The Dynamic Nature of the Folded and Unfolded States of the Villin Headpiece Helical Subdomain

A Dissertation Presented

by

Lauren Wickstrom

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Biochemistry and Structural Biology

Stony Brook University

May 2009

Stony Brook University

The Graduate School

Lauren Wickstrom

We, the dissertation committee for the above candidate for the
Doctor of Philosophy degree,
Hereby recommend the acceptance of this dissertation.

Carlos Simmerling, Ph. D., Advisor
Department of Chemistry, Stony Brook University

Daniel P. Raleigh, Ph. D., Advisor
Department of Chemistry, Stony Brook University

Steven O. Smith, Ph. D., Chair
Department of Biochemistry and Cell Biology, Stony Brook University

Robert C. Rizzo, Ph. D., Third Member
Department of Applied Mathematics and Statistics, Stony Brook University

Marivi Fernandez-Serra, Ph. D., Outside Member
Department of Physics and Astronomy, Stony Brook University

This dissertation is accepted by the Graduate School

Lawrence Martin
Dean of the Graduate School

Abstract of the Dissertation

The Dynamic Nature of the Folded and Unfolded States of the Villin Headpiece Helical Subdomain

by

Lauren Wickstrom

Doctor of Philosophy
in
Biochemistry and Structural Biology

Stony Brook University

2009

The symbiotic relationship between experiment and simulation is necessary for a complete understanding of biomolecular structure and dynamics. Computational approaches can provide structural populations and atomic detail, and describe motion not visible on the macroscopic level, which can be used to interpret the experimental average ensemble as well develop new experiments. In turn, simulations rely on experimental observables for validation of a particular model or method.

In this work, both tools are used collaboratively to study the structure of the folded and the unfolded states of proteins. Solution NMR and X-ray structures of the folded state are widely used as a reference for simulations and experiments. Recent work has shown that the denatured state contains structure that is important for understanding protein stability and the folding pathway. This knowledge can be utilized to understand and treat protein misfolding diseases.

One of the key model systems, the 36-residue villin headpiece helical subdomain (HP36), was chosen for these studies because of its simple topology, small size and fast folding properties. Structures of HP36 have been determined using X-ray crystallography and NMR spectroscopy, but the two structures exhibited clear differences. Molecular dynamics simulations and experimental double mutant cycles were used to show that the X-ray structure is the better representation of the folded state in solution.

Previous experimental evidence has suggested that there is residual structure in the denatured state of HP36. Fragment analysis has shown that the three individual helices of HP36 lack significant structure compared to a larger fragment containing the first two helices (HP21). These techniques, however, are low resolution and are unable to quantify low levels of helical structure and whether it occurs in the same regions as HP36. Simulations were used to quantify the structure in all of the fragments. The HP21 ensemble contains less helical structure than predicted by NMR experimental observables possibly due to deficiencies in sampling and the force field. To address these limitations, simulation methodology and models were investigated.

Table of Contents

List of Figures

List of Tables

List of Publications

1. Okur, A., **L. Wickstrom**, Layten, M., Geney, R., Song, K., Hornak, V., Simmerling, C. "Improved Efficiency of Replica Exchange Simulations through Use of a Hybrid Explicit/Implicit Solvation Model" *J. Chem. Theory Comput.* 2006, 420-433.

2. **L. Wickstrom**; Okur, A., Song, K., Raleigh, D.P., Simmerling, C. "The Unfolded State of the Villin Headpiece Helical Subdomain: Computational Studies of the Role of Locally Stabilized Structure" *J. Mol. Biol.* 2006, 360, 1094-1107.

3. Roe, D., Okur, A, **L. Wickstrom**, Hornak, V., Simmerling, C. "Secondary Structure in Generalized Born Solvent Models: Comparison of Conformational Ensembles and Free energy of Solvent Polarization from Explicit and Implicit Solvation" *J. Phys. Chem B.* 2007, 111, 1846-1857.

4. **L. Wickstrom**, Bi, Y, Hornak, V., Raleigh, D.P., Simmerling, C. "Reconciling the Solution and X-ray Structures of the Villin Headpiece Helical Subdomain: Molecular Dynamics Simulations and Double Mutant Cycles Reveal a Stabilizing Cation-pi Interaction" *Biochemistry* 2007, 46, 3624 – 3634.

5. Okur, A., **L. Wickstrom**, Simmerling, C. "Evaluation of Salt bridge Structure and Energetics in Peptides Using Explicit, Implicit, and Hybrid Solvation Models" *J. Chem. Theory Comput.* 2008, 4, 488 – 498.

6. **L. Wickstrom**, Ding, F., Patsalo, V., Okur, A,, Simmerling, C. "Improved Conformational Sampling in Explicit solvent: Application of the Reservior Replica Exchange Molecular Dynamics to Small Peptide Systems" 2008, **Submitted.**

7. **L. Wickstrom**, Okur, A., Simmerling, C. "Evaluating the Performance of the FF99SB Force Field Based on NMR Scalar Coupling Data" 2009 **In Press.**

8. **L. Wickstrom**, Balius, T., Okur, A., Maier, J., Duke, R., Simmerling, C. The Effect of Different Explicit Water Models on Peptide and Protein Conformational Preferences and Energetics" 2009 **Submitted.**

Acknowledgements

I would like to thank both of my advisors, Professor Carlos Simmerling and Professor Daniel Raleigh, for their guidance over the past 6 years. I would have not been able accomplish this work without their support. I also thank them for tolerating me during my "bumpy" period in graduate school and not giving up on me. If there is anything I have learned from this experience, it is that you get back what you put in. Lastly, I thank them for their friendship.

I want to also thank Professor Steven Smith, Professor Robert Rizzo and Professor Marivi Fernandez-Serra for serving as members on my committee. I thank them for their valuable suggestions and feedback over the years.

I would also like to thank the past and present members of the Simmerling lab. This lab has been like a 2nd family over the years to me. Everyone has been part of my life in so many aspects. I would like to thank Dr. Asim Okur for providing me with excellent training so I can continue his work. I would like to also thank Carlos's Computational Angels, Kerri Goldgraben and Dr. Salma Rafi. Lastly, I would like to thank AJ Campbell and Amber Carr for their friendship over the past few years. I would also like to thank the past and present members of the Raleigh lab, especially Dr. Andisheh Abedini for her friendship over the years. I would also like to thank Dr. Yuefeng Tang, Dr. Yuan Bi, and Wenli Meng who have been involved with the HP36 project. I also would like to thank Trent Balius and Vadim Patsalo for their contributions on the methodology projects.

I would like to thank all of the friends who supported me over the past few years outside of work. Lastly, I would like to thank my family. Your support has meant the world to me.

1. Introduction

1.1 Symbiotic Relationship between Experiments and Computational Studies

The symbiotic relationship between experiment and simulation is necessary for a complete understanding of biomolecular structure and dynamics. Molecular modelling can provide structural populations and atomic detail, describing motions not visible on the macroscopic level. The information can be used to interpret the experimental average ensemble as well as develop new experiments. Ab-initio protein folding simulations, protein design and comparative modeling methods have provided accurate models for amino acid sequences in cases where there were no experimentally determined structures [1-4]. In addition, simulations have provided flexibility to static models in regions which has been shown to be important in the docking field [5]. Lastly, simulations can provide a view of the interconverting structural populations in the experimental average. A recent comparison between molecular dynamics (MD) simulations and electron paramagnetic resonance (EPR) data has confirmed the existence of a wide open conformation and the dominance of the semi-open conformation in the unbound ensemble of HIV protease [6].

In turn, simulations rely on experimental observables for the development and validation of a particular method. Force fields rely heavily on experimental measurements such as bond lengths and angles and thermodynamic measurements for parameterization. Force fields are validated through the comparison of experimental and calculated values from simulations (ie. scalar coupling constants, order parameters, residual dipolar couplings and solvation free energies). This experimental data can also be used to optimize the computational model.

1.2 The Importance of Understanding Folded and Unfolded State Structure

1.2.1 What is the protein folding problem?

In this work, experiments and molecular modelling are used collaboratively to study the structure of the folded and the unfolded states involved in protein folding. The protein folding problem focuses on the propensity of amino acid sequences to quickly fold from the denatured state to the native 3-dimensional structure. In the 1960's, Anfinsen established that the only necessary information needed for a protein to fold was entirely contained in the amino acid sequence [7]. Furthermore, in 1968, Levinthal pointed out that the number of unfolded conformations is so enormous that a protein could not possibly find the native state by random sampling of all conformations. Together, these two insights suggest that nature finds a shortcut, or a folding pathway to find the most stable functional conformation in a reasonable manner [8]. This state is located at the free energy minimum in the funnel model for folding (Figure 1-1) [9].

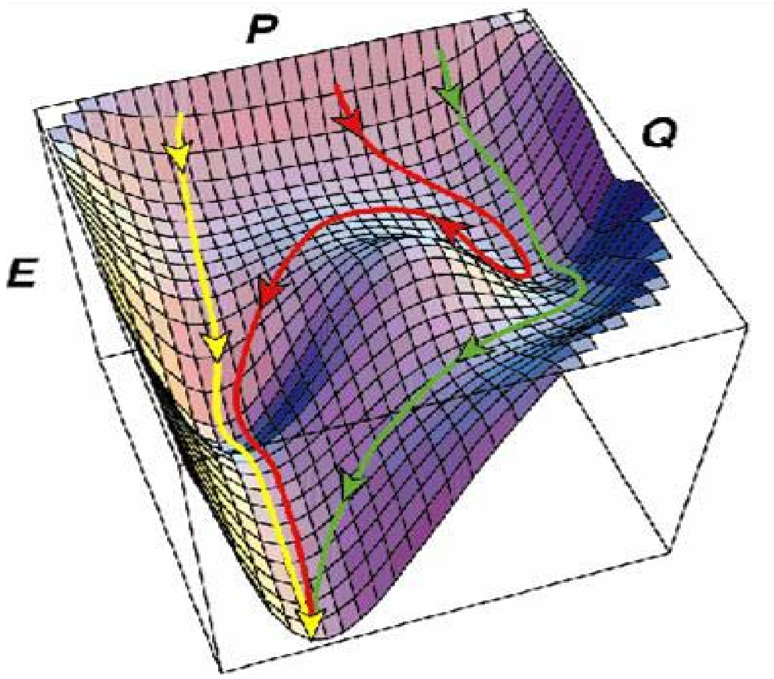


Figure 1-1. Folding funnel of lysozyme. E is the free energy of the system, Q is the proportion of native contacts formed and P is the configurational entropy. Adapted from Dobson et al. [9].

1.2.2 Importance of the folded state

It is important to understand the structure of the folded state of a protein. Proteins are involved in almost all processes within the body. If we are able to understand the structure of their functional form, we can start to dissect the parts of the protein important for biological activity. For example, HIV protease contains a catalytic triad which is responsible for cleaving the proteins involved in the HIV lifecycle. This knowledge can be used in areas such as drug design to create inhibitors that target the recognition pockets surrounding the active site of HIV protease.

The two main methods for structural determination of the folded state are nuclear magnetic resonance (NMR) and X-ray crystallography. Currently, the protein databank holds approximately 87 % X-ray structures and 13 % NMR structures. Presently, X-ray structures can provide a wealth of structural information due to the increasing amount of high resolution structures (346 structures are in the databank with a resolution between .5 and 1.0 Å). NMR provides an ensemble of structures which can take into account structural variability in regions like loops. Despite the advances, both approaches suffer from limitations which are discussed further in Chapter 2.

1.3 Studying the Unfolded State

1.3.1 Why is studying the unfolded state important?

Protein function is a vital part of the scientific community's interest. Because of this interest, the native (functional) state gets the most attention. The unfolded or denatured state had been assumed to have a random coil structure and was thought to have little significance in biological activities. Recent studies have shown that the denatured state can contain large amounts of structure which are of particular importance to protein stability, folding kinetics and mechanism [10]. Mutations can play a major role in altering the stability of the unfolded state and making a native state conformation more or less favored [11-13]. Identification of significant structure in the denatured state allows for a greater understanding of the protein folding pathway and the temporal ordering of folding events. This knowledge can be utilized to understand protein misfolding diseases such as prion and amyloid related illnesses [14, 15].

1.3.2 Difficulties in resolving the unfolded state ensemble experimentally

The denatured state is more difficult to study experimentally because of its low solubility, dynamic nature and the size of the structural ensembles compared to the native state. Experimental difficulties arise because of the short lifetime of the denatured state in refolding experiments and small populations of it at equilibrium. IR and CD experiments can only suggest basic information about residual secondary structure if the unfolded structural ensemble is well populated [16]. Small angle X-ray or neutron scattering (SAXS and SANS) can only produce information about properties of the average ensemble of the denatured state [17]. In more recent years, increasingly sophisticated NMR techniques have been developed that can resolve unfolded protein structure with more detail [18]. By shifting the equilibrium towards the unfolded state by using mild chemical denaturant, alteration of the pH [19], temperature [20] or mutation [21, 22], it is possible to study the denatured state [23]. While these techniques give accurate structural descriptions of highly denatured *in vitro* populations, they cannot give accurate descriptions of the denatured state under physiological conditions.

One indirect technique used to overcome the problems of studying the denatured state under native conditions is to analyze peptide fragments of secondary structure from the whole protein. According to the diffusion-collision theory of protein folding, secondary structure formation is followed by tertiary structure formation.[24] Peptide fragment analysis provides the local propensity for secondary structure formation and a potential look at structures in the early stages of folding [24, 25]. However, the lack of tertiary interactions with the rest of the protein chain can give an incomplete picture. In certain cases, larger fragments with more than 2 elements of secondary structure have suggested the presence of tertiary contacts in the denatured state [26, 27].

1.4 How Can Simulations Help Us Study the Protein Folding Problem?

In addition to experiments, MD simulations have investigated aspects of protein folding [28]. Simulations starting from a native state have been useful for testing stability of an experimental structural model [29-31], understanding hydration dynamics [32] and interactions in the folded state [29]. MD simulations starting from the unfolded state have been useful for understanding the protein folding pathways for various model systems [33-35]. These simulations enhance the understanding of more mechanistic details of protein folding and structure. They also allow us to examine sparsely populated unfolded and non-native structures that can influence experimental results. In addition, high temperature MD has been employed to study transition states in the unfolding pathway [22, 36].

1.4.1 The limitations of simulations

Simulations face a different set of limitations, which include the accuracy of the model, the force field, and time scale. Realistic, detailed simulations come at a high computational cost, forcing many to minimize and approximate properties of their system such as the representation of the protein or the solvent. Some force fields can predict wrong structures due to biases in parameters [37] or produce over stabilized structures for certain systems. Another problem is that computer power limits the simulation time. Even with the best resources, very few studies of protein folding have generated 1 μs of data [34]. Since proteins take from milliseconds to seconds to fold, much of the folding process can not be observed. Therefore, accurate populations at equilibrium cannot be generated. This time disadvantage also limits the size of the protein one is able to study and the number of structures sampled in a simulation. High temperature MD has been employed because the rates of denaturation are faster [22]. As the temperature increases. the more likely it is that a protein will unfold and have no significant structure or no physiologically relevant structures. In the next section, the focus will be on investigating MD methodology and its relevance to the work in this thesis.

1.5 Simulation Methodology

1.5.1 Force field

In molecular mechanics, force fields account for different kinds of energetic interactions a particle can experience. More complex force fields may be more accurate but will also require more computational time, further shortening simulations. The basic force field equation accounts for bond stretching, angle bending, torsions, electrostatic interactions and van der Waals forces (Equation 1-1). With initial Cartesian coordinates, it is possible to calculate the potential energy (U) of a protein.

$$U(r) = \sum_{bonds} K_r (r - r_o)^2 + \sum_{angles} K_\theta (\vartheta - \vartheta)^2 + \sum_{dihedrals} \frac{V_n}{2}(1 + \cos[n\phi - \lambda]) + \sum_{i<j}^{atoms} \frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \sum_{i<j}^{atoms} \frac{q_i q_j}{\varepsilon R_{ij}}$$

Equation 1-1. General force field equation.

The potential energy can be used to solve for the force (F) at a particular coordinate (x) (Equation 1-2).

$$F = -\frac{dU}{dx}$$

Equation 1-2. Relationship between force and potential energy.

The integration of Newton's second law provides the next coordinates and velocities for an atom of mass (m) at a particular time (t) (Equation 1-3) [38].

$$\frac{d^2x}{dt^2} = \frac{F_x}{m}$$

Equation 1-3. Newton's 2nd law of motion.

This process is repeated on the atoms at the new coordinates and velocities generating a trajectory for that particular molecule in the simulation.

Therefore, the accuracy of the force field is quite important for the generation of the correct ensemble of structures. Each force field has a unique set of parameters to input into the potential energy equation. The most commonly used force fields are available through AMBER [39], CHARMM [40], OPLS [41] and GROMOS [42]. Experimental comparison is a necessity to determine the accuracy of the energy function. Recent work has focused on using scalar coupling constants [43, 44] and NMR relaxation techniques [45-48] to verify force field quality. In Chapter 6, we evaluate the AMBER ff99SB with a variety of J-coupling constants.

1.5.2 Solvation

Accurate modeling of water is essential since it is involved with most biological interactions. Solvation properties are especially important at the solute-solvent interface compared to the bulk solvent. At solvent-solute interface, bridging waters play a key role in protein stability. The two solvation models used in MD simulations are the explicit solvent and implicit solvent water models.

1.5.2.1 Explicit solvent models

Explicit solvent is the more accurate choice for solvation effects. Under these conditions, the protein is solvated by many individual water molecules in a unit cell using periodic boundary conditions. The number of solvent molecules depends on the desired concentration of the system and the type of simulation. Simulations of a folded protein at room temperature will require a smaller box than at higher temperatures where unfolding may take place. With the increase in system size, these simulations are quite computationally expensive.

The accuracy of the water model will also play a role in the computational expense. The expense of a water model can vary depending on how many sites are included (ie TIP3P vs TIP4P) [49], and if certain effects are included in the model (ie. quantum effects or polarizability). If one is interested more in the solute behavior, this level of theory may not be necessary to observe accurate dynamics. In Chapter 7, we compare the effects of using two rigid water models, TIP3P [49] and TIP4P-Ew [50] on the interactions in small peptides and proteins.

1.5.2.2 Implicit models

Implicit models, such as Poisson Boltzmann (PB) and Generalized Born [51] (GB), have been used to reduce the computational expense of explicit water. The methods estimate the effect of water by calculating the average free energy of solvation for a solute molecule. This reduces the system size drastically as well as eliminates the solvent friction during the simulation. The result is a faster conformational search for your biomolecule of interest.

For accurate modeling, PB is the better choice for implicit solvation; however its implementation in molecular dynamics is computationally demanding [52]. Furthermore, GB is known to cause such artifacts such as the overstabilization of salt bridges [33, 53-57] and α-helices [58, 59]. There appears to be a need for the inclusion of the first explicit solvation shell to capture effects of the solute-solvent interface [57-61]. In Chapter 3, we compare the results of explicit and implicit water simulations.

1.5.3 Enhanced sampling with replica exchange molecular dynamics

Replica exchange molecular dynamics (REMD) or parallel tempering has been used to overcome the sampling problem and the high temperature problem [62, 63]. In temperature REMD, multiple non-interacting MD simulations are run over a range of different temperature (Figure 1-2). These replicas are allowed to exchange with each other according to a transition probability (Equation 1-4).

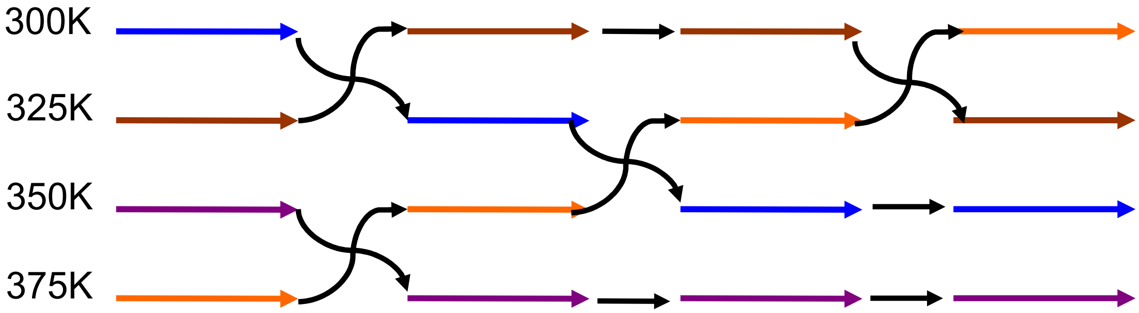


Figure 1-2. Replica exchange ladder of temperatures. Replica space is represented by different colors and temperature space is represented by each corresponding temperature.

$$\frac{\rho(X \to X')}{\rho(X' \to X)} = \exp\{(\frac{1}{k_B T_m} - \frac{1}{k_B T_n})(U(q^{[i]}) - U(q^{[j]}))$$

Equation 1-4. Transition probability for the neighboring replicas in REMD. $\rho(X \to X')$ is the exchange probability between states X and X' and $\rho(X' \to X)$ is the exchange probability between states X' and X. The exchanges are made between replicas i and j , which at located at the temperature Tm and Tn. kB is the Boltzmann constant. U is the potential energy at position q.

Through exchanges with a high temperature structures, lower temperature simulations can escape local minima allowing for the system to reach equilibrium. High temperature replicas can also exchange with lower temperature replicas producing a simulated annealing effect. This increases the amount of sampling of structures as compared to regular MD where structures may get trapped in local minima [53, 54, 64-70]. The weight factors in the REMD equations are Boltzmann weights, which drives the simulation to equilibrium, while the transition probability has been constructed to maintain canonical properties for each of the temperatures. REMD is used in Chapter 3-7 with different small peptide model systems.

1.5.3.1 Limitations of REMD

The REMD approach becomes especially challenging in explicit solvent. As the system size grows larger, the number of solvent molecules required increases. REMD rapidly becomes computationally unfeasible because the number of replicas needed to span a given temperature range increases with the square root of the number of degrees of freedom in the system [63, 71]. Solvent viscosity also slows the conformational search making it harder for a non-native structure to fold to a native conformation [72, 73]. Lastly, the benefits of high temperatures are limited to temperature dependent processes such as protein unfolding [74-78]. To our knowledge converged REMD simulations in explicit solvent from independent starting conformations have only been reported for short helical or unstructured peptides [58, 79, 80].

To overcome the problems encountered in standard REMD, new variations of REMD have been developed. One solution has been to discard some of the solvent degrees of freedom during the derivation of the REMD exchange probability. Explicit/Implicit Hybrid REMD and replica exchange with solute tempering methods are examples of this approach [58, 81]. While these methods reduce the amount of required replicas, convergence is still comparable to standard REMD. A second solution is to perform REMD with a converged structural reservoir to improve convergence. In Reservoir REMD (R-REMD) [82], a high temperature is used to generate a structural pool which eliminates many of the problems with temperature dependence if the optimum temperature is found. Nevertheless, there are difficulties in obtaining a converged ensemble for one temperature for larger systems. In Chapter 5, there is further discussion pertaining to R-REMD and its application to systems in explicit solvent. Another approach is Hamiltonian REMD, which uses a biasing functioning to scale the replicas. Several studies have applied this approach [71, 83, 84]. This approach eliminates any problems with temperature but finding a proper reaction coordinate can be challenging.

1.6 Model System Used to Study Protein Folding – Villin Headpiece Helical Subdomain

The system studied in half of this thesis is the villin headpiece helical subdomain (HP36). Villin is an actin regulatory protein located in the epithelial cells and microvilli of the gut and kidneys [85]. Villin is composed of seven domains: a gelsolin core and the headpiece. The gelsolin core is made up of six repeating homologous domains while the headpiece is made up of a single domain. The 76 residue headpiece is located at the C-terminus of the protein villin and contains one of the F-actin binding sites. Most studies use a 67 residue construct of the headpiece, HP67, missing the first 9 residues of the N-terminus, which maintains the properties of the entire headpiece. This fragment of the larger protein can fold and bind actin independently of the whole protein [86]. The N-terminus of HP67 (10-41) contains only one short helix while the C-terminus (35 residues) contains three helices and can fold independently of the whole headpiece domain (Figure 3) [86-88]. The N68H mutant of this 35 residue peptide has been studied by the Eaton group [89, 90]. The C-terminus has also been studied as a 36 residue construct due to the methionine used in the expression system in NMR [88] and X-ray crystallographic [91] studies. The methionine is located at the N-terminus and labeled as residue 41.

HP36 is one of the key model systems for experimental and computational protein folding studies [26, 34, 87, 88, 92-103] because of its simple topology, small size and fast folding properties. This small system is one of the fastest cooperatively folding proteins, folding on the time scale of microseconds [98-100]. The folded structure of this subdomain is made up of three α-helices and a hydrophobic core of three phenylalanines (Figure 1-4). The work in this thesis focuses on studying the folded state and unfolded state structure of HP36.

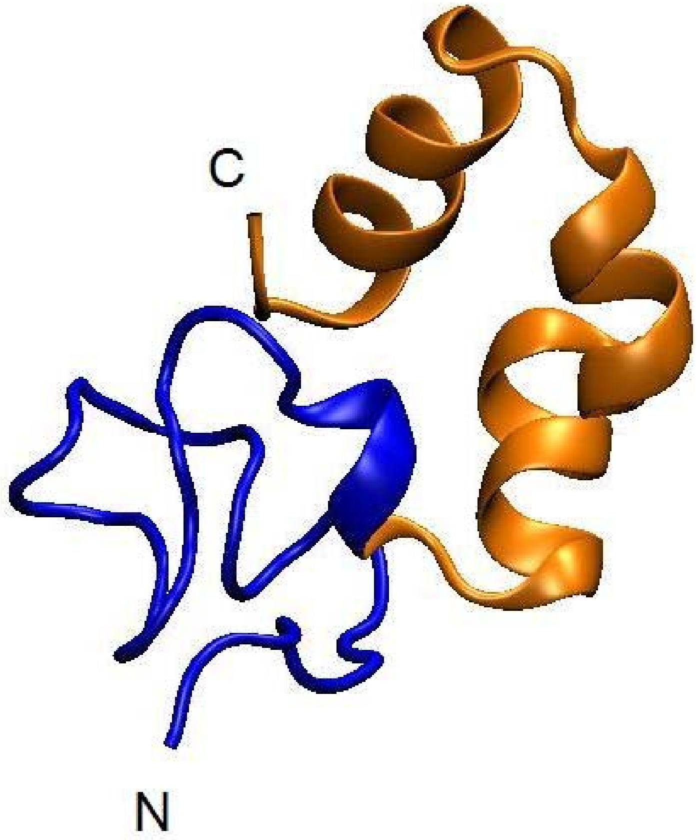


Figure 1-3. NMR structure of the villin headpiece (HP67) (pdb code 1QQV [86]). The section is blue represents the N-terminus and the section in orange is the C-terminus.

1.7 Aims of this Thesis

This thesis contains two sections. The first section focuses on studying the folded and unfolded state of the villin headpiece helical subdomain. The second section focuses on improving the simulation methodology used to study the problems presented in the first section. Chapter 2 investigates which structural model is the better representation of HP36 based on MD simulations and experimental validation. Chapter 3 focuses on supplementing low resolution experimental techniques with structural ensembles obtained from REMD for the isolated helices of HP36 to model the unfolded state under native conditions. Chapter 4 continues the unfolded state studies on the larger fragment containing helix-1 and helix-2. In Chapter 4, questions are raised about the effects of force field, sampling quality and the water model. Chapter 5 discusses the application of the R-REMD method to systems containing explicit solvent. Chapter 6 examines the performance of ff99SB with scalar coupling constants using two different solvent models with two polyalanine systems. Chapter 7 investigates the conformational preferences and energetics of small model peptides and a protein in TIP3P and TIP4P-Ew solvent models.

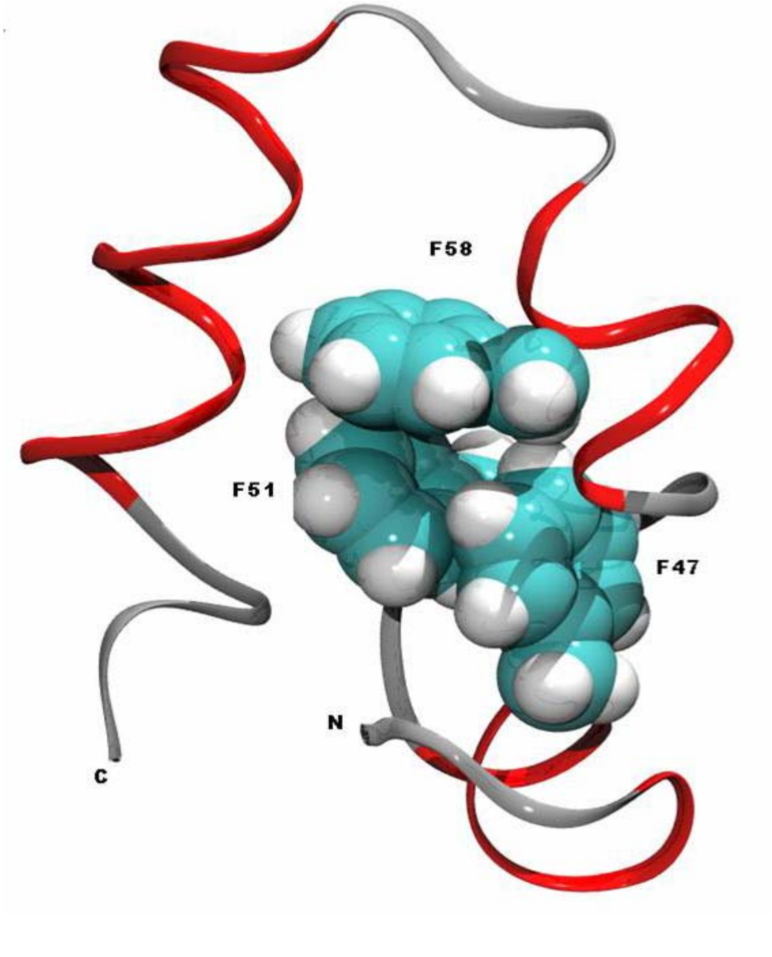


Figure 1-4. NMR structure of villin headpiece helical subdomain (HP36) (pdb code 1VII [88]). The backbone of the helices are in red. The rest of the backbone is colored silver. The phenylalanines, that makeup the hydrophobic core, are in cyan and white.

2. Reconciling the Solution and X-ray Structures of the Villin Headpiece Helical Subdomain: Molecular dynamics and Double Mutant Cycles Reveal a Stabilizing Cation-pi Interaction

Abstract

The 36 residue helical subdomain of the villin headpiece, HP36, is one of the smallest cooperatively folded proteins, folding on the microsecond timescale. The domain is an extraordinarily popular model system for both experimental and computational studies of protein folding. The structure of HP36 has been solved using X-ray crystallography and NMR spectroscopy, with the resulting structures exhibiting differences in helix packing, van der Waals contacts and hydrogen bonding. It is important to determine the solution structure of HP36 with as much accuracy as possible since this structure is widely used as a reference for simulations and experiments. We complement the existing data by using all-atom molecular dynamics simulations with explicit solvent to evaluate which of the experimental models is the better representation of HP36 in solution. After 50 ns of simulation initiated with the NMR structure, we observed that the protein spontaneously adopts structures with a backbone conformation, core packing and C-capping motif on the third helix that are more consistent with the crystal structure. We also examined hydrogen bonding and sidechain packing interactions between D44 and R55 and between F47 and R55 respectively, which were observed in the crystal structure but not present in the NMR-based solution structure. Simulations

showed large fluctuations in the distance between D44 and R55, while the distance between F47 and R55 remained stable, suggesting the formation of a cation-pi interaction between those residues. Experimental double mutant cycles confirmed that the F47/R55 pair has a larger energetic coupling than the D44/R55 interaction. Overall, these combined experimental and computational studies show that the X-ray crystal structure is the better reference structure for HP36 in solution at neutral pH. Our analysis also shows how detailed molecular dynamics simulations can help bridge the gap between NMR and crystallographic methods.

Acknowledgments

The material presented in this chapter has been published (Lauren Wickstrom*, Yuan Bi*, Viktor Hornak, Daniel P. Raleigh*, Carlos L. Simmerling 2007 *Biochemistry* 46, 3624-3634, “*“ represents joint first authors). This chapter contains direct excerpts from the manuscripts (presented with minor modifications) written by Lauren Wickstrom and Yuan Bi with suggestions and revisions from Professor Daniel P. Raleigh and Professor Carlos Simmerling. Yuan Bi conducted the experimental studies. Additional computer time and assistance was provided by Roberto Gomperts and the SGI Engineering group. We thank C. James McKnight for helpful discussions.

2.1 Introduction

The villin headpiece helical subdomain (HP36), the C-terminal portion of the villin headpiece, is the shortest naturally occurring sequence which has been shown to fold cooperatively (Figure 2-1). Infrared temperature jump [100], laser fluorescence [99, 104] and NMR lineshape analysis [98] techniques have measured the folding of HP36 to occur on the microsecond time scale. Its rapid folding, small size and simple topology of three helices have made this domain an extremely popular system for experimental [26, 87-90, 97-100, 104-107] and computational [34, 79, 93-95, 108-119] studies. Much of this work relies on using the folded structure as a reference and thus the accuracy of the known HP36 structure is of particular importance.

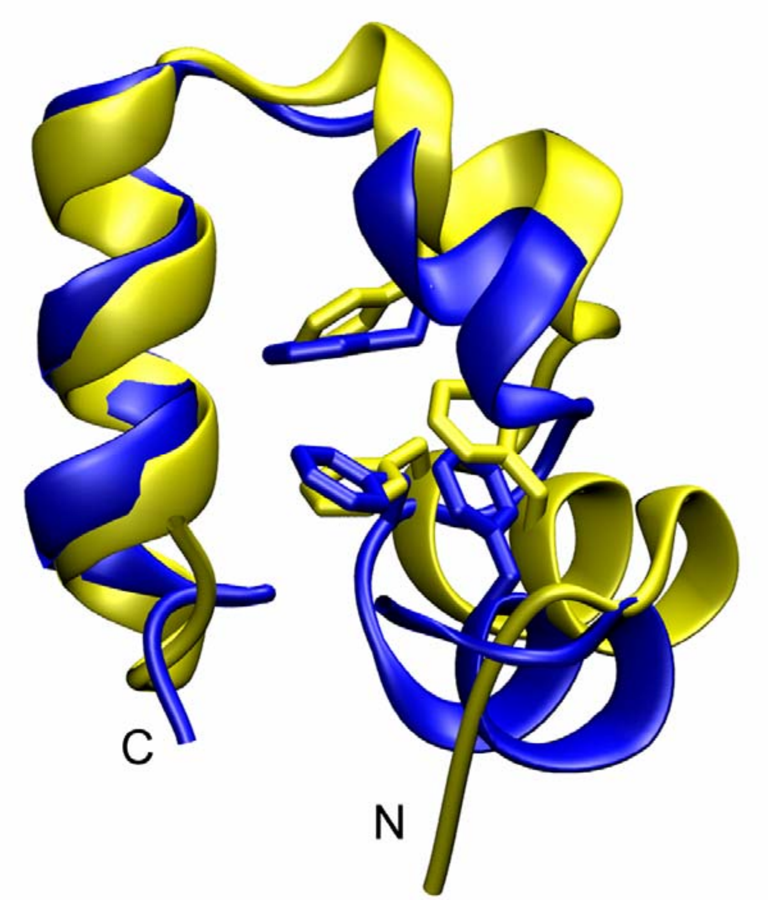


Figure 2-1. Two experimental structures of the villin helical subdomain showing only the backbone (ribbons) and heavy atoms for the 3 phenylalanines in the core (F47, F51 and F58). The NMR structure of HP36 (pdb code-1VII) is colored blue and the X-ray structure (pdb code – 1YRF) is colored yellow. Differences in the backbone and the phenylalanine core packing are highlighted using a best fit alignment on the backbone residues L62 to F76.

Several structures have been solved for HP36, one by NMR and the others by X-ray crystallographic methods [87, 89]. These structures vary in the hydrophobic core packing, interhelical H-bonds and in the length of the helices. In addition, two potentially important sidechain contacts differ significantly between the NMR and X-ray structures: F47/R55 (4.3 Å (X-ray) and 6.3 Å (NMR)) and D44/R55 (2.7 Å (X-ray) and 7.9 Å (NMR)) (Figure 2-2 A&B**)**. In the X-ray structure, the F47/R55 pair forms a van der Waals contact which could be particularly stabilizing as a cation-pi interaction, while D44/R55 form a hydrogen bond (D44-Oδ1 and R55-Nε). Neither contact is present in the NMR structures. These differences may arise from changes in the HP36 sequence used in the two sets of experiments, although this seems unlikely. The crystallographic study employed the N68H mutant of HP36 (to facilitate fluorescence studies) and also lacks the N-terminal methionine incorporated by the expression system used for the NMR study (note that we adopt the typical numbering convention [87, 88] for HP36, in which L42 follows the N-terminal methionine). Another possible reason for the structural differences could be the variation of experimental conditions such as pH or temperature. There was significant deviation in the pH between structural determinations; the NMR structure was solved at pH 3.7 in contrast to the more neutral conditions of the crystallography experiment (pH 6.7). An alternate explanation for the observed structural differences is that they arise from methodological limitations conditions; these frequently give rise to differences in structures of the same protein solved using different techniques. In general, NMR structures are less precise than X-ray structures, particularly if only homonuclear methods are used. Nevertheless, X-ray structures can suffer from effects due to crystal packing; the resulting contacts may have a local influence on conformational preferences.

The small size of HP36 and its correspondingly large surface area to volume ratio could make crystal contacts play an important role. On the other hand, crystallographic data is often collected at low temperatures which might result in the dampening of thermal motions that are present under physiological conditions.

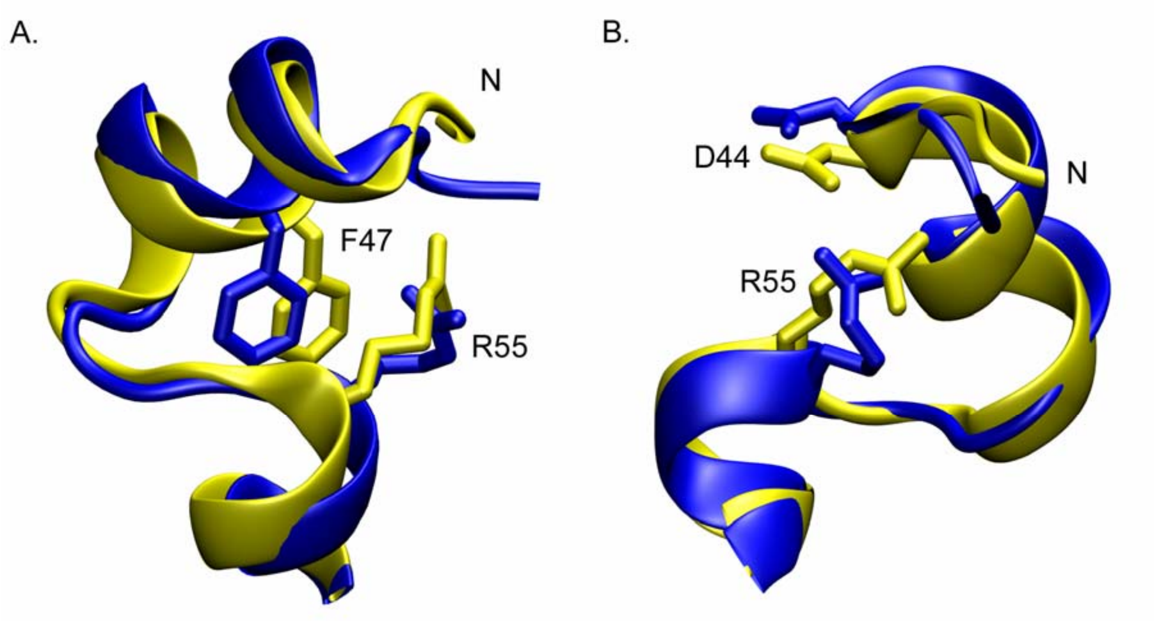


Figure 2-2. Comparison of sidechain interactions in the X-ray and NMR structures, using a best fit alignment on residues L42 to P62. (A) The R55 and F47 sidechains are shown in both the NMR (blue) and X-ray structure (yellow). In the X-ray structure, R55 is involved with a van der Waals contact with F47 and a hydrogen bond with D44. (B) In the X-ray structure, R55-Nε forms a hydrogen bond with D44-Oδ1 in contrast to the NMR structure where the atoms are almost 8 Å apart. The N-terminus is labeled.

Many computational studies have used HP36 as a model system for development and validation of protein folding methods and for optimization of force field parameters [34, 79, 92-95, 101, 108-110, 112-119]. If the native reference structure is not correct, the basis of these studies may not be valid. For example, the structure of the first helix and the C-terminus vary in the ensemble of NMR structures [88, 118] and many MD studies have therefore neglected these regions of the experimental structure when evaluating their success. Nevertheless, most simulations are performed at neutral pH and thus it is not clear if the simulations should be compared to the NMR structure from pH 3.7. A better structural model for neutral conditions would be invaluable for further work in understanding the folding and stability of this important model system for protein folding.

Accurate computational studies can provide an alternate method to study conformational behavior and alleviate the uncertainty about which structure is the better representation of the folded state in solution. In principle, molecular dynamics (MD) simulations can supply detailed information with spatial and time resolution that exceed the ability of NMR and X-ray experiments, providing insight into the role of specific interactions that may not be readily accessible through experiments that probe averages over rapidly interconverting ensembles.

Here, we conducted all-atom MD simulation in explicit solvent using the NMR structure of HP36 in order to gain insight into the details of the folded state in solution. The simulation diverges from the initial NMR structure and spontaneously adopts a structure with much greater similarity to the X-ray structure, arguing that the X-ray structure is a more accurate representation of the structure in solution at neutral pH. In addition, two residue pairs, D44/R55 and F47/R55, spontaneously formed contacts during the simulation, with the F47/R55 pair appearing to be more stable. These interactions were reported in the crystal structure but were not present in the ensemble of structures generated by the NMR studies. Thus we conclude that the F47/R55 may play an important role in stabilizing HP36 in solution. We acknowledge that simulation models can be limited in accuracy and any predictions should be tested through direct experimentation. In order to validate our computational observations, we employed an experimental double mutant cycle analysis. The results are consistent with our simulation data, and suggest that the interaction between F47 and R55 plays a role in stabilizing the native state through a cation-pi interaction. Overall, the results show how properly validated MD simulations can provide an avenue to test the stability and validity of structural models that were derived from experimental data.

2.2 Methods

2.2.1 Computational

The numbering system corresponds to that used for the full length villin headpiece, with the sequence M41–F76 (MLSDE DFKAV FGMTR SAFAN LPLWK QQNLKK EKGLF). HP36 has free N and C-termini that were modeled in the charged state. This sequence and termini correspond to those used in the experimental studies. All sidechains for Asp, Glu, Lys, and Arg were charged during the simulation. All calculations employed Amber version 8 [39] and used the ff99SB modification [45] of the Amber ff99 force field [120, 121]. SHAKE [122] was used to constrain bonds involving hydrogen. The time step was 2 fs. The temperature was maintained using the weak coupling algorithm [123] with a thermostat of 37 °C (310 K) and the pressure was equilibrated to 1 atm. All production simulations were performed using the NVT ensemble. An independent simulation using the NPT ensemble provided similar results (Data not shown).

Solvation plays a key role in biomolecular structural preferences and thus accurate treatment of solvation is essential for the investigation of structural propensities in simulations. Explicit solvent models can be highly effective, particularly when water has non-bulk properties and interacts directly with the solute [124]. Implicit models such as the semi-analytical Generalized Born model (GB) [51] are attractive because they are computationally less expensive and can converge more rapidly than simulations in explicit water due to lack of solvent viscosity. While GB has been widely used for protein folding studies by a number of groups, other investigators have reported poor results including secondary structural bias and ion pairing issues [53-55]. Our previous studies on fragments of HP36 have shown that the use of explicit water produced results which were much more consistent with experimental trends than those obtained with implicit solvent [79]. Consequently, we used explicit solvent in our simulations of HP36, in a truncated octahedral box using periodic boundary conditions with Particle Mesh Ewald (PME) [125] and a direct space cutoff of 8 Å. In order to investigate the influence of long-range periodicity, two additional simulations were run: one with the Isotropic Periodic Sum (IPS) [126] non-lattice method with a cutoff of 8 Å, and another with an atom-based nonbonded cutoff of 12 Å with no smoothing function. Simulations were initiated from the NMR structure (PDB ID 1VII) surrounded by 2327 TIP3P [49] waters molecules and equilibrated at 310 K for 50 ps with harmonic restraints on solute atoms, followed by minimization with gradually reduced positional restraints. The restraints were reduced from 5 kcal/mol*Å to 1 kcal/mol*Å to 0.5 kcal/mol*Å. After minimization, three 5 ps MD simulations were performed with the same gradually reduced restraints at constant pressure (1 atm) and temperature (310 K) in order to generate starting structures. The production simulations of the NMR structure were 50 ns in length for two PME simulations with different random seeds for assignment of velocities, and 30 ns for the IPS and cutoff simulations respectively. As a control, the X-ray structure (PDB ID 1YRF) was setup with the same amount of waters and equilibrated in a similar fashion. This simulation was run for 30 ns using PME.

2.2.2 Data analysis

The last 5 ns of the simulation were used for cluster analysis and DSSP calculations. Cluster analysis was performed with Moil-view [127] using all atoms as a similarity criterion with average linkage. Clusters were formed with the bottom-up approach using a similarity cutoff of 2.5 Å. DSSP analysis and calculation of distances, RMSD values, and radius of gyration were done using the ptraj module in Amber. Distances between sidechains were calculated using selected heavy atoms as indicated in the text. Potential mean forces (PMF) for the distances between the selected heavy atoms were calculated according to equation (Equation 2-1). Error bars were estimated for the PMF by averaging two independent simulations and subsequently subtracting the PMF of an individual simulation from the average PMF.

$$\Delta G = -RT \ln (N_i/N_0)$$

Equation 2-1. 1 Relative free energy calculated with histogram analysis. ΔG_i is the relative free energy bin i, N_i is the population of a particular histogram bin along the reaction coordinates that were employed and N_0 is the most populated bin.

2.3 Results

2.3.1 Simulations of the NMR structure

Figure 2-3 shows the backbone RMSD *versus* time and RMSD distributions calculated during the last 5 ns for selected regions of HP36 during the simulation. The RMSD is shown relative to both the NMR and X-ray structures. At the end of the equilibration period, the backbone RMSD (residues L42-L75) to each experimental structure was ~ 2.0 Å (Figure 2-3). At 8ns, a structural transition occurred causing the overall backbone RMSD (X-ray) to drop 1.0 Å below the RMSD (NMR). This greater similarity to the X-ray structure persisted throughout the remainder of the simulation.

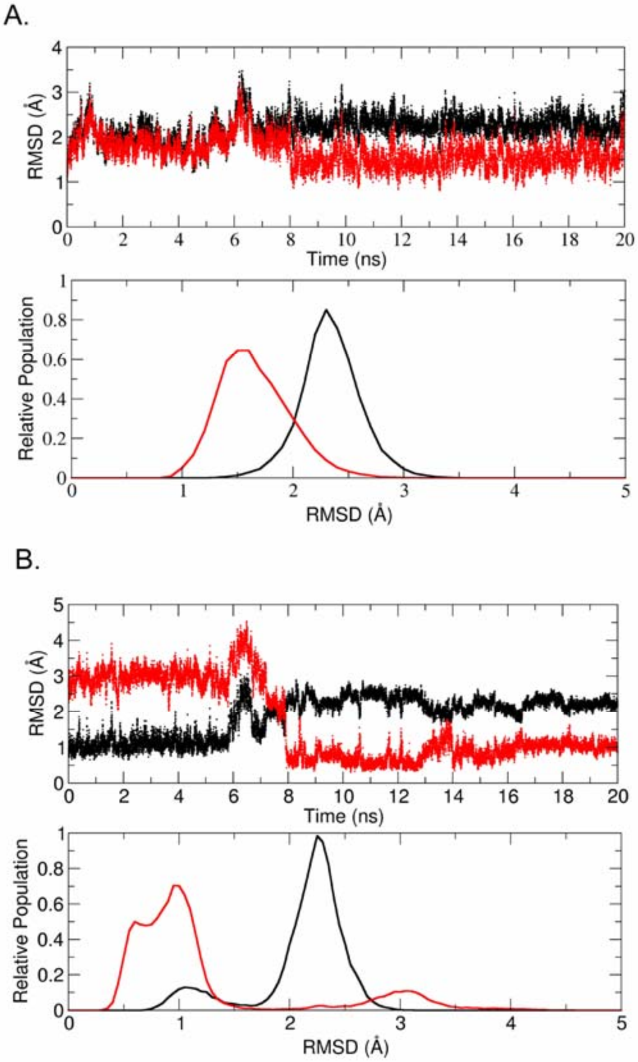


Figure 2-3. Time evolution and histogram distributions of the heavy atom backbone RMSD of (A) residues L42 to F76; and (B) residues P62 to F76 during the simulation of the NMR structure. Each calculation was performed using both the NMR (black) and X-ray (red) structures as the reference. A transition occurs near 8 ns, resulting in lower RMSD values compared to the X-ray structure. The C-terminal region (B) shows a particularly dramatic change from the initial NMR structure to one that much more closely matches the X-ray structure.

In Figure 2-3B, the RMSD relative to the X-ray structure of the region containing helix-3 (residues P62-F76) demonstrates even more clearly a switch during the simulation from similarity to the initial NMR structure to a greater similarity to the X-ray structure, as indicated by a reduction in the RMSD to the X-ray structure from 3 - 4 Å to 0.5 – 1.0 Å. Clearly, the simulation shows the inclination of HP36 to sample structures with a backbone similar to the X-ray structure despite being initiated with the NMR solution structure. The RMSD values for the two other helices remained stable and also showed a clear preference for the X-ray structure (Figure 2-4A and 2-4B).

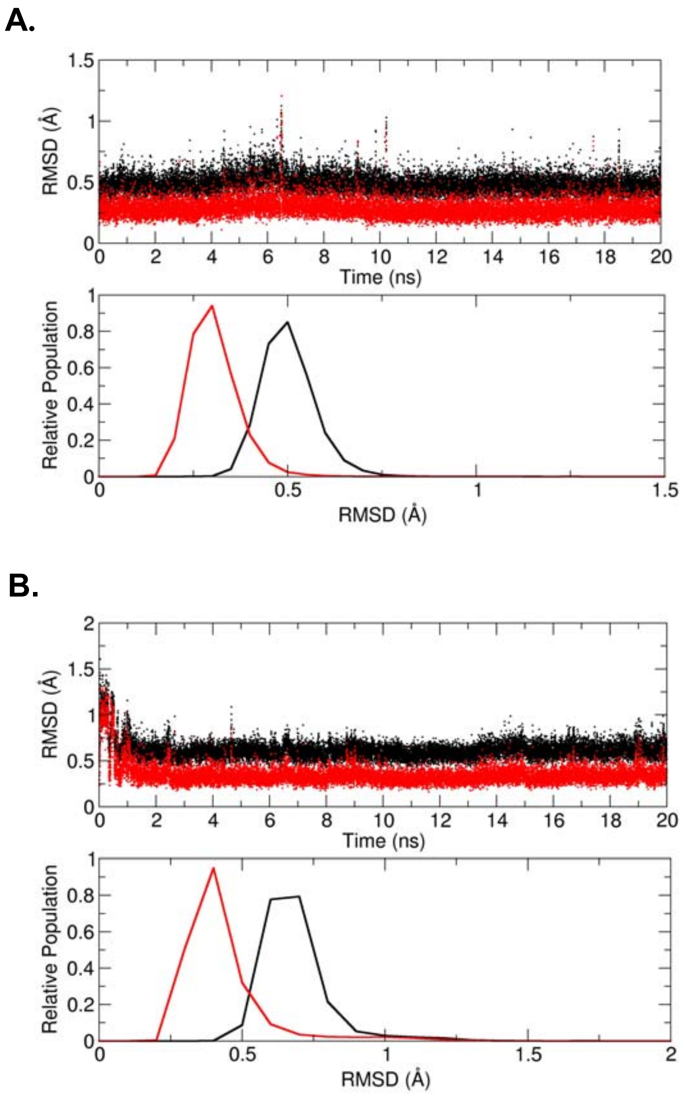


Figure 2-4. Time evolution and distributions of the heavy atom backbone RMSD of the (A) residues 43 to 49 (helix 1); and (B) residues 54 to 59 (helix 2). Each calculation was performed using the NMR (black) and X-ray (red) as a reference structure. The first and second helix remain quite stable during the simulation. Both helices have backbone structures are more structurally similar to the X-ray structure despite being initiated in the NMR structure.

In order to investigate the source of the large reduction in RMSD relative to the X-ray structure, a best fit alignment was performed on residues 61 to 74 to compare the differences before and after the structural transition. In Figure 2-5A, the NMR, X-ray and simulation structures are shown. The conformations of the C-terminus differ significantly between the X-ray and the NMR structure. The simulation structure spontaneously converts from the conformation in the NMR structure to that in the X-ray structure, concomitant with formation of three hydrogen bonds that stabilize the observed conformation. G74 forms a C-capping interaction with K70 and K71 at the end of helix-3, along with an additional hydrogen bond formed between K70 and L75. Figure 2-5B shows the time evolution of these hydrogen bond distances. In the beginning of the simulation, all three distances are 4 - 9 Å. At 8 ns, the distances are reduced to 2 - 3 Å, indicating formation of the hydrogen bonds that may play an important role in stabilizing the C-terminal helix. Importantly, all three hydrogen bonds are present in the X-ray structure but absent in the NMR structure (Figure 2-5A).

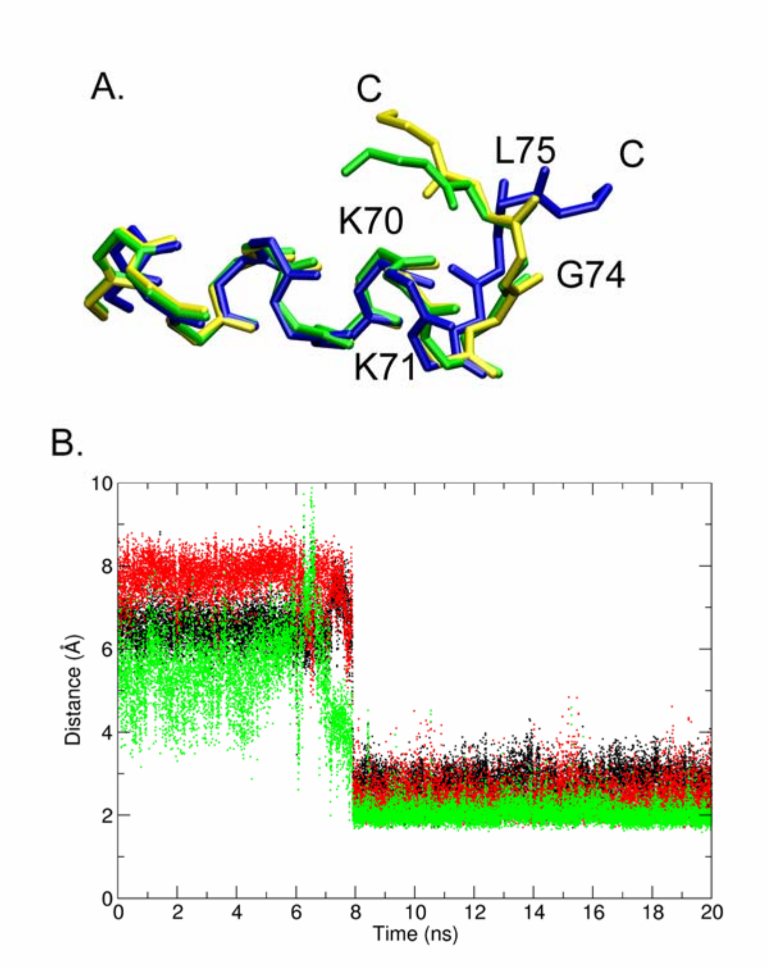


Figure 2-5. (A) Comparison of the C-terminal region (P62-F76) in the X-ray (yellow), NMR (blue) and simulation (green) structures. A key difference between the NMR and X-ray structures is the absence in the NMR structure of a C-capping motif on helix-3 observed in the X-ray structure. This motif is spontaneously adopted in the simulation. (B) The C-capping motif involves three backbone hydrogen bonds (black: K70-G74, red: K71-G74, green: K70-L75) that are formed at ~8 ns and stable throughout the remainder of the simulation.

Dictionary of secondary structural prediction (DSSP) [128] analysis was employed to characterize the secondary structure in the simulation in order to facilitate comparisons with the X-ray and NMR structures (Figure 2-6). In the simulations, helix-1 spans the same 8 residues as found in the X-ray structure (D44 to F51), while the NMR structure contained only a five residue helix from D44-K48. Thus the simulation significantly extends the length of the first helix, in agreement with the X-ray structure. Overall, the locations of the sequence of helices 2 and 3 are similar in the NMR and X-ray structures, although helix-2 is one residue shorter in the NMR structure, (residues R55 to F58 for the NMR vs. R55 to A59 for the X-ray). In the simulation, helix-2 appears consistent with both experimental structures; full α-helical content is sampled for residues 55 through 58, with partial helical content (~50%) observed for A59. This may indicate that the C-terminus of the longer helix in the X-ray structure frays at the temperature of the NMR experiment. In both the NMR and X-ray structures, the α–helical content is the same for helix-3 (L63-K72). The simulations sample the same helix, with residue K73 sampling a partial population of helical structures. As noted above, the simulation spontaneously adopts a C-capping motif for this helix that is present in the X-ray structure. Overall, the alpha helical structural content of the structures in the simulation is in much better agreement with the X-ray structure, particularly in helix-1.

