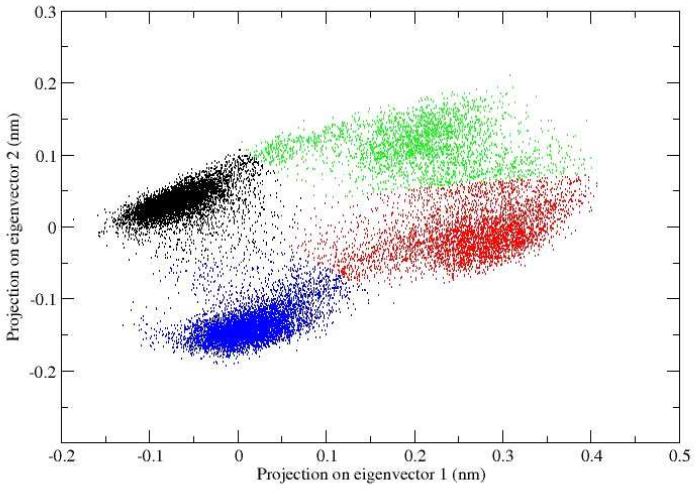

Figure 6.12: *k*-means conformational clustering of snapshots from accelerated MD of hepta-alanine.

simulation is predominately made up of the black and blue clusters. The blue cluster is primarily made up of configurations with backbone angle of the fourth residue in the alpha helical region and that of the third residue in the extended strand region. In addition, the accelerated molecular dynamics simulations extensively sample other conformations not sampled by the normal MD simulation that fall in two additional clusters shown in figure 6.12 as red and green. The green cluster represents structures with backbone conformations of the fourth residue in the extended strand region and that of the third residue in the alpha helical region. Fully extended structures with the backbone conformations of the fourth and third residues in the extend strand region on the phi/psi map fall in the red cluster. Therefore, from the above results, it can be seen that the accelerated molecular dynamics method samples the phase space more extensively than the normal molecular dynamics. Conformations that are not sampled by the normal molecular dynamics simulation are seen in the accelerated MD simulations.

The accelerated molecular dynamics approach falls into the class of enhanced sampling methods in which the energy barriers are effectively lower, so the system transi-

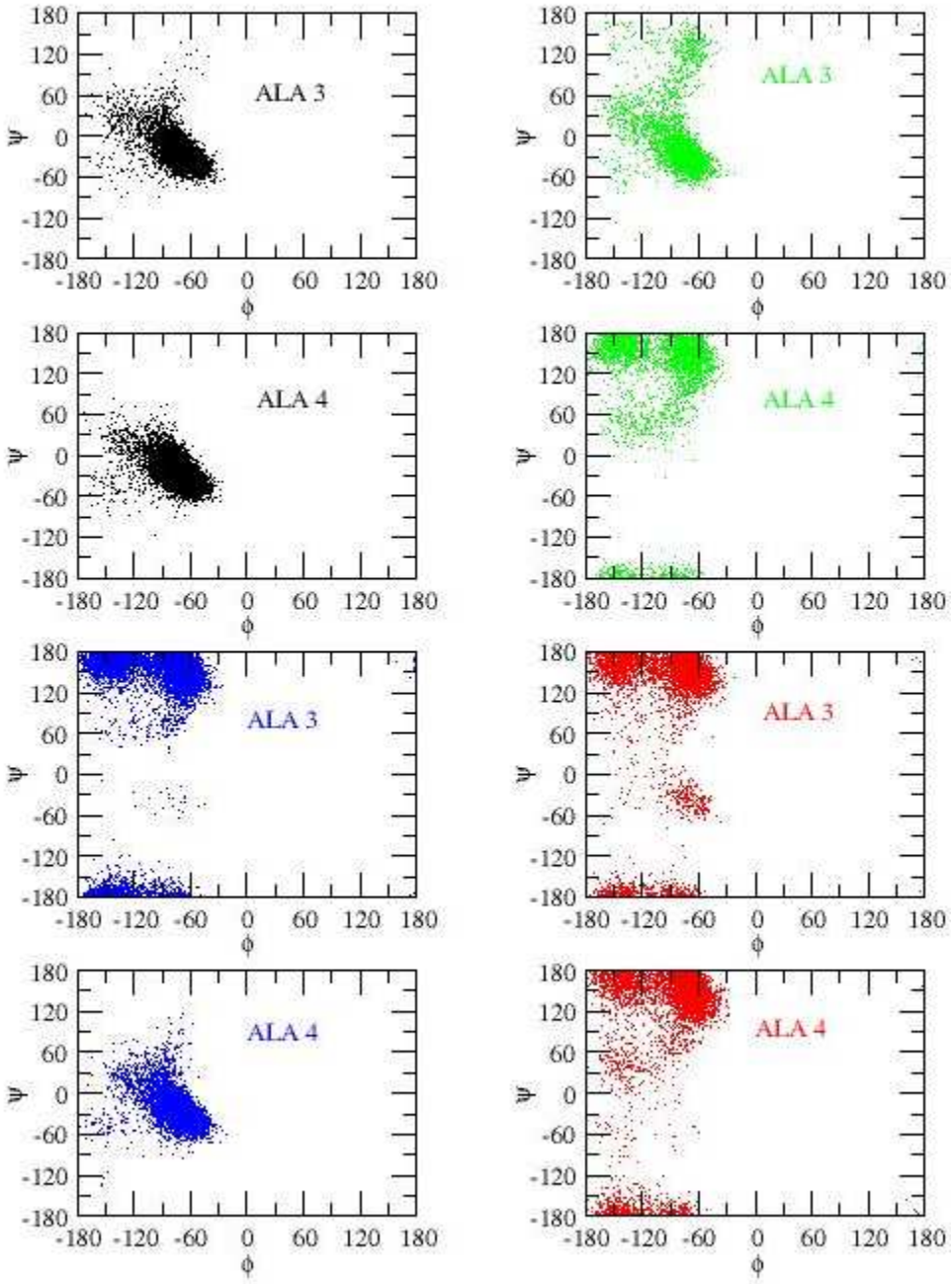

Figure 6.13: Backbone torsional angles sampled by ALA-3 and ALA-4 separated by conformational cluster.

tions between energy wells more rapidly. Two major advantages of the accelerated MD approach over many others are that little or no prior knowledge of the potential energy landscape is required, and that the boost is consistently applied throughout the potential energy surface. The latter property distinguishes this method from conformational flooding, wherein a "flooding potential"—similar to the bias potential—is added to the effective Hamiltonian of the system only around the configuration of the initial structure. This is done so as to only destabilize the initial configuration and allow the system to escape to another potential energy well in fewer computational steps.

Many simple techniques, including raising the temperature, have been devised to accelerate MD simulations and explore the conformational space of molecular systems. One of these techniques involves carrying out several short MD simulations on a particular system in which the starting conditions are different. Although some of these approaches may be quite useful to search for conformations, they may not be relied on to generate a Boltzmann distribution of conformations which is required for calculating thermodynamic quantities. Nonetheless, to fully assess the ability of the accelerated MD approach to explore the conformational space, multiple short normal MD simulations starting with the same initial structure but with different initial velocities have been carried out. The 5×10^6 steps of normal MD was split up into five 1×10^6 steps of normal MD simulations and then combined to analyze the backbone conformations of the third and fourth residues. This technique slightly improves the sampling of the conformational space, but not as extensively as the accelerated MD simulations (figures 8 and 9). The sampling of the conformational space can further be enhanced by starting the five different simulations with different conformations, but the combined trajectories will be unlikely to lead to a Boltzmann distribution of conformations. The accelerated molecular dynamics method not only has the ability to extensively sample the conformational landscape, but also results in the generation of a Boltzmann distribution of conformations.

6.5 Conclusion

Computational methods like molecular dynamics simulation are the only techniques that can be used to follow the time evolution of biological molecules, because there are presently no experimental techniques that can track precise dynamics in atomic detail. However, the timescale of molecular dynamics simulations is currently limited to nanoseconds, and simulations of biomolecules appear to be nonergodic because transitions between energy wells are rare, due to high energy barriers. This causes incomplete sampling for nanosecond length simulations. The approach presented here eases this problem by increasing the escape rate of a molecular system from potential energy wells while still accurately sampling the conformational space. By defining a simple and robust bias potential which raises the potential energy surface in regions where conventional molecular dynamics simulations spend many computational steps, it has been shown that molecular dynamics can be accelerated with a defined boost factor and converge to the correct canonical probability distribution. The definition of the bias potential echoes the shape of the potential energy landscape even at high boost energy, E_D, thus allowing the potential energy wells to be accurately sampled. Also, this approach is computationally efficient: a single step of accelerated molecular dynamics requires less than 5% more computation than a normal MD step, but on average represents far more conformational sampling. Therefore, present nanosecond timescale simulations of large biological systems can be accelerated greatly in approximately the same amount of computational time.

In this study, the dihedral torsions alone were accelerated, since conformational changes in proteins mostly involve changes in torsions. However, the acceleration could be applied to regions of the potential energy function that correspond to the degrees of freedom that are significantly responsible for changes to the configurations of the system under consideration.

This chapter is a reprint in full of material that appeared in *Accelerated Molecular Dynamics: A promising and efficient simulation method for biomolecules*. Donald Hamelberg, John Mongan and J. Andrew McCammon. Journal of Chemical Physics **120**(24), 11919-29, June 2004. I was the secondary author and researcher for this work.

Chapter 7

Interactive Essential Dynamics

ABSTRACT

Essential dynamics is a useful method for analyzing trajectories generated by molecular dynamics, but current tools are awkward to use, limiting the usefulness of the technique. This paper describes a new interactive graphical interface for visualization of essential dynamics results, including filtering a trajectory on an arbitrary set of eigenvectors and manipulation of a structure's projection along any eigenvector.

7.1 Introduction

Trajectories generated from molecular dynamics (MD) simulations provide a means to identify and study motions crucial for protein function.[155] Separating functionally important motions from random thermal fluctuations is a major challenge in analyzing MD trajectories. Principal component analysis of MD trajectory data, often called *essential dynamics* (ED),[128,156] is frequently used to separate large-scale correlated motions from local harmonic fluctuations.[157–162]

ED analysis constructs a new orthogonal basis set for the atomic coordinates in a trajectory, such that the greatest variance occurs along the first vector, with monotonically decreasing variance along successive vectors. These vectors are often called principal components or eigenvectors, since their derivation involves an eigen decomposition.

The eigenvalues from the eigen decomposition represent the relative amount of molecular motion that occurs along each eigenvector. The eigenvalue spectrum is sharply peaked for molecular trajectory data, indicating that most of the molecular motion can be described by displacements along the first few eigenvectors.[128,157–159,162] A trajectory can be projected onto a subset of selected eigenvectors so only motion along the selected vectors is allowed. The most commonly selected subset is the first *n* eigenvectors such that a given percentage of the molecular motion occurs within the subspace formed by the selected eigenvectors. Projection onto these vectors filters out thermal noise, making the functionally interesting motions easier to appreciate. Smaller subsets may be selected to isolate a particular aspect of the molecule's motion. One can also examine the functional meaning of a single eigenvector by generating a trajectory with atomic positions interpolated between extreme projections on the selected eigenvector.

ED is a standard method of analysis that is widely implemented in molecular simulation packages.[113,122,129,163] Tools in these packages take trajectory and eigenvector files as input and produce a new trajectory as output, which must be loaded into an integrated[122] or external[113,129,163] viewer. A more flexible approach,[164] implemented within a limited viewer, is not widely available. In the available tools, a separate trajectory file of interpolations between extreme projections must be generated for each eigenvector, and a separate filtered trajectory file must be generated for each set of eigenvectors selected for filtering. Some tools[122,163] are limited to filtering along a single eigenvector at a time, which may be problematic since rotational motion cannot be adequately represented with a single eigenvector.

Generating and loading a separate trajectory file for each aspect of the ED results is cumbersome and discourages complete understanding of the ED analysis. Interactive Essential Dynamics (IED) is a new program that addresses these problems, providing fully interactive analysis of ED results through a graphical interface. Filtering eigenvectors can be rapidly added or removed from within the viewer, even while the trajectory is being played. The functional meaning of an eigenvector can be examined from within the viewer by dragging the atomic positions along the eigenvector using a slider control. Arrows representing an atom's motion along an eigenvector can be drawn to provide a static representation of an eigenvector, as in the work of Huitema and van Liere.[164] IED

can calculate eigenvectors and projections directly, or read the results of calculations performed in GROMACS[129] or the ptraj module of AMBER 8.[113] IED also allows sets of vectors that do not have accompanying projections to be loaded so results of normal modes analysis performed by AMBER or GROMACS can be visualized. The Python scripting interface of Visual Molecular Dynamics (VMD)[131] is used for display. The extensive visualization, animation, rendering and analysis capabilities of VMD remain available while using IED.

7.2 Theory and Methods

To perform ED, coordinate data from each timestep is fitted to a reference structure to remove translational and rotational motion. The fitted trajectory data are used to construct a covariance matrix C according to equation 7.1

$$C = \left\langle (\mathbf{x} - \langle \mathbf{x} \rangle)(\mathbf{x} - \langle \mathbf{x} \rangle)^{\mathrm{T}} \right\rangle \tag{7.1}$$

where $\langle\,\rangle$ represents the mean across all timesteps, and the T superscript represents transpose. An eigen decomposition (or diagonalization) of the symmetric matrix C is performed to identify Λ, a diagonal matrix of eigenvalues and T, a matrix of column eigenvectors forming a new orthonormal basis set,[128] satisfying

$$C = T \Lambda T^{\mathrm{T}} \tag{7.2}$$

A zero-mean trajectory matrix, X, can be constructed by subtracting $\langle \mathbf{x} \rangle$ from the coordinate vector for each timestep to form the rows of X. The matrix of the projections of each timestep onto each eigenvector, P, is obtained by multiplying the trajectory matrix, X by T

$$P = X T \tag{7.3}$$

For use with IED, these calculations may be performed using the AMBER or GROMACS suites. IED is also capable of performing these calculations itself, but is less efficient than AMBER or GROMACS.

The trajectory matrix, X, can be reconstructed from the eigenvectors and projection matrices T and P, by right multiplying equation 7.3 by T^{T}

$$P T^{\mathrm{T}} = X T T^{\mathrm{T}} = X I = X \tag{7.4}$$

where I is an identity matrix. More usefully, a matrix of filtered trajectory data, F, can be calculated by multiplying a subset of the (column) projection vectors in P by the corresponding subset of the eigenvectors in T^{T}. This way F contains only motions that occur along the eigenvectors selected from P and T, since motions along other eigenvectors are represented by projections omitted from the calculation of F. IED employs this method to calculate filtered trajectories, adding $\langle \mathbf{x} \rangle$ to the coordinate vector in each row of F to translate the coordinates back to their original origins. When a single eigenvector is to be examined by interpolation between extreme projections, coordinates are calculated by varying the (scalar) projection value for the selected eigenvector at the current time step and recalculating the appropriate row from F for each value of the projection.

