Stony Brook University

Study of Biologically Relevant Phenomena Using Small Peptide Models

A Dissertation Presented

by

Daniel Robinson Roe

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Chemistry

Stony Brook University

August 2007

Stony Brook University

The Graduate School

Daniel Robinson Roe

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Carlos L. Simmerling – Dissertation Advisor
Associated Professor, Department of Chemistry

Daniel P. Raleigh – Chairperson of Defense
Professor, Department of Chemistry

David M. Hanson – Third Member
Professor, Department of Chemistry

Alexey Onufriev – Outside Member
Associated Professor, Departments of Computer Science and Physics, Virginia Tech

This dissertation is accepted by the Graduate School

Lawrence Martin
Dean of the Graduate School

Abstract of the Dissertation

Study of Biologically Relevant Phenomena Using Small Peptide Models

by

Daniel Robinson Roe

Doctor of Philosophy

in

Chemistry

Stony Brook University

2007

Understanding protein structure and dynamics is a central and important problem in structural biology. Small model peptides are useful for studying this problem as they reduce the complexity involved in studying the folding of larger proteins while providing important insights into the formation of protein secondary structure. The small size of model peptides makes them particularly amenable to study by molecular dynamics (MD) simulations, which can provide atomic-level detail of peptide dynamics. When this data is used in conjunction with that from experiment it can be used to explain certain experimental results, or make predictions that can then be tested by experiment. Agreement with experimental results is an important benchmark for the validation of simulation results.

One major problem in comparing MD simulations with experiment is that of convergence – the timescales available to MD simulations are typically orders of magnitude shorter than experimental timescales. Enhanced sampling techniques such as Replica Exchange Molecular Dynamics (REMD) can be used to improve convergence of simulations. Further improvements in sampling can be achieved through the use of implicit solvent models, which increase sampling by reducing solvent friction and improving the sampling of solvent configurations. However, it is extremely important to gauge the accuracy of implicit solvent models due to their approximate nature.

The work presented in this thesis is concerned with the study of various aspects of the protein folding problem via MD simulation methods, as well as the validation of such methods against experiment. The folding and unfolding kinetics of a model ß-hairpin are studied in detail via MD simulations, and compared to thermodynamic data obtained from REMD simulations. Folding is found to be kinetically partitioned into a fast and a slow phase, with the fast phase corresponding to a direct transition from the unfolded state to the folded state, and the slow phase corresponding to a transition from misfolded to unfolded structures.

The cooperativity of individual hairpin formation in a model 3-stranded ß-sheet was also studied via REMD simulations, and it was found that the actual cooperativity

was significantly larger than the previously estimated lower limit. Mutations were then made that affected the stability of the component hairpins of this ß-sheet, and a Ser to Val mutation was found that significantly stabilized the overall sheet from elimination of a non-native hydrogen bond that led to destabilization of the native structure.

Finally, the accuracy of several generalized Born (GB) implicit solvent models was studied via REMD simulations of a small polyalanine peptide. The GB solvent models were found to give incorrect secondary structure populations compared to simulations with explicit solvent and experimental results. This discrepancy in secondary structure was found to be related to incorrect estimation of the solvation free energy gap between conformations of the polyalanine peptide by these GB models. However, an implicit solvent model based on the Poisson Equation (PE) was found to give better results. Attempts to improve the accuracy of the GB models by fitting to PE were not successful, indicating there may be limitations to the improvement of current GB models.

Dedicated to the memory of my father, Gelston Grady Roe,
who taught me the importance of both knowing and understanding,
gave me courage, and showed me the true meaning of perseverance.

Table of Contents

List of Figures

List of Tables

List of Equations

Chapter 1
Introduction

1.1 Structural Biology and the Study of Proteins

Biology, simply defined, is the study of life. Although this is the direct meaning of the word, this belies the underlying complexity that such a study entails. It can be a simple thing to look at another organism such as an animal, plant, or an insect, and say that it is alive – but what are the underlying mechanisms and processes that sustain life? A simple analogy for an organism is an automobile. It is a simple thing to look at an automobile and say if it's running or not, but knowing what makes it run requires opening the hood and examining the various components that make the vehicle run: the engine, brakes, transmission, and so on. Similarly, an understanding of what makes an organism 'run' requires examining its components as well: the various organs and tissues, and the multitude of cells and biomolecules that they are composed of. Structural biology is concerned with the study of life at this level of detail; examining the numerous interactions between these components that compose a living organism in order to better understand the overall function of the organism.

One of these components that are essential to life as we know it is a class of biomolecules called proteins. The word protein is derived from the Greek *proteios*, meaning 'of first importance', and indeed proteins play a central role in many important biological processes. They function as enzymes, as structural components, and as means of transporting other molecules. They are also important in cellular signaling mechanisms, immune response, cell division, and a large variety of other processes essential to life[1].

Much as the study of the parts of a car engine allows one to comprehend how it runs, the study of proteins allow a more complete understanding of how the various processes in which they are involved in work in detail. Most proteins have a well-defined three-dimensional structure which largely determines the properties of the protein. However, unlike the various parts of a car engine, proteins are not static entities; they are dynamic molecules, and changes in this structure can and do occur. Proteins can be unfolded (denatured) and refolded, and in some cases even folded to an incorrect structure (misfolded). Both the structure and folding behavior of a protein can also be influenced by the surrounding environment: solvent, presence of ions, pH, and so on.

Understanding the structure of a protein and how it folds and unfolds can help explain the details of how that protein functions, which may then be used to make certain predictions about the behavior of that protein (such as, for example, what conditions may cause it to cease functioning properly). The work presented in this thesis is concerned with understanding the underlying forces which influence protein folding and stability.

1.1.1 Protein Structure and Stability

Proteins are composed of a linear sequence of molecular units called amino acids. The basic structure of an amino acid can be divided into two components. The first

component, the backbone, is common to every amino acid. It is composed of an amide and a carbonyl group flanking a central carbon, called the a-carbon. In a protein, amino acids are typically linked to each other through their amide and/or carbonyl groups. The second component, the sidechain, is what uniquely identifies an amino acid. The sidechain is connected to the a-carbon and determines the properties of the amino acid. There are 20 standard amino acids that compose the majority of all proteins.

Proteins contain three basic levels of structure: primary, secondary, and tertiary. The primary structure of a protein is simply its amino acid sequence. Secondary structure refers to common structural motifs that certain sequences of amino acids in a protein can adopt. Examples of this are the a-helix, the ß-hairpin, the ß-sheet, and reverse-turns. A typical protein is composed of numerous combinations of regions of secondary structure. Tertiary structure refers to how the various units of secondary structure relate to each other and form the overall fold of the protein. Examples of primary, secondary, and tertiary structure are given in Figure 1-1.

There are a variety of factors that influence the stability of proteins. The types of secondary structure that a given amino acid sequence can adopt are influenced by steric constraints imposed by the individual side-chains. Once formed, secondary structure is largely stabilized by hydrogen bonds that occur between the amide and carbonyl groups of the amino-acid backbone. Tertiary structure can be stabilized by ionic or hydrophobic interactions between side-chains. In fact, the hydrophobic effect (characterized by the burying of non-polar side-chains when in a polar solvent, *i.e.* water) is proposed to be one of the main forces that stabilize proteins[2].

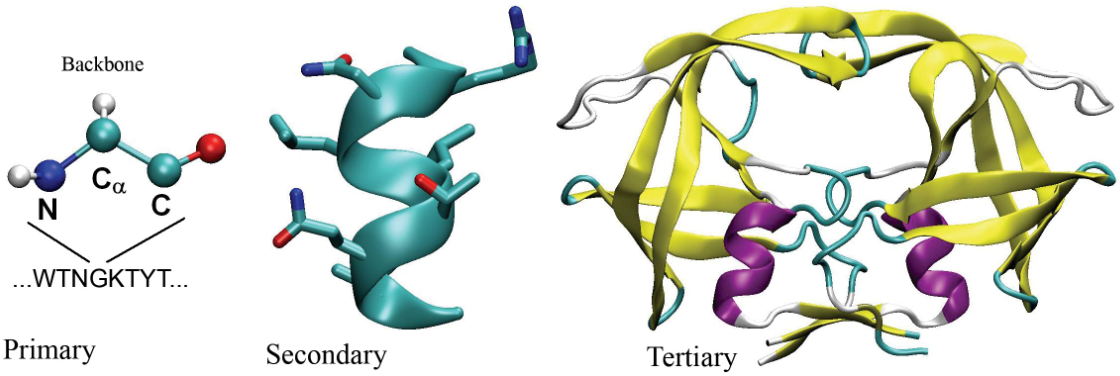

Figure 1-1. Primary, secondary, and tertiary structure in proteins. An amino acid sequence is shown as an example of primary structure, along with the backbone of an amino acid in a ball-and-stick representation with the side-chain omitted for clarity. Amino acids are linked through their N- and C- termini and can form regular secondary structure. An example of a-helical structure is shown here; the backbone is shown in Cartoon representation, the side-chains are shown in Licorice representation with hydrogen atoms omitted for clarity. The structure of HIV-1 Protease (PDB ID 1HVR) is shown as an example of tertiary structure. The entire protein is shown in Cartoon representation, colored by secondary structure type. Picture generated with VMD 1.8.4[3].

In vivo, proteins typically fold into a more or less well-defined three-dimensional structure. In 1950 Linus Pauling and Robert Corey proposed the structure of an a-helix based on possible hydrogen bonding patterns of amino acids[4], and later proposed the ß-sheet in 1951[5]. This was followed by the first full three-dimensional structure determination of the protein myoglobin, by John Kendrew *et al.*[6]. In 1973 Anfinsen

reported that bovine pancreatic ribonuclease, although non-functional when denatured with urea, was able to regain its functionality when the denaturing conditions were removed[7]. This indicated that the primary structure of a protein was implicitly related to its secondary and tertiary structure, *i.e.* all the information for the folding of a protein to its native structure is contained in its amino acid sequence.

1.1.2 Protein Folding

Understanding just how a protein folds into its final structure is one of the most important problems in structural biology. One of the main reasons for this is that protein folding is linked to protein function – an incorrect fold can result in malfunction or non-function of the protein. In fact, protein misfolding is implicated in many diseases[8], including Alzheimers and Parkinson's disease[9]. In addition to folding from the denatured state, understanding the dynamics of the folded protein itself can also be important for understanding how proteins interact with one another (such as conformational changes that can accompany the process of signaling[10]), or how certain drugs might interact with a target protein[11].

It was recognized and stated by Levinthal[12] that proteins could not arrive at their native structure from a completely random search, as this would take an astronomical amount of time due to the number of conformational degrees of freedom in a typical protein, and proteins are known to fold on the order of μs to seconds[13]. Levinthal concluded there must be certain pathways that would guide protein folding. This idea was later restated in terms of the free energy landscape of a protein being funnel-shaped[14]. In these terms, the energy gained during folding is represented by the depth of the funnel, and the number of states available at a given energy (*i.e.* entropy) is represented by the width of the funnel. Folding is then 'downhill', with losses in entropy compensated by favorable gains in energy.

A concept important to protein folding is that of cooperativity. A process can be thought of as cooperative if the next step in that process is easier than the previous step. Cooperativity can be seen at the level of tertiary structure in proteins; partially folded structures are usually less stable than the folded or denatured states[15]. Cooperativity is also present at the level of protein secondary structure. For example, a long α-helical segment is more stable than shorter α-helical fragments of total equivalent length[16]. The concept of cooperativity is explored further in Chapter 3.

1.1.3 Experimental Study of Proteins

Over the past century, a variety of tools and methods have become available to research the problems of protein structure and folding (for a brief review, see reference [17]). One method used to obtain protein structures is X-ray crystallography. In this method, crystals containing the protein of interest are grown and exposed to X-rays. The protein's structure can then be determined from the diffraction pattern of the X-rays. This method is capable of obtaining high resolution structures (around ~1 Å in the best cases), and is the most commonly used for protein structure determination (as of early 2007 ~85% of structures in the RCSB Protein Data Bank are from X-ray crystallography[18]).

However, there are several drawbacks to this method. It can be very difficult to obtain a crystal of the desired protein – in some cases mutations to the protein must be made in order to get it to crystallize properly. Also, since the protein must be immobilized within a crystal lattice, there is no information about the dynamics of the protein in solution, although there is some information regarding the flexibility of the crystallized conformation of the protein. In addition there is sometimes the question of whether the conditions under which the protein crystal was obtained has altered the conformation of the protein[19].

Another method commonly used to study proteins is nuclear magnetic resonance (NMR) spectroscopy. In this method, a sample of a protein is placed within a strong magnetic field. The nuclei of certain atoms in the protein can then be made to resonate after an applied pulse of electromagnetic radiation. The nuclei give off a characteristic signal that depends on their surrounding environment, which is called the chemical shift. This signal is also proportional to the number of nuclei present. Interactions between various nuclei from spin-spin coupling can result in further modifications of the signal. The chemical shift and spin-spin coupling result in a spectrum that is more or less unique for a molecule. This is especially useful for proteins, as it can distinguish between different conformations.

One major advantage of NMR over X-ray crystallography is that the protein can be observed in an aqueous environment which is close to what the protein actually experiences *in vivo*. In addition, since the protein is not bound by a crystal lattice and is able to move freely, information on the conformational dynamics of the protein can be obtained. For example, changes in conformation as a function of temperature change can be followed. The main disadvantage of NMR is in sample size; the larger the protein, the more overlap there is from a multitude of signals and the less one is able to assign portions of the NMR spectra to individual groups or atoms.

Circular dichroism (CD) is often used to measure secondary structure content in proteins based on absorption of circularly polarized light. Different secondary structure types, such as a-helices and ß-sheets for example, give rise to different signals, with the amplitude of the signal related to how much of the conformation is present. As in NMR spectroscopy, proteins can be studied in solution with CD, allowing changes in protein conformation to be studied as well. The main limitation of CD is that it can only give an overview of secondary structure, and not specifically where the secondary structure occurs in the protein.

Proteins which contain aromatic residues such as tryptophan can also be studied using UV fluorescence. Briefly, fluorescence is the process by which a photon of light is absorbed by a molecule, causing electronic excitation. Relaxation of the molecule to the ground state is accompanied by the emission of radiation, which can be detected. This technique is sensitive to the environment that a given residue is in; if the residue is solvent exposed, fluorescence will be relatively high, but if that residue becomes buried in the protein the fluorescence will be 'quenched'. The simplest application of this method is in proteins containing only one fluorophore, since there are complex electronic effects when multiple fluorophores are present.

1.2 Molecular Dynamics Simulations

Another method that has the potential to provide information on protein structure and folding that is complementary to that obtainable by experiment is computational simulation. In this theory-based method, a model of the given system and all the interactions contained within is created, and an attempt is made to predict certain properties of the system. One of the most popular computational methods is molecular dynamics (MD) simulation. Since the first MD simulation of 216 molecules of liquid water was performed by Stillinger and Rahman[20], and the first MD simulation of a protein soon after[21], MD simulation has become an important field of research (see Figure 1-2). Recent advances in MD simulation methods and computing power have further enhanced their reliability and utility (for a brief review, see reference [22]).

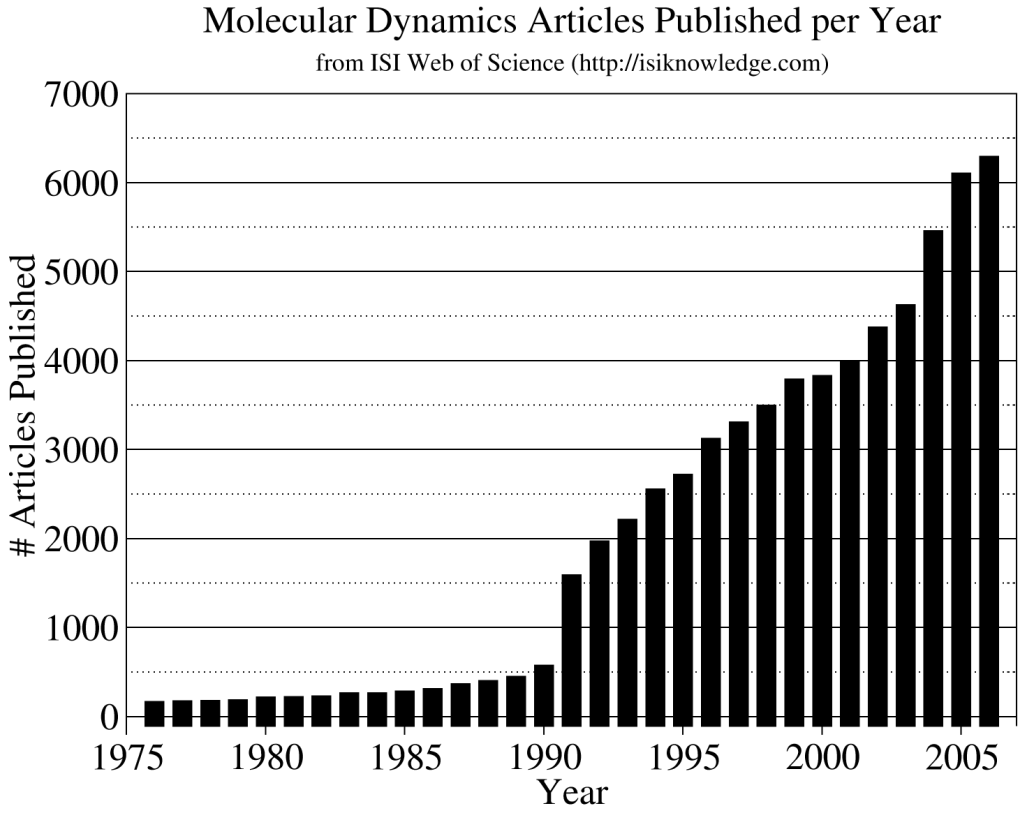

Figure 1-2. Number of articles with the topic "Molecular Dynamics" published per year, based on a search using ISI Web of Knowledge.

1.2.1 Basics of MD Simulation

In the most basic sense, MD simulation is the attempt to predict how a molecule will move over time by calculating the forces on the molecule derived from a specific representation of the molecular energy. Ideally, when running an MD simulation of a given protein the representation of that protein should be as complete and accurate as possible. In the most extreme case, this would mean having a quantum mechanical

representation of all atoms and electrons. Unfortunately this is not possible – due to the complexity of such calculations a complete quantum representation has only ever been given for a hydrogen atom. Larger systems can be solved for by use of some approximations, such as the Born-Oppenheimer approximation[23]. Although recently there have been advances in quantum dynamics simulation[24], quantum calculations themselves are quite time-consuming. Because of this, most MD remains non-quantum; that is to say, uses the classic laws of physics. Atoms are represented as spheres with a certain mass and radius, and their electrons are represented as point charges which are obtained from quantum calculations[25].

In MD simulations, the energy of a given configuration of atoms is described by a force field. The force field contains terms that represent all the various interactions in the system. The precise form of the force field can vary, but for protein simulations force fields will usually contain terms that describe bond bending and stretching, torsion angle rotation, and non-bonded interactions (van der Waals and Coulombic interactions). A simple example of a force field is shown in Equation 1-1. Once the energy of a molecule at a certain point in time is known, the forces on the molecule can be calculated and the future positions of the atoms in the protein can be predicted using Newton's second law of motion. While a full description of the detailed mechanics of MD simulations is outside the scope of this work, further details can be found in reference [23].

$$U(X_N) = \sum_{bonds} \frac{k_i}{2}(l_i - l_{i,0})^2 + \sum_{angles} \frac{k_i}{2}(q_i - q_{i,0})^2 + \sum_{torsions} \frac{V_n}{2}(1 + \cos(nv - g))$$
$$+ \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(4e_{ij} \left[\left(\frac{s_{ij}}{r_{ij}} \right)^{12} - \left(\frac{s_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4pe_0 r_{ij}} \right)$$

Equation 1-1. Example of a force field equation for use in MD simulations. U is the potential energy, X_N represents the coordinates of N atoms, the bond and angle terms are given by a simple harmonic potential, the torsions are represented by a periodic term with a certain number of wells, and the non-bonded interactions are represented by a Lennard-Jones 6-12 potential and a Coulombic term.

1.2.2 MD Simulations in Structural Biology

MD simulations can be very useful in not only obtaining data complementary to experiment, but also in obtaining data that can be difficult to obtain from experiments. One of the potential uses of MD simulations is the prediction of protein structure from an amino acid sequence. Theoretical prediction of protein structure is especially attractive in the case where the structure of a protein can be difficult to obtain by experimental methods (such as membrane proteins for example[26]). Anfinsen's discovery that the fold of a protein was linked to its amino acid sequence[7] implies that the three dimensional structure of a protein can be predicted. This has recently been demonstrated by Simmerling *et al.*, who in 2002 predicted the correct structure of the 20 residue Trp-cage mini-protein[27], which was later confirmed by experiment. However, all-atom structure prediction remains a challenge for larger peptide systems.

One advantage of MD simulation over experimental methods is it allows observation of a single molecule. In typical experiments large numbers of molecules are present, and so only the average behavior of the ensemble is observed (although there

have been advances in single molecule experiments, see [28]). MD simulations also have the advantage that molecules can be followed with atomic-level detail. While the resolution of experimental methods can be quite good, it is often the case that certain portions of the molecule cannot be resolved through experimental methods (for example, highly mobile loop regions in X-ray crystal structures). In fact, MD simulation techniques are used in refinement of structures obtained from experiment [29]. MD can also be used to explain discrepancies in protein structures obtained with different experimental methods [30].

MD can also be used for the fast screening of potential drugs that may bind to a protein target [31]. Processes that can be difficult to observe experimentally, such as the motion of individual ions through ion channels in cell membranes, can be studied in detail via MD simulation [32]. MD simulation also permits direct manipulation of physical parameters in order to gain insight into various phenomena, such as the adjustment of Van der waals interactions between solute and solvent to study the concept of hydrophobicity [33].

1.2.3 Sampling in MD Simulations

One major problem in obtaining thermodynamically relevant information on a system with MD simulations is that of sampling. Arguably the most accurate MD simulations are those in which all atoms are represented. Using such a representation, a system with N atoms would require N^2 calculations to be performed at each step of the simulation. For a relatively small (76 residue) protein like Ubiquitin (PDB ID 1UBQ) that has approximately 600 atoms, this means roughly 360,000 calculations each step. The number of calculations grows rapidly prohibitive as the number of atoms increases, especially if solvent atoms are explicitly represented. Because of this, all-atom simulations of even small proteins in explicit solvent have only currently reached into the μs range [34], which is only around the timescale of folding of small proteins [13]. Two methods commonly employed to increase sampling in MD simulations are the replica exchange method and the use of continuum solvation.