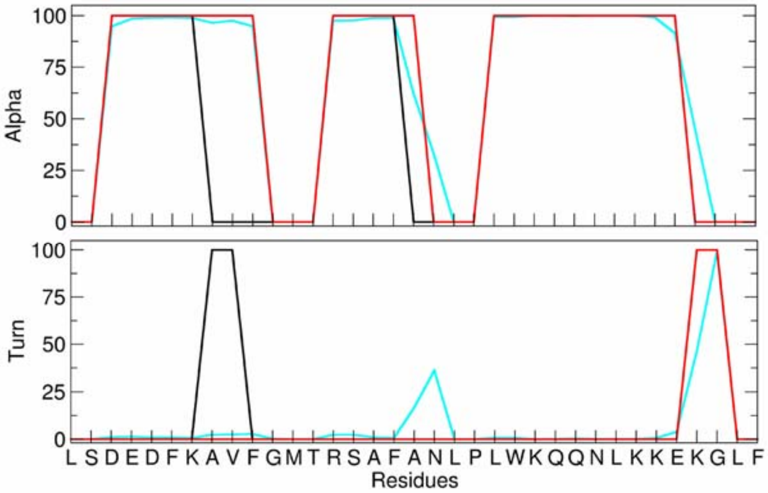


Figure 2-6. DSSP analysis of the NMR (black), X-ray (red); and simulation (cyan) structures of HP36. (A) Alpha helical content per residue. (B) Turn content per residue. Overall, helix-2 and helix-3 are nearly the same length in the X-ray and NMR structures, but helix-1 is 3 residues longer in the X-ray structure than in the NMR structure. The alpha helical content of the MD simulation is in very good agreement with the X-ray structure even though it was initiated from the NMR structure.

All-atom cluster analysis was used to generate a representative simulation structure using the last 5 nanoseconds of the trajectory. This structure has backbone and all-atom RMSD values relative to the X-ray structure of 1.5 and 2.7 Å (residues 42 to 75), while the RMSD values relative to the initial NMR structure were higher (2.3 Å (backbone) and 3.3 Å (all-atom)). Figure 2-7 shows all three structures after best-fit of the backbone from residues 42 to 62 (helices 1 and 2). Notably, the X-ray and simulation structure have a very similar spatial arrangement of their phenylalanine cores. In contrast to the X-ray and simulation structures, the NMR structure has F51 shifted more into the core. Thus, the backbone and core of the protein in the solution simulation possesses structural features that are much more similar to the X-ray structure despite being initiated from the NMR structure.

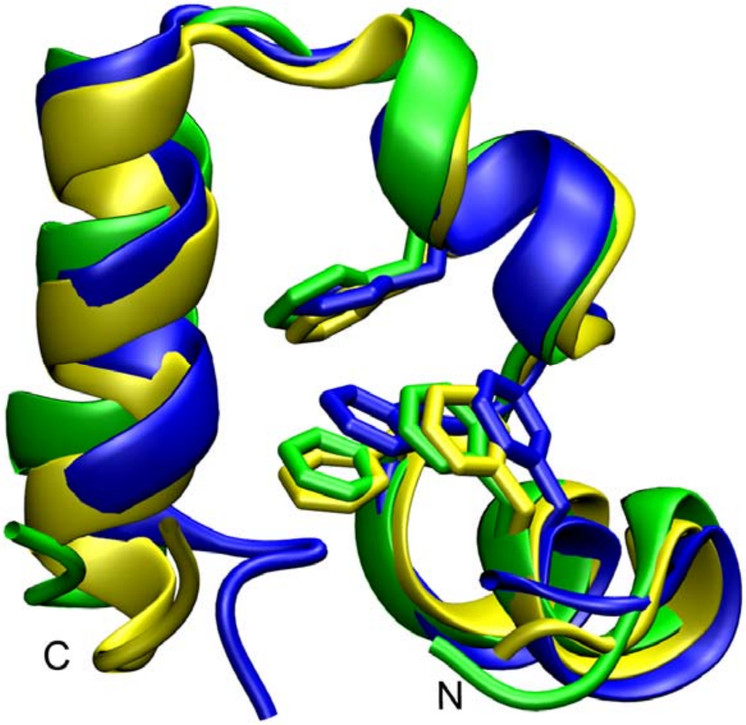


Figure 2-7. Comparison of backbone and core packing in the simulation (green), NMR (blue), and the X-ray (yellow) structure, highlighting the differences in the core packing. A best fit alignment was performed on residues L42 through P62. The packing of the phenylalanine core in the structure from the simulation structure is in much better agreement with the X-ray structure than with the NMR structure.

2.3.2 Structural similarities to the NMR Family

Given the diversity among the family of structures solved using the NMR data, it is reasonable to expect that some of them may be more similar than others to the X-ray structure. Figure 2-8 shows the backbone RMSD as compared to the X-ray, simulation and NMR average structures for each structure in the NMR family. Overall, the individual NMR structures are all more similar to the NMR average than to the X-ray structure (average RMSD values of 1.7 and 2.4 Å respectively). The RMSD of the three individual helices demonstrate similar differences. However, some of the individual members of the NMR family are similar to the X-ray and simulation structures, especially in helix 1. According the DSSP, 7 out of the 29 members of the NMR family sample alpha helical conformations at V50 (data not shown) which is outside of the helical region in the average NMR structure. This suggests that extension of helix-1 beyond the range seen in the average structure remains consistent with the NMR family. However, the overall backbone of the X-ray and the simulation structure differs from all of the structures in the NMR family (Figure 2-8).

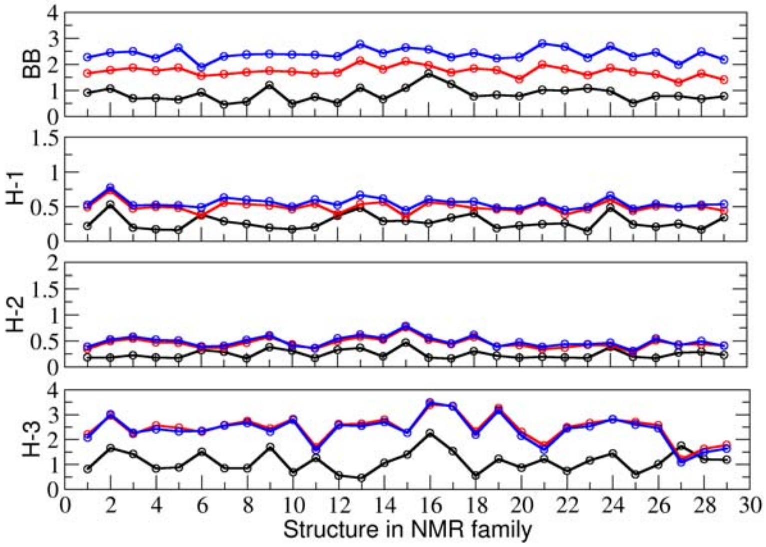


Figure 2-8. RMSD values of each structure in the NMR family, for backbone (BB – residues 42 to 75), helix-1 (H-1 - residues 3 to 9), helix-2 (H-2 – residues 54 to 59), and helix-3 (H-3–residues 62 to 76). Each calculation was performed using the NMR (black), X-ray (red) and the simulation (Blue) as a reference structure. Overall, the NMR structures are more similar to the NMR average structure than to the simulation or X-ray structure, although individual secondary structure elements are in good agreement with the X-ray structure for a few of the NMR models (e.g. H1 for structure #6).

2.3.3 Specific sidechain interactions

There are several specific sidechain interactions which differ in the NMR and X-ray structures. In the X-ray structure, R55 forms a van der Waals interaction with F47 and an interhelical sidechain-sidechain hydrogen bond with D44 (D44-Oδ1 and R55-Nε); both interactions are absent in the NMR structure. In Figure 2-9A & B, the simulation structure was aligned with the X-ray structure to highlight the similarities in the interaction of those particular sidechains. Since the simulation structure is a single snapshot, we also investigated the behavior of these contacts as a function of time during the MD run, observing fluctuations in both cases (Figure 2-10). In both the X-ray and the simulation structure, the H-bond distance between D44 and R55 is 2.7 Å, in contrast with the much longer distance of 7.9 Å in the NMR structure. This specific contact also samples a range of distances from 6.7 Å to 11.6 Å in the family of NMR structures (Figure 2-11). During the simulation, this hydrogen bond is broken and re-formed multiple times, suggesting that a reasonable description of the equilibrium distance distribution has been sampled (Figure 2-10A**)**. We used histogram analysis to calculate the potential mean force (PMF) for the pair to quantify the stability of the contact in the native state. While two free energy minima are located at the hydrogen bonding distance, two other local minima at 5.0 and 7.0 Å have relative energies of less than 0.6 kcal/mol compared to the contact pair (Figure 2-10B). Thus breaking this contact is expected to be a readily accessible thermal fluctuation. The stability of the contact between F47 and R55 was evaluated by measuring the distance from the Cγ of F47 to the Nε of R55 (Figure 2-10C). This distance had comparable values in the simulation and X-ray structures (4.7 and 4.3 Å, respectively), while a much longer distance of 6.3 Å is observed in the average

NMR structure. Only 2 structures in the entire NMR family sample a contact distance of less than 5.5 Å (Figure 2-11). In contrast with the D44/R55 pair, the PMF for formation of the F47/R55 contact shows only a single minimum at 5.5 Å (Figure 2-10D**)**. Overall, this suggests that R55 has a much more stable interaction with F47 than the salt bridge that it forms with D44.

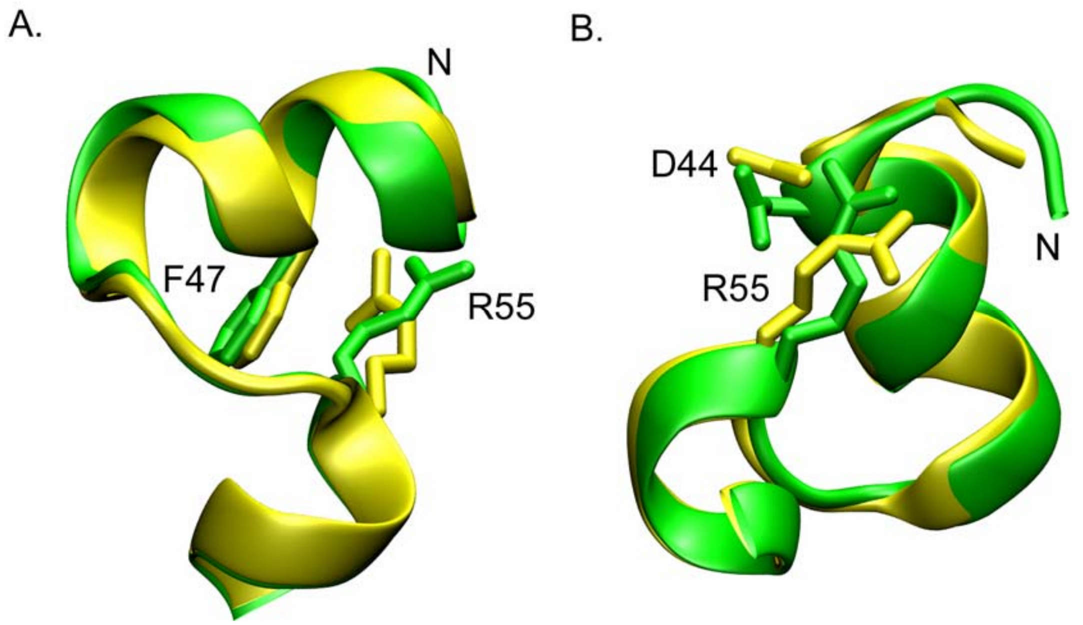


Figure 2-9. Comparison of selected sidechain interactions in the simulation structure (green) and the X-ray structure (yellow). A best fit alignment was performed on residues L42 through P62. In the simulation structure, R55 is 4.7 Å away from the base of the phenylalanine ring (A) and 2.7 Å away from the Oδ1 of D44 (B). This suggests that both contacts may play a role in the stability of the protein.

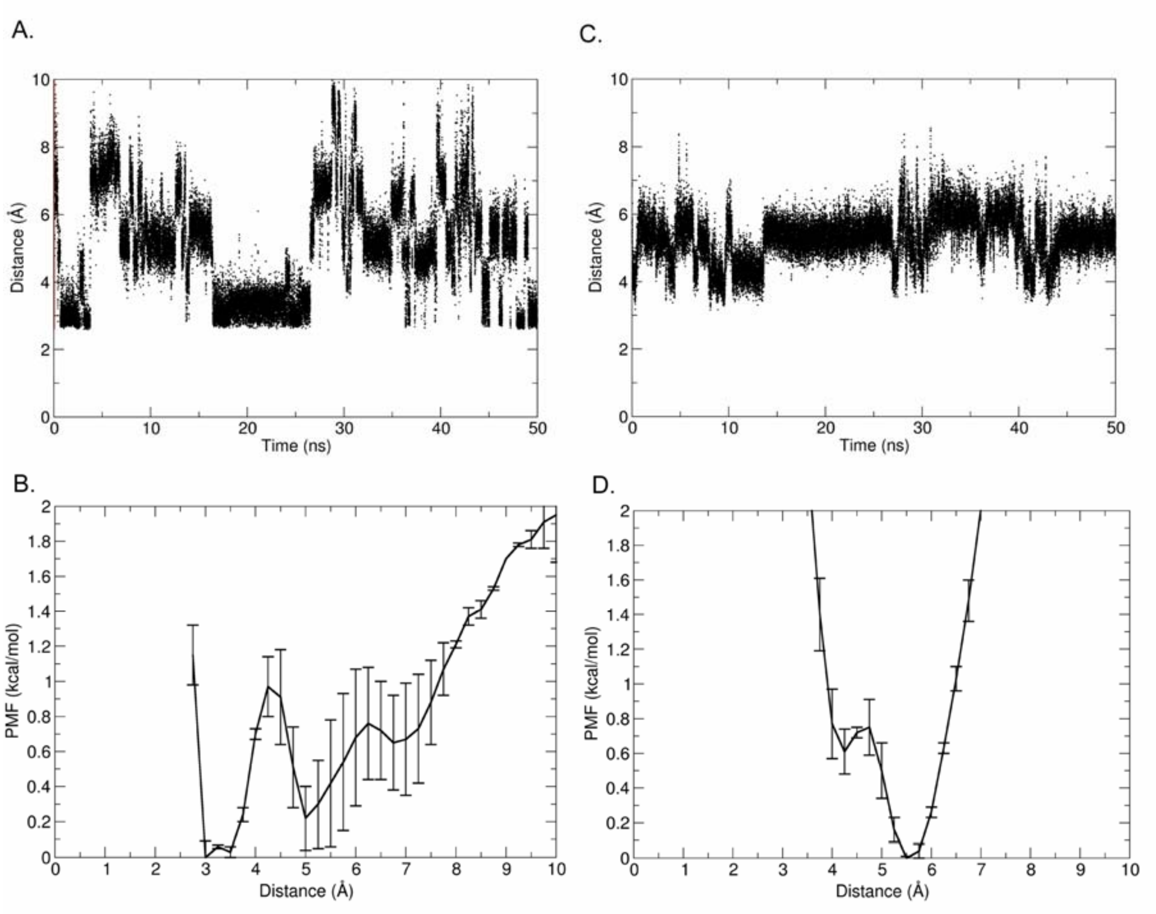


Figure 2-10. Time evolution (A and C) and PMFs (B and D) of specific contact distances involving R55 and D44 and R55 and F47. The distance between R55 and D44 fluctuates throughout the trajectory and shows multiple shallow free energy minima. In contrast, the distance measuring the contact between R55 and F47 is stable during the entire trajectory, with a single free energy minimum at 5.5 Ǻ. The results indicate that the R55/F47 contact is the more stable of these 2 residue pairs.

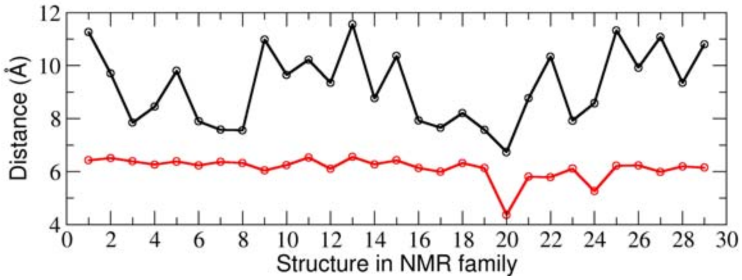


Figure 2-11. Specific contact distances involving R55 and D44 (black) and R55 and F47 (red) for each structure in the NMR family. The contact involving R55 and D44 ranged from 6.7 Å to 11.6 Å and the contact involving R55 and F47 ranged from 4.4 Å to 6.5 Å. However, the later contact has 2 structures with distances between the two residues less than 5.5 Å. For the most part, the NMR family does not contain the hydrogen bond and the van der Waals contact seen in both the X-ray and simulation structure.

2.3.4 Simulations of the X-ray structure

Figure 2-12 shows the backbone RMSD *versus* time and RMSD distributions calculated during the simulation starting from the X-ray structure. The RMSD is shown relative to the X-ray, NMR and simulation (from NMR) structures. After equilibration, the simulation samples backbone conformations (S43-L75) with an average RMSD relative to the X-ray structure of 1.5 Å and remains quite stable through the 30 ns duration. Overall, there is a preference to adopt structures comparable to the simulation structure discussed above rather than the NMR structure (RMSD compared to the the simulation-equilibrated NMR structure is 1.5 Å below the RMSD to the original NMR structure). Individual helices demonstrate comparable preferences for the X-ray and simulation structures (data not shown). Hence, the simulations starting from the NMR and X-ray structures both converge to a common simulation structure that is much closer to the X-ray structure than the NMR structure.

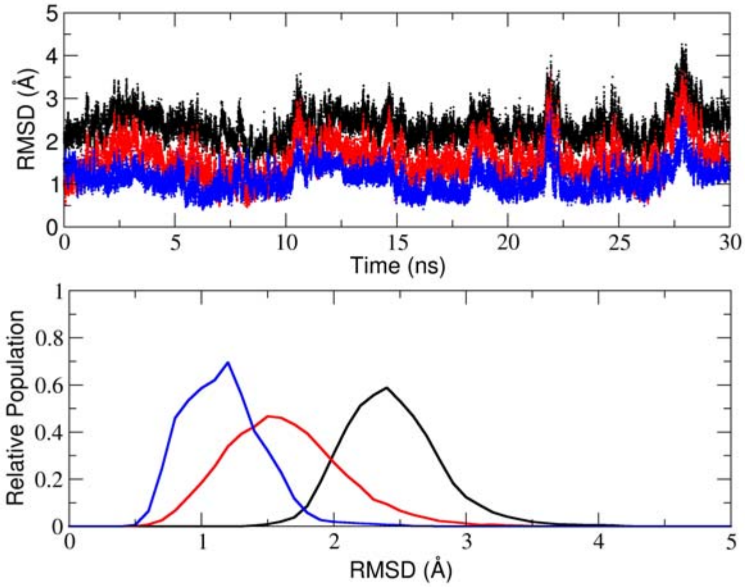


Figure 2-12. Time evolution and distributions of the heavy atom backbone RMSD of the residues 43 to 75 (heavy atoms of the backbone) during the simulation of the X-ray structure. Each calculation was performed using the NMR (black), X-ray (red) and the Simulation (Blue) as a reference structure. The simulation shows a preference for adopting a backbone structure similar to simulation and X-ray structure.

2.3.5 Experimental investigation of the putative sidechain interactions

While simulations can provide a detailed view of molecule structure and dynamics, many approximations are involved, necessitating validation through experimentation. A set of single mutants and double mutants were prepared in order to probe the putative sidechain interactions involving D44/R55 and F47/R55. D44 was mutated to Asn, F47 to Leu and R55 to Met. Thermal unfolding experiments were performed for wildtype HP36 (WT HP36) and for each of the mutants at pH 5.0 (Figure 2-13A, Table 2-1). The WT HP36 has a transition midpoint (T_m) of 73.0 ^{o}C, while all the variants show a lower melting temperature. The T_m of D44N, F47L, R55M, D44NR55M and F47LR55M are 57.8 ^{o}C, 45.6 ^{o}C, 67.3 ^{o}C, 55.4 ^{o}C and 35.3 ^{o}C, respectively. From the thermal unfolding curves, at 25 °C, 22 % of the population of F47L and 40 % of the population of F47LR55M are unfolded.

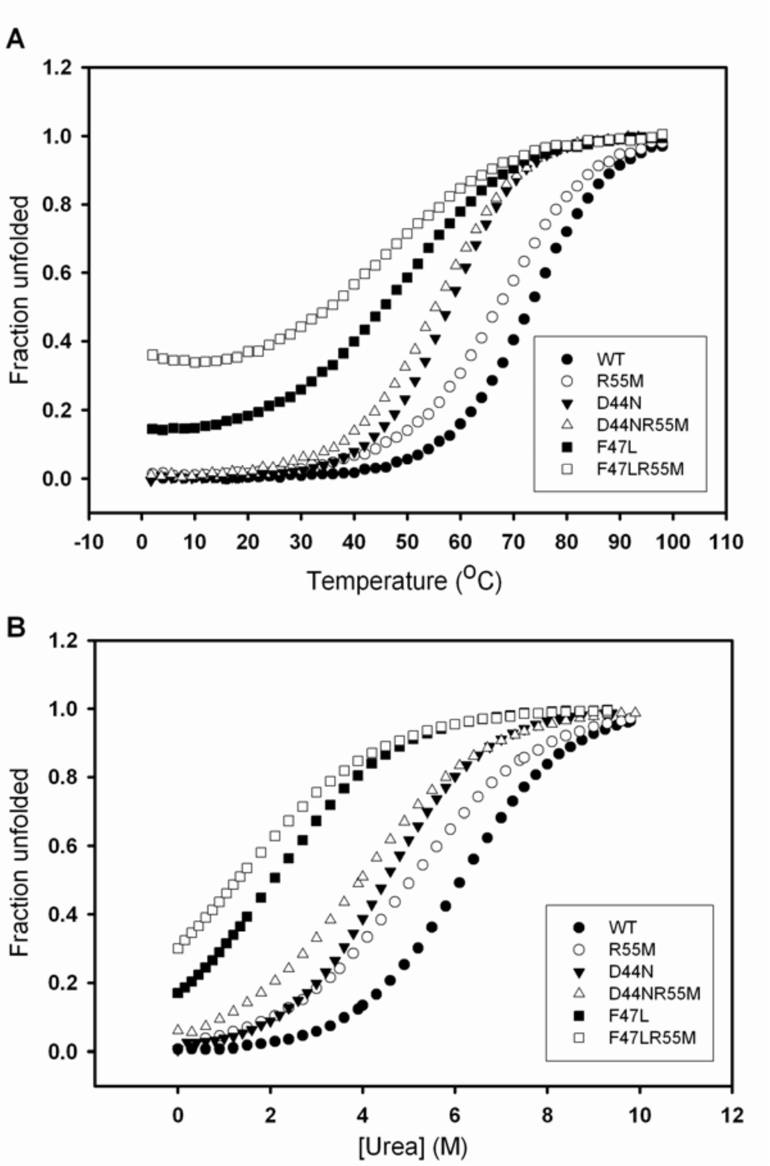


Figure 2-13. (A) Thermal unfolding curves for WT HP36 and its mutants; (B) Urea unfolding curves for WT HP36 and its mutants. Closed circles (●) represents the WT HP36, R55M is represented by open circles (○), D44N by closed triangle (▼), D44NR55M by open triangle (Δ), F47L by closed squares (■) and F47lR55M by open squares (□).

Table 2-1. Summary of equilibrium stability measurements for WT HP36 and its mutanin 10 mM sodium acetate, 150 mM sodium chloride, pH 5.0 at 25^{o}C.

Protein	T_m (^{o}C)	$\Delta H^{o}(T_m)$ (kcal mol^{-1})	$\Delta G^{o}{}_{U}(H_2O)$ (kcal mol^{-1})	M (kcal mol^{-1} M^{-1})
WT HP36	73.0	31.8	3.22	-0.52
D44N	57.8	32.1	2.48	-0.55
F47L	45.6	15.8	0.52[a]	-0.45[b]
R55M	67.3	26.3	2.19	-0.43
D44NR55M	55.4	27.4	1.74	-0.44
F47LR55M	35.3	9.8	0.19-0.28[c]	N/A

[a] $\Delta G^{o}{}_{U}(H_2O)$ of F47L is extrapolated from urea denaturation in different TMAO concentrations; [b] m is the average value of the m from urea denaturation in different TMAO concentrations; [c] $\Delta G^{o}{}_{U}(H_2O)$ of F47LR55M is calculated from Gibbs-Helmholtz equation using $\Delta C^{o}{}_{P}$ values ranging from 0.30-0.70 kcal mol^{-1} K^{-1}.

Urea denaturation experiments were also carried out in 10 mM sodium acetate and 150 mM sodium chloride at 25 ^{o}C to determine the free energy of unfolding. The estimated free energy for unfolding ($\Delta G^{o}{}_{U}$)was 3.22 kcal mol^{-1} for WT HP36, 2.48 kcal mol^{-1} for D44N, 2.19 kcal mol^{-1} for R55M and 1.74 kcal mol^{-1} for D44NR55M (Figure 2-13b, Table 2-1). The F47L and F47LR55M mutants were so unstable that the native baseline was not observed (Figure 2-13B) and the unfolding free energy could not be accurately measured by urea denaturation. Thermal and urea denaturation experiments showed that F47L and F47LR55M are partially unfolded in the absence of urea at 25 °C. Previous studies have shown that TMAO can stabilize partially or completely unfolded proteins [129]. Therefore, the combination of urea denaturation and TMAO stabilization can be utilized to estimate the stability of marginally stable proteins. In order to determine the unfolding free energy of F47L and F47LR55M, we performed urea denaturation experiments in increasing TMAO concentrations. For F47L, the titration curves show good pre- and post-transitions in different TMAO concentrations (Figure 2-14A). With increasing TMAO concentrations, the urea denaturation curves shifted to higher urea concentrations. The free energy of unfolding at each TMAO concentration was measured: $\Delta G^{o}{}_{U}$ ranges from 1.27 kcal mol^{-1} in 1.62 M TMAO to 1.67 kcal mol^{-1} in 2.50 M TMAO (Table 2-2). Mello and coworkers [129] have shown that the free energy of unfolding depends linearly on TMAO concentration. The extrapolated $\Delta G^{o}{}_{U}$ of F47L at 0 M TMAO was estimated to be 0.52 kcal mol^{-1} at 25^{o}C (Figure 2-14), which is in reasonable agreement with the value estimated from the thermal unfolding curve.

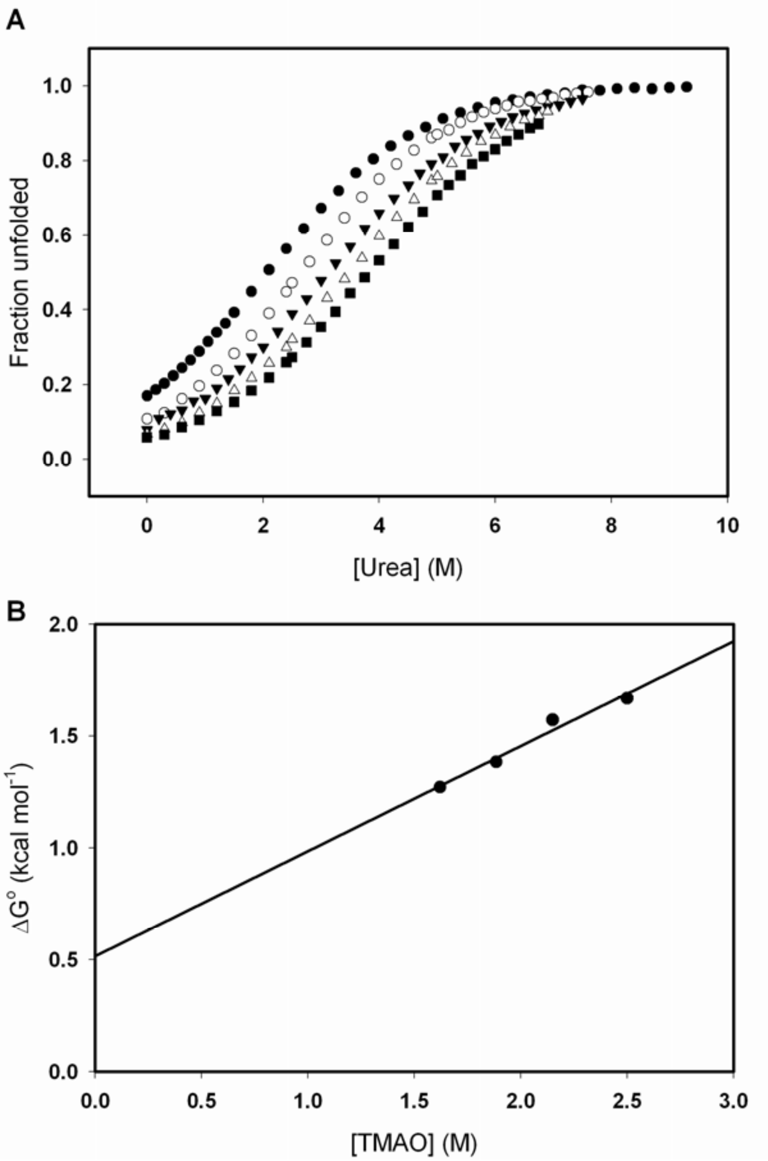


Figure 2-14. (A) Unfolding transitions of the F47L mutant in a mixed urea/ TMAO cosolvent monitored by circular dichroism at 222 nm. Urea denaturation in increasing TMAO concentrations (from left to right at 0, 1.60, 1.88, 2.15, 2.50M TMAO); (B) Dependence of unfolding free energy on TMAO concentration for the F47L mutant; parameters are obtained by fitting urea unfolding curves in different TMA004F concentrations. The straight line is the result of linear regression to each parameter.

Table 2-2. Summary of urea denaturation measurements in different TMAO concentrations for F47L in 10 mM sodium acetate, 150 mM sodium chloride, pH 5.0 at 25°C.

[TMAO] (M)	$\Delta G^{o}{}_{U}(TMAO)$ (kcal mol^{-1})	M (kcal mol^{-1} M^{-1})
1.62	1.27	-0.48
1.88	1.38	-0.44
2.15	1.57	-0.45
2.50	1.67	-0.44

Unfortunately the same strategy could not be applied to the F47LR55M double mutant. High TMAO concentrations are necessary to stabilize the protein to detect the pre-transition, but comparatively high urea concentrations are needed to observe the post-transition. Therefore, it is very difficult to find conditions where full unfolding curves could be measured. Thus, we extrapolated from the thermal unfolding data using the Gibbs-Helmholtz equation:

$$\Delta G^{o}{}_{U}(T) = \Delta H^{o}(T_m)\left(1 - \frac{T}{T_m}\right) - \Delta C^{o}{}_{p}\left[T_m - T + T\ln\left(\frac{T}{T_m}\right)\right]$$

Equation 2-2. 1 Gibbs-Helmholtz equation. Gibbs-Helmholtz equation. ΔG is the free energy, ΔH_m is the enthalpy of melting, T and T_m are the temperature and melting temperature, respectively, and ΔC_P is the heat capacity at constant pressure.

This calculation requires knowledge of the heat capacity change, $\Delta C^{o}{}_{P}$. HP36 is small, resulting in a very broad differential scanning calorimetry (DSC) transition, which makes it very difficult to calculate the heat capacity accurately by DSC. From the literature, the value of $\Delta C^{o}{}_{P}$ of unfolding is expected to be about 0.012 kcal mol^{-1} K^{-1} per residue of protein [130]. To a first approximation, the $\Delta C^{o}{}_{P}$ for HP36 can be calculated to be 0.43 kcal mol^{-1} K^{-1}. Another small 41-residue helical protein, the peripheral subunit-binding domain, has a $\Delta C^{o}{}_{P}$ value of 0.43 kcal mol^{-1} K^{-1} [131], suggesting that the estimate for HP36 is reasonable. In order to check whether the value of $\Delta C^{o}{}_{P}$ significantly affects the results, we use heat capacities ranging from 0.30 to 0.70 kcal mol^{-1} K^{-1} to calculate the $\Delta G^{o}{}_{U}$. The F47LR55M has a measured T_m of 35.3 ^{o}C and $\Delta H^{0}(T_m)$ of 9.5 kcal mol^{-1}, and the resulting calculated $\Delta G^{o}{}_{U}$ of F47LR55M at 25 ^{o}C ranged from 0.19 to 0.28 kcal mol^{-1} depending on the value of $\Delta C^{o}{}_{P}$ used (Table 2-3). The value of $\Delta G^{o}{}_{U}$ estimated from the Gibbs-Helmholtz equation is in good agreement with the fraction unfolded determined directly from the fit to the thermal melt.

Table 2-3. Calculation of $\Delta G^{o}{}_{U}(H_2O)$ of F47LR55M from Gibbs-Helmholtz equation using different $\Delta C^{o}{}_{P}$ values.

$\Delta C^{o}{}_{P}$ (kcal mol^{-1} K^{-1})	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70
$\Delta G^{o}{}_{U}(H_2O)$ (kcal mol^{-1})	0.28	0.25	0.24	0.23	0.22	0.214	0.206	0.197	0.19
$\Delta\Delta G^{o}{}_{coupling}$ [a] (kcal mol^{-1})	0.79	0.76	0.75	0.74	0.73	0.724	0.716	0.707	0.70

[a] The coupling free energy for WT HP36, F47L, R55M and F47LR55M double mutant cycle using different calculated the values of $\Delta G^{o}{}_{U}(H_2O)$ of F47LR55M.

The coupling free energy between the D44 or F47 sidechain and the R55 sidechain, $\Delta\Delta G^{o}{}_{coupling}$, was calculated using equation (2), where $\Delta G^{o}{}_{WT}$ and $\Delta G^{o}{}_{R55M}$ are the free energies of unfolding for wild type protein and R55M single mutant; and $\Delta G^{o}{}_{Single}$ represent D44N or F47L single mutants; and $\Delta G^{o}{}_{Double}$ represents the D44NR55M or F47LR55M double mutants.

$$\Delta\Delta G^{o}{}_{coupling} = (\Delta G^{o}{}_{WT} - \Delta G^{o}{}_{Double}) - [(\Delta G^{o}{}_{Single} - \Delta G^{o}{}_{Double}) + (\Delta G^{o}{}_{R55M} - \Delta G^{o}{}_{Double})]$$

Equation 2-3. Coupling free energy equation.

The relationship can be rearranged to a simpler form:

$$\Delta\Delta G^{o}{}_{coupling} = \Delta G^{o}{}_{WT} - \Delta G^{o}{}_{Single} - \Delta G^{o}{}_{R55M} + \Delta G^{o}{}_{Double} \qquad ($$

Equation 2-4. Simplified form of coupling free energy equation.

Using the $\Delta G^{o}{}_{U}$ values (Table 2-1) measured from experiments, the coupling free energy between the D44 sidechain and the R55 sidechain was close to zero (0.29±0.20 kcal mol^{-1}). In contrast, the coupling free energy between the F47 sidechain and R55 sidechain ranged from 0.70±0.20 to 0.79±0.20 kcal mol^{-1}. The different estimates arise from using different $\Delta C^{o}{}_{P}$ values to calculate $\Delta G^{o}{}_{Double}$. The analysis shows that there is a non-zero coupling between the F47 and R55.

2.4 Discussion

The explicit water MD simulation starting from the NMR structure showed a clear preference to sample structures with much greater similarity to the X-ray structure, as indicated by RMSD values, DSSP analysis, packing of the phenyalanine core, formation of a C-capping motif on helix-3 and adopting of specific contacts between side chains. Double mutant cycle experiments were performed and demonstrated clear coupling between F47 and R55. It is apparent that these residues are not interacting in the NMR structure but appear to do so in the X-ray structure. Based on free energies calculated from MD simulations and obtained experimentally through double mutant cycles, the F47/R55 contact appears to be a stronger interaction than the proposed salt bridge between D44 and R55. Hence, the van der Waals interaction seen in the X-ray structure appears to play an important role in stabilizing the solution structure of HP36. The coupling free energy between the D44 sidechain and the R55 sidechain is small, only 0.29±0.20 kcal mol^{-1}. The F47 sidechain to R55 sidechain coupling free energy is

0.70±0.20 to 0.79±0.20 kcal mol^{-1}. These results are consistent with the simulation results showing that F47/R55 interacts strongly and that the stability of the D44/R55 pair is lower than the thermal energy.

Previous studies by Frank et al. [97] have shown the importance of each phenylalanine in stabilizing the core of the protein. Interestingly, the F47LR55M double mutant is even less stable than these single Phe mutants, which suggests that the sidechain of R55 also plays a key role in stabilizing the structure. It is likely that the optimum packing of the three phenylalanines in the core is enhanced by R55 because it helps to shield the core with its long sidechain and also forms a cation-pi interaction. Cation-pi interactions can be important for folding and thermostability of various proteins and protein ligand systems [132-134]. For the single mutant, R55M, the Tm dropped 6 ^{o}C in thermal stability, showing that more than just a bulky sidechain it is required at position 55. In the majority of villin sequences, Lys is found as a conservative mutation in place of R55 [135]. This suggests that the charge is important for stabilizing the structure, but as the simulations and double mutant experiments indicate, the importance of this charge at position 55 does not arise from formation of an ion pair with D44 as observed in the crystal structure. It is worth noting, however that in the X-ray structure D44 appears to be involved in a network of interactions including a hydrogen bond to the backbone carbonyl L42. Backbone sidechain interactions cannot be probed by double mutant cycle analysis.

These simulations are models and as with any model there are limitations, especially in the interpretation of results. Realistic, detailed simulations come at a high computational cost that must often be balanced against the need for obtaining extensive conformational sampling. Computational models continue to improve; the Amber and CHARMm force fields have been used extensively enough to identify weaknesses [37, 45] such as overstabilization of secondary structure elements. In the simulations that we report here, this type of systematic error might contribute to the extension of alpha-helices that we observed, although we specifically addressed secondary structure bias in the development of the parameter set that was used for all of the present simulations [45]. It has also been noted that the use of PME to calculate long range electrostatics imposes long-range periodicity that that may result in artifacts from a crystal-like environment [136-138]. In the present case, simulations with two alternate treatments of long-range interactions (including undesirable direct truncation) provided essentially the same conclusion, that the simulations adopt a structure in better agreement with the crystal structure than with the NMR structure (Figure 2-15). Thus there is no evidence that the present results are an artifact of PME.

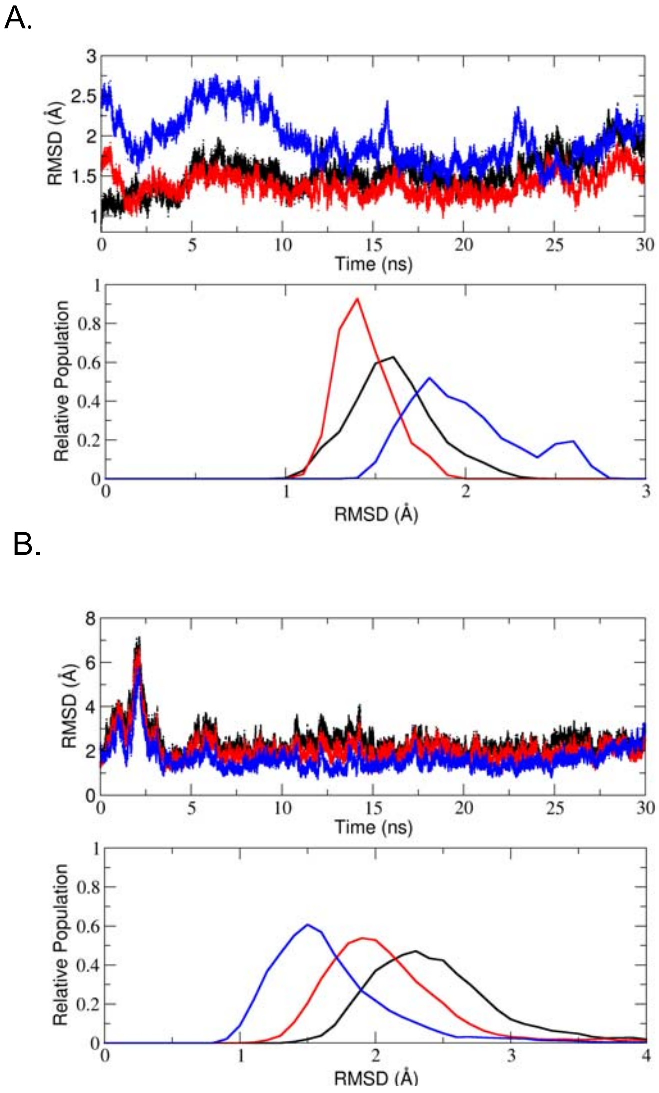


Figure 2-15. Time evolution and distributions of the heavy atom backbone RMSD of the (A) residues 42 to 75 (heavy atoms of the backbone) during the simulation of the NMR structure using (A) a cutoff of 12 Å with no smoothing and (B) IPS for the electrostatic treatment. Each calculation was performed using the NMR (black), X-ray (red) and the Simulation (Blue) as a reference structure. In both cases, the resulting structures are in better agreement with the X-ray structure than NMR structure (red histograms are shifted to lower RMSD values as compared to black).

Previous work by van der Spoel and Lindahl [109] reported a series of simulations of the NMR structure of HP36. These authors noted a modest degree of sensitivity to force field, water models, and protonation states. In their simulations, they noted larger structure fluctuations in the region connecting helices 1 and 2 as compared to the rest of the molecule. This observation is consistent with our results, which indicate this linker as one region in which the protein spontaneously adopts a conformation in the simulations more consistent with the crystal structure. At that time, there was no way for the van der Spoel and Lindahl to determine whether this larger fluctuation resulted from a conversion toward the crystal structure which was reported two years later. Importantly, van der Spoel and Lindahl also noted the importance of taking into account the pH of the experiment when running simulations of HP36. Upon protonation of the glutamic acid side chains in the starting structure, the resulting simulation displayed a greater correlation to the chemical shift and j-coupling results which were originally measured at a pH of 3.7. This observation further suggests that one must be cautious in the quantitative comparison of simulations at neutral pH to experimental data obtained at low pH.

In summary, the results from our simulations and experiments show that the recently published X-ray structure is a more accurate representation of the structure in solution at neutral pH than the NMR structure at low pH. Importantly, the simulations also indicated that a salt bridge between R55 and D44 observed in the low-temperature crystal structure was thermally unstable, in contrast to the stable interaction between R55 and F47 in the simulation. Experimental double mutant analysis confirmed that the interaction free energy of the salt bridge was small, and that the F47-R55 pair likely plays an important role in stabilizing the protein via a cation-pi interaction. The analysis presented here shows how the combination of molecular dynamics simulations and experimental measurements can be used to develop a better understanding of the structural properties of proteins in solution.

3. The Unfolded State of the Villin Headpiece Helical Subdomain: Computational Studies of the Role of Locally Stabilized Structure

Abstract

The 36 residue villin headpiece helical subdomain (HP36) is one of the fastest cooperatively folding proteins, folding on the microsecond timescale. HP36's simple three helix topology, fast folding and small size have made it an attractive model system for computational and experimental studies of protein folding. Recent experimental studies have explored the denatured state of HP36 using fragment analysis coupled with relatively low resolution spectroscopic techniques. These studies have shown that there is apparently only a small tendency to form locally stabilized secondary structure. In this study, we complement the experimental studies by using Replica Exchange Molecular Dynamics (REMD) with explicit solvent to investigate the structural features of these peptide models of unfolded HP36. To ensure convergence, two sets of simulations for each fragment were performed with different initial structures, and simulations were continued until these generated very similar final ensembles. These simulations reveal low populations of native-like structure and early folding events which cannot be resolved by experiment. For each fragment, calculated J-coupling constants and helical propensities are in good agreement with experimental trends. HP-1, corresponding to residues 41 to 53 and including the first α-helix, contains the highest helical population. HP-3, corresponding to residues 62 through 75 and including the third α-helix, contains a

small population of helical turn residing at the N-terminus while HP-2, corresponding to residues 52 through 61 and including the second α-helix, formed little to no structure in isolation. Overall, HP-1 was the only fragment to adopt a native-like conformation, but the low population suggests that formation of significant structure only occurs after formation of specific tertiary interactions.

Acknowledgments

The material presented in this chapter has been published (Lauren Wickstrom*, Asim Okur, Kun Song, Viktor Hornak, Daniel P. Raleigh, Carlos L. Simmerling 2006 *J. Mol. Biol.* 103, 1094-107). This chapter contains direct excerpts from the manuscript (presented with minor modifications) written by Lauren Wickstrom with suggestions and revisions from Professor Daniel P. Raleigh and Professor Carlos Simmerling. Dr. Asim Okur contributed simulation data for Ala10. Dr. Kun Song provided scripts for the J-coupling calculations. We also thank Roberto Gomperts and the SGI Engineering group for the computer time. The authors thank Ms. Yuefeng Tang for valuable discussions concerning HP36.

3.1 Introduction

Structure in the unfolded state may play a significant role in the rapid folding of proteins by limiting the conformational search. Recent experimental work from the Fersht and Oas labs has highlighted the role of unfolded state structure in the rapid folding of helical proteins [139, 140]. Other work has suggested the importance of polyproline II conformations (PPII) in the unfolded ensemble [141-144]. Unfortunately, direct experimental studies of the unfolded state are difficult because the most relevant unfolded state is that which is in equilibrium with the folded state under native conditions. The normal high cooperativity of folding together with the free energy balance of folding means that this state is only sparsely populated at equilibrium. Experimental difficulties also arise because of the short lifetime of the denatured state in refolding experiments. Consequently, indirect methods have to be employed but many approaches fail to examine the unfolded state under equilibrium conditions [18, 145-148].

One indirect approach to studying the denatured state under native conditions is to analyze peptide fragments corresponding to elements of secondary structure derived from the whole protein. Peptide fragment analysis provides the local propensity for secondary structure formation and a potential glimpse at structures that may form in the early stages of folding. Such locally stabilized structure can play a role in rapid folding by limiting the early stages of the conformational search. For example, one popular model for folding, the diffusion collision model, postulates a critical role for locally stabilized microdomains. The determination of these structural details are potentially of great importance for the folding of helical proteins [24, 25, 140].

The villin headpiece helical subdomain (HP36), the C-terminal portion of the villin headpiece, is the shortest naturally occurring sequence which has been shown to fold cooperatively (Figure 3-1). Its rapid folding, small size and simple topology of three helices have made this domain an extremely popular system for computational and theoretical studies [88, 93, 94, 101, 102, 108, 110, 112, 113, 116, 149]. These studies have largely focused on generation of the correct native topology and have not investigated the details of the folding mechanism or the role of residual structure in the unfolded state.

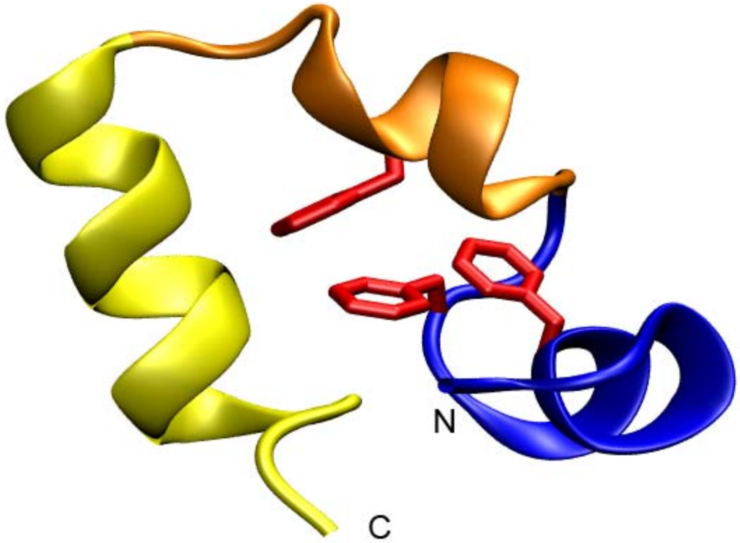


Figure 3-1. Structure of the villin headpiece subdomain (pdb code 1vii). HP-1 is in blue, HP-2 is in orange and HP-3 is in yellow. Phe47, Phe51, and Phe58 are shown in red. The N and C-termini are labeled.

Recent experimental work has explored the possibility of residual structure in the unfolded state of HP36 [26]. In that work, a set of fragments corresponding to the three α-helices were studied as well as a larger fragment containing the first two helices. None of the individual peptide fragments showed significant helical content as judged by Circular Dichroism (CD) spectroscopy. However, two of the helices in HP36 are quite small in fragments 1 (HP-1) and 2 (HP-2) and the CD spectra of short helices are not well understood [150-152]. Thus, it is not clear how best to interpret CD studies of the small helices that may be formed by these fragments, particularly when NMR studies hint at some tendency to form non-random structure. The experimentally measured ^{1}H-alpha chemical shift deviations from random coil (approximately 0.25 ppm upfield) observed for the HP36 fragments suggest either sparsely populated helical conformations or ring current effects in HP-1 and fragment 3 (HP-3). These potential ambiguities are due to the limitations that exist with experimental methods.

Simulations can help overcome these limitations and allow for the observation of structure at the level of individual molecules instead of the ensemble averages typically provided by experiments. Computational studies can also provide atomic level detail concerning specific interactions that may not be readily available from experimental studies of rapidly interconverting ensembles. This enhances the understanding of mechanistic details of protein folding and structure. However, conformational sampling remains a significant obstacle in molecular dynamics (MD) simulations. Generation of precise populations at equilibrium is difficult due to the protein folding time scale being much longer than is typically accessible to simulation. Hence, the study of partially populated states through simulation is hampered by poor convergence.

Replica exchange molecular dynamics (REMD) is an enhanced sampling technique [62-64] that can help overcome the limited time scale issues, yet it remains a challenging task to obtain converged results, particularly for large systems. Many different studies have used REMD to study folding in smaller model peptide systems [55, 66, 153-156] however studies of unfolded state structure have been more limited [157].

In this paper, we analyze the same set of short fragments of HP-36 that were studied experimentally in an attempt to clarify the extent of locally stabilized secondary structure. We conducted REMD simulations using both an implicit and explicit solvent model for each fragment. The results demonstrate that explicit solvent is the more accurate approach for studying these small peptides. We find HP-1 possesses the most native-like structure of the three fragments, and the potential role that locally stabilized structure may play a role in the fast folding of HP36 is discussed.

3.2 Methods

Three fragments were built from the sequence of HP36: M41–F76 (MLSDEDF KAVFGMTRSAFANLPLWKQQNLKKEKGLF). HP-1 (M41-M53) corresponds to the N-terminal helix of HP36. HP-2 (G52-L61) contains the second helix and HP-3 (P62-L75) contains the C-terminal helix. HP-1 has a free N-terminus while HP-2 and HP-3 have acetylated N-termini. All C-termini were amidated. These sequences and termini correspond to those used in the experimental studies [26]. All sidechains for Asp, Glu, Lys, and Arg were charged during the simulation. Ala_{10} was acetylated and amidated at the N and C termini respectively. All simulations were performed in Amber version 8 [39] and used the Amber ff99 force field [120, 121], with modifications to eliminate α-helical bias[45] . These parameters have been provided elsewhere [158], are denoted “ff99SB” in Amber version 9 and are available for download from the Amber web site (amber.scripps.edu). SHAKE [122] was used to constrain bonds involving hydrogen. The time step was 2 fs. Temperatures were maintained using weak Berendsen coupling [123]. All simulations were initiated from both an extended and a helical structure, with comparison of the two runs providing a lower bound for the uncertainty in resulting data.

3.2.1 Explicit solvent simulations

Explicit water simulations were performed in a truncated octahedral box using periodic boundary conditions and particle mesh Ewald [125] (PME) to calculate long range electrostatic interactions. The structures were equilibrated at 300 K for 50 ps with harmonic restraints on solute atoms, followed by minimization with gradually reduced positional restraints. The restraints were reduced from 5 kcal/mol*Å to 1 kcal/mol*Å to 0.5 kcal/mol*Å. After minimization, three 5 ps MD simulations with the same gradually reduced restraints at constant pressure (1 atm) and temperature (300 K) to generate starting structures. To improve sampling, we use REMD as implemented in Amber 8. Each system is represented by multiple simulations which are coupled to baths at different temperatures. Periodically, an exchange of replicas is attempted using a Metropolis-type criterion. The target exchange acceptance ratio for all simulations was approximately 20 % between temperatures ranging from 260 – 580 K. Exchanges between neighboring temperatures was attempted every 1 ps.

The initially extended and helical HP-1 fragments were surrounded by 1387 TIP3P waters and 1029 TIP3P waters respectively. The extended structure used 38 replicas ranging from 259 to 567 K (259, 264, 270, 276, 282, 288, 294, 300, 306, 313, 320, 327, 334, 341, 348, 355, 363, 371, 379, 387, 395, 404, 412, 421, 430, 439, 449, 458, 468, 478, 488, 499, 509, 520, 532, 543, 555, and 567 K) while the folded structure used 34 replicas ranging from 262 to 531 K (257, 262, 268, 274, 280, 287, 293, 300, 306, 313, 320, 327, 335, 342, 350, 358, 366, 374, 382, 391, 400, 408, 418, 427, 437, 446, 456, 467, 477, 488, 499, 510, 521 and 533 K). Each simulation was run for 42 ns. The extended and helical HP-2 fragments were surrounded by 1092 waters and 849 waters respectively. The extended structure used 34 replicas ranging from 269 to 548 K (269, 275, 281, 287, 294, 300, 307, 313, 320, 327, 334, 341, 349, 356, 364, 372, 380, 388, 397, 406, 414, 423, 433, 442, 452, 461, 471, 482, 492, 503, 514, 525, 537, and 548 K) while the helical structure used 30 replicas ranging from 265 to 543 K(269, 275, 281, 287, 294, 300, 307, 313, 320, 327, 334, 341, 349, 356, 364, 372, 380, 388, 397, 405, 414, 423, 432, 442, 452, 462, 472, 482, 492, 503, 514, 525, 536 and 548 K). Both were simulated for 32 ns. The extended and partially helical HP-3 fragments were surrounded by 1250 waters. Both were simulated with 40 replicas ranging from 266 to 578 K (266, 272, 277, 283, 288, 294, 300, 306, 312, 319, 325, 331, 338, 345, 352, 359, 366, 373, 381, 389, 396, 404, 413, 421, 429, 438, 447, 456, 465, 474, 484, 494, 503, 514, 524, 534, 545, 556, 567, and 579 K) for 41 ns. The first 10 ns of each simulation were discarded to reduce bias caused by the initial structure.

The Ala_{10} peptide in an α-helical and an extended conformation was solvated using 983 TIP3P water molecules for a total of 3058 atoms. 40 replicas were used at temperatures ranging from 267K to 571K (267, 272, 278, 283, 289, 294, 300, 306, 312, 318 324, 331, 337, 344, 351, 358, 365, 372, 379, 387, 394, 402, 410, 418, 426, 435, 443, 452, 461, 470, 479, 489, 498, 508, 518, 528, 539, 549, 560 and 571 K)., which were optimized to give a uniform exchange acceptance ratio of ~30%.

3.2.2 Implicit solvent simulations

The implicit solvent effects were calculated using the Generalized Born continuum model [51] using pairwise descreening [159] with mbondi radii [160]. Simulations were initiated with the same two initial conformation ensembles as were used for the explicit solvent REMD calculations. Both initial structures were minimized, followed by a brief equilibration. The same force field and target exchange ratios in the explicit solvent were implemented in the GB runs. The first 5 ns was discarded to remove initial structure bias in each run.

The HP-1 simulations used 8 replicas from 272 to 539 K (272, 300, 331, 365, 402, 443, 489, and 539 K.), for approximately 50 ns each. Simulations of HP-2 ran for 40 ns with 8 replicas between 268 and 587 K (268, 300, 336, 375, 420, 470, 525 and 587 K) and HP-3 simulations ran for 60 ns using 10 replicas between 277 and 571 K (277, 300, 325, 352, 382, 414, 449, 486, 527, and 571 K). For $Ala_{10,}$ 8 replicas were used at temperatures ranging from 270 to 571 K (270, 300, 334, 372, 414, 461, 513, 571 K). Exchanges were attempted every 1 ps and the REMD simulation was run for 50000 exchanges (50 ns).

3.2.3 Data analysis

Cluster analysis was performed on each simulation with MOIL-View [127], using backbone RMSD as a similarity criterion with average linkage. Clusters were formed with a bottom-up approach using a similarity cutoff of 2.5 Å; the populations of the resulting clusters for each fragment are discussed in the main text. The portions of the backbone were selected according to the region of the fragment where the HP36 native helix was located. DSSP analysis of the fragments confirmed that these regions were the most structured portions of the fragments. Conformational families were defined based on the combined set of structures from all simulations of the fragment (both initial structures and both solvent models), and the populations of each family were then calculated for the ensemble obtained from each simulation. Comparison of the populations of each structure type in the ensembles obtained from independent initial structures was used as a convergence metric.