When IED calculates ED directly, it can operate on trajectory data in any format that VMD is able to load. When loading results of ED analysis carried out in GROMACS, it requires a molecular topology file (in any VMD acceptable format), an eigenvectors file in GROMACS TRR binary format generated by g_covar, and a projections file generated by g_anaeig. The first timestep of the eigenvectors file is ignored, the second contains the molecule's average coordinates over the trajectory and the remaining timesteps contain eigenvectors in decreasing order of their eigenvalues. The projections file is formatted as text input to Grace, a plotting tool. Each eigenvector has a separate block of projection data within the file; within each block there is one projection per line, consisting of a time value followed by a projection value, separated by whitespace. When loading ED results from AMBER, the requirements are similar: a topology file, an eigenvectors file and a projections file. The eigenvectors file and projections file are both produced by ptraj and are in text format. The eigenvectors file contains two header lines, which are ignored, the average coordinates, and then the eigenvectors. Each eigenvector has a two line header consisting of a line containing 4 asterisks (****), followed by a line giving the ordinal number of the eigenvector and its eigenvalue. Numeric data

for the average coordinates and eigenvectors are whitespace delimited, with 7 values per line. The projections file has a two line header which is ignored. Each successive line contains a timestep number, followed by projection values onto each eigenvector for that timestep. The values are whitespace delimited.

Internally, IED represents the eigenvector data in a VMD trajectory object and the projection data in an Python Numeric array object. IED is an open source application, and is easily extended to other file formats by writing parsing routines to read data into the aforementioned data structures.

7.3 User Interface

IED is started either by selecting a trajectory in VMD for ED analysis, or by loading files containing the results of an ED analysis previously performed in ptraj or GROMACS. Once the ED data are loaded, a window is displayed with a checkbox and slider for each eigenvector (see figure 7.1). When necessary, the eigenvector slider area of the window can be scrolled to allow for arbitrarily large numbers of eigenvectors. Selecting a checkbox allows motion along the corresponding eigenvector and activates the eigenvector's slider, setting its position to the projection on the eigenvector for the current frame of the trajectory. Check boxes can be selected independently, allowing simultaneous analysis of any combination of eigenvectors. When the VMD animation controls are used to play the trajectory, the molecular display shows the filtered trajectory: the projection of the trajectory on the currently selected eigenvectors. Slider positions corresponding to selected eigenvectors are updated as each frame of the trajectory is displayed. The movement of the sliders provides an animated, graphical representation of the projection of the trajectory on each eigenvector. When the animation is stopped, the sliders for any selected eigenvector can be moved manually, which temporarily changes the projection value on the eigenvector for the displayed frame. The molecular display is updated as the slider is moved, making it easy to appreciate any eigenvector's contribution to the molecular motion. A comma delimited list of projections can be entered in the text box near the bottom of the window to rapidly set the projections along all

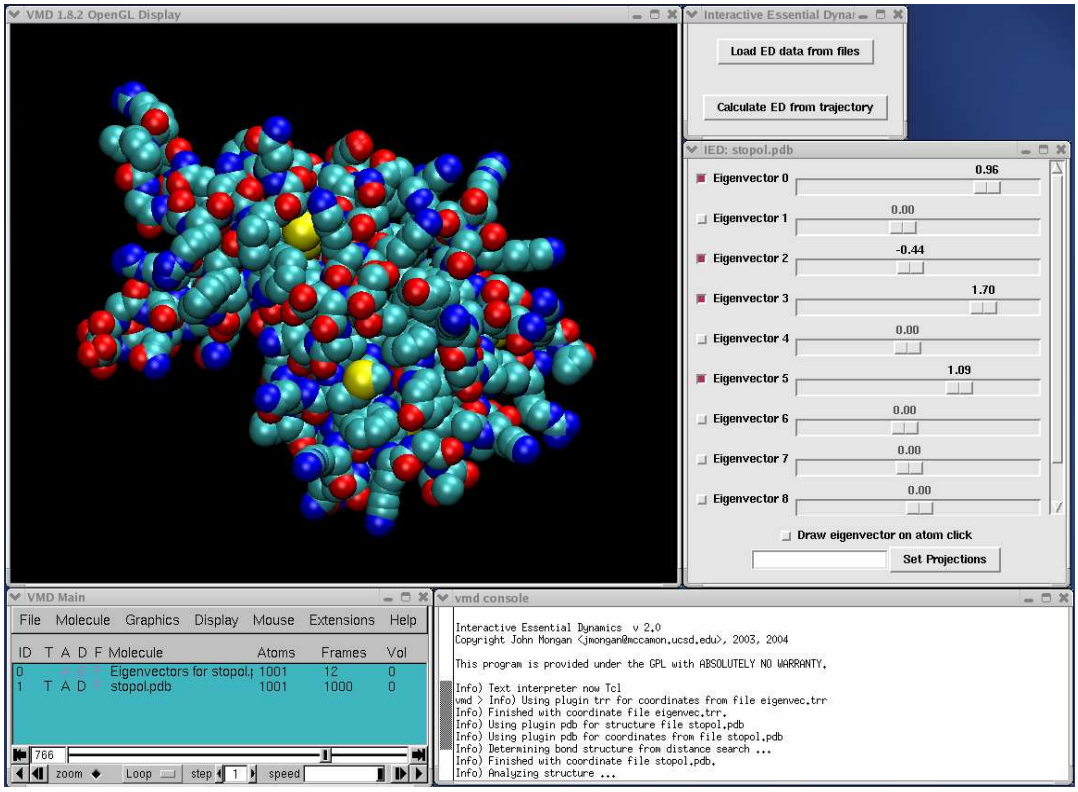

Figure 7.1: Screen shot of Interactive Essential Dynamics. Top right window is the main IED window, window immediately below contains the checkboxes and sliders for selecting eigenvectors and manipulating projections. Remaining windows are VMD windows: main window and animation controls at bottom left, console at bottom right and molecular display at top left.

eigenvectors.

Interactive manipulation of the molecule's projection along an eigenvector provides the clearest visualization of the eigenvector, but is not possible in cases where a static image is required for publication or presentation. Static visualizations can be produced by selecting a single eigenvector and clicking on representative atoms. An arrow is drawn through the clicked atoms, with the arrow's head representing the atom's position at the most positive projection and the tail representing the most negative projection.

When IED is used to visualize normal modes data, there is no associated trajectory, and no projections file is loaded. In this case, trajectory playing and filtering features are disabled, but all other features are available.

7.4 Summary

IED allows interactive visualization and manipulation of projections of protein motion on selected eigenvectors and easy selection and filtering on different discontinuous sets of eigenvectors. It increases efficiency in working with ED results and enables appreciation of aspects of the dynamics that might be missed with more limited tools.

IED is available at no charge under the Gnu Public License (GPL) at `http://mccammon.ucsd.edu/software.html`. The language, applications and libraries on which it depends are also freely available.

This chapter is a reprint in full of material that appeared in *Interactive Essential Dynamics.* John Mongan. Journal of Computer-Aided Molecular Design, **18**(6), 433-36, June 2004. I was the sole author and researcher for this work.

Chapter 8

Computational design of pyrone-based inhibitors of stromelysin-1

ABSTRACT

In an effort to develop alternatives to hydroxamate-based matrix metalloproteinase inhibitors (MPIs), we have utilized the drug discovery program LUDI enhanced with the structural coordinates of a bioinorganic model complex. This method has yielded the first pyrone-based MPIs. The inhibitors demonstrate nanomolar potency against MMP-3 and are selective for MMP-3 over MMP-2 and MMP-1. The potency and unusual selectivity profile of these MPIs are postulated to be attributable to the pyrone chelating group.

8.1 Introduction

The zinc(II)-dependent MMPs have been pursued as chemotherapeutic targets for the treatment of illnesses such as cancer, arthritis, and heart disease. Consequently, over the past two decades attempts to interfere with MMP activity have yielded numerous

inhibitors.[165] MMP inhibitors (MPIs) are based on a two-part strategy: chelation of the catalytic zinc(II) ion combined with non-covalent interactions within subsite pockets in the MMP active site.[165,166]

The majority of MPIs synthesized to date contain a hydroxamic acid as the chelating or zinc-binding group (ZBG).[165,166] Hydroxamate-based MPIs suffer from a number of drawbacks including low oral availability, and poor in vivo stability, and consequently have not succeeded in clinical trials.[167] These limitations have prompted the investigation of a small number of non-hydroxamate-based MPIs.[168–170] This chapter describes inhibitors that utilize a pyrone ZBG, which results in improved potency and novel selectivity relative to similar hydroxamate-based MPIs.[168]

Pyrones were selected for this study due to their synthetic versatility,[171] known biocompatibility,[172] and good aqueous solubility. An earlier study examining the use of maltol (3-hydroxy-2-methyl-4-pyrone) as a ZBG, indicated the 2-methyl substitutent was favorably oriented toward the hydrophobic S1′ pocket of stromelysin-1 (MMP-3).[173,174] Several studies show that targeting the S1′ pocket of MMPs yields potent and selective MPIs.[165,166] Therefore, simple aryl groups were attached to the 2-position of maltol in order to exploit this interaction.

8.2 Methods

To design pyrone-based inhibitors, the drug discovery program LUDI (Accelrys) augmented with parameters from a bioinorganic model complex was employed.[175] LUDI uses a constrained docking approach that identifies optimal fragments to link to the pyrone moiety at a specified point of attachment. Structural data of maltol bound to a tris(pyrazolyl)borate model complex[173] were integrated into a known MMP crystal structure to generate the initial receptor complex.[175] The point of attachment to the ZBG was defined as an N-H bond from an amide moiety on the 2-position of the maltol ring (the amide group was built in silico on the ZBG). Fragments were screened and ranked using a LUDI scoring function.[176]

PDB structures 1G4K (MMP-3) and 1QIB (MMP-2) were used for docking with

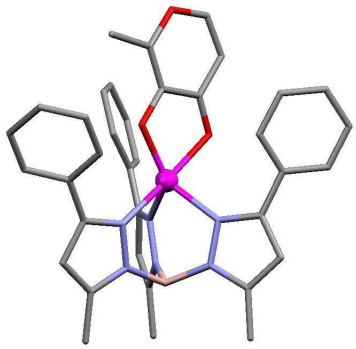

Figure 8.1: Crystal structure of maltol bound to ($Tp^{Ph,Me}$)Zn, a model of zinc(II) ion coordination in the MMP active site.

LUDI version 60a, as part of the InsightII 2000L framework. For structures having more than one protein in the asymmetric unit, the "A" chain was selected. Proteins were protonated using the "Hydrogens" command of the Biopolymer module of InsightII. Crystal waters and inhibitors were removed from each structure. The zinc binding group (ZBG) 3-hydroxy-2-methyl-4-pyrone (maltol) was positioned in the active site of each protein based on crystal structure coordinates of maltol bound to [($Tp^{Ph,Me}$)Zn] ($Tp^{Ph,Me}$ = hydrotris(3,5-phenylmethylpyrazolyl)borate), a model of the zinc(II) ion coordination in the MMP active site.[175] Positioning was performed by minimizing RMSD between the protein and model compound for the active site zinc and coordinating nitrogen atoms. Due to the rotational symmetry of the model compound, three alignments are possible. Only one alignment allows the R groups (see figure 8.3) access to the S1′ pocket;[177] this alignment was used in all docking studies. Hydrogen atoms were added to the aligned maltol molecule and an amide group was built at the 2-position of the ring using Cerius^2 4.8.1.

All docking was performed using LUDI link mode, where docked fragments are constrained such that a methyl group on the fragment must be aligned with a link site. The N–H amide bonds on the maltol ZBG were selected as the link sites. Initial docking used the default LUDI link library of fragments, and the following parameters: maximum alignment angle 20˚; maximum alignment RMSD 0.6 Å; search radius 11 Å; rotate bonds two at a time; preselect 4.0; minimum separation 3.0; lipophilic density 40; polar

density 40; minimum surface 0; link weight 1.0; lipophilic weight 1.0; H-bond weight 1.0; aliphatic aromatic off; reject bifurcated off; no unpaired polar off; electrostatic check off; minimum score 0; maximum fits 8000; maximum hits all; maximum unfilled cavity 0; energy estimate 1 scoring function;[176] and best fit. These parameters were chosen to maximize the quality and thoroughness of the docking. Despite this, it was found that results were somewhat dependent on the search sphere center, and favorable fragment poses could be missed with some search sphere centers, particularly for the larger fragments. To minimize this problem, multiple dockings were performed using different search sphere centers within the S1′ pocket; the results presented represent the union of these results.

Further docking was performed using a custom link library primarily based on the work of Hadjuk *et al.*,[178] consisting of the substituents illustrated in figure 8.3. Due to the limited ability of LUDI to handle rotational flexibility in fragments (only 120˚ or 180˚ rotations can be performed), all possible rotamers with 30˚ increments of rotation were generated for each substituent, and each rotamer was added to the library as a separate fragment. Bonds between phenyl groups were treated as non-rotatable, and rotamers with steric clashes or eclipsed conformations were excluded. Docking with this custom library was performed using the same parameters as above, except bond rotation was set to one at a time. It would seem that because the library already included all rotamers, bond rotation could be set to none, but using one at a time seemed to reduce the differences in results caused by using different search sphere centers.

8.3 Results and Discussion

Results from docking the custom library are presented in table 8.1. Reported LUDI scores represent the highest scoring pose for a fragment in the specified protein structure, rounded to the nearest 10. Higher scores indicate higher predicted affinity, with each 100 points representing a predicted order of magnitude decrease in IC_{50}.

The result of the LUDI docking for one of the high scoring compounds (**AM-5**, vide infra) is shown in figure 8.2. The fragment in figure 8.2 was found to reside in the S1′

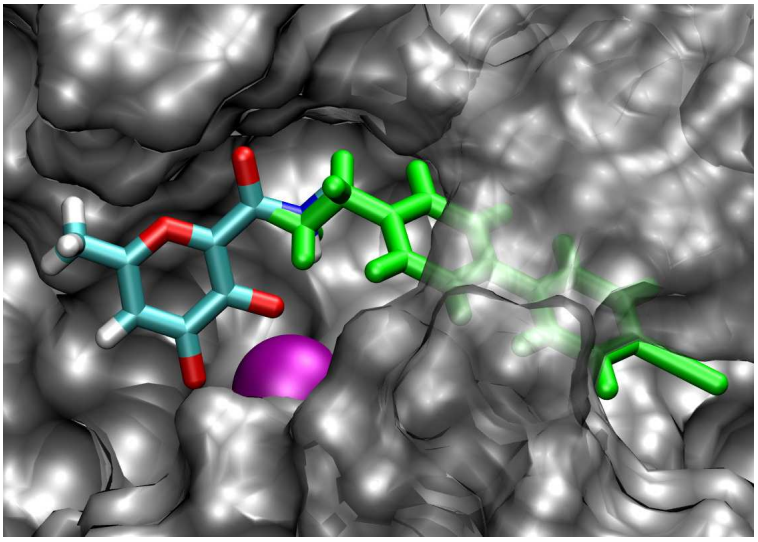

Figure 8.2: LUDI docking image of backbone fragment (green, in S1′ subsite) with pyrone ZBG (colored by element) in the active site of MMP-3 (gray). This fragment combination leads to the compound designated **AM-5** (see figure 8.3). The zinc(II) ion is shown as a magenta sphere.

pocket of MMP-3. The high and low scoring fragments from the custom library were similar in structure; therefore, all six compounds were synthesized to test the accuracy of the LUDI docking and scoring function.