1.2.3.1 Replica Exchange Molecular Dynamics

The problem of adequate sampling in MD simulations can be at least partially addressed through the use of enhanced sampling techniques such as parallel tempering [35] or replica exchange molecular dynamics (REMD) [36]. The work presented in this thesis makes extensive use of REMD to obtain well-converged ensembles of structures, which is required for calculation of thermodynamically relevant information such as free energy. REMD involves N non-interacting MD simulations of the desired system, each at a different temperature. Exchanges between the structures of neighboring replicas are attempted periodically and accepted with probability shown in Equation 1-2. As defined here, REMD allows the structure to make a random walk in temperature space. The main advantage to REMD is that it allows structures that may be kinetically trapped at low temperature to overcome potential energy barriers by exchanging to a higher temperature.

The temperature distribution of the replicas should be chosen so that 1) there is sufficient overlap between the potential energy distributions of neighboring replicas to ensure exchanges occur and 2) the highest temperature allows the structure to overcome any potential energy barriers in a reasonable amount of time. For some discussion on the selection of an optimal temperature distribution for replica exchange, see references [37] and [38].

$$p(E_0, b_0 \rightarrow E_1, b_1) = \min\{1, \exp[-(b_0 - b_1)(E_1 - E_0)]\}$$

Equation 1-2. Probability of accepting an exchange between neighboring replicas. E is potential energy, and ß = $(k_B T)^{-1}$, where kB is Boltzmann's constant and T is temperature.

Although the superior convergence of REMD over standard MD has been demonstrated for small peptides[39], there are still several issues that must be kept in minds when running REMD simulations. One important consideration is that although REMD is an enhanced sampling method, this in and of itself does not guarantee converged data. The default exchange criterion assumes Boltzmann-weighted ensembles, and this is generally not true at the beginning of a REMD simulation. A consequence of this is that until all replicas are converged, none are converged. Because of this, the best guarantee of convergence is still comparison of various order parameters obtained from 2 or more independent simulations; all work presented in this thesis using REMD uses this criteria for convergence.

1.2.3.2 Continuum Solvent Models

Another approach commonly used to address the problem of sampling in MD simulations is the use of continuum solvent models. An accurate description of the solvent surrounding a protein is essential in order to correctly describe its behavior. This can be done in a straightforward manner by explicitly including all solvent atoms as part of the system. However, this greatly increases the number of degrees of freedom available to the system, and obtaining relevant thermodynamic data requires sampling over these degrees of freedom. In a continuum solvent model there is no need to explicitly include solvent atoms – all solvent degrees of freedom are accounted for implicitly. While this can greatly enhance sampling in MD simulations, care must be taken to ensure that solvent effects are still accurately accounted for. This topic is covered in much more detail in Chapter 5.

1.3 Model Systems

Since the protein folding problem is quite complex, it is often desirable to study it on a simpler level. It has long been recognized that proteins contain a large amount of regular secondary structure in the form of a-helices and ß-sheets. Current theories of protein folding consider the formation of secondary structural elements as an important first step in the protein folding process[40-42]. Model peptides can be used to study the properties of these basic units of secondary structure and so provide insight into early stages of protein folding[43]. For example, several ß-hairpin peptide models have been used to explore various aspects of hairpin formation (for a review, see reference [44]).

An ideal model peptide is one that is small, but still exhibits properties of larger proteins such as secondary structure formation, tertiary structure formation, folding cooperativity, and so on. Model peptides generally fall into two categories: peptides that are derived from a larger, naturally occurring protein, and peptides that have been designed. Derived peptides have the advantage that they are 'all natural', and therefore probably are good models for actual folding events. One example is the N-terminal domain of ribosomal protein L9, which has been studied as a model of specific interactions in the denatured state of proteins [45]. Another example is the Trp-cage miniprotein, which is often studied via MD simulation since it is fast to fold and the smallest peptide (only 20 residues) that exhibits both secondary and tertiary structure [46].

Derived model peptides are not always available to study certain problems. In these cases it may be possible to design an appropriate model peptide. It is even possible to make use of unnatural amino acids to facilitate study of a given problem. For example, Schenck & Gellman designed DPDP, a 20-residue model ß-sheet peptide, to study cooperativity in ß-sheet formation. This peptide makes use of two unnatural amino acids (D-Proline) in the turn regions; replacement of these with L-Proline residues effectively 'turns off' hairpin formation. Designed peptides are also a useful benchmark for gauging current knowledge of the underlying forces which drive and stabilize protein folding and structure. For example, the very stable trpzip4 peptide was designed by replacing certain residues of the ß-hairpin fragment of protein G with tryptophan, creating a stronger hydrophobic core [47].

Since large protein systems can be costly to simulate, the small size of model peptide makes them ideal systems for study by MD simulations. For some examples of model peptides used to study protein folding via MD simulations, see reference [43]. The work presented in this thesis focuses on using model peptides to study protein folding, cooperativity, stability, and solvation.

1.4 Outline of Research Projects

The research outlined below is focused on the study of four topics relevant to protein folding and stability via MD simulations of model peptides. Chapter 2 covers the study of protein folding and unfolding pathways using the 12 residue ß-hairpin model Trpzip2. Chapter 3 and Chapter 4 cover the study of both cooperativity in folding and individual hairpin stability in the 20 residue 3-stranded ß-sheet model DPDP. Finally, Chapter 5 covers the study and comparison of various models for aqueous solvation using a 10 residue polyalanine peptide. The work in Chapter 3 and Chapter 5 has been published as references [48] and [49] respectively.

1.4.1 Folding and Unfolding Pathways Characterized in a Model ß-hairpin

Understanding how proteins fold to a well defined structure is a complex problem of great interest. The folding and unfolding behavior of a small ß-hairpin model peptide was studied via converged REMD simulations and MD simulations. Folding and unfolding kinetics were studied using non-equilibrium temperature jump simulations, and

results were validated against free energy data obtained from REMD simulations. The unfolded state is observed to have a high tendency to sample a ß-turn, along with non-specific hydrophobic contacts. Folding involves an increased specificity of these contacts and formation of native backbone hydrogen bonds, with both events occurring at the folding free energy barrier. While unfolding data can be fit to a single exponential implying a 2-state process, folding data requires a double exponential fit, suggesting the presence of kinetic partitioning. Further analysis reveals that folding involves a fast phase which involves direct transition to the native state, and a slow phase involving kinetic trapping in misfolded conformations. These same misfolded conformations are shown to be part of the free energy landscape obtained from REMD simulations. The combined kinetic and thermodynamic data describe a process of folding that is much more complex than the simple 2-state process typically ascribed to small systems.

1.4.2 Measurement of Folding Cooperativity between Two Hairpins of a 3-stranded ß-sheet

Cooperativity in ß-sheet formation is a problem of particular interest since certain diseases related to protein misfolding involve the formation of ß-sheet-like structures. The thermodynamic behavior of a previously designed three-stranded ß-sheet was studied via several µs of REMD simulations. The system is shown to populate at least four thermodynamic minima, including 2 partially folded states in which one hairpin is formed and the other hairpin is absent. Simulated melting curves show different profiles for the C and N-terminal hairpins, consistent with differences in secondary structure content in published NMR and CD/FTIR measurements, which probed different regions of the chain. Individual ß-hairpins that comprise the 3-stranded ß-sheet are observed to form cooperatively. Partial folding cooperativity between the component hairpins is observed, and good agreement between calculated and experimental values quantifying this cooperativity is obtained when similar analysis techniques are used. However, the structural detail in the ensemble of conformations sampled in the simulations permits a more direct analysis of this cooperatively than has been performed based on experimental data. The results indicate the actual folding cooperativity perpendicular to strand direction is significantly larger than the lower bound obtained previously.

1.4.3 Mutations Affecting Individual Hairpin Stability in a 3-stranded ß-sheet

Since certain diseases related to protein misfolding involve the formation of ß-sheet-like structures, it is important to understand the forces which stabilize ß-sheets and enhance sheet formation. In a previous study of a 3-stranded ß-sheet model peptide, the stability of the C-terminal hairpin was found to be significantly greater than the N-terminal hairpin. Based on observations of the hairpins from experiment and computational simulations, several mutants were simulated in an attempt to understand the underlying causes for the difference in hairpin stability. Mutation of a Tyr residue central to a hydrophobic cluster to a Thr was found to be destabilizing, but a Tyr to Val mutation was not, underlying the importance of a hydrophobic core. The addition of a salt-bridge to the N-terminal hairpin was found to have a negligible affect on stability.

Moving a Phe residue to the central strand of DPDP was found to be stabilizing through formation of a second hydrophobic core and stabilization of the first hairpin. A Ser to Val mutation near the turn region of the first hairpin was found to be even more stabilizing and significantly improved the stability of the N-terminal hairpin through elimination of an unfolding intermediate structure. However, a mutant combining both the Phe and Ser to Val mutations was found to have stability that was only in between the two mutants by themselves, indicating these mutations compete somehow. The results suggest that turn optimization is central to improving overall hairpin stability.

1.4.4 Evaluation of Implicit Solvent Model Accuracy via Detailed Free Energy Calculations

Computational simulations routinely make use of implicit solvent models. However, the accuracy of these models compared to explicit solvation is unclear. The effects of the use of three generalized Born (GB) implicit solvent models on the thermodynamics of a simple polyalanine peptide are studied via comparing several hundred ns of well-converged replica exchange molecular dynamics (REMD) simulations using explicit TIP3P solvent to REMD simulations with the GB solvent models. It is found that when compared to REMD simulations using TIP3P, the GB REMD simulations contain significant differences in secondary structure populations; most notably an over-abundance of a-helical secondary structure. This discrepancy is explored via comparison of the differences in the electrostatic component of the free energy of solvation (??GPol) between TIP3P (via Thermodynamic Integration calculations), the GB models, and an implicit solvent model based on the Poisson Equation (PE). The electrostatic component of the solvation free energies are calculated using each solvent model for four representative conformations of Ala10. Since PE is found to have the best performance with respect to reproducing TIP3P ??GPol values, effective Born radii from the GB models are compared to effective Born radii calculated with PE (so-called perfect radii), and significant and numerous deviations in GB radii from perfect radii are found in all GB models. The effect of these deviations on the solvation free energy is discussed, and it is shown that even when perfect radii are used the agreement of GB with TIP3P ??GPol values does not improve. This suggests a limit to the optimization of the effective Born radius calculation, and that future efforts to improve the accuracy of GB must extend beyond such optimizations. It also suggests that simulations with GB models will not be able to produce quantitative results without further optimization, although qualitative results may still be obtained.

Chapter 2
A Study of Folding and Unfolding Pathways of a Model ß-hairpin

2.1 Introduction

An important aspect of the protein folding problem lies in understanding the process by which proteins locate their native conformations from the vast available phase space. Small model peptides can provide insight into the dynamics of basic units of secondary structure such as a-helices and ß-sheets, the formation of which is thought to be important in the overall folding process[43]. The study of ß-hairpin and related ß-sheet structure is particularly interesting as the formation and aggregation of such structure is thought to be characteristic of many diseases in which protein misfolding is implicated[9]. This study focuses on ß-hairpin secondary structure formation, characterization of the native and unfolded ensemble and the changes that occur through the folding transition.

The results of previous computational studies of ß-hairpins have varied, and several folding mechanisms have been proposed. Bonvin *et al.* proposed a mechanism for the first ß-hairpin of tendamistat in which the turn is formed first, followed by hydrogen bond formation starting at the turn and subsequent stabilization by side-chain interactions[50]. The folding mechanism of the ß-hairpin fragment of protein G (GB1) proposed by Muñoz *et al.* also involves hydrogen bond formation initiating at the turn, then becoming stabilized by side-chain interactions[51]. Pande *et al.* proposed a slightly different mechanism for GB1 in which hydrophobic interactions between side-chains drive the strands together, after which the hydrogen bonds form to stabilize the structure[52]. Dinner *et al.* also proposed that folding was driven by hydrophobic collapse, after which hydrogen bonds propagate out from the hydrophobic core[53]. Klimov *et al.* proposed that following hydrophobic collapse, hydrogen bond formation proceeded rapidly from the turn, but this may depend on the position of the hydrophobic residues from the turn[54]. Zhou *et al.* proposed that hydrophobic collapse and hydrogen bond formation happen roughly at the same time[55]. In general, the difference in these mechanisms is the balance between hydrophobic collapse and hydrogen bond formation, and the order in which the hydrogen bonds form.

In contrast to experiments, a major drawback to MD simulations is the difficulty in obtaining well-converged ensemble-averaged data. Direct observation of folding events in unrestrained simulations may not be possible on timescales accessible to the simulation. An alternate approach is to generate thermodynamic properties with enhanced sampling techniques such as REMD, at the cost of losing explicit time-dependent behavior (*i.e.* direct observation of folding events) in the process. Using simulations in which the folding process was not observed to describe folding events usually relies on interpretation of free energy barriers observed in a reduced dimensionality and/or along pre-determined order parameters. These may not accurately reflect the actual barriers or even the minima encountered during folding of individual members of the ensemble.

In this study, the thermodynamics and kinetics of the trpzip2 ß-hairpin model[56] are studied via a combination of well-converged REMD and non-equilibrium T-jump MD simulations respectively. A significant tendency to form the ß-turn is found in the unfolded ensemble, along with the formation of non-specific hydrophobic contacts. Folding involves an increase in contact specificity coincident with formation of native backbone, and unfolding generally reflects the reverse process with differences in the sequence of events. While a single exponential describes the unfolding process, the folding process is described by a double exponential which partitions the folding ensemble into a slow and fast phase. It is shown that the fast phase is comprised of structures which transition directly to the native state from the unfolded state, and the slow phase results from a transition from the unfolded state to a misfolded structure. These misfolded structures are also seen in the converged REMD simuatlions. Finally, it is demonstrated that although unfolding occurs with single exponential kinetics, separate pathways are observed in the unfolding process.

2.2 Methods

2.2.1 Model System and Order Parameters

The model system chosen was the tryptophan zipper (trpzip) developed by Starovasnik *et al.*[56], shown in Figure 2-1. This ß-hairpin structural motif is stabilized through cross-strand tryptophan pairs. Trpzip2 (SWTWENGKWTWK, with a type I' ß-turn at NG) has the most cooperative melting curve and highest stability (~90% at 300K) among the trpzips; therefore, it was selected for use in this study. Thermodynamic properties for this peptide have been determined by NMR and CD spectroscopy, and a family of structures was refined using restraints from NMR experiments[56] (PDB code 1LE1). The N-terminal of the peptide was acetylated and the C-terminal was amidated, in accordance with the experimental system[56].

In addition to RMSD from native structure, two other measures of structure were used in subsequent analysis. HBlost is defined as the number of native backbone hydrogen bonds lost. Native backbone hydrogen bonds were defined based on the PDB structure and were considered between E5O-K8H, K8O-E5H, T3O-T10H, T10O-T3H, and S1O-K12H (shown as orange, brown, gold, blue, and violet lines respectively in Figure 2-1). These hydrogen bonds are also referred to in the results as Top, TopMid, BotMid, Bottom, and Lowest respectively, reflecting their proximity to the reverse turn. A hydrogen bond was considered present if the distance between the amide hydrogen and carbonyl oxygen was less than 2.9 Å.

The number of contacts between Trp side chains was also calculated. All six possible pairs of Trp residues were considered in order to obtain a measure of non-specific hydrophobic clustering; the two native pairs (W2-W11 and W4-W9, shown as black and cyan lines respectively in Figure 2-1) and four non-native pairs. A hydrophobic contact between the Trp residues was considered present if the distance between the center of mass of the Trp side-chains was less than 6.5 Å.

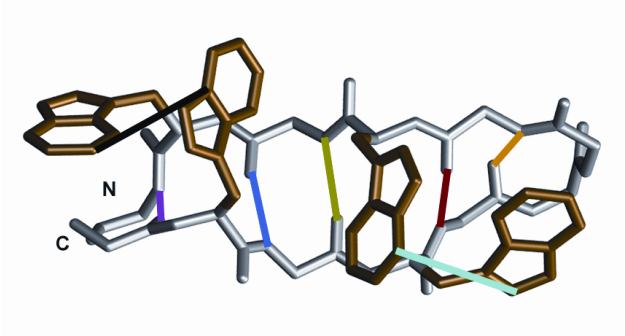

Figure 2-1. NMR-based conformation of trpzip2 (pdb code 1LE1). Side-chains are shown only for Trp residues. Native backbone hydrogen bonds and Trp packing contacts defined in the text are shown as color-coded lines, with the colors matching data curves for these contacts as shown in subsequent figures. The number of native backbone hydrogen bonds that are not present defines the "HBlost" order parameter.

2.2.2 Temperature Jump Simulation Details

To obtain the kinetics of trpzip2 folding/unfolding, simulations of an unfolded ensemble and a folded ensemble were subjected to simulations in which the temperature was instantaneously adjusted to the target temperature of 350 K (referred to hereafter as a T-jump simulation). All simulations were carried out using a locally modified version of Amber 6[57]. The systems were weakly coupled to an external bath[58] with a constant of 1.0 ps to maintain constant temperature. The SHAKE algorithm[59] was used to constrain all bond lengths, allowing a time step of 2.0 fs to be used. All non-bonded interactions were evaluated at each time step with no cut-offs. All simulations used the Generalized Born (GB) implicit solvent model[60] (igb = 1 in Amber), without additional friction terms. Although this lack of viscosity prevents direct comparison of simulated and experimental rate constants, conformational sampling is enhanced[61]. The force field used was ff94[62], with modifications made to reduce over-stabilization of a-helical conformations[63]. The total simulation time for folding and unfolding simulations was 2.1 μs.

Folding was studied using the following procedure. Non-native structures were generated as starting structures for the folding T-jump simulations from MD simulation at 800 K. Forty-nine snapshots with proper stereochemistry and trans peptide bonds were chosen randomly. The backbone RMSD values of these structures to the native structure ranged from 2 to 8 Å. This ensemble of structures was subjected to a temperature jump by instantaneously changing the bath temperature to 350 K. First passage times were calculated as the time at which the instantaneous backbone RMSD for residues 2-11 fell below 0.6Å to ensure that the native basin was reached. After folding, the simulations were terminated. The fraction of structures that had not yet folded was then calculated as a function of time.

Unfolding was studied using an analogous procedure. Initial structures for 53 unfolding trajectories were obtained by assigning different initial velocities corresponding to a distribution at 350 K to the native conformation that had been previously equilibrated at 300 K. First passage times were identified when the RMSD rose above 3.0 Å to ensure complete unfolding of the structure. The fraction of structures that had not yet unfolded was then calculated as a function of time.

It is noted here that these simulations are not strictly temperature jump simulations in the sense that the entire ensemble is not allowed to relax to a new equilibrium position. However, the termination of each simulation after a folding or unfolding event does allow direct calculations of the folding rate without contribution from the unfolding rate and *vice versa* and also eliminates the excessive time it would require each ensemble to relax to equilibrium (up to 5 μs[64]).

2.2.3 Replica Exchange Simulation Details

Converged equilibrium data was obtained with Replica Exchange Molecular Dynamics (REMD) simulations performed with Amber 8[65]. For each REMD simulation, 14 replicas at temperatures ranging from 251.7 K to 554.7 K were used. Extra replicas were added around the experimental melting temperature of 345 K to ensure better sampling for generation of the melting curve. Exchanges between replicas were attempted every ps and coordinates of each replica were saved at every exchange. An exchange acceptance ratio of about 15% was achieved. Two independent REMD simulations were run; one with all replicas starting from the experimental native conformation, and the other with all replicas starting from an unfolded conformation. Each REMD simulation was run for about 85000 exchanges (for a total simulation time of 85 ns per replica). All other details of the REMD simulations were the same as the T-jump MD simulations.

2.2.4 Thermodynamic Analysis

Data from the individual replicas in the REMD simulations were used to generate a melting curve for Trpzip2. Structures were classified as native when RMSD from the experimentally determined structure was under 1.7 Å (based on location of free energy barrier, see Figure 2-2). Fractions of folded and unfolded structures were calculated at each temperature below 373 K, and ?G was calculated assuming a two-state model of folding using Equation 2-1. These data were fit to the Gibbs-Helmholtz equation (Equation 2-2) to obtain values for melting temperature, enthalpy of melting, and heat capacity.

$$\Delta G = -RT \ln\left(\frac{1-F}{F}\right)$$

Equation 2-1. Gibbs free energy (?G) assuming a 2-state system. R and T are the gas constant and temperature respectively, and F is the fraction of folded structures.

$$\Delta G = \left[\Delta H_m\left(1-\frac{T}{T_m}\right)\right] - \Delta C_P\left[\left(T_m - T\right) + T\ln\left(\frac{T}{T_m}\right)\right]$$

Equation 2-2. Gibbs-Helmholtz equation. ?G is the free energy, $?H_m$ is the enthalpy of melting, T and T_m are the temperature and melting temperature respectively, and $?C_P$ is the heat capacity at constant pressure.

Free energies as a function of order parameters were obtained from histograms of converged REMD data; free energy values shown are relative to the most populated

histogram bin. Lower limits for uncertainties in thermodynamic values, melting curves, and contact fractions are reported as half the difference between the values obtained from each independent REMD simulation.

2.3 Results

2.3.1 Trpzip2 Thermodynamics: REMD simulations

Before carrying out a detailed analysis on any system, it is important to validate the model by ensuring that the simulations reproduce the experimentally determined structure and stability. Figure 2-2 shows free energy of trpzip2 at 350 K as a function of backbone RMSD from the native structure. The free energy minimum at RMSD = 0.8 corresponds to the native structure. The fraction of native structures based on an RMSD cutoff of 1.7 Å (location of the free energy barrier from native) is 0.40 ± 0.05, which seems reasonable given that this temperature is just 5 K over the experimentally determined melting temperature. The error bars, which are measured as half of the difference between values obtained with the separate REMD simulations, show that good convergence has been achieved; the largest errors are less than 0.5 kcal mol^{-1}.