DSSP analysis [128] and calculation of distances, RMSDs, and radius of gyration were done using the ptraj module in Amber. Distances between sidechains were calculated using heavy atoms of the charged atoms. Free energy histograms were calculated at 300 K according to equation 3-1.

ΔG_i = -RT $\ln(N_i/N_0)$

Equation 3.1. Relative free energy calculated with multidimensional histogram analysis. ΔG_i is the relative free energy bin I, N_i is the population of a particular histogram bin along the reaction, and N_0 is the most populated bin. R and T are the gas constant and temperature.

Lifson-Roig (LR) analysis was implemented to calculate the probability of forming helices of a particular length. Backbone torsion (φ/Ψ) angles were used to evaluate whether a residues was helical or non-helical. Using the Garcia and Sanbonmatsu definition [80], a residue was considered helical if $\varphi = 60 \pm 30$ and $\Psi = 47 \pm 30$. Helical populations were calculated using equation 3-2. We note that this provides absolute helical content and formation of several short helices in single structure is possible.

$Hp = (H_l/N)$

Equation 3.2. Equation to calculate helical populations. Hp is the population of a helix at a particular length, H_l is the amount of that helix of a particular length L and N is the total number of structures in the ensemble.

J-coupling constants were calculated using a version of the Karplus equation (Equation 3-3) previously employed for analysis of small peptides:

$^3J(H_N,H_\alpha) = A\cos^2(\varphi - 60) + B\cos(\varphi - 60) + C$

Equation 3.3. Karplus equation for the calculation of $^3J(H_N,H_\alpha)$ scalar couplings. A, B, and C are constants.

where A = 6.51, B = -1.76, and C = 1.60 [161]. All calculations were performed on the combined data set including simulations started with the extended and folded starting structures.

3.3 Results

3.3.1 Measuring convergence of the REMD simulations

When the goal of a simulation study is simply to identify a low-energy conformation, it is typically unnecessary to generate a Boltzmann-weighted ensemble with conformations populated according to relative energies. However, when one wishes to use these results to gain insight into the relationship of the unfolded state to the folding process, it is necessary to obtain a reliable and quantitative estimation of the extent to which any residual structure is present in the unfolded state, with well defined limits.

In the present case, we investigate the role of locally stabilized structure in the unfolded state ensemble of the villin headpiece protein. In order to ensure that the simulations are robust and that the populations that we observe are precise, all of the simulations were repeated with two different initial starting structures. For each fragment, one simulation was initiated from a fully extended structure while another was started from a helical structure. Since it has been demonstrated that different properties converge at different rates [162], we use as our convergence metric the fractional populations of alternate conformations which are the main focus of our analysis. As described in the Methods, conformation families are defined based on the combined set of structures from all simulations, and the populations of each conformation family are then calculated for the ensemble obtained from each of the two alternate simulations. For all fragments, the absolute populations sampled in the two independent runs in the TIP3P [49] explicit water model demonstrated a high correlation (r = 0.994 (HP-1), 0.993 (HP-2), 0.863 (HP-3)), indicating that the populations of each conformational basin are independent of initial coordinates (Figure 3-2). For the simulations in implicit solvent, there was a high correlation between runs 1 and 2 for HP-1 (0.999) and HP-3 (0.996), but HP-2 showed poor convergence (0.279). The explicit solvent simulations clearly provide more data precision for all three fragments; this may arise from slower convergence in the implicit model due to high barriers to conformational change arising from salt bridges that are too strong in the implicit model [56].

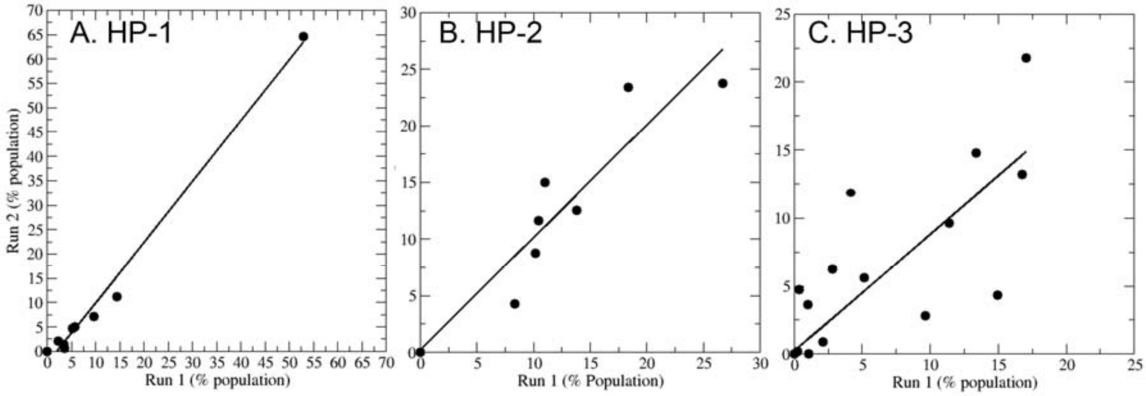


Figure 3-2. Comparison of cluster populations between two independent REMD simulations at 300 K in explicit solvent. (A) HP-1 (slope = 1.24, r = 0.994) (B) HP-2 (slope = 0.923, r = 0.993) (C) HP-3 (slope = 0.859, r = 0.86). Cluster families were defined based on a combined data set of all trajectories of the fragment. The high correlation between the populations for each fragment suggests that the REMD simulations are well converged and that the populations of individual structure types are reliable.

3.3.2 Comparison of structural ensembles obtained using explicit and implicit solvent models

Accurate treatment of solvation is essential for meaningful simulation of biological molecules in solution. Explicit solvent models can be highly effective, particularly when water has non-bulk properties and interacts directly with the solute [124]. Implicit models such as the semi-analytical generalized Born model (GB) [51] are attractive because they are computationally less expensive and can converge more rapidly than simulations in explicit water due to lack of solvent viscosity. While GB has been widely used for protein folding studies, others have reported poor results including secondary structural bias and ion pairing issues [53-55]. We note, however, that many variants of the GB model exist and relatively few studies comparing their performance for protein folding have been published [163-165].

In this study, both the GB and TIP3P solvent models share the same largest cluster in HP-1, indicating the same most preferred structure (Figure 3-3). Nevertheless they differ significantly in the contribution of this conformation to the overall ensemble (90 % in GB vs. 25 % in TIP3P). Overall the populations of the conformation families for HP-1 show a poor correlation between TIP3P and GB ensembles (r = 0.67 and a slope of 0.26). This arises primarily from a ~1.5 kcal/mol overstabilization of the α-helical region of the Ramachandran region in the GB simulations as compared to TIP3P. The other fragments showed similarly poor agreement between the solvent models. We therefore focus on results obtained using TIP3P and discuss GB data only to illustrate specific shortcomings observed with that model in the discussions.

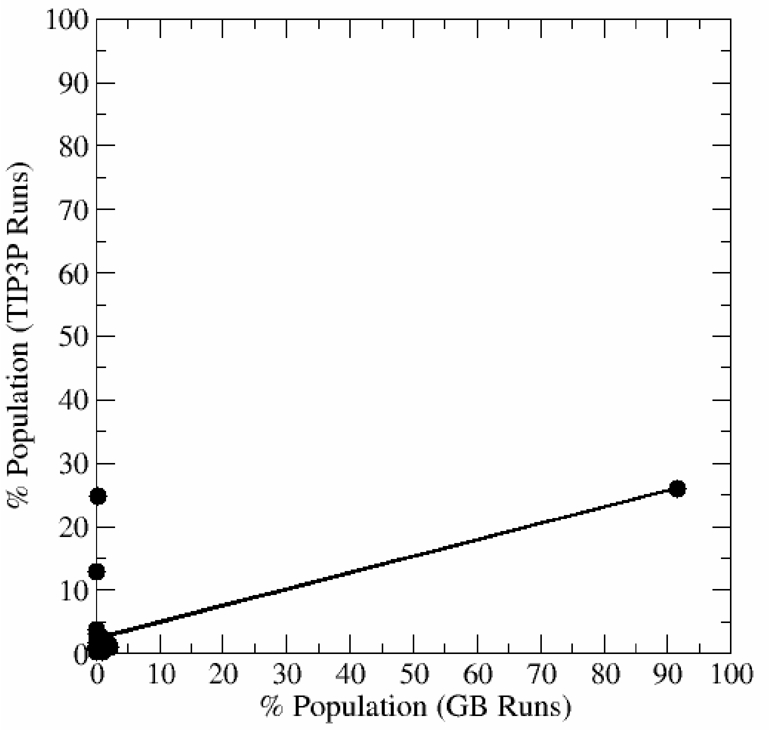


Figure 3-3. Comparison of cluster populations for HP-1, similar to Figure 3-2 but comparing populations between simulations in explicit and GB implicit solvent. The correlation between the populations sampled in the different solvent models was quite low(slope = 0.26 r = 0.67), suggesting that the implicit solvent model samples the structure families in very different amounts than sampled in explicit solvent. The correlations for the other fragments is similarly poor.

3.3.3 Summary of data analysis approaches

Data from the two independent sets of simulations were combined for analysis of each fragment as described in Methods. Differences between the data sets provide a low bound to the actual uncertainty. Analysis included calculation of root mean square deviation (RMSD) of structures as compared to the conformation of the fragment in intact HP36, radius of gyration (Rg), torsion angles, secondary structure types using the dictionary of secondary structure prediction (DSSP) algorithm [128], Lifson-Roig (LR) analysis of the distribution of helix lengths [166, 167], and conformational cluster analysis.

3.3.4 Simulations of Ala$_{10}$

A simulation of Ala$_{10}$ in explicit solvent was run as a control. This particular system was chosen due to a similar size to the three fragments and the fact that Ala has the smallest sidechain (other than Gly). Comparison of the Ala$_{10}$ structural ensemble to those from the fragments provides insight into the role of specific side-chain interactions in the HP36 fragments.

The central residue of Ala$_{10}$ samples local backbone conformations which are located in all 4 basins of the Ramachandran plot: α helix, PPII, anti-parallel beta sheet and the left handed helical basin (Figure 3-4). Furthermore, DSSP analysis of Ala$_{10}$ resulted in an average α-helical content of only 1.3 %. Since Ala10 lacks any intrinsic structure, we conclude that any helical content observed in the fragments are the result of the sequence. In addition, the lack of significant helical content in Ala$_{10}$ suggests that the force field employed does not suffer from over-stabilization of α-helices, as has been reported for previous versions of the Amber force field [37, 168].

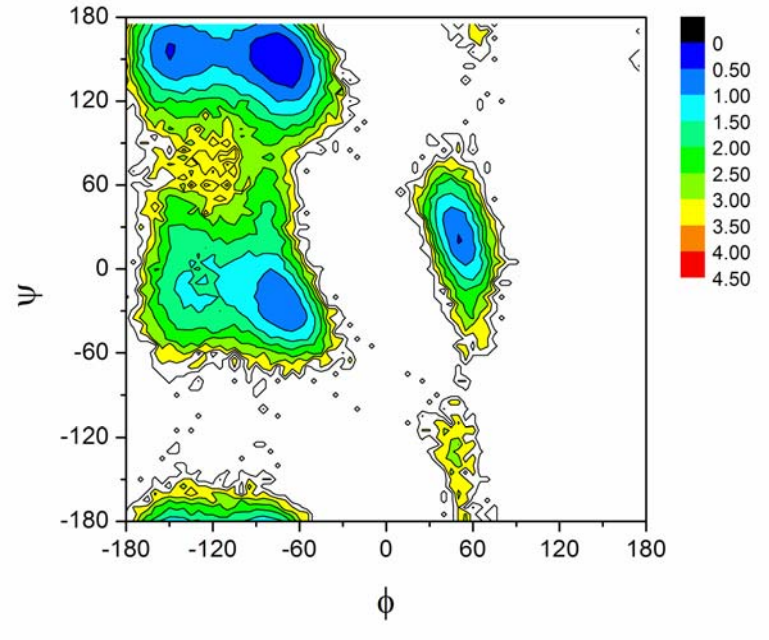


Figure 3-4. Free energy profile at 300 K for a central Ala residue (Ala5) in Ala_{10} from REMD using the TIP3P water model. Basins corresponding to the major secondary structure types are all similar in free energy for models using explicit solvent, suggesting that there is no secondary structural bias in the force field. The units are in kcal/mol.

3.3.5 Conformational preferences for fragment HP-1

HP-1 **(**M*41*LSDEDFKAVFGM*53*) contains the sequence that forms the N-terminal helix of HP36, (between D44 and K48), located near the center of the fragment, with 3 backbone α-helical hydrogen bonds in intact HP36. The fragment includes a stretch of predominantly basic and acidic residues and several residues that can perform helix N-capping. We first present properties of the entire ensemble of structures sampled, followed by more detailed discussion of specific preferred conformations.

Figure 3-5 shows the free energy landscape at 300 K for HP-1 along coordinates of the Rg and the RMSD to the backbone of the NMR structure of HP36. The global free energy minimum has a low RMSD (1.0 Å) and an Rg value of 7.0 Å, similar to HP-1 in the native state of HP36 (Rg = 6.7 Å). The broad shape of the minimum with respect to

Rg indicates that the compactness of the fragment is variable, while the relatively narrow shape with respect to RMSD suggests that the structures remain quite native-like during these fluctuations. Overall, the consistency between the RMSD and native like Rg in this landscape indicate that at least half (discussed later in cluster analysis) of the ensemble of structures populated by HP-1 have a high similarity to the conformation adopted by the fragment in intact HP36, suggesting that residual native structure in this region is fairly well populated in the unfolded state of HP36.

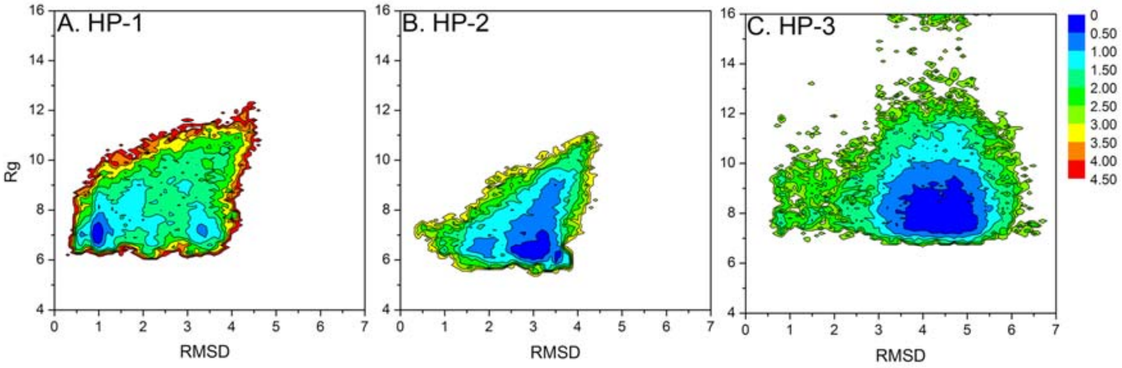


Figure 3-5. Free energy landscapes of three fragments at 300 K from REMD explicit water simulations. (A) Rg vs RMSD (backbone 43 – 49) to the native HP36 structure for HP-1 (B) Rg vs RMSD (backbone 54-59) to the native HP36 for HP-2 (C) Rg vs. RMSD (backbone 64-70) to the native HP36 for HP-3. While all three fragments remain compact as in HP36, only HP-1 has a free energy minimum located at a low RMSD. The other two fragments occupy minima with higher RMSDs and more broad minima than HP-1. The units are kcal/mol.

Figure 3-6A shows the Ramachandran free energy surfaces at 300 K for three residues in the HP-1 ensemble, selected from the terminal regions and the center of the fragment. As expected, the termini are more flexible, with the N-terminal Leu42 predominantly sampling the PPII and helical basins, while the C-terminal V50 samples shallow, broad minima in the PPII, and α-helical basins (approximately 1.0 – 1.5 kcal/mol in depth). In contrast, the central D46 is stabilized in the helical region by approximately 2.0 kcal/mol relative to other basin. D46 adopts local backbone angles which correspond to the ones seen in the native state of HP36.

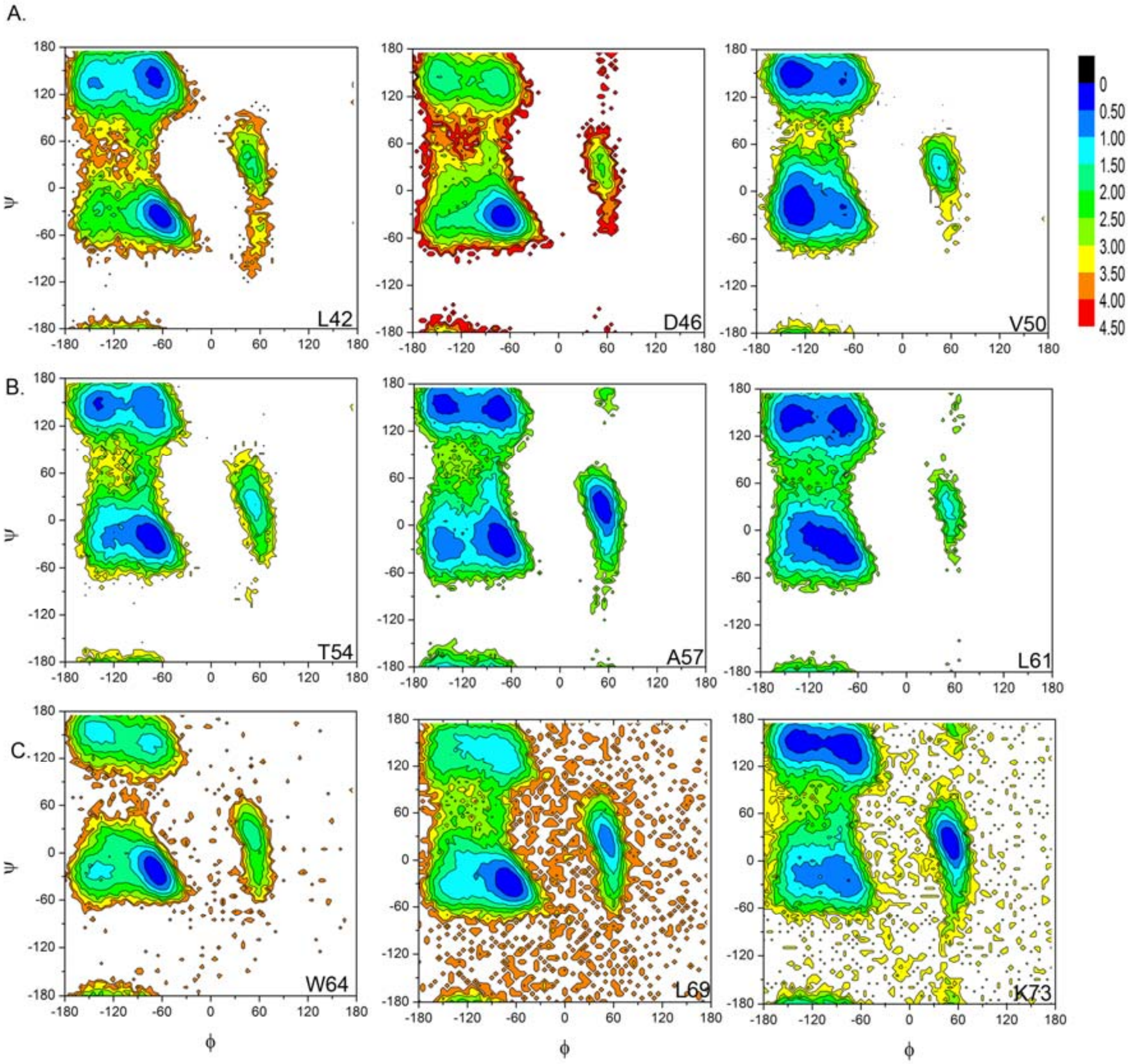


Figure 3-6. Free energy profiles of residues at 300 K from REMD explicit water simulations in (A) HP-1 (L42, D46, V50); (B) HP-2 (T54, A57, L61) and (C) HP-3 (W64, L69, K73). HP-1 and HP-3 have global free energy minima in the helical region for residues D46 in HP-1 and W64 in HP-3. All three residues in HP-2 occupy shallow local minima, which suggest significant conformational flexibility. The units are kcal/mol.

LR and DSSP analysis were employed to evaluate the relationship between local and long range helical structure in HP-1 (Figure 3-7A & 3-8A). The most probable helix length is 3 to 4 residues, but longer helices of 5-6 residues are still found in 5 - 10% of the structures in the ensemble. This is similar to the length of the first helix in the native HP36. DSSP shows that the α-helical content of the ensemble increases rapidly from the N-terminus toward the center of the fragment, then drops sharply towards the C-terminus (average α-helicity of 35 % and 8 % at N and C-termini respectively). This is consistent with the trend seen in the free energy surfaces which show that the relative depth of the helical basin increases towards the center of the sequence and decreases toward the C-terminus. Importantly, the high α-helical propensity observed in the center of the fragment (35 %) is in the same region as the α-helix in the native HP36 structure (D44 – K48). Other types of secondary structure are less prevalent and the propensities are fairly consistent across the sequence, in contrast to the increase in helical content in the middle of the fragment.

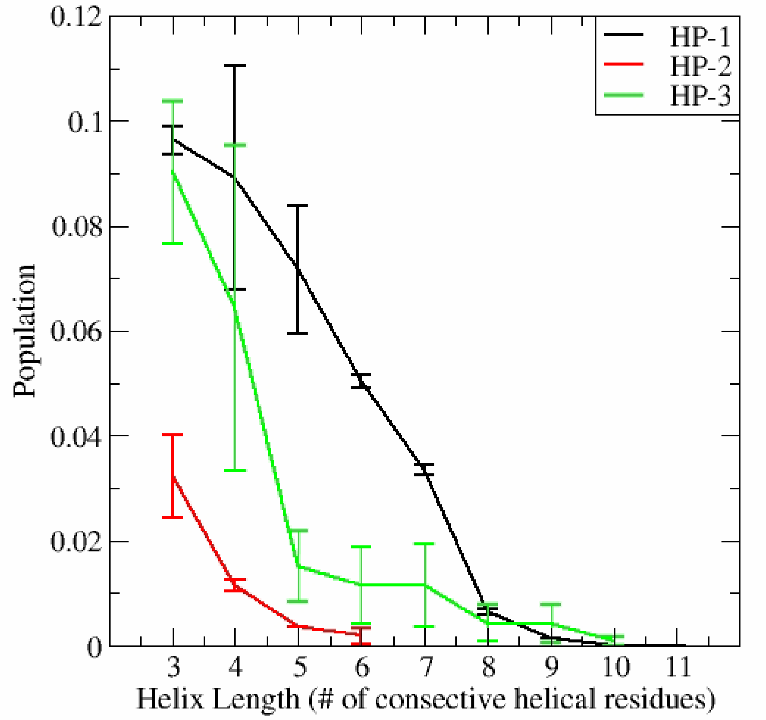


Figure 3-7. Probability of finding helices of a particular length in the ensembles at 300 K for (A) HP-1(black), (B) HP-2 (red), and (C) HP-3 (green). The populations show that HP-2 has little probability (2-4%) of forming even a single turn of helix with 3 residues. HP-3 does form very short helices, but only HP-1 shows significant population of helices that are 5-7 residues in length.

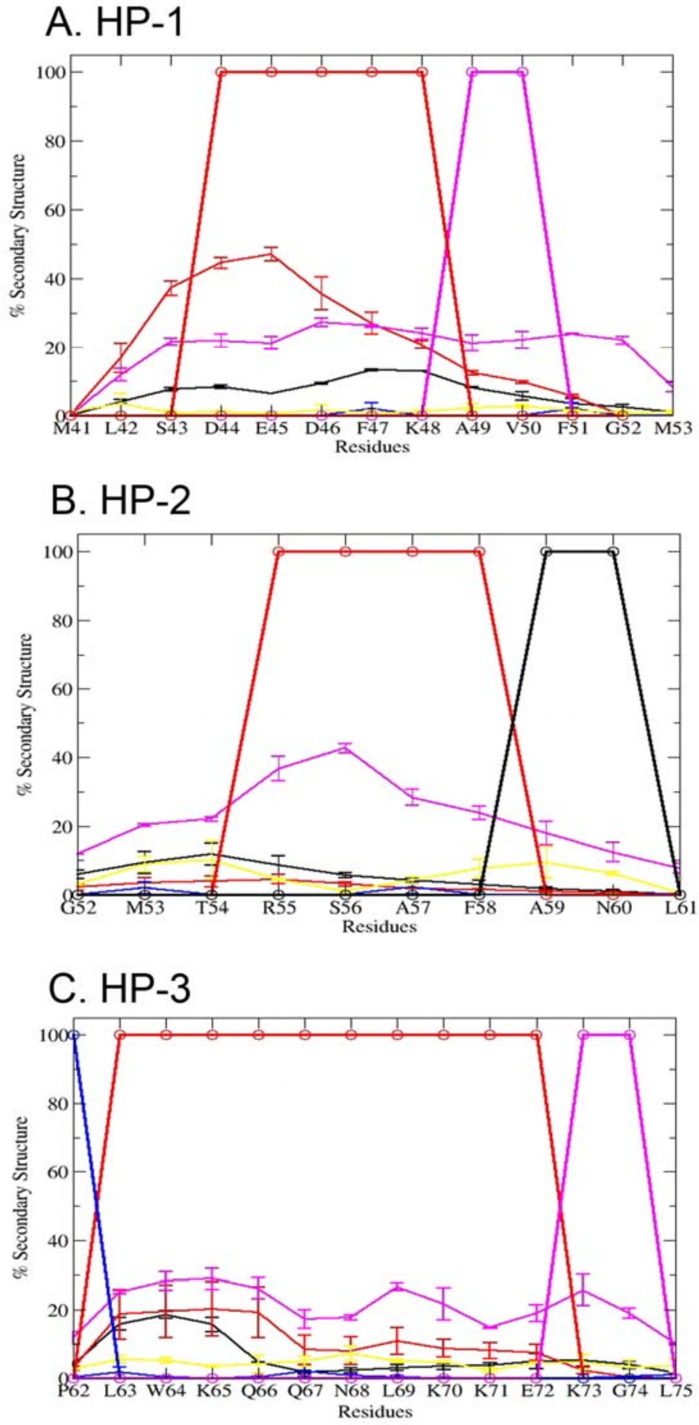


Figure 3-8. DSSP Analysis of (A) HP-1; (B) HP-2; (C) HP-3 at 300 K. (Black = 3_{10}; Red = α-helix; Green = π-helix; Blue = parallel β-strand; Yellow = antiparallel β-strand; Magenta = turn). Circles represent the NMR structure. HP-1 contains the most helical content of all three fragments.

Cluster analysis was performed on the backbone of HP-1 from residues 43 to 49 (Figure 3-9). The two most populated conformational clusters, which account for approximately 59 and 13 % of the structures respectively, contain the most α-helical structure. The first cluster contains structures with modest population of α-helix from S43 to A49, while cluster 2 features a shorter α-helical turn from L42 to E45. The population of cluster 3 (9 % of the ensemble) is made up of helix-turn-helix structures displaying a helical content of 20 and 15 % at the N and C-termini respectively, however there is greater variation in the structures within the entire cluster than in the first two clusters. These structural populations are in agreement with the DSSP results which showed that the N-terminus is more helical than the C-terminus. It is impressive to note that the most populated cluster forms the full helical conformation seen in HP36.

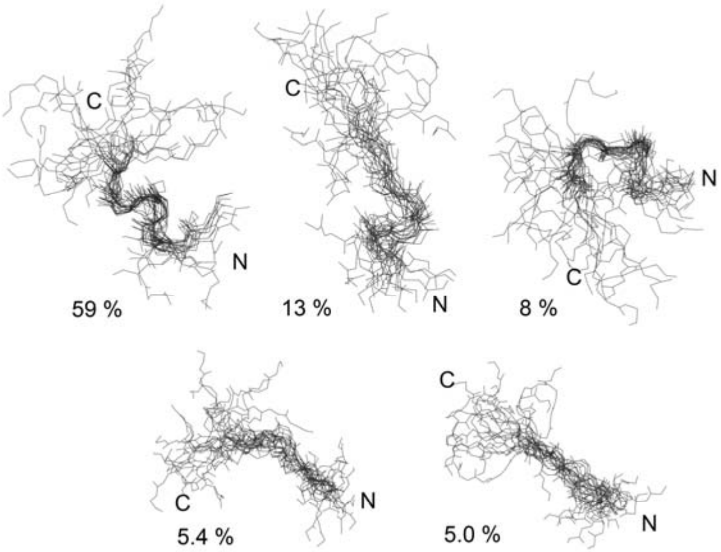


Figure 3-9. Representative structures for the five most populated clusters of HP-1. (A) 1st cluster (59 ± 6 %) (B) 2nd cluster (13 ± 2 %) (C) 3rd cluster (8 ± 1 %) (D) 4th cluster (5.4 ± .4 %) (E) 5th cluster (5.0 ± .3 %). The most populated cluster contains helical content similar to helix 1 in the native HP36.

3.3.6 HP-2 shows no significant residual structure

HP-2 (G*52*MTRSAFANL*61*) contains the shortest helix in HP36, comprising residues R55 through F58 and containing only one α-helical hydrogen bond between T54-F58. This fragment contains N-capping residues that might favor helix formation, however the small number of residues in HP-2 may not be sufficient to stabilize the native helical conformation. Figure 3-5B shows the free energy surface at 300 K in explicit solvent along the Rg and RMSD to the HP36 structure. The distribution of the minima is much broader and not as well defined as was observed for HP-1. This landscape shows one broad, shallow global free energy minimum centered at RMSD of 2.5 Ǻ and Rg of ~ 6.0 Ǻ. Similar to HP-1, this Rg value is comparable to that seen for this compact region in native HP36 (6.0 Å). In contrast to our HP-1 results, however, the structures sampled for HP-2 are quite different than the structure in the corresponding helical region in native HP36, averaging an RMSD of 2.9 Å for the residues comprising the native helix. These observations show that HP-2 is less structured than HP-1 and does not form an appreciable amount of the structure seen in HP36.

Figure 3-6B shows Ramachandran free energy surfaces for three residues of HP-2 at 300K. Thr54 and Ala57 sample all 4 major basins, while Leu61 occupies a broader basin in the α-helical region, and to a lesser extent the PPII and anti-parallel β-strand regions. DSSP analysis shows that the N-terminal portion of HP-2 is more helical than the C-terminal region, however the total helical content is significantly lower for HP-2 (2.2 ± .7 %) than for HP-1 (19.8 %) (Figure 3-8B). LR analysis indicates that HP-2 does not adopt significant helical content, with less than 5% population even for short 3-residue helices (Figure 3-7). Instead, HP-2 contains a modest population of turn (22.0 ±

1.8 %) in the center of its sequence and a small population of anti-parallel β-strand at each of the termini. The high turn population could indicate nascent helix initiation or transient β−turn structure.

Cluster analysis was performed on the backbone of HP-2 from residues 55 to 59 since this region contains the native helix of HP36 and DSSP indicates significant turn population in the fragment. In Figure 3-10, we show representative structures for the top five clusters, each of which makes up 10–30 % of the entire ensemble of HP-2. This is unlike HP-1, in which more than half of the entire ensemble is comprised of a single structure cluster. All of the clusters have very low α-helicity, except for the second cluster which has structures that are most helical in the center (approximately 10 %). The smaller clusters appear to sample random coil and turn-like structures.

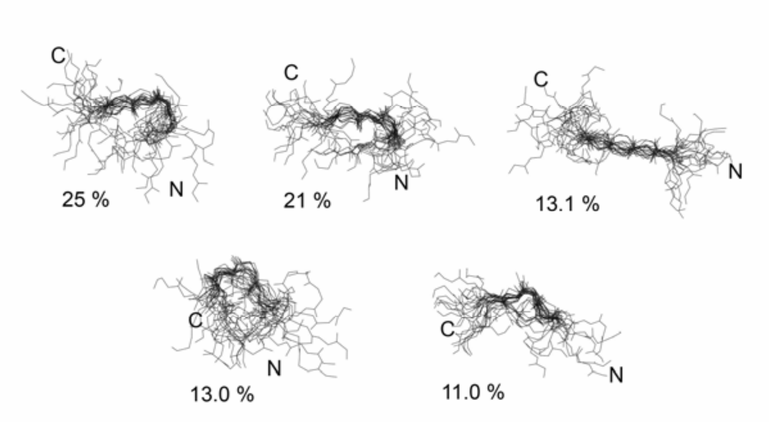


Figure 3-10. Representative structures for the five most populated clusters of HP-2. (A) 1st cluster (25.2 %) (B) 2nd cluster (21 ± 3 %) (C) 3rd cluster (13.1 ± .6 %) (D) 4th cluster (13 ± 2 %) (E) 5th cluster (11.0 ± .6 %). All five clusters sample random coil and turn-like structure.

Overall, HP-2 does not show strong evidence of any well-defined structure. It appears that the residues in HP-2 are more flexible and this region of HP36 relies on tertiary contacts and packing constraints to stabilize secondary structure. Vermeulen et al.

have suggested that the second helix should not be stable without residues from the third helix which favorably interact with the second helix's dipole [135].

3.3.7 Analysis of HP-3

HP-3 (P62LWKQQNLKKEKGL75) contains the longest helix in the native structure of HP36 (L63 to E72), with 7 stable α-helical hydrogen bonds, a proline that promotes helix initiation in native HP36, and a patch of acidic and basic residues that serve as part of the actin binding domain. Figure 3-5C shows the free energy landscapes for the entire ensemble of HP-3 at 300 K along the radius of gyration (Rg) and RMSD to the HP36 structure. Despite the relatively large number of helical hydrogen bonds in the HP36 structure, the distribution of the minima for HP-3 is the broadest of all three fragments, centered at an RMSD of 4.0 Å and Rg of 8.0 Å. The structures are also somewhat less compact than in HP36 (Rg of 9.0 ± 2.0 Å compared to 7.0 Å in native HP36). While HP-1 showed significant sampling of native-like backbone structure (RMSD = 1.0 Å), and HP-2 showed a larger average RMSD of 2.9 Å, HP-3 shows even larger deviations from the HP36 native structures with an average RMSD of 4.5 ± 1.9 Å. These larger values may indicate further deviation from native structure as compared to HP-2, or they may arise from the larger size of this fragment.

Unlike HP-1 but similar to HP-2, HP-3 shows little residual helical content despite being the longest helix in HP36. Figure 3-6C shows Ramachandran free energy surfaces for three residues (W64, N68, and K73) selected from different parts of HP-3. W64 has the global free energy minimum in the helical region of the free energy surface while N68 occupies all 4 major basins at nearly equal free energies, unlike the well-defined α-helical conformation observed for this residue in the middle of helix 3 in HP36. Similar shallow minima are sampled by K73, which is consistent with the low level of structure seen at the C-termini of all three fragments.

Figure 3-7 shows results of LR analysis on HP-3, which indicates a 5-10% probability of forming short 3-4 residue helices, larger than was seen in HP-2. This is in agreement with DSSP calculations that show limited population of helical turns in the N-terminal region (W64 is approximately 20 % α-helical). This lack of helical structure in the C-terminus is consistent with the free energy surfaces of L69 and K73 which show shallow local minima sampled in the 4 major basins (Figure 3-6C). Unlike HP-1, nearly no propensity to form longer helices of 5-6 residues is observed. Instead, many structures in the ensemble of HP-3 have a high local turn population (20 ± 2 %) (Figure 3-8C). In the intact HP36, helix 3 makes extensive hydrophobic contacts with residues from helix 1 and helix 2. In the absence of these interactions, the helical structure is not stable.

Further evidence of significant conformational variability is obtained from cluster analysis, where the five largest clusters account for only 69% of the ensemble (Figure 3-11). DSSP analysis shows that the center of HP-3 samples a significant amount of turn conformation, with the only significant helical content being 3_{10} structure near the N-terminus that is present in clusters 1, 3 and 4. Clusters 1 and 4 are made up of α- helical structures at the N-terminus from L63 to Q66 while cluster 5 appears to be somewhat native-like, sampling a long helix between residues L63 and K72, with an average α-helicity per residue of 34 %. However, cluster 5 only comprises only 10 % of the structures and therefore it does not make a significant contribution to the ensemble average. Unlike HP-2, however, it does indicate that HP-3 has a small propensity to adopt a native-like conformation.

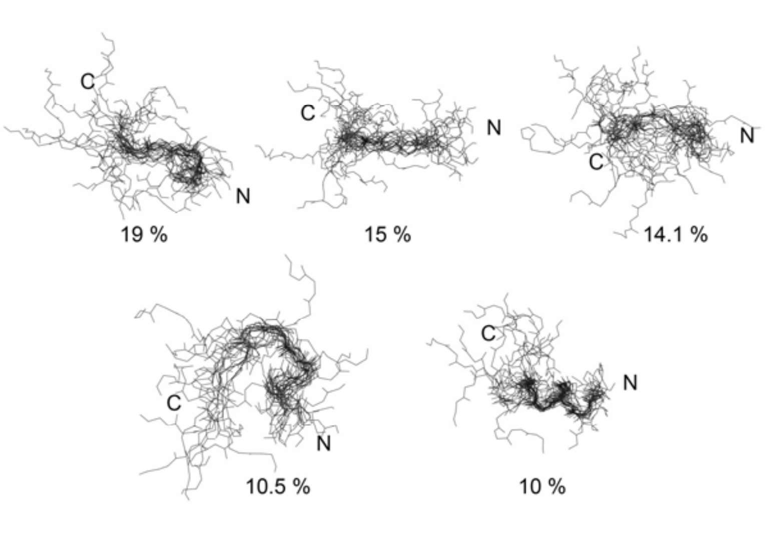


Figure 3-11. Representative structures for the five most populated clusters of HP-3. (A) 1st cluster (19 ± 2 %) (B) 2nd cluster (15 ± 2 %) (C) 3rd cluster (14.1 ± .7 %) (D) 4th cluster (10.5 ± .9 %) (E) 5th cluster (10 ± 5 %). Cluster 5 contains the most native-like helix, however the population is quite small.

3.4 Discussion

3.4.1 What may stabilize the high population of helical structure in HP-1?

We examined the entire HP-1 ensemble to identify contacts that may be playing a part in stabilizing the helical structure. Approximately 50 % of the ensemble had ion-pair contacts between D44 - K48 (27 ± 4%), E45 - K48(8 ± 2 %), and both D44 and E45 with K48 (14.0 ± .03 %). Another contact was present (85.0 ± .5 %) between the D44 backbone carbonyl and the charged sidechain of K48; this is present alone (56 ± 4 %) and with the charged sidechain of D44 (30 ± 6 %) (Figure 3-12). These contacts are not observed in the NMR and the X-ray structures where the charged groups of these residues are more than 6 Å apart in space. These interactions appear to desolvate the backbone hydrogen bonds, resulting in stabilization of α-helical structure. These types of interactions have been shown to favor helical structure in various peptide systems [169].

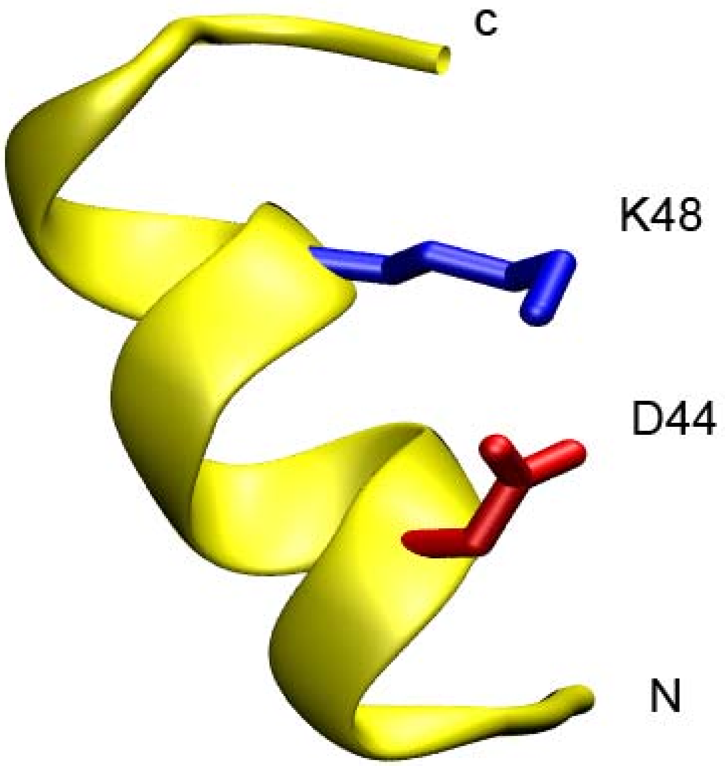


Figure 3-12. Representative structure from the most populated cluster of the HP-1 ensemble. This image shows the interaction between the sidechain of K48 and D44, which may play a role in stabilizing the helical structure in HP-1.

3.4.2 Comparison with experimental data

The tendency of the three fragments to adopt helical structure in simulations is in good agreement with the trends seen in CD experiments (HP-1 > HP-3 > HP-2). These differences observed in the experiment are small due to the length and the low population of helix. Figure 3-13 shows the theoretical and experimental J-coupling values for the residues in all three fragments. The calculated J-couplings match most of the experimental trends with the exception of a few residues whose deviation from experiment is quite small (<1.5 Hz). HP-1 and HP-3 have calculated J-couplings that are lower than 7 Hz (shifted to the helical region) in the N-terminal region, consistent with analysis of their ensembles. The calculated J-couplings of HP-2 are 7.0 ± 0.5 Hz, consistent with an average ensemble that has no specific structural preference. While the results do not show any strong conformational preferences, they show that the ensembles generated in the simulations are able to reasonably reproduce experimental parameters.

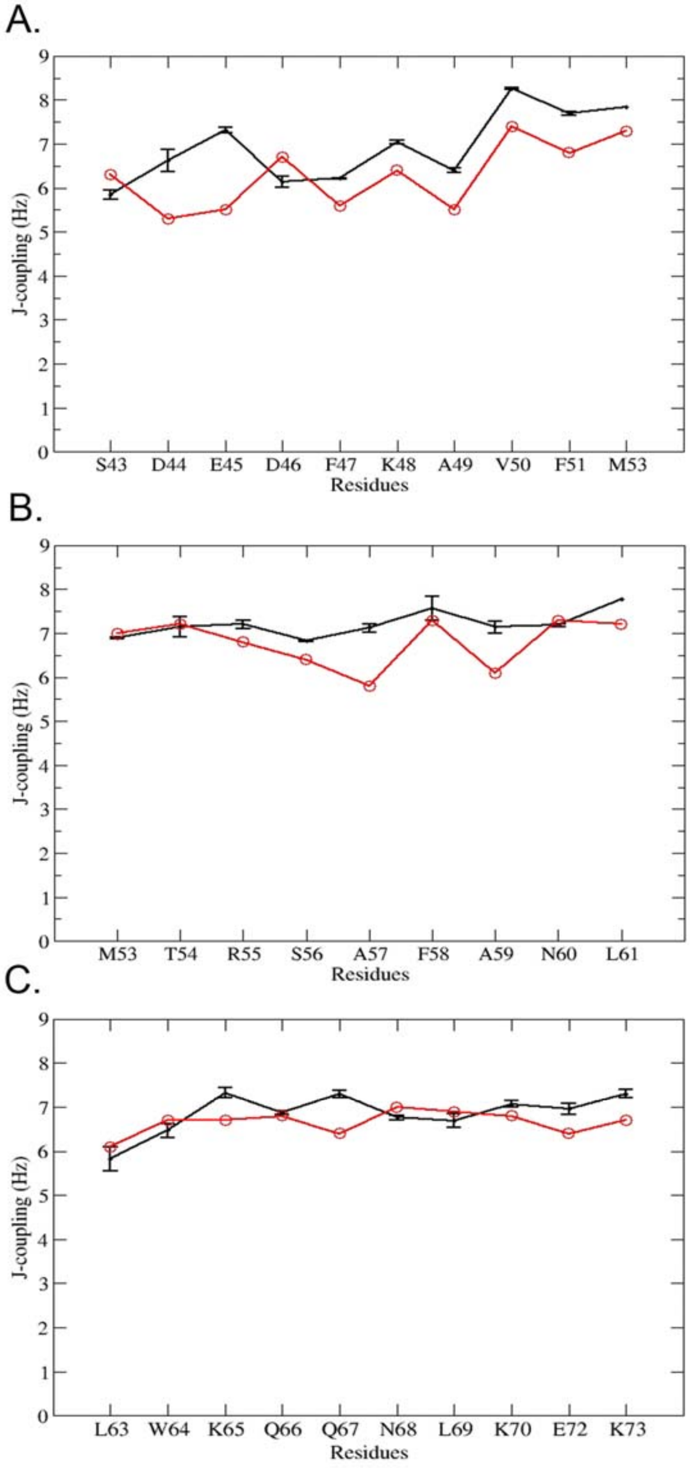


Figure 3-13. J-coupling values for residues in (A) HP-1 (B) HP-2; (C) HP-3. Experimental values are shown in red and calculated values are shown in black. Calculations for all three fragments followed the relative trends and were in good agreement of the experimental data. Some residues are missing because either the J-coupling constants were not measured or the results were ambiguous.

Some differences exist between the simulation data and previous solid state NMR freeze quench studies that also suggested some non-random structure [106]. Those studies are consistent with V50 in HP-1 adopting a relatively well ordered local polyproline II (PPII) conformation; A57 in HP-2 is more conformationally disordered, but retains significant helix content; and L69 in HP-3 is the most disordered of these three labeled residues in unfolded 35-residue villin headpiece subdomain (HP35). Those experiments suggested that local structure is present for HP-1 and HP-2 but only disordered structures are populated for HP-3 [106]. Our study showed no backbone conformational preference at 300 K for V50 in HP-1, A57 in HP-2, and L69 in HP-3 (Figures 3-6). However, the solid state NMR experiments are performed at cold temperatures which might induce these residues to adopt more a rigid backbone conformation. Our analysis evaluated the ensembles generated at 300 K, conditions that were similar to the original fragment study, and the structural populations are more relevant to folding at this temperature.

3.4.3 Implications for folding

This study suggests that the HP-1 has the highest tendency to adopt helical structure among the three fragments, and these have high similarity to the structure of the fragment in HP36. HP-3 also samples fully formed structure as adopted in the native state of HP36, but at a significantly smaller overall probability than HP-1. HP-2 contains the least residual structure and samples a wide variety of conformations, all in low population. These isolated structures need to be stabilized by other contacts to form native helical structure. All of the fragments are more helical than Ala_{10}, suggesting that the side chains play an important role and that observed tendency to form helices does not arise from over-stabilization of helical structure that has been reported for earlier versions of the Amber force field.

Overall, the geometric ensemble properties of HP-1 are remarkably similar to that found in the HP36 native state, but with a much lower overall propensity to form well-defined structure. In previous studies, the Pande lab has proposed the "mean-structure hypothesis" which states that the geometry of the collapsed unfolded state of small peptides and proteins in an average sense corresponds to the native equilibrium state even though individual structures in the ensemble demonstrate unfolded, random coil properties [93]. The presence of a significant population of HP-1 structures with low RMSD values suggests that at least some of the individual structures sampled may also be highly (although locally) native-like.

Preformed structure in the unfolded state has been implicated for potentially favoring very rapid folding. Residual structure might help guide proteins to fold into the native state. Recent studies of model helical bundles have suggested that such residual structure is essential in aiding the protein folding process [139, 148]. The presence of low levels of highly native-like structure in the HP-1 fragment may play a role in the fast folding of HP36. This residual helical content in the HP-1 fragment varies only weakly with temperature, with average α-helical content of 19% at 300K and 17% at 340K (considering all helix lengths). This is consistent with experimental studies for larger fragments and with the intact protein, which have shown that there is considerable structure in the unfolded state at higher temperatures [26]. We note, however, that force fields of the type used in this study are not parameterized to quantitatively reproduce temperature-dependent behavior.

The diffusion collision model has often been applied to helical proteins. However, Islam et al. has noted that this model is ineffective at describing the folding of HP36 due to the relatively small size of the helices in the subdomain [94]. Residual structure in isolated helices may not be enough to drive the folding process. Gianni et al. [170] and Daggett et al. [171] have suggested that some proteins can form unstable secondary structure that will become stable once tertiary contacts are secured around a nucleus of hydrophobic contacts. This mechanism seems to relate better to HP36.

Much experimental work has focused on the contacts that stabilize the native state. The work by McKnight's group has stressed the importance of three phenylalanines in maintaining the hydrophobic core [97]. Experimental fragment studies [26] have shown the fragment containing the first two helices of HP36 and all three of these phenylalanines maintains a considerable amount of residual structure, presumably due to these hydrophobic interactions. The crystal structure of HP35 has also suggests some hydrogen bonding interactions between the first two helices that may influence the compactness of the structure in HP36's unfolded state [89]. These interactions stabilize the first and second helix and allow them to form more structure than seen in the individual fragments.

In summary, REMD simulations using explicit solvent have been used as a method for studying the propensity to populate locally stabilized unfolded state structure in HP36. Two simulations using explicit solvent were run for each fragment and both converged to the same population of structures. HP-1 was shown to contain the most helical structure with a low RMSD to the native HP36 structure, implying that this region may be partially structured in the unfolded state of HP36. The low tendency to adopt helical structure in the other two fragments indicates that these rely on contacts from each other for stability. This is in agreement with experimental studies which demonstrate that tertiary contacts are necessary to form stable, detectable structure.

4. Simulations of the Larger Peptide Model of the Unfolded State of HP36

Abstract

Recent experimental and computational studies of HP36 have been carried out with the goal of trying to understanding the role of residual structure in the unfolded state of this subdomain. A large fragment made up of helix-1 and helix-2 (HP21) of HP36 has been studied with NMR and CD and has shown more helical structure than the isolated fragments and native-like tertiary contacts between Phe residues. Several NMR experiments suggest this may be a reasonable model for the denatured state of HP36. In order to further characterize the structure of this peptide, we ran standard REMD simulations from unfolded and folded conformations. For HP21, we found that the region corresponding to the first helix in HP36 contained the most native-like structure, which is consistent with the isolated fragment HP-1. There also appears to be a small part of the ensemble which indeed forms the phenylalanine core. Nevertheless, comparisons of experimental and calculated J-coupling constants and chemical shifts show that the ensemble obtained from the simulation is not as helical as suggested by the experiment. A subset of structures within the ensemble containing the phenylalanine core showed better agreement between the calculated and NMR observables in regions of HP36 containing helix-1 and helix-2. Some approaches are mentioned that may be possible solutions to the issues with this peptide.

Acknowledgements

The material presented in this chapter was written by Lauren Wickstrom with suggestions and revisions from Carlos Simmerling and Daniel Raleigh. Roberto Gomperts assisted in running simulations at SGI.

4.1 Introduction

HP36 is one of the fastest folding model systems studied experimentally and computationally [34, 87, 93, 94, 97, 101, 102, 108, 112, 113, 116, 149]. One major focus of these studies has been to understand the role of residual structure in the unfolded state and how it affects the fast folding of HP36 [26, 34, 87, 95, 100, 107, 118, 172-174]. There have been a few views regarding this matter. H/D NMR experiments on the native state of HP36 displayed several slowly exchanging amide resonances with protection factors that were larger than predicted based upon $\Delta G_{unfolding}$ that were mainly located in helix-1 and helix-2. This could be due to structure in the unfolded state in that region [87]. FRET studies in 8 M urea, however, have suggested that the region between helix-2 and helix-3 remain compact in the denatured state [173]. This result is also in accord with MD simulations which show that folding is initiated between helix-2 and helix-3 [34, 95, 118, 172, 174, 175]. Yang et al. [175] has noted that despite the initial formation of these helices, there is indeed native helical content in helix-1 and helix-2 in the transition state ensemble . Simulation studies of a double mutant of HP36 have stressed the formation of helix-1 and the phenylalanine core as being important for the its fast folding [176].

Recent studies by the Raleigh group have explored the denatured state expermentally using fragment analysis [26, 107]. While none of the individual fragments corresponding to the α-helices in HP36 showed residual structure, a 21-residue fragment with the first two helices of HP36, known as HP21, displayed secondary structure (20-25 % helicity) and possible tertiary interactions between the aromatic residues and Val50. In addition, Val50 methyl chemical shifts were similar in HP21 and estimated thermally denatured state chemical shifts of HP36 obtained from temperature dependent NMR folding analysis [98]. H/D NMR exchange experiments of HP21 revealed the presence of six of the seven most protected amides seen previously in the native state studies of HP36 (F47, K48, A49, V50, F51, F58, L61) [87, 107]. If this is a good model for the unfolded HP36, these results could imply that there is significant structure in the denatured state of HP36 in this region.

NMR studies have elucidated more structural details of this peptide [107]. J-coupling constants ($^3J(H_N,H_\alpha)$), backbone-sidechain NOEs, backbone-backbone NOEs and chemical shifts ($^1H^\alpha$, $^1H^N$, ^{15}N, $^{13}C^\alpha$, $^{13}C^\beta$ and ^{13}CO) suggest that HP21 populates an ensemble of structures which are helical in regions that are helical in the crystal structure of HP36 [177]. In addition, aromatic-sidechain NOEs suggest both native and non-native hydrophobic clustering. Native-like contacts are formed between the phenylalanines and other hydrophobic residues in the peptide based on those NOEs. It is difficult to calculate one representative structure because it is likely to be sampling many interconverting conformations. Secondly, there are not enough NOEs to accurately define a specific structure. There are extensive amount of short and medium range NOEs (approximately 50) however there are less than 10 long range NOEs to define tertiary contacts of the peptide.

Hence, there are still many questions that remain about this peptide.

1) What are the different conformations that make up the ensemble of structures?

2) What are the correct populations of structures?

In this study, we apply a similar approach to that used in our studies of the isolated fragments of secondary structure. In the previous work, we used REMD in explicit solvent to obtain ensembles for HP-1, HP-2 and HP-3 and found that HP-1 contained the largest population of native-like structure of all of the fragments [79]. This would suggest that helix-1 could form local native-like structure in the unfolded state of HP36. We conducted REMD simulations of HP21 and those results are described in this chapter. We characterize the structure formation with DSSP, RMSD and phenylalanine contact distances. Further analysis evaluates the accuracy of the ensemble by comparison of experimental and calculated scalar couplings and C_α chemical shifts. We find that HP21 contains helical structure in helix-1 and has a low structural population with phenylalanine core formation. In addition, helix-2 displays the same helical content in HP21 and HP-2. The ensemble, however, contains less helical structure than predicted by experiment. These results suggest that there are still issues with the simulation methodology such as sampling, water model or the force field which need to be investigated.

4.2 Methods

4.2.1 Preparation of structure

HP21 was built using the first 21 residues from the sequence of HP36: M41-F76 (MLSDEDFKAVFGMTRSAFANLPLWKQQNLKKEKGLF). HP21 was simulated with a free N-terminus and amidated at the C-terminus in order to correspond to the system used by the Raleigh group. These sequences and termini correspond to the those used in the experimental studies [26, 107]. All sidechains for Asp, Glu, Lys and Arg were charged during the simulation. Simulations were performed using the ff99SB [45] force field in Amber 9 [39]. SHAKE [122] was used to constrain bonds with hydrogen. The time step was 2 fs. Temperatures were maintained with Berendsen coupling [123] . Simulations were initiated from a folded and a collapsed conformation. The folded structure was built by deleting residues 62 through 76 from the NMR structure (pdb code 1VII [88]). The collapsed structure was obtained from the 449 K temperature trajectory of the REMD simulation started from the folded structure. This conformation was completely lacked significant secondary structure.

4.2.2 Explicit solvent simulations

Simulations in explicit water were performed in a truncated octahedral box with periodic boundary conditions and particle mesh Ewald [125] (PME) to calculate long-range electrostatic interactions. The water box contained 3970 TIP3P [49] waters. The structures were equilibrated at 300 K for 50 ps with harmonic restraints on solute atoms, followed by minimization with gradually reduced restraints. The restraints were reduced from 5 kcal/mol*Å to 1 kcal/mol*Å to 0.5 kcal/mol*Å. After minimization, three 5 ps MD simulations were performed with the same gradually reduced restraints at constant pressure (1 atm) and temperature (300 K) in order to generate starting structures. We implemented the REMD version in Amber 9. The target exchange acceptance ratio was approximately 13 %. Exchanges between neighboring temperatures were made every 1 ps.

The collapsed and the folded conformation were surrounded by 3970 waters. Both simulations required 54 replicas ranging from 276 to 518 K (276.1, 279.4, 282.7, 286.1, 289.5, 293.0, 296.5, 300.0, 303.6, 307.2, 310.9, 314.6, 318.4, 322.2, 326.0, 329.9, 333.9, 337.8, 341.9, 346.0 350.1, 354.3, 358.5, 362.8, 367.1, 371.5, 376.0, 380.5, 385.0, 389.6 394.3, 399.0, 403.7, 408.6, 413.5, 418.4, 423.4, 428.5, 433.6, 438.8, 444.0, 449.3, 454.7, 460.1, 465.6, 471.2, 476.8, 482.5. 488.3, 494.1, and 500.0 K) The simulation starting from the folded structure was run for 38 ns while the simulations starting from the extended conformation was run for 100 ns. The first 10 ns of each run was discarded.