Synthesis of the pyrone-based MPIs was performed as illustrated in figure 8.3. Two synthetic routes were utilized, based on the commercial availability of the desired amine backbones. 2-Carboxy-3-benzyloxy-6-methyl-pyran-4(1H)-one (**1**) was prepared by a literature method.[171] Compound **1** was then activated with NHS, followed by coupling to the desired amine, and removal of the benzyl protecting group to yield compounds **AM-1**, **AM-2**, **AM-3**, and **AM-4**. The synthesis of **AM-5** and **AM-6** was accomplished similarly, but required the Suzuki coupling of 3-benzyloxy-6-methyl-pyran-4(1H)-one-2-carboxy-N-(4-iodobenzylamide) (**2**) with 4-cyanophenylboronic acid and 4-biphenylboronic acid, respectively, as an intermediate step.

The inhibitory activity of compounds **AM-1** through **AM-6** was evaluated using a fluorescence-based assay;[179] the IC_{50} values are listed in table 8.1. **AM-2**, **AM-5**, and **AM-6** were the most potent compounds against MMP-3, with IC_{50} values in the nanomolar range. The IC_{50} values against MMP-3 correlate well with the scores ob-

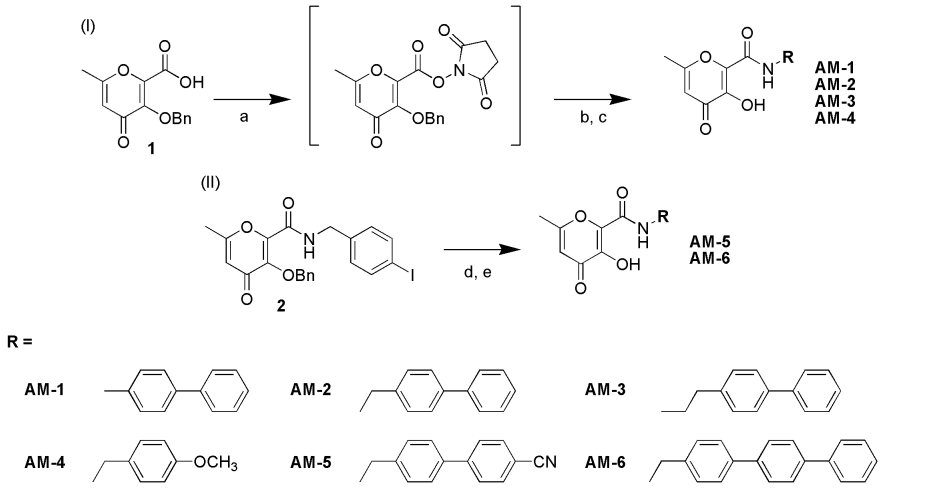

Figure 8.3: Synthetic scheme for MPIs. Key: (I) a) NHS, DCC, dry THF; b) 'amine', dry THF, 88%; c) 10% Pd/C, H_2 35psi, MeOH or 1:1 HCl:CH_3COOH, 60–89%. (II) d) $ArB(OH)_2$, 2M K_2CO_3, $Pd(C_2H_3O_2)_2$, PPh_3, toluene, 135 ^oC, 40–85%; e) 10% Pd/C, H_2 35psi, MeOH or 1:1 HCl:CH_3COOH, 60–91%.

tained for each fragment using the program LUDI. Although the LUDI scores do not perfectly parallel the relative inhibitory activity, the approach used here does clearly distinguish between poor, moderate, and exceptional MPIs.

Interestingly, the pyrone-based MPIs presented here are more potent than the analogous hydroxamate-based inhibitors,[180] which is contrary to the accepted dogma that hydroxamic acids are the best ZBGs.[181] As expected, the effects of linker length (compare **AM-1**, **AM-2**, and **AM-3**) and backbone substituents (**AM-5** relative to **AM-2**) are consistent with analogous hydroxamate-based MPIs.[178] These results strongly support the concept that ZBGs equal or superior to hydroxamates can be identified and utilized in novel MPI designs.[170,180]

The observed trends in the IC_{50} values of the MPIs described here against MMP-3 suggest that the large aromatic backbone substituents of these compounds occupy the S1′ subsite. This hypothesis was further examined by determining the selectivity of these compounds against different MMPs. Traditionally, the incorporation of bulky groups directed toward the S1′ pocket results in selectivity over MMP-1, which has a shallow S1′ pocket.[165] All six MPIs were found to be poor inhibitors of MMP-1 (table 8.1). The poor activity of these compounds against MMP-1 is wholly consistent with

Table 8.1: IC_{50} Values (μM) and LUDI scores for MPIs against MMP-1, MMP-2 and MMP-3. Higher LUDI scores indicate better predicted affinity (lower IC_{50}).

Inhibitor	MMP-1	MMP-2		MMP-3	
		IC_{50}	LUDI	IC_{50}	LUDI
AM-1	>50	36(5)	—	>50	—
AM-2	>50	9.3(0.5)	530	0.24(0.01)	600
AM-3	>50	27(2)	—	36(1)	—
AM-4	>50	>50	440	2.4(0.2)	440
AM-5	>50	0.61(0.01)	570	0.010(0.002)	640
AM-6	>50	>50	690	0.019(0.002)	700

the aryl backbone groups occupying the S1′ pocket, which supports the LUDI results (figure 8.2) and ZBG orientation predicted by the bioinorganic modeling studies.[173]

The inhibitors were also tested for potency against MMP-2. Like MMP-3, MMP-2 has a deep S1′ pocket and potency against these two enzymes is expected to be comparable, as found with hydroxamate-based MPIs.[165,166] Interestingly, although **AM-2**, **AM-4**, **AM-5**, and **AM-6** showed a range of potencies against MMP-3, all four compounds were substantially less potent against MMP-2. Indeed, **AM-5** showed >2500-fold selectivity for MMP-3, which is the highest selectivity reported for an MPI for MMP-3 over MMP-2.

The observed selectivity of these compounds for MMP-3 over MMP-2 is in contrast to the selectivity observed for most deep S1′ pocket MPIs. Hydroxamate-based MPIs that occupy the S1′ pocket are almost exclusively more potent for MMP-2 than MMP-3, with few exceptions.[165,166,182] MPIs reported to be selective for MMP-3 over MMP-2 generally target the S3′ subsite;[182] however, based on the LUDI docking, the MPIs presented here have no significant interactions in the S3′ subsite, and indeed give similar LUDI scores when docked to MMP-2 or MMP-3 (Table 8.1). Therefore, it is plausible that the observed selectivity originates from the pyrone ZBG. It has been reported that more acidic ZBGs, such as carboxylates (a weaker ZBG than the hydroxamate),[181] are generally more potent for MMP-3 than MMP-2,[168,182] which is attributed to the difference in the optimal pH for the two enzymes. MMP-3 prefers a more acidic environment (pH ~6.0) compared with other MMPs (including MMP-2), which favor a higher pH

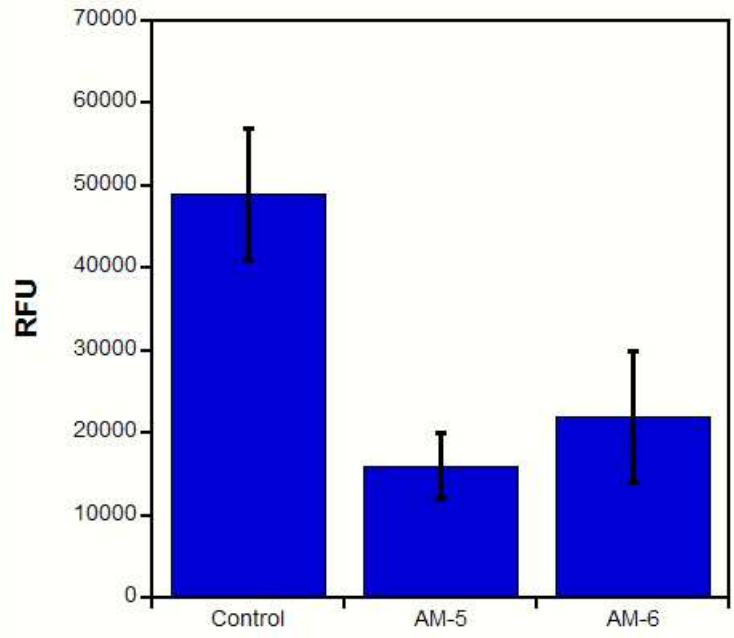

Figure 8.4: Neonatal cardiac fibroblast (CF) invasion assay results. Fluorescent measurement (in RFUs) of lysed cells after invasion with: no inhibitor (Control), 250 nM **AM-5**, and 250 nM **AM-6**. Increased RFUs indicates increased cell invasion.

(~7.5).[183] By analogy, the selectivity of the MPIs reported here is likely due to the greater acidity of the pyrone versus hydroxamate chelator (ΔpK_a ~1).[184] These results suggest that the ZBG, and not only the MPI backbone, can provide selectivity between different MMPs without compromising potency. Furthermore, the selectivity of the pyrone MPIs could be advantageous in the environment of hypoxic tumors, where a more acidic ZBG may prove more effective.

The ability of **AM-5** and **AM-6** to inhibit invasion of neonatal rat cardiac fibroblasts through a collagen membrane was examined, as a gauge of the in vivo potential of these MPIs. At a concentration of 250 nm, the two inhibitors were found to reduce invasion by 67% (**AM-5**) and 55% (**AM-6**) (figure 8.4). In summary, the use of pyrone ZBGs results in more potent inhibitors than those produced with the widely employed hydroxamate group. These results also indicate that the use of a non-hydroxamate ZBG reveals a novel route to MMP inhibitor selectivity. Overall, the findings reported here suggest a chelator-driven approach to metalloprotein drug design can produce potent and selective metalloprotein inhibitors.

This chapter is a reprint in full of material that appeared in *Potent, Selective Pyrone-Based Inhibitors of Stromelysin-1*. David T. Puerta, John Mongan, Ba L. Tran, J. Andrew McCammon, and Seth M. Cohen. Journal of the American Chemical Society, **127**(41), 14148-49, October 2005. I was the secondary author and conducted the computational portion of the research for this work. Co-authors of the article conducted the synthesis and assays or supervised and directed the work.

Chapter 9

Evaluation and binding mode prediction of thiopyrone-based inhibitors of anthrax lethal factor

ABSTRACT

Anthrax lethal factor (LF) is one of three proteins involved in anthrax pathogenesis and lethality. Inactivation of the LF gene in *B. anthracis* leads to a thousand-fold or greater reduction in virulence, which suggests that anthrax pathology is highly dependent on LF.[185] This chapter presents an effective inhibitor of anthrax lethal factor based on a heterocyclic chelator scaffold, computational predictions of the binding mode for this inhibitor and evidence that accurate prediction of binding modes requires use of a molecular surface-like boundary between solute and solvent.

9.1 Introduction

Anthrax LF is a zinc(II)-dependent, hydrolytic enzyme that cleaves the N-terminus of the D-domain of mitogen-activated protein kinase kinases (MAPKK), which impairs essential signal transduction pathways and results in macrophage apoptosis along with other harmful consequences for the host.[186–188] The potential for bioterrorism

and inadequate treatments for anthrax, especially at late stages of infection, have amplified interest in finding effective anthrax lethal factor inhibitors (LFi). Several approaches have led to the identification of a variety of LFi[189–192] including library screening and optimization,[193,194] fragment-based NMR screening (**BI-11B3**, figure 9.1),[195] mass spectrometry-based screening,[196] and re-examination of inhibitors of other metalloproteases and related hydroxamate-based compounds.[193,197–199] An example of the latter class is the broad spectrum matrix metalloproteinase (MMP) inhibitor **GM6001** (figure 9.1), which was found to be an effective inhibitor of LF in vitro and in cell culture.[193] Structural characterization of **GM6001** in the LF active site shows that the hydroxamate group of the inhibitor chelates the catalytic zinc(II) ion.[193] Indeed, the direct binding of the active site zinc(II) ion is proposed to be important in the majority of LFi described to date.[191,193–195] Based on the use of hydroxamate-based inhibitors, a bioinorganic approach to the design of LFi is applied here. The strategy focuses on the metal-ligand interactions of a metalloprotein inhibitor,[170] which appear to be central to the inhibition of LF. The effectiveness of several previously described ligands as zinc-binding groups (ZBGs)[181] for incorporation into LFi are reported. The ZBGs, shown in figure 9.1, were selected based on their inhibition of MMPs, as well as their potential to overcome the limitations associated with hydroxamate-based inhibitors.[170] Finally, the potency and binding mode of a novel LFi based on a thiopyrone chelator (**AM-2S**, figure 9.1) is reported.