There are two other well-defined free energy minima in Figure 2-2 located at RMSD = 2.4 and 3.2 Å. Visual examination of structures extracted from these minima based on RMSD cutoffs show they represent misfolded conformations of trpzip2, termed invertedTrp, wrongTrp, and GKturn, shown in . The minimum at RMSD = 2.4 Å contains a mixture of invertedTrp and wrongTrp structures. Based on an RMSD cutoff of 1.7 Å the fraction of invertedTrp and wrongTrp structures are 0.023 ± 0.004 and 0.031 ± 0.004 at 350 K respectively. In the wrongTrp structure one of the strands has inverted so that Trp residues that would normally stack are on opposite sides of trpzip2. Good Trp packing is no longer possible due to the strand inversion. In the invertedTrp structure both strands have inverted so that all of the Trp residues are on the opposite side of trpzip2 relative to where they are in the native structure. Although Trp packing is still possible in this case, the inversion of the strands puts strain on the turn region, as evidenced by the kinked turn in this structure. In both cases, the turn itself is no longer type II'. Backbone hydrogen bonding is still possible, although the hydrogen bonding pattern obviously differs from native due to the strand inversions.

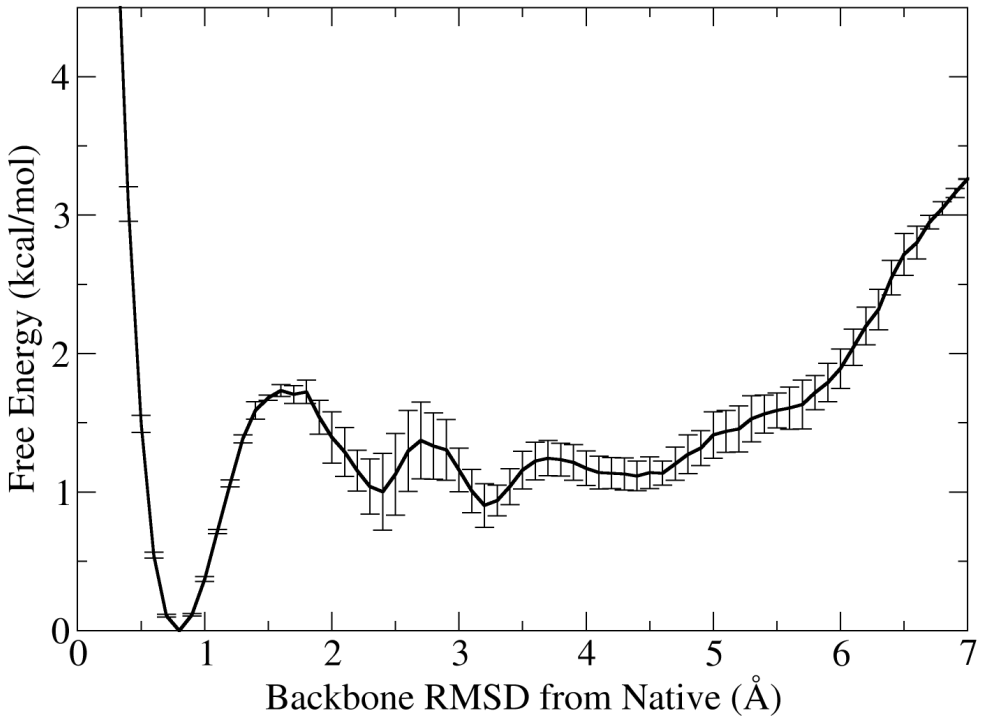

Figure 2-2. Free energy for Trpzip2 as a function of backbone RMSD from the native structure, calculated from two independent REMD simulations. The minimum located at RMSD=0.8 corresponds to native structure. The other minima located at RMSD = 2.4 and 3.2 correspond to misfolded structures. Error bars reflect the difference between the two REMD simulations.

The minimum at RMSD = 3.2 Å consists of the GKturn structure, in which the reverse turn has shifted towards the C-terminus. In addition, the strands are inverted as in the invertedTrp structure. Based on an RMSD cutoff of 1.7 Å the fraction of GKturn structures at 350 K is 0.049 ± 0.005. Although there are less potential backbone hydrogen bonds in GKturn, Trp2 and Trp9 are positioned close enough to still pack. It should be noted that these three structures do not represent an exhaustive search of all possible misfolded structures (such as from cluster analysis); however they likely represent the majority of well-defined misfolded structures due to their location in free energy minima.

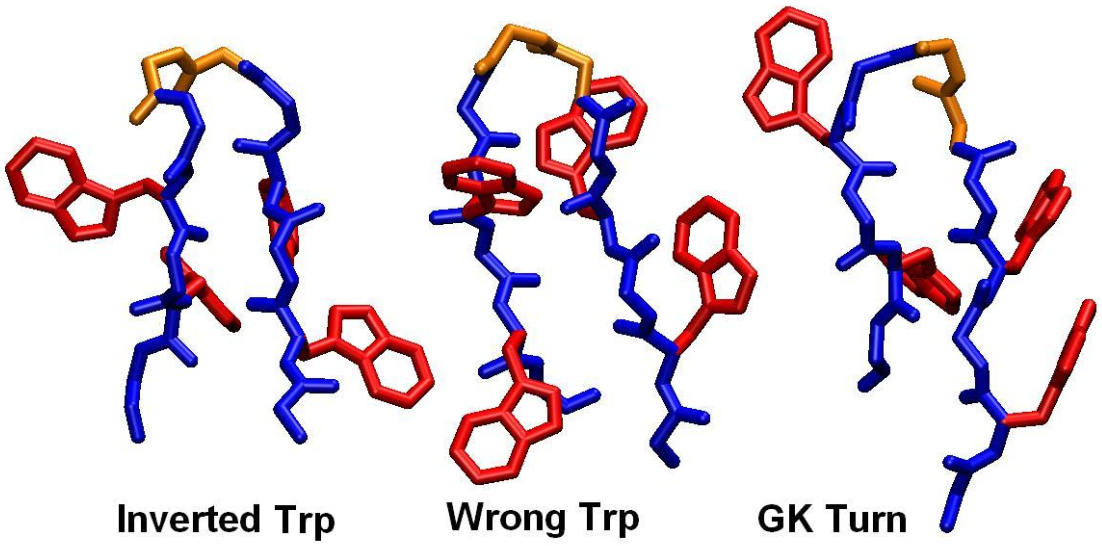

Figure 2-3. Misfolded structures extracted from the non-native free energy minima shown in Figure 2-2. Structures are shown with only the backbone (blue) and Trp residues (red) in licorice representation. The backbone of the NG turn is colored orange. Hydrogens are omitted for clarity. In the invertedTrp structure, both strands are inverted so that the Trp residues are on the opposite face of trpzip2 compared to the native structure. In wrongTrp only one strand is inverted. In the GKturn structure, the turn has been shifted one residue towards the C-terminus. Picture generated with VMD 1.8.4 [3].

Although there is a clear barrier to folding to and unfolding from the native state at RMSD = 1.7 Å in Figure 2-2, it is impossible to ascertain from the limited information available in this plot whether the non-native free energy minima represent folding intermediates or off-pathway kinetic traps. Since kinetic information is not available from REMD simulations, folding pathways of trpzip2 were studied using standard MD as described later in this chapter.

In order to estimate the thermal stability of trpzip2, the fraction of native structure present (based on a 1.7 Å RMSD cutoff) was calculated from the REMD simulation data. The resulting melting curve is shown in Figure 2-4, along with an analogous curve generated from experimental data[56]. Agreement from about 320 K to 360 K is quite good, indicating that simulations in this temperature range are probably reliable. As in Figure 2-2, the size of the error bars indicate good convergence has been achieved between the individual REMD simulations. However, the curves begin to diverge at lower temperatures. This is likely because the solvent model used in this study is unable to reproduce the effect of cold denaturation seen in the experimental melting curve.

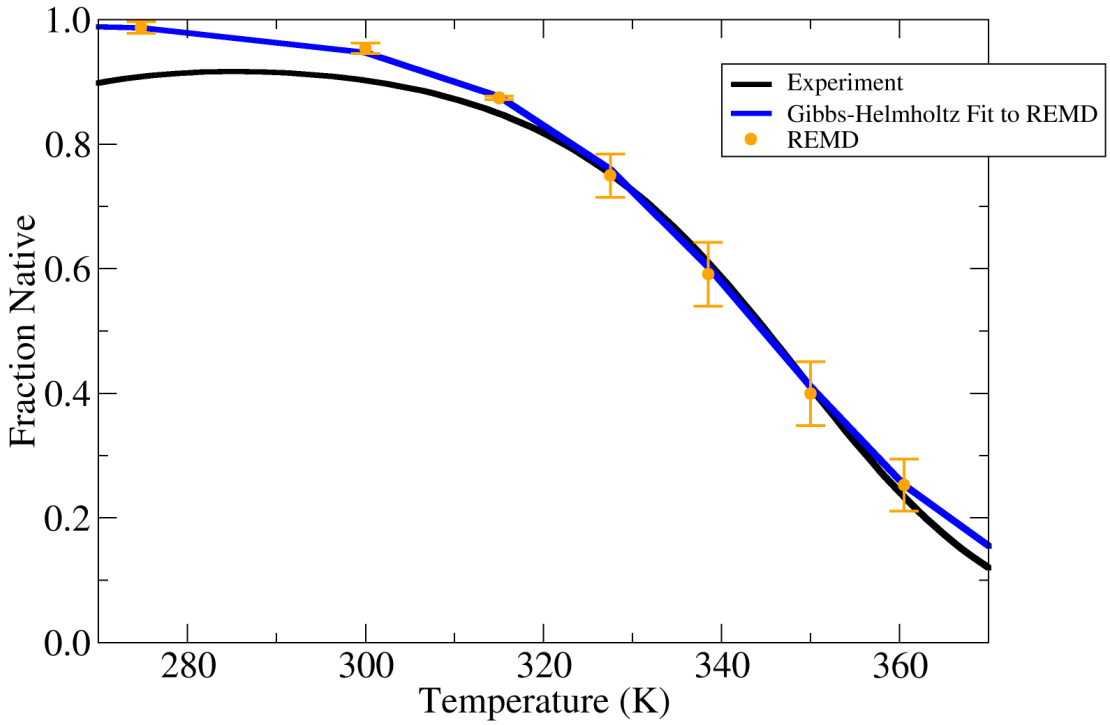

Figure 2-4. Melting curves of trpzip2, reproduced from experimental parameters[47] (black curve) and the average values from two independent REMD simulations calculated using an RMSD cutoff of 1.7 Å (orange circles). The blue curve is a fit of the Gibbs-Helmholtz equation (Equation 2-2) to the REMD data. Although the curves begin to diverge at low temperature, agreement in the range from 320 to 360 K is quite good.

The melting curve data obtained from the REMD simulations was fit to the Gibbs-Helmholtz equation (see Methods on page 13), and the melting temperature T_m, enthalpy of melting $?H_m$, and heat capacity $?C_P$ were calculated. The calculated values of T_m and $?H_m$ are in very good agreement with experiment. The calculated $?H_m$ is about 10% too small, while the calculated T_m of 344 K is just 1 K below the experimentally determined melting temperature of 345 K. In contrast, the calculated value for $?C_P$ is much smaller than the experimental value, and the uncertainty is far greater. This is because the heat capacity, when calculated from a melting curve using Equation 2-2, depends greatly on the curvature at the extreme ranges of temperature. As shown in Figure 2-4, the agreement between experiment and the REMD simulations at these temperatures is poor and likely reflects the inability of the solvent model used in this study to properly account for the properties of aqueous solvation at these temperatures.

The ability to closely reproduce the native hairpin structure and several key thermodynamic parameters suggests that the simulations provide a useful model for folding of this peptide, particularly at temperatures near T_m. An additional validation of our approach was provided by a more detailed analysis of the packing of the Trp side chains that stabilize the native hairpin. In particular, the face-to-face stacking of the

indole rings originally observed in the NMR-based conformations (PDB code 1HRX, now withdrawn) differs from that obtained after a further refinement stage that included chemical shift data[47] (PDB code 1LE1). In our simulations, the native conformations adopted the edge-to-face packing seen in the more accurate NMR-based conformations, even though all simulations that began with "native" conformations used 1HRX.

	Experiment	REMD
?Hm (cal/mol)	16770 ± 60	15998 ± 504
?Cp (cal/mol K)	281 ± 2	85 ± 50
Tm (K)	345 ± 0.1	344 ± 3

Table 2-1. Enthalpy of melting, heat capacity, and melting temperature from experiment[47] and calculated from REMD simulation data. Errors are half the difference between values obtained from each independent REMD simulation.

2.3.2 Characterization of the Non-native Ensemble

The interaction of the indole rings of the Trp side chains was suggested to be a dominant stabilizing factor for the trpzip hairpin[56]. The fraction of all possible Trp residue pairs present as a function of HBlost is shown in Figure 2-5. The plot shows that the packing in the native state (HBlost=0) is highly specific for native Trp pairs 2:11 and 4:9. The bottom Trp pair (2:11) appears particularly well packed. The middle Trp pair (2:9) shows some packing in the native structure, but not to the degree of the native pairs. The total fraction of Trp packing in the native state (sum of all Trp-Trp packing fractions) is 1.2, indicating that on average at least one well packed Trp pair is present in this state. These observations are all consistent with the NMR-derived conformation (Figure 2-1).

As more backbone hydrogen bonds are lost (HBlost increase), Trp packing becomes non-specific in nature. In addition, there is a decrease in the total fraction of Trp packing (0.44 at HBlost=5). This is consistent with a study done by Yang *et al.* which also showed residual Trp packing in the unfolded state that could only be disrupted with high concentrations of denaturant and high temperature[66]. The increase in Trp packing specificity from the unfolded state to the native state, along with the entropy loss from formation of backbone hydrogen bonds, likely represents a significant contribution to the entropic component of the free energy barrier for folding trpzip2.

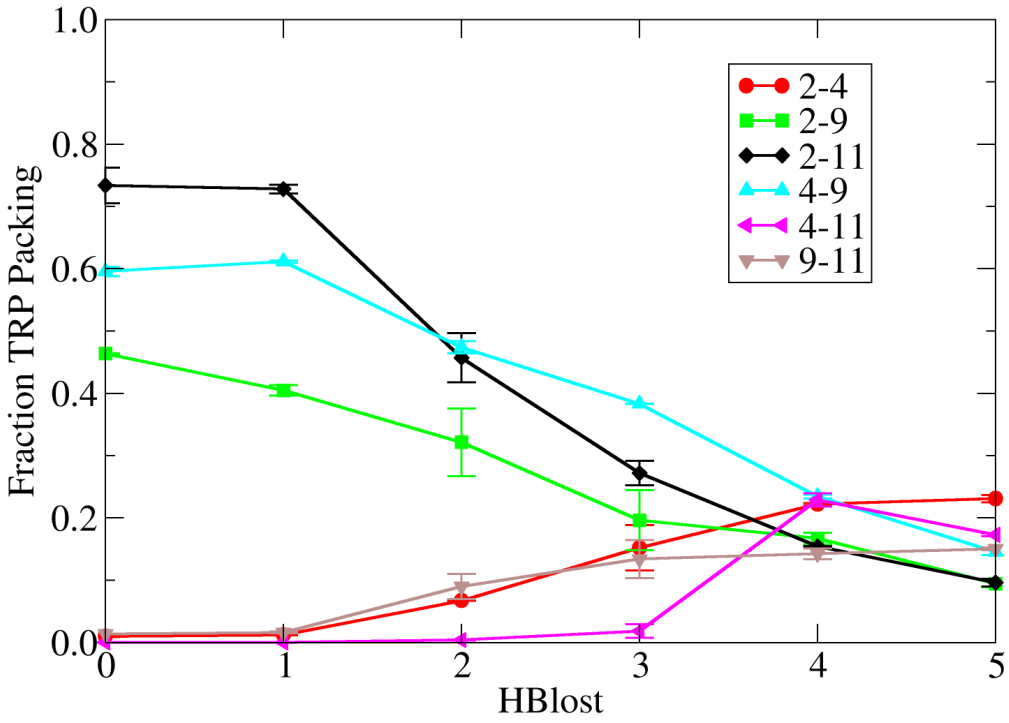

Figure 2-5. Fraction of Trp residue packing present at 350 K as a function of backbone hydrogen bonds lost (HBlost). Packing between two Trp residues was calculated using the distance between the center of mass of the two Trp residues and a distance cutoff of 6.5 Å. Packing is high for Trp2-Trp11 and Trp4-Trp9 in the native state (HBlost=0), consistent with the NMR structure shown in Figure 2-1. Packing becomes much less specific as hydrogen bonds are lost, and overall Trp-Trp packing decreases.

Figure 2-6 shows the fraction of each native backbone hydrogen bond present as a function of HBlost. According to this plot, the backbone hydrogen bond most likely to form first (~96% of the time in fact) when going from the unfolded state (HBlost=5) to the native state is the hydrogen bond closest to the reverse turn (labeled Top). This is a sensible result, and is consistent with the previously stated purpose of the reverse turn; to reduce the entropic cost of bringing the individual strands of the hairpin together. However, the Top hydrogen bond also tends to be the first hydrogen bond lost from the native state (about 88% of the time). This can be rationalized due to the large twist[56] in the trpzip2 hairpin conformation which weakens this hydrogen bond, perhaps in favor of improving interactions between the aromatic rings. In fact, it can be seen in Figure 2-5 that packing of the Trp4-Trp9 pair does increase slightly from HBlost=0 to HBlost=1. The remaining hydrogen bonds tend to be lost in order of their proximity to the reverse turn, with those hydrogen bonds closest to the turn being lost last.

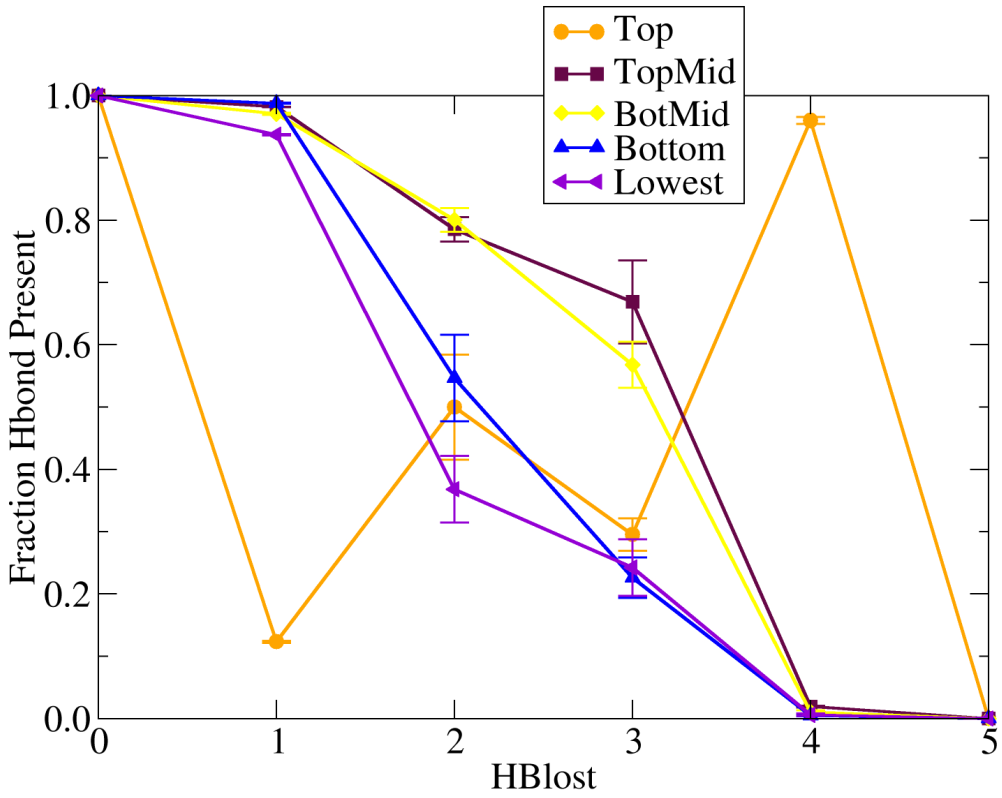

Figure 2-6. Fraction of backbone hydrogen bonds present at 350 K as a function of HBlost. The hydrogen bond nearest the turn (Top) tends to be the first one formed from the unfolded state (at HBlost=4) as well as the first one lost from the Native (at HBlost=1). The other hydrogen bonds tend to be lost starting from the turn region towards the termini of the hairpin. However, the exact pathway is not available from the thermodynamic data.

Thus, Figure 2-5 and Figure 2-6 provide a potential picture of folding. In the unfolded state, tprzip2 consists of non-specific Trp packing. The turn then serves to bring the individual strands together, stabilized by the formation of the Top hydrogen bond. This is followed by the formation of more hydrogen bonds near the turn region and more specific Trp packing. During this process the Top hydrogen bond is destabilized, and when Trp packing has become native-like this hydrogen bond tends to be lost, perhaps to accommodate more favorable arrangement of the Trp residue pairs.

Unfortunately, there is no information about actual folding pathways in the REMD simulation data because of the non-continuous nature of REMD simulation trajectories. It is also not possible to discern the relationship between the native state and misfolded structures shown in Figure 2-3. It is therefore desirable to obtain kinetic data to link these various thermodynamic data together and obtain a clear picture of the actual folding process that this peptide undergoes.

2.3.3 Temperature-jump Simulations

Ensembles of folded and unfolded structures were subjected to a temperature-jump as described in Methods and simulations were continued until first passage was recorded for 85% of the folding ensemble and 100% of the unfolding ensemble. It is critical to observe folding/unfolding events in a large percentage of each ensemble for two reasons: 1) to ensure that all relevant pathways have been sampled, and 2) the earliest observed events (rapid folding events for example) may not be relevant to the behavior of the majority of an ensemble[67].