4.2.3 Analysis

Analysis was performed at 286 K unlike the isolated fragment studies which were performed at 300 K since the NMR experimental studies of HP21 were performed at 285 K [26]. DSSP analysis [128] and distances, and RMSDs were calculated using the ptraj module in Amber 9. RMSD calculations used the X-ray structure (pdb code 1YRF [89]) as a reference structure because previous studies have shown that this is the better representation of the folded state at neutral conditions [79]. Distances were calculated between the center of mass of the heavy atoms of the phenylalanine rings of F47 and F51 and F47 and F58. Relative free energy histograms were calculated at 300 K according to equation 4-1.

$$\Delta G_i = -RT \ln(N_i/N_0)$$

Equation 4-1. Relative free energy calculated with multidimensional histogram analysis. ΔG_i is the relative free energy bin I, N_i is the population of a particular histogram bin along the reaction, and N_0 is the most populated bin. R and T are the gas constant and temperature respectively.

J-coupling constants were calculated using a version of the Karplus equation (Equation 4-2) previously employed for analysis of small peptides:

$$^{3}J(H_N,H_\alpha) = A\cos^2(\varphi - 60) + B\cos(\varphi - 60) + C$$

Equation 4-2. Karplus equation for the calculation of $^{3}J(H_N,H_\alpha)$ scalar couplings. A, B, and C are constants.

where A = 6.51, B = -1.76, and C = 1.60 [161]. These parameters were used in the previous fragment study [161]. We also calculated the scalar constants with another set of Karplus parameters from Brushweiler et al. [178] (A = 9.5, B = -1.4, C = 0.3). These parameters do not include motional averaging which is normally incorporated into the fitting of empirical J-coupling parameters. Two parameter sets were used in order to test the sensitivity of the data. The average and standard deviation were calculated for each set of data. Chemical shifts for the C_α chemical shifts were calculated using the SHIFTS [179, 180] and SHIFTX programs. Chemical shift deviations were calculated using the random coil values from Wishart et al. [181]. The average and standard deviation were calculated for each set of data. Since there were issues with the agreement of the experimental and calculated data for the initial ensemble calculations, other NMR observables were not calculated for the ensembles or subsets. Cluster analysis was performed for the subset of structures with the phenylalanine core with MOIL-View [127], using backbone RMSD as a similarity criterion with average linkage. Clusters were formed with a bottom-up approach using a similarity cutoff of 2.5 Ǻ.

4.3 Results

4.3.1 Structural properties of the HP21 ensemble

HP-21 (M41LSDEDFKAVFGMTRSQFANL61) consists of the sequence that forms the N-terminal and central helix of intact HP36 (Figure 4-1). In the NMR structure of HP36 [88], the first helix extends from D44-K48, while in contrast the first helix is three residues longer in the X-ray structure [89], ranging from D44-F51. The lengths of helix-2 are quite similar in length, extending from R55-F58 in the NMR model and R55-A59 in the X-ray structure. The HP21 fragment also contains the three phenylalanines (F47, F51, F58) which play an essential role in the hydrophobic core and stability of HP36. We first present properties of the entire ensemble which is followed by a more detailed discussion of preferred conformations.

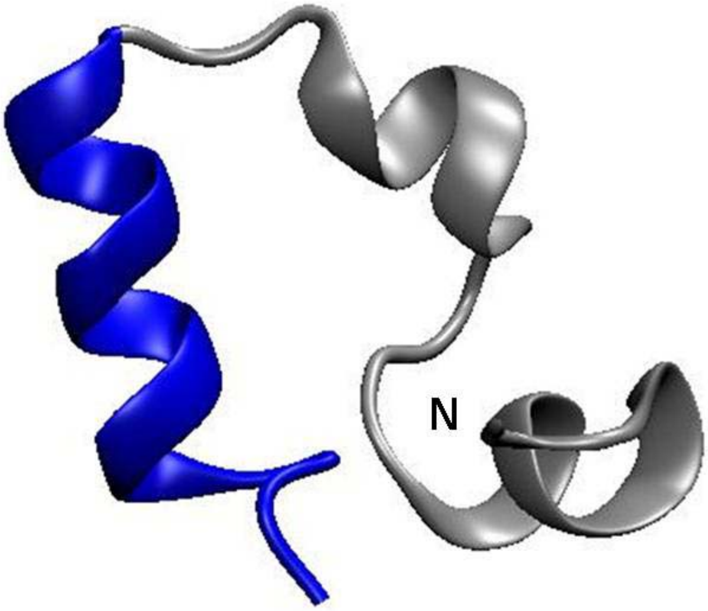


Figure 4-1. Structure of the villin headpiece subdomain (pdb code 1VII [88]). HP21 is in silver. The N-terminus is labeled.

DSSP analysis was employed to evaluate helix formation in the HP21 peptide (Figure 4-2). In both simulations, the ensemble samples a higher helicity in the N-terminus of the peptide than in the C-terminus (average helicities of 21.41 +/- 0.08 % and 8.14 +/- 0.91 % at the N and C-termini respectively). The highest helical content is centralized around E45 which is consistent with the location of the first helix in X-ray and NMR structures of HP36. These results are consistent with the previous simulation studies of the isolated fragments, in which the region where helix-1 is located in HP36 contains the most helical structure.

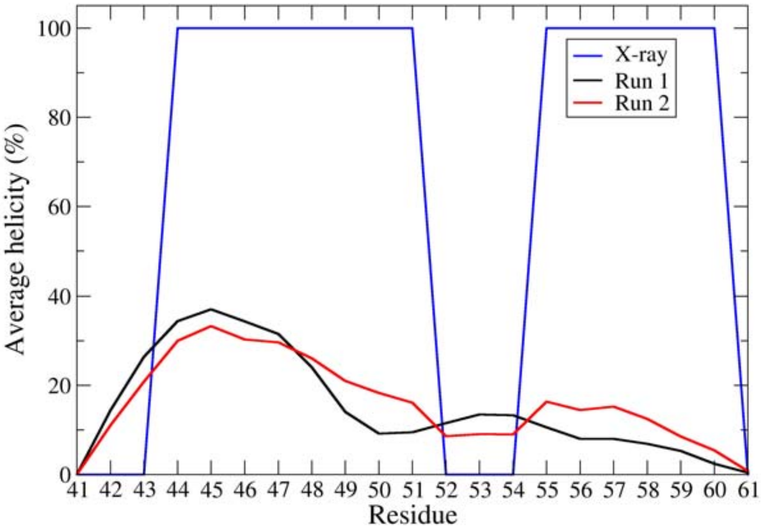


Figure 4-2. Average helicity per residue for the simulations of HP21 starting from a collapsed (black) and a folded (red) conformation at 286 K. The helicity for the X-ray structure [89] of HP35 is shown in blue. Helicity was calculated with DSSP analysis combining 3_{10} and α-helical content.

Figure 4-3 shows the free energy landscape at 286 K for HP21 along the coordinates of RMSD of helix 2 and helix-1 to the X-ray structure of HP36. The global minima are different in the simulations starting from independent conformations. Since the simulations have not reached equilibrium, it is not formally correct to consider these in terms of free energies. In this discussion, we will assume that they are equilibrated in order to understand the most favored conformational states of the peptide. In the simulation starting from the collapsed conformation, the global minimum is located at an RMSD (helix-1) of 3.2 Å and an RMSD (helix-2) of 3.1 Å which would suggest that the structures have sampled non-native conformations. In contrast, the global minimum is located at an RMSD (helix-1) of 1.0 Å and an RMSD (helix-2) of 0.2 Å for the simulation started from a folded conformation. Nevertheless, there is a local minimum located at an RMSD (helix-1) of 1.0 Å and an RMSD (helix-2) of 3.0 Å that is quite shallow and has a relatively small free difference from the global minimum (0.4 kcal/mol). There is a small tendency of forming both the native helix-1 and helix-2 at the same time (approximately 1 % and 7 % of the structures have an RMSD < 1.0 Å for helix-1 and helix-2 in the simulations starting from the collapsed and folded conformations respectively). HP21 forms more native-like structure in the region of the first helix in the full length HP36 than in the region of the second helix. Approximately 16 % and 17 % of the structures have an RMSD < 1.0 Å for helix-1 in the simulations starting from the collapsed and folded conformations respectively. In contrast, approximately 2 % and 9 % of the structures have an RMSD < 1.0 Å for helix-2 in the simulations starting from the collapsed and folded conformations respectively. From the landscape, the barriers are higher for the formation of structures with a low RMSD (helix-2) which would suggest that the region of helix-1 should be native-like in order for helix-2 to form native-like structure. Nevertheless, these results agree with our previous results [79] that helix-1 contains the most native-like structure.

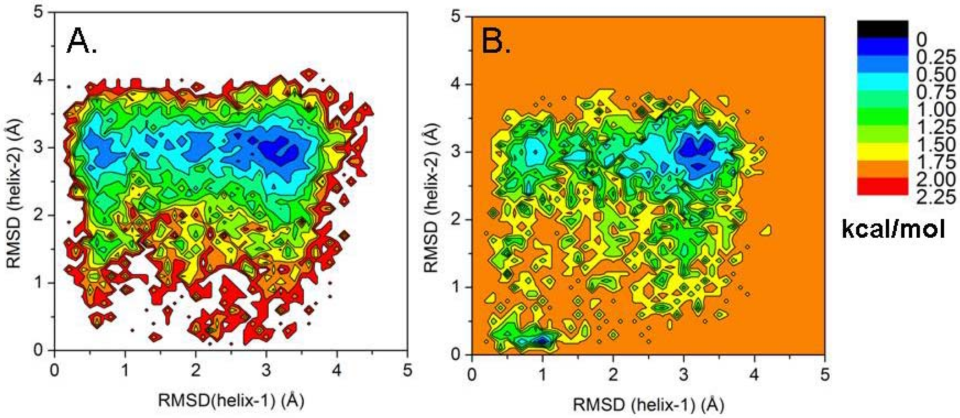


Figure 4-3 Free energy landscapes of the RMSD(helix-2) vs RMSD(helix-1) for the simulations of HP21 starting from a collapsed (A) and a folded (B) conformation at 286 K.

To investigate phenylalanine contacts in this peptide, we calculated free energy surfaces for the distances between F47/F51 and F47/F58 (Figure 4-4). Contacts are formed at distances less than 7.5 Å. This cutoff was selected based on distance histograms of the phenylalanine distances. To form a phenylalanine core similar to the folded state of the intact HP36, both phenylalanine contacts should be present. The global minima in each of these landscapes are different for both simulations. In the simulation starting from the collapsed conformation, the global minimum is located at a distance of 13.0 Å and 17.0 Å for F47/F51 and F47/F58. In contrast, the global minimum is located at a distance of 5.0 Å and 5.0 Å for F47/F51 and F47/F58, which is similar to the distances in the X-ray structure (4.68 Å and 5.28 Å for F47/F51 and F47/F58) in the simulation starting from the folded conformation. This minimum is also on the free energy surface of the run started from the collapsed conformation (0.7 kcal/mol higher than the global minimum on the surface). In both simulations, these structures sample this hydrophobic core (approximately 1 % and 9 % of the structures form a phenylalanine core in simulations starting from the collapsed and folded conformation respectively).

Overall, HP21 preferred to form the F47-F51 contact (approximately 16 % and 25 % of these structures form this contact starting from the collapsed and folded conformation respectively) compared to the F47-F58 contact (approximately 6 % and 14 % of these structures form this contact starting from the collapsed and folded conformation respectively). From these results, we can conclude that HP21 is forming a well populated native-like contact between F47-F51 and a minor population of structures which contain a phenylalanine core.

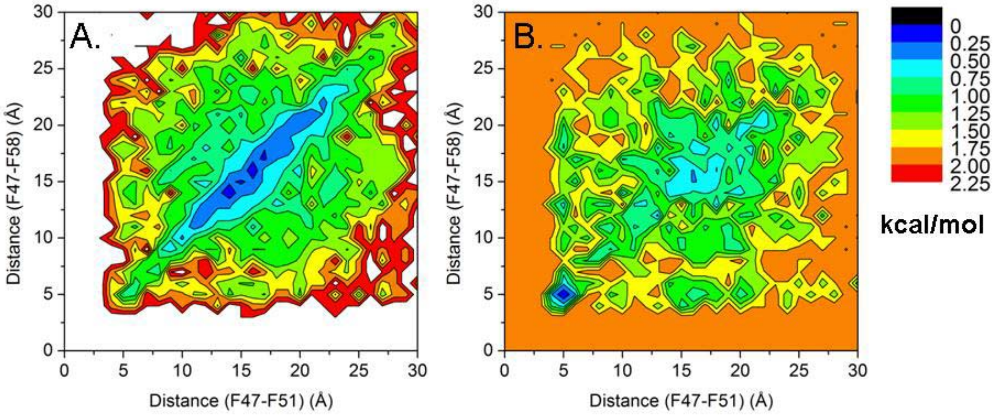


Figure 4-4 Free energy landscapes of the distances between F47 and F58 versus the distance between F47 and F51 for the simulations of HP21 starting from a collapsed (A) and folded (B) conformation at 286 K.

4.3.2 Comparison of calculated and experimental NMR observables for the HP21 ensemble

Although DSSP and RMSD are excellent structural measures, they can not be directly compared with experimental values in order to interpret the accuracy of a simulation because they are not a direct measure of populations measured in the experiment. To evaluate our structural populations, we calculated and compared $^3J(H_N,H_\alpha)$ scalar coupling constants using two Karplus parameter sets to experimental scalar constants for both simulations at 286 K (Figure 4-5). Overall, the calculated scalar coupling constants for both simulations appear to be shifted to more random coil values compared to the experimental constants. The scalar coupling trends show that the helical content of N-terminus is higher than the C-terminus consistent with the DSSP analysis. The sensitivity of the Karplus parameters is small with an average RMS to experimental values of 1.4 and 1.8 Hz using the Vuister and Brushweiler parameters [161, 178]. The ensemble appears to contain less helical structure than predicted by the experimental values.

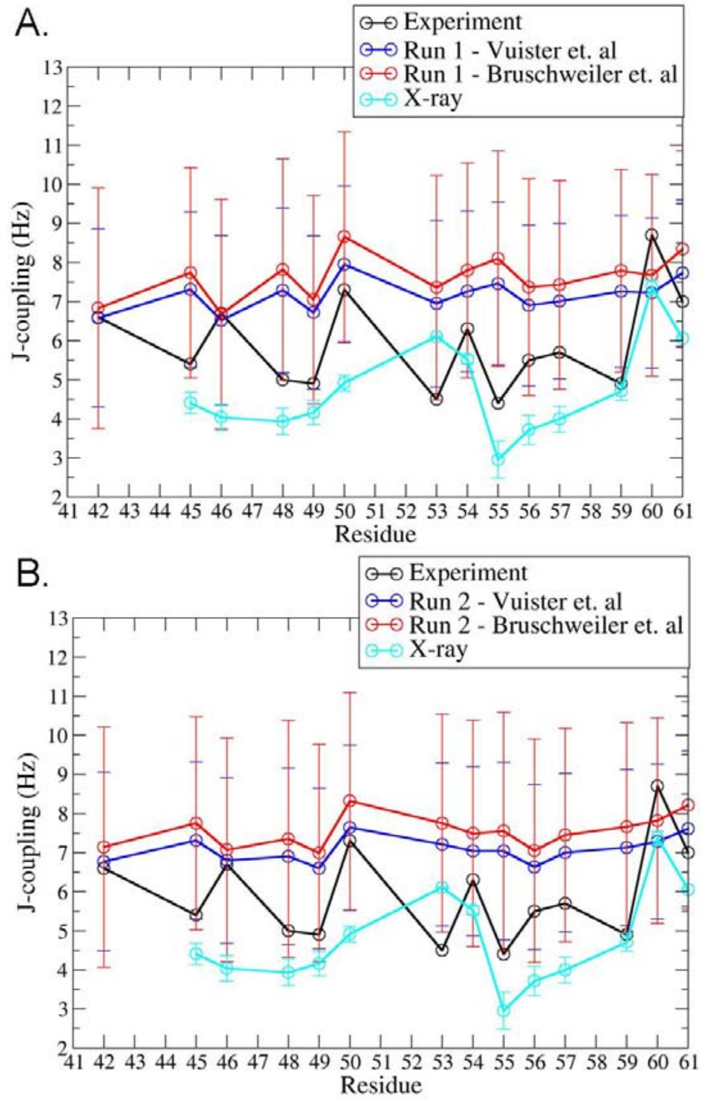


Figure 4-5. Comparison of calculated and experimental $^3J(H_N,H_\alpha)$ scalar coupling constants for the simulations of HP21 starting from a collapsed (A) and a folded (B) conformation at 286 K. Experimental values are shown in black. Each run used the two Karplus parameter sets for the scalar coupling calculations. Some residues are missing because either the J-coupling constants were not measured or the results were ambiguous in the experiment. The standard deviation was shown for each calculated constant. The scalar coupling constants were also calculated for the X-ray structure[89] of HP35 with both parameter sets.

In addition, we also calculated C_α chemical shifts for both simulations using SHIFTS and SHIFTX. In Figure 4-6, we compared chemical shift deviations for the calculated and the experimental values. C_α CSDs greater than zero correspond to α-helical structure while C_α CSDs less than zero correspond to β-structure. C_α CSDs are considered random coil if they are equal to zero. Similar to the scalar constant comparison, the α-helical populations appear to be higher in the experiments compared to the simulations. The results from the SHIFTS and SHIFTX calculations are almost identical (average deviation of .05 +/- .01). Overall, the ensembles contain a smaller helical population for HP21 than predicted by the experiment.

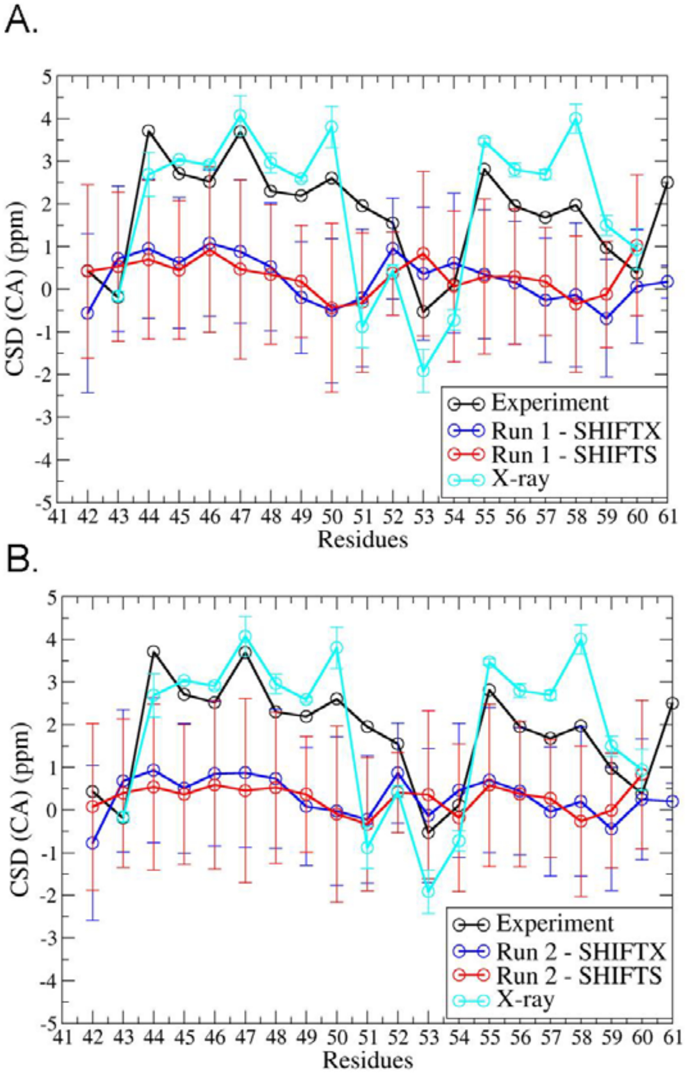


Figure 4-6. Comparison of calculated and experimental C_α chemical shift deviations for the simulations of HP21 starting from a collapsed (A) and a folded (B) conformation at 286 K. Experimental values are shown in black. Chemical shifts were calculated with SHIFTX (blue) and SHIFTS (red). The standard deviation was shown for each calculated scalar coupling constant. The scalar coupling constants were calculated for the X-ray structure of HP35 [89] with both parameter sets.

4.3.3 Comparison of calculated and experimental NMR observables for a subset

The next question investigated was whether any structures within the ensemble resemble the experimental measurement. Since the results deviated significantly from the experimental values, there are questions about if the correct structures are being generated or if there is an issue with not having the correct population of structures. A subset of structures was collected based on the criteria of containing a phenylalanine core. This criteria was selected because previous NOE data [107] has suggested that HP21 forms contacts between F47/F51 and F47/F58. The subset of structures was selected based on phenylalanine core formation. A core was formed if both phenylalanine contacts were < 7.5 Å. This group was approximately 3 % of the entire ensemble collected from both simulations.

Figure 4-7 shows a comparison of the calculated and experimental $^3J(H_NH_\alpha)$ scalar coupling constants for the subset of structures. Both sets of calculated J-coupling constants (RMS(Brushweiller) = 1.27 Hz and RMS(Vuister) =1.05 Hz) are in better agreement with the experimental values than the entire ensemble (RMS(Brushweiller) = 1.40 Hz and RMS(Vuister) =1.80 Hz). The trends of the calculated constants match the experimental trends showing the scalar coupling fluctuations through the sequence unlike the ensemble, which showed a flat sequence dependence (Figure 4-5). It should be noted that the subset standard deviations have become smaller and include the experimentally measured values (with the exception of M53). These results imply that the formation of the phenylalanine core is correlated with helix formation around the regions of native helix in HP36. Cluster analysis was performed on the backbone of the subset of structures (Figure 4-8). In the subset, approximately 63 % of the structures sample a backbone similar to the X-ray structure. In addition, DSSP analysis of the subset showed an increase in helical content in the regions which correspond to helix-1 and helix-2 in the X-ray structure compared to the entire ensemble (Figure 4-9).

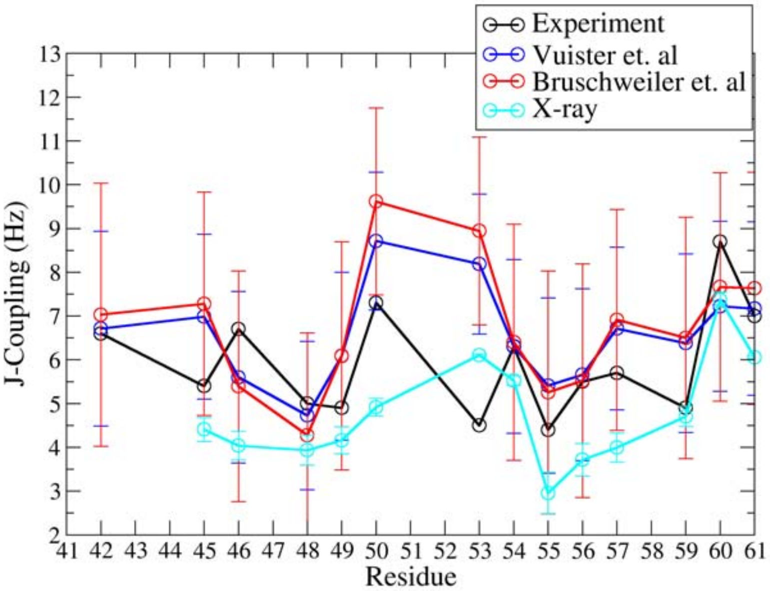


Figure 4-7. Comparison of experimental and calculated $^3J(H_N,H_\alpha)$ scalar coupling constants from a set of structures containing the phenylalanine core at 286 K. Experimental values are shown in black. Two Karplus parameter sets were used for the scalar coupling calculation. Some residues are missing because either the J-coupling constants were not measured or the results were ambiguous in the experiment. The standard deviation was shown for each calculated scalar coupling constant. The scalar coupling constants were also calculated for the X-ray structure with both parameter sets.

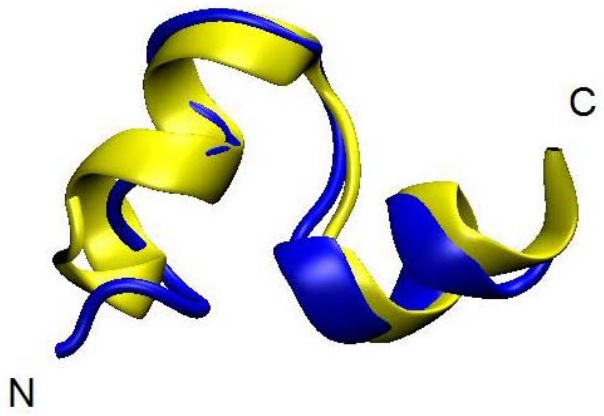


Figure 4-8. Comparison of the representative structure of the most populated cluster of the subset of structures containing the phenylalanine core (blue) and the X-ray structure (yellow). A best fit alignment is performed on residues 44 to 62. The backbone of the representative structure is similar to the backbone of the X-ray structure.

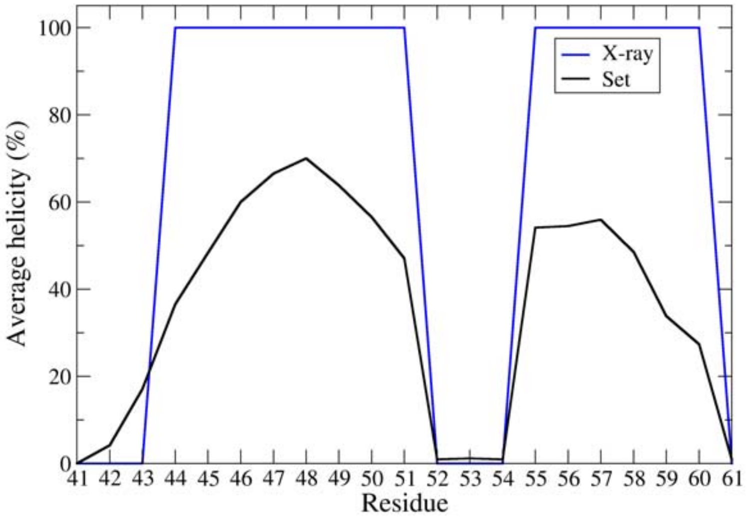


Figure 4-9. Average helicity per residue for the subset of structures (black) containing the phenylalanine core at 286 K. The helicity for the X-ray structure is shown in blue.

C_α CSDs were calculated for the subset of structures and compared to experimental CSDs (Figure 4-10). The agreement between the experimental and calculated CSDs has improved for both methods similar to the improvement noted with the scalar couplings

(RMS = .91 for SHIFTS and .93 for SHIFTX) compared to the entire ensemble of structures (RMS = 1.65 +/- 0.02 for SHIFTS and 1.65 +/- .07 for SHIFTX). The subset is shifted to more α-helical CSDs than the entire ensemble. Notable agreement is seen in the region where the helix-2 is located in the full length HP36. The biggest deviation between calculated and experimental values is between V50 and G52 which is also consistent with the scalar coupling results. In addition, these experimental shifts also deviate from the chemical shifts of the X-ray structure which may suggest the formation of a non-native backbone conformation in this region. Nevertheless, the subset appears to show an improved agreement with experiment.

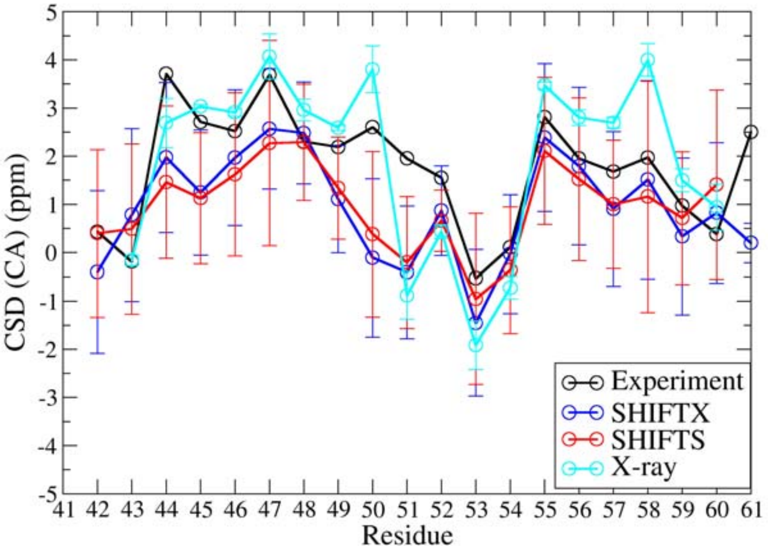


Figure 4-10. Comparison of calculated and experimental C_α chemical shift deviations from a set of structures containing the phenylalanine core at 286 K. Experimental values are shown in black. Two Karplus parameter sets were used for the scalar coupling calculation. Some residues are missing because either the J-coupling constants were not measured or the results were ambiguous. The standard deviation was shown for each calculated scalar coupling constant. The chemical shifts were also calculated for the X-ray structure (cyan).

Table 4-1. RMS values of the calculated shifts compared to the experimental shifts.

C_α CSDs	SHIFTS(ppm)	SHIFTX(ppm)
Ensemble (Collapsed)	1.67	1.72
Ensemble (Fold)	1.64	1.58
Set A	0.91	0.93

4.4 Conclusions

In this chapter, we studied the HP21 ensemble generated by REMD in explicit solvent and characterized the ensemble's structural properties. DSSP and RMSD analysis suggests there is a greater preference for structure in the region of the native helix-1 as compared to native helix-2 of HP36. The additional residues compared to HP-1, the isolated fragment, have little effect on the formation of additional secondary structure. In that region, there is actually a decrease in the helicity in the larger fragment (26 % in HP21 and 21 % in HP-1) [79]. Compared to isolated fragment HP-2, HP21 has a similar amount of structure in the region of native helix-2 (approximately 8 % in both fragments). These trends are different that the experimental trends which show an increase in helical content [107]. Our next focus was the formation of the phenylalanine core in HP21. All three phenylalanine residues are present in HP21 which may allow for the formation of a phenylalanine core similar to the native state of HP36. In the HP21 ensemble, there appears to be a small part of the ensemble which indeed forms this core (3 % of the structures from both simulations). Based on these simulations, there appears to be the formation of native-like structure in this fragment.

To evaluate the quality of the ensemble, we calculated $^3J(H_NH_\alpha)$ scalar coupling and Cα CSDs and compared with experimental values. Based on this comparison, the populations appear to be quite different from the experiment which suggests the formation of more α-helical structure in regions of helix 1 and helix 2 in the native structure of HP36. We selected a subset of the ensemble based on the formation of two contacts based on previous NOE data in order to test. If both contacts are formed, it should form phenylalanine core similar to the native state. This improved the agreement between calculated and experimental observables in the more helical regions of structure. Nevertheless, the region between the two native helices showed the biggest deviation from the experiment which suggests possible non-native behavior in this region.

There are a few possible reasons for the deviations of the ensemble from the experiment. REMD is a useful method nevertheless it has its disadvantages especially in explicit solvent. One major issue is convergence of all the replicas especially in water where the number of replicas is higher and issues with viscosity are present. Our results suggest convergence from the DSSP analysis, however as we look at other features such as contact formation, the uncertainty becomes greater. In addition, there are issues with using high temperatures for enhanced sampling. Folding can exhibit non-Arrhenius behavior. Work by Levy et al. [75] as well as several experimental studies [76-78] have shown that the temperature dependence of folding decreases after a certain temperature. Therefore, it would be useful to test a reservoir approach to this problem. Molecular dynamics simulations would be performed at an optimum temperature (where the folding rate is still temperature dependent) for sampling in order to obtain a converged Boltzmann weighed ensemble. This would be followed by REMD where exchanges could be made between the reservoir and the replica with the highest temperature. This approach is discussed in further detail in Chapter 5.

Another reason for the lack of agreement with experiment is issues with the ff99SB force field. Recent findings from our lab have shown that there are issues with the helical content in marginally stable peptides. Chapter 6 explores the possible problems with the current force field. The main problem appears to be the φ torsional potential that has caused the barrier between the PP_{II} and β basin to be too small. This would affect the helical basin which thus may not be as populated as it should be.

Another possible cause for these deviations could be weak hydrophobic effect. This could hinder the burial of key hydrophobic contacts which could lead to more structure in the backbone. The TIP3P water model, however, has shown reasonable agreement between calculated and experimental solvation free energies of small non-polar compounds [60, 124]. In Chapter 7, there will be a further discussion of the effects of water models.

Several strategies could be used to continue this project. The development of a better backbone potential would allow for this project to be revisited with a better force field. Another approach would be to use the current ensemble of structures with a reweighing scheme for the populations. This would emphasize the agreement of the populations with experimental values. A similar approach called ENSEMBLE uses a similar reweighing scheme with structures generated from high temperature simulations [182, 183]. While there are many other possible solutions to this problem, they should focus on altering the helical content of the ensemble for better agreement with experimental observables.

5. Improved Conformational Sampling in Explicit Solvent: Application of Reservoir Replica Exchange Molecular Dynamics to Small Peptide Systems

Abstract

One of the greatest challenges for simulations of biomolecules is sampling the entire free energy landscape. Folding and unfolding events occur on a slow time scale that is typically not accessible by simulations under biologically relevant conditions. Replica exchange molecular dynamics (REMD) can often overcome these obstacles by using high temperatures to facilitate escape from kinetic traps. However, obtaining converged data with REMD remains a challenge, especially for large systems with complex topologies or simulations in explicit solvent. A relatively new method, Reservoir REMD (R-REMD) improves efficiency by allowing the REMD simulations to exchange conformations with a pool of structures that were previously generated at high temperature. This can decouple the slow conformational search from the expensive simulation of many replicas, as compared to the typical approach of simulating all replicas during the time that only the high temperature simulations are effectively exploring phase space. The reservoir approach has been shown to be beneficial in simulations in the gas phase or using implicit solvent. Nevertheless, many current implicit solvent models have been shown to cause secondary structural bias and overstabilized ion-pair effects between charged residues. Here, the R-REMD approach is applied to two model peptides in explicit solvent for which we were also able to obtain

converged ensembles at 300 K using standard REMD. It is shown that coupling to the high temperature reservoir results in low temperature R-REMD ensembles that are in excellent agreement with results from standard REMD. This suggests that structure reservoirs can be successfully exploited even with periodic systems in explicit water, which are known to be problematic with standard REMD.

Acknowledgments

This chapter contains direct excerpts from the manuscript (presented with minor modifications) written by Lauren Wickstrom with suggestions and revisions from Professor Carlos Simmerling and Dr. Daniel Roe. Fangyu Ding assisted in running and analyzing simulations of HP-1. Dr. Asim Okur and Vadim Patsalo contributed simulation data for Ala10.

5.1 Introduction

Adequate conformational sampling in molecular dynamics simulations remains a major obstacle for obtaining accurate structural populations for biomolecules at equilibrium. Local minima can restrict the movement along the free energy terrain and leave large areas of unexplored conformational space. Several reviews have discussed the recent progress and remaining challenges of conformational sampling [64, 184].

One popular approach for overcoming insufficient sampling in simulations is the replica exchange method (also known as parallel tempering) [62, 63, 185-188]. In temperature replica exchange molecular dynamics (REMD) [62], multiple non-interacting simulations are performed for the same system, which are coupled to thermostats at different temperatures. Periodically, an exchange is attempted between neighboring replicas using a Metropolis-type criterion. Typically, replicas range from experimentally accessible and biologically relevant temperatures to higher temperatures such as 600 K. Through exchanges with a high temperature replica, lower temperature simulations can escape kinetic traps allowing for the system to reach equilibrium faster than normal MD. Furthermore, the transition probability is formulated to ensure that canonical ensemble properties for each replica are maintained, which in turn results in correct temperature dependent observables (within the limits of the model). Many different groups have applied REMD to studies of peptide and small protein folding [53-55, 62, 63, 79, 80, 149, 153, 154, 168].

Nevertheless, REMD simulations do suffer from some major drawbacks. High temperatures may not be beneficial for the conformational search, especially in cases where the temperature dependence of the folding rate is weak or even non-Arrhenius [75]. Thus, simulations started from non-native conformations may struggle to find the native state even at higher temperatures. Another issue is that once the replica does sample a favorable low energy structure, it is exchanged to lower temperatures and the conformational search must begin again. This becomes problematic since multiple folding events are necessary to achieve the correct population of folded structures below the melting temperature of a peptide. Implicit solvent simulations require fewer replicas than corresponding simulations in explicit solvent, and folding events also occur more frequently, leading to better REMD convergence. Lastly, the exchange probability is derived under the assumption that structures being considered for an exchange are already Boltzmann weighted. This is not true in the beginning of the REMD simulations and the exchange function must play a role in driving the simulations towards convergence.

Previously, we and others [82, 84, 189] have developed the Reservoir-REMD (R-REMD) method to help overcome these issues with normal REMD. Similar to the J-walking method [190], an ensemble is generated at one high temperature for the reservoir, where convergence is more rapid but the desired thermodynamic ensemble is not obtained due to temperature. Standard REMD is subsequently run below this temperature, providing an annealing ladder to optimize reservoir structures and re-weight the high-temperature ensemble. Advantages of this method are that the simulations start with the correct exchange criterion due to the Boltzmann weighting of the reservoir and there is no reliance on folding events within the replicas themselves. Individual structure in the reservoir can seed multiple MD replicas, meaning that fewer folding events are needed. This is especially important in explicit solvent, where many replicas are required due to system size, and thus many folding events are required to populate the replicas below the thermal transition temperature in standard REMD. Successful application of this method has been shown using the trpzip2 β-hairpin and the dPdP three stranded antiparallel β-sheet, both in implicit solvent. R-REMD simulations were more efficient and resulted in highly accurate melting profiles and absolute populations of structures compared to standard REMD [82]. In similar work, Li et al. [84] performed finite reservoir replica exchange method (FRREM), a version of Hamiltionian REMD (H-REMD), where the reservoir was collected using a scaling parameter of $\lambda = 0.1$ which was subsequently coupled to a production run using a scaling parameter of $\lambda = 1$. On a test case, FRREM was shown to be five times more efficient at sampling than normal H-

REMD on a butane-like molecule in the gas phase. Lyman et al. [189] developed the resolution exchange method in which the reservoir was made up of coarse grained structures and exchanges were performed with all-atom REMD simulations. This approach decreased the computer time by 15-fold on a dileucine peptide in implicit solvent.

Despite these impressive results, implicit solvent models suffer from inaccuracies due to the approximation of the free energy of solvation. Implicit models such as the semi-analytical generalized Born model (GB) [51] are attractive because they are computationally less expensive and can converge more rapidly than simulations in explicit water due to lack of solvent viscosity. While GB has been widely used for protein folding, we and others have reported weaknesses such as secondary structural bias and the overstablization of ion pairs [53-56, 59]. On the other hand, explicit solvent models are essential, particularly in cases where water has non-bulk properties and interacts directly with the solute in such cases as bridging water. Previously, we have shown that explicit solvent is necessary to obtain structural populations for short model peptides in qualitative agreement with experiment [58, 79].

One of the remaining challenges for REMD is efficient application to systems in explicit solvent. The number of solvent waters plays a role in increasing the computational expense of REMD. REMD rapidly becomes computationally unfeasible, because the number of replicas needed to span a given temperature range increases with the square root of the number of degrees of freedom in the system [63, 71]. Furthermore, solvent viscosity slows the conformational search for all of the involved replicas, thus requiring long simulations for many replicas. To our knowledge converged REMD simulations in explicit solvent from independent starting conformations have only reported for short helical or unstructured peptides [58, 79].

We believe that modified REMD approaches, such as R-REMD, will be crucial for overcoming the problems associated with performing REMD in explicit solvent. The decoupling of conformational search and reweighing aspects of REMD may avoid the costly simulation of large numbers of replicas in explicit solvent during the time that only the highest temperature replicas are effectively sampling the energy landscape. By using R-REMD, only one temperature would be simulated for the long time required to sample the landscape in explicit water. The others temperatures would be run subsequently, and use the information gained from extensive search at a single temperature to more rapidly obtain the temperature dependence of the ensemble using the efficient Monte Carlo of REMD and the conformational annealing of the temperature ladder.

Although this approach seems reasonable, R-REMD has not been demonstrated in explicit solvent. It is important to first ensure that accuracy is maintained, particularly since the reservoir is a small subset of the actual ensemble, and the potential energies used in the exchange with the reservoir are likely dominated by the solvent, rather than the solute of interest. In the present study, we have validated this new application of R-REMD to two small model peptides in explicit water, using as a reservoir a limited subset of the overall high-temperature ensemble. These peptides were selected because our previous simulations have shown excellent convergence of their populations from two independent runs, providing a precise data set against which the R-REMD data can be judged. Our first test case was an alanine polypeptide containing 10 residues (Ala10) using R-REMD in explicit solvent. Previous studies have shown that the use of GB models with Ala10 results in an overestimation of helical content compared to the explicit solvent simulations which primarily sample PPII conformations [58, 59, 80], in agreement with experimental findings for short Ala peptides [44, 141, 191-193]. These R-REMD simulations were then compared to the corresponding standard REMD simulations, obtaining excellent agreement with absolute populations and structural properties between REMD and R-REMD.

To further validate our approach, we applied the R-REMD method in explicit solvent to a larger and more complex system, HP-1, corresponding to the isolated first helix of the villin headpiece helical subdomain [26, 79]. Previous work on HP-1 showed that GB simulations incorrectly sample large populations of α-helix and a salt bridge interaction as compared to explicit solvent simulations [79]. We calculated secondary structure content, structural properties (RMSD, Rg), absolute populations of cluster families and melting curves, and compared these results with the standard REMD simulations. The observable distributions and the sequence dependent secondary structural trends of the REMD and R-REMD runs in explicit water are quite similar and the populations of various conformational families also demonstrate a high degree of similarity between the methods ($r = 0.890$). The melting curve of the R-REMD simulations was quite consistent with melting profiles of independent REMD simulations. Future work will investigate applying this sampling method to larger systems in explicit solvent where convergence of REMD and reservoirs are more challenging.

5.2 Methods

HP-1 (M41LSDEDFKAVFGM53) corresponds to the N-terminal helix of HP36.

HP-1 has a free N-terminus and an amidated C-terminus in accordance with previous experimental and computational studies [26, 79]. We adopt the typical numbering convention for HP36, in which L42 follows the N-terminal methionine [87, 88]. All sidechains for Asp, Glu, Lys, and Arg were charged during the simulation. Ala_{10} was acetylated and amidated at the N and C termini respectively. All calculations employed Amber version 9 [39] and used the ff99SB modification [45] of the Amber ff99 force field [120, 121]. SHAKE [122] was used to constrain bonds to hydrogen. The time step was 2 fs. Temperatures were maintained using weak Berendsen coupling [123]. Explicit water simulations were performed with the TIP3P [49] water model, truncated octahedral periodic boundary conditions and particle mesh ewald (PME) [125] to calculate long-range electrostatic interactions. Simulations were run in the NVT ensemble.

5.2.1 Explicit water reservoir REMD (R-REMD)

Reservoir REMD simulations (R-REMD) were run using the same simulations parameters and temperatures as previously published for the standard REMD simulations of these peptides [58, 79]. For Ala10, 20 replicas were used ranging from 267 to 387 K (267, 272, 278, 283, 289, 294, 300, 306, 312, 318, 324, 331 337, 344, 351, 358, 365, 372, 379 and 387 K) while using 394 K as a reservoir in similar fashion to our previous work [58]. This simulation was started from a collapsed structure and run for 36 ns. For HP-1, 18 replicas were used ranging from 276 to 391K (276, 282, 288, 294, 300, 306, 313, 320, 327, 334, 341, 348, 355, 363, 371, 379, 387, and 395 K) while using 404 K as a reservoir. Two independent simulations were run 18 ns for each replica using all replicas in a collapsed conformation in one run and folded conformation for the other. Exchange success rates for all simulations are provided as Table 5-1 and Table 5-2, showing that exchanges with the reservoir indeed occur with similar success rates as for the standard replicas.

Table 5-1. Success rate of exchanges for each temperature in the Ala10 R-REMD simulation.

Ala10	**Run 1**
Temperature (K)	**Success rate (%)**
267	25.25
272	25.49
278	24.61
283	26.21
289	24.88
294	25.46
300	25.65
306	25.27
312	26.00
318	25.32
324	24.35
331	25.92
337	25.03
344	25.90
351	24.42
358	25.64
365	26.49
372	26.13
379	26.09
387	24.91

Table 5-2. Success rate of exchanges for each temperature in the HP-1 R-REMD simulations

HP-1	**Run 1**	**Run 2**
Temperature (K)	**Success rate (%)**	**Success Rate (%)**
276	13.81	14.75
282	12.93	13.51
288	13.23	13.27
294	13.26	13.58
300	12.97	12.66
306	13.36	12.98
313	13.75	13.08
320	12.88	13.23
327	12.68	13.61
334	14.80	12.81
341	13.73	12.73
348	13.53	14.98
355	14.62	13.98
363	14.16	13.72
371	13.98	13.79
379	13.77	13.92
387	14.91	14.62
395	13.86	14.37

5.2.3 Generation of reservoir structures

We previously demonstrated that reservoirs could be generated from multiple independent MD runs at high temperature [82]. As discussed above, the present work aims to validate the accuracy of R-REMD with a limited structure set in explicit water, and thus we extracted reservoir structures from our previous standard REMD ensembles of Ala10 and HP-1 at 399 and 404 K respectively. These ensembles were reduced to

10,000 structures by selecting equidistant snapshots from the REMD temperature trajectories. The trajectory files were standard Amber ASCII format and therefore the coordinates have limited precision (0.001 Å). Velocities were not saved in the REMD run, therefore velocities for reservoir structures were assigned from a Maxwell-Boltzmann distribution following each exchange. For these reasons, as well as the small size of the reservoirs as compared to the true ensemble, it is important to investigate the accuracy of the low-temperature ensembles obtained using R-REMD.

5.2.4 Analysis

Cluster analysis was performed with MOIL-View [127], using backbone RMSD as a similarity criterion with average linkage. Selections of backbone regions were between Ala2 and Ala9 for Ala10 and between residues 43 and 51 for HP-1. Clusters were formed with a bottom-up approach using a similarity cutoff of 2.5 Å. Cluster analysis was performed on trajectories combined from standard and reservoir R-REMD simulations, and the normalized populations for each cluster type were calculated for each of the original simulations. The populations of each conformation family were then calculated for the ensemble obtained from first and the second half of the Ala10 simulations and for each of the two simulations starting from different conformations for HP-1 [162].

DSSP analysis [128], end to end distances, RMSDs, and radius of gyration were all done using the ptraj module in Amber. Melting curves were constructed by calculating the average helicity (over time and sequence) for each temperature. Helical residues were selected based on DSSP criterion.

5.3 Results

5.4.1 Ala10

We first tested R-REMD in explicit solvent using Ala10. In Figure 5-1, we show end-to-end distributions at 300 K obtained from standard REMD and R-REMD simulations. Clearly, the structural ensembles exhibit similar broad end-to-end distance profiles ranging from 4 to 25 Å. Both simulations sample structures with the same global properties (within error bars) and appear to have no strong conformational preference, as expected for short polyalanines.

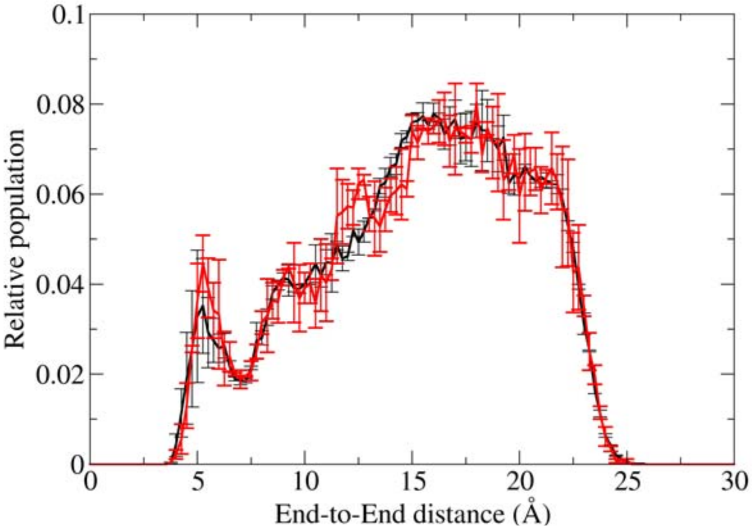


Figure 5-1. Ala10 end-to-end distributions at 300 K obtained with standard REMD (black) and R-REMD(red). Error bars were obtained from 2 simulations for the standard REMD and using the first and second half of the data for R-REMD.

Following our previous published work [58, 79, 82], we evaluated the populations of each cluster to determine whether independent simulations give similar ensembles. All structures from both methods were combined and used to define a common set of families, then the population of each family was computed for each trajectory and compared. This is important because we want to be confident that the populations of each conformational basin are independent of initial coordinates, and REMD results are reproduced by R-REMD even though the reservoir had low precision, lacked velocities and was generated at high temperature. Previously, independent standard REMD simulations of Ala10 showed a high correlation (r = 0.974) suggesting that these simulations are well converged [58]. In Figure 5-2, we compared both the net R-REMD and REMD ensembles and observe remarkable agreement between the cluster populations with the correlation and regression coefficient of 0.986 and 0.972 respectively. It is evident from this analysis that R-REMD samples structural families at 300K in excellent agreement with standard REMD.

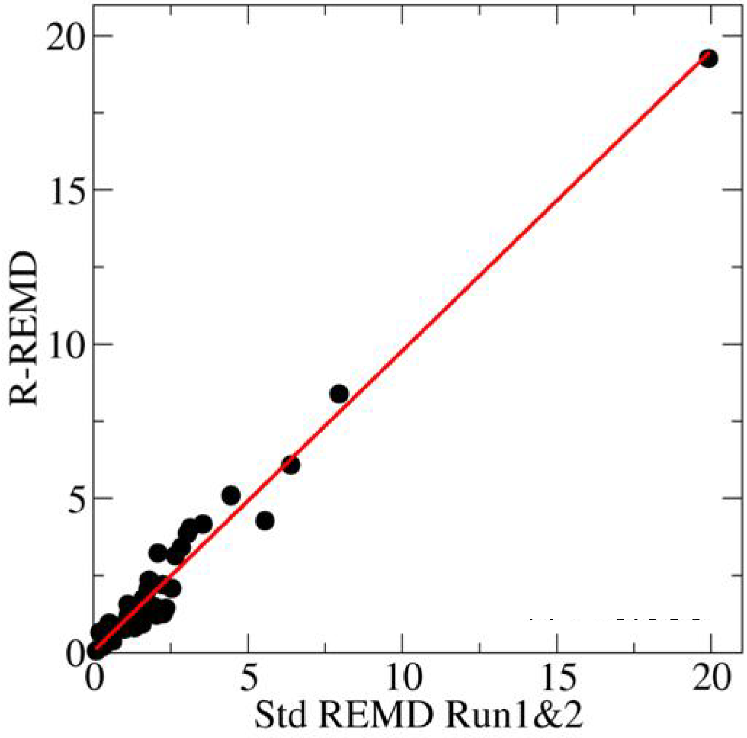


Figure 5-2. Comparison of populations for Ala10 structure familes sampled in different simulations at 300 K. Clusters are defined using the combined data set. Populations in R-REMD and REMD simulations sample very similar populations (r = 0.986).

The most populated cluster of the R-REMD and REMD simulations corresponds to an extended PPII structure, which is consistent with experimental work on short Ala peptides [44, 141, 191-193]. To investigate the efficiency of R-REMD compared to REMD, we monitored the population of this cluster over the duration of the simulation at 300 K (Figure 5-3). In the first 5 ns, all three simulations undergo fluctuations as they approach their equilibrium values. At approximately 2 ns, the R-REMD simulations converge to a population of 20 - 25 %, similar in both value and rate to standard REMD. The fast convergence of the standard simulations suggests that more complex topologies should be studied in the future to explore efficiency gains of R-REMD, and the present study will focus on accuracy of coupling to limited reservoirs generated at high temperature.

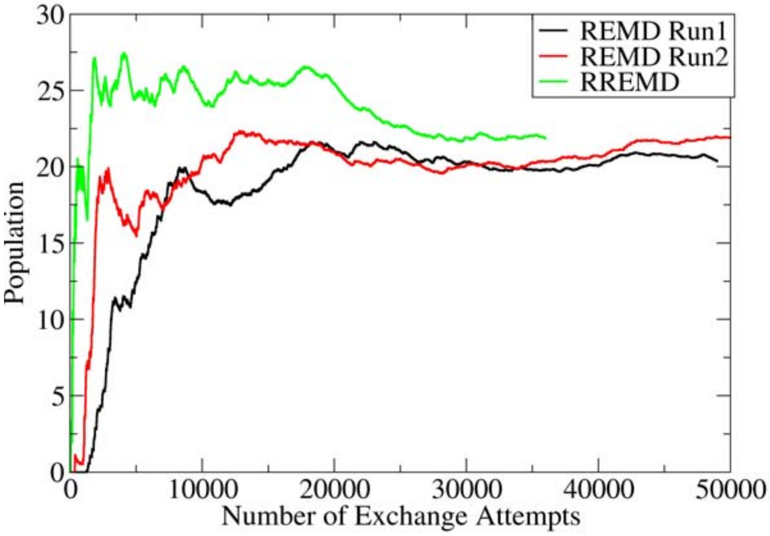


Figure 5-3. Population of cluster corresponding to polyproline II helix as a function of time for REMD simulations, with the REMD simulations in 2 independent simulations in black/red and the R-REMD shown in green at 300 K.

5.4.2 HP-1

To validate this approach on a more complex system with non-trivial sidechains, we applied this method to HP-1, corresponding to the isolated first helix of HP36. We have previously shown with standard REMD that HP-1 contains a modest amount of helical structure in the region where the native helix is located in full sequence of HP36 [79]. In addition, we obtained very similar final ensembles starting from two different initial structures with REMD. Therefore, the ensemble is converged and suitable for the validation of the R-REMD approach. Here, we compare those results to new two simulations performed using R-REMD, starting from two different folded and collapsed conformations.

Figure 5-4A shows RMSD distributions relative to the backbone of the NMR structure [88] of full length HP36, between residues 43 and 49. Both REMD and R-REMD ensembles have similar RMSD distributions ranging from 0.5 to 4.5 Å with the most populated region centered around 1.1 Å. R-REMD simulations are again observed to be within the error bars of the REMD simulations. In Figure 5-4B, radius of gyration (Rg) distributions are shown for both the REMD and R-REMD simulations. Both sample a range of structures with an Rg between 6.0 and 12 Å and contain their highest structural populations between 7.0 and 7.5 Å. Dictionary of secondary structural prediction (DSSP) [128] analysis was employed to characterize the secondary structure (Figure 5-5). For both REMD and R-REMD, secondary structure profiles demonstrate a high α-helical propensity in the center of the fragment. This is in the same region as the first α-helix occurring in the NMR and X-ray structures of HP36 (D44 – K48 in the NMR structure)

[88, 89]. Overall, R-REMD simulations reproduce similar structural observables compared to REMD simulations.

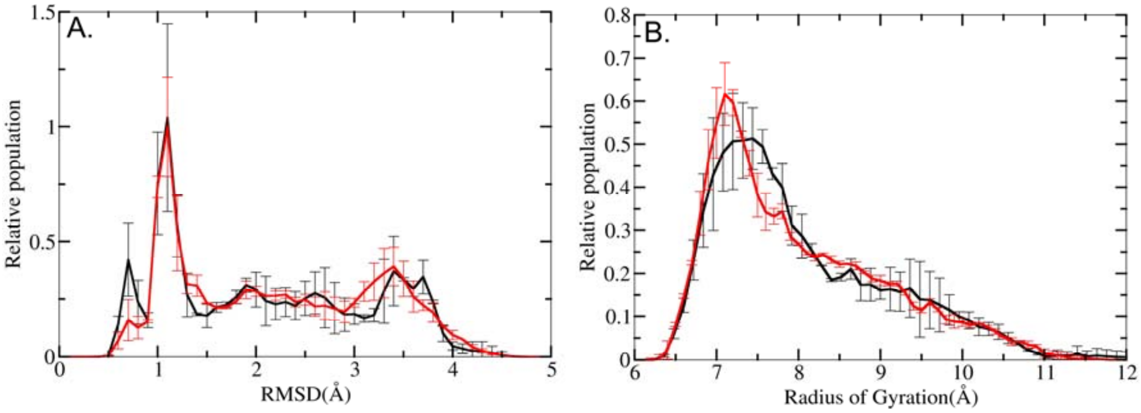


Figure 5-4. Histogram of the (A) RMSD of the backbone from residues 43 to 49 and (B) Radius of gyration for the R-REMD(black) and standard REMD (Red) simulation. Error bars are obtained from two independent simulations.