9.2 Experimental assays

The *in vitro* potency of compounds **1**–**11** (figure 9.1) against LF was evaluated in an assay based on established procedures using a fluorescent peptide substrate (table 9.1).[200] Compounds **1**–**11** were compared relative to the hydroxamate group used in many LFi and metalloproteinase inhibitors, as represented by acetohydroxamic acid (**AHA**, figure 9.1). It is important to recognize that compounds **1**–**11** represent only the ZBG portion of a metalloproteinase inhibitor (figure 9.1). These ZBGs will be used as a platform to which a backbone substituent can be added to provide additional po-

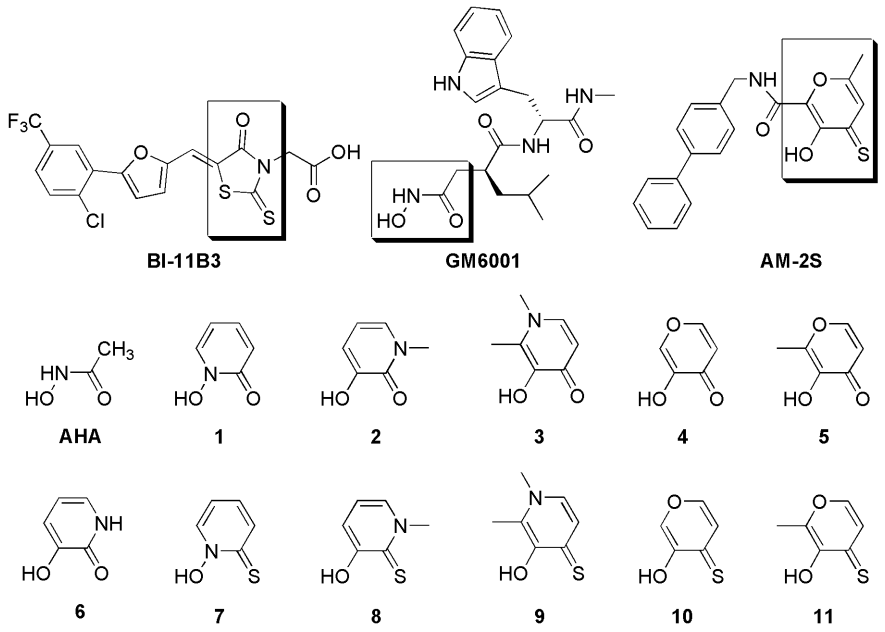

Figure 9.1: New (**AM-2S**) and previously described (**BI-11B3**, **GM6001**) inhibitors of LF (top). **AM-2S** is based on one of the heterocyclic zinc-binding groups (ZBG, **1** – **11**) examined in this study (bottom). The ZBG of each full-length inhibitor is highlighted in a raised box.

tency and selectivity, resulting in a complete LFi. The O,O donor ligands **1**, **2**, **4**, and **6**, on average, showed comparable inhibition to **AHA** (compound **3** was not soluble and compound **5** showed no inhibition up to its solubility limit of ~6 mM). This suggests that inhibitors based on these heterocycles will have comparable potency to hydroxamate-based LFi, but may avoid some of the clinical shortcomings of hydroxamate-based metalloproteinase inhibitors.[170] Furthermore, O,S donor ligands (**7**–**11**) showed improved LF inhibition over **AHA**. This is consistent with earlier findings, which show that sulfur-containing ligands inhibit zinc-dependent metalloproteinases more effectively than their O,O analogues.[181]

To obtain potent, selective metalloproteinase inhibitors, the ZBG must be appended to a backbone moiety to target the protein of interest. **GM6001** (figure 9.1) has a hydroxamate ZBG that is attached to a complex backbone with a hydrophobic leucine mimetic at the P1′ position;[193] the result is that **GM6001** is a very potent, broad spectrum MMPi as well as a potent LFi. This motif, of a ZBG appended with a hydrophobic P1’ sub-

stituent, has been suggested as a general strategy for obtaining potent LFi.[193] Based on this construct, a biphenyl backbone was attached to a thiopyrone ZBG (**10**, **11**) to obtain the inhibitor **AM-2S**. Evaluation of **AM-2S** in the LF assay gave an IC_{50} value of ~14 μM against LF (table 9.1). This value is comparable to several reported LFi,[201] including **GM6001**, which has been found to have an IC_{50} value of 2–20 μM.[193,195]

Table 9.1: IC_{50} values for ZBGs (**1**, **2**, **4**, **6**–**11**) and **AM-2S** against LF measured using a fluorescence-based assay. IC_{50} values are based on at least three independent experiments.

ZBG	IC_{50} against LF (μM)	Potency vs. AHA
AHA	11400 $\pm$ 1000	n/a
1	6570 $\pm$ 160	1.7-fold
2	32000 $\pm$ 3000	0.36-fold
4	27000 $\pm$ 3000	0.42-fold
6	6100 $\pm$ 500	1.9-fold
7	3900 $\pm$ 200	2.9-fold
8	690 $\pm$ 70	16-fold
9	1460 $\pm$ 60	7.8-fold
10	204 $\pm$ 16	56-fold
11	260 $\pm$ 30	43-fold
AM-2S	13.9 $\pm$ 0.3	820-fold

9.3 Computational methods

Binding modes of **AM-2S** within the LF active site were investigated computationally using AMBER.[58] Docking was accomplished by using an extension of the previously described restrained exhaustive minimization approach.[175] The work presented here extends the previously described method in three important ways. First, a brief initial minimization was conducted to attempt to resolve steric conflicts, rather than eliminating all starting structures with conflicts. Second, solvent effects were represented during the minimization using the generalized Born model. Third, due to the increase in computational cost involved with generalized Born, structures were minimized to an intermediate level of convergence, clustered, and then a single representative from each cluster was minimized to full convergence, to avoid duplicating computational efforts.

LF coordinates were taken from Protein Data Bank structure 1PWQ. Chain A of the asymmetric unit was selected. Residues numbered less than 302 were removed to reduce the atom count. All deleted atoms are on the opposite face of the protein from the active site and are more than 30 Å away from the active site zinc(II) ion, so this simplification is expected to have minimal effects on the results.

The zinc-binding group (ZBG) 3-hydroxy-2-methyl-4-thiopyrone (thiomaltol) was positioned in the active site of LF based on crystal structure coordinates of thiomaltol bound to the active site model [($Tp^{Ph,Me}$)ZnOH] ($Tp^{Ph,Me}$ = hydrotris(3,5-phenylmethylpyrazolyl)borate).[181] Positioning was performed by minimizing the RMSD between the protein and model compound for the active site zinc ion and coordinating nitrogen atoms. Due to the rotational symmetry of the model compound, three alignments with the protein are possible. Two of these alignments lead to reasonable inhibitor poses, while the third produces unresolvable steric clashes with the protein. A methyl group, the peptide biphenyl unit, and hydrogen atoms not present in the model compound crystal structure were added to the **AM-2S** model based on standard equilibrium bond lengths and angles using Cerius^2 (Accelrys).

For each of the two plausible alignments of the ZBG in the LF active site, a set of all possible rotamers of the backbone portion of the inhibitor was generated. The two bonds adjacent to the methylene carbon and the bond connecting the two phenyl groups were rotated in 15˚ increments, while the bond connecting the ZBG with the carbonyl carbon was rotated in 180˚ increments (to maintain ring conjugation). Rotamers having steric clashes within the inhibitor were eliminated, leaving approximately 12,000 unique inhibitor positions.

Energy minimizations were conducted using the sander module of the AMBER 8 suite. The LF protein was represented using the ff99 force field, modified with the phi/psi potential of Simmerling, Strockbine, and Roitberg.[44] The inhibitor was modeled using the GAFF force field,[202] with parameter assignments conducted by the antechamber module of AMBER. Inhibitor partial charges were determined using the AM1-BCC method, based on a net inhibitor charge of -1.

For each inhibitor starting position, an initial gas phase minimization (10 steps steepest descent followed by 90 steps conjugate gradient) was performed to attempt to alle-

viate steric clashes with the protein. During this minimization, positions of the protein, active site zinc(II) ion, and ZBG portion of the inhibitor were fixed by applying harmonic restraints with a force constant of 100 kcal/mol/Å^2. Structures that successfully completed this minimization yielding a favorable (negative) van der Waals energy were passed on to the second stage of minimization.

In the second minimization stage, harmonic restraints were relaxed to 10 kcal/mol/Å^2, and solvent effects were introduced using the *OBC* generalized Born model (igb=5)[47] with Bondi radii[16] to define the dielectric boundary. Minimization began with 10 steps of steepest descent and continued with conjugate gradient steps until the root mean square of the Cartesian elements of the gradient were less than 5×10^{-3} kcal/mol/Å. This fairly lenient convergence criterion was employed because full minimization of every structure was computationally prohibitive.

After the second minimization, minimized inhibitor structures were clustered based on their Cartesian coordinates with the *k*-means clustering algorithm using Euclidian distances and arithmetic means to define cluster centroids.[130] Because the purpose of clustering was to eliminate redundant computation rather than to identify an optimal clustering, cluster number was adjusted such that the representative (lowest energy) member of each cluster had an RMSD of no more than 1 Å from the representative member of the nearest cluster. 25 clusters were sufficient to achieve this maximum separation.

For each cluster, the lowest energy member was selected, and minimization was continued separately under the previously used *OBC* GB model as well as the *GBn* model described in chapter 3. For this final stage of minimization, the convergence criterion was 1×10^{-4} kcal/mol/Å.

The two different GB models that were employed, *OBC* GB[47] and *GBn* (see chapter 3), differ in how they define the dielectric boundary between the solute and solvent. Their boundaries are related to two of the most commonly used surface definitions, the van der Waals surface (vdWS) and the molecular surface (MS). With a vdWS, any point not inside a solute atom is defined as solvent; with an MS only points outside the surface defined by rolling a solvent sphere over the solute are defined as solvent. The MS has the attractive property of excluding any space smaller than a water molecule

from the solvent region, while the vdWS defines these small crevices between solute atoms as containing water. vdWS (or related smoothed surfaces) are commonly used in implicit solvent models because they are computationally more tractable than MS, but as discussed in chapter 2, the non-physical solvent pockets allowed by vdWS can cause errors in protein solvation free energies and hydrogen bonding potentials. Both the *OBC* GB and *GBn* models are based on a vdWS with correction terms that attempt to emulate the properties of a MS. *OBC* GB employs a geometry independent correction that makes corrections on an averaged basis and has little effect on surface atoms, while *GBn* uses a geometrically-based pairwise correction that corrects for solvent pockets near the surface as well as within the core of the protein.

The method described here is among the most physically rigorous methods for restrained docking, but is very computationally intensive. Total computer time was approximately 100 processor-weeks using 3.2 GHz Intel Xeon processors. The vast majority of this time was spent in the second minimization stage.

Poisson-Boltzmann calculations were conducted with a pre-release version of APBS 0.4.0.[11] The linearized PB model was employed along with the multiple Debye-Huckel boundary condition. Charge was discretized using the cubic B-spline method (spl2). Dielectric values were 1.0 for solute and 80.0 for solvent regions. Calculations were perfomed initially on a coarse grid with a resolution of 0.8 Å and then on a smaller grid with resolution of 0.1 Å using the coarse grid potential as boundary conditions. Molecular surface calculations used a minimum sampling point density (sdens) of 50 points per Å.[186]

9.4 Computational results and discussion

Both GB models identified the structure shown in figure 9.2 as having the lowest free energy. This structure is 3.4 kcal/mol lower in energy than the next best conformer when calculated by the *OBC* GB model and 2.5 kcal/mol lower when calculated by *GBn*. The biphenyl group is found in the substrate binding groove of LF, near the binding location of that observed for the LF20 peptide.[193] In contrast, for the second alignment of the

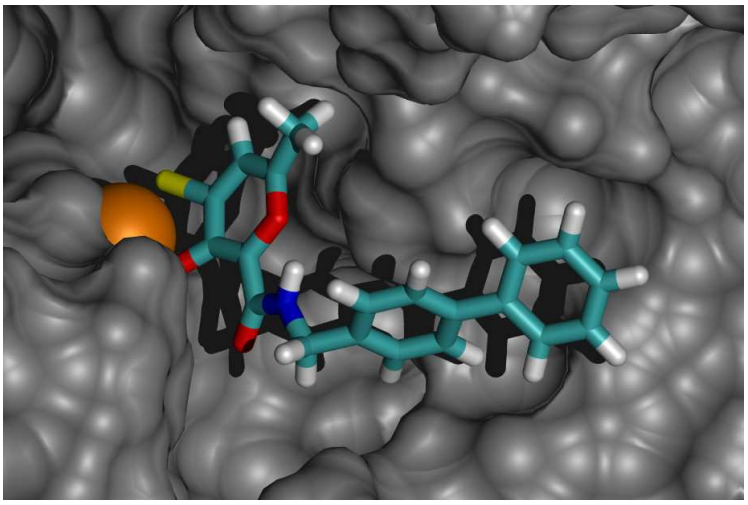

Figure 9.2: Lowest energy configuration of **AM-2S** in LF active site, identified by both GB models. Orange sphere is zinc(II) ion.

ZBG, the inhibitor conformation with the lowest free energy was dependent on which dielectric boundary was used. The lowest energy conformers calculated by *GBn* and *OBC* GB are shown in figure 9.3. The lowest energy conformer calculated using *OBC* GB places the biphenyl group into a narrow groove, which is highlighted in figure 9.4.

Considering only the non-solvation components of the energy, configuration I identified by *OBC* GB (figure 9.3, purple) is lower energy than configuration II calculated using *GBn* (figure 9.3, colored by atom type), due mostly to more favorable van der Waals interactions between the biphenyl group and the protein. However, configuration I highlighted in figure 9.4 entails a significant energetic penalty for desolvating the polar groups in the groove by displacing water with the non-polar biphenyl group. When using the more vdWS-like *OBC* GB model, this penalty is underestimated due to the crevices between the biphenyl group and protein (figure 9.4), and configuration I is calculated to have a solvation energy only 2.8 kcal/mol higher than that of configuration II, such that configuration I is incorrectly identified as having the lowest total energy. When using the more MS-like *GBn* model, solvent is excluded from these crevices leading to a larger difference in solvation energies of 5.8 kcal/mol, which is sufficient to identify configuration II as having the lowest energy.

Because GB models involve approximations of both the dielectric boundary and

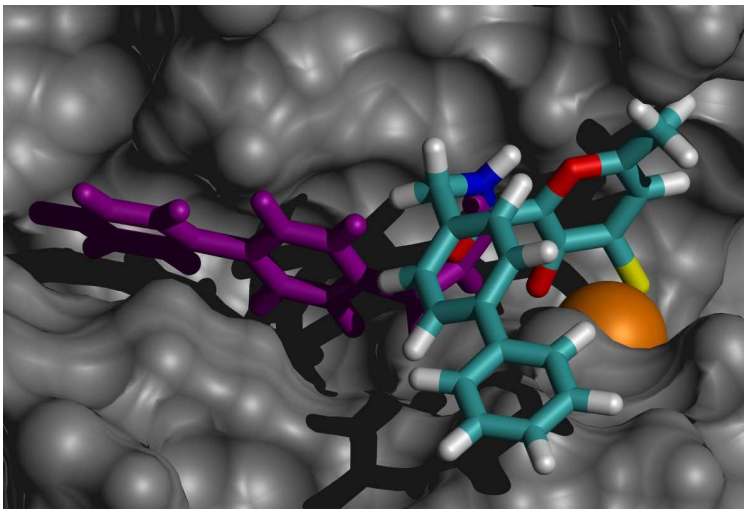

Figure 9.3: Lowest energy configurations of **AM-2S** in LF active site for alternate ZBG orientation. Configuration I (purple) appears to be lower in energy only when using vdWS-like solvent-solute boundary such as that employed by *OBC* GB; configuration II (colored by atom type) is the correct configuration, identified by *GBn.*

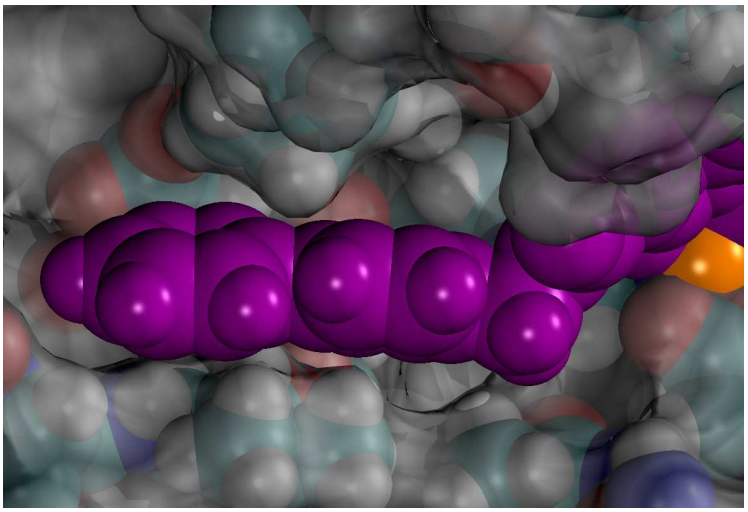

Figure 9.4: Detail view of positioning of biphenyl group (purple) for configuration I from figure 9.3. Atom type colored spheres are LF atoms, with transparent grey LF MS superimposed to aid visualization. Note the small crevices between the biphenyl group and the LF binding groove that are defined as solvent with a vdWS-like boundary but not with a MS, as well as the partially obscured polar oxygens (red spheres) near the bottom of the groove.

electrostatic interactions, the results were confirmed using the less efficient but more physically rigorous Poisson-Boltzmann (PB) method implemented in APBS.[11] PB calculations with an uncorrected vdWS show little difference in solvation energy, with configuration I more favorable than II by 0.4 kcal/mol. When the PB calculation employed a true MS, the solvation energy of configuration II is 36.6 kcal/mol lower than that of I, confirming the selection of configuration II by *GBn* as the correct result.