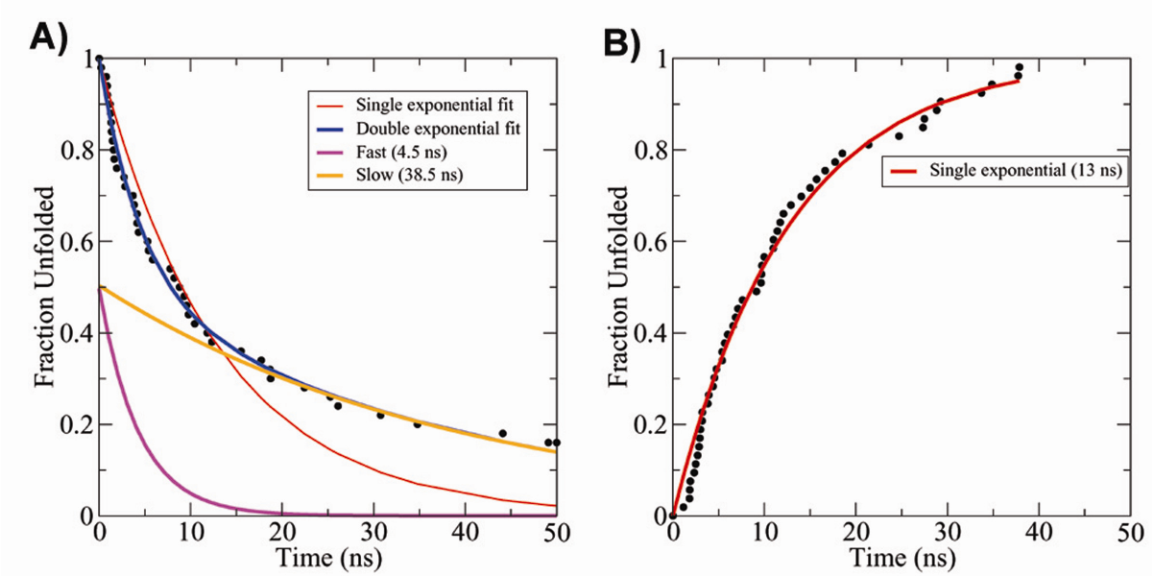

Figure 2-7. A) Fraction of unfolded structures as a function of simulation time at 350 K for the unfolded ensemble. Black dots represent a folding event in the ensemble. When the data is fit with a single exponential (red line), the fit is quite poor. However, when fit with a double exponential (blue line), the fit is improved significantly. The double exponential can be separated into a fast phase (purple line) and a slow phase (orange line) representing unfolded to native and misfolded to unfolded transitions respectively (see text for details). B) Fraction of unfolded structures as a function of simulation time at 350 K for the folded ensemble. Here, the data can be fit to a single exponential.

We initially attempted to fit the data with a single exponential, assuming simple 2-state kinetics[68]. The decay (folding) or rise (unfolding) in the fraction of non-native conformations is shown in Figure 2-7. Unfolding can be fit by a single exponential with a relaxation time of 13 ns (Figure 2-7b). In contrast, folding data requires at least two exponentials with approximately equal weights (Figure 2-7a), and relaxation times differing by nearly an order of magnitude (4.5 ns and 38.5 ns). The double exponential fit suggests the presence of kinetic partitioning[69], in this case through at least two single exponential processes. Similar partitioning has been encountered in the folding kinetics of proteins[70]. It should be noted here that a single exponential fit does not imply a single unfolding pathway (or even a single rate constant); parallel reactions initiated from the same basin will always give rise to single-exponential behavior.

In order to ascertain the reason for the double exponential fit of the folding data, trajectories that proceeded past the timescale for which the faster process is nearly complete (>20ns, Figure 2-7) were inspected for potential kinetically trapped structures

that might contribute to the slower phase. This revealed sampling of non-native metastable hairpin conformations; in fact, the same conformations shown in Figure 2-3 that were obtained from the REMD simulations (wrongTrp, GKturn, and invertedTrp). None of these misfolded structures were present in the unfolded ensemble (at 800 K) before the T-jump simulations.

In order to confirm that these non-native hairpin structures were responsible for the slow folding phase, simulations sampling misfolded structures were separated from the rest of the ensemble of T-jump refolding trajectories. Simulations that sampled misfolded hairpins showed single-exponential folding to the native state with a relaxation time of 31 ns, very similar to the slower phase of the double exponential behavior seen for the entire ensemble (38 ns).

The misfolded hairpin structures represent off-pathway intermediates (Figure 2-8). The transition from the unfolded to a misfolded state occurs on a timescale similar to the transition from unfolded to native state (~4 ns in each case). In addition, none of the unfolding trajectories sampled the misfolded structures prior to their transition into the unfolded state. These results demonstrate that the transition into the misfolded structures is not responsible for the slow folding behavior. Examination of refolding trajectories revealed that the transition between misfolded and native conformation is not direct; misfolded structures always show significant unfolding prior to reaching the native state (see Figure 2-9 for an example of such a trajectory). The misfolded structures have RMSD values near 2.4 and 3.2 Å, yet RMSD values rise to ~5 Å before successful folding occurs. This serves to illustrate how folding pathways cannot be directly observed from thermodynamic data; the apparent free energy barrier between misfolded and native structures observed in Figure 2-2 obscures the fact that transitions do not directly occur between those respective minima.

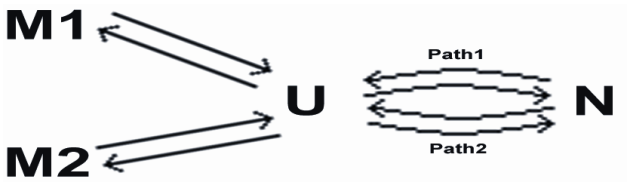

Figure 2-8. Potential folding scheme for trpzip2 based on kinetic information from T-jump simulations. Structures either fold from the Native state (N) directly to the Unfolded state (U), or fold to a misfolded structure (M). Misfolded structures are required to unfold before they can reach the native state, giving rise to the double exponential behavior seen in folding. Unfolding simply consists of a transition to the unfolded state, giving rise to the single exponential behavior seen in unfolding.

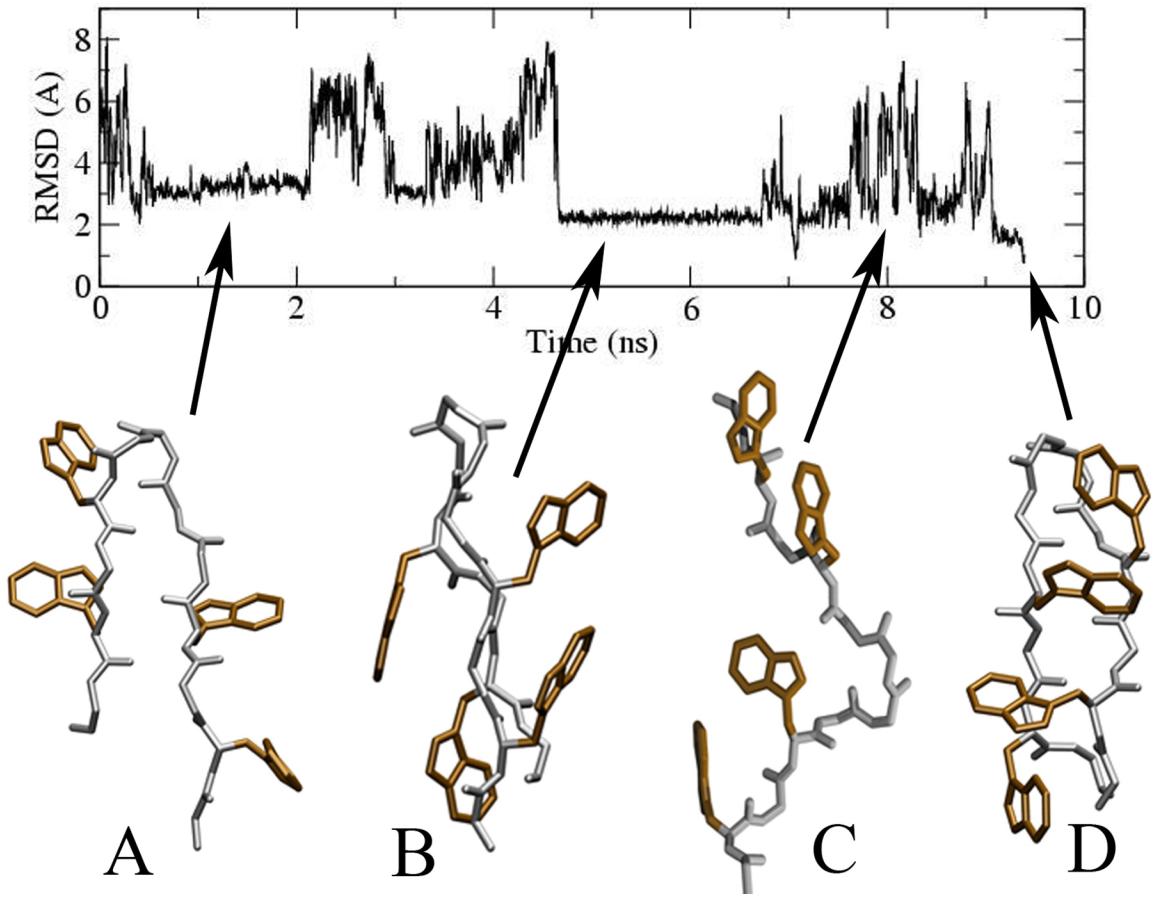

Figure 2-9. An example of a folding trajectory of one of the Tjump simulations. The unfolded structure at the beginning rapidly relaxes to the GKturn structure (A), where it remains for some time. Unfolding occurs at ~2 ns, and the structure refolds to the WrongTrp structure at ~4.4 ns (B). After unfolding again, the native structure is finally found (D). This trajectory serves to illustrate that there is no direct transition between misfolded states and the native state.

First passage times to the folded state were then recalculated starting from the first snapshot after leaving a misfolded basin; in other words, the transition from unfolded to misfolded conformation was eliminated from those trajectories. In this case, folding behavior was single exponential with a timescale nearly identical to the fast folding phase, consistent with the observation that a transition from the unfolded state to the native state is rapid. Thus the slow folding is not due to entering the misfolded basins from the unfolded state, nor to folding after leaving the misfolded basins; rather the rate limiting step is the unfolding of the misfolded structures.

Finally, the ensemble of trajectories that never sampled a misfolded conformation (~50%) shows single exponential folding with a relaxation time nearly identical to the fast phase of the double-exponential fit to all simulations seen in Figure 2-7. This indicates that the trajectories that sampled a misfolded structure give rise to the entire slow phase of folding. The fact that the unfolding T-jump simulations can be fit to a single exponential is consistent with no direct transitions between the native state and misfolded conformations.

2.3.4 Analysis and Comparison of Folding and Unfolding Pathways

As we noted above, folding events were not observed to originate directly from the misfolded hairpins. Thus, folding pathways were examined using the ensemble of folding trajectories that did not become kinetically trapped in a misfolded basin. For unfolding, the entire ensemble of unfolding trajectories was used. For each ensemble, the time-dependent fraction of 7 native contacts (the 5 native backbone hydrogen bonds and 2 Trp-packing contacts) at 350 K was calculated (Figure 2-10). Comparison of relative rates of forming each contact provides insight into the sequence of contact formation (in an ensemble-averaged manner). Comparison of folding and unfolding also highlights any differences in the two processes.

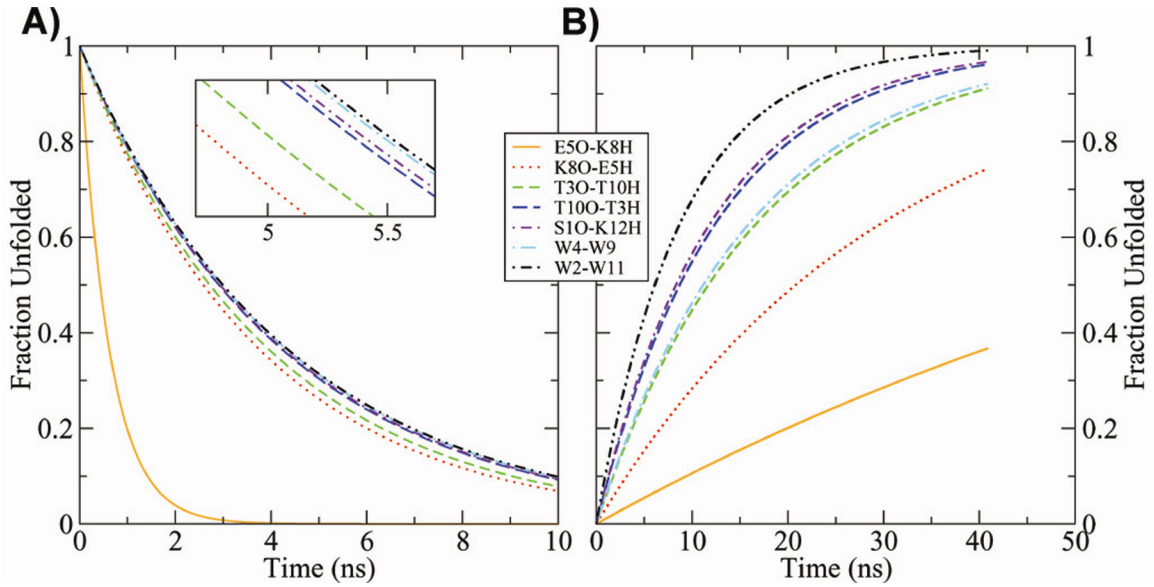

Figure 2-10. A) Fraction of backbone hydrogen bonds and Trp-packing vs simulation time at 350 K for the portion of the unfolded ensemble that did not sample any misfolded structures. The turn forms almost immediately, followed later by rapid formation of backbone hydrogen bonds and Trp-packing in that order. Backbone hydrogen bonds form proceeding from the turn region towards the termini. B) Fraction of backbone hydrogen bonds and Trp-packing vs simulation time at 350 K for the folded ensemble. The bottom (relative to the turn region) Trp pair (W2-W11) breaks first, followed by the lowest hydrogen bonds, followed by the top Trp pair (W4-W9), followed by the rest of the hydrogen bonds. The top two hydrogen bonds break much slower than the rest of the contacts.

The most apparent feature of this data is that the contact corresponding to the ß-turn forms on a much more rapid timescale than any of the other contacts. During unfolding, this contact was lost most slowly and was retained by a large fraction of the ensemble even after the remaining contacts were nearly completely lost. Both sets of observations imply a high tendency to form the turn in the unfolded state, consistent with analysis of the REMD data.

For the remainder of the contacts, it is interesting to note that the rates of contact formation during folding vary by less than 15%, while nearly 300% variation is seen during unfolding under the same conditions. In both cases, however, the ordering of backbone hydrogen bond formation or loss is consistent, with zipping occurring from the turn out, and unzipping from the termini toward the turn.

Although unfolding is the reverse of folding with respect to the backbone hydrogen bonds, Trp packing shows an important difference. During folding (Figure 2-10a), native Trp-Trp contacts formed only after the hairpin was complete, with Trp4-Trp9 forming before Trp2-Trp11. In contrast, unfolding (Figure 2-10b) occurs by initial loss of a single Trp pair contact (usually Trp2-Trp11), followed by loss of the adjacent backbone hydrogen bonds, then the second Trp pair contact, and finally the last set of hydrogen bonds. Thus formation of the two native Trp pairs is the last step during folding, but loss of both pairs is not the first step during unfolding.

These trends in the ensemble data were confirmed by visual inspection of multiple individual trajectories. It was found that unfolding actually occurs simultaneously by two very different pathways. In the predominant pathway (90% of the unfolding simulations) unfolding proceeds by initial loss of the Lowest hydrogen bond (S1O-K12H) followed by successive loss of backbone hydrogen bonds from the termini towards the turn, consistent with the order of contact loss seen in Figure 2-10b. However, a second minor unfolding pathway also exists (10%) in which the TopMid hydrogen bond (K8O-E5H) is initially lost, destabilizing the turn region, and unzipping proceeds away from the turn towards the termini. This pathway is not apparent from the contact loss curves obtained these simulations, presumably due to the lower weight of this unfolding pathway in the ensemble data. It is also of interest to note that no reverse of the minor unfolding pathway was seen for any member of the folding ensemble (hydrogen bonds for the open end of the hairpin never formed before those near the turn).

A free energy landscape was constructed using the Lowest and TopMid hydrogen bond distances from the converged REMD simulation data at 350 K (Figure 2-11a). As the data is presented here there are two free energy barriers to cross in order for trpzip2 to unfold. The first is the initial escape from the native free energy minimum, and corresponds to breaking either the Lowest or TopMid hydrogen bond. The second corresponds to breaking all remaining contacts; this barrier is significantly higher if the TopMid hydrogen bond is broken first. There is also another free energy barrier for this pathway that occurs at a TopMid distance ~5.5 Å. Since pulling apart the two strands near the turn region is inherently more difficult than pulling them apart at the termini, this likely corresponds to breaking another contact – another hydrogen bond or a Trp pair. Higher free energy barriers for this pathway are consistent with the fact that this pathway only occurs in ~10% of unfolding simulations.

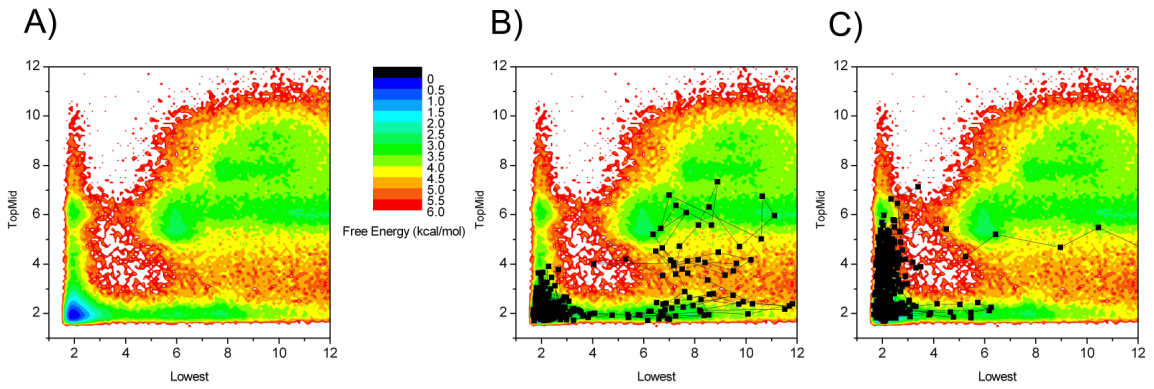

Figure 2-11. A) Free energy landscape as a function of the Lowest and TopMid hydrogen bond distances from REMD simulations of Trpzip2 at 350 K. B) Unfolding pathway observed during T-jump simulation overlaid onto the free energy landscape – unfolding starts at the termini and proceeds towards the turn. This was the predominant unfolding pathway (~90%). C) Alternate unfolding pathway – unfolding starts near the turn and proceeds towards the termini. This was a minor pathway (~10%).

In order to better visualize these pathways, the last 500 frames from unfolding simulation trajectories sampling each pathway were projected onto the landscape (Figure 2-11b and Figure 2-11c). Each unfolding pathway begins in the free energy minimum. It is seen here that structures can travel as much as 4 Å along either coordinate without then crossing into the unfolded basin; this simply corresponds to partial unfolding events. In Figure 2-11b the Lowest hydrogen bond is eventually lost completely, and the structure then spends some time in the second free energy basin before finally crossing into the unfolded basin. Likewise, in Figure 2-11c the TopMid hydrogen bond is lost completely, but the structure then spends much more time in the second free energy basin before crossing into the unfolded basin, most likely because the barriers are much higher.

The landscape thus suggests an explanation for the more cooperative folding process as compared to unfolding (Figure 2-10). Multiple barriers are encountered during both of the unfolding pathways, yet no significant free energy barrier to folding is present once either native hydrogen bond has formed. It should be noted however that since this landscape employs backbone hydrogen bond order parameters, it does not provide insight into the different coupling of these parameters to Trp pair contact formation observed between folding and unfolding simulations.

2.4 Conclusions

The thermodynamics and kinetics of the model ß-hairpin peptide trpzip2 were studied via well-converged REMD simulations and several µs of T-jump MD simulations. A free energy plot of the REMD data as a function of backbone RMSD from native trpzip2 showed a well-defined native minimum. There were also 2 well-defined non-native minima. Analysis of these minima revealed they were composed of various types of misfolded structure. The melting curve of trpzip2 calculated with data from REMD simulations was in good agreement with experiment. Several thermodynamic parameters were also calculated and found to agree with experimental values, except for

heat capacity; this was the result of the poor behavior of the GB solvent model used at temperature extremes.

The unfolded state of trpzip2 was studied with respect to Trp-Trp residue packing and backbone hydrogen bond formation. It was found that Trp packing is highly specific in the native state. Trp packing is also present in the unfolded state, although it becomes unspecific. The backbone hydrogen bond nearest the turn was found to have a tendency to be the first formed from the unfolded state. However, this hydrogen bond was found to have a tendency to be broken in the native state.

Unfolding in T-jump simulations was found to be single exponential in nature. However, folding was found to be double exponential, and comprised a fast and a slow phase. The fast phase was found to correspond to the transition from unfolded directly to native trpzip2. The slow phase was found to correspond to kinetic trapping of trpzip2 into the misfolded states observed in the REMD simulations. Misfolded structures were required to unfold before they could fold to the native state.

Analysis of hydrogen bond and Trp packing formation *vs.* time for folding pathways that did not encounter misfolded structures showed almost immediate formation of the turn hydrogen bond, followed by rapid formation of the hydrogen bonds from the turn region towards the termini and then formation of Trp-Trp cross-strand pairs. However, unfolding was found to proceed by disruption of the Trp pair nearest the termini, followed by disruption of adjacent hydrogen bonds, followed by disruption of the next Trp pair, and then disruption of remaining hydrogen bonds. This analysis reflects the behavior of the ensemble as a whole, so minor pathways may not be reflected.

In fact, two unfolding pathways were identified for trpzip2 based on order of backbone hydrogen bond loss. In the predominant pathway (~90%), hydrogen bonds were lost from the termini towards the turn, but in another minor pathway (~10%) hydrogen bond loss started from the turn and proceeded towards the termini.

These data present a complex picture for the folding and unfolding of trpzip2. In the unfolded state, trpzip2 most likely resembles a molten globule-like structure with non-specific Trp packing. The Trp packing serves along with the turn in keeping the two strands of the hairpin in close proximity. The strands eventually come together and rapidly form backbone hydrogen bonds; however, the backbone hydrogen bond pattern may be non-native and a misfolded structure can result. If the backbone hydrogen bonding pattern is native, the native Trp-pairs also form rapidly and the folded structure is reached. Unfolding requires the breaking of a Trp-pair, which is followed by breaking of adjacent hydrogen bonds. Unfolding most often begins at the termini, but in certain cases can proceed from the turn region.

The implication of the folding pathway for trpzip2 being different from the unfolding pathway is that different free energy barriers are encountered during each process. This is not to say that microscopic equilibrium is not maintained – it is possible that folding could occur by formation of hydrogen bonds at the termini first, but the reverse turn makes this so unlikely that it is simply never observed.

It should be noted that the accuracy of the data obtained in this study is sensitive to choice of forcefield and solvent model. In particular, the solvent model used in this study has shown a tendency to over-stablize a-helical conformations[71-74]. However, the forcefield may be used to compensate for solvent model inadequacies through modification of torsional parameters[75]. While this does not address the underlying

problem of solvent model accuracy, it can be used to obtain useful data as long as one is careful to maintain agreement with experimental results. The forcefield used in this study was developed based on trpzip2 with same solvent model as in this study[63], and it is noted that good agreement with experimental data was obtained in this study.