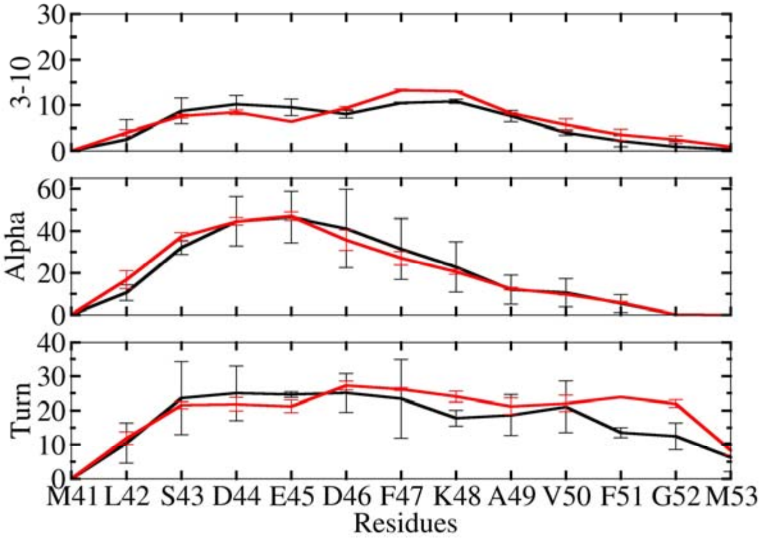


Figure 5-5. DSSP analysis of HP-1 as a function of sequence for the standard REMD (black) and R-REMD (red) simulations. Error bars are obtained from two independent simulations. Each secondary structure profile of the R-REMD simulation overlaps well with the standard REMD simulation.

Similar to our analysis on Ala10, we compared the populations of cluster families to evaluate how precise our results were for the standard REMD and R-REMD, and how accurate the R-REMD results were compared to regular REMD. Standard REMD simulations starting from different initial conformations showed a high correlation between cluster populations using only backbone residues from 43 to 49 at 300 K (r = 0.994) [79]. This region was originally selected because it contained the most helical structure. For our current studies, we selected the larger backbone region (between residues 43 and 51) to perform our cluster analysis to ensure that the flexible N- and C-terminal ends of the HP-1 fragment were converged as well as the middle part of the sequence. In order to reduce bias of a dominant conformer on the correlation statistics, we analyzed populations in a higher temperature ensemble (355 K). Standard REMD simulations again demonstrated a high correlation between families of structures (r = 0.890) (Figure 5-6A). The two independent R-REMD simulations showed exceptional agreement with r = 0.968 and a slope of 1.024 (Figure 5-6B). Both sets of simulations demonstrate reliable precision of their populations of structures. Using these precise ensembles obtained by REMD and R-REMD, we observe very good agreement between the REMD and R-REMD data sets (r = 0.897 (0.791 without the biggest cluster), comparable to the 0.890 obtained comparing individual standard REMD runs). These results indicate that R-REMD performs well at reproducing the ensembles obtained from standard REMD.

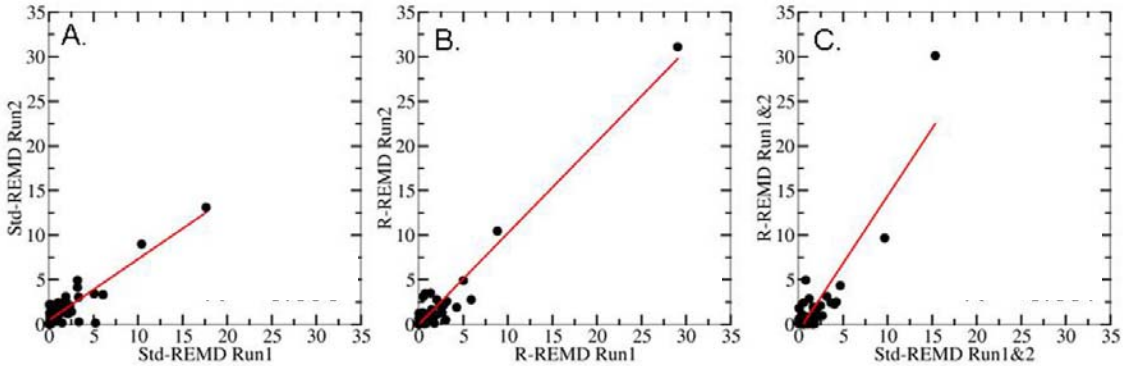


Figure 5-6. Comparison of a set of HP-1 structures sampled in different simulations at 355 K. (A) Comparison of standard REMD from folded vs standard REMD from extended (r = 0.890). (B) Comparison of R-REMD from different initial structures (r =0.968). (C) Comparison of the combined data of the standard REMD and the combined data of the R-REMD (r = 0.897).

To compare the accuracy of the temperature dependence of the R-REMD simulations compared with the REMD, we calculated melting curves for the standard REMD and R-REMD simulations. We calculated the average helicity through DSSP analysis and compared the helical content of the fragment at each temperature (Figure 5-7). Overall, the melting curves exhibit highly similar profiles, with the two methods providing essentially identical results within the (small) error bars. The size of the error bars also appear to decrease the closer the temperature trajectory is to the reservoir. This is expected since higher temperatures exchange with the reservoir earlier in the simulations than the lower temperatures and will converge faster. Nevertheless, the R-REMD demonstrates excellent convergence at multiple temperatures and is essentially in quantitative agreement with R-REMD along the full temperature range.

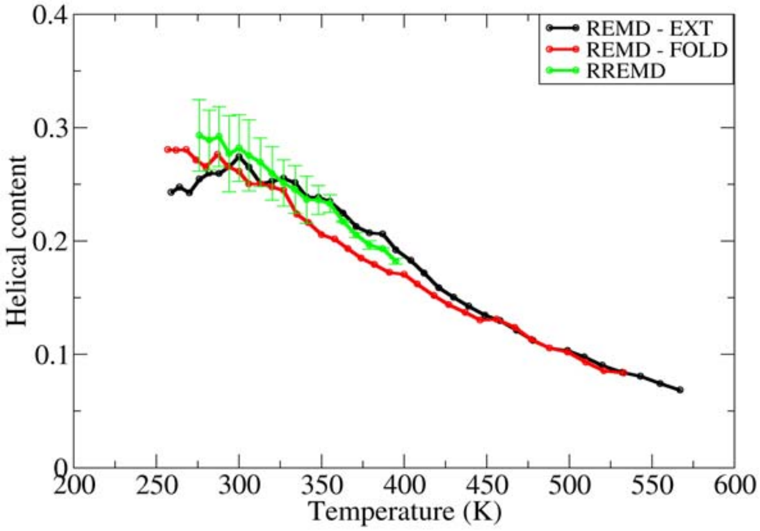


Figure 5-7. Comparison of melting profiles for the HP-1 fragment. The R-REMD simulations exhibit similar melting behavior to the REMD simulations starting from the extended and folded conformation respectively. Two melting curves are shown for the standard REMD simulations because those two runs used slightly different temperature sets.

5.5 Conclusions

We implemented the R-REMD approach with explicit solvent using two small peptide systems, Ala10 and HP-1, for which well converged results have been obtained with standard REMD. For the reservoir, we used a temperature trajectory of one of the standard REMD simulations closest to 400 K. The goal of this work was to demonstrate that this method is able to obtain accurate results compared with standard REMD. Ala10 was run for 36 ns using 20 replicas and HP-1 was run for 18 ns for each replica, repeated from two initial structures. R-REMD simulations sample similar structural properties as the standard REMD simulations. We also achieve excellent precision for R-REMD simulations when comparing either the first and second half of the simulation or independent simulations. There is also a high correlation between the absolute structural populations of REMD and R-REMD ensembles, indicating good accuracy. This is expected since the method is formally rigorous, within the restrictions of a reservoir that represents an incomplete subset of the actual ensemble, the lack of velocities in the reservoir and the limited precision of the reservoir coordinate files.

This implementation is an example of the continued progress in enhanced sampling methods. In explicit solvent, the R-REMD approach is beneficial because it requires only one converged ensemble at high temperature. This is especially important in explicit water where folding events are slow and solvent viscosity impedes the conformational search. Standard REMD requires multiple folding events, while R-REMD uses one converged reservoir to perform the sampling which is similar to pseudoexchange simulations [189]. In this report, R-REMD is slightly more efficient in sampling; however, these are small systems and may not benefit from these methods as much as larger complex peptides since they contain a relatively low amount of structure and quickly reach equilibrium.

There are still remaining challenges with obtaining a Boltzmann weighted reservoir for systems that require microseconds and beyond to achieve folding events, even at high temperature. A promising approach might be to generate a reservoir using implicit solvent since decreased viscosity may facilitate the crossing of barriers that can dominate relaxation times for the system in explicit water [72]. Likewise, conformations arising from a structure prediction protocol could be used, with explicit solvent added. The reservoir would not represent a Boltzmann weighted population due to the change in Hamiltonian and representation; thus one would need to employ a correction to the exchange calculations using our non-Boltzmann reservoir REMD method [194]. In the present report, we have demonstrated for two non-trivial peptides that use of a limited set of high-temperature structures in explicit solvent is practical and provides accurate results at low temperatures of interest, paving the way for such future developments.

6. Evaluating the Performance of the FF99SB Force Field Based on NMR Scalar Coupling Data

Abstract

Force field validation is essential for the identification of weaknesses in current models and development of more accurate models of biomolecules. NMR coupling and relaxation methods have been used to effectively diagnose strengths and weaknesses of many existing force fields. Studies using the ff99SB force field have shown excellent agreement between experimental and calculated order parameters and residual dipolar calculations. Nevertheless, recent studies have suggested that ff99SB demonstrates poor agreement with J-coupling constants for short polyalanines. We performed extensive replica exchange molecular dynamics simulations on Ala_3 and Ala_5 in TIP3P and TIP4P-Ew solvent models. Our results suggest that the performance of ff99SB is among the best of currently available models. In addition, scalar coupling constants derived from simulations in the TIP4P-Ew model show a slight improvement over the ones using the TIP3P model. Despite the overall excellent agreement, the data suggest areas for possible improvement.

Acknowledgments

This chapter contains direct excerpts from the manuscript (presented with minor modifications) written by Lauren Wickstrom with suggestions and revisions from Professor Carlos Simmerling. Dr. Asim Okur contributed simulation data for Ala_3. We

thank Dr. Gerhard Hummer and Dr. Robert Best for J-coupling scripts and feedback, and Dr. Adrian Roitberg for helpful discussions.

6.1 Introduction

A significant challenge in the use of computation to study complex biomolecular systems is force field accuracy. Force fields are made up of a molecular mechanics (MM) energy function with empirical parameters, which are typically obtained from fitting to experimental or high level quantum mechanical (QM) data. These approximations may lead to inaccuracies in calculated kinetic and/or thermodynamic properties. ff94 [120] is one of the examples, with a strong bias favoring helical content. While not always apparent in short simulations, ff94 leads to overstabilization of helical systems and the adoption of stable helices for sequences that have non-helical experimentally determined structures [37]. Even in cases where the force field matches well to the QM data that was used in parameter development, errors can arise from inconsistencies in the model. For example, many non-polarizable force fields employ partial charges that are intended to reproduce the enhanced dipoles found in aqueous solution, yet the dihedral potentials are fit to reproduce gas-phase QM energy profiles using these charges. These effects, combined with the relatively small size of the systems used for parameter development indicate that validation against experimental data is vital.

ff99SB was developed to improve the secondary structure balance of the previous AMBER protein force fields and also to improve the description of glycine residues [45]. Although the parameters were fit using QM data, the development relied on the validation of candidate parameters against experiment. Decoy sets of helical peptides, hairpins and miniproteins demonstrated the correct energy minima. Calculated NMR order parameters for ubiquitin and lysozyme also showed better agreement with experiment than previous force fields. Showalter et al. [47] demonstrated that ubiquitin dynamics as measured by residual dipolar couplings obtained from ff99SB simulation are "comparable to or better than the best static structural models and the NMR ensemble". Other work has shown similarly good agreement between ff99SB simulations and NMR structural and relaxation data [46, 79, 195, 196]. Overall, these studies have suggested that ff99SB is in at least reasonable agreement with experiment for a variety of proteins.

One disadvantage of these studies on complex systems is that it can be difficult to decompose inaccuracies into the specific force field terms that need improvement. Short polyalanines have become useful simple model systems for studying conformational variability of unfolded states where structural preferences are weak and therefore the system is highly sensitive to small inaccuracies [142, 197-199]. A recent study by Graf et al. [44] of Ala_n ($n = 3$ to 7) showed that significant differences exist between the experimental and calculated J-coupling constants from unweighted simulation data. Building on this availability of extensive experimental data, Best et al. [43] performed a follow up study on Ala_5 using variations of the AMBER [39], CHARMM [40], GROMOS [42] and OPLS [41] force fields using various sets of Karplus parameters to calculate the scalar coupling constants. Force fields were evaluated using a χ^2 value, which calculated the sum of deviations of each calculated J-coupling constant compared to the experimental values, normalized by a factor related to the assumed systematic error in the coupling constant calculations. Among the parameter sets tested, ff99SB was ranked among the worst for this data set. An erratum [200] corrected key aspects of the data, with the result that the ff99SB ranking significantly improved. We present here a more detailed analysis of ff99SB performance using two water models and different length peptides.

We performed replica exchange molecular dynamics simulations [62, 63] for 50 ns/replica of Ala_3 and Ala_5 in two explicit water models. Precision was quantified using fully independent simulations from different initial structures. Our studies address the performance of the ff99SB force field, compare the effects of using different water models, and suggest improvements to ff99SB that may improve agreement with experiment.

6.2 Methods

6.2.1Simulation details

We simulated Ala_3 and Ala_5 with a free N- and protonated C-terminus. These sequences and termini correspond to conditions used in the experimental studies [44]. All simulations were performed in Amber version 9 [39] and used the ff99SB [45] force field. SHAKE [122] was used to constrain bonds to hydrogen. The time step was 2 fs. Temperatures were maintained using weak Berendsen coupling [123]. Explicit water simulations were performed in a truncated octahedron box with the TIP3P [49] and TIP4P-Ew [50] water models. Simulations were run in the NVT ensemble and particle mesh Ewald [125] was used to calculate long-range electrostatic interactions.

6.2.2 Ala_3

For both water models, an extended structure of Ala_3 was solvated with approximately 500 water molecules (498 for TIP4P and 525 for TIP3P). The structures were equilibrated at 300 K for 50 ps with harmonic restraints on solute atoms. The restraints were reduced from 5 kcal/mol*Å to 1 kcal/mol*Å to 0.5 kcal/mol*Å. After minimization, three 5 ps MD simulations were performed with the same gradually reduced restraints at constant pressure (1 atm) and temperature (300 K) to generate starting structures.

To improve sampling, we used replica exchange molecular dynamics [62, 63] as implemented in Amber 9. The target exchange acceptance ratio for all simulations was approximately 20 % between temperatures ranging from 291 – 580 K (291, 300, 310, 320, 330, 340, 351, 362, 374, 386, 398, 411, 424, 438, 451, 466, 481, 496, 512, 528, 545 and 562 K). Exchanges between neighboring temperatures was attempted every 1 ps. In order to evaluate convergence, an additional simulation was run using a structure which started from an α-helical conformation in the 2nd residue. The simulations were run for 50 ns exchange attempts. The first 5 ns of each simulation was discarded.

6.2.3 Ala$_5$

For both water models, an extended structure of Ala$_5$ was solvated with 891 water molecules. The structures were equilibrated at 300 K for 50 ps with harmonic restraints on solute atoms. The restraints were reduced from 5 kcal/mol*Å to 1 kcal/mol*Å to 0.5 kcal/mol*Å. After minimization, three 5 ps MD simulations were performed with the same gradually reduced restraints at constant pressure (1 atm) and temperature (300 K) to generate starting structures. REMD simulations were run using a target acceptance ratio of approximately 20 % between the temperatures 293 to 415 K (293, 300, 307, 314, 322, 329, 337, 345, 353, 361, 370, 378, 387, 396, 406 and 415 K).

Exchanges between neighboring temperatures were attempted every 1 ps. In order to evaluate convergence, we ran an additional simulation starting from an α-helical conformation of Ala_5. Both simulations were run for 50 ns. The first 5 ns of each simulation was discarded.

6.2.4 Analysis details

6.2.4.1 Karplus Parameter Details

Equation 6.1 was used for the calculation of the J coupling constants.

$$J(\theta) = A\cos^2(\theta + \Delta) + B\cos(\theta + \Delta) + C$$

Equation 6-1. Karplus equation.

A, B, C and Δ are listed in Tables 6-1 through 6-3 except in the case of $^3J(H_{N,}C_\alpha)$ which uses equation 6-2.

$$^3J(H_{N,}C_\alpha)\ (\varphi_i,\psi_{i-1}) = -0.23 \cos \varphi_i - 0.20 \cos \psi_{i-1} + 0.07 \sin \varphi_i + 0.08 \sin \psi_{i-1} + 0.07 \cos \varphi_i \cos \psi_i + 0.12 \cos \varphi_i \sin \psi_{i-1} - 0.08 \sin \varphi_i \cos \psi_{i-1} - 0.14 \sin \varphi_i \sin \psi_{i-1} + 0.54$$

Equation 6-2. Equation used to calculate the $^3J(H_{N,}C_\alpha)$ scalar coupling constant.

These calculations were done comparably to the work by Graf et al. and Best et al. [43, 44].

Table 6-1. Original (“Orig”) parameters used in the Karplus equation [44] from Graf et al. [44].

Coupling	Torsion	A (Hz)	B (Hz)	C (Hz)	Δ(°)
$^3J(H_N,H_\alpha)$	φ_i	7.09	-1.42	1.55	-60
$^3J(H_N,C')$	φ_i	4.29	-1.01	0.00	180
$^3J(H_\alpha,C')$	φ_i	3.72	-2.18	1.28	120
$^3J(C,C')$	φ_i	1.36	-0.93	0.60	0
$^3J(H_N,C_\beta)$	φ_i	3.06	-0.74	0.13	60
$^1J(N,C_\alpha)$	ψ_i	1.70	-0.98	9.51	0
$^2J(N,C_\alpha)$	ψ_{i-1}	-0.66	-1.52	7.85	0

Table 6-2. “DFT1” parameters used in the Karplus equation [201]. Parameters for unlisted J coupling constants used parameters in S1.

Coupling	Torsion	A (Hz)	B (Hz)	C (Hz)	Δ(°)
$^3J(H_N,H_\alpha)$	φ_i	9.44	-1.53	-0.07	-60
$^3J(H_N,C')$	φ_i	5.58	-1.06	-0.30	180
$^3J(H_\alpha,C')$	φ_i	4.38	-1.87	0.56	120
$^3J(C,C')$	φ_i	2.39	-1.25	0.26	0
$^3J(H_N,C_\beta)$	φ_i	5.15	0.01	-0.32	60

Table 6-3. “DFT2” parameters used in the Karplus equation [201]. Parameters for unlisted J coupling constants used parameters in S1.

Coupling	Torsion	A (Hz)	B (Hz)	C (Hz)	Δ(°)
$^3J(H_N,H_\alpha)$	φ_i	9.14	-2.28	-0.29	-64.51
$^3J(H_N,C')$	φ_i	5.34	-1.46	-0.29	172.49
$^3J(H_\alpha,C')$	φ_i	4.77	-1.85	0.49	118.61
$^3J(C,C')$	φ_i	2.71	-0.91	0.21	-2.56
$^3J(H_N,C_\beta)$	φ_i	4.58	-0.36	-0.31	58.18

Phi and psi dihedrals for the central residue of the Ala peptides were calculated using the ptraj module in Amber 10 [39] .

6.2.5 Error analysis

The agreement between the experimental and calculated constants was evaluated using the equation 6-3, following the procedure previously reported [43].

$$\chi^2 = \mathrm{N}^{-1} \sum_{j=1}^{N} (\langle J_j \rangle_{sim} - J_{j,\exp})^2 / \sigma_j^2$$

Equation 6-3. Equation used for error analysis.

$\langle J_j \rangle_{sim}$ is the average coupling constant j obtained from the simulation while $J_{j,\exp}$ is the experimental coupling constant for J. The average was calculated using the scalar coupling constants $^3J(H_N,H_\alpha)$, $^3J(H_N,C')$, $^3J(H_\alpha,C')$, $^3J(C,C')$, $^3J(H_N,C_\beta)$, $^1J(N,C_\alpha)$, $^2J(N,C_\alpha)$, and $^3J(H_N,C_\alpha)$ where N is the total number of J values. The systematic error σ_j was included to account for possible substituent effects neglected in the Karplus equation for each coupling constant (Table 6-4). The estimates in Table 6-4 of this document were used for this work. We note that these are identical to those used by Best et al. in reference [43] but that they do not match the values provided in Table 6-4 of that publication (G. Hummer, pers. comm.).

Table 6-4. Estimates of errors σ_j for each scalar coupling reported in Best et al. [43].

Coupling	$\boldsymbol{\sigma_j}$
$^3J(H_N,H_\alpha)$	0.91
$^3J(H_N,C')$	0.59
$^3J(H_\alpha,C')$	0.38
$^3J(C,C')$	0.22
$^3J(H_N,C_\beta)$	0.39
$^1J(N,C_\alpha)$	0.59
$^2J(N,C_\alpha)$	0.50
$^3J(H_N,C_\alpha)$	0.10

6.2.6 Populations of secondary structure for the central residue of Ala$_5$

Populations of secondary structure were calculated using the basin definitions in the previous work [43]. Secondary structure basin populations for central residues were calculated based on phi/psi dihedral angle pairs. The definitions of the four principle regions were as follows: right handed helix (α_R), (φ,ψ) ~ (-160 to -20, -120 to +50); extended β-strand conformation, (-180 to -110, +50 to +240; or +160 to +180, +110 to +180); and polyproline II, (-90 to -20, +50 to +240). Error bars were constructed from the independent runs. Dictionary of secondary structural prediction (DSSP) [128] analysis was performed by the ptraj module of Amber 10 [39].

6.3 Results

Scalar coupling constants were calculated for each polyalanine simulation with the three Karplus parameter sets identical to the Best et al. study (Table 6-5). We also employed the same χ^2 calculation as Best et al. to evaluate the deviation of the J-coupling constants from the experimental values. For Ala$_3$, the χ^2 values varies from 1.57 to 2.17 depending on the solvent model and the parameter set. The Ala$_5$ J-coupling constants were also quite sensitive to these different variables, but even with the larger peptide size, the χ^2 values remained < 2.0. Importantly, the χ^2 values for Ala_5 were at least as low as any of the other force fields evaluated by Best et al. [43].

Table 6-5. Scalar coupling χ^2 values for Ala_3 and Ala_5 using three different Karplus equation parameter sets.

Peptide	**Water**	**DFT1 [201]**	**DFT2 [201]**	**Original [44]**
Ala_3	TIP3P	1.60 +/- 0.04	1.89 +/- 0.01	2.17 +/- 0.01
	TIP4P-Ew	1.57 +/- 0.01	1.75 +/- 0.04	2.05 +/- 0.04
Ala_5	TIP3P	1.44 +/- 0.02	1.62 +/- 0.03	1.81 +/- 0.01
	TIP4P-Ew	1.36 +/- 0.01	1.36 +/- 0.01	1.55 +/- 0.01

In addition to the protein force field, the water model may also be expected to have a significant effect on the accuracy for these short, solvent exposed peptides. In order to compare solvent effects, we generated simulations using the TIP3P [49] and TIP4P-Ew [50] solvent models. Simulations using TIP4P-Ew have shown better agreement between calculated and experimental NMR observables [195, 202] due to more realistic diffusion and tumbling in this water model. However, TIP3P has been shown to be better than TIP4P-Ew at reproducing solvation free energies of small molecules [124]. Our data for both polyalanines indicate that the deviations from experiment are modestly smaller (3 to 16 % reduction in the χ^2 value) when the TIP4P-Ew solvent model is used (Table 6-5). This data, and those from the previous studies, suggest that the combination of using the ff99SB force field with the TIP4P-Ew solvent model is one of the best combinations currently available, at least for short peptides.

To address the remaining issues with the force field, we compare the J-coupling constants across the sequence for the Ala peptides. We selected $^3J(H_N,H_\alpha)$ and the $^2J(N,C_\alpha)$ constant due their sensitivity to the φ and ψ angles of the backbone (Figure 6-1). For the middle residues of Ala_5, the calculated $^3J(H_N,H_\alpha)$ values show deviations ranging from 1.2 to 1.7 Hz from the experimental observables and (ranging from 6.8 to 7.4 Hz depending on the parameter set compared to the experimental values between 5.6 to 6.0 Hz). These trends are observed in the other polyalanine simulations as well (Figure 6-2); the coupling constants from the simulations are too large, indicating too much sampling of β-like local backbone conformations as compared to PP_{II}, although the latter remains the dominant conformation. In contrast, the calculated $^2J(N,C_\alpha)$ constants are in excellent agreement with the experimental values with the largest deviation in residue 2 which shifts to values that suggest slightly too much α-helical conformation (which are generally around 6.50 Hz on the Karplus curve). Therefore, the most apparent issue for the local backbone conformations in these simulations is that the ensembles are shifted slightly away from favored PP_{II} conformations.

A.

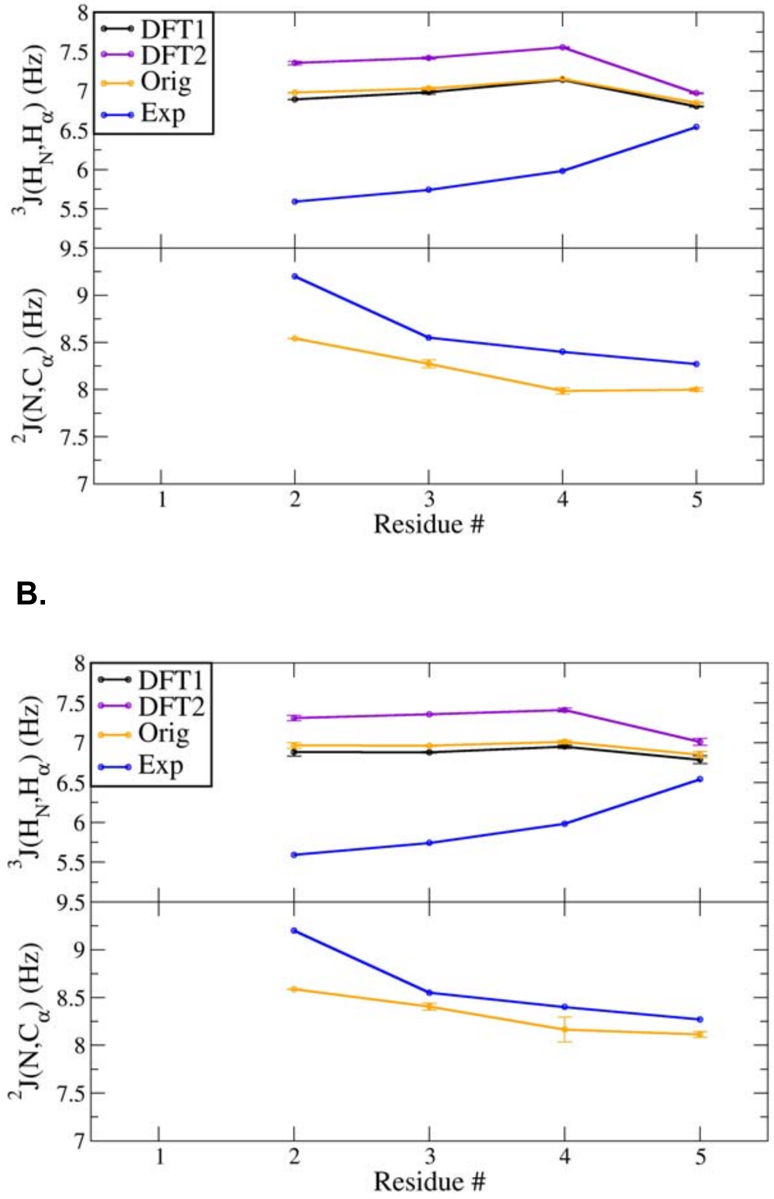


Figure 6-1. Average $^3J(H_N,H_\alpha)$ and $^2J(N,C_\alpha)$ scalar constants for simulations of Ala_5 in TIP3P (A) and TIP4P-Ew (B) solvent models. These constants are calculated with the original DFT1 (black), DFT2 (purple) and the original (orange) Karplus parameters respectively. Experimental scalar values are plotted in each graph in blue. Error bars are calculated from average difference between two simulations.

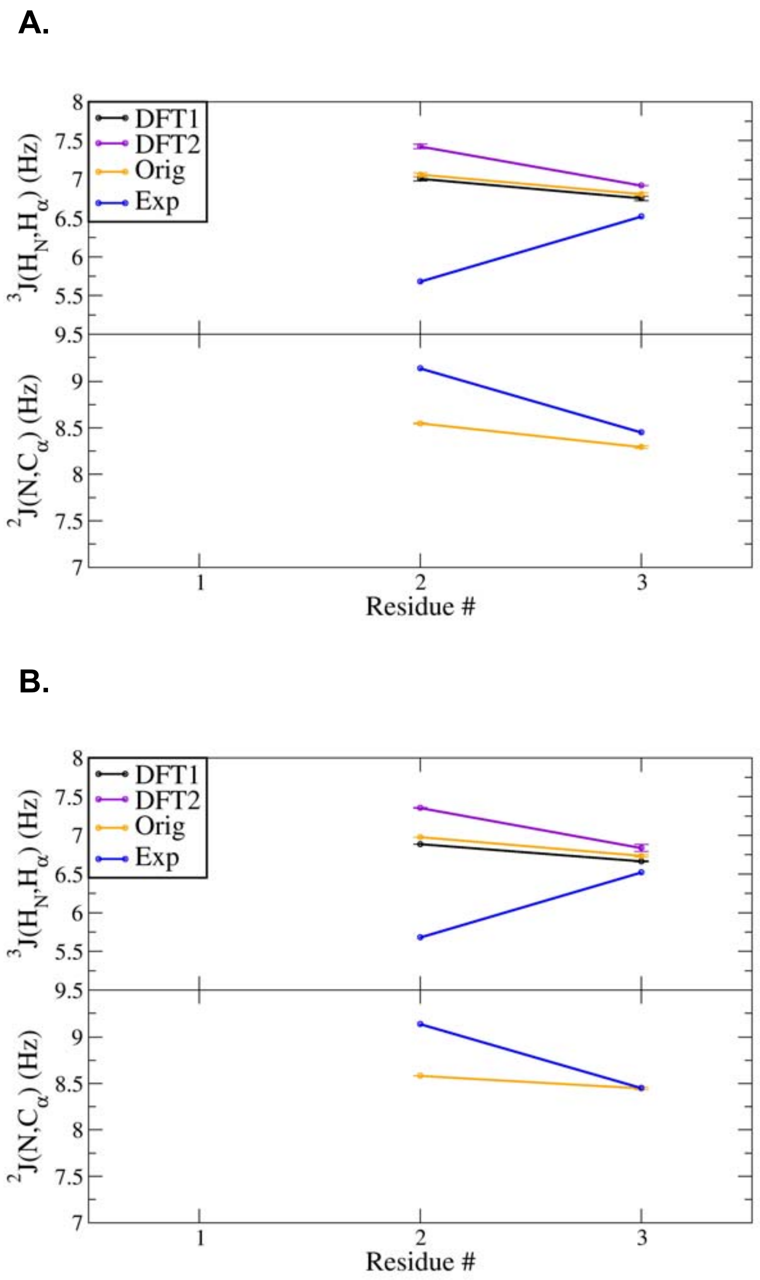


Figure 6-2. Average $^3J(H_N,H_\alpha)$ and $^2J(N,C_\alpha)$ coupling constants of each residue for the simulations of Ala_3 and Ala_5 in TIP3P (A/C) and TIP4P-Ew (B) solvent models at 300 K. Ala_5 simulations in TIP4P-Ew are included in the main text. DFT1, DFT2 and Original (Orig) correspond to the Karplus parameter set used in the calculation. The experimental values are also included on each graph [44]. Error bars were calculated from the average difference of the two independent simulations.

Our results show that scalar coupling calculations are sensitive to the implemented Karplus parameter sets (Table 6-5 and Figure 6-1 and 6-2). Based on the calculated

$^3J(H_N,H_\alpha)$ values, the DFT2 [201] parameters should have the worst χ^2 value; however the Orig set [44] produces the worst results for Ala_5 in the TIP4P-Ew model. In Figure 6-3, the average J-coupling scalar constants are shown for the other scalar constants involved in the χ^2 analysis. The parameter sets show similar trends for most of the scalar constants except for $^3J(H_N,C_\beta)$ where the DFT2 set performs the best, compensating the errors in the $^3J(H_N,H_\alpha)$ and resulting in lower overall χ^2. Thus, the χ^2 data should be interpreted with caution and the influence of Karplus parameters on individual errors must be considered. Furthermore, the Orig parameters implicitly include the effects of motional averaging, and one would therefore expect worse agreement with experiment when scalar couplings are back-calculated from the full ensembles using empirical parameters fit to experimental data [178]. Since this is not apparent, it suggests that the effect of force field inaccuracies on the simulated ensembles may exceed the effects of including dynamic fluctuations both implicitly in the Karplus parameters and explicitly in the MD ensembles.

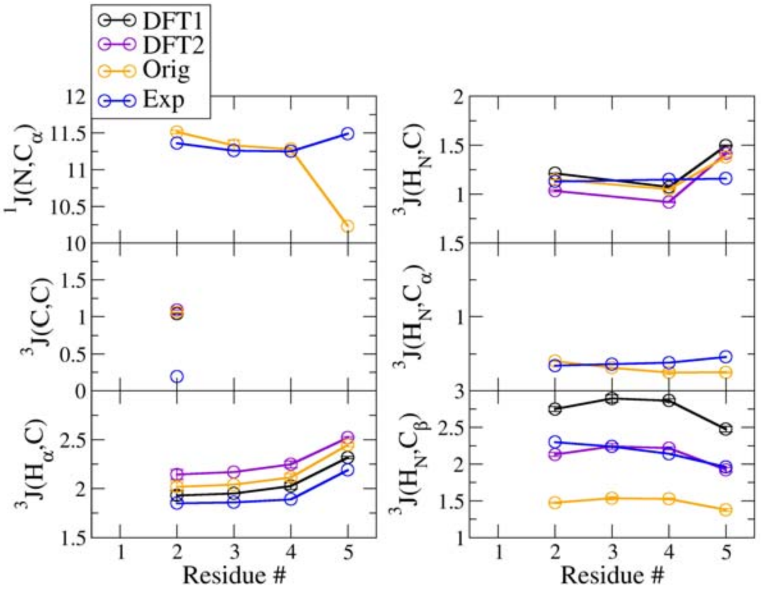


Figure 6-3. Average ^{1}J(N,C$_{\alpha)}$, ^{3}J(C,C), ^{3}J(Hα,C), ^{3}J(H$_N$,C), ^{3}J(H$_N$,Cα), and ^{3}J(H$_{N,}$C$_\beta$) coupling constants of each residue for the simulations of Ala$_5$ in TIP4P-Ew solvent model at 300 K. DFT1, DFT2 and Original (Orig) correspond to the Karplus parameter set used in the calculation. The experimental values are also included on each graph [44]. Error bars were calculated from the average difference of the two independent simulations.

Helical structural bias has been a problem associated with previous Amber force fields. We thus focused on Ala$_5$ since its sequence is long enough to permit an α-helical hydrogen bond. We calculated the percentage of α-helical conformations sampled by the central residue of Ala$_5$ with the definition used by Best et al. [43] (Table 6-6). In the TIP3P and TIP4P-Ew solvent models, the α-basin populations are 20 % and 15 % respectively, a significant improvement over the ff94 and ff99 force fields (90-95 % of the ensembles sample an α-helical population depending on simulation conditions) [43]. Nevertheless, one must use caution in interpreting these results in terms of helix formation. These calculations measure only the ϕ/ψ Ramachandran basin at the residue level; the structures may not sample an actual α-helical hydrogen bond. To test this, we repeated our calculations employing the dictionary for secondary structure prediction (DSSP) [128] definition for helicity which resulted in populations of 0.4 % and 0.0 % of 3_{10} and α-helix in both solvent models. Hence, the ff99SB ensemble does not suffer from the ailments of the previous force fields since it samples local α-helical conformations only in the random coil state, and no measurable amount of helical conformations.

Table 6-6. Populations of α, β and PPII basins on the Ramachandran map for the central residue of Ala_5. Error bars were calculated from the average difference of each basin population for two independent simulations.

Peptide	**Water Model**	**α**	**β**	**PP_{II}**
Ala_5	TIP3P	19.6 +/- 1.4	34.2 +/- 0.4	41.0 +/- 0.8
	TIP4P-Ew	15.1 +/- 4.6	36.6 +/- 2.7	45.1 +/- 2.0

6.4 Conclusions

In conclusion, comparisons of the Ala_3 and Ala_5 ensembles in both water models exhibit an excellent agreement between experimental and calculated scalar couplings. We also find that these calculations are somewhat, though not strongly, solvent model dependent, indicating that the TIP4P-Ew water model is the better choice for comparisons with NMR scalar coupling data. The deviations of the calculated and the experimental $^3J(H_N,H_\alpha)$ scalar constants with all of the parameter sets suggest that deviations are the largest in the φ-torsional potential which could have effects on larger systems. Nevertheless, ff99SB does not face the helical bias issues of the previous force fields. Future work will move towards using this experimental data as a reference for the further improvement of our force field parameters.

7. The Effect of Different Explicit Water Models on Peptide and Protein Conformational Preferences and Energetics

Abstract

Because of its importance in many biological processes, accurate modeling of water is one of the primary challenges in biomolecular simulations. Despite the availability of potentially more accurate water models, fixed charged models remain a popular choice due to their relatively low computational expense. Among other factors, water model performance is dependent on its compatibility with the force field, temperature, and long range electrostatic method used in a simulation. Two of the most popular fixed charged models are the TIP3P and TIP4P water models. Recent parameterization of TIP4P for use with the Ewald method, TIP4P-Ew, has resulted in an improved agreement with experiment for properties of bulk water. Simulations using TIP4P-Ew have shown better agreement between calculated and experimental NMR observables while TIP3P has been shown to be better than TIP4P-Ew at reproducing solvation free energies of small molecules. Questions still remain about which model is the better choice for simulations of peptides and proteins. In this work, we investigate the effect of using the TIP3P and TIP4P-Ew water models on the conformational preferences and energetics using model systems of various sizes, including Ala_3 and Ala_5, a short peptide with an ion pair, and the protein lysozyme. We studied local backbone dihedrals, conformational transition rates, radial distribution functions, ion-pairing, temperature dependence of structural

properties, and water density near the solute surface. We found that all except ion pairing and transition rates are relatively insensitive to the choice of water model.

Acknowledgments

This chapter contains direct excerpts from the manuscript (presented with minor modifications) written by Lauren Wickstrom with suggestions and revisions from Professor Carlos Simmerling. Dr. Asim Okur contributed simulation data for Ala_3. Trent Balius contributed the TIP3P simulation data for lysozyme, the crystal water occupancy analysis and appendix section 6. James Maier provided normalization scripts for the water density calculations. Robert Duke wrote the PME code in the AMBER package. We thank Holger Gohlke for order paramater scripts.

7.1 Introduction

Accurate modeling of water is essential since it is involved with most biological interactions. A few noteworthy examples include desolvation to form receptor-ligand interactions, expulsion of water from the hydrophobic core during protein folding and water mediated reactions in the catalytic sites of enzymes. The stability of these biomolecules is strongly influenced by the solvent-solute interface. Continuum models such as Poisson Boltzmann (PB) and Generalized Born (GB) [51] have been used to reduce the computational expense of explicit water. For accurate modeling, PB is often the better choice for implicit solvation, however its implementation in molecular dynamics is computationally demanding [52]. Furthermore, many GB implementations are known to cause such artifacts such as the overstabilization of salt bridges [33, 53-57] and α-helices [58, 59]. There appears to be a need for the inclusion of the first explicit solvation shell to capture effects for these biological molecules [57-61].

Various different water models exist for use in simulations and have been extensively reviewed [203-205]. There are a variety of water models which include quantum effects [206], explicit polarization [207-215], flexibility [216-218] as well as rigid fixed charged water models such as the TIPnP [49, 219] and SPC [220, 221] models. Proper water model selection is dependent upon how accurately one wishes to model the bulk solvent in a simulation and its computational expense. Work by Gerber's group has emphasized how electronic structure/ab initio methods are necessary to capture anharmonic effects seen in vibrational spectra between glycine and water [222-224]. If one is interested more in the solute behavior, this level of theory may not be necessary to observe accurate dynamics. As a result, rigid molecular mechanical water models remain widely used due to their low computational expense. The focus of this work will be on examining the effects of these rigid models.

Rigid water model performance is dependent on the force field, temperature, and treatment of long range electrostatics in a particular simulation. Typically, TIP3P is used with CHARMM [40] and AMBER [225], while SPC [220] and SPC/E [221] are often used with GROMOS [226]. Since OPLS [41] was developed from TIP3P/TIP4P parameters, it is often used with all three TIP models. The utilization of a force field with an incompatible water model may cause problems with transferability for parameters such as partial charges, and may lead to an imbalance between the solute-solvent and solvent-solvent interactions. Despite this reasoning, recent work by Nutt et al. [227] has shown that use of all TIPnP models with CHARMM resulted in similar results for all of the water models. While this thorough study examined solute-solvent properties, solvation free energies and protein dynamics, the work did not address the effects on the solute interactions and conformational preferences. In addition to the force field transferability issues, there are questions about how these water models will behave over a range of temperatures. Most fixed-charge water models have been developed for use at room temperature with a few recent exceptions [50, 219]. Therefore, accuracy may diminish significantly at temperatures other than 300 K.

Furthermore, parameters of many of these early water models were fit using a truncated cutoff method for long range electrostatics. Due to the inaccuracies of this treatment [228, 229] and the availability of increasing computational resources, more sophisticated methods such as Ewald summation and reaction field techniques are now preferred for treating long range interactions. Nevertheless, parameters for these water models were fit with a truncation of long-range nonbonded interactions, which has resulted in changes to their thermodynamic and kinetic properties when these interactions are included [50, 230]. Recent efforts have been made to reparameterize these models for more modern methods like PME [50, 231, 232].

The scope of this work will focus on two very popular fixed charged water molecules: TIP3P and TIP4P, which differ in their topology, thermodynamic, and kinetic properties. TIP3P is a three site model with one oxygen (negatively charged) and two hydrogen (positively charged) atoms, while TIP4P is a four site model with one oxygen (no charge), a dummy atom (negatively charged) and two hydrogen atoms (positively charged). The dummy atom is shifted along the bisector of the HOH angle in the direction of the hydrogens (Figure 7-1). This additional atom increases the computational expense of TIP4P compared to TIP3P. Nevertheless, TIP4P is more consistent with experimental properties (ie. dipole, heat of vaporization, diffusion coefficient, long range structure in the radial distribution functions, temperature of maximum density, and self-diffusion coefficient) [49, 233] than is TIP3P except for the dielectric constant (dielectric is reported as 82 and 56 in TIP3P and TIP4P respectively). Subsequent reparameterization has been performed on the TIP4P model for PME electrostatics by minimizing the experimental error for the enthalpy of vaporization and density from 235 – 400 K [50]. This work resulted in improved structural properties such as dipole, diffusion coefficient and improved radial distributions compared to the previous TIP4P model for a range of temperatures and a slight improvement for the dielectric constant (63.9). Despite the higher computational expense, these encouraging results suggest that TIP4P-Ew is an attractive alternative to the TIP3P model.

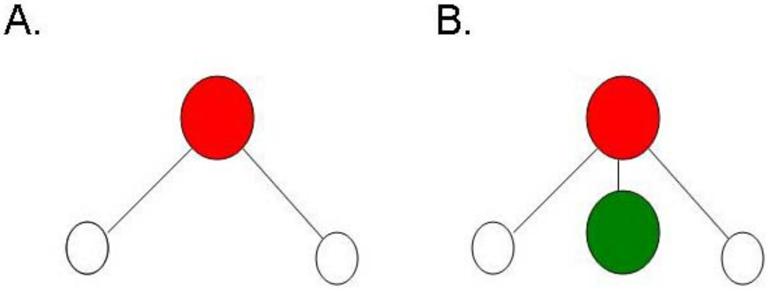


Figure 7-1. Two dimensional representations of the topologies of the TIP3P (A) and TIP4P-Ew (B) water models. The red circles represent the oxygen atoms, the white circles represent the hydrogens; and the green circle represents the dummy atom.

Several previous studies have also compared effects of using the TIP3P and TIP4P-Ew solvent models in biomolecular simulations [61, 124, 195, 202]. Notably, TIP4P-Ew has shown improvements over TIP3P for simulation studies involving comparisons to

NMR structural and relaxation data due to more realistic diffusion and tumbling in this water model [195, 202]. Wong et al. [202] suggests that poor diffusion properties in TIP3P will have an impact on the hydrogen bond dynamics of the solvent-solute interface. In contrast, solvation free energy calculations on small molecules have shown that TIP3P is in better agreement with the experimental values than the TIP4P-Ew water model [124, 234]. Shirts et al. [124] presented the argument that TIP4P-Ew was optimized for reproducing properties of water rather than solute-solvent properties.

Questions still remain about how each water model affects specific structural preference and stability in small peptides and proteins. In this study, we examine the effect of using TIP3P and TIP4P-Ew water model on conformational preferences of biomolecules. First, we focused on the possible differences in the backbone structure and local secondary structural conformations caused by the different explicit water models. Recent work fitting secondary structure populations to J-coupling constants has suggested that the short polyalanine peptides, Ala_3 and Ala_5 mainly sample PP_{II} structure with small amounts of local β conformations in solution [44] while a study using two-dimensional IR spectroscopy [235] with MD simulations and DFT calculations [236] has suggested that Ala_3 is primarily made up PP_{II} structure with minor populations of α_R helix. In contrast, Raman, FTIR and CD spectroscopy has suggested that there is less preference for PP_{II} structure and more β-strand type structure in this trialanine system (50 % each for the PP_{II} and β populations) [237, 238]. We performed replica exchange molecular dynamics (REMD) [62, 63] simulations on two small polyalanine polymers, Ala_3 and Ala_5. We examined φ/ψ distributions and secondary structural populations of the central residue, populations of cluster families, water density around the largest populated cluster, radial distribution functions and the temperature dependent properties of the backbone dihedrals for both Ala peptides. For both systems, the conformational populations at various temperatures and water structure were similar using both water models. We also ran 100 ns of molecular dynamics simulations of Ala_3 in order to evaluate transition rates in the TIP3P and TIP4P-Ew water models. In TIP4P-Ew, the transition rate between local secondary structural basins for the central residue of Ala_3 was lower than TIP3P, however this had a small effect overall on the structural populations.

Our second focus was to examine the possible effect of the water model on interactions between oppositely charged sidechains. We ran REMD simulations on a small model peptide containing a potential salt bridge between an Arg and Glu. Previously, this system was used to study ion pairing in an explicit vs implicit water study [57]. Here, we compare the potential mean of force for salt bridge formation, salt bridge geometries and cluster populations for both TIP3P and TIP4P-Ew solvent models. From this analysis, it is evident that the ion pairing is ~0.6 kcal/mol less stable in TIP4P-Ew than in TIP3P.

Last, we ran molecular dynamics simulations of hen egg white lysozyme in order to see if the effects seen in the model peptides were translated to this larger system. In both water models, lysozyme demonstrated similar structural trends in the backbone (except for the more the more flexible loop regions) while the salt bridge again appeared to differ in its stability. We also calculated the water density for the simulations in both water models and compared high density regions to the crystal water locations. Previous simulation studies have used this approach in order to determine importance of structural waters around protein surfaces [239-242]. In the lysozyme simulations, the regions of high water density in both models correspond well to structured waters in the X-ray structure.

In conclusion, the TIP3P and TIP4P-Ew models are quite comparable for structural properties of non-charged residues and water however transition rates and salt bridging interactions appear to be sensitive to the different water models.

7.2 Methods

We simulated Ala_3 and Ala_5 with a free N-terminus and a protonated C-terminus. These sequences and termini correspond to the low pH experimental studies [44]. Counter ions were not used. All simulations were performed in Amber version 9 [39] and used the ff99SB [45] force field. SHAKE [122] was used to constrain bonds to hydrogen. The time step was 2 fs. Temperatures were maintained using Berendsen thermostat [123] using a coupling constant of 0.1 ps. Explicit water simulations were performed in a truncated octahedron box with the TIP3P and TIP4P-Ew water models. Polyalanine REMD simulations were run in the NVT ensemble and particle mesh Ewald (PME) [125] was used to calculate long-range electrostatic interactions. A cutoff of 8.0 Å was used for real space electrostatics and Leonard-Jones calculations with a tolerance of 0.100e-4. All standard MD simulations used the *pmemd [243]* module of Amber, while REMD simulations used *sander.*

7.2.1 Ala_3

For both water models, an extended structure of Ala_3 was solvated with approximately 500 water molecules (498 for TIP4P and 525 for TIP3P). The structures were equilibrated at 300 K for 50 ps with harmonic restraints on solute atoms. The restraints were reduced from 5 kcal/mol*Å to 1 kcal/mol*Å to 0.5 kcal/mol*Å. After minimization, three 5 ps MD simulations were performed with the same gradually reduced restraints at constant pressure (1 atm) and temperature (300 K) to generate starting structures. Molecular dynamics simulations were run for 100 ns for each water model.

To improve sampling, we have performed REMD [62, 63] as implemented in Amber 9 in a similar fashion to previous studies [58]. The target exchange acceptance ratio for all simulations was approximately 20 %, with temperatures ranging from 260 – 580 K (291, 300, 310, 320, 330, 340, 351, 362, 374, 386, 398, 411, 424, 438, 451, 466, 481, 496, 512, 528, 545 and 562 K). Exchanges between neighboring temperatures was attempted every 1 ps. In order to evaluate convergence, an additional simulation was run with all the replicas initiated in an α-helical conformation in the 2nd residue. The simulations were run for 50 ns and the first 5 ns were discarded.

7.2.2 Ala$_5$

For both water models, an extended structure of Ala$_5$ was solvated in a truncated octahedron box using 891 water molecules. The restraints were reduced from 5 kcal/mol*Å to 1 kcal/mol*Å to 0.5 kcal/mol*Å. After minimization, three 5 ps MD simulations were performed with the same gradually reduced restraints at constant pressure (1 atm) and temperature (300 K) to generate starting structures. REMD simulations were run using a target acceptance ratio of approximately 20 % with the temperatures of 293 to 415 K (293, 300, 307, 314, 322, 329, 337, 345, 353, 361, 370, 378, 387, 396, 406 and 415 K). Exchanges between neighboring temperatures were attempted every 1 ps. In order to evaluate convergence, we ran an additional simulation with all the replicas initiated starting in an α-helical conformation for all the residues of Ala_5. Both simulations were run for 50 ns and the first 5 ns were discarded.

7.2.3 Arg-Ala-Ala-Glu Model Peptide

The setup and details for these simulations were the same as the previous work on this system in TIP3P water [57]. The peptide was solvated with 2286 TIP4P-Ew waters. REMD simulations were restrained to the representative conformation obtained from the highest populated TIP3P cluster using weak positional restraints on the backbone atoms (1.0 kcal/mol*Å). These simulations were run for 30 ns and the first 5 ns were discarded as equilibration.

7.2.4 Lysozyme

We simulated conformation A of hen egg lysozyme (PDB code 1IEE[244]) with a free N- and C-terminus. The crystal waters were removed and the structure was solvated with approximately 4998 waters in a truncated octahedron box. The system was first minimized for 1000 steps using positional restraints of 5 kcal/(mol Å) on the heavy atoms under constant volume. This system was equilibrated at 300 K for 15 ps with the same restraints, followed by two 15 ps MD simulations with gradually reduced restraints at 300 K under constant pressure of 1 atm. The MD simulation used a time step of 1 fs. The temperature was maintained with a Berendsen thermostat [123] with a coupling constant of 1 ps. Simulations were run with both TIP3P and TIP4P-Ew water models for 50 ns. Two simulations were run for each water model starting from two initial random number seeds.

7.2.5 Analysis

7.2.5.1 Ala peptides

Phi (φ) and psi (ψ) dihedrals for the central residue of the Ala peptides were calculated using the ptraj module in Amber 9. Free energy surfaces for the backbone dihedrals of the central residue were calculated at 300 K according to equation 7-1.

$$\Delta G_i = -RT \ln(N_i/N_0)$$

Equation 7-1. Relative free energy calculated with multidimensional histogram analysis. ΔG_i is the relative free energy bin I, N_i is the population of a particular histogram bin along the reaction, and N_0 is the most populated bin. R and T are the gas constant and temperature.

Secondary structure basin populations for central residues were calculated based on φ/ψ dihedral angle pairs. The definitions of the four principle regions were as follows: right handed helix (α_R), (φ,ψ) ~ (-160 to -50, -60 to +30); left handed helix (α_L), (+20 to + 70, -30 to +70); extended β-strand conformation (β), (-180 to -110, +110 to 180); and polyproline II (PP_{II}), (-110 to -40, +110 to +180).

Cluster analysis was performed with MOIL-View[127], using backbone RMSD as a similarity criterion with average linkage. The structures were clustered using the entire backbone for Ala_3 and residues 2 to 4 for Ala_5. Clusters were formed with a bottom-up approach using a similarity cutoff of 0.5 Å for Ala_3 and Ala_5 respectively. Cluster analysis was performed on trajectories combined from TIP3P and TIP4P-Ew REMD simulations, and the normalized population for each cluster was calculated for each of the original simulations. This ensures consistent cluster definitions in all runs. The populations of each conformation family were calculated for the TIP3P and TIP4P-Ew ensemble [162].

Water density calculations were performed using the ptraj module in Amber 9. The bin spacing was 0.5 Å. Density grids were normalized by dividing the values by a grid normalization constant (4.103e-3 (avg # of waters in bulk water/bin)) and the number of frames in the trajectory. The normalized water grids map the number of waters relative to bulk water. Radial distribution functions were calculated with the ptraj module using a bin size of 0.2 Å.

7.2.5.2. Peptide with Ion Pair

Salt bridge PMFs were calculated using histogram analysis along a reaction coordinate defined using the distance between Cζ of Arg2 and Cδ of Glu5 for the model peptide. To investigate salt bridge orientations, cluster analysis was performed on atoms of the Arg and Glu sidechains for the TIP3P and TIP4P-Ew ensembles with a similarity cutoff of 1.2 Å. Populations were compared using the same procedure as the Ala peptides.

To probe the sensitivity of the specific H-bond donor to the guanidinium group, distances between Cζ of Arg2 and each of the two Oε of Glu were calculated. A correlation plot with those distances was constructed using the same procedure as the φ/ψ free energy surface. Both analysis methods were used in the previous work comparing the salt bridge strength between explicit solvent, hybrid and GB models [57]. Cluster analysis was performed on the structures occupying the most populated basins on the free energy surface using a similarity cutoff of 1.2 Å.

7.2.5.2 Lysozyme

Root mean square deviations (RMSD), distances for salt bridge donor and acceptors and water density were calculated with the ptraj module. Order parameters were calculated using the isotropic reorientation eigenmode dynamics approach [245] in the ptraj module and a script used in previous work by Koller et al. [46]. Crystal water occupancy was calculated by assigning Cartesian coordinates to the water density in the trajectory. Structured waters in the X-ray structure were subsequently mapped onto a grid and compared to the high regions of density from the simulation. In order to compare the regions of high density from the simulation to X-ray structure waters, we summed the density at all grid points within 0.5 Å away from each crystal water location and associated this value with the crystal water. The occupancy values were averaged over the trajectory for each water model. Thereby, we quantified which crystal waters position was highly populated during the simulation. Note that this method does not take into account the highly populated regions that are not near any crystal water. This is determent through visual inspection of the densities.

7.3 Results and Discussion

7.3.1 Characterizing populations of backbone conformations

7.3.1.1 Ala$_3$

We first examined the local conformational preferences for the smaller polyalanine, Ala$_3$. Histogram analysis was employed to calculate Ramachandran free energy profiles of the central residue of Ala$_3$ at 300 K (Figure 7-2). In both explicit water models, Ala$_3$ samples the PP$_{II}$ conformation as its global minimum while sampling β, α_R and α_L helical conformations as local minima. The shape of the landscapes are also quite similar for both water models and consistent with previous results testing the FF99SB on Ala$_3$ with an amidated N-terminus and N-methylated C-terminus [45]. One minor difference is the barriers between PP$_{II}$ and α_R secondary structural basins which are slightly higher in the TIP4P-Ew model relative to the TIP3P model. The free energy barriers are 0.4 kcal/mol higher between $\varphi < -90$ and $-60 < \psi < 0$ (Figure 7-2B) which would suggest that α_R helical conformations sampled in the TIP4Pew simulations have longer lifetimes.