The configuration illustrated in figure 9.2 is approximately 28 kcal/mol lower in energy than the best alternative with the alternate ZBG orientation illustrated in figure 9.3. Nevertheless, figure 9.3 may represent a relevant configuration as there are unfavorable steric interactions between the ZBG and protein for both ZBG orientations, but they are considerably worse in figure 9.3. It was necessary to keep the protein fairly rigid so that the minimizations would be computationally tractable, so these unfavorable interactions could not be relaxed in the computational modeling. Although it seems likely that the protein would be sufficiently flexible to reduce or eliminate these interactions, there is no straight-forward way to calculate the resulting conformations or energies. Therefore, while it is reasonable to compare the relative energies of different configurations with the same ZBG orientation, no fair comparison can be made between poses having different ZBG orientations on the basis of the results presented here. Until further studies have been conducted that incorporate greater protein flexibility or alternate protein conformations that do not clash with the ZBG, the poses illustrated in Figures 9.2 and 9.3 should both be considered valid possibilities.

The results presented here illustrate the importance of a physically realistic solvent-solute boundary in docking and binding studies. The work with small molecule hydrogen bonding and salt bridge models presented in chapter 2 showed that errors due to solvent pockets were mostly for high energy configurations (*i.e.* transition states), but these results show that for small molecule-protein interactions, even minimum energy conformations can be sufficiently affected to produce erroneous ranking of binding modes. It may be noted that a non-polar small molecule and a polar, concave protein surface, as found in the system examined here, are likely to produce the greatest disparity between MS and vdWS results. However, these properties are sufficiently common in systems of interest that the problems caused by non-physically small solvent pockets

cannot be safely ignored.

This chapter is a preprint in full of *Evaluation and binding mode prediction of thiopyrone-based inhibitors of anthrax lethal factor*. Jana A. Lewis, John Mongan, J. Andrew McCammon and Seth M. Cohen. Submitted to Angewandte Chemie International. I was co-primary author and conducted the computational portion of the research for this work. Co-authors of the article conducted the assays or supervised and directed the work.

Chapter 10

Future Directions

All the results and discoveries in the preceding chapters have built on the work of others, as evidenced by the over 200 bibliography entries that follow. It is my hope that this work will likewise be a foundation for myself and others to build on. This chapter explores some of the directions that further advances in these areas might take.

10.1 Implicit solvation and generalized Born models

The importance of a Lee-Richards-like[1] molecular surface boundary between solvent and solute was illustrated in chapter 2, and the generalized Born (GB) model developed in chapter 3 provides a proof of concept that key aspects of such a boundary can be incorporated into a fast, analytical GB model. The results in chapter 3 show that this new model is an improvement over earlier GB models on a wide variety of quality measures, and chapter 9 illustrates an example where the degree of improvement is sufficient to effect a qualitative change in the calculated result. Nevertheless, the improvements of the *GBn* model presented in chapter 3 are somewhat more modest than might be hoped, given the substantial differences between molecular and atom-centered or van der Waals-like surfaces illustrated in chapter 2.

Like most GB models, the *GBn* model is based on the Coulomb field approximation (CFA). The CFA provides a conceptually simple basis for GB that achieves fairly good results for small molecules, but proves to be a very poor approximation for macro-

molecules. To achieve reasonable results for large molecules, the free parameters of a CFA GB model must be carefully fit to experimental or Poisson-Boltzmann (PB) results to reduce the effect of error introduced by the CFA. This fitting process substantially reduces the generality of the resulting GB model, and errors can only be partially corrected, not eliminated. A far better approach is to start with a more accurate approximation of electrostatic field density. Charlie Brooks, Michael Lee, Michael Feig, Freddie Salsbury and Wonpil Im, have pioneered this, with a series of refinements of an empirically derived correction term to the CFA.[5,8,36] Tomasz Grycuk has derived a different non-CFA expression[53] with physical basis in the Kirkwood electrostatic model.[81] Extension of the molecular surface-like integration method developed here to these more accurate approximations should provide a marked improvement in GB accuracy with very little loss of efficiency.

It has recently been suggested that as PB methods improve in computational efficiency and ability to calculate numerically stable and accurate forces, they may soon become a preferred alternative to GB.[25] With further advances in PB methods, this may yet occur. However, at the present time this suggestion fails to recognize that most of these advances in PB have been enabled by the adoption of the atom-centered dielectric boundaries discussed in chapter 2. As illustrated by the results in that chapter, the differences between results generated with molecular and atom-centered surfaces may be significantly greater than the differences between PB and a high-quality GB model. Until this problem is solved, these new PB methods may be both less efficient and less accurate than the best GB methods. Early efforts to address this issue by Lu and Luo illustrate the difficulty of the problem, making it clear that solution cannot be achieved by a simple force field reparameterization.[4] Rather than PB supplanting GB, it seems more likely that both families of methods will evolve, developing into multiple implementations that span the spectrum of performance-accuracy tradeoffs.

10.2 Constant pH molecular dynamics

Perhaps the biggest challenge facing constant pH molecular dynamics methods is convergence. Convergence is a difficult issue for molecular dynamics (MD) methods in general, and the addition of protonation state degrees of freedom makes the problem worse. No current method shows fully converged protonation state populations for proteins, and some methods apparently do not even produce converged results for small molecule model compounds in reasonable amounts of computer time. Since the most common and straightforward validation method for constant pH schemes is comparison of protonation state populations (or the related predicted pK_a values) to experimental data, it is difficult to assess the relative merits of different approaches: legitimate systematic differences can be lost in a sea of random sampling error due to poor convergence.

Correct averages are achieved from simulation only when the number of transitions between energy wells is sufficient that the time average of conformations over the simulation approximates the ensemble average. Therefore calculating correct ensemble average protonation populations will require reaching convergence multiple times for the electrostatic environment of each of a large number of conformations.

Improving the rate at which protonation populations converge in a single electrostatic environment may yield some gains, but at some point convergence is limited by the rate of conformational sampling. Use of implicit solvent MD accelerates sampling while reducing computational cost, and forms the basis of many of the most successful constant pH methods, including the model introduced in chapter 5. It may also be fruitful to combine accelerated sampling methods (*e.g.* replica exchange,[203] locally enhanced sampling,[204] accelerated MD[157]) with constant pH to increase rates of conformational sampling.

Implicit solvent MD is generally more computationally efficient than explicit solvent MD, and provides faster conformational sampling.[27] However, these methods are relatively recent, and have not been parameterized as carefully and extensively as their explicit solvent counterparts. Protonation of a titratable group is very sensitive to the strength of salt bridges and hydrogen bonds involving the titratable group. The strength

of these non-bonded interactions is determined largely by the solvent model. Both of GB-based constant pH methods reviewed in chapter 4, model developed by Lee *et al.*[102] and the model presented in chapter 5, encountered significant difficulties involving hydrogen bonding. Substantial improvements in protonation state populations may be achieved with implicit solvent models that employ the advances described in the preceding section to model the strengths of non-bonded interactions more accurately.

The most interesting aspect of constant pH MD will not be the development or improvement of methods, but the ability to address unanswered and previously inaccessible biological questions that is provided by sufficiently advanced methods. Examples of the many important systems where pH plays a key role in triggering conformational rearrangements are the GALA peptide, engineered as a model of a viral fusion peptide; the c subunit of the F_0 part of F_0F_1 ATP synthase, responsible for driving ATP synthesis; hemoglobin and hemagglutinin, the influenza protein that mediates fusion of the viral envelope with the cell membrane.

10.3 Accelerated molecular dynamics

The biased-potential sampling method developed in chapter 6 provides a simple means for accelerating conformational sampling on potential energy landscapes where the locations of important minima may not be known. The method as described in chapter 6 provides for accelerated convergence to Boltzmann population distributions (after reweighting), but does not allow for the recovery of kinetic information from the accelerated trajectory. A recent extension of the method by Hamelberg, Shen and McCammon shows that kinetics data can, in fact, be extracted from accelerated MD simulations with appropriate analysis and characterization of the roughness of the energy landscape.[205]

Both chapter 6 and the recent Hamelberg *et al.* paper involve application of accelerated MD to small peptide systems. There are a number of open questions regarding scaling the method to larger proteins. It remains to be determined what values of the tunable energy threshold and well depth parameters, E and α, are optimal for achieving the most efficient sampling. Additional studies are also needed to examine the tradeoffs

between modifying the potential energy for the entire system with a single threshold energy and modifying components of the potential separately (*e.g.* a separate threshold for each torsion).

The importance and effectiveness of accelerated MD will be best judged in application to problems where sampling with conventional MD is too slow to reach acceptable levels of convergence, such as in the constant pH methods described above and in chapters 4 and 5.

10.4 Zinc(II) protease inhibitor design

The drug design process described in chapter 8 was successful in producing inhibitors with IC_{50} values in the nanomolar range, as generally required for lead compounds. A number of avenues remain open for improvement of both the inhibitors and the methods used to design them. While the affinities are quite good, the specificities between different matrix metalloproteinase (MMP) subtypes is fairly low for most of the inhibitors, with the notable exception of **AM-6**. This is not unexpected, as all of the inhibitors designed as part of this work interact primarily with the zinc and the hydrophobic S1' pocket, which have highly conserved structure across the MMP subtypes. The MMPs comprise a large family of enzymes with distinct expression patterns, so effective therapeutic applications will likely require specific inhibitors. Improved specificity may be obtained by designing inhibitors that interact with binding sites opposite the S1' pocket, which show greater variability between MMP subtypes.

LUDI, the program used for design of the MMP inhibitors described in chapter 8, yielded predictions having good correlation with experimental assays, but it may not be the best choice for future work. The program often yields inconsistent results, and as a close-source commercial application cannot be improved, extended or easily debugged. Furthermore, there is no clear path for improving the knowledge-based scoring function employed by LUDI. A better approach may involve the use of physically-based implicit solvent methods such as the GB method developed in chapter 3. While GB is probably too slow for general docking problems, the constrained docking approach developed

in chapter 8 eliminates all six external degrees of freedom, narrowing the search space sufficiently that it should be possible to use GB.

The lethal factor (LF) inhibitor studied in chapter 9 has surprisingly high affinity considering that it was not originally designed for LF. The relatively polar nature of the binding pockets near the LF catalytic zinc suggest that polar substituents with the ability to form hydrogen bonds may be more successful than the purely hydrophobic groups that have been successful in MMP inhibitor design. Future inhibitors stand to benefit from efforts targeted specifically at LF using the methods and improvements outlined above.

An additional major methodological improvement will come from incorporating protein flexibility into the design efforts. All the drug design work described here has held the protein rigid, since full protein flexibility is currently computationally infeasible. An intermediate approach may be taken, where the protein is still held fixed, but calculations are repeated for multiple rigid structures representing the protein's conformational ensemble. Such structures may be taken from MD simulations. Since proteins often sample significantly different conformational ensembles in ligand-bound and unbound states, it is desirable that some or all of these structures be drawn from simulations of ligand-bound proteins. The difficulty of determining accurate forcefield parameters for ligand-metal interactions around the zinc has hampered these simulations. However, forthcoming improvements in quantum mechanical/molecular mechanical (QM/MM) methods should allow for these troublesome interactions to be treated with semi-empirical QM methods, so metal-ligand forcefield parameters will be unnecessary.

10.5 Conclusion

This dissertation has explored the importance of appropriately defining the solvent-solute boundary in implicit solvent models, and the improvements that may be realized by incorporating more correct boundary definitions into GB models. Additional methods and applications made feasible by the computational efficiency provided by implicit

solvation have been examined, including fast, accurate constant pH MD and prediction of protein-ligand binding for rational drug design.

Most computational scientific methods involve approximations that are made to gain computational efficiency, and implicit solvation clearly belongs to this class. It is occasionally suggested that approximate methods are really only an interim solution, useful only until the rising tide of computational power makes it possible to eliminate the errors inherent in the approximation through the use of more fundamental methods. I believe that this is short-sighted, for a number of reasons. First, it should be recognized that while computational power continues to increase, it may do so more slowly in the future. While the work presented here was being conducted, all of the world's major microprocessor manufacturers shifted their focus from increasing the speed of individual processors to increasing the number of processors in a computer. Further advances in computational power will be realized not by loading the same program onto a newer, faster processor, but by the far more difficult process of designing new programs that can effectively make use of very large numbers of processors. Second, the history of biomolecular simulation is less one of increasingly accurate simulations of the same systems and much more one of applying increasing computational power to ever larger systems and longer simulations. There is a very very long way to go before feasibly simulated systems become large enough and timescales long enough to encompass most questions of biological interest. If the target is set at simulations of a system the size of a cell for a period of one second, then there are approximately 9 orders of magnitude in duration and 4 orders of magnitude in each of three dimensions in size beyond what can be easily accomplished today.

Without question, there are limitations and errors inherent in any approximation. Some instances where these limitations are particularly noticeable for implicit solvation involve cases where specific interactions between solute atoms and individual solvent molecules are important, such as in solvent-mediated hydrogen bonds, and simulation of high salt concentrations, especially those containing divalent ions. While these sorts of difficulties may limit the application of implicit solvent, the errors introduced seem to be small enough to be negligible in most cases.