Chapter 3
Folding Cooperativity in a 3-stranded ß-sheet Model

3.1 Introduction

How a protein folds into its final three-dimensional structure based on the information contained in a linear chain of amino acids is one of the most important problems in molecular biology. At the basic level of this process is the formation of units of protein secondary structure, a-helices and ß-sheets. ß-sheet formation is more complex than a-helix formation; ß-sheets are made up of two distinct structural elements, strands and turns. Interactions between strands can occur between residues that are quite distant from each other in the protein chain, and whether these residues are in hydrogen-bonded sites or not can change the stability of their interaction[76]. The type of turn linking strands together can also have a profound effect on stability[77]; a strongly turn-promoting sequence reduces the entropy cost of bringing two ß-strands together.

The folding process of many proteins is thought to be cooperative and consisting of two-states[68]: the native or folded state, and the unfolded state. Cooperativity can also be seen at the level of secondary structure. The formation of isolated a-helix structure has been shown to be cooperative[78-80]. Unlike a-helices however, ß-sheets have the ability to exhibit cooperativity in two dimensions: parallel to strand direction and perpendicular to strand direction[81].

Cooperativity in ß-sheet formation along the direction of the strand can be thought of in terms of native backbone hydrogen bond formation between two adjacent strands (essentially ß-hairpin formation). If the process is cooperative, the formation of each hydrogen bond between strands has a lower free energy cost than the formation of the preceding hydrogen bond. Stanger *et al.* observed that the stability of several designed ß-hairpins was increased when the strand length was increased from 5 to 7 residues[82]. In a model study of backbone hydrogen bond formation in ß-sheets, Guo *et al.* found there was a sequence-independent cooperativity parallel to strand direction inherent in ß-sheet structure hydrogen bond formation[83].

Cooperativity perpendicular to strand direction can be thought of in terms of multiple hairpin formation: the formation of one hairpin will reduce the free energy cost of forming another hairpin. In studies of a designed three-stranded ß-sheet, Beta3s, de Alba *et al.* found little cooperativity if any between strands, *i.e.* perpendicular to strand direction[84]. However, the reference peptides used to calculate cooperativity in Beta3s contained charged termini while Beta3s did not, and the overall population values were quite low (13-31%), so a cooperative effect (if present) may not have been seen. In another study, Sharman & Searle found that a designed three-stranded ß-sheet was more stable than its isolated C-terminal hairpin in aqueous methanol[85], indicating some cooperativity. Griffiths-Jones & Searle obtained similar results for 3ß, another designed three-stranded ß-sheet, and calculated that adding the N-terminal strand stabilized 3ß by 0.26 kcal mol^{-1}[86].

Syud *et al.* observed a larger effect in studies of yet another designed three-stranded ß-sheet, DPDP, and calculated that the formation of the N-terminal hairpin

stabilized the C-terminal hairpin by as much as 0.42 kcal mol^{-1}[87]. DPDP was originally designed by Schenck & Gellman[81] specifically to study the cooperativity of ß-sheets perpendicular to strand direction. DPDP takes its name from the D-Pro residues in its two hairpin turns. D-Pro residues are strong promoters of type I' and II' turns[88] which strongly favor ß-hairpin formation[89]. For example, significant stabilization of Beta3s occurred when its turn sequences were replaced with D-Pro-Gly turns[90]. DPDP was observed to fold into the expected three-stranded ß-sheet structure by NMR spectroscopy[87]. The ß-sheet population of DPDP ranges from 75-83% at 277 K as estimated from NMR data by Ha chemical shift deviations, to 42-59% at 278 K as estimated by CD and FTIR[91]. We note that the NMR experiment probed the amount of secondary structure present only in the C-terminal hairpin, while the CD/FTIR data reflects overall secondary structure content for the entire peptide. Melting curves calculated for DPDP via these methods lacked the sigmoidal characteristic expected for two-state folding. The varying ß-sheet populations given by the two different probes are also consistent with a non-two-state process. Syud *et al.* have suggested that a four-state model is more reasonable for DPDP[87].

In order to study cooperativity perpendicular to strand direction in DPDP, Schenck *et al.* created LPDP, an analogue of DPDP in which the turn D-Proline in the N-terminal hairpin was replaced with L-Proline, effectively abolishing N-terminal hairpin formation. A subsequent analysis by Syud *et al.* compared the free energy of formation of the C-terminal hairpin in DPDP (with the N-terminal hairpin at least partially present) to the free energy of formation of the C-terminal hairpin in LPDP (N-terminal hairpin absent)[87]. The resulting data established that having the N-terminal hairpin present in DPDP stabilized the formation of the C-terminal hairpin by about 0.4 kcal mol^{-1}. Syud *et al.* also found cooperativity in similar experiments with a designed four-stranded ß-sheet[87].

In this study, the thermodynamic behavior of DPDP was studied via more than 2.4 µs of standard molecular dynamics (MD) and replica exchange molecular dynamics[35, 36] (REMD) simulations. REMD (also known as parallel tempering MD) is a simulation technique that is able to cross high energy barriers in a shorter amount of time and provide improved sampling at lower temperatures than standard MD. REMD works by simulating N non-interacting replicas at N different temperatures, where N is chosen so that there is sufficient energy overlap between replicas. A replica at low temperature is given a greater chance to cross energy barriers by being exchanged with a replica at a higher temperature. This exchange is accepted or rejected based on a Metropolis criterion. Further details of the method have been presented elsewhere[35, 36].

Unlike standard MD, REMD simulations were able to provide reproducible free energy surfaces, illustrating the improved convergence of REMD over normal MD. The global free energy minimum obtained with REMD (from which a representative structure is shown in Figure 3-1) is shown to adopt the expected 3-stranded ß-sheet conformation (an atomic-detail structure for the 3-stranded sheet form of DPDP has not been published). The simulation data shows that the C-terminal hairpin is significantly more stable than the N-terminal hairpin, consistent with the higher ß-sheet population observed via NMR experiments (which focused on the C-terminal hairpin) as compared CD/FTIR measurements, which probed the entire peptide.

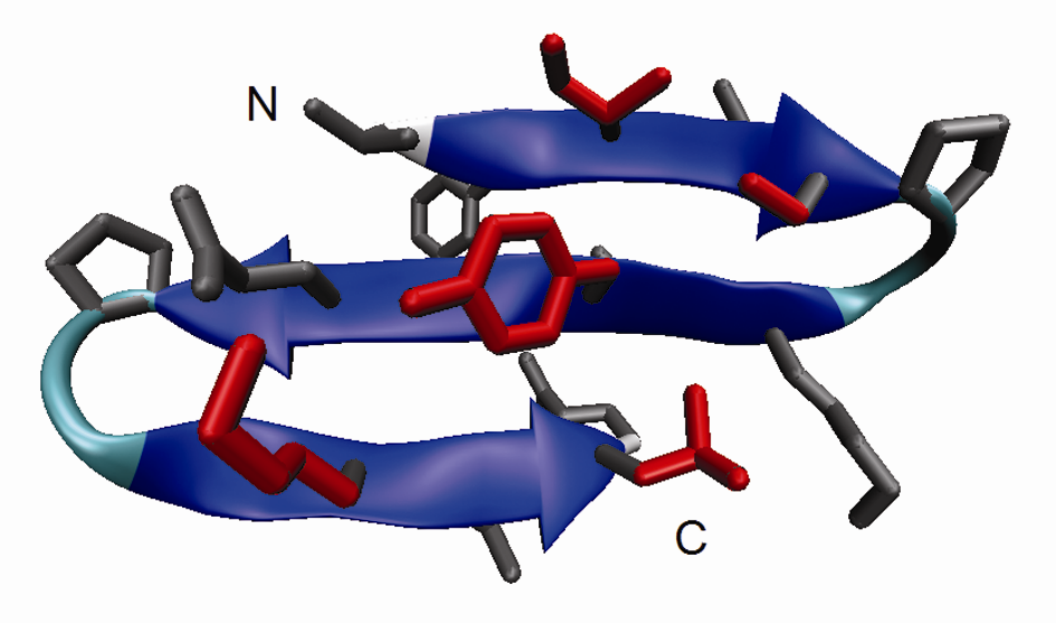

Figure 3-1. Three stranded ß-sheet model of DPDP as determined through simulation. Backbone is shown as a cartoon, sidechains are shown in a 'licorice' representation. Residues that form the hydrophobic cluster (I3, S5, Y10, K17, L19) are shown in red. Picture generated with VMD 1.8.3 [3].

Previous experiments have only provided a lower bound for the cooperativity between the component hairpins of DPDP due to the uncertainty of the state of the N-terminal hairpin in fully folded DPDP. When we analyze our full ensemble of structures, we are able to reasonably reproduce this lower limit to folding cooperativity. However, due to the atomic-level resolution provided by the computational data, the state of each hairpin is known for every structure in the ensemble. The cooperativity can therefore be calculated directly by comparing the free energy of formation of a hairpin among subsets of the ensemble in which the state of the other hairpin is well-defined (either present or absent). The results of this work indicate that the actual cooperativity perpendicular to strand direction is about 2 kcal mol^{-1} larger than previously estimated lower limit. Similar values of cooperativity are obtained for both hairpins, suggesting this may be a general effect in ß-sheets.

An alternate approach to quantifying the extent of cooperativity could be investigated through comparing hairpin stabilities in full-length DPDP to those in peptides containing only the region corresponding to each hairpin. This type of fragment analysis has been employed to study the unfolded state of proteins under conditions in which is it poorly populated [92]. We chose not to investigate that route in the present case since the truncation of the sequence could lead to end effects such as fraying, resulting in an additional level of uncertainty that we believe can be avoided through detailed structural analysis of for the full-length peptide. Our approach of examining only the full sequence is also consistent with published DPDP cooperativity analysis based on experimental data [87], thus permitting a direct validation of our results.

3.2 Methods

3.2.1 Model System

A model system was created from the sequence of DPDP (V_1FITSdPGKTY$_{10}$TEVdPGOKILQ$_{20}$, dP=D-proline, O=ornithine) except that a Lysine was substituted for the Ornithine. Replacing Ornithine with a Lysine in a related peptide analogous to the C-terminal hairpin of DPDP caused no detectable effect on the structure[93]. The termini were amidated and acetylated in accordance with experiments. DPDP was designed with a net charge of +2 to prevent aggregation[81], and our model retains this net charge.

3.2.2 Simulation Details

Simulations were carried out using Amber version 8.0[65]. All hydrogen atom bond lengths were constrained using the SHAKE algorithm[59]. All nonbonded interactions (ie. without cutoff) were calculated at each time step. All systems were modeled with the AMBER ff99 force-field with modified backbone parameters to reduce α-helical bias[63]. Steepest descent energy minimization was performed on all structures for 500 steps prior to simulation. All simulations were carried out using an implementation[60, 94] of the Generalized Born (GB) implicit solvent model available in Amber (igb=1). Explicit counterions were not included, consistent with most studies that employ continuum solvation models. Some studies have questioned the validity of implicit solvent models[71, 95] for protein folding studies, however, implicit solvent models have been used successfully in studies of both β-hairpin and β-sheet systems[53, 66, 96-99]. Also, previous studies of β-sheets in explicit solvent have not shown that water plays a specific structural role in the folding/unfolding process[100, 101]. Obtaining well-converged thermodynamic data in explicit solvent remains a significant challenge even for systems of the size of DPDP; therefore we employed an implicit solvent model with the understanding that careful validation against available experimental data must be obtained.

Simulations were performed using REMD as implemented in Amber version 8. A total of 12 replicas were used for REMD simulations of DPDP at the following temperatures: 260.1, 279.3, 300.0, 322.2, 346.0, 371.6, 399.1, 428.7, 460.4, 494.5, 531.0, and 570.3 K. The number of replicas was chosen so that sufficient energy overlap would be achieved between replicas. The temperature for each replica was generated based on an exponential distribution and chosen so that an exchange acceptance ratio of 0.15 would be achieved. Exchanges between replicas at neighboring temperatures were attempted at an interval of 1 ps. Temperatures were maintained between exchanges by coupling to an external bath using Berendsen's scheme[58].

3.2.3 Native Contacts and Data Analysis

A list of contacts in the β-sheet state (Table 3-1) were defined from analysis of a low temperature (277 K) 10 ns normal MD run starting from the model 3-stranded sheet structure. Hydrogen bond contacts were included if they existed for more than 70% of the

simulation, and side-chain contacts were considered extant if the center of mass of the side-chain (Ca for the glycines) was less than 6.5 Å from another non-neighboring side-chain more than 60% of the time. This procedure is similar to that followed in a study of a different three-stranded ß-sheet[100]. Based on these criteria 28 contacts were defined, as listed in Table 3-1. Cutoffs for QH1 and QH2 were chosen based on the boundaries of the free energy basins obtained for these order parameters in DPDP (Figure 3-3). Hydrogen bonds were defined as a distance between hydrogen donor and acceptor of less than 2.5Å. No angle cutoff was used.

Hairpin	Contact	Type	Hairpin	Contact	Type
1	V1-I3	Sidechain	2	T9-Q20	Sidechain
1	V1-E12	Sidechain	2	Y10-K17	Sidechain
1	F2-T11	Sidechain	2	Y10-L19	Sidechain
1	I3-S5	Sidechain	2	T11-V13	Sidechain
1	I3-E12	Sidechain	2	T11-I18	Sidechain
1	T4-G7	Sidechain	2	T11-Q20	Sidechain
1	T4-T9	Sidechain	2	E12-dP14	Sidechain
1	S5-Y10	Sidechain	2	E12-G15	Sidechain
1	V1H-E12O	Hbond	2	E12-K17	Sidechain
1	V1O-E12H	Hbond	2	V13-I18	Sidechain
1	I3H-Y10O	Hbond	2	T9O-Q20H	Hbond
1	I3O-Y10H	Hbond	2	T11H-I18O	Hbond
2	K8-Y10	Sidechain	2	T11O-I18H	Hbond
2	K8-Q20	Sidechain	2	V13H-K16O	Hbond

Table 3-1. Native contact list for DPDP. Contacts obtained from 10 ns of standard MD simulation at 277K.

Free energy landscapes were calculated from multidimensional histograms according to ?Gi = - RTln(Ni/N0), where Ni is the population of a particular histogram bin along the desired coordinates (*e.g.* fraction of contacts and radius of gyration), and N0 is the most populated bin, making 0.0 kcal mol^{-1} the lowest free energy state. The free energy curves in Figure 3-6 were calculated in a similar manner using 1-dimensional histograms. Hairpin X was considered 'present' when QHX was greater than 0.5.

?G values for each hairpin state in Table 3-2 were calculated at 279 K using ?G = - RTln(X/N-X). For the data corresponding to the experiments in which the state of the neighboring hairpin was not determined, X is the number of structures with the hairpin of interest present and N is the total number of structures. This corresponds to the net ?G for formation of the hairpin in the entire ensemble. For the data in which the state of the neighboring hairpin is considered, X is the number of structures with the hairpin of interest present and also with the neighboring hairpin in the specified state (either present or absent). N is the total number of structures with the neighboring hairpin in the specified state. For example, in the case of ?G for formation of hairpin 2 with hairpin 1 present, X is the number of structures with both QH2>0.5 and QH1>0.5, while N is the number of structures with QH1>0.5.

Uncertainties in these ?G values were calculated as half the difference between the values obtained from the linear and collapsed simulations. Uncertainties in ??G values were obtained from the square root of the sum of squares of the individual ?G uncertainties.

3.3 Results

Since no atomic-resolution structure has yet been published for DPDP, several different conformations were generated as starting points for simulations: a model of the fully folded 3-stranded ß-sheet, a completely extended structure generated by the Leap module of AMBER using only the amino acid sequence (referred to hereafter as linear), and a compact structure (referred to hereafter as collapsed). The collapsed structure was chosen at random from an ensemble of structures generated with standard MD simulations starting from the linear system at 350 K (see Methods for details, total simulation time ~161 ns), with the sole criteria that no ß-sheet backbone hydrogen bonding be present. A structure candidate for the fully formed sheet was selected from the same simulations as a structure that exhibited the backbone hydrogen bonding scheme expected based on analysis of NMR data for DPDP[87] (Figure 3-1). Although this structure was chosen based on backbone hydrogen bonding atom distances and not energy, further MD and REMD simulations verify that this structure falls within the global free energy minimum at 279 K (average RMSD <1.0 ?), and is therefore representative of the 3-stranded ß-sheet state in the simulation model in addition to possessing the secondary structure indicated by experimental measurements.

A list of 28 "native" contacts were defined, as listed in Table 3-1 (see Methods). Q_{Total} is defined as the fraction of 3-stranded sheet contacts formed, QH1 is the fraction of N-terminal hairpin contacts formed, and QH2 is the fraction of C-terminal hairpin contacts formed. Hereafter we refer to the N-terminal hairpin (residues 1-13) as hairpin 1 and the C-terminal hairpin (residues 8-20) as hairpin 2. For each of these contact parameters 1.0 means all contacts were present and 0.0 means no contacts were present. Hairpin 1 was considered folded if QH1 > 0.5, hairpin 2 was considered folded if QH2 > 0.5, and DPDP was considered folded only if both cutoffs were satisfied. In addition to contacts, the radius of gyration (RG) for the side-chains of residues I3, S5, Y10, K17, and L19 was used as an alternate order parameter. A clustering of these residues on one face of DPDP was inferred from NOE data[87].

3.3.1 Molecular Dynamics Simulations

It is important to validate the extent of sampling convergence before performing thermodynamic analysis of simulation data. A reasonable measure of convergence is obtained by comparing data obtained from simulations initiated from different conformations to look for evidence of kinetic trapping[69] on the timescale that is simulated. Any differences in the resulting data provide a lower bound for true uncertainty. Two normal MD simulations of DPDP were performed at 350 K starting from both the 3-stranded sheet structure and the linear structure (referred to hereafter as the ß-sheet and linear MD simulations) for 235 and 218 ns respectively. Two-dimensional population histograms of these simulations calculated from various coordinates are shown in Figure 3-2.

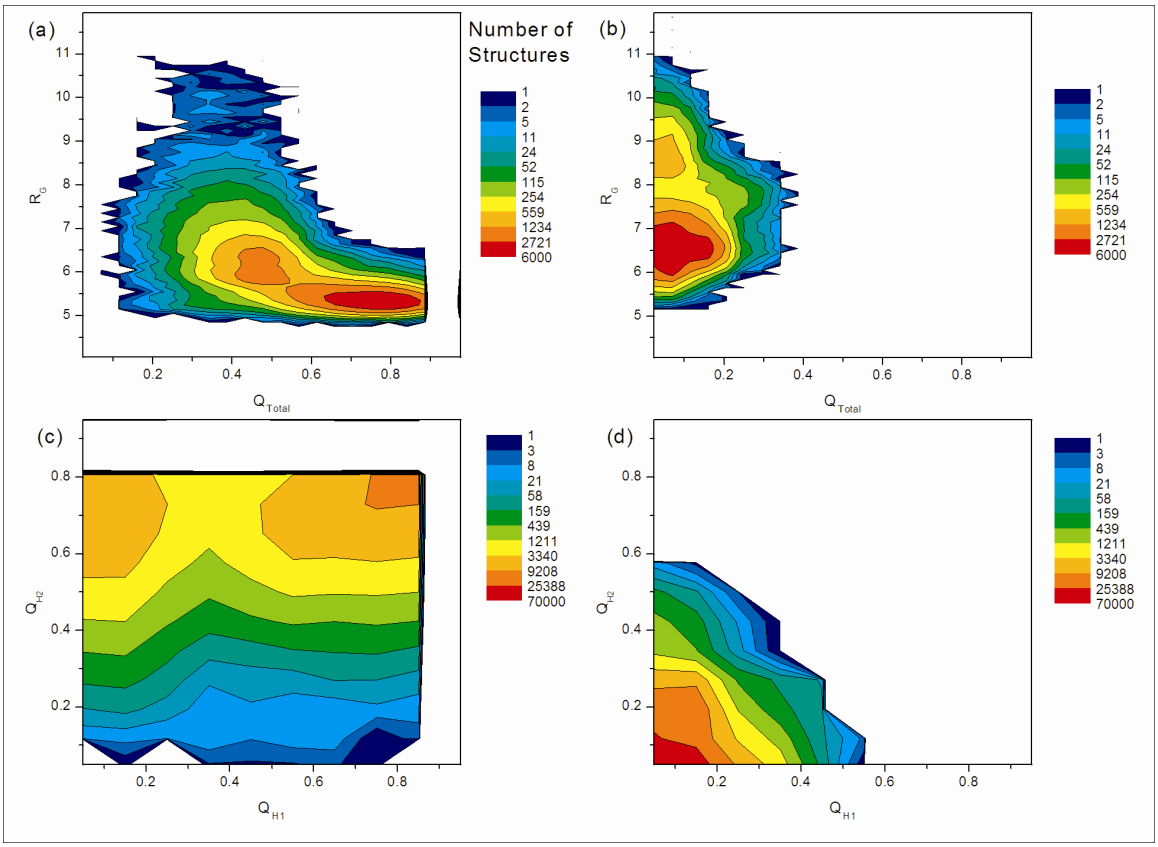

Figure 3-2. Two-dimensional population histograms of DPDP from standard MD simulations at 350 K representing about 230,000 structures. A logarithmic contour scale is used. (a) and (b) show QTotal vs. RG for simulations starting from the 3-stranded sheet model and linear structures respectively. The linear simulation (b) is trapped and never forms the three-stranded sheet. The ß-sheet model simulation (a) explores some conformational space but never fully unfolds. (c) and (d) show QH1 vs. QH2 for simulations starting from ß-sheet model and linear structures respectively. Again, the linear simulation (d) is trapped. While hairpi n 1 shows a tendency to unfold in the simulation starting from the model three stranded sheet (c), hairpin 2 never unfolds. Overall, data is poorly converged.

Figure 3-2a and Figure 3-2b show population as a function of the Q_{Total} and RG order parameters. In the ß-sheet MD simulation (Figure 3-2a) DPDP mainly stays in a wide (Q_{Total}) and narrow (RG) peak centered on a Q_{Total} of about 0.80, which corresponds to the fully formed sheet. In this state the core tends to be compact, with RG around 5.5 ?. A second peak located around Q_{Total}=0.45 and RG=6.0 ? corresponds to partially unfolded structures. A fully unfolded state is never reached. In contrast, DPDP in the linear simulation remains in a peak located below Q_{Total}=0.20, indicating that few ß-sheet contacts form and the folded state is never reached. In this unfolded state, RG fluctuates between 6.5 ? and 8.5 ?, indicating the hydrophobic core is much less compact than when the sheet is fully formed.