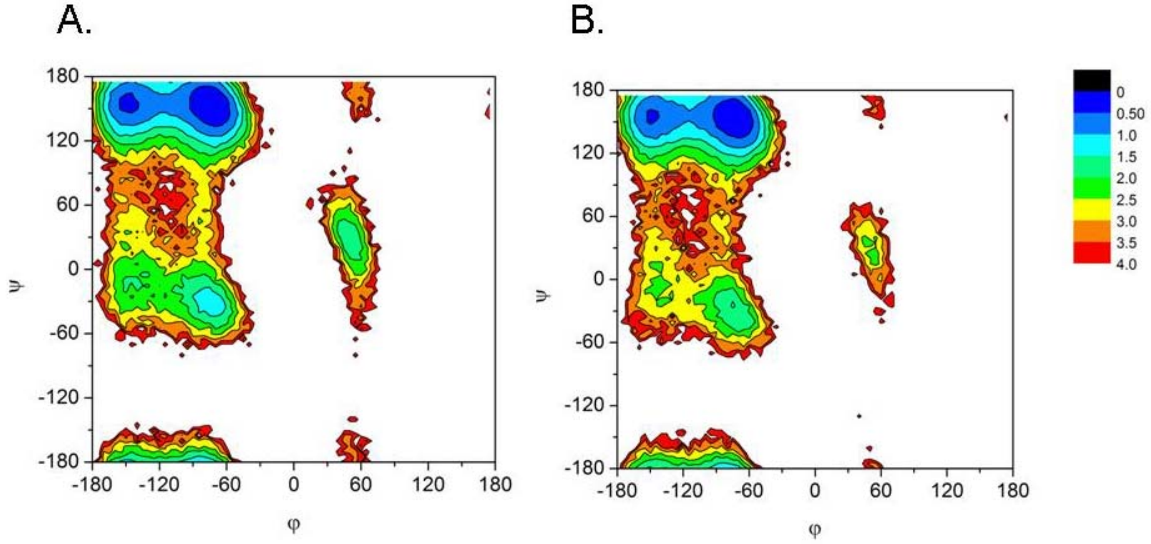


Figure 7-2. Free energy profiles for the central residue of Ala$_3$ from REMD in TIP3P (A) and TIP4P-Ew (B) solvent models. Free energies were calculated from populations as described in Methods. Contour levels are spaced 0.5 kcal/mol apart.

In order to quantitatively interpret the free energy Ramachandan plots, we analyzed the Ala$_3$ REMD simulation data obtained with each explicit water model in terms of fractional population of local conformational basins corresponding to the four secondary structure elements (PP$_{II}$, β, α_L,α_R) (Table 7-1). Error bars were obtained from averaging the two runs started from two initial conformations. There is an overall preference for

PP_{II} structure in the central residues of the Ala_3. Hornak et al. [45] and Okur et al. [58] have used neutral Ala_3 which have resulted in a smaller population of PP_{II} structure (approximately 40 % vs 53 % for the neutral and charged Ala_3 respectively in TIP3P) showing a weak effect of the charge on the overall ensemble of this short peptide. Using TIP4P-Ew explicit model, Ala_3 shows a slightly greater preference toward PP_{II} conformations compared to the simulations using TIP3P water (57 % ± 0.1 and 53 % ± 0.1 respectively). The relative fractions of β conformations are similar for both water models while populations of α_R and α_L are higher for the TIP3P water model (approximately 1.5 and 3 times more than TIP4P-Ew). These populations are quite different from previously reported values [44] and may be due to the definition of the basin or the preferences of the force field or both. Nevertheless, the populations in the secondary structural basin are similar for both water models.

Table 7-1. Populations of Basins on the Alanine Tetrapeptide φ/ψ Energy Landscapes Corresponding to Alternate Secondary Structures at 300 K

Solvent	$\boldsymbol{\alpha_R}$	**β**	$\mathbf{PP_{II}}$	$\boldsymbol{\alpha_L}$
TIP3P	9.9 ± 0.0	28.3 ± 0.3	52.5 ± 0.0	2.4 ± 0.5
TIP4P-Ew	6.2 ± 0.6	29.0 ± 0.5	57.1 ± 0.1	0.8 ± 0.1

Following our previously published work [58, 79, 82], we evaluated the populations of each cluster to determine whether independent simulations give similar ensembles. All structures from both methods were combined and used to define a common set of families, then the population of each family was computed for each trajectory and compared. This is important because we want to be confident that the populations of each conformational basin are independent of initial coordinates. By using this method, we can also evaluate the similarity between the ensembles sampled by two different solvent models similar to our previous work [58]. The REMD ensembles sampled from duplicate runs with same solvent models demonstrated excellent agreement for the TIP3P and TIP4P-Ew solvent model respectively with a correlation coefficient of 0.991 and 0.990 at 300 K (data not shown). Figure 7-3A shows the comparison of the Ala_3 ensembles in TIP3P and TIP4P-Ew water. It is evident that the ensembles are quite similar with correlation and regression coefficients of 0.995 and 1.125 respectively. For Ala_3, there appears to be very little difference in the ensemble of backbone conformations sampled in the two solvent models. The most populated conformation sampled by both explicit water simulations was a fully PP_{II} conformation (Figure 7-4A) at 300 K which is the same as our previous work on Ala_{10} [58]. This conformation made up approximately 19 % of the TIP3P ensemble and 22 % of the TIP4P-Ew ensemble.

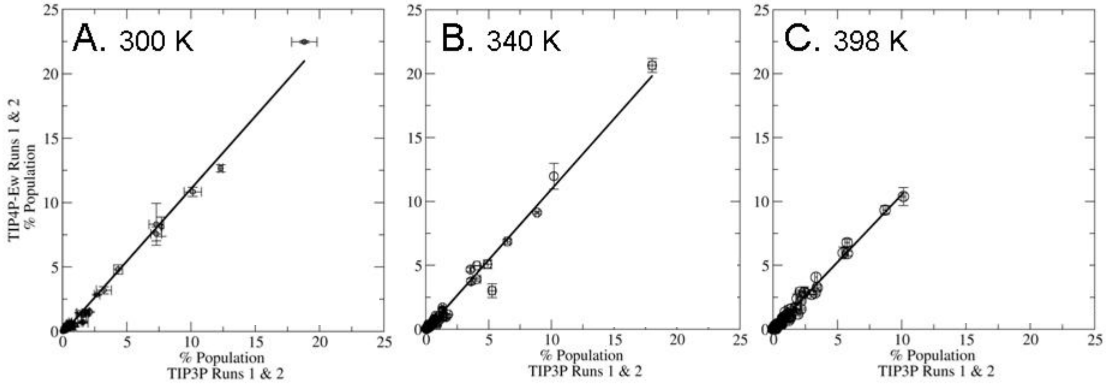


Figure 7-3. Comparison of populations for Ala_3 structure families sampled using TIP3P and TIP4P-Ew explicit solvent models. The ensembles are compared at 300 K (A) (r = 0.995), 340K (r = 0.991) (B) and 398 K (r = 0.992) (C) respectively. Clusters are defined using the combined data set. Populations are similar for the two solvent models at each temperature.

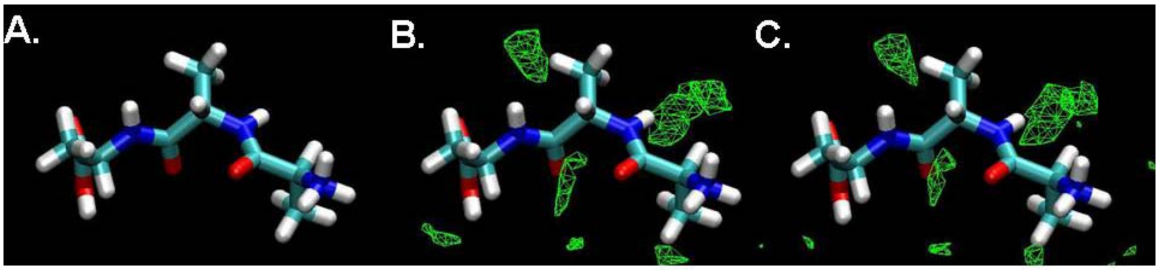


Figure 7-4. Representative structure and solvent density for the most populated cluster for Ala_3 in TIP3P and TIP4P-Ew solvent models (A). The PP_{II} conformation is the most populated in both solvent models. Solvent density is shown for the TIP3P (B) and TIP4P-Ew (C) models. The density for each model is quite similar

Next, we investigated how much the water structure around the most populated conformation differed in each solvent model. Normalized water density grids were calculated for the most populated cluster for each water model and overlapped on the representative structure for that cluster. The oxygen density of the water surrounding the PP_{II} conformation is shown for TIP3P and TIP4P-Ew solvent models (Figure 7-4B and C). The positions of the density are quite similar in both solvent models, where the amide groups tend to point towards highly populated regions of water molecules. The radial distribution functions between the oxygen of the carbonyl of the central residue and the oxygen of the water are almost identical (Figure 7-5A) while radial distributions for the solvent-solvent interactions are quite different for both solvent models (consistent with previous results [49, 50]) (Figure 7-5B). It is clear that TIP3P and TIP4P-Ew solvent models have little effect on the structuring of water around the most populated PP_{II} conformation.

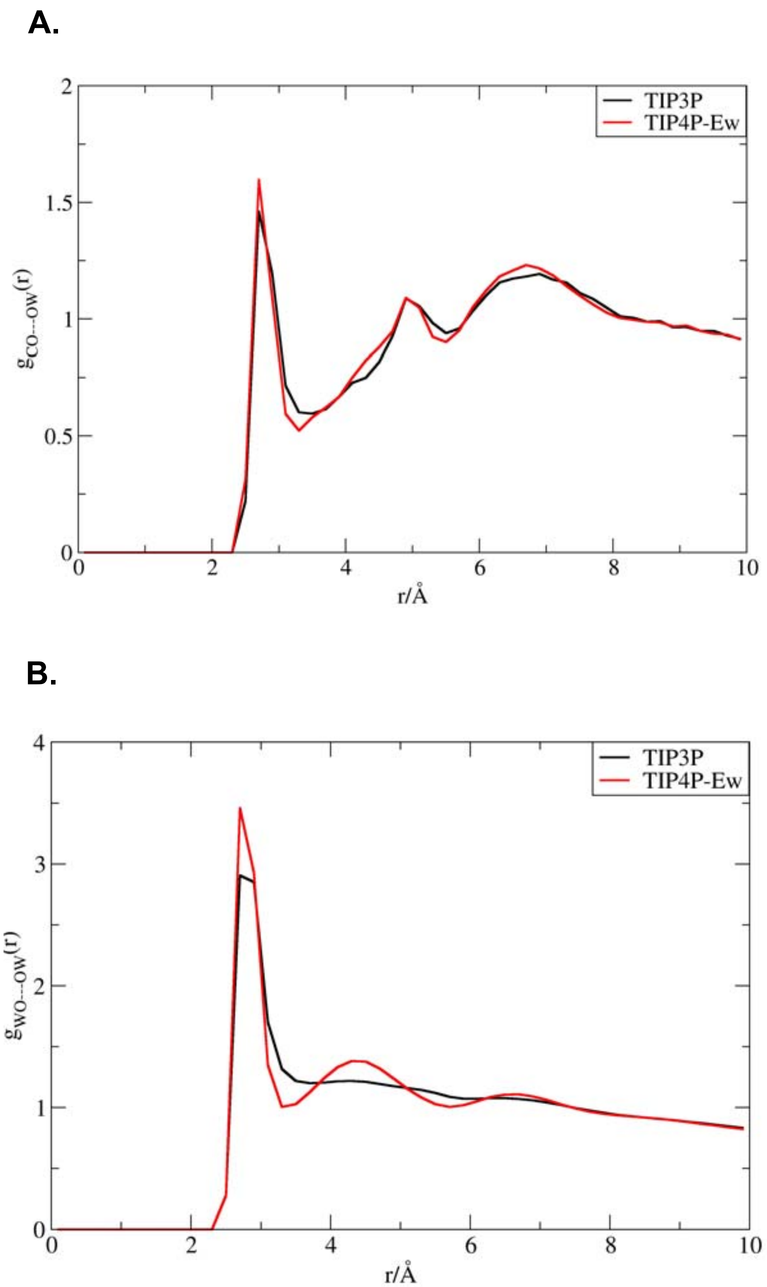


Figure 7-5. Radial distribution functions for (a) $g_{OW---OC}(r)$ and (a) $g_{OW---OW}(r)$ where OC is the oxygen(O) on the carbonyl(C) on the central residue and OW is the oxygen(O) of the water (W) in TIP3P (black) and TIP4P-Ew (red) models at 300 K. The $g_{OW---OC}(r)$ is similar for both water models while the $g_{OW---OW}(r)$ distributions differ in each solvent model.

Figure 7-6 shows the temperature dependence of PP_{II}, β, α_L,and α_R secondary structural basins of the central residue of Ala_3. As temperature increases, the populations of PP_{II}, and β conformations decrease while the α_R and α_L helical conformation increase in each of the solvent models. Throughout the range of temperatures, the population of the PP_{II}, and β conformations are higher in the TIP4P-Ew solvent model than in the TIP3P model. In Figure 7-3B and 3C, the ensembles of Ala_3 in TIP3P and TIP4P-Ew sample similar populations for the cluster families of the backbone at 340 K and 398 K respectively (correlation coefficient of 0.991 and 0.992 for the 340 K and 398 K ensemble comparisons of TIP3P vs TIP4P-Ew). Although there is a slight shift in the secondary structural populations, the temperature dependent trends for the backbone of Ala_3 are quite similar in both water models.

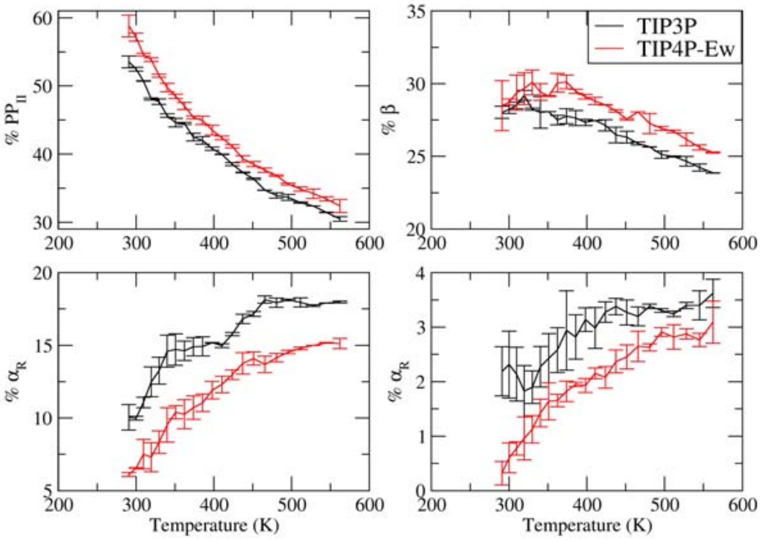


Figure 7-6. Temperature dependence of the secondary structural populations of the central residue of Ala_3. The different secondary structural basins shown are PP_{II} (a), β (b), α_R (c),and α_L(d). TIP3P is in black and TIP4P-Ew is in red. The temperature dependent behavior is similar for both water models.

7.3.1.2 Comparison of Conformational Transition Rates using TIP3P and TIP4P-Ew solvent models

Anomalously high water self-diffusion rates may allow for more transitions in the TIP3P water model and hence result in the sampling of alternative conformations. In order to look at transition rates of the backbone dihedral angles, we ran standard molecular dynamics (MD) simulations of Ala_3 in TIP3P and TIP4P-Ew water models. Figure 7-7 shows the time evolution of both the φ and ψ dihedral angles of the central residue of Ala_3 for the two different explicit water models. While the φ dihedral angles remain stable in the negative region during these MD simuations, the ψ angles sample a range of different configurations. The simulations using the TIP3P water model appear to be making more frequent structural transitions throughout the 100 ns than the TIP4P-Ew water model.

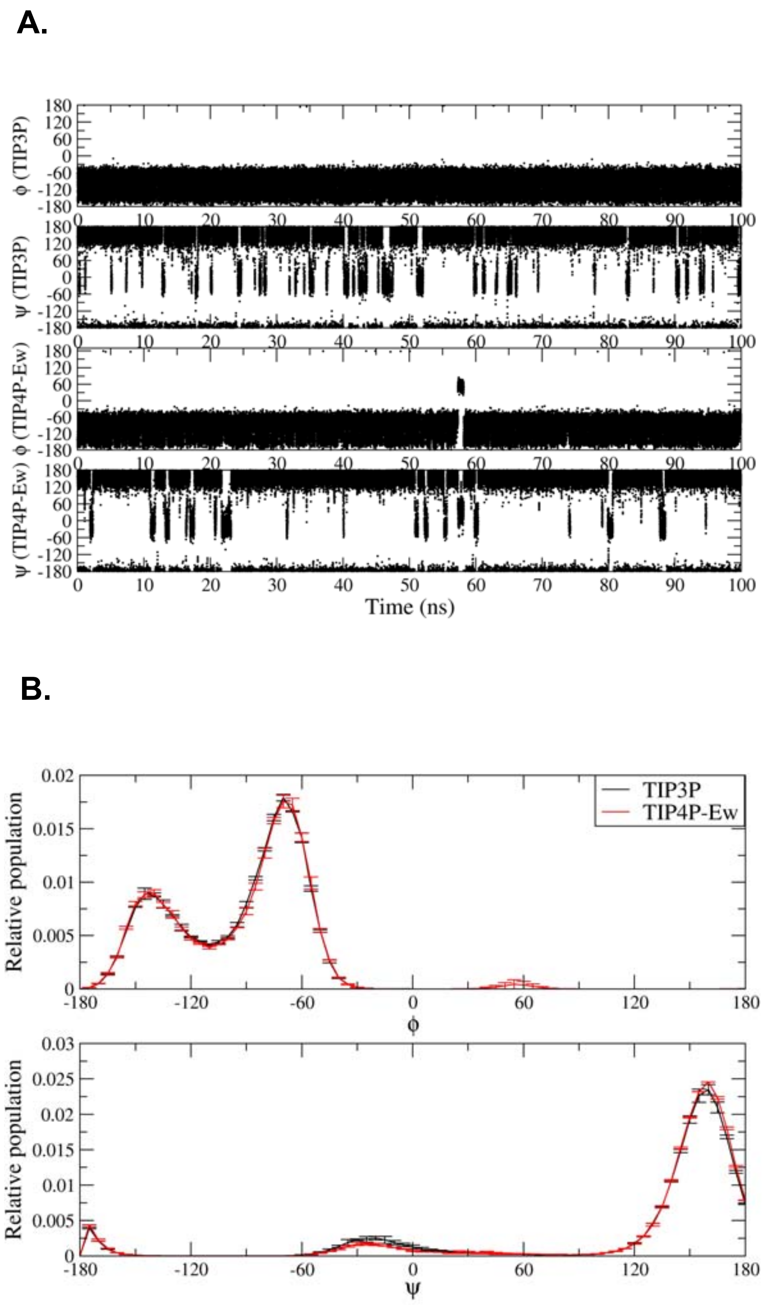


Figure 7-7. Time evolution (A) and relative histogram (B) distributions of the φ and ψ angles of the central residue of Ala_3 using the TIP3P (black) and TIP4P-Ew (red) water models. The ψ angle of residue 2 appears to be making more structural transitions however this appears to have small effect on the relative population.

We further analyzed these Ala_3 MD simulations by calculating the number of transitions between the different secondary structural elements. In Table 7-2, the possible basin transitions are listed with their corresponding frequencies during the 100 ns simulation. In both explicit water model simulations, the most frequent transitions were made between the PP_{II} and the β conformations. The only other secondary structural transitions were between α_R and PP_{II} and β and α_R and both occurred at a rate of 0.3 ns^{-1} or smaller during both simulations. Transitions between any other secondary structural basins were not made due to infrequent sampling and higher free energy barriers (Figure 7-2). In TIP3P, Ala_3 makes 1.3 times as many transitions between the PP_{II} and β basins and 4 times as many transitions between the β and α_R basins as compared to the simulations in the TIP4P-Ew water model. More backbone transitions occur in the TIP3P model because the viscosity is less than in the TIP4P-Ew model [50] which is most likely due to the higher self-diffusion constant [233]. This appears to have a very minor effect on the φ/ψ populations which are comparable in both models (Figure 7-7B).

Table 7-2. Amount of Secondary Structural Transitions in Ala_3 MD simulations in TIP3P and TIP4P-Ew explicit water models

Initial Basin	**Final Basin**	**Transition rate in TIP3P (ns^{-1})**	**Transition rate in TIP4P-Ew (ns^{-1})**
PP_{II}	β	35.9	28.0
PP_{II}	α_R	0.26	0.17
PP_{II}	α_L	0.00	0.00
B	PP_{II}	35.9	28.1
B	α_R	0.23	0.06
B	α_L	0.00	0.00
α_R	PP_{II}	0.22	0.12
α_R	β	0.27	0.11
α_R	α_L	0.00	0.00
α_L	PP_{II}	0.00	0.01
α_L	B	0.00	0.00
α_L	α_R	0.00	0.00

7.3.1.3 Ala_5

We extended our analysis to the longer Ala peptide, Ala_5. Similar to Ala_3, the PP_{II} basin is the free energy minimum for both solvent models and is sampled only slightly more frequently in TIP4P-Ew (52 ± 2.0 % vs 48 ± 1.0 %) (Figure 7-8 and Table 7-3). Figure 7-8 shows that the free energy barriers between the α_R and PP_{II} conformation have decreased compared to the barriers seen in the free energy profiles of the central residue of Ala_3 (approximately 0.35 to 0.49 kcal/mol in TIP3P and TIP4P-Ew respectively) (Figure 7-2). This barrier has caused a population shift from the PP_{II} and β local conformation to an increased number of α_R conformations for the central residue (18 ±1 vs 14 ± 4, Table 7-3). These results contradict J-coupling studies that suggest that there are no significant changes in structure caused by increasing length Ala peptide chain, though the differences may be within the data uncertainties [44]. We also looked at the secondary structural propensities for the neighboring residues of the central residues. Across the sequence, the population of PP_{II} and β decreases as the α_R population increased. The solvent conformational preferences remained the same as the central residue for those basins (Figure 7-9).

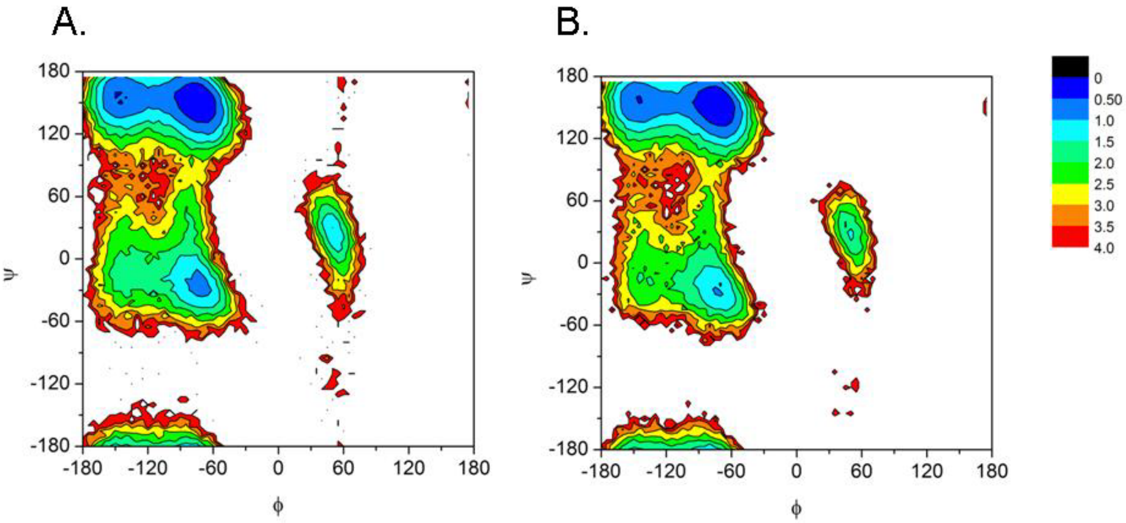


Figure 7-8. Free energy profiles for the central residue of Ala_5 in TIP3P (a) and TIP4P-Ew (b) solvent models. Free energies were calculated from populations as described in Methods. Contour levels are spaced 0.5 kcal/mol apart.

Table 7-3. Populations of Basins on the Ala_5 Energy Landscapes Corresponding to Alternate Secondary Structures at 300 K

Solvent	$\boldsymbol{\alpha_R}$	**β**	$\mathbf{PP_{II}}$	$\boldsymbol{\alpha_L}$
TIP3P	18.0 +/- 1.4	22.8 +/- 0.3	47.7 +/- 1.0	4.4 +/- 0.2
TIP4P-Ew	13.7 /- 4.3	24.6 +/- 2.2	51.9 +/- 2.0	2.6 +/- 0.1

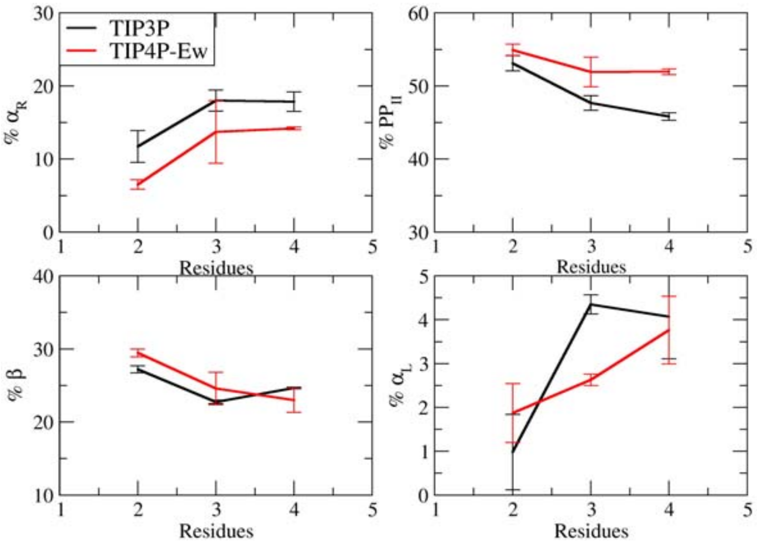


Figure 7-9. Secondary structural populations of the central residues in Ala_5. The different secondary structural basins shown are PP_{II} , β , α_R and α_L. TIP3P is in black and TIP4P-Ew is in red.

Similar to our analysis on Ala_3, we compared the populations of cluster families of Ala_5 in the different water models. Once again, there exists an excellent correlation between the structural populations sampled by the TIP3P and TIP4P-Ew models with correlation and regression coefficients of 0.988 and 1.193 respectively (Figure 7-10). Figure 7-11A shows the representative structure for the most populated cluster in both the TIP3P and TIP4P-Ew water models. The top cluster makes up 13 % of the TIP3P ensemble and 17 % of the TIP4P-Ew ensemble. This cluster is made up of primarily local PP_{II} conformations in residue 3 and 4 (residue 3 also samples some β conformations) in simulations using both water models (Figure 7-12). Figures 7-11B and Figure 7-11C show the oxygen density of the water sampled in this cluster, which is centered on the representative structure in both water models. Similar to Ala_3, both water models sample similar densities around the amide groups.

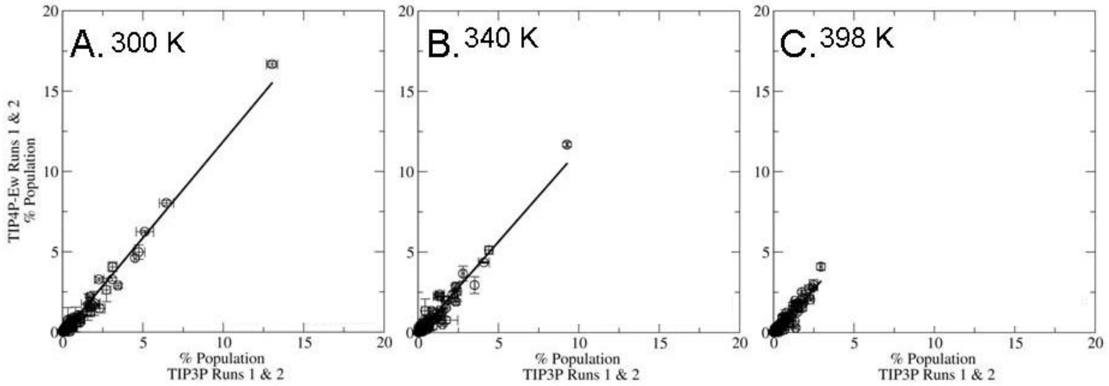


Figure 7-10. Comparison of populations for Ala_5 structure families sampled using TIP3P and TIP4P-Ew explicit solvent models. The ensembles are compared at 300 K (r = 0.988) (A), 340K (r = 0.971) (B) and 398 K (r = 0.957) (C) respectively. Clusters are defined using the combined data set. Populations are similar for the two solvent models at each temperature.

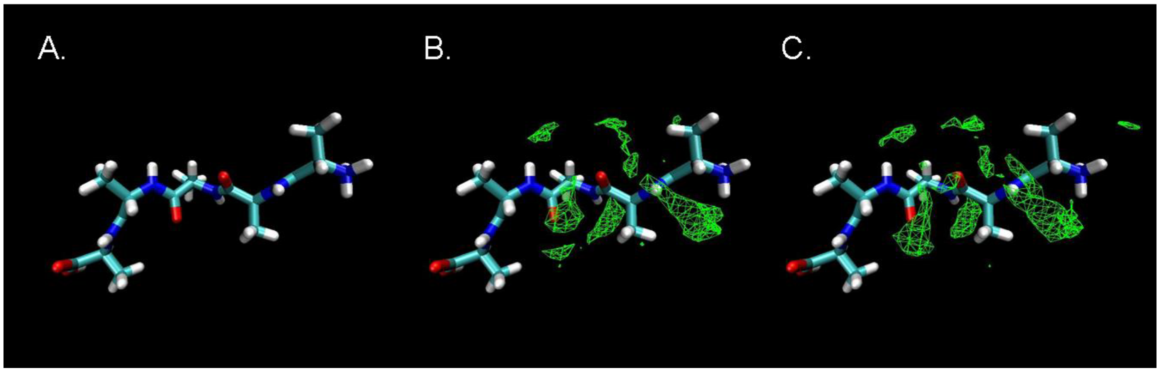


Figure 7-11. Representative structure and solvent density for the most populated cluster for Ala_3 in TIP3P and TIP4P-Ew solvent models. The PPII conformation is the most populated in both solvent models (A). Solvent density is shown for the TIP3P (B) and TIP4P-Ew (C) models. The density for each model is quite similar.

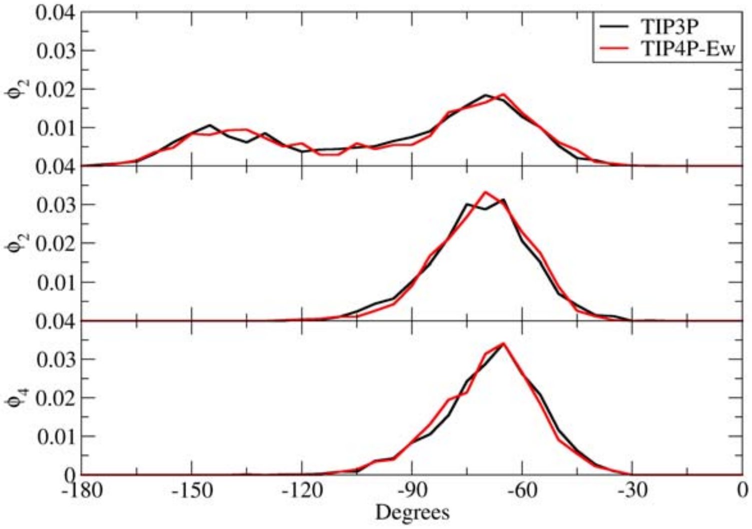


Figure 7-12. Relative population of the φ angles for residue 2 (φ_2), residue 3 (φ_3), and residue 4 (φ_4) for the most populated cluster of the Ala_5 ensemble at 300 K. The structures sampled in TIP3P and TIP4P-Ew are shown in black and red respectively. In both solvent models, each dihedral angle samples a similar distribution of structures.

Figure 7-13 shows the temperature dependence of the populations of the PP_{II}, β, α_L, and α_R basins of the central residue of Ala_5. Similar to Ala_3, as temperature increases, the PP_{II} conformations decrease while the α_R and α_L helical conformation increase in each of the solvent models. However, there appears to be very little temperature dependence in the local β conformations. Nevertheless, the Ala_5 ensembles sample similar populations for the cluster families of the backbone at 340 K and 398 K respectively (correlation coefficient of 0.971 and 0.957 for the 340 K and 398 K ensemble comparisons of TIP3P vs TIP4P-Ew) (Figure 7-10). Through the range of temperatures, both water models demonstrate similar trends similar to the Ala_3 ensembles. It is evident that like Ala_3, Ala_5 is relatively insensitive to these water models.

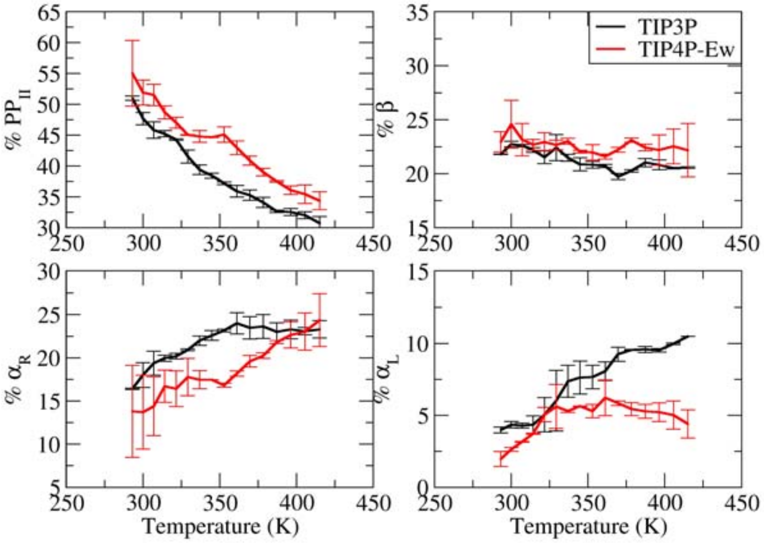


Figure 7-13. Temperature dependence of the secondary structural populations of the central residue of Ala_5. The different secondary structural basins shown are PP_{II}, β, α_R, and α_L TIP3P is in black and TIP4PEW is in red.

7.3.2 Model Peptide with ion pair

In order to investigate the effect of different explicit water models on sidechain interactions, we ran REMD simulations of a model peptide containing a potential ion pair. The sequence of this system was Ace-Arg-Ala-Ala-Glu-NH_2, with both Arg and Glu modeled in the charged state. Previous studies with this peptide were used to compare simulations using explicit water (TIP3P), different GB implicit solvent models and a hybrid explicit/implicit model [57]. In the same fashion, we investigate the effect of the different explicit solvent models on salt bridge strength and geometry. The backbone was restrained to a PP_{II} conformation to eliminate potential effects of different backbone conformations on the side chain interaction.

Salt bridge strength was evaluated through the calculation of the potential mean force for the distance between Cζ of Arg and Cδ of Glu as sampled in the simulated ensemble (Figure 7-14). The data demonstrates that the simulations using the TIP3P solvent model samples salt bridges that are moderately more stable (by approximately 0.6 kcal/mol) than the simulations using the TIP4P-Ew. The shape of the free energy minimum also varies using these different explicit solvent models. TIP4P-Ew simulations sample a broader minimum than the REMD TIP3P simulations. This suggests that the TIP4P-Ew simulations may sample multiple sidechain conformations in the minimum.

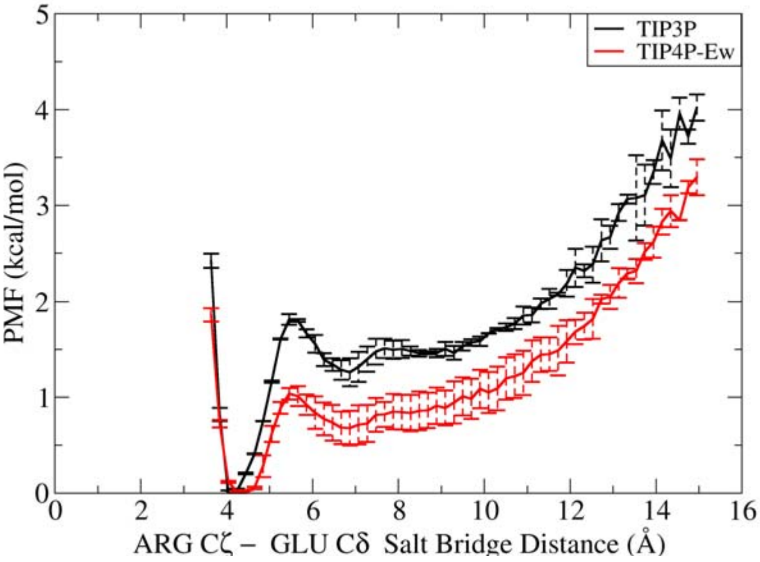


Figure 7-14. Potentials of mean force for the distance between Cζ of Arg and Cδ of Glu sidechains to compare the free energies for salt bridge formation for different solvent models. TIP3P is shown in black and TIP4P-Ew is shown in red. Salt bridges appear less stable in TIP4P-Ew than in TIP3P.

Similar to our analysis on the Ala peptides, we compared the populations of cluster families to evaluate how precise our results were for the REMD TIP3P and TIP4P-Ew simulations, and how similar the populations are in the different water models. The correlation coefficient for the independent REMD simulations in the different solvent models at 300 K was excellent for both models (0.995 for TIP3P and 0.888 for TIP4P-

Ew) (data not shown). In Figure 7-15, we compared the populations of sidechain conformations for TIP3P and TIP4P-Ew water models. **The populations are quite different in both water models,** which is demonstrated by a correlation coefficient of 0.731 and a regression coefficient of 0.461. It is clear from this analysis that the populations of sidechain geometries sampled by this model peptide vary in the different solvent models much more than the shorter polyalanine backbone.

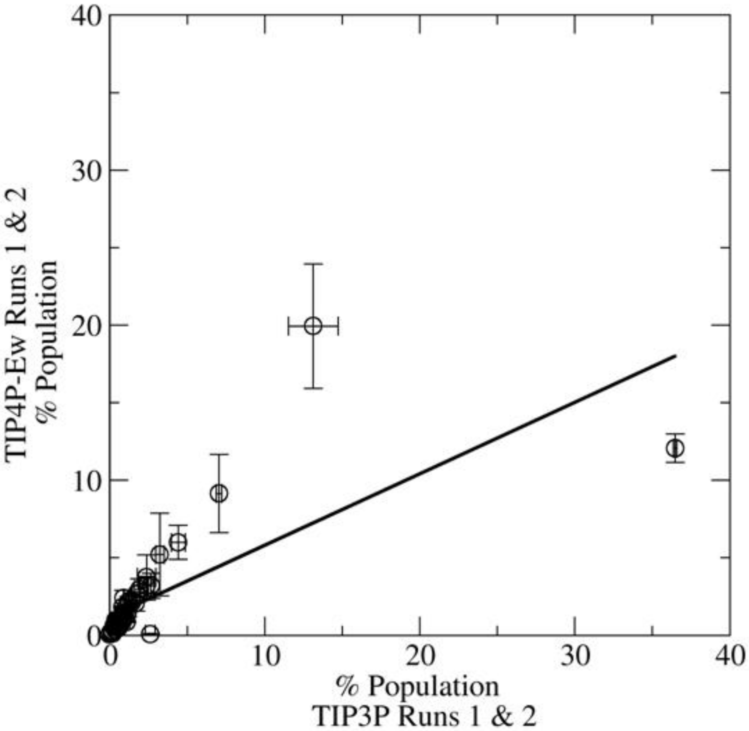


Figure 7-15. Population comparison of sidechain conformations for the salt bridge model peptide at 300 K. The plot is comparing the similarities of the conformations of the TIP3P and TIP4P-Ew ensemble. There is a relatively low correlation between the populations of both solvent models (r = 0.731).

Figure 7-16 shows the representative conformations for the most populated clusters for each water model. The salt bridge adopts a sidechain geometry which makes up 36 % of the TIP3P population and only 12 % of the TIP4P-Ew population of structures. The position of these sidechains allows for bifurcated hydrogen bonding between the Nε/NH1 and Oε1/Oε2 pairs. This conformation is also the most populated cluster observed for Arg-Glu pairs in proteins according the Atlas of Protein Sidechain Interactions. In contrast, the TIP4P-Ew ensemble prefers to adopt an alternate conformation (approximately 20 % of the TIP4P-Ew ensemble versus 13 % in the TIP3P) in which the χ_4 dihedral angle of the Arg sidechain is flipped 180° compared to the preferred structures in TIP3P (Figure 7-16B). The TIP3P ensemble appears to have a dominant sidechain geometry compared to the TIP4P-Ew model which suggests that there is more conformational entropy in the guanidinium group of the Arg sidechain in the TIP4P-Ew solvent compared to TIP3P solvent. It has been suggested that Arg salt bridges are too rigid using the TIP3P model as compared to NMR observables [246].

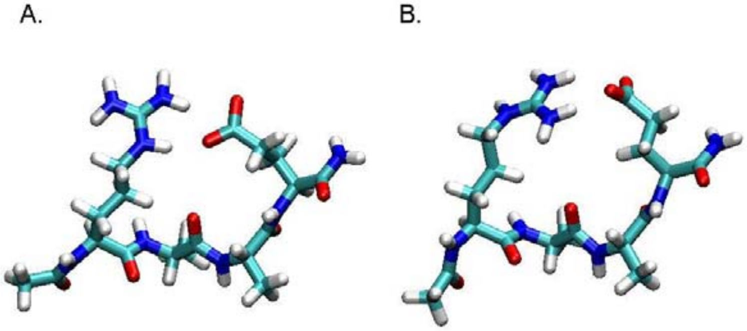


Figure 7-16 Representative structure of the most populated salt bridge geometries in standard using (A) TIP3P and (B) TIP4P-Ew.

In order to further characterize the geometry of the salt bridge, we investigated the hydrogen bond orientation between the Arg and Glu sidechains by calculating 2-dimensional free energy profiles for the distances between the Arg Cζ and the two Glu oxygens (Oε1 and Oε2) for the entire ensemble. The TIP3P ensemble prefers to sample a salt bridge geometry where both Glu oxygens simultaneously form hydrogen bonds with the Arg (Figure 7-17A) which is in agreement with the preferred conformations obtained through cluster analysis. The free energy profile for these hydrogen bonds also has a broad free energy minimum where each Glu oxygen is a comparable distance away from the Arg Cζ. The TIP4P-Ew appears to sample this conformation as well (Figure 7-17B). However, it appears to prefer one hydrogen bond between Arg and a single Glu oxygen, which shifts the other Arg C ζ to Glu oxygen distances to longer distances. This results in two minima on the surface due to the symmetry of the Glu carboxyl groups.

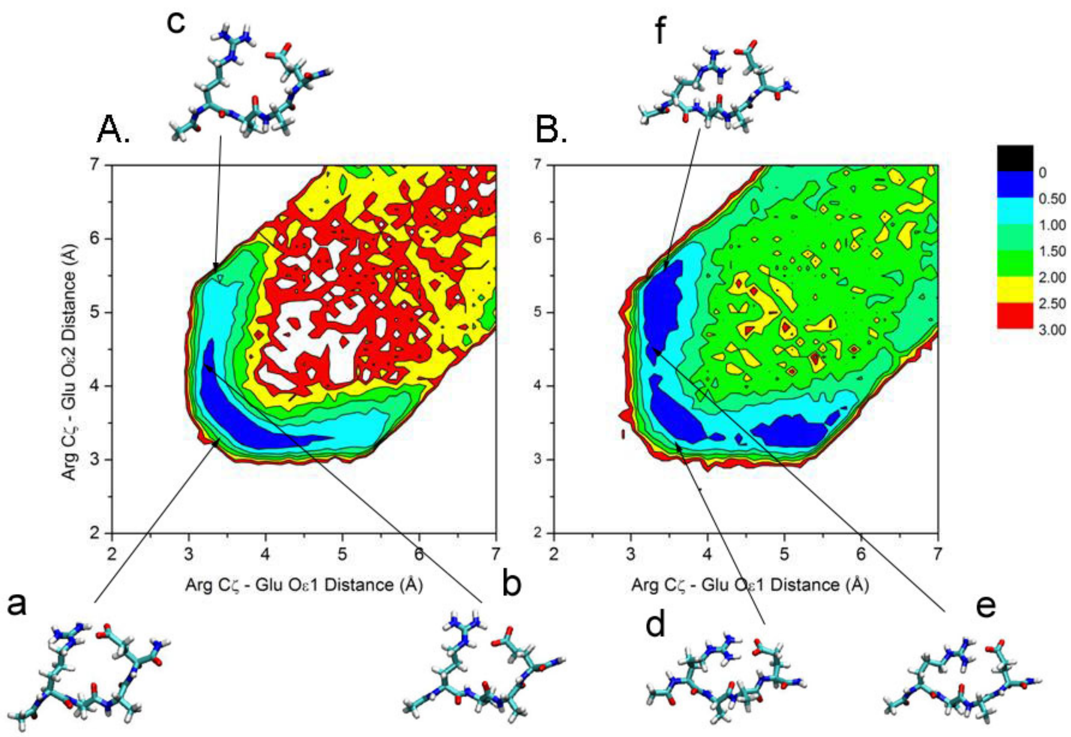


Figure 7-17. Free energy surface describing the geometry of salt bridge formation. The axes show the distance Arg Cζ and Glu $O\varepsilon_1$ versus the distance Arg Cζ and Glu $O\varepsilon_2$. Cluster analysis was performed on the structures in the major basins of the free energy surface. The representative structures of the top two clusters are mapped on the surface for each basin. Upper left and lower right sections of surfaces indicate convergence due to the symmetry of the sidechains. Contour levels are spaced 0.5 kcal/mol apart.

Cluster analysis was performed on only structures populating the major basins of the free energy landscapes shown in Figure 7-17 to compare the populations of different sidechain conformations defining these local minima. The three basins selected were between 3.0 and 6.0 Å for the Arg Cζ – Glu $O\varepsilon_2$ distance and between 3.0 and 4.0 Å for the Arg Cζ – Glu $O\varepsilon_1$ distance. Table 7-4 lists the top clusters for each basin for the TIP3P and TIP4P-Ew water models. For both water models, the free energy minimum is located in basin 1 at a Arg Cζ – Glu $O\varepsilon_1$ distance of 3.7 Å and a Arg Cζ – Glu $O\varepsilon_2$ of 3.4 Å however the structural preferences are quite different within this basin. The representative structures in the two most populated clusters (Cluster A and B) in this

TIP3P basin 1 prefer a sidechain geometry similar to the structure in Figure 10A while the most populated cluster in the TIP4P-Ew basin 1 prefers a geometry similar to the structure in Figure 7-17B but samples a bifurcated hydrogen bond between NH1/NH2 and Oε_1/Oε_2 pair. This geometry is the second most populated cluster observed in the Atlas of Protein Sidechain Interactions. This structure makes up a small part of the overall population in the TIP4P-Ew ensemble (approximately 3.0 %). The sidechain geometry in basin 1 is also observed in basins 2 and 3 in each solvent model. Overall, none of the highly populated clusters in basins 2 and 3 dominate a major part of the ensemble (less than 10 % of the overall populations) especially for the TIP4P-Ew model. In addition, we looked at the populations of the Arg χ_4 dihedral angle and found that TIP3P and TIP4P-Ew sample different distributions of structures (Figure 7-18). In TIP3P, conformations prefer to sample dihedral angles around 90° while TIP4P-Ew structures slightly prefer to sample structures with a χ_4 angle of -90. Overall, there is significantly more conformational variability and flexibility in TIP4P-Ew ensemble compared to TIP3P.

Table 7-4. Cluster populations for structural basins on the free energy surface describing salt bridge formation for the TIP3P and TIP4P-Ew water models.

Cluster	population in basin (%)	population in ensemble (%)
TIP3P		
A	68.7	13.4
B	75.1	9.4
C	50.3	3.5
TIP4P-Ew		
D	38.8	3.0
E	61.5	4.7
F	73.3	5.6

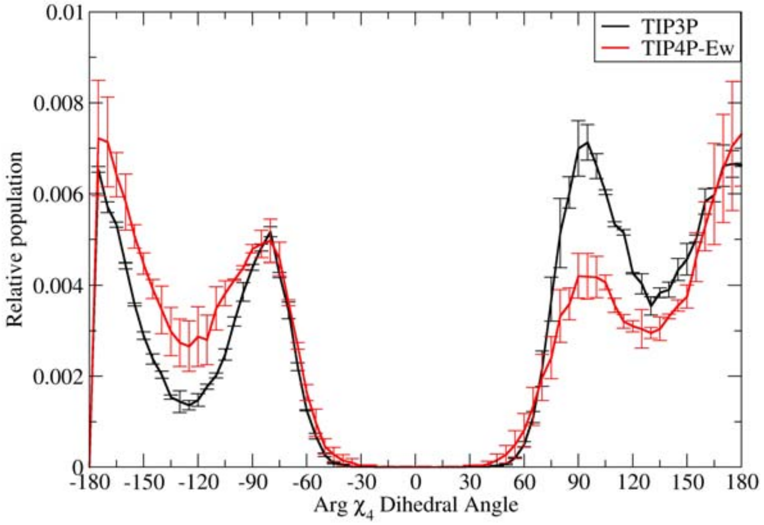


Figure 7-18. Relative populations of Arg χ_4 angle sampled in TIP3P (black) and TIP4P-Ew (red) sampled by the model peptide with an ion pair at 300 K.

Figure 7-19 shows the temperature dependence of salt bridge formation for the TIP3P and TIP4P-Ew water models. For TIP3P, salt bridge formation peaks at 64 % at 296 K and decreases to 38 % at 584 K, in contrast, in TIP4P-Ew the profile is nearly constant (43-45 %) for TIP4P-Ew throughout the same temperature range. **It is apparent that the salt bridge formation is much more temperature dependent in the TIP3P than in TIP4P-Ew solvent model.** We investigated further details of this behavior by looking at the distance distribution between Arg Cζ and Glu $O\varepsilon_1$ at eight different temperatures between 300 and 584 K (Figure 7-20). In both water models, there are two major peaks that are sampled at less than 6 Å. The larger peak is composed of two peaks located at approximately 3.5 and 3.8 Å which corresponds to structures with a bifurcated and a single hydrogen bond formation respectively. The smaller peak at 5.5 Å is the distance sampled when one hydrogen bond is formed by the alternative Arg Cζ and Glu $O\varepsilon_2$ distance. In the TIP3P water model, the populations at 3.5 and 3.8 Å decrease as the temperature increases. In contrast, the populations remain the same for the TIP4P-Ew water model at each of the temperatures. This TIP4P-Ew temperature dependent behavior is consistent with previous simulation work in the SPC model which showed similar populations salt bridge contacts at a range of temperatures [247]. Salt bridges are also known to be important in thermal stability of hyperthermophiles [247-253] hence these results seem to suggest reasonable behavior, but clear differences exist between the models.

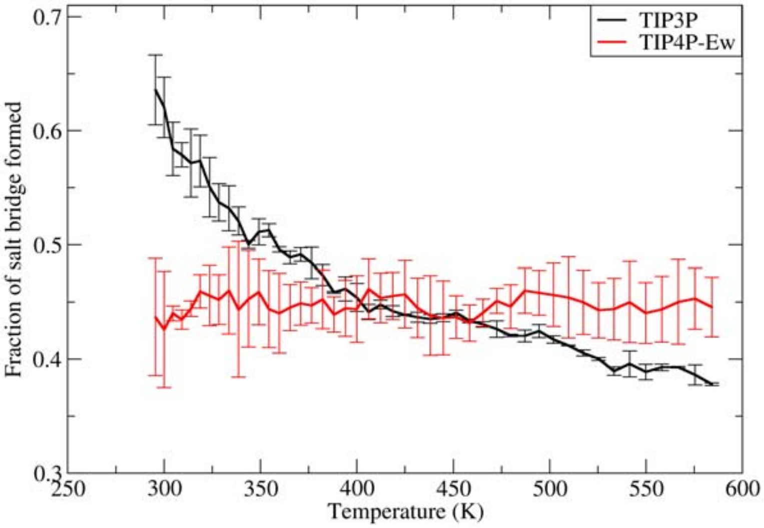


Figure 7-19. Melting curve for salt bridge formation in the TIP3P (black) and TIP4P-Ew (red) solvent models. In TIP3P, the salt bridge appears to melt while there appears to no temperature dependence with the salt bridge in TIP4P-Ew.

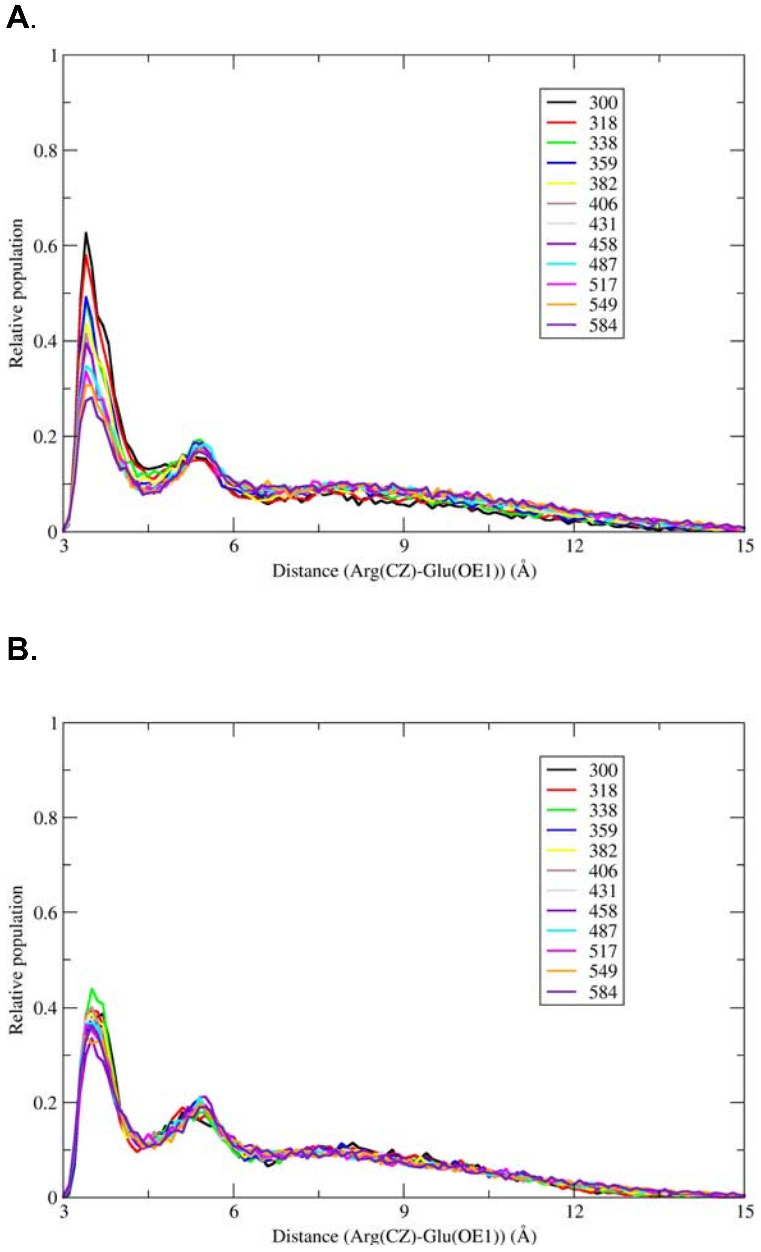


Figure 7-20. Relative population histograms of the distance between Arg Cζ and Glu $O\varepsilon_1$ in TIP3P (A) and TIP4P-Ew (B) for different temperatures. As the temperature increases, the TIP3P salt bridge melts while the TIP4P-Ew distance distributions remained constant.

7.3.3 Lysosyme

We chose the Hen egg lysozyme protein as a larger system (129 residues) to evaluate the structural differences caused by using these different water models. Due to its stability, lysozyme is a very frequently studied protein with multiple X-ray and NMR structures. Since there is so much experimental data available for this system, lysozyme is often used to benchmark the quality of molecular dynamics simulation data [45, 46, 254-258]. In this study, we ran simulations starting from the high resolution (0.94 Å) structure of lysozyme (PDB code 1IEE[244]) in both the TIP3P and TIP4P-Ew water models. Each simulation was run for 50 ns and repeated with a different initial velocity distribution.

To evaluate the stability in each water model, we compare the backbone RMSD for each of the simulations. In Figure 7-21, the time evolution of the backbone RMSD is shown for two simulations in the TIP3P and TIP4P-Ew water models. In both graphs, the RMSD stays below 1.8 Å and remains quite stable. The simulations of lysozyme in TIP4P-Ew appear to fluctuate slightly more than in TIP3P. Figure 13B shows the relative histograms of the backbone RMSD for TIP3P and TIP4P-Ew. Although the maximum is similar for both water models (0.9 and 0.95 Å for TIP3P and TIP4P-Ew respectively), TIP4P-Ew samples a broader distribution of structures.