I envision a future of biomolecular simulation where implicit solvation does not

become obsolete, but plays an important role as part of a hierarchy of methods employed in hybrid simulations. Such hybrid simulations would be in the spirit of today's QM/MM simulations, where a region of interest is simulated using quantum mechanics while the remainder of the system is treated with less accurate but more efficient molecular mechanics. Expanding on this theme, future hybrid simulations may employ a wider spectrum of methods. In addition to QM and MM, treatment of the solute might be extended to include coarse-grained modeled regions, where whole residues are modeled as a single particle with simplified force calculations. Likewise, hybrid solvent modeling could involve a combination of methods with implicit methods used to treat most of the system and explicit, or even more fundamental empirical valence bond or quantum mechanical methods, used in regions where greater resolution and fidelity is required. In this way, maximum accuracy can be achieved in the parts of the system that require it, while maximum efficiency is achieved in the portions of the system that allow it, so that the greatest possible use of the available computational power can be made.

Bibliography

1. Lee B, Richards FM; *Interpretation of Protein Structures: Estimation of Static Accessibility*; Journal of Molecular Biology; **55**(3) (1971) 379.

2. Grant JA, Pickup BT, Nicholls A; *A smooth permittivity function for Poisson-Boltzmann solvation methods*; Journal of Computational Chemistry; **22**(6) (2001) 608–640.

3. Im W, Beglov D, Roux B; *Continuum Solvation Model: computation of electrostatic forces from numerical solutions to the Poisson-Boltzmann equation*; Computer Physics Communications; **111**(1-3) (1998) 59–75.

4. Lu Q, Luo R; *A Poisson-Boltzmann dynamics method with nonperiodic boundary condition*; The Journal of Chemical Physics; **119**(21) (2003) 11035–11047.

5. Lee MS, Feig M, Salsbury FR, Brooks CL; *New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations*; Journal of Computational Chemistry; **24**(11) (2003) 1348–56.

6. Lee MS, Olson MA; *Evaluation of Poisson Solvation Models Using a Hybrid Explicit/Implicit Solvent Method*; The Journal of Physical Chemistry B: Condensed matter, materials, surfaces, interfaces and biophysical chemistry; **109**(11) (2005) 5223–5236.

7. Friedrichs M, Zhou RH, Edinger SR, Friesner RA; *Poisson-Boltzmann analytical gradients for molecular modeling calculations*; The Journal of Physical Chemistry B: Condensed matter, materials, surfaces, interfaces and biophysical chemistry; **103**(16) (1999) 3057–3061.

8. Im W, Lee MS, Brooks CL; *Generalized Born model with a simple smoothing function*; Journal of Computational Chemistry; **24**(14) (2003) 1691–702.

9. Prabhu NV, Zhu PJ, Sharp KA; *Implementation and testing of stable, fast implicit solvation in molecular dynamics using the smooth-permittivity finite difference Poisson-Boltzmann method*; Journal of Computational Chemistry; **25**(16) (2004) 2049–2064.

10. Wagoner J, Baker NA; *Solvation forces on biomolecular structures: A comparison of explicit solvent and Poisson-Boltzmann models*; Journal of Computational Chemistry; **25**(13) (2004) 1623–1629.

11. Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA; *Electrostatics of nanosystems: Application to microtubules and the ribosome*; Proceedings of the National Academy of Sciences of the United States; **98**(18) (2001) 10037–10041.

12. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M; *CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations*; Journal of Computational Chemistry; **4**(2) (1983) 187–217.

13. Nina M, Beglov D, Roux B; *Atomic radii for continuum electrostatics calculations based on molecular dynamics free energy simulations*; The Journal of Physical Chemistry B: Condensed matter, materials, surfaces, interfaces and biophysical chemistry; **101**(26) (1997) 5239–5248.

14. Nina M, Im W, Roux B; *Optimized atomic radii for protein continuum electrostatics solvation forces*; Biophysical Chemistry; **78**(1-2) (1999) 89–96.

15. Swanson JMJ, Adcock SA, McCammon JA; *Optimized radii for Poisson-Boltzmann calculations with the AMBER force field*; Journal of Chemical Theory and Computation; **1**(3) (2005) 484–493.

16. Bondi A; *van der Waals Volumes and Radii*; Journal of Physical Chemistry; **68**(3) (1964) 441–451.

17. Masunov A, Lazaridis T; *Potentials of mean force between ionizable amino acid side chains in water*; Journal of the American Chemical Society; **125**(7) (2003) 1722–30.

18. Hawkins GD, Cramer CJ, Truhlar DG; *Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium*; Journal of Physical Chemistry; **100**(51) (1996) 19824–19839.

19. Gilson MK, Davis ME, Luty BA, McCammon JA; *Computation of Electrostatic Forces On Solvated Molecules Using the Poisson-Boltzmann Equation*; Journal of Physical Chemistry; **97**(14) (1993) 3591–3600.

20. Honig BH, Nicholls A; *Classical Electrostatics in Biology and Chemistry*; Science; **268**(5214) (1995) 1144–1149.

21. Madura JD, Briggs JM, Wade RC, Davis ME, Luty BA, Ilin A, Antosiewicz JM, Gilson MK, Bagheri B, Scott LR, McCammon JA; *Electrostatics and diffusion of molecules in solution: simulations with the University of Houston Brownian Dynamics program*; Computer Physics Communications; **91**(1-3) (1995) 57–95.

22. Beroza P, Case DA; *Calculations of proton-binding thermodynamics in proteins*; Energetics of Biological Macromolecules, Part B; **295** (1998) 170–189.

23. Scarsi M, Apostolakis J, Caflisch A; *Continuum electrostatic energies of macromolecules in aqueous solutions*; Journal of Physical Chemistry A; **101**(43) (1997) 8098–8106.

24. Cramer CJ, Truhlar DG; *Implicit solvation models: Equilibria, structure, spectra, and dynamics*; Chemical Reviews; **99**(8) (1999) 2161–2200.

25. Baker NA; *Improving implicit solvent simulations: a Poisson-centric view*; Current Opinion in Structural Biology; **15**(2) (2005) 137–143.

26. Luo R, David L, Gilson MK; *Accelerated Poisson-Boltzmann calculations for static and dynamic systems*; Journal of Computational Chemistry; **23**(13) (2002) 1244–1253.

27. Feig M, Brooks CL; *Recent advances in the development and application of implicit solvent models in biomolecule simulations*; Current Opinion in Structural Biology; **14**(2) (2004) 217–24.

28. Still WC, Tempczyk A, Hawley RC, Hendrickson T; *Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics*; Journal of the American Chemical Society; **112**(16) (1990) 6127–6129.

29. Hawkins GD, Cramer CJ, Truhlar DG; *Pairwise Solute Descreening of Solute Charges From a Dielectric Medium*; Chemical Physics Letters; **246**(1-2) (1995) 122–129.

30. Schaefer M, Karplus M; *A comprehensive analytical treatment of continuum electrostatics*; Journal of Physical Chemistry; **100**(5) (1996) 1578–1599.

31. Qiu D, Shenkin PS, Hollinger FP, Still WC; *The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii*; Journal of Physical Chemistry A; **101**(16) (1997) 3005–3014.

32. Edinger SR, Cortis C, Shenkin PS, Friesner RA; *Solvation free energies of peptides: Comparison of approximate continuum solvation models with accurate solution of the Poisson-Boltzmann equation*; The Journal of Physical Chemistry B: Condensed matter, materials, surfaces, interfaces and biophysical chemistry; **101**(7) (1997) 1190–1197.

33. Jayaram B, Liu Y, Beveridge DL; *A modification of the generalized Born theory for improved estimates of solvation energies and pK shifts*; The Journal of Chemical Physics; **109**(4) (1998) 1465–1471.

34. Ghosh A, Rapp CS, Friesner RA; *Generalized Born model based on a surface integral formulation*; The Journal of Physical Chemistry B: Condensed matter, materials, surfaces, interfaces and biophysical chemistry; **102**(52) (1998) 10983–10990.

35. Bashford D, Case DA; *Generalized Born Models of Macromolecular Solvation Effects*; Annual Review of Physical Chemistry; **51** (2000) 129–52.

36. Lee MS, Salsbury FR, Brooks CL; *Novel generalized Born methods*; The Journal of Chemical Physics; **116**(24) (2002) 10606–10614.

37. Felts AK, Harano Y, Gallicchio E, Levy RM; *Free energy surfaces of beta-hairpin and alpha-helical peptides generated by replica exchange molecular dynamics with the AGBNP implicit solvent model*; Proteins; **56**(2) (2004) 310–321.

38. Romanov AN, Jabin SN, Martynov YB, Sulimov AV, Grigoriev FV, Sulimov VB; *Surface Generalized Born Method: A Simple, Fast, and Precise Implicit Solvent Model beyond the Coulomb Approximation*; Journal of Physical Chemistry A; **108**(43) (2004) 9323–9327.

39. Dominy BN, Brooks CL; *Development of a generalized Born model parametrization for proteins and nucleic acids*; The Journal of Physical Chemistry B: Condensed matter, materials, surfaces, interfaces and biophysical chemistry; **103**(18) (1999) 3765–3773.

40. David L, Luo R, Gilson MK; *Comparison of generalized Born and Poisson models: Energetics and dynamics of HIV protease*; Journal of Computational Chemistry; **21**(4) (2000) 295–309.

41. Tsui V, Case DA; *Molecular Dynamics Simulations of Nucleic Acids with a Generalized Born Solvation Model*; Journal of the American Chemical Society; **122**(11) (2000) 2489–2498.

42. Calimet N, Schaefer M, Simonson T; *Protein molecular dynamics with the generalized Born/ACE solvent model*; Proteins; **45**(2) (2001) 144–158.

43. Spassov VZ, Yan L, Szalma S; *Introducing an implicit membrane in generalized Born/solvent accessibility continuum solvent models*; The Journal of Physical Chemistry B: Condensed matter, materials, surfaces, interfaces and biophysical chemistry; **106**(34) (2002) 8726–8738.

44. Simmerling C, Strockbine B, Roitberg AE; *All-atom structure prediction and folding simulations of a stable protein*; Journal of the American Chemical Society; **124**(38) (2002) 11258–11259.

45. Wang T, Wade RC; *Implicit solvent models for flexible protein-protein docking by molecular dynamics simulation*; Proteins; **50**(1) (2003) 158–169.

46. Nymeyer H, Garcia AE; *Simulation of the folding equilibrium of alpha-helical peptides: A comparison of the generalized Born approximation with explicit solvent*; Proceedings of the National Academy of Sciences of the United States; **100**(24) (2003) 13934–13939.

47. Onufriev A, Bashford D, Case DA; *Exploring protein native states and large-scale conformational changes with a modified generalized Born model*; Proteins; **55**(2) (2004) 383–94.

48. Gallicchio E, Levy RM; *AGBNP: An analytic implicit solvent model suitable for molecular dynamics simulations and high-resolution modeling*; Journal of Computational Chemistry; **25**(4) (2004) 479–499.

49. Lee MC, Duan Y; *Distinguish protein decoys by using a scoring function based on a new AMBER force field, short molecular dynamics simulations, and the generalized Born solvent model*; Proteins; **55**(3) (2004) 620–34.

50. Sigalov G, Scheffel P, Onufriev A; *Incorporating variable dielectric environments into the generalized Born model*; The Journal of Chemical Physics; **122**(9) (2005) 094511.

51. Onufriev A, Case DA, Bashford D; *Effective Born radii in the generalized Born approximation: the importance of being perfect*; Journal of Computational Chemistry; **23**(14) (2002) 1297–304.

52. Onufriev A, Bashford D, Case DA; *Modification of the generalized Born model suitable for macromolecules*; The Journal of Physical Chemistry B: Condensed matter, materials, surfaces, interfaces and biophysical chemistry; **104**(15) (2000) 3712–3720.

53. Grycuk T; *Deficiency of the Coulomb-field approximation in the generalized Born model: An improved formula for Born radii evaluation*; The Journal of Chemical Physics; **119**(9) (2003) 4817–4826.

54. Wojciechowski M, Lesyng B; *Generalized Born model: Analysis, refinement, and applications to proteins*; The Journal of Physical Chemistry B: Condensed matter, materials, surfaces, interfaces and biophysical chemistry; **108**(47) (2004) 18368–18376.

55. Feig M, Onufriev A, Lee MS, Im W, Case DA, Brooks CL; *Performance comparison of generalized born and Poisson methods in the calculation of electrostatic solvation energies for protein structures*; Journal of Computational Chemistry; **25**(2) (2004) 265–84.

56. Swanson JMJ, Mongan J, McCammon JA; *Limitations of Atom-Centered Dielectric Functions in Implicit Solvent Models*; The Journal of Physical Chemistry B: Condensed matter, materials, surfaces, interfaces and biophysical chemistry; **109**(31) (2005) 14769 –14772.

57. Nelder JA, Mead R; *A simplex method for function minimization*; Computer Journal; **7** (1965) 308–15.

58. Case DA, Darden T, III TEC, Simmerling C, Wang J, Merz KM, Wang B, Pearlman DA, Duke RE, Crowley M, Brozell S, Luo R, Tsui V, Gohlke H, Mongan J, Hornak V, Caldwell JW, Ross WS, Kollman PA; *AMBER 9*. March 2006.

59. Jones E, Oliphant T, Peterson P, *et al.*; *SciPy: Open source scientific tools for Python*. 2001–; URL `http://www.scipy.org/`.

60. Onufriev A, Case DA, Bashford D; *Structural details, pathways, and energetics of unfolding apomyoglobin*; Journal of Molecular Biology; **325**(3) (2003) 555–67.

61. Wyman J, Gill S; *Binding and linkage*; University Science Books, Mill Valley, CA. 1990.

62. Alberty R; *Thermodynamics of Biochemical Reactions*; John Wiley, New York. 2003.

63. Matthew JB, Gurd FRN, García-Moreno EB, Flanagan MA, March KL, Shire SJ; *pH-dependent Processes in Proteins*; CRC Critical Reviews in Biochemistry; **18**(2) (1985) 91–197.

64. Bashford D; *Macroscopic electrostatic models for protonation states in proteins*; Frontiers in Bioscience: A Journal and Virtual Library; **9** (2004) 1082–1099.

65. García-Moreno EB, Fitch CA; *Structural interpretation of pH and salt-dependent processes in proteins with computational methods*; Energetics of Biological Macromolecules, Part B; **380** (2004) 20–51.

66. Warshel A, Sussman F, King G; *Free-energy of Charges in Solvated Proteins - Microscopic Calculations Using a Reversible Charging Process*; Biochemistry; **25**(26) (1986) 8368–8372.

67. Pratt L, Hummer G; *Simulation and Theory of Electrostatic Interactions in Solution*; American Institute of Physics, Melville, NY. 1999.

68. Bergdorf M, Peter C, Hünenberger PH; *Influence of cut-off truncation and artificial periodicity of electrostatic interactions in molecular simulations of solvated ions: A continuum electrostatics study*; The Journal of Chemical Physics; **119**(17) (2003) 9129–9144.

69. Darden T, Pearlman DA, Pedersen LG; *Ionic charging free energies: Spherical versus periodic boundary conditions*; The Journal of Chemical Physics; **109**(24) (1998) 10921–10935.