Next, the fractional structure sampled by each of the component hairpins was examined. Two-dimensional population histograms for order parameters QH1 vs.QH2 are shown in Figure 3-2c and Figure 3-2d. The ß-sheet simulation (Figure 3-2c) contains two major peaks: the fully formed sheet (top right) and hairpin 1 unfolded (top left). This suggests that hairpin 1 is less stable than hairpin 2. In contrast, structures in the linear

simulation (Figure 3-2d) never form a significant fraction of either hairpin; QH1 and QH2 mainly stay at about 0.0 with no other peaks present.

It should be noted that calculation of free energies from population data requires sampling of a proper Boltzmann weighted ensemble, and the large differences between the landscapes in Figure 3-2 indicate that the simulations are kinetically trapped on this timescale (~200ns). Thus, any free energies obtained from these simulations would have extremely large uncertainties and convey little insight into the thermodynamic properties of DPDP, and so were not calculated. In addition, the complete lack of correspondence between the overall topology of the landscapes shown in Figure 3-2 suggests that they may not even provide a reliable indication of the positions or characteristics of important local minima that may be traversed during the folding process. However, the relatively high stability of the ß-sheet structure during the 235 ns simulation suggests that our initial structure model is at least reasonable. In order to further probe the conformational space of DPDP, verify that the 3-stranded ß-sheet model is indeed the "native" conformation under these simulation conditions, and to calculate free energy landscapes employing various order parameters, more extensive sampling was obtained using REMD.

3.3.2 Replica Exchange Molecular Dynamics

Two sets of REMD simulations were run, with all 12 replicas in one simulation starting from the extended linear structure and all 12 replicas in the other simulation starting from the collapsed conformation (see Methods for details). The alternate starting structures were used in order to obtain convergence estimates, as described above. Neither simulation included any of the ß-sheet model structure, nor did any initial structure have any ß-hairpin hydrogen bonds. Each simulation was carried out for ~130,000 exchange attempts, for a total simulation length of ~1.5 µsec for each initial structure.

Figure 3-3 shows the free energy landscapes for Q_{Total} vs. RG and QH1 vs. QH2 at 346 K. This temperature was chosen to enable comparison to the standard MD results at 350 K shown in Figure 3-2. In addition, since this temperature is above the melting point (Tm) of the 3-stranded sheet conformation (see below), minima corresponding to non-native conformations are reasonably well populated. This facilitates visual interpretation of the folding landscape. The qualitative features of the landscape are consistent with other temperatures, and the temperature-dependent populations of the various minima will be discussed in more detail below. In order to permit direct comparison to experimental data obtained at 277 K, quantitative folding cooperativity analyses are performed using the ensemble sampled at 279 K.

Despite some quantitative differences in relative depth of the minima, it appears that the major features of the free energy landscape of DPDP are qualitatively consistent between both REMD simulations. This is a dramatic difference from the poorly converged data obtained using standard MD. Both REMD simulations were started from unfolded structures, yet both found the fully formed ß-sheet within a reasonable amount of simulation time (<15 ns, data not shown). This is quite good compared with our standard MD simulation starting from an unfolded state, which showed no ß-sheet-like structure even after 200 ns of simulation time. The simulations were able to reproducibly find three-stranded ß-sheet structures that were consistent with experimental observations

of DPDP as a fully formed sheet. These structures were the global free energy minimum at low temperatures under our simulation conditions and should be good representatives of the 3-stranded sheet state of DPDP, for which an atomic-resolution structure has not yet been published. The agreement between free energy landscapes calculated from REMD simulations is also much better than that obtained from normal MD and illustrates the superior convergence of the REMD method as has been observed by others[96].

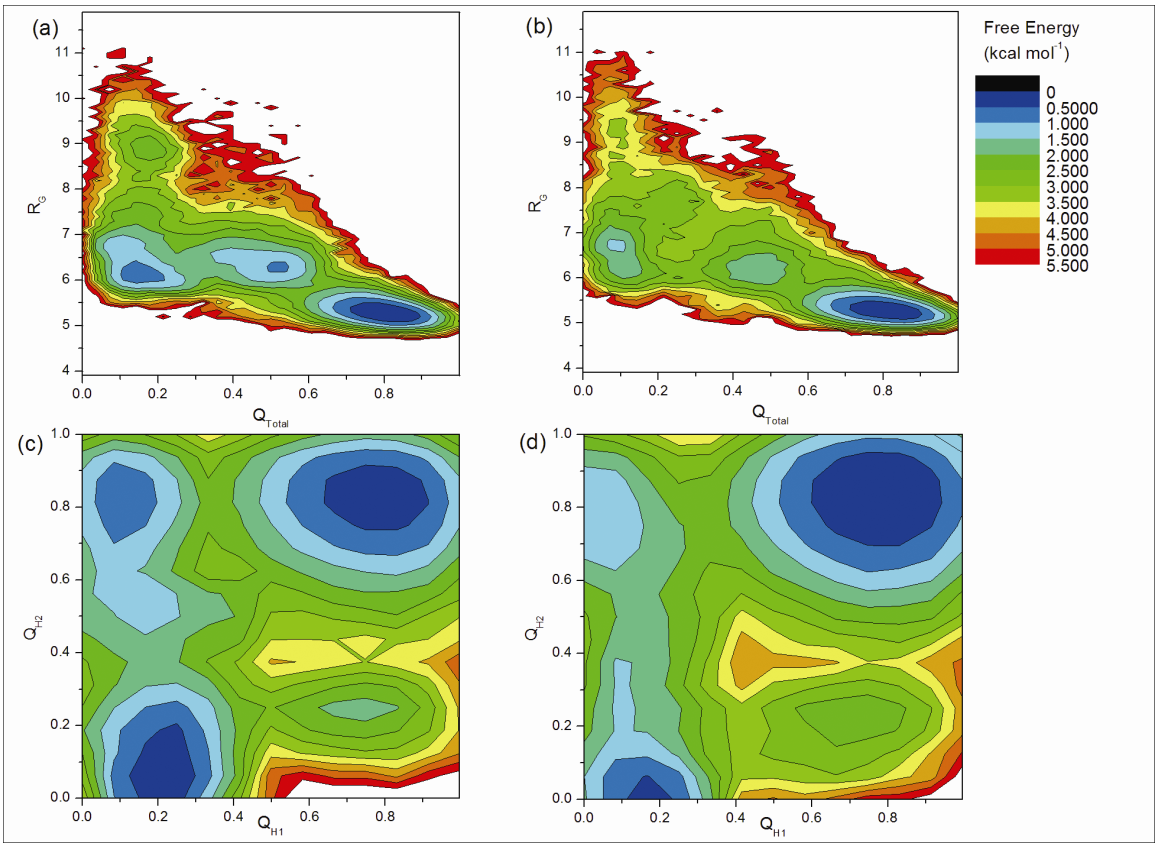

Figure 3-3. Free energy landscapes of DPDP from REMD simulations at 346K representing about 130,000 structures. Data is much better converged than that obtained from standard MD (Figure 3-2), as seen from the similarity of the landscapes from simulations starting from different structures. (a) and (b) show QTotal vs. RG for simulations starting from collapsed and linear structures respectively. There are at least three major minima in these landscapes, showing that DPDP is a non-two-state system. (c) and (d) show QH1 vs. QH2 for simulations starting from collapsed and linear structures respectively. These landscapes indicate DPDP behaves more like a four-state system, with minima corresponding to (clockwise from top-right) fully formed ß-sheet, only hairpin 1 folded, unfolded, and only hairpin 2 folded. Both (c) and (d) show that hairpin 2 alone is about 1.0 kcal mol^{-1} more stable than hairpin 1 alone.

It is apparent from the Q_{Total} vs. RG landscapes shown in Figure 3-3a and Figure 3-3b that DPDP is not a two-state system. Both landscapes contain at least three well-populated free energy minima (with depths ranging from 0-2 kcal mol^{-1}) centered at Q_{Total}=0.80, 0.45, and 0.15. The global free energy minimum, located at Q_{Total}=0.80 in both REMD data sets, corresponds to the fully formed ß-sheet under the simulation conditions, with high similarity (average RMSD <1.0) to the 3-stranded model structure we selected from our MD simulation (see Methods). This is particularly encouraging

because no ß-sheet hydrogen bonds were present in any of the initial structures for these REMD simulations, yet both predict the 3-stranded sheet as the global free energy minimum. The minimum is centered slightly lower than Q_{Total}=1.00, reflecting some fraying at the ends of each hairpin due to thermal fluctuations; this is expected as contacts were defined at the lower temperature of 277 K. The broad oval shape of the minimum with respect to Q_{Total} is also consistent with a moderate degree of conformational flexibility in the fully formed sheet, though the narrow range of RG values sampled in this basin indicates that the hydrophobic core remains substantially intact during these fluctuations. The next of the three minima, centered at Q_{Total}=0.45, corresponds to an ensemble of partially folded structures, the nature of which will be discussed below.

The minimum at Q_{Total}=0.15 corresponds to the unfolded state. The core is much less compact in this state, as seen from the RG coordinate which ranges from about 6.5 to 9.0 ?. The non-zero Q_{Total} indicates that this ensemble retains at least some structure elements of the ß-sheet, although this data does not reveal whether the residual contacts are consistent among all structures in this state. Analysis of the fraction of each of the contacts sampled in this fairly well-defined minimum reveals that the residual contacts arise predominantly from turn formation, which is not surprising considering the strong propensity of D-proline to form type I' and II' ß-turns[88, 89]. It is this drive toward turn formation that provides DPDP with significant stability as compared to peptides of similar length without D-proline residues[81, 87, 90], as a strong turn is related to greater hairpin stability[77, 102, 103].

Next, the stability of hairpin 1 compared to hairpin 2 was examined. Figure 3-3c and Figure 3-3d show free energy landscapes calculated from the QH1 and QH2 order parameters. These landscapes suggest that DPDP populates at least 4 states, with the four basins corresponding to formation of (clockwise from top right) both hairpins (3-stranded sheet), only hairpin 1, neither hairpin (unfolded), and only hairpin 2. This 4-state model is consistent with that proposed by Syud *et al.*[87] and Griffiths-Jones & Searle[86] for three-stranded anti-parallel ß-sheet folding. It is apparent from these figures that the basin with hairpin 2 present and hairpin 1 absent is lower in free energy than that with hairpin 1 present and hairpin 2 absent, *i.e.* hairpin 2 is thermodynamically more stable than hairpin 1. Based on relative free energies of these basins, it is estimated that formation of hairpin 2 alone is ~1.0 kcal mol^{-1} more stable than hairpin 1 alone in both simulations. Difference in hairpin stability has been observed for other 3-stranded ß-sheets[96, 104].

In order to further examine the nature of the structures in the partially unfolded ensemble (Q_{Total}=0.45), free energy landscapes were plotted as a function of Q_{Total} and either QH1 or QH2 (Figure 3-4). These landscapes show that the minima in the previous Q_{Total} vs. RG landscapes (Figure 3-3a and Figure 3-3b) centered at Q_{Total}=0.45 are themselves made up of two separate minima corresponding to having either hairpin 1 or hairpin 2 folded. This data matches the QH1 vs. QH2 landscapes shown in Figure 3-3c and Figure 3-3d, showing that four states are indeed present under these conditions. Also, it is seen that at Q_{Total} =0.45 it is more favorable (again ~1.0 kcal mol^{-1} in both simulations) to have hairpin 2 folded than hairpin 1. Since partially formed hairpins are not well-populated, each of the two hairpins shows significant cooperativity parallel to the strand direction. This is consistent with NMR data[87] that indicates similar populations of hairpin 2 when measured using different residues in the C-terminal strand.

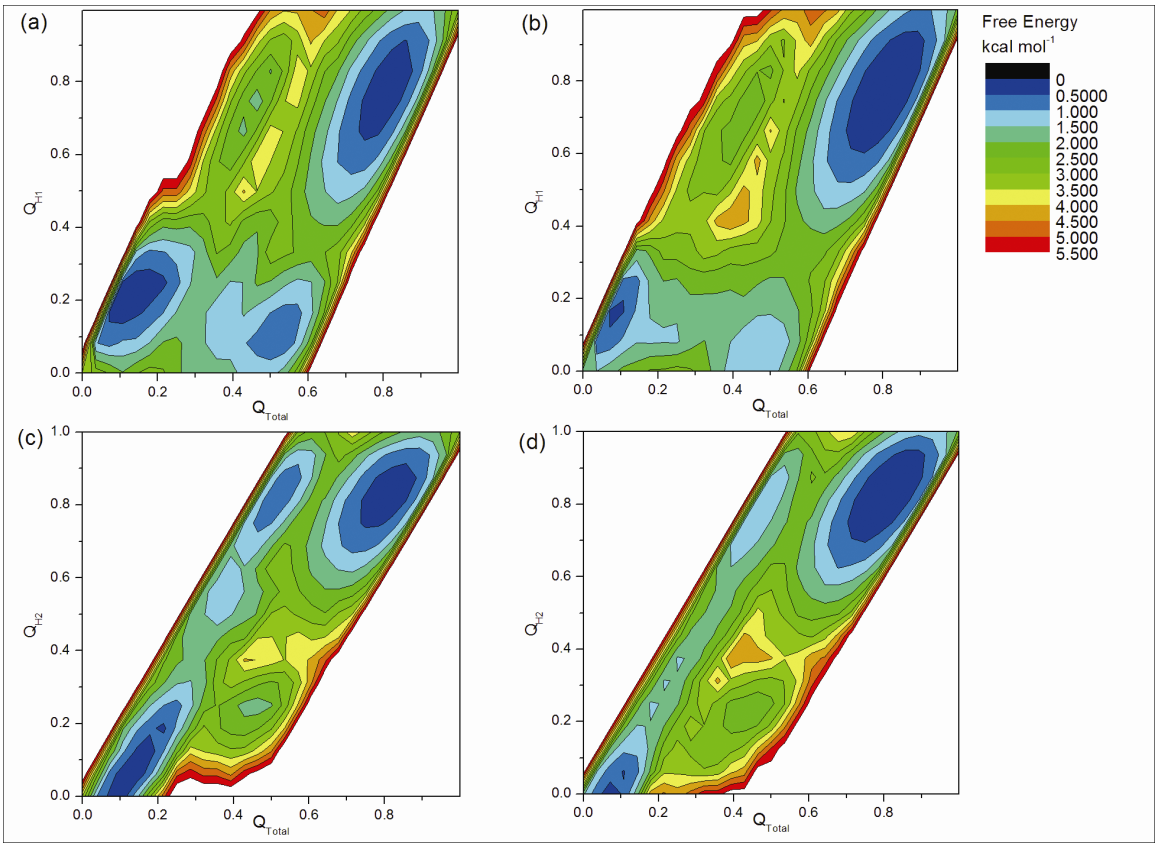

Figure 3-4. Free energy landscapes from REMD simulations at 346 K representing approximately 130,000 structures. (a) and (b) show QTotal vs QH1 for simulations starting from collapsed and linear structures respectively. (c) and (d) show QTotal vs QH2 for simulations starting from collapsed and linear structures respectively. The minima at QTotal=0.45 previously seen in Figure 3-3a and Figure 3-3b are seen here to be made up of partially folded hairpin 1 and hairpin 2 structures. Again, hairpin 1 is less stable than hairpin 2; at QTotal=0.45 structures with high QH1 are about 1.0 kcal mol^{-1} higher in free energy than structures with high QH2.

Melting curves for DPDP and its component hairpins are shown in Figure 3-5. However, it should be noted that these curves are not intended to quantitatively reproduce experimental values since the order parameter used to calculate fraction folded in this case (Q) may be quite different from the data obtained from various experimental probes, particularly since folding is not two-state; such behavior has been reported previously[64]. As described above, we use a contact fraction cutoff of 0.5 since it reasonably matches the boundaries of the free energy basins observed in Figure 3-3 and Figure 3-4. However, melting curves obtained with this cutoff slightly overestimate the hairpin and sheet stability as compared to NMR and CD measurements. This may indicate that our simulation model overestimates the stability of ß-sheet formation, although the differences correspond to only a few tenths of kcal mol^{-1} in free energy, well below our expectation of error for an additive molecular mechanics force field with a continuum solvent model. The difference may also reflect the challenge in comparing fractional population based on atomic coordinates to those based on experimental observables, particularly for systems that do not fold cooperatively. If a more restrictive cutoff value of 0.75 is used, the agreement with both sets of experimental data is significantly improved, with the melting curve for hairpin 2 falling within the range based

on NMR data and the melting curve for the 3-stranded sheet falling within the range indicated by the CD data. In any case, one must always exercise caution when interpreting data obtained using a strict cutoff based on structural properties.

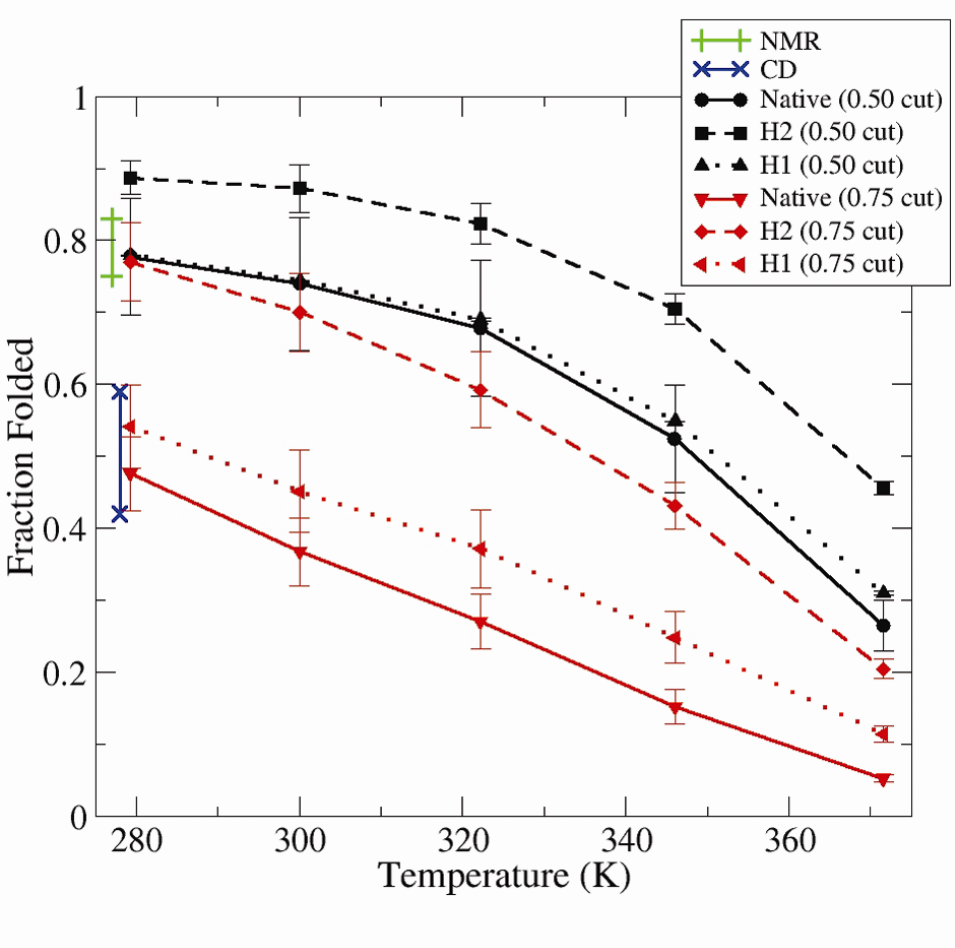

Figure 3-5. Average melting curves for DPDP hairpin 1, hairpin 2, and overall ensemble from linear and collapsed REMD simulations. The data is quite sensitive to choice of cutoff. The black lines represent a less restrictive cutoff of 0.50, while the red lines represent a more restrictive cutoff of 0.75. Hairpin 1 structures were considered folded when QH1 was greater than the cutoff, hairpin 2 structures were considered folded when QH2 was greater than the cutoff, and fully formed ß-sheet structures were considered formed when both cutoffs were satisfied. Using a more restrictive cutoff (0.75), the hairpin 2 melting curve intersects with NMR data (green line) and overall melting behavior intersects with CD data (blue line). Regardless of cutoff, hairpin 2 is more stable than hairpin 1 in each case. Error bars reflect differences between the linear and collapsed REMD data sets.

The calculated curves are also quite useful in comparing the stability of the individual hairpins to that of the full ß-sheet. It is clear from Figure 3-5 that the relative temperature-dependent stabilities do not strongly depend on the choice of cutoff; for both values tested, hairpin 2 is much more stable as compared to hairpin 1.

3.3.3 Cooperativity

Cooperativity perpendicular to strand direction in our simulations was measured in two ways using the ensemble of structures at 279 K from the REMD simulations. Since our cooperativity calculations are based solely on simulation data for DPDP, there is the additional advantage that the effect of the L-Pro substitution (LPDP) need not be considered.

The first calculation was intended to generate data comparable to that which had been obtained through experiment [87], in which hairpin 2 population was compared between DPDP and LPDP (in which hairpin 1 is expected to be absent). For this case, we calculated the free energy for formation (?G) of hairpin 2 in two sets of DPDP structures: our entire ensemble (to match DPDP in experiments) and only the subset of structures which lacked hairpin 1 (QH1<0.5, analogous to LPDP in experiments). Differences in the free energy of formation of hairpin 2 (??G) between these sets of structures are presumably attributable to the influence of the conformation of hairpin 1. Since hairpin 1 is only partially folded in DPDP, the effect on hairpin 2 will depend on the extent of hairpin 1 folding (which was not determined in the experiments and is therefore not evaluated for the purpose of this particular calculation). It is for this reason that Syud *et al.* acknowledged that their values for cooperativity represent only a lower limit.