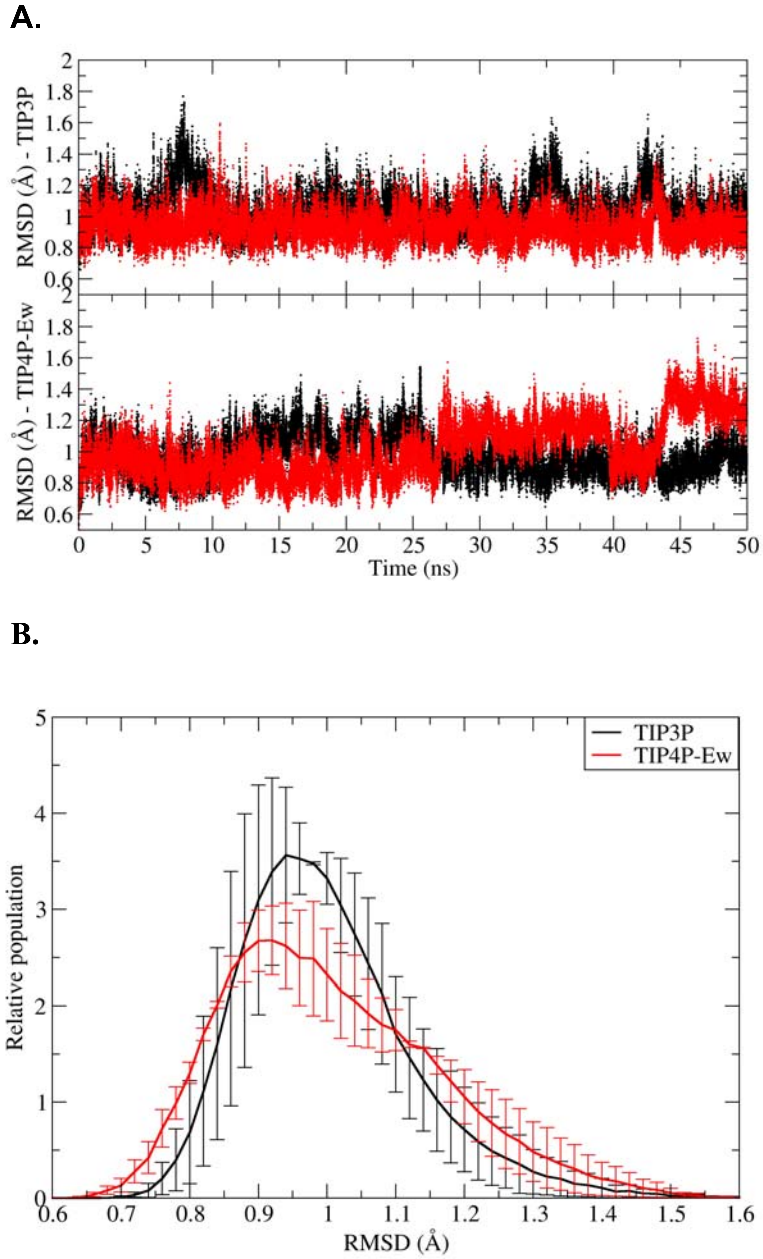


Figure 7-21. Time evolution (A) and relative population histogram (B) of the backbone RMSD of lysozyme. In figure A, run1 is in black and run 2 is in red. In figure B, the histogram is black is tip3p and the histogram in red is TIP4P-Ew. In both water models, the simulations seem quite stable however TIP4P-Ew shows more fluctuations.

In Figure 7-22, calculated S^2 parameters are shown for the different water models.

In the TIP3P and TIP4P-Ew simulations, the most flexible parts of the backbone are the loop regions. The largest fluctuations occur in L1 in the TIP3P water model while L2 fluctuates the most in the TIP4P-Ew model; the latter is more consistent with experiment. The TIP4P-Ew model shows slightly better agreement between calculated and experimental order parameters than TIP3P.

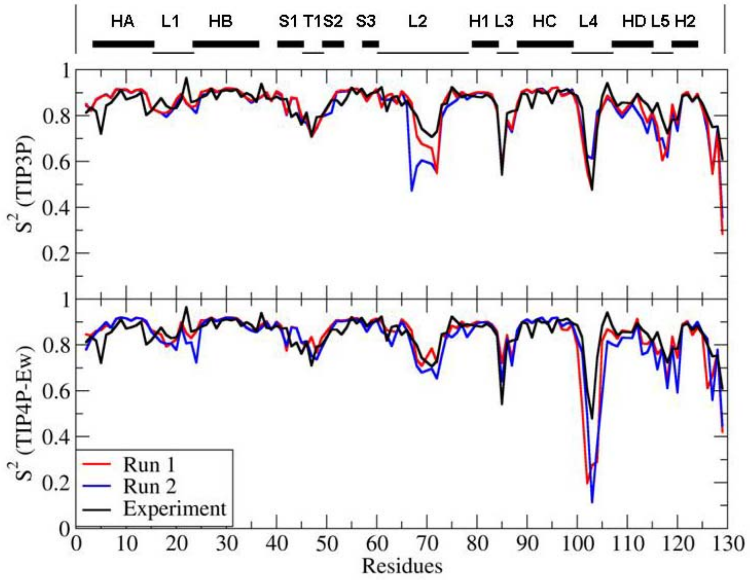


Figure 7-22 Comparison of experimental and calculated S^2 parameters for the simulations in the (a) TIP3P and (b) TIP4P-Ew water models. Run 1 and run 2 are shown in red and blue respectively. Experimental values are shown in black. Secondary structures of lysozyme: helix A (HA: residues 4-15), loop 1 (L1: 16-23), helix B (HB: 24-36), strand 1 (S1: 41-45), turn 1 (T1: 46-49), strand 2 (S2: 50-53), strand 3 (S3: 58-60), long loop 2 (L2: 61-78), 3_{10} helix 1 (H1: 80-84), loop 3 (L3: 85-89), helix C (HC: 89-99), loop 4 (L4: 100-107), helix D (HD: 108-115), loop 5 (L5: 116-119) and 3_{10} helix 2 (H2: 120-124).[45] The biggest deviations are seen in the loop regions. The calculated S^2 parameters from the TIP4P-Ew model are in slightly better agreement to experimental values[256] than the TIP3P S^2 parameters.

Further analysis investigated the effect of the solvent models on behavior of salt bridges in this larger protein. A salt bridge involving Arg was selected in order to be consistent with the previous analysis on the model peptide (Figure 7-23). The salt bridge pair was formed between Asp48 and Arg61 in the X-ray structure (3.6 Å) and connects the turn region of a β hairpin to the distant long loop 2 region (Figure 7-24). We monitored the salt bridge contact by measuring the distance between Asp48(Cγ) and Arg61(Cζ) in the simulations in TIP3P and TIP4P-Ew (Figure 7-23A). During the simulation, the contact appears to fluctuate more in the TIP4P-Ew model than in TIP3P model. We used histogram analysis to calculate the relative stability of the salt bridge contact in both water models (Figure 7-23B). The free energy minimum for salt bridge formation is located at 4.2 Å and 4.5 Å in the TIP3P and the TIP4P-Ew water model respectively. In the TIP4P-Ew model, the contact between Asp48 and Arg61 appears 1 kcal/mol less stable than in the TIP3P model. Consistent with the results of the model peptide, these salt bridge PMF and location of the free energy minimum suggest that salt bridges are stronger in the TIP3P model.

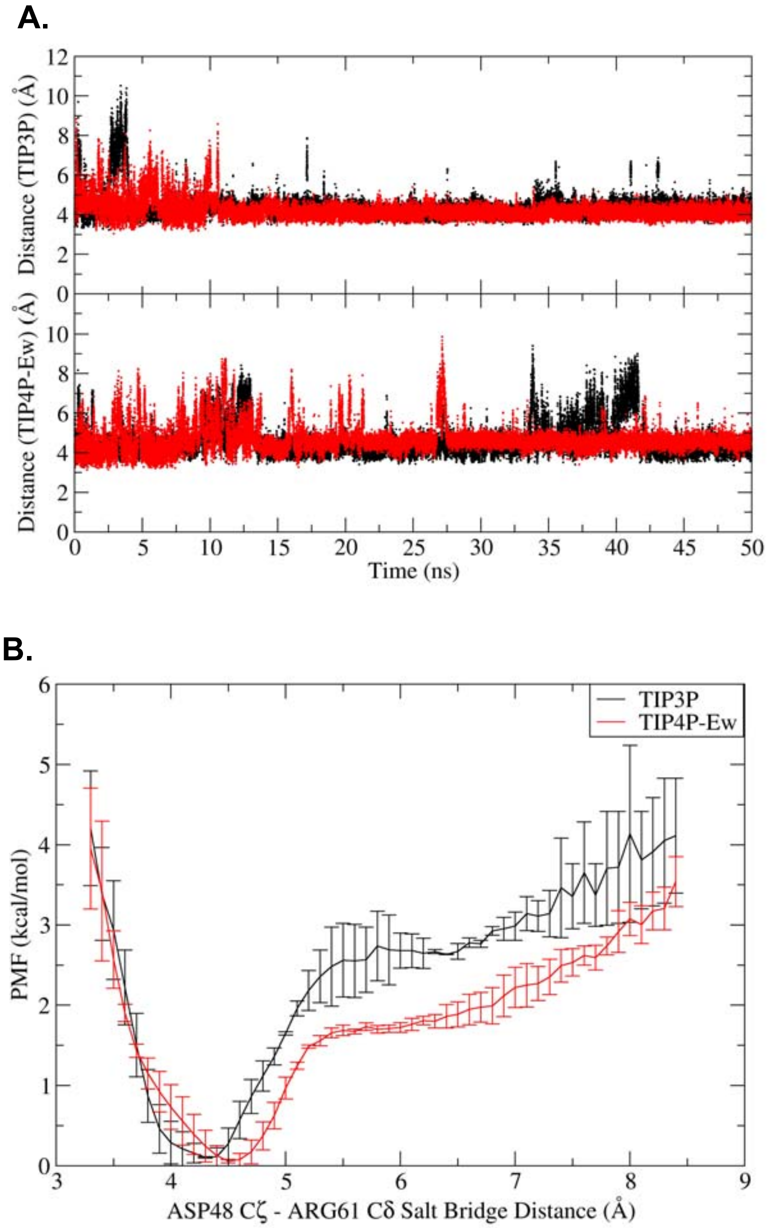


Figure 7-23. Time evolution (A) and potentials of mean force (B) for the distance between Asp48(CG) and Arg61(Cζ). These results suggest that salt bridges are less stable in the simulations using the TIP4P-Ew water model.

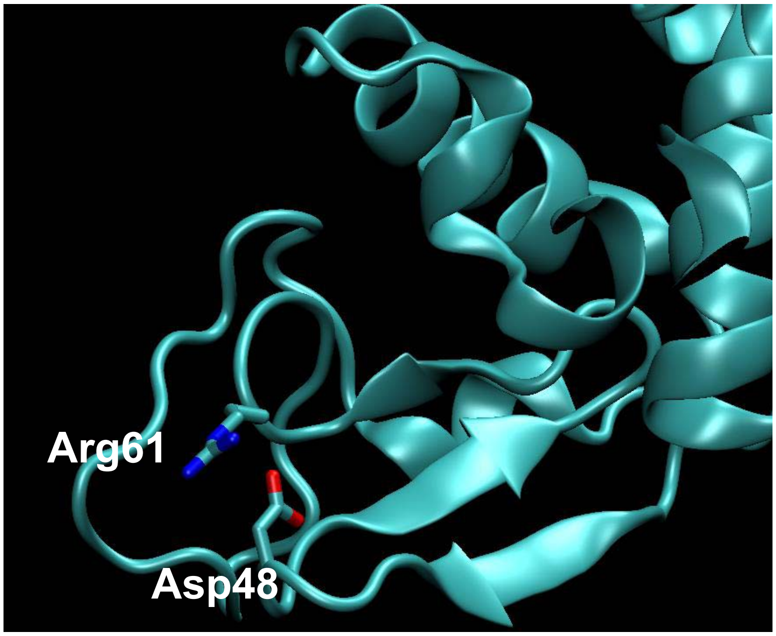


Figure 7-24. Salt bridge formed by Asp48 and Arg61 in the X-ray structure of Lysozyme.

Similar to the small peptides, we investigated the differences in the water density sampled by lysozyme in the two different solvent models. Figure 7-25 shows the water density sampled by run1 and run2 in the TIP3P (A) and TIP4P-Ew (B) models. The regions of high density were similar for both runs in each water model. We combined both runs and compared the water densities for the simulations in the different water models (Figure 7-26). Many of the regions of high oxygen density correspond to the locations of crystal waters. In order to compare the accuracy of each model, we calculated the occupancy of water for the regions containing waters in the X-ray structure. We note that all crystal waters were removed before MD, thus we are testing whether the simulation in each model can properly locate these positions. For both water models, the highest occupancy was located in the same region, which corresponded to a buried water in the X-ray structure (76.2 ± 6.9 % in TIP3P and 69.0 ± 0.5 %). These results are consistent with the results on polyalanine described above which suggest that the water density does not differ significantly in TIP3P and TIP4P-Ew water model.

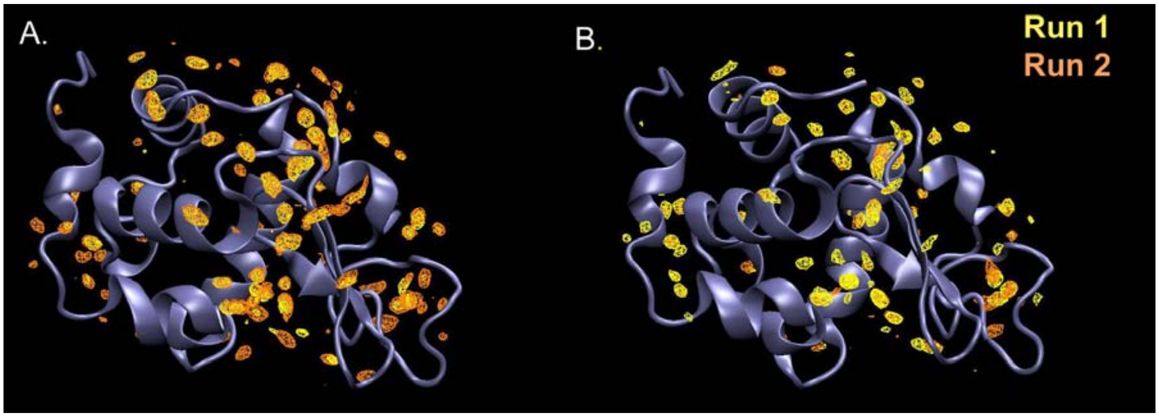


Figure 7-25. Water density observed in the TIP3P (A) and TIP4Pew (B) simulations mapped onto the X-ray structure. The density from run 1 and run 2 are in yellow and orange respectively. The regions of high density are similar in both runs in each solvent model.

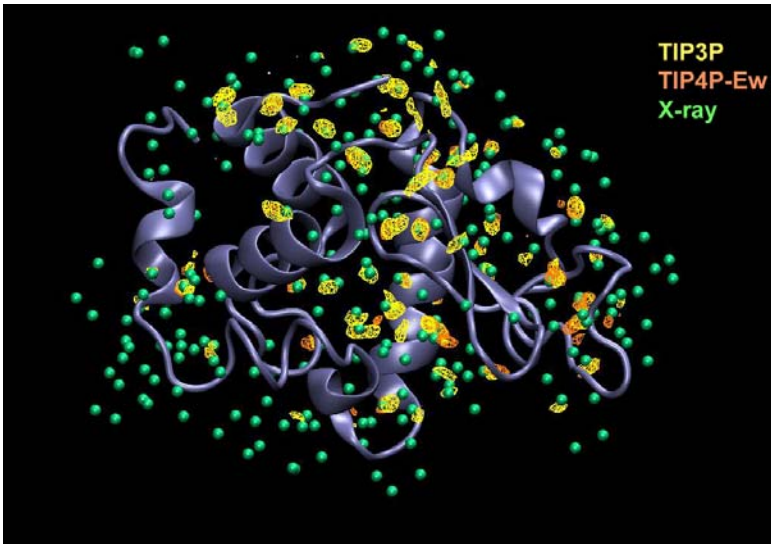


Figure 7-26. Water density seen in the TIP3P (yellow) and TIP4P-Ew (orange) simulation mapped on the X-ray structure of lysozyme. The crystal waters (green) are mapped on the density grid in order to look at fraction occupancy.

7.4 Conclusions

In this work, we compared conformational preferences and energetics in the TIP3P and TIP4P-Ew solvent model using short polyalanines, a model peptide with an ion pair and lysozyme as our larger test case. For Ala_3, Ala_5 and the model peptide with an ion pair, we ran REMD simulations in order to obtain equilibrium populations for ensembles in both water models. Standard molecular dynamics simulations of lysozyme were run in order to determine if the same effects were seen with the smaller peptide systems were translated to a larger protein. We also ran standard molecular dynamics of Ala_3 in order to investigate the effect of water models of conformational transition rates.

For the small polyalanines, we found that φ/ψ populations of the backbone and regions of higher water density were relatively similar in both water models. The temperature dependent properties of the backbone were also similar in both TIP3P and TIP4P-Ew. Transitions rates for the backbone of Ala_3 were up to four times higher in

TIP3P than in TIP4P-Ew. This did not appear to have an effect on the overall populations of dihedral angles. In the peptide with the ion-pair, we observed different behavior in both water models. Salt bridges in the TIP4P-Ew model were less stable than in TIP3P and did not exhibit dependent temperature behavior. In lysozyme, the trends were consistent with the small model peptides. The backbone appeared to be stable throughout all of the regions except for fluctuations in the loop of the protein in both water models. A native salt bridge between Asp48 and Arg61 appeared to be less stable in TIP4P-Ew than in TIP3P model by 1 kcal/mol. In addition, the water occupancy of both models was comparable, occupying the highest density around the location of a buried crystal water.

From these studies, the differences in peptide conformations and energetics are quite small in the TIP3P and TIP4P-Ew water models. Charged residues appear to be more sensitive to the choice of water model. Recent NMR relaxation experiments from Tribovic et al. [246] have suggested that salt bridges in TIP3P with ff99SB are too rigid; our results suggest that simulations in TIP4P-Ew may be better at reproducing the details of these sidechain interactions. Other NMR structural and relaxation studies have also confirmed that TIP4P-Ew gives better agreement with experimental results than TIP3P. Nevertheless, TIP3P appears to perform quite well on non-charged residues and has advantages such as reduced computational cost and more rapid convergence of thermodynamic properties (although likely at the expense of reduced accuracy for kinetics analysis). These properties may make TIP3P desirable for some studies.

References

1. Fiser, A., Feig, M., Brooks, C.L., and Sali, A., *Evolution and physics in comparative protein structure modeling.* Accounts of Chemical Research, 2002. **35**(6): p. 413-421.

2. Park, S., Xi, Y., and Saven, J.G., *Advances in computational protein design.* Current Opinion in Structural Biology, 2004. **14**(4): p. 487-494.

3. Bradley, P., Misura, K.M.S., and Baker, D., *Toward high-resolution de novo structure prediction for small proteins.* Science, 2005. **309**(5742): p. 1868-1871.

4. Schueler-Furman, O., Wang, C., Bradley, P., Misura, K., and Baker, D., *Progress on modeling of protein structures and interactions.* Science, 2005. **310**(5748): p. 638-642.

5. Totrov, M. and Abagyan, R., *Flexible ligand docking to multiple receptor conformations: a practical alternative.* Current Opinion in Structural Biology, 2008. **18**(2): p. 178-184.

6. Ding, F., Layten, M., and Simmerling, C., *Solution structure of HIV-1 protease flaps probed by comparison of molecular dynamics simulation ensembles and EPR experiments.* Journal of the American Chemical Society, 2008. **129**(30): p. 11004–11005.

7. Anfinsen, C.B., *Principles that govern the folding of protein chains.* Science, 1973. **181**: p. 223.

8. Levinthal, C., *Are there pathways for protein folding?* Journal de Chimie Physique, 1968. **85**: p. 44.

9. Dobson, C.M., Sali, A., and Karplus, M., *Protein Folding: A Perspective from Theory and Experiment.* Angewandte Chemie International Edition, 1998. **37**: p.

868-893.

10. Shortle, D., *The denatured state (the other half of the folding equation) and its role in protein stability.* Faseb, 1996. **10**: p. 27.

11. Cho, J.H., Sato, S., and Raleigh, D.P., *Thermodynamics and kinetics of non-native interactions in protein folding: A single point mutant significantly stabilizes the N-terminal domain of L9 by modulating non-native interactions in the denatured state.* Journal of Molecular Biology, 2004. **338**(4): p. 827-837.

12. Anil, B., Song, B.B., Tang, Y.F., and Raleigh, D.P., *Exploiting the right side of the ramachandran plot: Substitution of glycines by D-alanine can significantly increase protein stability.* Journal of the American Chemical Society, 2004. **126**(41): p. 13194-13195.

13. Anil, B., Craig-Schapiro, R., and Raleigh, D.P., *Design of a hyperstable protein by rational consideration of unfolded state interactions.* Journal of the American Chemical Society, 2006. **128**(10): p. 3144-3145.

14. Moore, R.A., Taubner, L.M., and Priola, S.A., *Prion protein misfolding and disease.* Current Opinion in Structural Biology, 2009. **19**(1): p. 14-22.

15. Chiti, F. and Dobson, C.M., *Protein misfolding, functional amyloid, and human disease.* Annual Review of Biochemistry, 2006. **75**: p. 333-366.

16. Fersht, A., *Structure and mechanism in protein science : a guide to enzyme catalysis and protein folding*. 1999, W.H. Freeman, New York. xxi, 631 p.

17. Plaxco, K. and Gross, M., *Unfolded, yes, but random? Never!* Nature Structural Biology, 2001. **8**(8): p. 659-660.

18. Shortle, D. and Ackerman, M.S., *Persistence of Native-like Topology in a Denatured Protein in 8M Urea.* Science, 2001. **293**: p. 487-489.

19. Li, Y., Horng, J.C., and Raleigh, D.P., *pH dependent thermodynamic and amide exchange studies of the C-terminal domain of the ribosomal protein L9: Implications for unfolded state structure.* Biochemistry, 2006. **45**(28): p. 8499-8506.

20. Li, Y., Shan, B., and Raleigh, D.P., *The cold denatured state is compact but expands at low temperatures: Hydrodynamic properties of the cold denatured state of the C-terminal domain of L9.* Journal of Molecular Biology, 2007. **368**(1): p. 256-262.

21. Anil, B., Li, Y., Cho, J.H., and Raleigh, D.P., *The unfolded state of NTL9 is compact in the absence of denaturant.* Biochemistry, 2006. **45**(33): p. 10110-10116.

22. Mayor, U., Guydosh, N.R., Johnson, C.M., Grossmann, J.G., Sato, S., Jas, G.S., Freund, S.M., Alonso, D.O., Daggett, V., and Fersht, A.R., *The complete folding pathway of a protein from nanoseconds to microseconds.* Nature, 2003. **421**(6925): p. 863-867.

23. Brockwell, D.J., Smith, D.A., and Radford, S.E., *Protein folding mechanisms: new methods and emerging ideas.* Current Opinion in Structural Biology, 2000. **10**(1): p. 16-25.

24. Karplus, M. and Weaver, D.L., *Protein folding dynamics: the diffusion-collision model and experimental data.* Protein Science 1994. **3**(4): p. 650-668.

25. Islam, S.A., Karplus, M., and Weaver, D.L., *Application of the diffusion-collision model to the folding of three-helix bundle proteins.* Journal of Molecular Biology, 2002. **318**(1): p. 199-215.

26. Tang, Y., Rigotti, D.J., Fairman, R., and Raleigh, D.P., *Peptide models provide evidence for significant structure in the denatured state of a rapidly folding protein: the villin headpiece subdomain.* Biochemistry, 2004. **43**(11): p. 3264-3272.

27. Horng, J.C., Moroz, V., Rigotti, D.J., Fairman, R., and Raleigh, D.P., *Characterization of large peptide fragments derived from the N-terminal domain of the ribosomal protein L9: Definition of the minimum folding motif and characterization of local electrostatic interactions.* Biochemistry, 2002. **41**(45): p. 13360-13369.

28. Schaeffer, R.D., Fersht, A., and Daggett, V., *Combining experiment and simulation in protein folding: closing the gap for small model systems.* Current Opinion in Structural Biology, 2008. **18**(1): p. 4-9.

29. Wickstrom, L., Bi, Y., Hornak, V., Raleigh, D.P., and Simmerling, C., *Reconciling the solution and X-ray structures of the villin headpiece helical subdomain: molecular dynamics simulations and double mutant cycles reveal a stabilizing cation-pi interaction.* Biochemistry, 2007. **46**(12): p. 3624-3634.

30. Fan, H. and Mark, A.E., *Relative stability of protein structures determined by X-ray crystallography or NMR spectroscopy: A molecular dynamics simulation study.* Proteins: Structure, Function and Genetics, 2003. **53**(1): p. 111-120.

31. Layten, M., Hornak, V., and Simmerling, C., *The open structure of a multi-drug-resistant HIV-1 protease is stabilized by crystal packing contacts.* Journal of the American Chemical Society, 2006. **128**(41): p. 13360-13361.

32. Bagchi, B., *Water dynamics in the hydration layer around proteins and micelles.* Chemical Reviews, 2005. **105**(9): p. 3197-3219.

33. Simmerling, C., Strockbine, B., and Roitberg, A.E., *All-atom structure prediction and folding simulations of a stable protein.* Journal of the American Chemical Society, 2002. **124**(38): p. 11258-11259.

34. Duan, Y. and Kollman, P.A., *Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution.* Science, 1998. **282**(5389): p. 740-744.

35. Snow, C.D., Sorin, E.J., Rhee, Y.M., and Pande, V.S., *How well can simulation predict protein folding kinetics and thermodynamics?* Annu Rev Biophys Biomol Struct, 2005. **34**: p. 43-69.

36. Day, R., Bennion, B.J., Ham, S., and Daggett, V., *Increasing temperature accelerates protein unfolding without changing the pathway of unfolding.* J Mol Biol, 2002. **322**(1): p. 189-203.

37. Okur, A., Strockbine, B., Hornak, V., and Simmerling, C., *Using PC clusters to evaluate the transferability of molecular mechanics force fields for proteins.* Journal of Computational Chemistry, 2003. **24**(1): p. 21-31.

38. Leach, A.R., *Molecular Modeling*. 3rd ed. 1998, Addison Wesley Longman Limited, London. 595.

39. Case, D.A., Cheatham III, T., Darden, T., Gohlke, H., Luo, R., Merz Jr, K.M., Onufriev, A., Simmerling, C., Wang, B., and R.W., W., *The Amber biomolecular simulation programs.* Journal of Computational Chemistry, 2005. **26**(16): p. 1668-1688.

40. Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., and Karplus, M., *Charmm - a Program for Macromolecular Energy, Minimization, and Dynamics Calculations.* Journal of Computational Chemistry, 1983. **4**(2): p. 187-217.

41. Jorgensen, W.L., Maxwell, D.S., and Tirado Rives, J., *Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids.* Journal of the American Chemical Society, 1996. **118**(45): p. 11225-11236.

42. Eising, A.A., H., H.P., Mark, A.E., Scott, W.R.P., and Tironi, I.G., *Biomolecular simulation: the GROMOS96 manual and user guide*. 1996, Vdf Hochschulverlag. ETH Zurich.

43. Best, R.B., Buchete, N.V., and Hummer, G., *Are current molecular dynamics force fields too helical?* Biophysical Journal, 2008. **95**(1): p. L7-L9.

44. Graf, J., Nguyen, P.H., Stock, G., and Schwalbe, H., *Structure and Dynamics of the Homologous Series of Alanine Peptides: A Joint Molecular Dynamics/NMR Study.* Journal of the American Chemical Society, 2007. **129**(5): p. 1179-1189.

45. Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., and Simmerling, C., *Comparison of multiple amber force fields and development of improved protein backbone parameters.* Proteins- Structure, Function, and Bioinformatics, 2006. **65**(3): p. 712-725.

46. Koller, A.N., Schwalbe, H., and Gohlke, H., *Starting structure dependence of NMR order parameters derived from MD simulations: Implications for judging force-field quality.* Biophysical Journal, 2008. **95**(1): p. L4-L6.

47. Showalter, S.A. and Bruschweiler, R., *Quantitative molecular ensemble interpretation of NMR dipolar couplings without restraints.* Journal of the American Chemical Society, 2007. **129**(14): p. 4158-4159.

48. Trbovic, N., Kim, B., Friesner, R.A., and Palmer, A.G., 3rd, *Structural analysis of*

protein dynamics by MD simulations and NMR spin-relaxation. Proteins: Structure, Function and Genetics, 2008. **71**(2): p. 684-694.

49. Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W., and Klein, M.L., *Comparison of simple potential functions for simulating liquid water.* Journal of Chemical Physics, 1983. **79**(2): p. 926-935.

50. Horn, H.W., Swope, W.C., Pitera, J.W., Madura, J.D., Dick, T.J., Hura, G.L., and Head-Gordon, T., *Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew.* Journal of Chemical Physics, 2004. **120**(20): p. 9665-9678.

51. Still, W.C., Tempczyk, A., Hawley, R.C., and Hendrickson, T., *Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics.* Journal of the American Chemical Society, 1990. **112**(16): p. 6127-6129.

52. Onufriev, A., *Implicit Solvent Models in Molecular Dynamics Simulations*, in *Annual Reports in Computational Chemistry*, D.C. Spellmeyer, Editor. 2008, Elsevier: Amsterdam. p. 125-137.

53. Zhou, R., *Free energy landscape of protein folding in water: explicit vs. implicit solvent.* Proteins: Structure, Function and Genetics, 2003. **53**(2): p. 148-161.

54. Zhou, R.H. and Berne, B.J., *Can a Continuum Solvent Model Reproduce the Free Energy Landscape of a Beta-hairpin Folding in Water?* Proceedings from the National Academy of Science of the United States of America, 2002. **99**(20): p. 12777-12782.

55. Pitera, J.W. and Swope, W., *Understanding folding and design: replica-exchange simulations of "Trp-cage" miniproteins.* Proceedings from the National Academy of Science of the United States of America, 2003. **100**(13): p. 7587-7592.

56. Geney, R., Layten, M., Gomperts, R., Hornak, V., and Simmerling, C., *Investigation of salt bridge stability in a generalized born solvent model.* Journal of Chemical Theory and Computation, 2006. **2**(1): p. 115-127.

57. Okur, A., Wickstrom, L., and Simmerling, C., *Evaluation of salt bridge structure and energetics in peptides using explicit, implicit, and hybrid solvation models.* Journal of Chemical Theory and Computation, 2008. **4**(3): p. 488-498.

58. Okur, A., Wickstrom, L., Layten, M., Geney, R., Song, K., Hornak, V., and Simmerling, C., *Improved Efficiency of Replica Exchange Simulations through Use of a Hybrid Explicit/implicit Solvation Model.* Journal of Chemical Theory and Computation, 2006. **2**(2): p. 420-433.

59. Roe, D.R., Okur, A., Wickstrom, L., Hornak, V., and Simmerling, C., *Secondary structure bias in generalized born solvent models: Comparison of conformational ensembles and free energy of solvent polarization from explicit and implicit solvation.* Journal of Physical Chemistry B, 2007. **111**(7): p. 1846-1857.

60. Mobley, D.L., Bayly, C.I., Cooper, M.D., Shirts, M.R., and Dill, K.A., *Small Molecule Hydration Free Energies in Explicit Solvent: An Extensive Test of Fixed-Charge Atomistic Simulations.* Journal of Chemical Theory and Computation, 2009. **5**(2): p. 350-358.

61. Mobley, D.L., Barber, A.E., Fennell, C.J., and Dill, K.A., *Charge asymmetries in hydration of polar solutes.* Journal of Physical Chemistry B, 2008. **112**(8): p. 2405-2414.

62. Sugita, Y. and Okamoto, Y., *Replica-exchange molecular dynamics method for protein folding.* Chemical Physics Letters, 1999. **314**(1-2): p. 141-151.

63. Hansmann, U.H.E., *Parallel tempering algorithm for conformational studies of biological molecules.* Chemical Physics Letters, 1997. **281**(1-3): p. 140-150.

64. Tai, K., *Conformational sampling for the impatient.* Biophysical Chemistry, 2004. **107**(3): p. 213-220.

65. Rhee, Y.M. and Pande, V.S., *Multiplexed-replica exchange molecular dynamics method for protein folding simulation.* Biophysical Journal, 2003. **84**(2 Pt 1): p. 775-786.

66. Zhou, R., Berne, B.J., and Germain, R., *The free energy landscape for beta hairpin folding in explicit water.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(26): p. 14931-14936.

67. Sugita, Y. and Okamoto, Y., *Replica-exchange molecular dynamics method for protein folding. Chem. Phys. Lett.*, 1999. **314**(1-2): p. 141-151.

68. Machta, J., Newman, M.E., and Chayes, L.B., *Replica-exchange algorithm and results for the three-dimensional random field ising model.* Physical Review E Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics, 2000. **62**(6 Pt B): p. 8782-8789.

69. Sikorski, A., Kolinski, A., and Skolnick, J., *Computer simulations of protein folding with a small number of distance restraints.* Acta Biochimica Polonica, 2002. **49**(3): p. 683-692.

70. Jang, S., Shin, S., and Pak, Y., *Replica-exchange method using the generalized effective potential.* Physical Review Letters, 2003. **91**(5): p. 058305.

71. Fukunishi, H., Watanabe, O., and Takada, S., *On the Hamiltonian Replica Exchange Method for Efficient Sampling of Biomolecular Systems: Application to Protein Structure Prediction.* Journal of Chemical Physics, 2002. **116**(20): p. 9058-9067.

72. Zagrovic, B. and Pande, V., *Solvent viscosity dependence of the folding rate of a small protein: Distributed computing study.* Journal of Computational Chemistry, 2003. **24**(12): p. 1432-1436.

73. Jacob, M. and Schmid, F.X., *Protein folding as a diffusional process.* Biochemistry, 1999. **38**(42): p. 13773-13779.

74. Rhee, Y.M. and Pande, V.S., *Solvent viscosity dependence of the protein folding dynamics.* Journal of Physical Chemistry B, 2008. **112**(19): p. 6221-6227.

75. Zheng, W., Andrec, M., Gallicchio, E., and Levy, R.M., *Simulating Replica Exchange Simulations of Protein Folding with a Kinetic Network Model.* Proceedings of the National Academy of Science of the United States of America, 2007. **104**(39): p. 15340-15345.

76. Oliveberg, M., Tan, Y.J., and Fersht, A.R., *Negative activation enthalpies in the kinetics of protein folding.* Proceedings of the National Academy of Sciences of the United States of America, 1995. **92**(19): p. 8926-8929.

77. Munoz, V., Thompson, P.A., Hofrichter, J., and Eaton, W.A., *Folding dynamics and mechanism of beta-hairpin formation.* Nature, 1997. **390**(6656): p. 196-199.

78. Scalley, M.L. and Baker, D., *Protein folding kinetics exhibit an Arrhenius temperature dependence when corrected for the temperature dependence of protein stability.* Proceedings of the National Academy of Science of the United States of America, 1997. **94**(20): p. 10636-10640.

79. Wickstrom, L., Okur, A., Song, K., Hornak, V., Raleigh, D.P., and Simmerling, C.L., *The unfolded state of the villin headpiece helical subdomain: Computational studies of the role of locally stabilized structure.* Journal of Molecular Biology, 2006. **360**(5): p. 1094-1107.

80. Garcia, A.E. and Sanbonmatsu, K.Y., *Alpha-helical stabilization by sidechain shielding of backbone hydrogen bonds.* Proceedings from the National Academy of Science of the United States of America, 2002. **99**(5): p. 2782-2787.

81. Liu, P., Kim, B., Friesner, R.A., and Berne, B.J., *Replica exchange with solute tempering: A method for sampling biological systems in explicit water.* Proceedings of the National Academy of Science of the United States of America, 2005. **102**(39): p. 13749-13754.

82. Okur, A., Roe, D.R., Cui, G.L., Hornak, V., and Simmerling, C., *Improving Convergence of Replica-exchange Simulations through Coupling to a High-temperature Structure Reservoir.* Journal of Chemical Theory and Computation, 2007. **3**(2): p. 557-568.

83. Hritz, J. and Oostenbrink, C., *Hamiltonian replica exchange molecular dynamics using soft-core interactions.* Journal of Chemical Physics, 2008. **128**(14): p. 144121.

84. Li, H., Li, G., Berg, B.A., and Yang, W., *Finite Reservoir Replica Exchange to Enhance Canonical Sampling in Rugged Energy Surfaces.* Journal of Chemical Physics, 2006. **125**(14): p. 144902.

85. Khurana, S. and George, S.P., *Regulation of cell structure and function by actin-binding proteins: villin's perspective.* FEBS Letters, 2008. **582**(14): p. 2128-2139.

86. Vardar, D., Buckley, D.A., Frank, B.S., and McKnight, C.J., *NMR structure of an F-actin-binding "headpiece" motif from villin.* Journal of Molecular Biology, 1999. **294**(5): p. 1299-1310.

87. McKnight, C.J., Doering, D.S., Matsudaira, P.T., and Kim, P.S., *A thermostable*

35-residue subdomain within villin headpiece. Journal of Molecular Biology, 1996. **260**(2): p. 126-134.

88. McKnight, C.J., Matsudaira, P.T., and Kim, P.S., *NMR structure of the 35-residue villin headpiece subdomain.* Nature Structural Biology, 1997. **4**(3): p. 180-184.

89. Chiu, T.K., Kubelka, J., Herbst-Irmer, R., Eaton, W.A., Hofrichter, J., and Davies, D.R., *High-resolution x-ray crystal structures of the villin headpiece subdomain, an ultrafast folding protein.* Proceedings of the National Academy of Science of the United States of America, 2005. **102**(21): p. 7517-7522.

90. Kubelka, J., Chiu, T.K., Davies, D.R., Eaton, W.A., and Hofrichter, J., *Sub-microsecond protein folding.* Journal of Molecular Biology, 2006. **359**(3): p. 546-553.

91. Bi, Y., Cho, J.H., Kim, E.Y., Shan, B., Schindelin, H., and Raleigh, D.P., *Rational design, structural and thermodynamic characterization of a hyperstable variant of the villin headpiece helical subdomain.* Biochemistry, 2007. **46**(25): p. 7497-7505.

92. Zagrovic, B., Snow, C.D., Shirts, M.R., and Pande, V.S., *Simulation of folding of a small alpha-helical protein in atomistic detail using worldwide-distributed computing.* Journal of Molecular Biology, 2002. **323**(5): p. 927-937.

93. Zagrovic, B., Snow, C.D., Khaliq, S., Shirts, M.R., and Pande, V.S., *Native-like mean structure in the unfolded ensemble of small proteins.* Journal of Molecular Biology, 2002. **323**(1): p. 153-164.

94. Islam, S.A., Karplus, M., and Weaver, D.L., *Application of the diffusion-collision model to the folding of three-helix bundle proteins.* J Mol Biol, 2002. **318**(1): p. 199-215.

95. Shen, M.Y. and Freed, K.F., *All-atom fast protein folding simulations: the villin headpiece.* Proteins, 2002. **49**(4): p. 439-445.

96. Srinivas, G. and Bagchi, B., *Folding and unfolding of chicken villin headpiece: Energy landscape of a single-domain model protein.* Current Science, 2002. **82**(2): p. 179-185.

97. Frank, B.S., Vardar, D., Buckley, D.A., and McKnight, C.J., *The role of aromatic residues in the hydrophobic core of the villin headpiece subdomain.* Protein Science, 2002. **11**(3): p. 680-687.

98. Wang, M., Tang, Y., Sato, S., Vugmeyster, L., McKnight, C.J., and Raleigh, D.P., *Dynamic NMR line-shape analysis demonstrates that the villin headpiece subdomain folds on the microsecond time scale.* Journal of the American Chemical Society, 2003. **125**(20): p. 6032-6033.

99. Kubelka, J., Eaton, W.A., and Hofrichter, J., *Experimental tests of villin subdomain folding simulations.* Journal of Molecular Biology, 2003. **329**(4): p. 625-630.

100. Brewer, S.H., Vu, D.M., Tang, Y., Li, Y., Franzen, S., Raleigh, D.P., and Dyer, R.B., *Effect of modulating unfolded state structure on the folding kinetics of the villin headpiece subdomain.* Proceedings of the National Academy of Science of the United States of America, 2005. **102**(46): p. 16662-16667.

101. Fernandez, A., Shen, M.Y., Colubri, A., Sosnick, T.R., Berry, R.S., and Freed, K.F., *Large-scale context in protein folding: villin headpiece.* Biochemistry, 2003. **42**(3): p. 664-671.

102. Jang, S., Kim, E., Shin, S., and Pak, Y., *Ab initio folding of helix bundle proteins using molecular dynamics simulations.* Journal of the American Chemical Society, 2003. **125**(48): p. 14841-14846.

103. Lin, C.Y., Hu, C.K., and Hansmann, U.H., *Parallel tempering simulations of HP-36.* Proteins, 2003. **52**(3): p. 436-445.

104. Buscaglia, M., Kubelka, J., Eaton, W.A., and Hofrichter, J., *Determination of ultrafast protein folding rates from loop formation dynamics.* Journal of Molecular Biology, 2005. **347**(3): p. 657-664.

105. Vugmeyster, L., Trott, O., McKnight, C.J., Raleigh, D.P., and Palmer, A.G., 3rd, *Temperature-dependent dynamics of the villin headpiece helical subdomain, an unusually small thermostable protein.* Journal of Molecular Biology, 2002. **320**(4): p. 841-854.

106. Havlin, R.H. and Tycko, R., *Probing site-specific conformational distributions in protein folding with solid-state NMR.* Proceedings of the National Academy of

Science of the United States of America, 2005. **102**(9): p. 3284-3289.

107. Tang, Y., Goger, M.J., and Raleigh, D.P., *NMR characterization of a peptide model provides evidence for significant structure in the unfolded state of the villin headpiece helical subdomain.* Biochemistry, 2006. **45**(22): p. 6940-6946.

108. De Mori, G.M., Colombo, G., and Micheletti, C., *Study of the Villin headpiece folding dynamics by combining coarse-grained Monte Carlo evolution and all-atom molecular dynamics.* Proteins, 2005. **58**(2): p. 459-471.

109. van der Spoel, D. and Lindahl, E., *Brute-Force Molecular Dynamics Simulations of Villin Headpiece: Comparison with NMR Parameters.* Journal of Physical Chemistry B, 2003. **107**(40): p. 11178-11187.

110. Duan, Y., Wang, L., and Kollman, P.A., *The early stage of folding of villin headpiece subdomain observed in a 200-nanosecond fully solvated molecular dynamics simulation.* Proceedings of the National Academy of Science of the United States of America, 1998. **95**(17): p. 9897-9902.

111. Fernandez, A., Shen, M.Y., Colubri, A., Sosnick, T.R., Berry, R.S., and Freed, K.F., *Large-scale context in protein folding: villin headpiece.* Biochemistry, 2003. **42**(3): p. 664-71.

112. Sullivan, D.C. and Kuntz, I.D., *Conformation spaces of proteins.* Proteins, 2001. **42**(4): p. 495-511.

113. Sullivan, D.C. and Kuntz, I.D., *Protein folding as biased conformational diffusion.* Journal of Physical Chemistry B, 2002. **106**(12): p. 3255-3262.

114. Ripoll, D.R., Vila, J.A., and Scheraga, H.A., *Folding of the villin headpiece subdomain from random structures. Analysis of the charge distribution as a function of pH.* Journal of Molecular Biology, 2004. **339**(4): p. 915-925.

115. Bandyopadhyay, S., Chakraborty, S., Balasubramanian, S., and Bagchi, B., *Sensitivity of polar solvation dynamics to the secondary structures of aqueous proteins and the role of surface exposure of the probe.* Journal of the American Chemical Society, 2005. **127**(11): p. 4071-4075.

116. Herges, T. and Wenzel, W., *Free-energy landscape of the villin headpiece in an*

all-atom force field. Structure, 2005. **13**(4): p. 661-668.

117. Trebst, S., Troyer, M., and Hansmann, U.H.E., *Optimized parallel tempering simulations of proteins.* The Journal of Chemical Physics, 2006. **124**(17): p. 174903.

118. Jayachandran, G., Vishal, V., and Pande, V.S., *Using massively parallel simulation and Markovian models to study protein folding: examining the dynamics of the villin headpiece.* Journal of Chemical Physics, 2006. **124**(16): p. 164902.

119. Zagrovic, B. and van Gunsteren, W.F., *Comparing atomistic simulation data with the NMR experiment: how much can NOEs actually tell us?* Proteins, 2006. **63**(1): p. 210-218.

120. Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz Jr, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W., and Kollman, P.A., *A second generation force field for the simulation of proteins, nucleic acids, and organic molecules.* Journal of the American Chemical Society, 1995. **117**(19): p. 5179-5197.

121. Wang, J., Cieplak, P., and Kollman, P.A., *How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?* Journal of Computational Chemistry, 2000. **21**(12): p. 1049-1074.

122. Ryckaert, J.-P., Ciccotti, G., and Berendsen, H.J.C., *Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes.* Journal of Computational Physics, 1977. **23**(3): p. 327-341.

123. Berendsen, H.J.C., Postma, J.P.M., van Gunsteren, W.F., DiNola, A., and Haak, J.R., *Molecular dynamics with coupling to an external bath.* Journal of Chemical Physics, 1984. **81**(8): p. 3684-3690.

124. Shirts, M.R. and Pande, V.S., *Solvation free energies of amino acid side chain analogs for common molecular mechanics water models.* The Journal of Chemical Physics, 2005. **122**(13): p. 134508.

125. Darden, T., York, D., and Pedersen, L., *Particle mesh Ewald: an N.log(N) method for Ewald sums in large systems.* Journal of Chemical Physics, 1993. **98**(12): p. 10089-10092.

126. Wu, X.W. and Brooks, B.R., *Isotropic periodic sum: A method for the calculation of long-range interactions.* Journal of Chemical Physics, 2005. **122**(4): p. 1-18.

127. Simmerling, C., Elber, R. and Zhang, J., , *MOIL-View - A Program for Visualization of Structure and Dynamics of Biomolecules and STO- A Program for Computing Stochastic Paths, in Modelling of Biomolecular Structure and Mechanisms*, in *Modeling of Biomolecular Structure and Mechanism*, A. Pullman, Editor. 1995: Kluwer, Netherlands. p. 241-265.

128. Kabsch, W. and Sander, C., *Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features.* Biopolymers, 1983. **22**(12): p. 2577-2637.

129. Mello, C.C. and Barrick, D., *Measuring the stability of partly folded proteins using TMAO.* Protein Science, 2003. **12**(7): p. 1522-1529.

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

131. Spector, S., Young, P., and Raleigh, D.P., *Nativelike structure and stability in a truncation mutant of a protein minidomain: the peripheral subunit-binding domain.* Biochemistry, 1999. **38**(13): p. 4128-4136.

132. Shi, Z., Olson, C.A., Bell, A.J.J., and Kallenbach, N.R., *Stabilization of alpha-helix structure by polar side-chain interactions: Complex salt bridges, cation-pi interactions, and C-H...O H-bonds.* Biopolymers, 2001. **60**(5): p. 366-380.

133. Gallivan, J.P. and Dougherty, D.A., *Cation-pi interactions in structural biology.* Proceedings of the National Academy of Science of the United States of America, 1999. **96**(17): p. 9459-9464.

134. Dougherty, D.A., *Cation-pi interactions in chemistry and biology: a new view of benzene, Phe, Tyr, and Trp.* Science, 1996. **271**(5246): p. 163-168.

135. Vermeulen, W., Vanhaesebrouck, P., Van Troys, M., Verschueren, M., Fant, F., Goethals, M., Ampe, C., Martins, J.C., and Borremans, F.A., *Solution structures of the C-terminal headpiece subdomains of human villin and advillin, evaluation of headpiece F-actin-binding requirements.* Protein Science, 2004. **13**(5): p. 1276-1287.

136. Hunenberger, P.H. and McCammon, J.A., *Effect of artificial periodicity in simulations of biomolecules under Ewald boundary conditions: a continuum electrostatics study.* Biophysical Chemistry, 1999. **78**(1-2): p. 69-88.

137. Hunenberger, P.H. and McCammon, J.A., *Ewald artifacts in computer simulations of ionic solvation and ion-ion interaction: A continuum electrostatics study.* Journal of Chemical Physics, 1999. **110**(4): p. 1856-1872.

138. Weber, W., Hunenberger, P.H., and McCammon, J.A., *Molecular dynamics simulations of a polyalanine octapeptide under Ewald boundary conditions: Influence of artificial periodicity on peptide conformation.* Journal of Physical Chemistry B, 2000. **104**(15): p. 3668-3675.

139. Religa, T.L., Markson, J.S., Mayor, U., Freund, S.M., and Fersht, A.R., *Solution structure of a protein denatured state and folding intermediate.* Nature, 2005. **437**(7061): p. 1053-1056.

140. Myers, J.K. and Oas, T.G., *Preorganized secondary structure as an important determinant of fast protein folding.* Nature Structural Biology, 2001. **8**(6): p. 552-558.

141. Shi, Z., Olson, C.A., Rose, G.D., Baldwin, R.L., and Kallenbach, N.R., *Polyproline II structure in a sequence of seven alanine residues.* Proceedings of the National Academy of Science of the United States of America, 2002. **99**(14): p. 9190-9195.

142. Mezei, M., Fleming, P.J., Srinivasan, R., and Rose, G.D., *Polyproline II helix is the preferred conformation for unfolded polyalanine in water.* Proteins, 2004. **55**(3): p. 502-507.

143. Kentsis, A., Mezei, M., Gindin, T., and Osman, R., *Unfolded state of polyalanine is a segmented polyproline II helix.* Proteins, 2004. **55**(3): p. 493-501.

144. Asher, S.A., Mikhonin, A.V., and Bykov, S., *UV Raman demonstrates that alpha-helical polyalanine peptides melt to polyproline II conformations.* Journal of the American Chemical Society, 2004. **126**(27): p. 8433-8440.

145. Shortle, D., *The denatured state (the other half of the folding equation) and its role in protein stability.* Journal of the Federation of American Societies for Experimental Biology, 1996. **10**: p. 27.

146. Cho, J.H. and Raleigh, D.P., *Mutational analysis demonstrates that specific electrostatic interactions can play a key role in the denatured state ensemble of proteins.* Journal of Molecular Biology, 2005. **353**(1): p. 174-185.

147. Zhang, O. and Forman-Kay, J.D., *NMR studies of unfolded states of an SH3 domain in aqueous solution and denaturing conditions.* Biochemistry, 1997. **36**(13): p. 3959-3970.

148. Mayor, U., Grossmann, J.G., Foster, N.W., Freund, S.M., and Fersht, A.R., *The denatured state of Engrailed Homeodomain under denaturing and native conditions.* Journal of Molecular Biology, 2003. **333**(5): p. 977-991.

149. Kinnear, B.S., Jarrold, M.F., and Hansmann, U.H., *All-atom generalized-ensemble simulations of small proteins.* Journal of Molecular Graphics and Modelling, 2004. **22**(5): p. 397-403.

150. Manning, M.C. and Woody, R.W., *Theoretical CD studies of polypeptide helices: examination of important electronic and geometric factors.* Biopolymers, 1991. **31**(5): p. 569-586.

151. Chin, D.H., Woody, R.W., Rohl, C.A., and Baldwin, R.L., *Circular dichroism spectra of short, fixed-nucleus alanine helices.* Proceedings of the National Academy of Science of the United States of America, 2002. **99**(24): p. 15416-15421.

152. Siedlecka, M., Goch, G., Ejchart, A., Sticht, H., and Bierzyski, A., *Alpha-helix nucleation by a calcium-binding peptide loop.* Proceedings of the National Academy of Science of the United States of America, 1999. **96**(3): p. 903-908.

153. Gnanakaran, S. and Garcia, A.E., *Helix-coil transition of alanine peptides in water: force field dependence on the folded and unfolded structures.* Proteins, 2005. **59**(4): p. 773-782.

154. Roe, D.R., Hornak, V., and Simmerling, C., *Folding cooperativity in a three-stranded beta-sheet model.* Journal of Molecular Biology, 2005. **352**(2): p. 370-381.

155. Yang, W.Y., Pitera, J.W., Swope, W.C., and Gruebele, M., *Heterogeneous folding of the trpzip hairpin: full atom simulation and experiment.* Journal of Molecular Biology, 2004. **336**(1): p. 241-251.

156. Nymeyer, H. and Garcia, A.E., *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 Science of the United States of America, 2003. **100**(24): p. 13934-13939.

157. Rao, F. and Caflisch, A., *Replica exchange molecular dynamics simulations of reversible folding.* The Journal of Chemical Physics, 2003. **119**(7): p. 4035-4042.

158. Hornak, V., Okur, A., Rizzo, R.C., and Simmerling, C., *HIV-1 protease flaps spontaneously open and reclose in molecular dynamics simulations.* Proceedings of the National Academy of Science of the United States of America, 2006. **103**(4): p. 915-920.

159. Hawkins, G.D., Cramer, C.J., and Truhlar, D.G., *Pairwise solute descreening of solute charges from a dielectric medium.* Chemical Physics Letters, 1995. **246**(1-2): p. 122-129.

160. Bondi, A., *Van Der Waals Volumes + Radii.* Journal of Physical Chemistry, 1964. **68**(3): p. 441-451.

161. Vuister, G.W., Wang, A.C., and Bax, A., *Measurement of 3-Bond Nitrogen Carbon-J Couplings in Proteins Uniformly Enriched in N-15 and C-13.* Journal of the American Chemical Society, 1993. **115**(12): p. 5334-5335.

162. Smith, L.J., Daura, X., and van Gunsteren, W.F., *Assessing equilibration and convergence in biomolecular simulations.* Proteins, 2002. **48**(3): p. 487-496.

163. Feig, M. and Brooks, C.L., 3rd, *Recent advances in the development and application of implicit solvent models in biomolecule simulations.* Current Opinion in Structural Biology, 2004. **14**(2): p. 217-224.

164. Fan, H., Mark, A.E., Zhu, J., and Honig, B., *Comparative study of generalized Born models: protein dynamics.* Proceedings of the National Academy of Science of the United States of America, 2005. **102**(19): p. 6760-6764.

165. Onufriev, A., Bashford, D., and Case, D.A., *Exploring protein native states and large-scale conformational changes with a modified generalized born model.* Proteins, 2004. **55**(2): p. 383-394.

166. Hong, Q. and Schellman, J.A., *Helix-Coil Theories - a Comparative-Study for Finite Length Polypeptides.* Journal of Physical Chemistry, 1992. **96**(10): p. 3987-3994.

167. Lifson, S. and Roig, A., *On the Theory of Helix-Coil Transition in Polypeptides.* Journal of Chemical Physics, 1961. **34**(6): p. 1963-1974.

168. Garcia, A.E. and Sanbonmatsu, K.Y., *Exploring the energy landscape of a beta hairpin in explicit solvent.* Proteins, 2001. **42**(3): p. 345-354.

169. Ghosh, T., Garde, S., and Garcia, A.E., *Role of backbone hydration and salt-bridge formation in stability of alpha-helix in solution.* Biophysical Journal, 2003. **85**(5): p. 3187-3193.

170. Gianni, S., Guydosh, N.R., Khan, F., Caldas, T.D., Mayor, U., White, G.W., DeMarco, M.L., Daggett, V., and Fersht, A.R., *Unifying features in protein-folding mechanisms.* Proceedings of the National Academy of Science of the United States of America, 2003. **100**(23): p. 13286-13291.

171. Daggett, V. and Fersht, A.R., *Is there a unifying mechanism for protein folding?* Trends in Biochemical Science, 2003. **28**(1): p. 18-25.

172. Jayachandran, G., Vishal, V., Garcia, A.E., and Pande, V.S., *Local structure formation in simulations of two small proteins.* Journal of Structural Biology, 2007. **157**(3): p. 491-499.

173. Glasscock, J.M., Zhu, Y., Chowdhury, P., Tang, J., and Gai, F., *Using an amino acid fluorescence resonance energy transfer pair to probe protein unfolding: application to the villin headpiece subdomain and the LysM domain.* Biochemistry, 2008. **47**(42): p. 11070-11076.

174. Lei, H., Wu, C., Liu, H., and Duan, Y., *Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations.* Proceedings of the National Academy of Science of the United States of America, 2007. **104**(12): p. 4925-4930.

175. Yang, J.S., Wallin, S., and Shakhnovich, E.I., *Universality and diversity of folding mechanics for three-helix bundle proteins.* Proceedings of the National Academy of Science of the United States of America, 2008. **105**(3): p. 895-900.

176. Ensign, D.L., Kasson, P.M., and Pande, V.S., *Heterogeneity even at the speed limit of folding: large-scale molecular dynamics study of a fast-folding variant of the villin headpiece.* Journal of Molecular Biology, 2007. **374**(3): p. 806-816.

177. Meng, W., Shan, B., Tang, Y., and Raleigh, D.P., *Native Like Structure in the Unfolded State of the Villin Headpiece Helical Subdomain, an Ultrafast Folding Protein.* Proteins- Structure, Function and Bioinformatics, In Press.

178. Bruschweiler, R. and Case, D.A., *Adding Harmonic Motion to the Karplus Relation for Spin-Spin Coupling.* Journal of the American Chemical Society, 1994. **116**(24): p. 11199-11200.

179. Osapay, K. and Case, D.A., *Analysis of Proton Chemical-Shifts in Regular Secondary Structure of Proteins.* Journal of Biomolecular NMR, 1994. **4**(2): p. 215-230.

180. Xu, X.P. and Case, D.A., *Automated prediction of N-15, C-13(alpha), C-13(beta) and C-13 ' chemical shifts in proteins using a density functional database.* Journal of Biomolecular NMR, 2001. **21**(4): p. 321-333.

181. Wishart, D.S., Bigam, C.G., Holm, A., Hodges, R.S., and Sykes, B.D., *H-1, C-13 and N-15 random coil NMR chemical-shifts of the common amino-acids .1. Investigations of nearest-neighbor effects.* Journal of Biomolecular NMR, 1995. **5**(1): p. 67-81.

182. Choy, W.Y. and Forman-Kay, J.D., *Calculation of ensembles of structures representing the unfolded state of an SH3 domain.* Journal of Molecular Biology, 2001. **308**(5): p. 1011-1032.