70. Figueirido F, Delbuono GS, Levy RM; *Prediction of pK(a) shifts without truncation of electrostatic interactions: An explicit solvent calculation for succinic acid*; Journal of Physical Chemistry; **100**(16) (1996) 6389–6392.

71. Simonson T, Carlsson J, Case DA; *Proton binding to proteins: pK(a) calculations with explicit and implicit solvent models*; Journal of the American Chemical Society; **126**(13) (2004) 4167–4180.

72. Levy RM, Belhadj M, Kitchen DB; *Gaussian Fluctuation Formula for Electrostatic Free-energy Changes in Solution*; The Journal of Chemical Physics; **95**(5) (1991) 3627–3633.

73. Simonson T; *Gaussian fluctuations and linear response in an electron transfer protein*; Proceedings of the National Academy of Sciences of the United States; **99**(10) (2002) 6544–6549.

74. Simonson T; *Electrostatics and dynamics of proteins*; Reports on Progress in Physics; **66**(5) (2003) 737–787.

75. Baker NA; *Poisson-Boltzmann methods for biomolecular electrostatics*; Energetics of Biological Macromolecules, Part B; **383** (2004) 94–118.

76. Lim C, Bashford D, Karplus M; *Absolute pKa Calculations with Continuum Dielectric Methods*; Journal of Physical Chemistry; **95**(14) (1991) 5610–5620.

77. Richardson WH, Peng C, Bashford D, Noodleman L, Case DA; *Incorporating solvation effects into density functional theory: Calculation of absolute acidities*; International Journal of Quantum Chemistry; **61**(2) (1997) 207–217.

78. Klicic JJ, Friesner RA, Liu SY, Guida WC; *Accurate prediction of acidity constants in aqueous solution via density functional theory and self-consistent reaction field methods*; Journal of Physical Chemistry A; **106**(7) (2002) 1327–1335.

79. Chipman DM; *Computation of pK(a) from dielectric continuum theory*; Journal of Physical Chemistry A; **106**(32) (2002) 7413–7422.

80. Tomasi J; *Thirty years of continuum solvation chemistry: a review, and prospects for the near future*; Theoretical Chemistry Accounts: Theory, Computation, and Modeling; **112**(4) (2004) 184–203.

81. Tanford C, Kirkwood JG; *Theory of Protein Titration Curves .1. General Equations for Impenettrable Spheres*; Journal of the American Chemical Society; **79**(20) (1957) 5333–5339.

82. Bashford D, Karplus M; *pKas of Ionizable Groups in Proteins - Atomic Detail From a Continuum Electrostatic Model*; Biochemistry; **29**(44) (1990) 10219–10225.

83. Schutz CN, Warshel A; *What Are the Dielectric "Constants" of Proteins and How To Validate Electrostatic Models?*; Proteins; **44**(4) (2001) 400–417.

84. Warshel A; *Computer Modelling of Chemical Reactions in Enzymes and Solutions*; John Wiley & Sons, New York. 1991.

85. Sham YY, Chu ZT, Warshel A; *Consistent calculations of pK(a)'s of ionizable residues in proteins: Semi-microscopic and microscopic approaches*; The Journal of Physical Chemistry B: Condensed matter, materials, surfaces, interfaces and biophysical chemistry; **101**(22) (1997) 4458–4472.

86. Shurki A, Warshel A; *Structure/function correlations of proteins using MM, QM/MM, and related approaches: Methods, concepts, pitfalls, and current progress*; Advances in Protein Chemistry; **66** (2003) 249–313.

87. Alexov EG, Gunner MR; *Incorporating protein conformational flexibility into the calculation of pH-dependent protein properties*; Biophysical Journal; **72**(5) (1997) 2075–2093.

88. Georgescu RE, Alexov EG, Gunner MR; *Combining conformational flexibility and continuum electrostatics for calculating pK(a)s in proteins*; Biophysical Journal; **83**(4) (2002) 1731–1748.

89. Alexov E; *Role of the protein side-chain fluctuations on the strength of pair-wise electrostatic interactions: Comparing experimental with computed pK(a)s*; Proteins; **50**(1) (2003) 94–103.

90. Koumanov A, Karshikoff A, Friis EP, Borchert TV; *Conformational averaging in pK calculations: Improvement and limitations in prediction of ionization properties of proteins*; The Journal of Physical Chemistry B: Condensed matter, materials, surfaces, interfaces and biophysical chemistry; **105**(38) (2001) 9339–9344.

91. Kuhn B, Kollman PA, Stahl M; *Prediction of pK(a) shifts in proteins using a combination of molecular mechanical and continuum solvent calculations*; Journal of Computational Chemistry; **25**(15) (2004) 1865–1872.

92. Onufriev A, Case DA, Ullmann GM; *A novel view of pH titration in biomolecules*; Biochemistry; **40**(12) (2001) 3413–3419.

93. Poland D; *Free energy of proton binding in proteins*; Biopolymers; **69**(1) (2003) 60–71.

94. Tanford C, Roxby R; *Interpretation of Protein Titration Curves - Application to Lysozyme*; Biochemistry; **11**(11) (1972) 2192–2198.

95. Bashford D, Karplus M; *Multiple-site Titration Curves of Proteins - An Analysis of Exact and Approximate Methods for Their Calculation*; Journal of Physical Chemistry; **95**(23) (1991) 9556–9561.

96. Beroza P, Fredkin DR, Okamura MY, Feher G; *Protonation of interacting residues in a protein by a Monte Carlo method: Application to lysozyme and the photosynthetic reaction center of Rhodobacter sphaeroides*; Proceedings of the National Academy of Sciences of the United States; **88**(13) (1991) 5804–5808.

97. Mertz JE, Pettitt BM; *Molecular Dynamics at a Constant pH*; International Journal of Supercomputer Applications and High Performance Computing; **8**(1) (1994) 47–53.

98. Börjesson U, Hünenberger PH; *Explicit-solvent molecular dynamics simulation at constant pH: Methodology and application to small amines*; The Journal of Chemical Physics; **114**(22) (2001) 9706–9719.

99. Baptista AM; *Comment on "Explicit-solvent molecular dynamics simulation at constant pH: Methodology and application to small amines"*; The Journal of Chemical Physics; **116**(17) (2002) 7766–7768.

100. Börjesson U, Hünenberger PH; *pH-dependent stability of a decalysine alpha-helix studied by explicit-solvent molecular dynamics simulations at constant pH*; The Journal of Physical Chemistry B: Condensed matter, materials, surfaces, interfaces and biophysical chemistry; **108**(35) (2004) 13551–13559.

101. Baptista AM, Martel PJ, Petersen SB; *Simulation of protein conformational freedom as a function of pH: Constant-pH molecular dynamics using implicit titration*; Proteins; **27**(4) (1997) 523–544.

102. Lee MS, Salsbury FR, Brooks CL; *Constant-pH molecular dynamics using continuous titration coordinates*; Proteins; **56**(4) (2004) 738–52.

103. Khandogin J, Brooks CL; *Constant pH molecular dynamics with proton tautomerism*; Biophysical Journal; **89**(1) (2005) 141–157.

104. Bürgi R, Kollman PA, van Gunsteren WF; *Simulating proteins at constant pH: An approach combining molecular dynamics and Monte Carlo simulation*; Proteins; **47**(4) (2002) 469–80.

105. Baptista AM, Teixeira VH, Soares CM; *Constant-pH molecular dynamics using stochastic titration*; The Journal of Chemical Physics; **117**(9) (2002) 4184–4200.

106. Walczak AM, Antosiewicz JM; *Langevin dynamics of proteins at constant pH*; Physical Review E: Statistical, nonlinear, and soft matter physics; **66**(5 Pt 1) (2002) 051911.

107. Dlugosz M, Antosiewicz JM, Robertson AD; *Constant-pH molecular dynamics study of protonation-structure relationship in a heptapeptide derived from ovomucoid third domain*; Physical Review E: Statistical, nonlinear, and soft matter physics; **69**(2 Pt 1) (2004) 021915.

108. Dlugosz M, Antosiewicz JM; *Constant-pH molecular dynamics simulations: a test case of succinic acid*; Chemical Physics; **302**(1-3) (2004) 161–170.

109. Dlugosz M, Antosiewicz JM; *The impact of protonation equilibria on protein structure*; Journal of Physics: Condensed Matter; **17**(18) (2005) S1607–S1616.

110. Mongan J, Case DA, McCammon JA; *Constant pH molecular dynamics in generalized Born implicit solvent*; Journal of Computational Chemistry; **25**(16) (2004) 2038–48.

111. Nielsen JE, McCammon JA; *On the evaluation and optimization of protein X-ray structures for pKa calculations*; Protein Science: A Publication of the Protein Society; **12**(2) (2003) 313–326.

112. Wlodek ST, Antosiewicz JM, McCammon JA; *Prediction of titration properties of structures of a protein derived from molecular dynamics trajectories*; Protein Science: A Publication of the Protein Society; **6**(2) (1997) 373–382.

113. Case D, Darden T, TE Cheatham I, Simmerling C, Wang J, Duke R, Luo R, Merz K, Wang B, Pearlman D, Crowley M, Brozell S, Tsui V, Gohlke H, Mongan J, Hornak V, Cui G, Beroza P, Schafmeister C, Caldwell J, Ross W, Kollman P; *AMBER 8*. March 2004.

114. Wang JM, Cieplak P, Kollman PA; *How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?*; Journal of Computational Chemistry; **21**(12) (2000) 1049–1074.

115. Wang J, Wang W, Kollman PA; *Antechamber: An accessory software package for molecular mechanical calculations*; Abstracts of Papers of the American Chemical Society; **220**(Part 1) (2001) 135–COMP.

116. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Zakrzewski VG, J A Montgomery J, Stratmann RE, Burant JC, Dapprich S, Millam JM, Daniels AD, Kudin KN, Strain MC, Farkas O, Tomasi J, Barone V, Cossi M, Cammi R, Mennucci B, Pomelli C, Adamo C, Clifford S, Ochterski J, Petersson GA, Ayala PY, Cui Q, Morokuma K, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Cioslowski J, Ortiz JV, Baboul AG, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Gomperts R, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Gonzalez C, Challacombe M, Gill PMW, Johnson BG, Chen W, Wong MW, Andres JL, Head-Gordon M, Replogle ES, Pople JA; *Gaussian 98 (Revision A.11.4)*. 1998.

117. Bashford D, Case DA, Dalvit C, Tennant L, Wright PE; *Electrostatic Calculations of Side-chain pK(a) Values in Myoglobin and Comparison With NMR Data for Histidines*; Biochemistry; **32**(31) (1993) 8045–8056.

118. Kyte J; *Structure in Protein Chemistry*; Garland Publishing, Inc. 1995; page 64.

119. Wolfram Research Inc; *Mathematica, Version 5*. 2003.

120. Chen JL, Noodleman L, Case DA, Bashford D; *Incorporating Solvation Effects Into Density-functional Electronic-structure Calculations*; Journal of Physical Chemistry; **98**(43) (1994) 11059–11068.

121. Demchuk E, Wade RC; *Improving the continuum dielectric approach to calculating pK(a)s of ionizable groups in proteins*; Journal of Physical Chemistry; **100**(43) (1996) 17373–17387.

122. Vriend G; *WHAT IF: A molecular modeling and drug design program*; Journal of Molecular Graphics; **8**(1) (1990) 52–56.

123. Hooft RWW, Sander C, Vriend G; *Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures*; Proteins; **26**(4) (1996) 363–376.

124. Gilson MK, Honig BH; *Energetics of Charge-Charge Interactions in Proteins*; Proteins; **3**(1) (1988) 32–52.

125. Nielsen JE, Vriend G; *Optimizing the hydrogen-bond network in Poisson-Boltzmann equation-based pK(a) calculations*; Proteins; **43**(4) (2001) 403–412.

126. Warshel A; *Calculations of Enzymatic Reactions: Calculations of pKa, Proton Transfer Reactions, and General Acid Catalysis Reactions in Enzymes*; Biochemistry; **20**(11) (1981) 3167–3177.

127. Mehler EL, Guarnieri F; *A self-consistent, microenvironment modulated screened Coulomb potential approximation to calculate pH-dependent electrostatic effects in proteins*; Biophysical Journal; **77**(1) (1999) 3–22.

128. Amadei A, Linssen ABM, Berendsen HJC; *Essential Dynamics of Proteins*; Proteins; **17**(4) (1993) 412–425.

129. Lindahl E, Hess B, van der spoel D; *GROMACS 3.0: a package for molecular simulation and trajectory analysis*; Journal of molecular modeling; **7**(8) (2001) 306–317.

130. de Hoon MJL, Imoto S, Nolan J, Miyano S; *Open source clustering software*; Bioinformatics; **20**(9) (2004) 1453–1454.

131. Humphrey W, Dalke A, Schulten K; *VMD: Visual molecular dynamics*; Journal of Molecular Graphics; **14**(1) (1996) 33–38.

132. Karplus M, McCammon JA; *Molecular dynamics simulations of biomolecules*; Nature Structural Biology; **9**(9) (2002) 646–652.

133. Cheatham TE, Brooks BR; *Recent advances in molecular dynamics simulation towards the realistic representation of biomolecules in solution*; Theoretical Chemistry Accounts: Theory, Computation, and Modeling; **99**(5) (1998) 279–288.

134. Cheatham TE, Young MA; *Molecular dynamics simulation of nucleic acids: Successes, limitations, and promise*; Biopolymers; **56**(4) (2000) 232–256.

135. Cheatham TE, Kollman PA; *Molecular dynamics simulation of nucleic acids*; Annual Review of Physical Chemistry; **51** (2000) 435–471.

136. Hamelberg D, Mcfail-Isom L, Williams LD, Wilson WD; *Flexible structure of DNA: Ion dependence of minor-groove structure and dynamics*; Journal of the American Chemical Society; **122**(43) (2000) 10513–10520.

137. Hamelberg D, Williams LD, Wilson WD; *Influence of the dynamic positions of cations on the structure of the DNA minor groove: Sequence-dependent effects*; Journal of the American Chemical Society; **123**(32) (2001) 7745–7755.

138. Hamelberg D, Williams LD, Wilson WD; *Effect of a neutralized phosphate backbone on the minor groove of B-DNA: molecular dynamics simulation studies*; Nucleic acids research; **30**(16) (2002) 3615–3623.

139. Palmer RG; *Broken Ergodicity*; Advances in Physics; **31**(6) (1982) 669–735.

140. Voter AF; *A method for accelerating the molecular dynamics simulation of infrequent events*; The Journal of Chemical Physics; **106**(11) (1997) 4665–4677.

141. Voter AF; *Hyperdynamics: Accelerated molecular dynamics of infrequent events*; Physical Review Letters; **78**(20) (1997) 3908–3911.

142. Grubmüller H; *Predicting slow structural transitions in macromolecular systems: Conformational flooding*; Physical Review E: Statistical physics, plasmas, fluids, and relat; **52**(3) (1995) 2893–2906.