The resulting free energy profiles for the sub-ensembles are shown in Figure 3-6, and the free energies of hairpin formation and cooperativity are listed in Table 3-2 (obtained from the ensembles at 279 K as described in Methods). It is immediately apparent that the free energy profile of formation for a given hairpin is strongly influenced by the presence or absence of the neighbor, indicating at least partial cooperativity. The formation of hairpin 2 in the overall ensemble (Figure 3-6a, comparable to DPDP in experiments) becomes 1.1 ± 0.7 kcal mol^{-1} less favorable when hairpin 1 is known to be absent (Figure 3-6e, comparable to LPDP in experiments). This is in reasonable agreement with the values of 0.38 to 0.42 kcal mol^{-1} obtained by Syud *et al.*[87]. Although only hairpin 2 was used as an experimental probe of cooperativity, we also calculated the effect of (partial) hairpin 2 formation on the stability of hairpin 1. This effect is much greater than was calculated for hairpin 2; the formation of hairpin 1 becomes 3.0 ± 0.5 kcal mol^{-1} less favorable when hairpin 2 is absent. It is apparent that without hairpin 2, the formation of hairpin 1 is very unfavorable. It is important to note that this approach resulted in a ß-sheet cooperativity value that depends significantly on which hairpin is considered in the analysis (- 1.1 vs. - 3.0 kcal mol^{-1}).

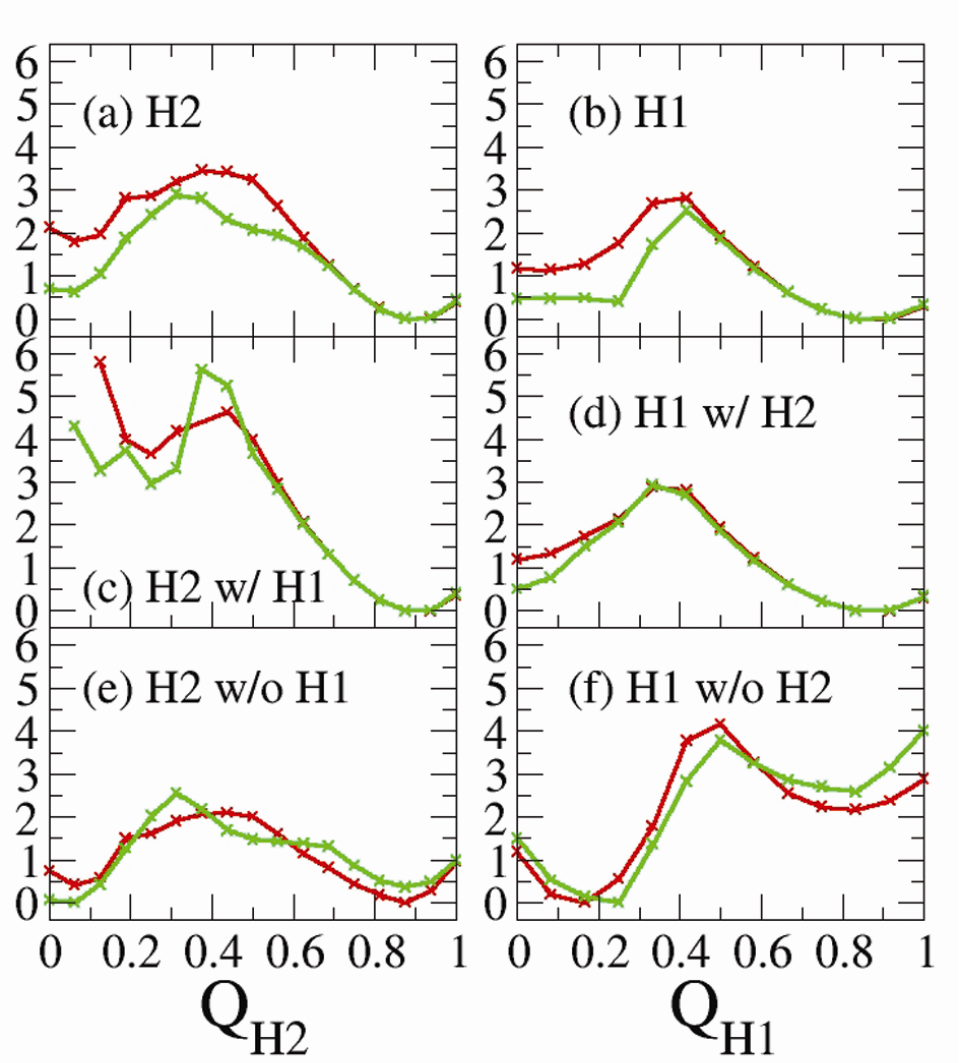

Figure 3-6. Free energy (in kcal mol^{-1}) of individual hairpin formation in DPDP at 279 K for varying states of the other hairpin. The red curves were calculated from the REMD simulation starting from the linear structure, and the green curves were calculated from the REMD simulation starting from the collapsed structure. Hairpin 2 formation is shown as a function of QH2 (a) with the state of hairpin 1 undetermined (equivalent to DPDP in experiments), (c) with hairpin 1 present, and (e) with hairpin 1 absent (equivalent to LPDP in experiments). Hairpin 1 formation is shown as a function of QH1 (b) with the state of hairpin 2 undetermined, (d) with hairpin 2 present, and (f) with hairpin 2 absent. Hairpin X was considered present if QHX>0.50 and absent if QHX=0.50. The noise at low values of QH2 in (a) reflects the fact that there is a low population of structures with only hairpin 1 folded.

Our next measure of cooperativity was intended to be more direct than was possible by comparing DPDP and LPDP. In this case, the free energy of formation of a given hairpin in the ensemble of structures lacking the other hairpin (Figure 3-6e and Figure 3-6f) was compared to the free energy of formation of that hairpin when the other hairpin is known to be present (Figure 3-6c and Figure 3-6d). For hairpin 2, this ??G was calculated by comparing the free energy of formation of hairpin 2 in two sets of structures; those with hairpin 1 absent (QH1 less than 0.50) (Figure 3-6e) and those with

hairpin 1 present (QH1 greater than 0.50) (Figure 3-6c). This corrects the cooperativity value for the partial hairpin folding within DPDP. In this case the formation of hairpin 2 when hairpin 1 is present becomes 3.2 ± 0.5 kcal mol^{-1} more favorable than the formation of hairpin 2 when hairpin 1 is absent. This ??G value is nearly three times that obtained when the state of hairpin 1 was not taken into account (- 1.1 ± 0.7 kcal mol^{-1}).

a) ?G H2	-1.3 ±	0.6	**d) ?G H1**	-0.8 ±	0.4
b) ?G H2 w/ H1	-3.4 ±	0.4	**e) ?G H1 w/ H2**	-1.1 ±	0.3
c) ?G H2 w/o H1	-0.2 ±	0.3	**f) ?G H1 w/o H2**	2.2 ±	0.2
??G H2exp (a-c)	-1.1 ±	0.7	**??G H1exp (d-f)**	-3.0 ±	0.5
??G H2sim (b-c)	-3.2 ±	0.5	**??G H1sim (e-f)**	-3.3 ±	0.4

Table 3-2. Average cooperativity values calculated from both REMD simulations in kcal mol-1 at 279 K for various hairpin states of DPDP. ??Gexp refers to cooperativity values obtained using ensembles corresponding to those studied experimentally, while ??Gsim uses ensembles that correct for partial formation of the neighbor hairpin in ensemble (a). A detailed discussion of this difference is presented in the text, and uncertainty calculations are described in Methods.

In contrast, the values obtained for hairpin 1 formation when hairpin 2 is known to be present are quite similar to those obtained when the state of hairpin 2 is not taken into consideration (- 3.3 ± 0.4 kcal mol^{-1}). This is because no significant fraction of structures having hairpin 1 without hairpin 2 is sampled, as stated above. In each case, the end result is that having one hairpin present stabilizes the formation of the other hairpin by ~3 kcal mol^{-1}, indicating significant cooperativity perpendicular to strand direction.

We noted that the cooperativity values obtained for the two hairpins differed significantly when the partial folding of the partner hairpin is not considered. In contrast, when the data is corrected for partial folding of the neighboring hairpin the values become quite similar (- 3.2 ± 0.5 vs. - 3.3 ± 0.4 kcal mol^{-1}). This sequence independence of the resulting cooperativity measure indicates that including the partial folding correction results in a more robust analysis method, and also suggests that the value may reflect general properties of ß-sheets; further study on other sequences should be carried out to confirm this hypothesis.

3.3.4 Relative Hairpin Stability

The overall ensemble of structures consists of four states: folded, unfolded, and a partially folded state that is composed of structures with either one of the two hairpins formed, in agreement with the non-two-state folding behavior of DPDP observed by Syud *et al.*[87]. There exists some population of structures with only hairpin 1 or only hairpin 2 folded, resulting in a four-state folding model as has been proposed for 3-stranded ß-sheets by Griffiths-Jones & Searle [86].

Our simulations indicate that hairpin 2 is more stable than hairpin 1. The melting curve in Figure 3-5 shows the ß-sheet population ranges obtained from NMR[87] and CD/FTIR[91] experiments. The difference between these ranges agrees well with the observed difference between our simulated melting curves for hairpin 2 and the fully formed ß-sheet. Syud *et al.* measured the change in Ha chemical shifts of residues in hairpin 2, using a cyclic peptide that corresponded to a DPDP hairpin 2 as the fully folded reference state and LPLP as the random coil reference state (replacing D-proline with L-proline abolishes the hairpin turn[81]). They noted that, since folding is not 2-

state, their NMR measurements only provided the population of hairpin 2, which were used to estimate cooperativity values. In contrast, the CD/FTIR measurements of Kuznetsov *et al.*[91] probe the entire peptide and thus the average ß-sheet content for all states.

A previous computational study of DPDP by Wang & Sung[105] indicated that hairpin 1 was more stable than hairpin 2. However, their result was based on 100 ns of standard MD data at 297 K; our standard MD simulations were very poorly converged on this timescale even though they were run at a higher temperature of 350 K. In addition, this observation does not appear to be consistent with the results from NMR and CD experiments; if hairpin 1 were more stable, then the NMR experiment should report a lower population than CD, which was not observed.

Why is hairpin 2 so much more stable than hairpin 1? Our computational observation appears to be consistent with experimental studies of DPDP. Syud *et al.* found that replacing the second D-Pro in DPDP with Asn did not alter the folding pattern or cooperativity values of that hairpin significantly[87] (although overall stability was reduced). Chen *et al.* reported that mutating D-Pro in hairpin 1 to Asp resulted in formation of a ß-bulge instead of a ß-turn, while the same mutation in hairpin 2 retained the ß-hairpin shape[102]. The fact that hairpin 2 appears less sensitive to mutations may indicate that it is intrinsically more stable than hairpin 1, though this is by no means a comprehensive analysis. It is also possible that having Ser5 in the *i* position of hairpin 1 ß-turn detracts from the stability of hairpin 1. Amino acid propensities[106] indicate that Ser is strongly turn-promoting, which may lead it to compete for turn geometry with D-Pro and Gly in hairpin 1, reducing the population of the native hairpin. In contrast, the *i* position in hairpin 2 is occupied by Val, which is considered turn-breaking. This suggests that mutating Ser5 to Val may confer additional stability to hairpin 1, and perhaps also to the 3-stranded sheet state due to the cooperativity between the hairpins. Furthermore, analysis of contact melting profiles (data not shown) reveals that hairpin 2 in DPDP contains two strong non-turn contacts: a salt-bridge between Glu13 and Lys18 and a hydrophobic cross-strand interaction between Tyr11 and Leu20. Cross-strand salt-bridging is known to have a stabilizing effect in ß-hairpins[107, 108]. Rao & Caflisch suggested that the C-terminal hairpin of another designed 3-stranded ß-sheet was more stable than the N-terminal hairpin due to strong hydrophobic interactions, particularly a contact between aromatic Trp and Tyr[96]. In contrast, hairpin 1 has no salt bridge and weaker hydrophobic interactions overall. Mutations targeting hairpin 1 and Tyr11 specifically are explored in Chapter 4.

The experimentally observed cooperativity between the individual hairpins is reproduced in our simulations, although our data is more amenable to direct calculation of the actual magnitude of the effect. Schenck & Gellman studied the population of hairpin 2 in two sequences: DPDP and LPDP, in which the D-Pro in hairpin 1 was replaced by L-Pro, effectively "turning off" formation of hairpin 1[81]. DPDP demonstrated increased population of hairpin 2 as a result of partial hairpin 1 formation. Subsequent NMR experiments by Syud *et al.*[87] estimated that the additional stability conferred upon DPDP by "turning on" hairpin 1 to be from -0.42 ± 0.22 kcal mol^{-1} to -0.38 ± 0.13 kcal mol^{-1} at 277 K, depending on the Ha resonance chosen. Importantly, they noted that this value is a lower limit on the true cooperativity since hairpin 1 is not

always in the hairpin state in DPDP, thus limiting the effect on hairpin 2. In other words, hairpin 1 in DPDP is not fully "turned on".

Experimental cooperativity values are not available for hairpin 1, but our calculations show a similar effect; the free energy difference between a) having hairpin 1 irrespective of the state of hairpin 2 and b) having hairpin 1 without hairpin 2 is - 2.9 ± 0.5 kcal mol^{-1} at 279 K. Interestingly and unlike hairpin 2, this value is relatively insensitive to whether we consider the state of hairpin 2; the ??G value for formation of hairpin 1 in ensembles of structures with hairpin 2 fully present or fully absent is - 3.3 ± 0.4 kcal mol^{-1}. This insensitivity most likely reflects the relative instability of hairpin 1 with respect to hairpin 2 and the relatively low population of hairpin 1 in the absence of the full 3-stranded ß-sheet.

It is interesting to note that all of our calculations place the value for hairpin cooperativity at about - 3.3 ± 0.5 kcal mol^{-1} at 279 K. Further investigation is needed to determine whether this value reflects general properties of ß-sheet systems.

3.4 Conclusions

Several aspects of the thermodynamic behavior of the three-stranded ß-sheet DPDP were investigated via several microseconds of well-converged REMD simulation of DPDP. It was found that the replica exchange method was superior to standard MD for exploring the thermodynamic landscape of this system. DPDP was found to behave in a four-state manner, consistent with expectations based on experimental evidence. Hairpin 2 was more stable than Hairpin 1, consistent data obtained from NMR measurements of Hairpin 2 and CD/FTIR measurements of the overall peptide. The folding cooperativity perpendicular to strand direction was found to be about -3.3 kcal mol^{-1}, significantly larger than the previously estimated lower limit. Similar values were obtained for each hairpin, suggesting this may be a general effect in ß-sheet systems, although further experiments would be needed to confirm the generality of this result. Cooperativity has been observed for a 4 stranded ß-sheet sequence related to DPDP[87], so it is possible this effect may be extended to the addition of several strands, and is consistent with the tendency for ß-sheet structures to aggregate[8].

It should be noted that the accuracy of the data obtained in this study is sensitive to choice of forcefield and solvent model. In particular, the solvent model used in this study has shown a tendency to over-stablize a-helical conformations[71-74]. However, the forcefield may be used to compensate for solvent model inadequacies through modification of torsional parameters[75]. While this does not address the underlying problem of solvent model accuracy, it can be used to obtain useful data as long as one is careful to maintain agreement with experimental results. The forcefield used in this study was developed based on a ß-hairpin peptide with same solvent model as in this study[63], and was used to generate results for that ß-hairpin that matched experiment (See Chapter 2), so it might be expected to be transferable to other small ß systems using this solvent model. The good agreement of hairpin cooperativity in DPDP calculated in this study with that obtained by experiment is a good indication that these results are reliable and likely within acceptable error (~1 kcal mol^{-1}).

Chapter 4
Effect of Mutations on Individual Hairpin Stability in a 3-stranded ß-sheet Model

4.1 Introduction

An important aspect of the protein folding problem is that of protein stability. A protein can be made to fold faster or more efficiently by either stabilizing the native state, or destabilizing the unfolded state[109] or a misfolded conformation[92]. Understanding the origins of stability in proteins may therefore give insight into the folding process. In particular, understanding stability in ß-sheet structures can be particularly important, as the formation of these types of structures are implicated in diseases related to protein misfolding[8].

ß-hairpins can garner stability from two main sources; interactions between the strands, which include hydrogen bonds and hydrophobic clustering of side-chains, and turns, which can reduce the entropic penalty of bringing two strands together. In previous work on the three-stranded anti-parallel ß-sheet model peptide DPDP, it was noted that despite both turns having DPro and Gly in the *i+1* and *i+2* positions of their reverse turns, the C-terminal hairpin was more stable than the N-terminal hairpin by about 1.0 kcal mol^{-1}. For simplicity, the N-terminal and C-terminal hairpins are hereafter referred to as Hairpin 1 and Hairpin 2 respectively.

In this study, contributions to stability from the strands and turn regions of DPDP are probed by several well-converged REMD simulations of various mutants of DPDP. The affect of perturbing the hydrophobic core of DPDP is explored by mutating a Tyr residue central to this core. Several mutations focus specifically on improving the stability of Hairpin 1 by either improving hydrophobic contacts between the two strands of Hairpin 1, optimizing the *i* position of the reverse turn, or introducing a salt bridge.

Of all the mutations made, three were seen to improve the stability of the first hairpin (as judged from the melting profile of each mutant). One mutant, referred to as FT, improved stability by moving a hydrophobic Phe residue from the first strand in the N-terminal hairpin to the central strand. The other mutant, referred to as S5V, mutated a Ser in the i position of the N-terminal hairpin reverse turn to a Val. It is shown this improves stability by destabilizing a potential unfolding pathway, and so is an example of negative design. The last mutant, referred to as EVK, was just as stable as S5V despite having the introduction of a salt-bridge in addition to the turn optimization. An attempt was made to combine the FT and S5V mutations; this mutant is referred to as FTV. However, FTV showed a stability only between that of S5V and FT, indicating these mutations are not additive and perhaps compete with each other.

4.2 Methods

4.2.1 Model Systems

All systems simulated here were based off of the 3-stranded ß-sheet model peptide DPDP (sequence Ace-V_2FITSdPGKTY_{10}TEVdPGOKILQ$_{20}$-NH_2, dP=D-proline, O=Ornithine), except that a Lysine was substituted for the Ornithine. Replacing Ornithine with a Lysine in a related peptide analogous to the C-terminal hairpin of DPDP caused no detectable effect on the structure[93]. The termini were amidated and acetylated in accordance with experiments. DPDP was designed with a net charge of +2 to prevent aggregation[81], and our model retains this net charge.

Several mutants were made of DPDP by simply deleting the side-chain atoms of the target residue except for the Cß atom, then building in the side-chain of the desired type using the Leap module of Amber. Six different mutants were simulated grouped into 3 classes: mutants of a Tyr residue central to the hydrophobic core of DPDP, mutants focusing on improving the stability of Hairpin 1, and combinations of different mutations. Mutated residues in each mutant are shown in Table 4-1.

	2	4	5	8	9	10	11
Wildtype	Phe	Thr	Ser	Lys	Thr	Tyr	Thr
Tyrosine Mutants							
Y10V						Val	
Y10T						Thr	
Hairpin 1 Mutants							
FT	Thr						Phe
S5V			Val				
Combination Mutants							
EVK		Glu	Val		Lys		
FTV	Thr		Val				Phe

Table 4-1. Summary of DPDP mutants.

4.2.2 Simulation Details

Simulations were carried out using Amber version 8.0[65]. All hydrogen atom bond lengths were constrained using the SHAKE algorithm[59]. All nonbonded interactions (ie. without cutoff) were calculated at each time step. All systems were modeled with the AMBER ff99 force-field with modified backbone parameters to reduce a-helical bias[63]. Steepest descent energy minimization was performed on all structures for 500 steps prior to simulation. All simulations were carried out using an implementation[60, 94] of the Generalized Born (GB) implicit solvent model available in Amber (igb=1). Explicit counterions were not included, consistent with most studies that employ continuum solvation models.

Simulations were performed using REMD as implemented in Amber version 8. A total of 12 replicas were used for all REMD simulations at the following temperatures: 260.1, 279.3, 300.0, 322.2, 346.0, 371.6, 399.1, 428.7, 460.4, 494.5, 531.0, and 570.3 K. The number of replicas was chosen so that sufficient energy overlap would be achieved between replicas. The temperature for each replica was generated based on an exponential distribution and chosen so that an exchange acceptance ratio of 0.15 would

be achieved. Exchanges between replicas at neighboring temperatures were attempted at an interval of 1 ps. Temperatures were maintained between exchanges by coupling to an external bath using Berendsen's scheme[58].

For each mutant, two REMD simulations were run; one starting from a ß-sheet conformation, and one starting from a linear conformation. Each simulation was run for at least 80,000 exchanges, until good convergence was seen as measured by the differences in the melting curves obtained with each simulation. The total simulation time for all mutants was ~ 1.2 µs/replica.

4.2.3 Order Parameters and Melting Curve Calculation

Since none of the mutations studied here change the structure of the backbone of DPDP, the fraction of backbone hydrogen bonds formed (X) was chosen as the main gauge of stability that would be common to all structures. Four native backbone hydrogen bonds were defined for each hairpin (see Table 3-1). Overall melting curves were calculated using the criteria that for a structure to be considered folded, each hairpin must have X>0.5. Melting curves for individual hairpins were calculated using the same cutoff (X>0.5). Error bars represent half the difference observed between the independent REMD simulations.

4.3 Results

4.3.1 Tyrosine mutants

In their studies of DPDP, Syud *et al.* identified a clustering of residues (I3, S5, Y10, K17, and L19) as a possible stabilizing force in DPDP[87]. Examination of the folded structure of DPDP reveals the reason these residues are able to come into close contact is the twist of the ß-sheet, which creates a concave space on either side of DPDP where residues can cluster (Figure 4-1a). The central residue of this cluster is Y10, and so mutations of this residue were made to explore what factors may be important in this position. Two mutations were considered: Y10V and Y10T. These mutations were chosen for several reasons. Both Val and Thr, like Tyr, prefer ß-sheet conformations according to statistical amino acid preferences[106, 110]. Val and Thr are different enough from Tyr structurally that changes in stability can be expected, and similar enough to each other that direct comparison is relatively simple; the change from Val to Thr amounts to exchanging a methyl group for a hydroxyl.