183. Marsh, J.A., Neale, C., Jack, F.E., Choy, W.Y., Lee, A.Y., Crowhurst, K.A., and Forman-Kay, J.D., *Improved structural characterizations of the drkN SH3 domain unfolded state suggest a compact ensemble with native-like and non-native structure.* Journal of Molecular Biology, 2007. **367**(5): p. 1494-1510.

184. Roitberg, A. and Simmerling, C., *Special issue: Conformational sampling.* Journal of Molecular Graphics and Modelling, 2004. **22**(5): p. 317.

185. Geyer, C.J. and Thompson, E.A., *Annealing markov-chain monte-carlo with applications to ancestral inference.* Journal of the American Statistical Association, 1995. **90**(431): p. 909-920.

186. Hukushima, K. and Nemoto, K., *Exchange monte carlo method and application to spin glass simulations.* Journal of the Physical Society of Japan, 1996. **65**(6): p. 1604-1608.

187. Swendsen, R.H. and Wang, J.S., *Replica monte carlo simulation of spin glasses.* Physical Review Letters, 1986. **57**(21): p. 2607-2609.

188. Tesi, M.C., van Rensburg, E.J.J., Orlandini, E., and Whittington, S.G., *Monte carlo study of the interacting self-avoiding walk model in three dimensions.* Journal of Statistical Physics, 1996. **82**(1-2): p. 155-181.

189. Lyman, E., Ytreberg, F.M., and Zuckerman, D.M., *Resolution exchange simulation.* Physical Review Letters, 2006. **96**(2): p. 028105.

190. Frantz, D.D., Freeman, D.L., and Doll, J.D., *Reducing quasi-ergodic behavior in monte-carlo simulations by j-walking - Applications to atomic clusters.* Journal of Chemical Physics, 1990. **93**(4): p. 2769-2784.

191. Chen, K., Liu, Z., and Kallenbach, N.R., *The polyproline II conformation in short alanine peptides is noncooperative.* Proceedings of the National Academy of Science of the United States of America, 2004. **101**(43): p. 15352-15357.

192. McColl, I.H., Blanch, E.W., Hecht, L., Kallenbach, N.R., and Barron, L.D., *Vibrational raman optical activity characterization of poly(l-proline) II helix in alanine oligopeptides.* Journal of the American Chemical Society, 2004. **126**(16): p. 5076-5077.

193. Schweitzer-Stenner, R. and Measey, T.J., *The alanine-rich XAO peptide adopts a heterogeneous population, including turn-like and polyproline II conformations.* Proceedings of the National Academy of Science of the United States of America, 2007. **104**(16): p. 6649-6654.

194. Roitberg, A.E., Okur, A., and Simmerling, C., *Coupling of replica exchange simulations to a non-boltzmann structure reservoir.* Journal of Physical Chemistry B, 2007. **111**(10): p. 2415-2418.

195. Fawzi, N.L., Phillips, A.H., Ruscio, J.Z., Doucleff, M., Wemmer, D.E., and Head-Gordon, T., *Structure and dynamics of the AB(21-30) peptide from the interplay of NMR experiments and molecular simulations.* Journal of the American Chemical Society, 2008. **130**(19): p. 6145-6158.

196. Showalter, S.A., Johnson, E., Rance, M., and Bruschweiler, R., *Toward quantitative interpretation of methyl side-chain dynamics from NMR by molecular dynamics simulations.* Journal of the American Chemical Society, 2007. **129**(46): p. 14146-14147.

197. Shi, Z.S., Chen, K., Liu, Z.G., and Kallenbach, N.R., *Conformation of the backbone in unfolded proteins.* Chemical Reviews, 2006. **106**(5): p. 1877-1897.

198. Zagrovic, B., Lipfert, J., Sorin, E.J., Millettt, I.S., van Gunsteren, W.F., Doniach, S., and Pande, V.S., *Unusual compactness of a polyproline type II structure.* Proceedings of the National Academy of Science of the United States of America, 2005. **102**(33): p. 11698-11703.

199. Makowska, J., Rodziewicz-Motowidlo, S., Baginska, K., Vila, J.A., Liwo, A., Chmurzynski, L., and Scheraga, H.A., *Polyproline II conformation is one of many local conformational states and is not an overall conformation of unfolded peptides and proteins.* Proceedings of the National Academy of Science of the United States of America, 2006. **103**(6): p. 1744-1749.

200. Best, R.B., Buchete, N.V., and Hummer, G., *Correction.* Biophysical Journal, 2008. **95**(9): p. 4494.

201. Case, D.A., Scheurer, C., and Bruschweiler, R., *Static and dynamic effects on vicinal scalar J couplings in proteins and peptides: A MD/DFT analysis.* Journal of the American Chemical Society, 2000. **122**(42): p. 10390-10397.

202. Wong, V. and Case, D.A., *Evaluating rotational diffusion from protein MD simulations.* Journal of Physical Chemistry B, 2008. **112**(19): p. 6013-6024.

203. Guillot, B., *A reappraisal of what we have learnt during three decades of computer simulations on water.* Journal of Molecular Liquids, 2002. **101**(1-3): p. 219-260.

204. Dick, T.J. and Madura, J.D., *A Review of the TIP4P, TIP4P-Ew, TIP5P, TIP5P-E Water Models*, in *Annual Reports in Computational Chemistry*, D.C. Spellmeyer, Editor. 2005, Elsevier: Amsterdam. p. 59-74.

205. Jorgensen, W.L. and Tirado-Rives, J., *Potential energy functions for atomic-level simulations of water and organic and biomolecular systems.* Proceedings of the National Academy of Science of the United States of America, 2005. **102**(19): p.

6665-6670.

206. Paesani, F., Zhang, W., Case, D.A., Cheatham, T.E., and Voth, G.A., *An accurate and simple quantum model for liquid water.* Journal of Chemical Physics, 2006. **125**(18): p. 184507

207. Rick, S.W., Stuart, S.J., and Berne, B.J., *Dynamical fluctuating charge force fields: Application to liquid water.* Journal of Chemical Physics, 1994. **101**(7): p. 6141-6156.

208. Chen, B., Xing, J., and Siepmann, J.I., *Development of Polarizable Water Force Fields for Phase Equilibrium Calculations.* Journal of Physical Chemistry B, 2000. **104**(10): p. 2391-2401.

209. Saint-Martin, H., Hernandez-Cobos, J., Bernal-Uruchurtu, M.I., Ortega-Blake, I., and Berendsen, H.J.C., *A mobile charge densities in harmonic oscillators (MCDHO) molecular model for numerical simulations: The water-water interaction.* Journal of Chemical Physics, 2000. **113**(24): p. 10899-10912.

210. Iuchi, S., Morita, A., and Kato, S., *Molecular dynamics simulation with the charge response kernel: Vibrational spectra of liquid water and N-methylacetamide in aqueous solution.* Journal of Physical Chemistry B, 2002. **106**(13): p. 3466-3476.

211. Burnham, C.J. and Xantheas, S.S., *Development of transferable interaction models for water. I. Prominent features of the water dimer potential energy surface.* Journal of Chemical Physics, 2002. **116**(4): p. 1479-1492.

212. Yu, H.B., Hansson, T., and van Gunsteren, W.F., *Development of a simple, self-consistent polarizable model for liquid water.* Journal of Chemical Physics, 2003. **118**(1): p. 221-234.

213. Lamoureux, G., MacKerell, A.D., and Roux, B., *A simple polarizable model of water based on classical Drude oscillators.* Journal of Chemical Physics, 2003. **119**(10): p. 5185-5197.

214. Ren, P.Y. and Ponder, J.W., *Polarizable atomic multipole water model for molecular mechanics simulation.* Journal of Physical Chemistry B, 2003. **107**(24): p. 5933-5947.

215. Donchev, A.G., Ozrin, V.D., Subbotin, M.V., Tarasov, O.V., and Tarasov, V.I., *A quantum mechanical polarizable force field for biomolecular interactions.* Proceedings of the National Academy of Science of the United States of America, 2005. **102**(22): p. 7829-7834.

216. Toukan, K. and Rahman, A., *Molecular-Dynamics Study of Atomic Motions in Water.* Physical Reviews B: Condensed Matter, 1985. **31**(5): p. 2643-2648.

217. Dang, L.X. and Pettitt, B.M., *Simple intramolecular model potentials for water.* Journal of Physical Chemistry, 1987. **91**(12): p. 3349-3354.

218. Lawrence, C.P. and Skinner, J.L., *Flexible TIP4P model for molecular dynamics simulation of liquid water.* Chemical Physics Letters 2003. **372**(5-6): p. 842-847.

219. Mahoney, M.W. and Jorgensen, W.L., *A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions.* Journal of Chemical Physics, 2000. **112**(20): p. 8910-8922.

220. Berendsen, H.J.C., Postma, P.M., van Gunsteren, W.F., and Hermans, J., *Interaction Models for Water in Relation to Protein Hydration*, in *Intermolecular Forces*, B. Pullmann, Editor. 1981, D. Reidel Publishing Company: Reidel, Dordrecht. p. 331-342.

221. Berendsen, H.J.C., Grigera, J.R., and Straatsma, T.P., *The missing term in effective pair potentials.* Journal of Physical Chemistry 1987. **91**(24): p. 6269-6271.

222. Chaban, G.M. and Gerber, R.B., *Anharmonic vibrational spectroscopy of the glycine-water complex: Calculations for ab initio, empirical, and hybrid quantum mechanics/molecular mechanics potentials.* Journal of Chemical Physics, 2001. **115**(3): p. 1340-1348.

223. Gerber, R.B., Brauer, B., Gregurick, S.K., and Chaban, G.M., *Calculation of anharmonic vibrational spectroscopy of small biological molecules.* PhysChemComm, 2002. **5**: p. 142-150.

224. Gerber, R.B., Chaban, G.M., Gregurick, S.K., and Brauer, B., *Vibrational spectroscopy and the development of new force fields for biological molecules.* Biopolymers, 2003. **68**(3): p. 370-382.

225. Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta, S., and Weiner, P., *A New Force-Field for Molecular Mechanical Simulation of Nucleic-Acids and Proteins.* Journal of the American Chemical Society, 1984. **106**(3): p. 765-784.

226. Hermans, J., Berendsen, H.J.C., Vangunsteren, W.F., and Postma, J.P.M., *A Consistent Empirical Potential for Water-Protein Interactions.* Biopolymers, 1984. **23**(8): p. 1513-1518.

227. Nutt, D.R. and Smith, J.C., *Molecular dynamics simulations of proteins: Can the explicit water model be varied?* Journal of Chemical Theory and Computation, 2007. **3**(4): p. 1550-1560.

228. Bader, J.S. and Chandler, D., *Computer-Simulation Study of the Mean Forces between Ferrous and Ferric Ions in Water.* Journal of Physical Chemistry 1992. **96**(15): p. 6423-6427.

229. Steinbach, P.J. and Brooks, B.R., *New Spherical-Cutoff Methods for Long-Range Forces in Macromolecular Simulation.* Journal of Computational Chemistry, 1994. **15**(7): p. 667-683.

230. Lisal, M., Kolafa, J., and Nezbeda, I., *An examination of the five-site potential (TIP5P) for water.* Journal of Chemical Physics, 2002. **117**(19): p. 8892-8897.

231. Rick, S.W., *A reoptimization of the five-site water potential (TIP5P) for use with Ewald sums.* Journal of Chemical Physics, 2004. **120**(13): p. 6085-6093.

232. Price, D.J. and Brooks, C.L., *A modified TIP3P water potential for simulation with Ewald summation.* Journal of Chemical Physics, 2004. **121**(20): p. 10096-10103.

233. Mahoney, M.W. and Jorgensen, W.L., *Diffusion constant of the TIP5P model of liquid water.* Journal of Chemical Physics, 2001. **114**(1): p. 363-366.

234. Mobley, D.L., Dumont, E., Chodera, J.D., and Dill, K.A., *Comparison of charge models for fixed-charge force fields: small-molecule hydration free energies in explicit solvent.* Journal of Physical Chemistry B, 2007. **111**(9): p. 2242-2254.

235. Woutersen, S. and Hamm, P., *Structure determination of trialanine in water using*

polarization sensitive two-dimensional vibrational spectroscopy. Journal of Physical Chemistry B, 2000. **104**(47): p. 11316-11320.

236. Woutersen, S., Pfister, R., Hamm, P., Mu, Y.G., Kosov, D.S., and Stock, G., *Peptide conformational heterogeneity revealed from nonlinear vibrational spectroscopy and molecular-dynamics simulations.* Journal of Chemical Physics, 2002. **117**(14): p. 6833-6840.

237. Eker, F., Cao, X.L., Nafie, L., and Schweitzer-Stenner, R., *Tripeptides adopt stable structures in water. A combined polarized visible Raman, FTIR, and VCD spectroscopy study.* Journal of the American Chemical Society, 2002. **124**(48): p. 14330-14341.

238. Eker, F., Griebenow, K., and Schweitzer-Stenner, R., *Stable conformations of tripeptides in aqueous solution studied by UV circular dichroism spectroscopy.* Journal of the American Chemical Society, 2003. **125**(27): p. 8178-8185.

239. Steinbach, P.J. and Brooks, B.R., *Protein hydration elucidated by molecular dynamics simulation.* Proceedings of the National Academy of Science of the United States of America, 1993. **90**(19): p. 9135-9139.

240. Brunne, R.M., Liepinsh, E., Otting, G., Wuthrich, K., and van Gunsteren, W.F., *Hydration of proteins. A comparison of experimental residence times of water molecules solvating the bovine pancreatic trypsin inhibitor with theoretical model calculations.* Journal of Molecular Biology, 1993. **231**(4): p. 1040-8.

241. Baron, R. and McCammon, J.A., *Dynamics, hydration, and motional averaging of a loop-gated artificial protein cavity: the W191G mutant of cytochrome c peroxidase in water as revealed by molecular dynamics simulations.* Biochemistry, 2007. **46**(37): p. 10629-10642.

242. Dolenc, J., Baron, R., Missimer, J.H., Steinmetz, M.O., and van Gunsteren, W.F., *Exploring the conserved water site and hydration of a coiled-coil trimerisation motif: A MD simulation study.* ChemBioChem, 2008. **9**(11): p. 1749-1756.

243. Duke, R.E. and Pedersen, L.G., *PMEMD 3.* 2003, University of North Carolina-Chapel Hill.

244. Sauter, C., Otalora, F., Gavira, J.A., Vidal, O., Giege, R., and Garcia-Ruiz, J.M., *Structure of tetragonal hen egg-white lysozyme at 0.94 angstrom from crystals*

grown by the counter-diffusion method. Acta Crystallographica Section D: Biological Crystallography, 2001. **57**: p. 1119-1126.

245. Prompers, J.J. and Bruschweiler, R., *General framework for studying the dynamics of folded and nonfolded proteins by NMR relaxation spectroscopy and MD simulation.* Journal of the American Chemical Society, 2002. **124**(16): p. 4522-4534.

246. Trbovic, N., Cho, J., Abel, R., Friesner, R., Rance, M., and Palmer, A., *Protein Side-Chain Dynamics and Residual Conformational Entropy.* Journal of the American Chemical Society, 2009. **131**(2): p. 615-622.

247. Thomas, A.S. and Elcock, A.H., *Molecular simulations suggest protein salt bridges are uniquely suited to life at high temperatures.* Journal of the American Chemical Society, 2004. **126**(7): p. 2208-2214.

248. Perutz, M.F., *Electrostatic Effects in Proteins.* Science., 1978. **201**(4362): p. 1187-1191.

249. Elcock, A.H., *The stability of salt bridges at high temperatures: implications for hyperthermophilic proteins.* Journal of Molecular Biology, 1998. **284**(2): p. 489-502.

250. Xiao, L. and Honig, B., *Electrostatic contributions to the stability of hyperthermophilic proteins.* Journal of Molecular Biology, 1999. **289**(5): p. 1435-1444.

251. Vieille, C. and Zeikus, G.J., *Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability.* Microbiology and Molecular Biology Reviews, 2001. **65**(1): p. 1-43.

252. Zhou, H.X., *Toward the physical basis of thermophilic proteins: linking of enriched polar interactions and reduced heat capacity of unfolding.* Biophysical Journal, 2002. **83**(6): p. 3126-3133.

253. Dominy, B.N., Minoux, H., and Brooks, C.L., 3rd, *An electrostatic basis for the stability of thermophilic proteins.* Proteins, 2004. **57**(1): p. 128-141.

254. Soares, T.A., Daura, X., Oostenbrink, C., Smith, L.J., and van Gunsteren, W.F.,

Validation of the GROMOS force-field parameter set 45Alpha3 against nuclear magnetic resonance data of hen egg lysozyme. Journal of Biomolecular NMR, 2004. **30**(4): p. 407-422.

255. Stocker, U. and van Gunsteren, W.F., *Molecular dynamics simulation of hen egg white lysozyme: a test of the GROMOS96 force field against nuclear magnetic resonance data.* Proteins, 2000. **40**(1): p. 145-153.

256. Buck, M., Boyd, J., Redfield, C., MacKenzie, D.A., Jeenes, D.J., Archer, D.B., and Dobson, C.M., *Structural determinants of protein dynamics: analysis of 15N NMR relaxation measurements for main-chain and side-chain nuclei of hen egg white lysozyme.* Biochemistry, 1995. **34**(12): p. 4041-4055.

257. Buck, M. and Karplus, M., *Internal and Overall Peptide Group Motion in Proteins: Molecular Dynamics Simulations for Lysozyme Compared with Results from X-ray and NMR Spectroscopy.* Journal of the American Chemical Society, 1999. **121**(41): p. 9645-9658.

258. Buck, M., Bouguet-Bonnet, S., Pastor, R.W., and MacKerell, A.D., Jr., *Importance of the CMAP correction to the CHARMM22 protein force field: dynamics of hen lysozyme.* Biophysical Journal, 2006. **90**(4): p. L36-38.

Appendix 1 – Explicit Solvent Equilibration

This procedure was the standard setup/equilibration for systems in explicit solvent in this thesis. This is different than the standard setup/equilibration implemented in lab currently.

1) Structures were built with leap using either TIP3P or TIP4P-Ew.
2) The structures were equilibrated at 300 K for 50 ps with harmonic restraints (10 kcal/mol*Å) on the heavy atoms at constant pressure. The reference structure was the starting structure for the restraining procedure.

Sample input file used for Ala$_3$

```
Md1.in
 &cntrl
      imin = 0, ntx = 1, nstlim = 25000,
      ntc = 2, ntf = 1, tol=0.0000001, ntt = 1, dt = 0.001,
      ntb = 2, ntp = 1, tautp = 0.5, taup = 0.05,
      ntwx = 1000, ntwe = 0, ntwr = 1000, ntpr = 500,
      scee = 1.2, cut = 8.0,
      ntr=1, tempi = 300.0, temp0 = 300.0,
      nscm = 1000, iwrap = 1, restraintmask=":1-3", restraint_wt = 10.0,
 &end
 &ewald
 &end
```

3) The next step was minimization using the steepest descent method. This was done for 500 steps. Harmonic restraints remained on the heavy atoms using the same restraint from the equilibration (10 kcal/mol*Å). The reference structure was the last structure from the equilibration.

Sample input file for the first step of minimization

```
min1.in
 &cntrl
      imin = 1, ntx = 1, maxcyc = 1000, ntmin = 2,
      ntc = 1, ntf = 1,
      ntb = 1, ntp = 0, nsnb = 20,
      ntwx = 500, ntwe = 0, ntpr = 50,
      scee = 1.2, cut = 8.0,
      ntr = 1, restraintmask=":1-3", restraint_wt = 10.0,
 &end
 &ewald
 &end
```

Four other minimization steps were run gradually reducing the restraints on the structure. In the input file min2.in, the restraint weight was lowered to 5 kcal/mol*Å. In the input file min3.in, the restraint weight was lowered to 2 kcal/mol*Å. In the input file, min4.in the restraint weight was lowered to 1 kcal/mol*Å. The last input file, min5.in, had no restraints on the structure. During each minimization step, the reference structure was the last restart structure generated from the previous simulation.

4) The minimization was followed by three short MD simulations (5 ps each) in order to equilibrate the structure at a particular temperature and pressure. The first MD simulation (md2.in) was performed with 5 kcal/mol*Å harmonic restraints on the solute at 300 K under constant pressure. For all three MD steps, the reference structure was the last restart structure from the minimization.

Sample of md2.in

```
antibody
 &cntrl
      imin = 0, ntx = 1, nstlim = 5000,
      ntc = 2, ntf = 1, tol=0.0000001, ntt = 1, dt = 0.001,
      ntb = 2, ntp = 1, tautp = 0.5, taup = 0.05,
      ntwx = 1000, ntwe = 0, ntwr = 1000, ntpr = 50,
      scee = 1.2, cut = 8.0,
      ntr=1, tempi = 300.0, temp0 = 300.0,
      nscm = 1000, iwrap = 1, restraintmask=":1-3", restraint_wt = 5.0,
 &end
 &ewald
 &end
```

During the 3rd MD step, the restraints were lowered to 1.0 kcal/mol*Å. During the last MD step, the restraints were turned off.

Appendix 2 – Explicit Solvent REMD Setup

1) A structure should be setup and equilibrated according to the Simmerling lab equilibration procedure under **constant pressure**. The first replica tutorial should be reviewed before proceeding with this current tutorial.

2) After structure equilibration, one must use the tslop3 program to calculate the number of replicas.

3) A few things must be considered before selecting a temperature range.
a) A 20 % exchange ratio has been shown to be optimum in most cases. This is obtained by using 0.1 (half of 0.2) in the tslop3 program.
b) The replicas should span the temperatures where there is experimental data.
c) 300 K should be included as one of the temperatures.
d) The highest temperature should be around 400 K. The folding rate begins to decrease at a particular temperature so you may not benefit from using extremely high temperatures. We have also shown that sampling with a reservoir generated at 400 K is sufficient in the R-REMD approach.
e) An even number of replicas must be selected.
f) The number of replicas should be adjusted according to the computer system (ie. on a bluegene partition) and the goals of the project. You do not want to use a partition with 1024 processors and only use 600 of those processors. You should try to optimize the replica number to fit that partition (ie. 64 replicas using 16 processors/replica).

4) The groupfile should be setup with scripts from the implicit solvent REMD tutorial. The input files can be generated with the script below that is similar to the one found in the implicit solvent tutorial. There are differences in some of the parameters in the rem.in file that should be noted.

```
#!/bin/csh

    set i=0

    while ($i < 32)
       @ j = $i + 1
       set ext=`printf "%3.3i" $i`
       echo $ext
       set temp=`head -$j temperature.dat | tail -1`
       echo $temp
    cp md.r rem.r.$ext
    cat > rem.in.$ext <<EOF
    Sample explicit solvent input file
    md.in - trpzip2 solvateoct
     &cntrl
          imin = 0, ntx = 5, nstlim = 500,
          ntc = 2, ntf = 1, tol=0.0000001, ntt = 1, dt = 0.002,
          ntb = 1, ntp = 0, irest = 1,
          ntwx = 500, ntwe = 0, ntwr = 500, ntpr = 500,
          scee = 1.2, cut = 6.0,
          ntr = 0, temp0 = $temp, tempi = 0.0,
          nscm = 500, iwrap = 1,
          nsnb = 20,
          tautp = 0.1, offset = 0.09,
          numexchg = 40000, repcrd = 0,
          irest = 1, ntave = 0,
     &end
     &ewald
     &end
     EOF
      @ i++
    End
```

(a) Explicit solvent REMD is run under constant volume conditions (ntb = 1) in AMBER. There are other implementations with constant pressure [1].

(b) This is set to run with a Berendsen thermostat (ntt = 1). In REMD, the temperature is equilibrated at the current temperature after the exchange. This must be done with a stronger coupling constant than normal MD (tautp = 0.1). This should be changed if you plan on performing exchanges more frequently during REMD.

(c) The other parameters that vary between implicit and explicit solvent REMD input files are highlighted in bold.

5) The most important step is benchmarking your simulations so you can determine the number of exchanges for each run. This will allow for you to estimate how long a simulation will take and eliminate the possibility of a run not completing all the exchanges specified in the input files. It is best to underestimate the amount of exchanges according to your benchmark. For example, a 30 min benchmark produces 22 exchanges. If you run a job for 96 hours, that should produce 4224 exchanges. The best choice would be to set your job for 4000 exchanges since computer speed could vary during each run. If the job does not finish, it is possible that all the exchanges will not be processed. This will result in restarts at different temperatures and timesteps. This should be avoided!!!

6) A sample run script is provided in the Simmerling lab equilibration procedures which can be adjusted for the particular simulation.

References

1) Paschek, D., Hempel, S., and Garcia, A.E., *Computing the stability diagram of the Trp-cage miniprotein.* Proceedings from the National Academy of Science of the United States of America, 2008. **105**(46): p. 17754-17759.

Appendix 3 – Clustering

Introduction

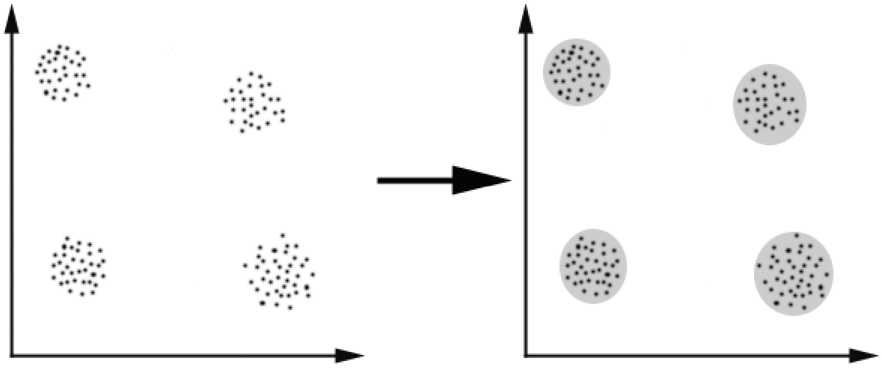
Why cluster?

Clustering is a technique used in many fields to dissect data into discrete sets based on similarity.

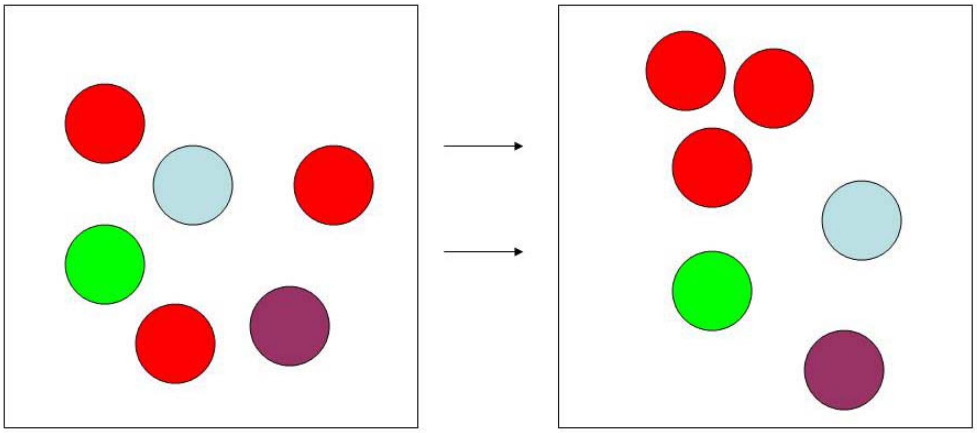
There are 2 main types of clustering:

a) distance based
b) conceptual clustering

Distance based clustering relies on judging similarity by a given geometric distance as seen in example 1.



Conceptual clustering focuses on grouping objects that have a descriptive concept in common. Example 2 illustrates how color can be used as the concept for clustering of the circles.



The type of clustering performed in this lab is distance similarity clustering. We use this tool to observe populations of different structures in a particular trajectory of data. Root Mean Squared Deviation (RMSD) is used as the similarity cutoff in this procedure. A region of the protein is selected based on what one is interested in, whether it be the whole backbone of a protein or specific sidechain conformations. The algorithm used in these calculations is referred to as the average linkage method. More information about this method and other methods can be found on this website:
http://www.elet.polimi.it/upload/matteucc/Clustering/tutorial_html/index.html

How to perform cluster analysis?
You need two files to run this program:

1. inp
2. runanalysis

How to setup the inp file?

It is extremely important that the format be the same as below. Any missed lines will cause the program to malfunction.

```
vilfrag1.top              (Line 1)
6                         (Line 2)
pick #bac 1 13 done       (Line 3)
vilfrag1.combo.x          (Line 4)
1                         (Line 5)
64000                     (Line 6)
20                        (Line 7)
2.5                       (Line 8)
y                         (Line 9)
y                         (Line 10)
n                         (Line 11)
y                         (Line 12)
```

```
indexfile               (Line 13)
99   (last line)        (Line 14)
```

1. Definition of Lines

Line 1 – Topology file
Line 2 – Picks the cluster trajectory program
Line 3 – The region of your molecule which you want to cluster on. This command must be in a certain format. Please refer to the moilview manual which addresses the pick syntax.
http://morita.chem.sunysb.edu/~carlos/viewpage/manual.v10.html
Line 4 – The name of your trajectory file. If you want to look at multiple trajectories, you need to combine them before you run this program.
Line 5 – Start frame number
Line 6 – End frame
Line 7 – Stride – This is the number of frames that are skipped. There can only be 5000 frames run on the 1st pass. The program will crash if there are more frames. In my example, only 3200 frames are read in on the first pass.
Line 8 – This is the for similarity cutoff.
Line 9 – This allows for a 2nd pass to be performed. (keep y)
Line 10 – Write clusters to trajectory (either y or n). This allows you to have the trajectories of similar structures so you can visualize or analyze them.
Line 11 – Asking for you to try another cutoff (keep n)
Line 12 – Saves the cluster indexfile (keep y)
Line 13 – Filename for saving the cluster indexfile
Line 14 – Exits (99)

How to setup up the runanalysis file?

1) Use one line in a file

```
/mnt/raidb/lwicky/bin/ANALYSIS.V3.1.cluster/source/analysis  < inp
```

This allows the program analysis to use inp data to run the program.

2)chmod u+x runanalysis

This makes the program executable.

3)./runanalysis – runs the cluster analysis program

What kind of outputs do we get from running cluster analysis?

1) clustertraj files
2) indexfile
3) phi.flatwell
4) chiral flatwell
5) fort.7 file

****** If you run another cluster analysis in same directory, you need to remove all the files into another directory. The program will not run if those files are still there.

****** Personally, I like to rename fort.7 because if you use a postprocessing program it is likely that this file will be overwritten.

What data is important besides the clustertraj files?

(A) Fort.7 – This file is an output for your whole cluster analysis. If there are any errors, you will be able to find it in this file. The end of file has a summary of the populations for each cluster after the 2nd pass of clustering.

```
Cluster #   1  best structure  46301) has   1028 members
Cluster #   2  best structure  61861) has   8997 members
Cluster #   3  best structure  42061) has   2040 members
```

From this batch of clusters, cluster 2 is the most populated. The best representative structure for the cluster is the frame number in the original trajectory with the lowest RMSD to that cluster.

(B) Indexfile - Tell you the frame number of the original trajectory and the cluster it is located in.

Example (Frames 20-30)

```
  20      2
  21      2
  22      2
  23      2
  24      2
  25      2
  26     15
  27     15
  28      3
  29     15
  30      3
```

Frame 20 is located in cluster 2 while frame 29 is located in cluster 15.

How to perform population analysis using results from cluster analysis?

This analysis enables you to compare the populations of structures from different runs. This is useful when comparing two independent runs from different initial starting conditions (structures, solvent models, or sampling techniques…etc).

First you must combine both trajectories. In order to look at populations, each trajectory must be weighted the same.

Ex.

```
trajin ../../ANAL/lauren.traj.0wat 12001 48000 2 (REMD)
trajin ../../ANAL/rem.x.006.0wat 12001 48000 2 (REMD)
trajin ../../ANAL/rem.1.x.006.strip 1 18000 1 (RREMD)
trajin ../../ANAL/rem.2.x.006.strip 1 18000 1 (RREMD)
trajout traj.combined nobox
go
```

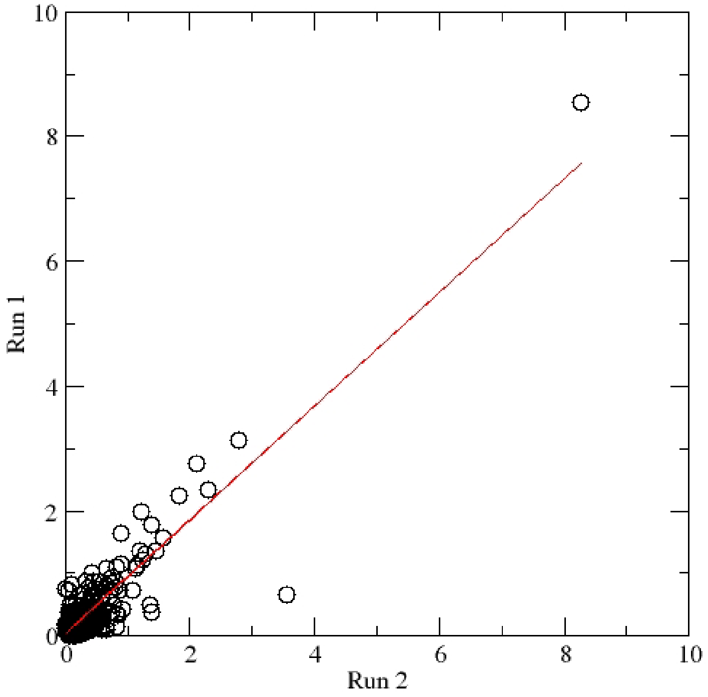
In this example, four trajectories are being combined however they represent two individual sampling methods (REMD and RREMD). Both sampling methods have 36,000 structures each. Using the same amounts of structures for each individual run allows you to perform comparisons for the different individual runs (ie. Compare rem.1 with rem.2).

Second, run cluster analysis. You will obtain the indexfile with a list of structures that belong to each cluster. In order to compare the populations, you must use this script. Using vi make file called pop.sh and paste the following:

```
set i = 0
while ( $i < 346)
awk -v i="$i" '{if($1 <= 36000 && $2 == i) x++; if ($1 > 36001 && $1 <= 72000 &&
$2 == i) y++}END{print x/360, y/360}' indexfile
@ i++
End
```

The first line does not need to be changed. The second line is the number of clusters (if it is 345 – you must use 346) obtained from cluster analysis (check fort.7 file). The third line calculates the populations of structures in each cluster from the first half (column x) and second half (column y) of the trajectory. The print statements prints out the percentage of structures in each cluster (each column should total up to 100).

Then you plot!



Appendix 4 – Reservoir REMD (R-REMD)

Reservoir REMD is implemented by coupling standard REMD to a pregenerated Boltzmann-populated reservoir (Figure 1). Previous implementations have generated the converged structural ensemble by running MD around 400 K. After these structures are generated, standard REMD is subsequently run below this temperature, providing an annealing ladder to optimize reservoir structures and re-weight the high-temperature ensemble. The difference between REMD and R-REMD is that the simulations start with the correct exchange criterion due to the Boltzmann weighting of the reservoir and there is no reliance on folding events within the replicas themselves. For more details, refer to reference 82 in Chapter 5.

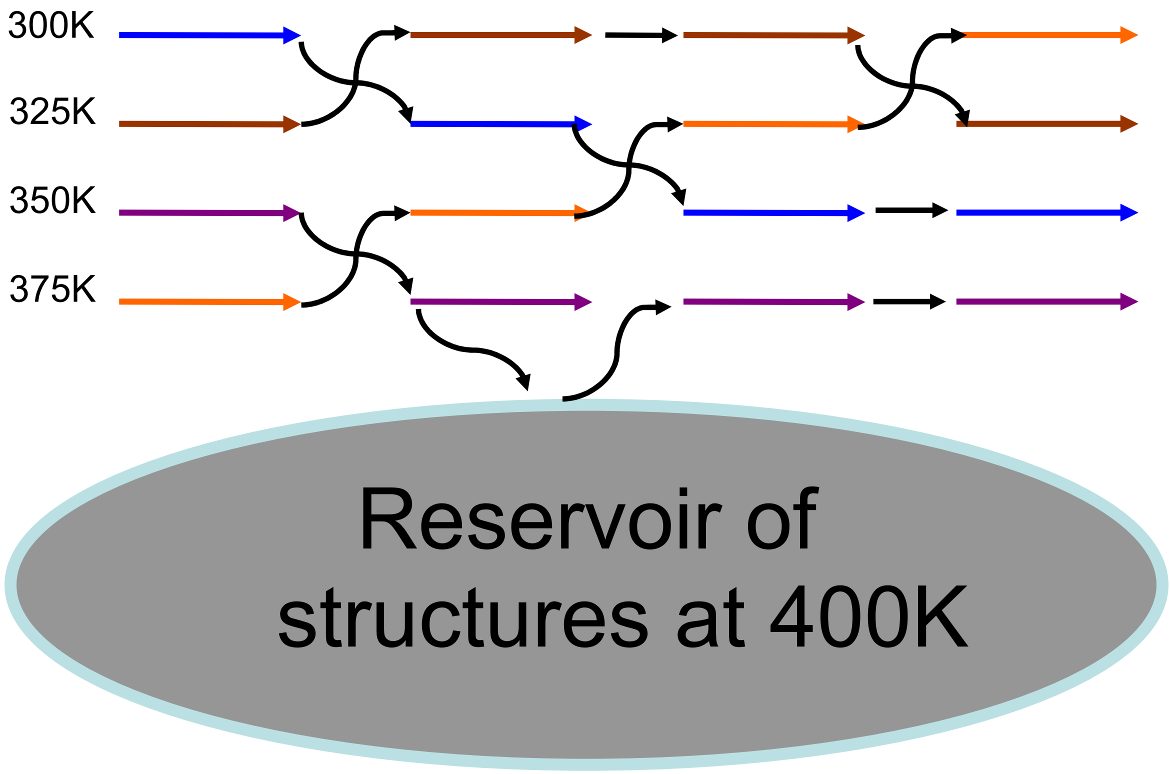


Figure 1 – Schematic diagram of the implementation of R-REMD. Structures are allowed to j-walk between the reservoir and the highest temperature replica. Structures from the highest temperature replica are not put into reservoir after the exchanges (pseudoexchange). The structures remain the same in the reservoir during R-REMD.

Appendix 5 – J-Coupling Tutorial

J-coupling constants are a measure of the coupling effect of nuclear spins due to the bonding electrons in the magnetic field. There are various J-coupling constants that can be measured for small peptide systems [1]. The scalar constant between 3 bonds (^3J) can be calculated using the Karplus equation. This equation relates the dihedral angle between the protons (θ) to the coupling constant.

This procedure will discuss the calculation of $^3J(H_N,H_\alpha)$ scalar coupling constants. This coupling constant probes local secondary structure. This is calculated using the backbone φ angle and the Karplus equation.

$$^3J(H_N,H_\alpha) = A\cos^2(\theta) + B\cos(\theta) + C$$

Equation 1: Karplus equation for the calculation of $^3J(H_N,H_\alpha)$ scalar couplings. A, B, and C are constants. θ is dihedral angle. Typically, θ is equal to φ-60.

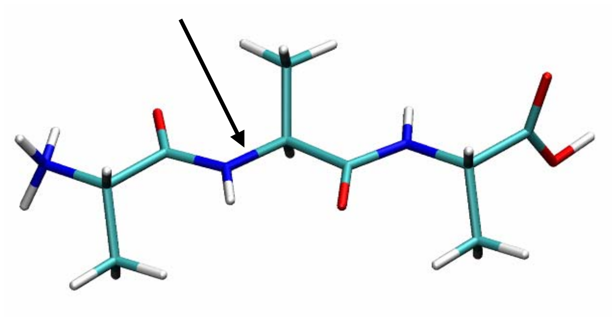


Figure 1: Backbone φ angle in Ala$_3$.

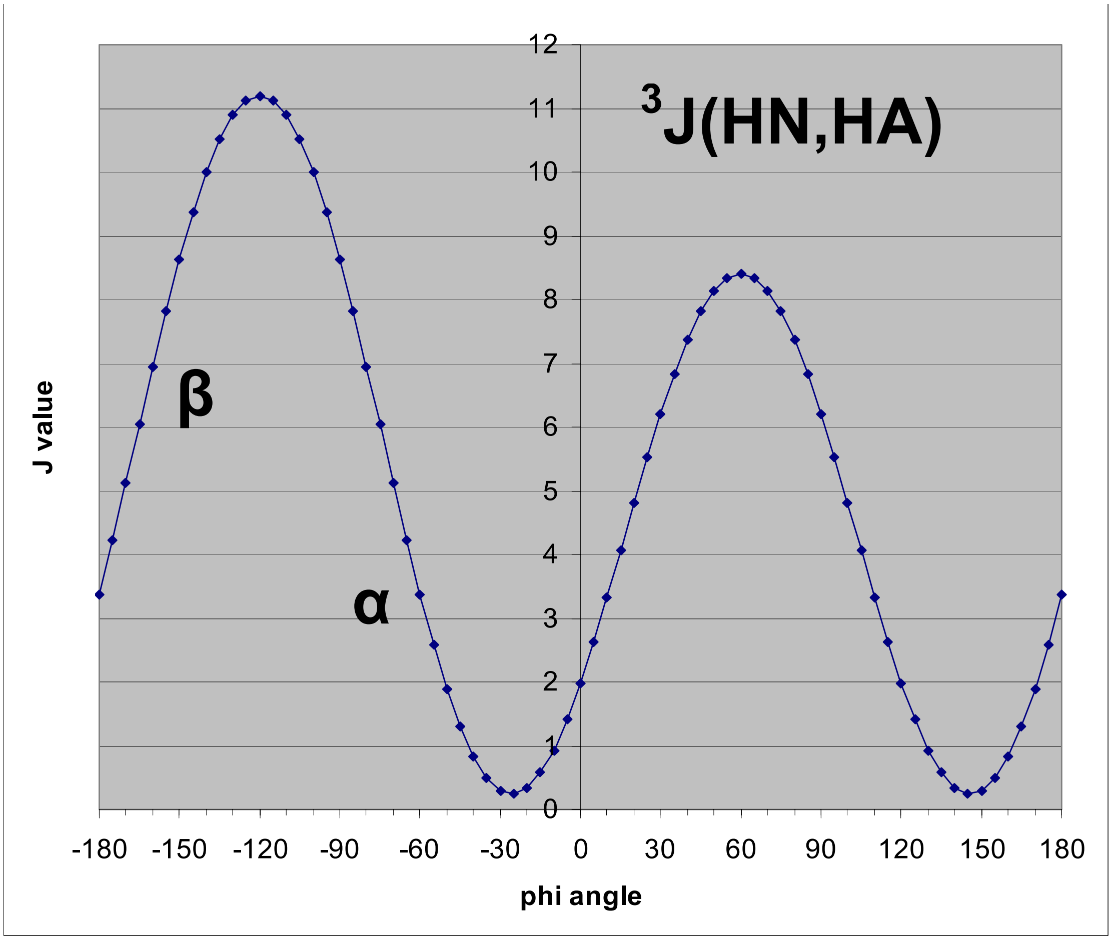


Figure 2: Example of a Karplus curve.

$^3J(H_{N,}H_\alpha)$ scalar coupling constants located around 3-5 Hz correspond to local α-helical structures (Figure 2). Coupling constants higher than 7 Hz correspond to local β-structure. PP_{II} conformations can have similar scalar coupling constants compare to α-helical structures (typically around 5.0 Hz).

1) The first step in this procedure is to calculate the phi angle for each residue in your system.
This can be done using this script (located in /mnt/raidc/lwicky/tutorial/jcoupling)

```
#!/bin/tcsh

rm dihed.in


set i=1

while ($i < 20)
   @ j = $i + 1

echo "dihedral phi$i :$i@C :$j@N :$j@CA :$j@C out $j.PHI" >> dihed.in

@ i++
end

echo "trajin traj.combo.phecore.traj" > tmp.1
cat tmp.1 dihed.in > ptraj.in

ptraj vilfrag1+2.top ptraj.in
```

Things to consider

a) In this script, the loop is echoing a command for the dihedral calculation. You must look at your system and count the number of phi dihedrals. This number will determine the value you use in the loop. In this example, 20 dihedrals can be calculated so it will run the loop until i = 21.

b) You need to change the trajectory and topology in this script. This is specific for HP21. This example can be found in: /mnt/raidc/lwicky/tutorial/jcoupling.

2) The next step is to use the Karplus equation to calculate the scalar couplings. You must obtain the A, B, C, and θ values relevant to your system. The script used to calculate the constants is **JCoupling.bax.pl** (located in mnt/raidc/lwicky/tutorial/jcoupling). In this example, the parameters are A = 6.51, B = -1.76, C = 1.6 and $\theta = \varphi - 60$. Other parameters for different J-coupling constants can be found in Best et al [2]. If you open up the script, these values can be changed in these lines of the script.

```
  $theta =  (abs($cur - 60))*3.14159/180;
  $costh = cos($theta);
  $jc = 6.51*$costh*$costh - 1.76*$costh +1.6;
```

3) Next step is to calculate the scalar couplings for every residue and calculate the average value for the trajectory. The average is calculated with a script called stat.pl located in the same directory. The **jcoupling.bax.sh** script is setup to calculate the J-couplings using **JCoupling.bax.pl**. This is followed by the average calculation. All of the J-couplings are put into “jcoupling.dat”.

```
#!/bin/csh

rm jcoupling.dat

set i=2

while ($i < 22)

./JCoupling.bax.pl $i.PHI 2 > jcoup.$i
./stat.pl jcoup.$i 1  | grep "Mean value:" | awk '{print$3}' > avg.jcoup.bax.$i
cat avg.jcoup.bax.$i >> jcoupling.dat

echo $i

@ i++
end
```

References

1) Graf, J., Nguyen, P.H., Stock, G., and Schwalbe, H., *Structure and Dynamics of the Homologous Series of Alanine Peptides: A Joint Molecular Dynamics/NMR Study.* Journal of the American Chemical Society, 2007. **129**(5): p. 1179-1189.

2) Best, R.B., Buchete, N.V., and Hummer, G., *Are current molecular dynamics force fields too helical?* Biophysical Journal, 2008. **95**(1): p. L7-L9.

Appendix 6 – Water Density Calculations

This tutorial was contributed by Trent Balius
Using Ptraj to calculate water densities:
See script at the bottom for easy to run file. Make changes where specified.

Step I: center reference

```
trajin restart_file.crd #restart file
center :1-129 origin mass #change mask
image origin center familiar
trajout new_restart_file.crd restart
go
```

Step II: Reorient and center the molecule over the entire trajectory to the reference structure.

```
trajin trajectory.crd

center :1-129 origin mass #change mask
image origin center familiar
reference new_restart_file.crd.1 #file generated above
#rms
#orient all
rms reference out output_file.rms @N,C,CA #change mask
trajout new_trajectory.crd
go
```

One may want to calculate the average structure here to compare to the density. See attached script.

Step III: calculate densities using grid.

```
trajin new_trajectory.crd
grid grid_output.xplor 100 0.5 100 0.5 100 0.5 :WAT@O origin max 0.3
go
```

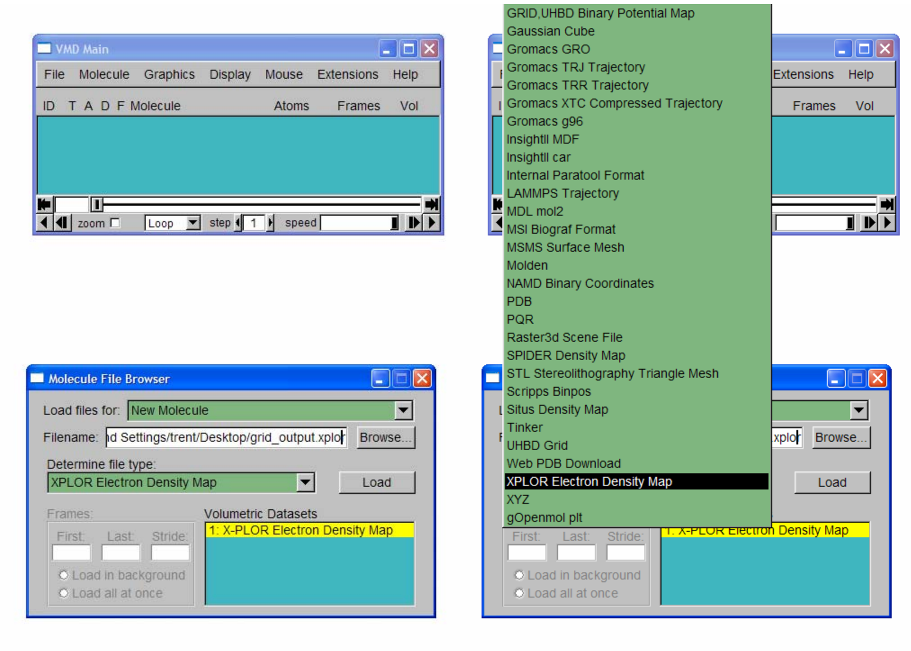
The grid command calculates a three dimensional histogram. In this case, the command is calculated with in a 50 by 50 by 50 $Å^3$ box as specified. The protein and the box are centered about the same point. The grid spacing is specified to be 0.5x0.5x0.5 $Å^3$.

Isovalues are obtained as output from the grid function: This is located in the file ‘grid_output.xplor’. The isovalue is incremented every time an oxygen of a water molecule is found in the grid space.

Using VMD to visualize water densities:

Step I: load density histogram:

Go to ‘>file>new molecule. Select file ‘`grid_output.xplor`’. import this file as a xplor



Load files for: New Molecule

Filename: d Settings/trent/Desktop/grid_output.xplor

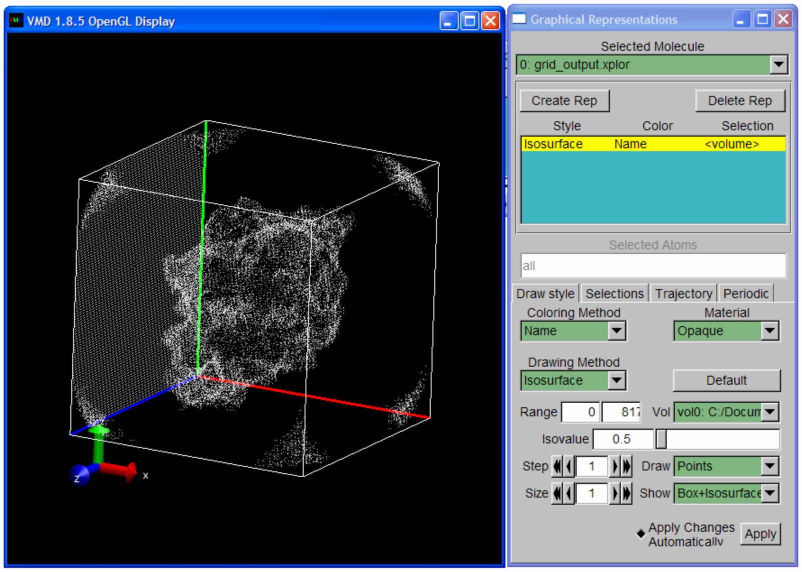
Determine file type: XPLOR Electron Density Map

Volumetric Datasets: 1: X-PLOR Electron Density Map

Step II: adjust isovalues: The isovalues are the number of water that pass through a given grid box during the entire simulation.

In VMD, we can specify the threshold to what level of density we want to display. If the density threshold is low then the entire space that contained water will be shown. As you increase the threshold only progressively higher density regions will be shown.

1)Go to ‘>>Graphics>>representations’.



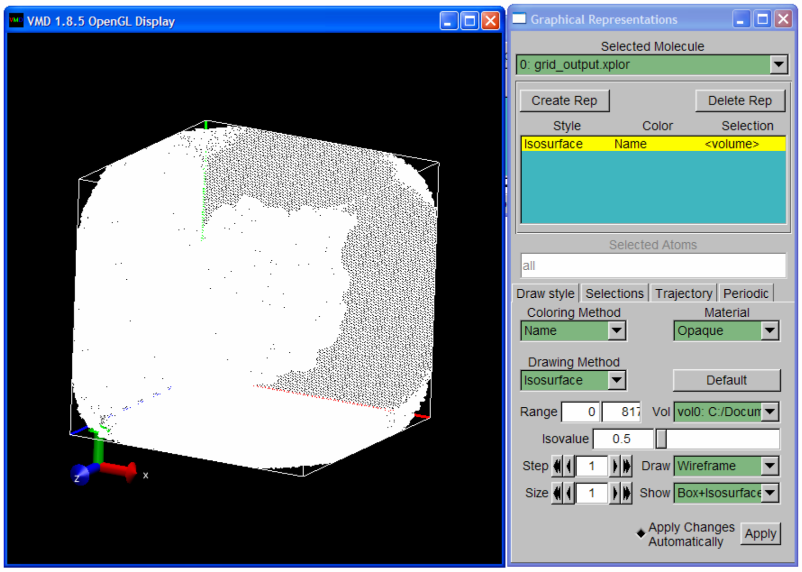
Selected Molecule: 0: grid_output.xplor

Style	Color	Selection
Isosurface	Name	<volume>

Selected Atoms: all

Coloring Method: Name
Material: Opaque
Drawing Method: Isosurface
Range: 0 817
Vol: vol0: C:/Docum
Isovalue: 0.5
Step: 1
Draw: Points
Size: 1
Show: Box+Isosurface
Apply Changes Automatically

2) Change draw from points to wireframe



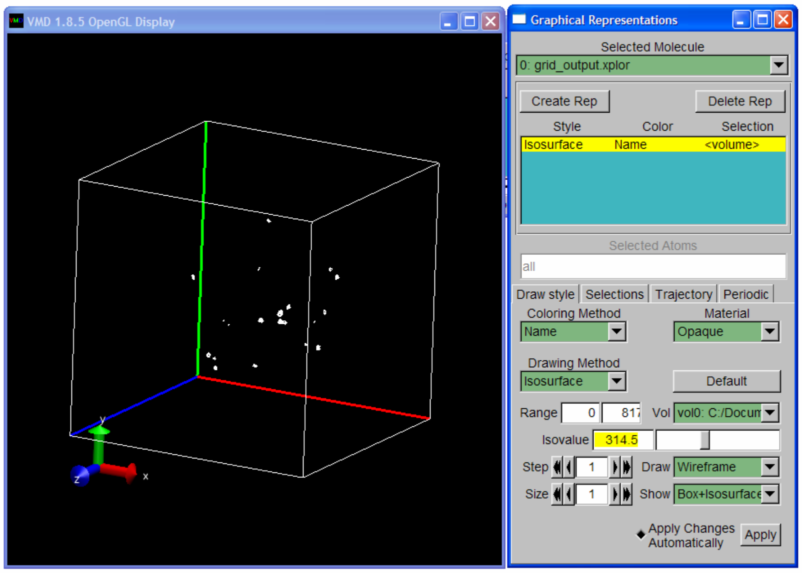
Selected Molecule: 0: grid_output.xplor

Style	Color	Selection
Isosurface	Name	<volume>

Selected Atoms: all

Coloring Method: Name
Material: Opaque
Drawing Method: Isosurface
Range: 0 817
Vol: vol0: C:/Docum
Isovalue: 0.5
Step: 1
Draw: Wireframe
Size: 1
Show: Box+Isosurface
Apply Changes Automatically

3) Change isovalue to higher thresholds



Selected Molecule: 0: grid_output.xplor

Style	Color	Selection
Isosurface	Name	<volume>

Selected Atoms: all

Coloring Method: Name
Material: Opaque
Drawing Method: Isosurface
Range: 0 817
Vol: vol0: C:/Docum
Isovalue: 314.5
Step: 1
Draw: Wireframe
Size: 1
Show: Box+Isosurface
◆ Apply Changes Automatically

Step III: Import molecule (reference or average).

If you wish to compare your densities with a crystal structure or another structure that has not be properly centered and oriented. You should align that structure to one that has been properly centered and oriented such as the reference structure.

1) Import the reference structure (SI)
2) Import the non-centered structure (SII)
3) Align SII to SI by using >>Extensions>>Analysis>>RMSD calculator.
    a. Change residues
    b. Make sure that you are aligning SII to SI and not the other way around. If you do align them the wrong way round delete SI and SII reload them and try again.

Script file:

```
#!/bin/tcsh

set ptraj = "/mnt/raidb/programs/amber9.ifort9/exe/ptraj"
# you may need to change path
# specify amber ptraj location.
set parm = "../sol.8ogc.parm7"
# change
# specify parameter/topology file name (input)
set rst = "../md3_8og.rst"
# change
# specify reference file name (input)
set traj = "../md3_8og.x"
# change
# specify trajectory file name (input)
```

```
#goto water

$ptraj $parm <<EOF

trajin $rst

center :1-32 mass origin
image origin center familiar

trajout reference.rst7 restart # change. Also change other file name
                               # below at (*)
go
EOF

water:

$ptraj $parm <<EOF

trajin  $traj 100 500 #change. Number specify what part of trajectory
                      #       to input and perform analysis on.

center :1-32 mass origin # change mask
image origin center familiar

reference reference.rst7.1 #change. Same as above at (*)
rms reference mass :1-32

trajout water.x

grid grid_wat.xplor 100 0.5 100 0.5 100 0.5 :WAT@O
grid grid_wath.xplor 100 0.5 100 0.5 100 0.5 :WAT@H?
#best look at O's density

translate x -0.25 y -0.25 z -0.25
average avg.pdb pdb # outputs average structure as pdb.

EOF
```