143. Sugita Y, Okamoto Y; *Replica-exchange molecular dynamics method for protein folding*; Chemical Physics Letters; **314**(1-2) (1999) 141–151.

144. Torrie GM, Valleau JP; *Non-physical Sampling Distributions in Monte Carlo Free Energy Estimation: Umbrella Sampling*; Journal of Computational Physics; **23**(2) (1977) 187–199.

145. Wu XW, Wang SM; *Enhancing systematic motion in molecular dynamics simulation*; The Journal of Chemical Physics; **110**(19) (1999) 9401–9410.

146. Berne BJ, Straub JE; *Novel methods of sampling phase space in the simulation of biological systems*; Current Opinion in Structural Biology; **7**(2) (1997) 181–189.

147. Steiner MM, Genilloud PA, Wilkins JW; *Simple bias potential for boosting molecular dynamics with the hyperdynamics scheme*; Physical Review B: Condensed matter; **57**(17) (1998) 10236–10239.

148. Rahman JA, Tully JC; *Puddle-skimming: An efficient sampling of multidimensional configuration space*; The Journal of Chemical Physics; **116**(20) (2002) 8750–8760.

149. Pal S, Fichthorn KA; *Accelerated molecular dynamics of infrequent events*; Chemical Engineering Journal; **74**(1-2) (1999) 77–83.

150. Gong XG, Wilkins JW; *Hyper molecular dynamics with a local bias potential*; Physical Review B: Condensed matter; **59**(1) (1999) 54–57.

151. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA; *A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules*; Journal of the American Chemical Society; **117**(19) (1995) 5179–5197.

152. Loncharich RJ, Brooks BR, Pastor RW; *Langevin Dynamics of Peptides: The Frictional Dependence of Isomerization Rates of N-acetylalanyl-N'-methylamide*; Biopolymers; **32**(5) (1992) 523–535.

153. Pearlman DA, Case DA, Caldwell JW, Ross WS, Cheatham TE, Debolt S, Ferguson D, Seibel G, Kollman PA; *AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules*; Computer Physics Communications; **91**(1-3) (1995) 1–41.

154. Ryckaert JP, Ciccotti G, Berendsen HJC; *Numerical Integration of Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of N-alkanes*; Journal of Computational Physics; **23**(3) (1977) 327–341.

155. Shen TY, Tai KH, Henchman RH, McCammon JA; *Molecular dynamics of acetylcholinesterase*; Accounts of Chemical Research; **35**(6) (2002) 332–340.

156. Garcia AE; *Large-Amplitude Nonlinear Motions in Proteins*; Physical Review Letters; **68**(17) (1992) 2696–2699.

157. Hamelberg D, Mongan J, McCammon JA; *Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules*; The Journal of Chemical Physics; **120**(24) (2004) 11919–29.

158. de Groot BL, Daura X, Mark AE, Grubmüller H; *Essential dynamics of reversible peptide folding: memory-free conformational dynamics governed by internal hydrogen bonds*; Journal of Molecular Biology; **309**(1) (2001) 299–313.

159. Xiong B, Huang XQ, Shen LL, Shen JH, Luo XM, Shen X, Jiang HL, Chen KX; *Conformational flexibility of beta-secretase: molecular dynamics simulation and essential dynamics analysis*; Acta pharmacologica Sinica; **25**(6) (2004) 705–13.

160. Lee J, Suh SW, Shin S; *Computational studies of essential dynamics of Pseudomonas cepacia lipase*; Journal of Biomolecular Structure and Dynamics; **18**(2) (2000) 297–309.

161. Crabbe MJC, Cooper LR, Corne DW; *Use of essential and molecular dynamics to study gammaB-crystallin unfolding after non-enzymic post-translational modifications*; Computational Biology and Chemistry; **27**(4-5) (2003) 507–10.

162. Arcangeli C, Bizzarri AR, Cannistraro S; *Concerted motions in copper plastocyanin and azurin: an essential dynamics study*; Biophysical Chemistry; **90**(1) (2001) 45–56.

163. Kendall RA, Apra E, Bernholdt DE, Bylaska EJ, Dupuis M, Fann GI, Harrison RJ, Ju JL, Nichols JA, Nieplocha J, Straatsma TP, Windus TL, Wong AT; *High performance computational chemistry: An overview of NWChem a distributed parallel application*; Computer Physics Communications; **128**(1-2) (2000) 260–283.

164. Huitema H, van Liere R; *Interactive Visualization of Protein Dynamics*; IEEE Visualization 2000 Proceedings; **11**.

165. Skiles JW, Gonnella NC, Jeng AY; *The design, structure, and clinical update of small molecular weight matrix metalloproteinase inhibitors*; Current medicinal chemistry; **11**(22) (2004) 2911–77.

166. Whittaker M, Floyd CD, Brown P, Gearing AJ; *Design and therapeutic application of matrix metalloproteinase inhibitors*; Chemical Reviews; **99**(9) (1999) 2735–76.

167. Coussens LM, Fingleton B, Matrisian LM; *Matrix metalloproteinase inhibitors and cancer: trials and tribulations*; Science; **295**(5564) (2002) 2387–92.

168. Breuer E, Frant J, Reich R; *Recent non-hydroxamate matrix metalloproteinase inhibitors*; Expert Opinion On Therapeutic Patents; **15**(3) (2005) 253–269.

169. Hajduk PJ, Shuker SB, Nettesheim DG, Craig R, Augeri DJ, Betebenner D, Albert DH, Guo Y, Meadows RP, Xu L, Michaelides M, Davidsen SK, Fesik SW; *NMR-based modification of matrix metalloproteinase inhibitors with improved bioavailability*; Journal of medicinal chemistry; **45**(26) (2002) 5628–39.

170. Puerta DT, Cohen SM; *A bioinorganic perspective on matrix metalloproteinase inhibition*; Current topics in medicinal chemistry; **4**(15) (2004) 1551–73.

171. Liu ZD, Piyamongkol S, Liu DY, Khodr HH, Lu SL, Hider RC; *Synthesis of 2-amido-3-hydroxypyridin-4(1H)-ones: novel iron chelators with enhanced pFe3+ values*; Bioorganic and Medicinal Chemistry; **9**(3) (2001) 563–73.

172. Finnegan MM, Rettig SJ, Orvig C; *A Neutral Water-soluble Aluminum Complex of Neurological Interest*; Journal of the American Chemical Society; **108**(16) (1986) 5033–5035.

173. Puerta DT, Cohen SM; *Examination of novel zinc-binding groups for use in matrix metalloproteinase inhibitors*; Inorganic chemistry; **42**(11) (2003) 3423–30.

174. Chen L, Rydel TJ, Gu F, Dunaway CM, Pikul S, Dunham KM, Barnett BL; *Crystal structure of the stromelysin catalytic domain at 2.0 A resolution: inhibitor-induced conformational changes*; Journal of Molecular Biology; **293**(3) (1999) 545–57.

175. Puerta DT, Schames JR, Henchman RH, McCammon JA, Cohen SM; *From model complexes to metalloprotein inhibition: a synergistic approach to structure-based drug discovery*; Angewandte Chemie-international Edition; **42**(32) (2003) 3772–4.

176. Bohm HJ; *On the Use of Ludi to Search the Fine Chemicals Directory for Ligands of Proteins of Known 3-dimensional Structure*; Journal of Computer-Aided Molecular Design; **8**(5) (1994) 623–632.

177. Puerta DT, Cohen SM; *Elucidating drug-metalloprotein interactions with tris(pyrazolyl)borate model complexes*; Inorganic chemistry; **41**(20) (2002) 5075–82.

178. Hajduk PJ, Sheppard G, Nettesheim DG, Olejniczak ET, Shuker SB, Meadows RP, Steinman DH, Carrera GM, Marcotte PA, Severin J, Walter K, Smith H, Gubbins E, Simmer R, Holzman TF, Morgan DW, Davidsen SK, Summers JB, Fesik SW; *Discovery of potent nonpeptide inhibitors of stromelysin using SAR by NMR*; Journal of the American Chemical Society; **119**(25) (1997) 5818–5827.

179. Knight CG, Willenbrock F, Murphy G; *A novel coumarin-labelled peptide for sensitive continuous assays of the matrix metalloproteinases*; FEBS letters; **296**(3) (1992) 263–6.

180. Castelhano AL, Billedeau R, Dewdney N, Donnelly S, Horne S, Kurz LJ, Liak TJ, Martin R, Uppington R, Yuan ZY, Krantz A; *Novel Indolactam-based Inhibitors of Matrix Metalloproteinases*; Bioorganic and medicinal chemistry letters; **5**(13) (1995) 1415–1420.

181. Puerta DT, Lewis JA, Cohen SM; *New beginnings for matrix metalloproteinase inhibitors: identification of high-affinity zinc-binding groups*; Journal of the American Chemical Society; **126**(27) (2004) 8388–9.

182. Fray MJ, Dickinson RP, Huggins JP, Occleston NL; *A potent, selective inhibitor of matrix metalloproteinase-3 for the topical treatment of chronic dermal ulcers*; Journal of medicinal chemistry; **46**(16) (2003) 3514–25.

183. Johnson LL, Pavlovsky AG, Johnson AR, Janowicz JA, Man CF, Ortwine DF, Purchase CF, White AD, Hupe DJ; *A rationalization of the acidic pH dependence for stromelysin-1 (Matrix metalloproteinase-3) catalysis and inhibition*; The Journal of biological chemistry; **275**(15) (2000) 11026–33.

184. Gorden AEV, Xu J, Raymond KN, Durbin P; *Rational design of sequestering agents for plutonium and other actinides*; Chemical Reviews; **103**(11) (2003) 4207–82.

185. Pezard C, Berche P, Mock M; *Contribution of individual toxin components to virulence of Bacillus anthracis*; Infection and immunity; **59**(10) (1991) 3472–7.

186. Collier RJ, Young JAT; *Anthrax toxin*; Annual review of cell and developmental biology; **19** (2003) 45–70.

187. Agrawal A, Pulendran B; *Anthrax lethal toxin: a weapon of multisystem destruction*; Cellular and molecular life sciences: CMLS; **61**(22) (2004) 2859–65.

188. Park JM, Greten FR, Li ZW, Karin M; *Macrophage apoptosis by anthrax lethal factor through p38 MAP kinase inhibition*; Science; **297**(5589) (2002) 2048–51.

189. Dell'Aica I, Donà M, Tonello F, Piris A, Mock M, Montecucco C, Garbisa S; *Potent inhibitors of anthrax lethal factor from green tea*; EMBO reports; **5**(4) (2004) 418–22.

190. Fridman M, Belakhov V, Lee LV, Liang FS, Wong CH, Baasov T; *Dual effect of synthetic aminoglycosides: antibacterial activity against Bacillus anthracis and inhibition of anthrax lethal factor*; Angewandte Chemie-international Edition; **44**(3) (2005) 447–52.

191. Lee LV, Bower KE, Liang FS, Shi J, Wu D, Sucheck SJ, Vogt PK, Wong CH; *Inhibition of the proteolytic activity of anthrax lethal factor by aminoglycosides*; Journal of the American Chemical Society; **126**(15) (2004) 4774–5.

192. Numa MMD, Lee LV, Hsu CC, Bower KE, Wong CH; *Identification of novel anthrax lethal factor inhibitors generated by combinatorial Pictet-Spengler reaction followed by screening in situ*; Chembiochem: A European Journal of Chemical Biology; **6**(6) (2005) 1002–6.

193. Turk BE, Wong TY, Schwarzenbacher R, Jarrell ET, Leppla SH, Collier RJ, Liddington RC, Cantley LC; *The structural basis for substrate and inhibitor selectivity of the anthrax lethal factor*; Nature Structural and molecular biology; **11**(1) (2004) 60–6.

194. Panchal RG, Hermone AR, Nguyen TL, Wong TY, Schwarzenbacher R, Schmidt J, Lane D, McGrath C, Turk BE, Burnett J, Aman MJ, Little S, Sausville EA, Zaharevitz DW, Cantley LC, Liddington RC, Gussio R, Bavari S; *Identification of small molecule inhibitors of anthrax lethal factor*; Nature Structural and molecular biology; **11**(1) (2004) 67–72.

195. Forino M, Johnson S, Wong TY, Rozanov DV, Savinov AY, Li W, Fattorusso R, Becattini B, Orry AJ, Jung D, Abagyan RA, Smith JW, Alibek K, Liddington RC, Strongin AY, Pellecchia M; *Efficient synthetic inhibitors of anthrax lethal factor*; Proceedings of the National Academy of Sciences of the United States; **102**(27) (2005) 9499–504.

196. Min DH, Tang WJ, Mrksich M; *Chemical screening by mass spectrometry to identify inhibitors of anthrax lethal factor*; Nature Biotechnology; **22**(6) (2004) 717–23.

197. Menard A, Papini E, Mock M, Montecucco C; *The cytotoxic activity of Bacillus anthracis lethal factor is inhibited by leukotriene A4 hydrolase and metallopeptidase inhibitors*; The Biochemical Journal; **320**.

198. Shultz CS, Dreher SD, Ikemoto N, Williams JM, Grabowski EJJ, Krska SW, Sun Y, Dormer PG, Dimichele L; *Asymmetric hydrogenation of N-sulfonylated-alpha-dehydroamino acids: toward the synthesis of an anthrax lethal factor inhibitor*; Organic Letters; **7**(16) (2005) 3405–8.

199. Tonello F, Seveso M, Marin O, Mock M, Montecucco C; *Screening inhibitors of anthrax lethal factor*; Nature; **418**(6896) (2002) 386.

200. Cummings RT, Salowe SP, Cunningham BR, Wiltsie J, Park YW, Sonatore LM, Wisniewski D, Douglas CM, Hermes JD, Scolnick EM; *A peptide-based fluorescence resonance energy transfer assay for Bacillus anthracis lethal factor protease*; Proceedings of the National Academy of Sciences of the United States; **99**(10) (2002) 6603–6.

201. Montecucco C, Tonello F, Zanotti G; *Stop the killer: how to inhibit the anthrax lethal factor metalloprotease*; Trends in biochemical sciences; **29**(6) (2004) 282–5.

202. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA; *Development and Testing of a General Amber Force Field*; Journal of Computational Chemistry; **25**(9) (2004) 1157–74.

203. Nymyer H, Gnanakran S, Garc A; *Atomic simulations of protein folding, using the replica exchange algorithm*; Meth. Enzymol.; **383** (2004) 119–149.

204. Simmerling C, Fox T, Kollman PA; *Use of locally enhanced sampling in free energy calculations: Testing and application to the alpha ->beta anomerization of glucose*; Journal of the American Chemical Society; **120**(23) (1998) 5771–5782.

205. Hamelberg D, Shen T, McCammon JA; *Relating kinetic rates and local energetic roughness by accelerated molecular-dynamics simulations*; The Journal of Chemical Physics; **122**(24) (2005) 241103.