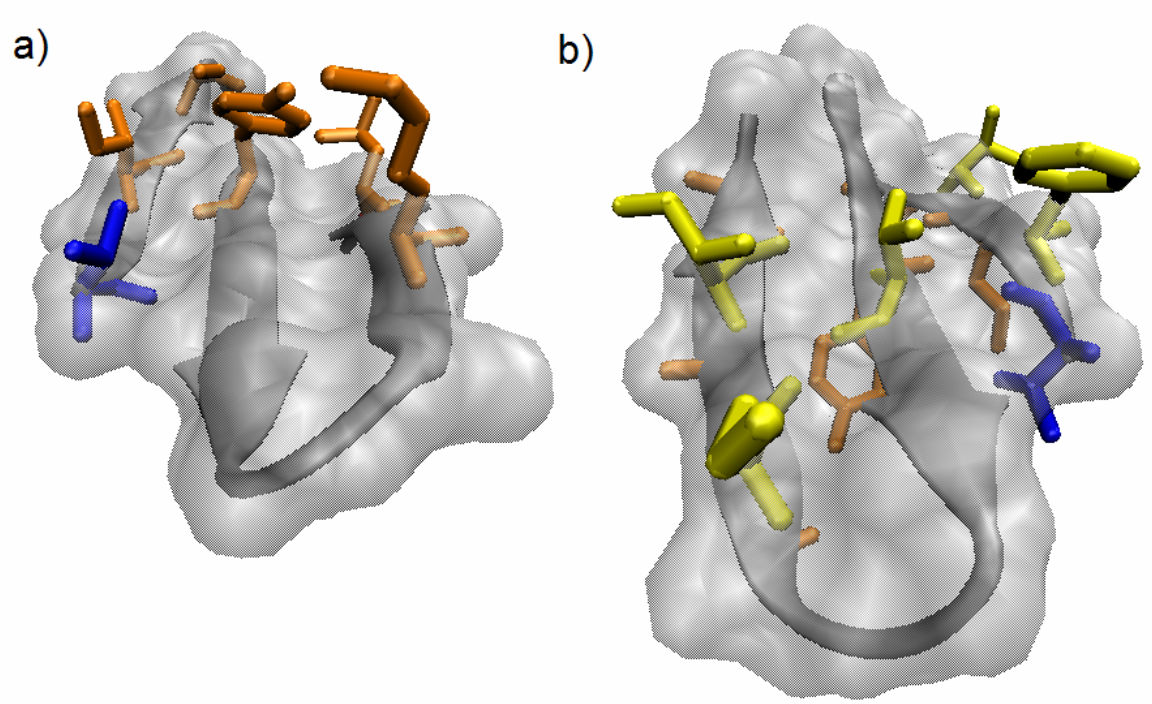

Figure 4-1. Twist and hydrophobic clustering in DPDP. Solvent Accessible Surface Area of the backbone calculated with VMD 1.83[3] shown in transparent grey, N-terminal acetyl group shown in blue. a) Hydrophobic cluster of residues (colored orange) inferred from NMR experiments. The twist of the sheet serves to bring these residues closer to each other. Tyr10, the central residue of this cluster is mutated to Val and Thr in Y10V and Y10T respectively. b) Potential hydrophobic cluster of residues (colored yellow) on the opposite face of DPDP. The central residue of this cluster is Thr11. Note Phe2 in the more solvent exposed position in strand 1. Thr11 and Phe2 are swapped in the FT mutant.

Two REMD runs of each mutant were performed starting from a folded structure based on the folded structure of DPDP and an extended linear structure. Mutants were run until reasonable convergence (comparable to wildtype DPDP) in melting curve plots was achieved. All other mutants were run in the same fashion. The total simulation times for each of these mutants (both from folded and linear structures) were Y10V=200ns, and Y10T=140ns. Since the definition of native contacts will vary from mutant to mutant, comparisons were made using X (fraction of backbone hydrogen bonds formed) instead of native contacts, as was done in Chapter 3.

Melting curves calculated from X for DPDP wildtype, Y10V, and Y10T are shown in Figure 4-2a. Y10T is less stable than DPDP, which is not surprising considering this mutation involves the loss of many side chain atoms. In addition, the shape of the curve is much more linear, indicating that the cooperativity observed in DPDP has been all but eliminated in Y10T. What is surprising is that the Y10V melting curve shows almost no change from the DPDP curve. The fact that mutation of Tyr to Val (net loss of 5 atoms and the aromatic ring) produces little change is interesting, but even more significantly the further change from Val to Thr (net change of two atoms) results in a comparatively drastic change in stability.

A plot of the average distribution of radius of gyration (RG) values for the DPDP, Y10V, and Y10T simulations is shown in Figure 4-2b. Each distribution shows two peaks

located at approximately RG=5.5 Å and RG=6.4 Å. However, it is clear that the residue cluster is on average much more compact in the DPDP and Y10V simulations than in the Y10T simulations. It had been seen previously in Figure 3-3a and Figure 3-3b in Chapter 3 that transition from the folded minimum to the intermediate minimum was accompanied by an increase in RG, *i.e.* disruption of the hydrophobic core. Essentially, mutation of one of the methyl groups of Val to a hydroxyl results in disruption of the hydrophobic core in DPDP, as it is unfavorable to bury the more polar hydroxyl group. This results in a lower overall stability.

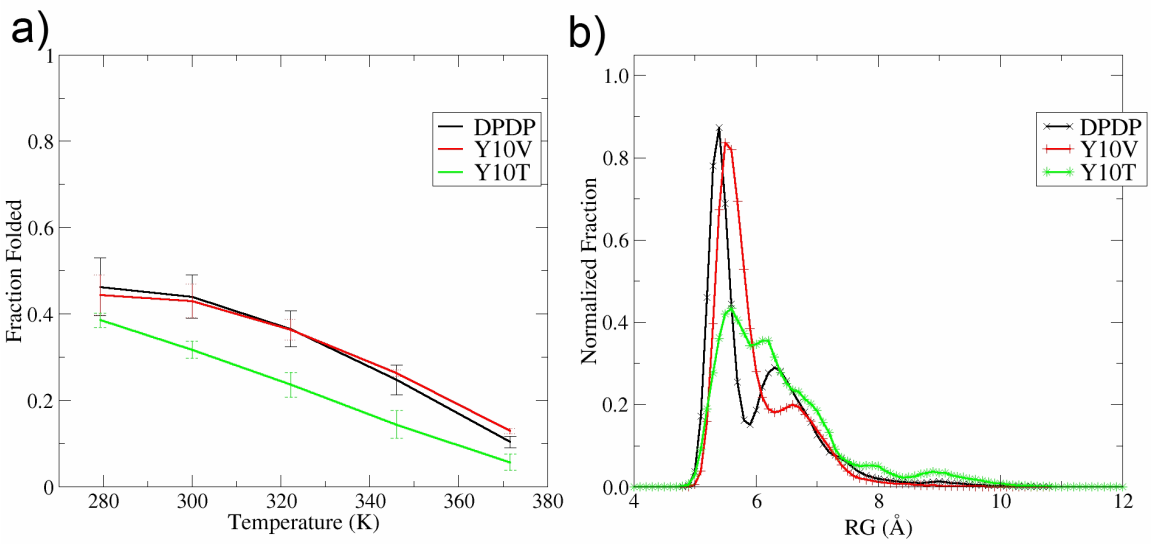

Figure 4-2. a) Melting curves calculated using fraction backbone hydrogen bonds present for DPDP, Y10V, and Y10V. Y10T is destabilized compared to DPDP, Y10V is relatively unchanged. b) Distribution of radius of gyration (RG) of the hydrophobic cluster (depicted in Figure 4-1a) for DPDP, Y10V, and Y10T. The hydrophobic core is disrupted far more often in the Y10T mutant compared with DPDP or Y10V, causing the lower stability seen in Figure 4-2a. Error bars represent half the difference between values obtained from two independent REMD simulations starting from different structures.

4.3.2 Hairpin 1 mutants

Since the previous simulations of native DPDP indicated that H1 was significantly less stable compared to the H2, several mutations were made to the sequence of H1 to try and improve stability and understand the differences between the two hairpins. Three mutants were considered: F2T/T11F (referred to as FT), S5V, and T4E/S5V/T9K (EVK). The total simulation times for the mutants are FT=140 ns, S5V=260 ns, and EVK=270 ns.

The first mutant, FT, was made with the idea of improving hydrophobic contacts in DPDP. Figure 4-1a shows the hydrophobic cluster on one side of DPDP thought to provide stability; simulations showed disruption of this cluster to be destabilizing. It was envisioned that residues F2, T4, T11, K16, and I18 on the other side of DPDP may be in a position to form an analogous hydrophobic cluster (referred to hereafter as the alternate hydrophobic cluster, Figure 4-1b). If these residues were to form a similar cluster, then T11 would be in the key position occupied by Tyr in the original cluster, and previous

simulations showed that Thr was destabilizing in such a position. Therefore a double mutant was made, F2T/T11F, essentially swapping the Phe and Thr, which are located across from each other in the hydrogen bond registry of DPDP. This mutation seemed sensible for several reasons. First, the mutation conserves the overall composition of DPDP because only the positions of residues are being swapped. Second, it places Phe in the position analogous to the one occupied by Tyr in the other cluster. Third, it moves the more hydrophobic Phe closer to the interior of the protein where it might have more solvent protection, and the more polar Thr to a more solvent exposed position in the first strand.

An examination of the melting curves for DPDP and FT (Figure 4-3) show that the FT mutant is overall more stable than DPDP. In addition, the stability of H1 is significantly increased, while the stability of H2 is relatively unchanged. This suggests that the majority of improvement occurs in H1 contacts. Since it was theorized that stability would be achieved by forming an alternate hydrophobic cluster of residues 2, 4, 11, 16, and 18, the radius of gyration of these residues was plotted for DPDP, the FT mutant, and the S5V mutant (Figure 4-4a). From this plot it is seen that the distribution of radius of gyration values for the alternate hydrophobic cluster of FT are shifted lower than the distributions for DPDP and S5V (a turn 1 mutant), indicating the alternate cluster is more compact in the FT mutant. For comparison, the distribution of the RG values for the original hydrophobic cluster are shown for the same systems in Figure 4-4b. Here it is seen that the distributions hover around the same value, indicating that the original cluster is not perturbed in these mutants.

Different turn types are known to have a significant affect on hairpin stability. The DPro/Gly sequence in DPDP has a strong propensity to form type I' and type II' reverse turns. Despite both turns having this sequence, each hairpin has a different stability. To explore how strongly the identity of the residues in the turn sequence affect stability, two mutants of the H1 turn region were made. The first, S5V, makes it so that the four residues in the i through i+3 positions are the same in both turns (VdPGK). The second, EVK, makes it so that 6 residues including the turn are the same (EVdPGKK). This mutation also creates the potential for a salt bridge between residues E4 and K9.

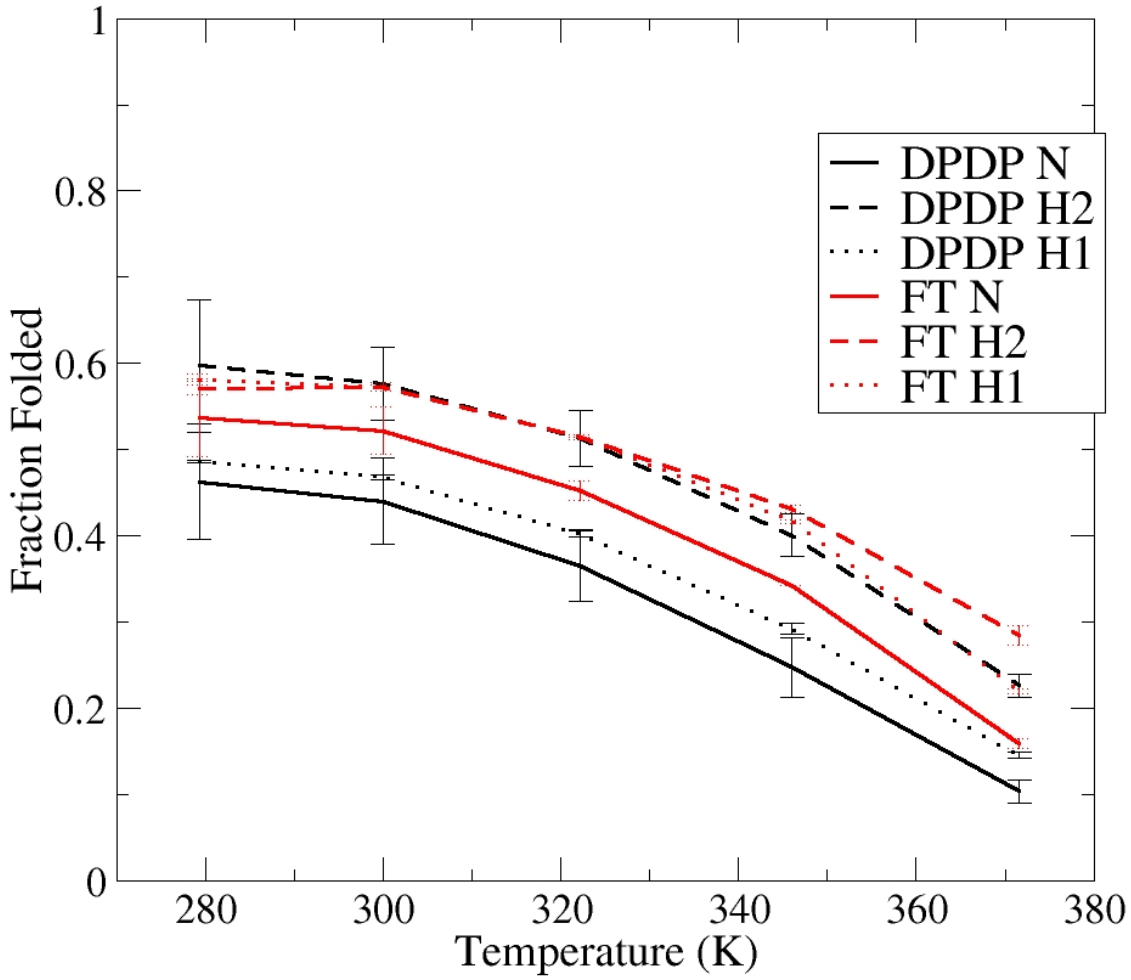

Figure 4-3. Melting curves for wildtype DPDP and the FT mutant. Melting curves are shown for overall, only Hairpin 1 (H1) and only Hairpin 2 (H2). In DPDP Hairpin 2 is more stable than Hairpin 1, but in the FT mutant the hairpins are approximately equal in stability, resulting in a net gain in stability for FT over wildtype DPDP.

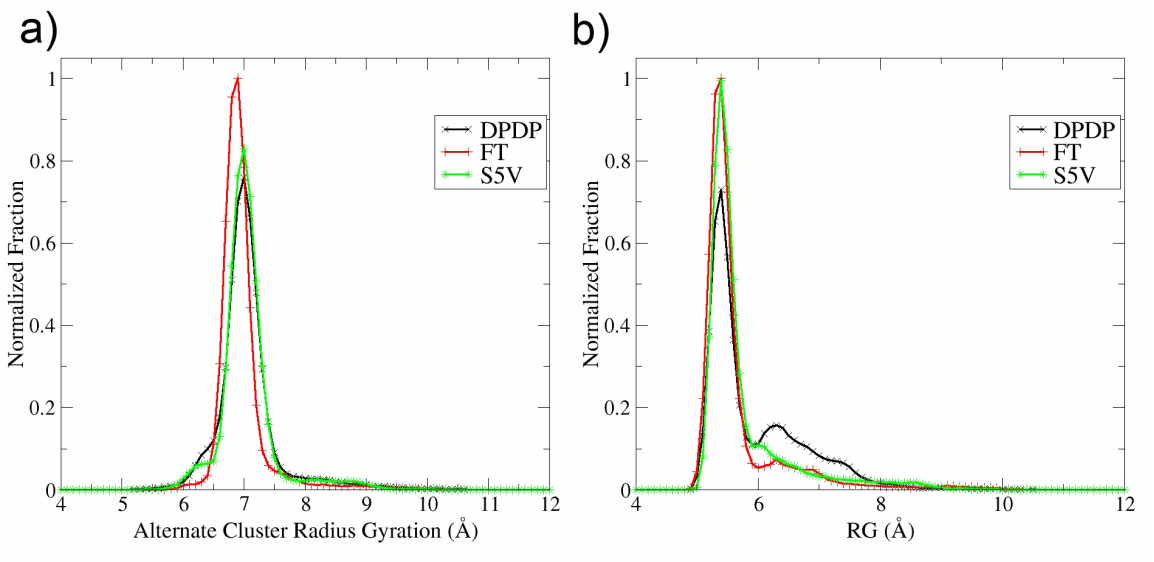

Figure 4-4. Radius of gyration of various hydrophobic clusters in DPDP, FT, and S5V. a) Radius of gyration of the 'alternate' cluster depicted in Figure 4-1b. The average value is shifted slightly lower in FT, indicating the alternate cluster is more compact in this mutant. b) Radius of gyration of hydrophobic cluster depicted in Figure 4-1a. The average value is similar for all systems, indicating that these mutations do not perturb this cluster.

Figure 4-5 shows melting curves for DPDP, EVK, and S5V. Both the EVK and S5V mutants are much more stable than DPDP. In addition, H2 stability is now approximately equal to H1 stability for the EVK and S5V simulations. Since the EVK and S5V mutations are overall comparable in stability, it can be inferred that most of the stabilization is occurring from the S5V mutation, while salt bridging between E4 and K9 provides only a slight increase in stability. Since the addition of the salt bridge appears to marginally affect stability it is likely that overstabilization of salt-bridging, known to be a problem in some simulations with GB solvation[72, 95, 111-113], is not a factor in this system, although ideally a T4E/T9K mutation would be available to observe the actual effects of adding a salt bridge in this position.

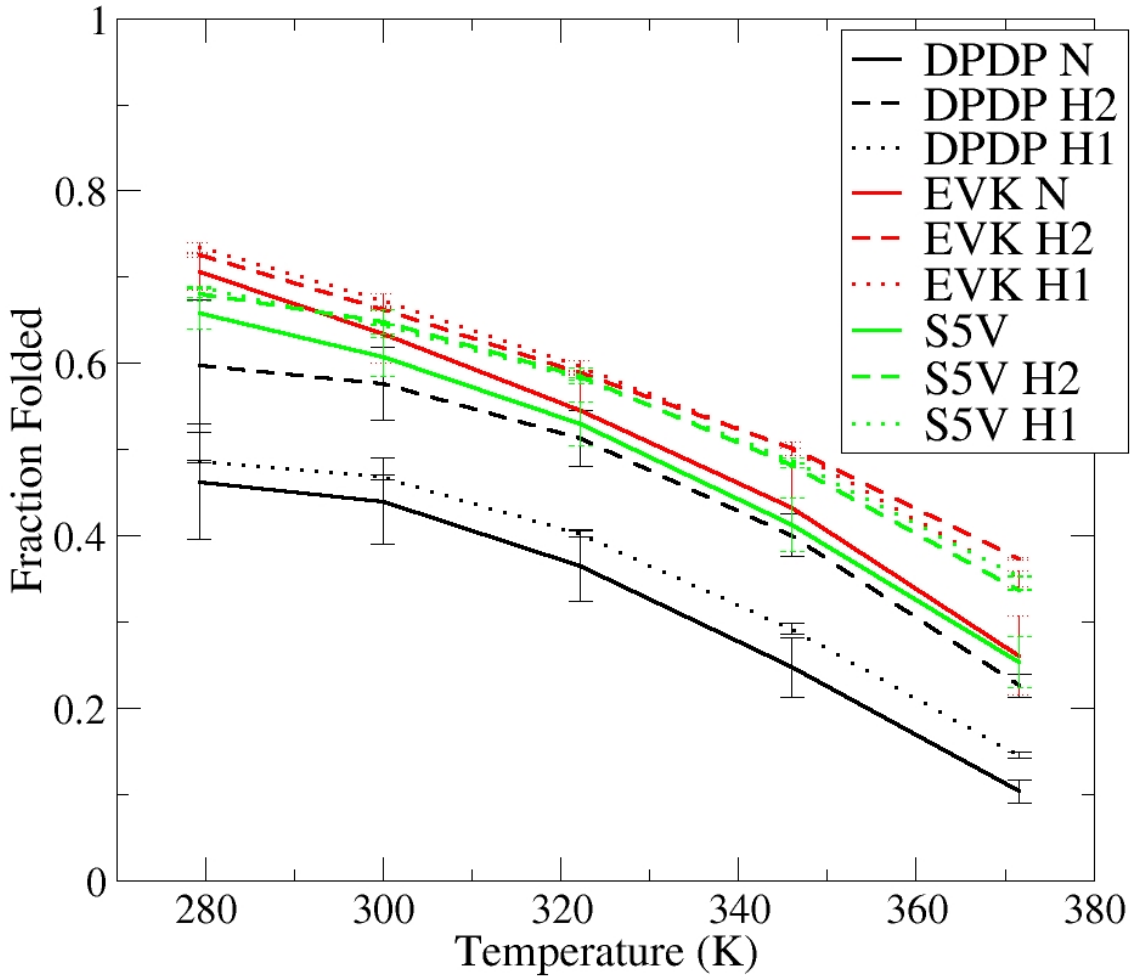

Figure 4-5. Melting curves for DPDP, EVK, and S5V. EVK and S5V are much more stable than DPDP. Since the stability of S5V and EVK is approximately equal, the majority of the stabilization comes from the S5V mutation. Note also how H1 and H2 stabilities are approximately equal in the EVK and S5V mutants.

Unlike previous mutations, there is no readily apparent reason why the S5V mutation should offer any stability benefits other than the fact that the same sequence is in H2 in DPDP. Figure 4-6 shows free energy landscapes as a function of f/? dihedral angles for residues 5, 4, and 3 in both DPDP and S5V. It is seen that the S5V mutation seems to drastically reduce the propensity of certain residues to access the left-handed helical region of the f /? plot, particularly residue 5. This effect is seen to be translated down to residues T4 and I3. It was thought that structures with residue 5 dihedrals in the left-handed helical regions accessible to Ser5 in DPDP but not to Val5 in S5V were perhaps responsible for the lower stability of DPDP.

Structures that fell within either of the two wildtype DPDP residue 5 free energy landscape minima (Figure 4-6) centered at f/?=130, 95 (the 'normal' ß-sheet

conformation) and f/?=40, 65 (the left-handed a-helical region) were analyzed. H1 from representative structures from these clusters are shown in Figure 4-7. It is immediately apparent that strand 1 is ‘kinked’ in the structures located in the left-handed a-helical region. The native backbone hydrogen bond between S5H and K8O has been disrupted and a non-native backbone hydrogen bond between T4O and K8H forms, stabilizing this ‘kinked’ conformation and forming a 3-residue bulge in place of the 2-residue turn. In DPDP it is known that mutation of DPro7 in turn 1 to Asp resulted in the formation of a 5 residue bulge, while the same mutation in turn 2 had no effect[102], indicating that perhaps the sequence of turn 1 is predisposed to form bulge-like structures.

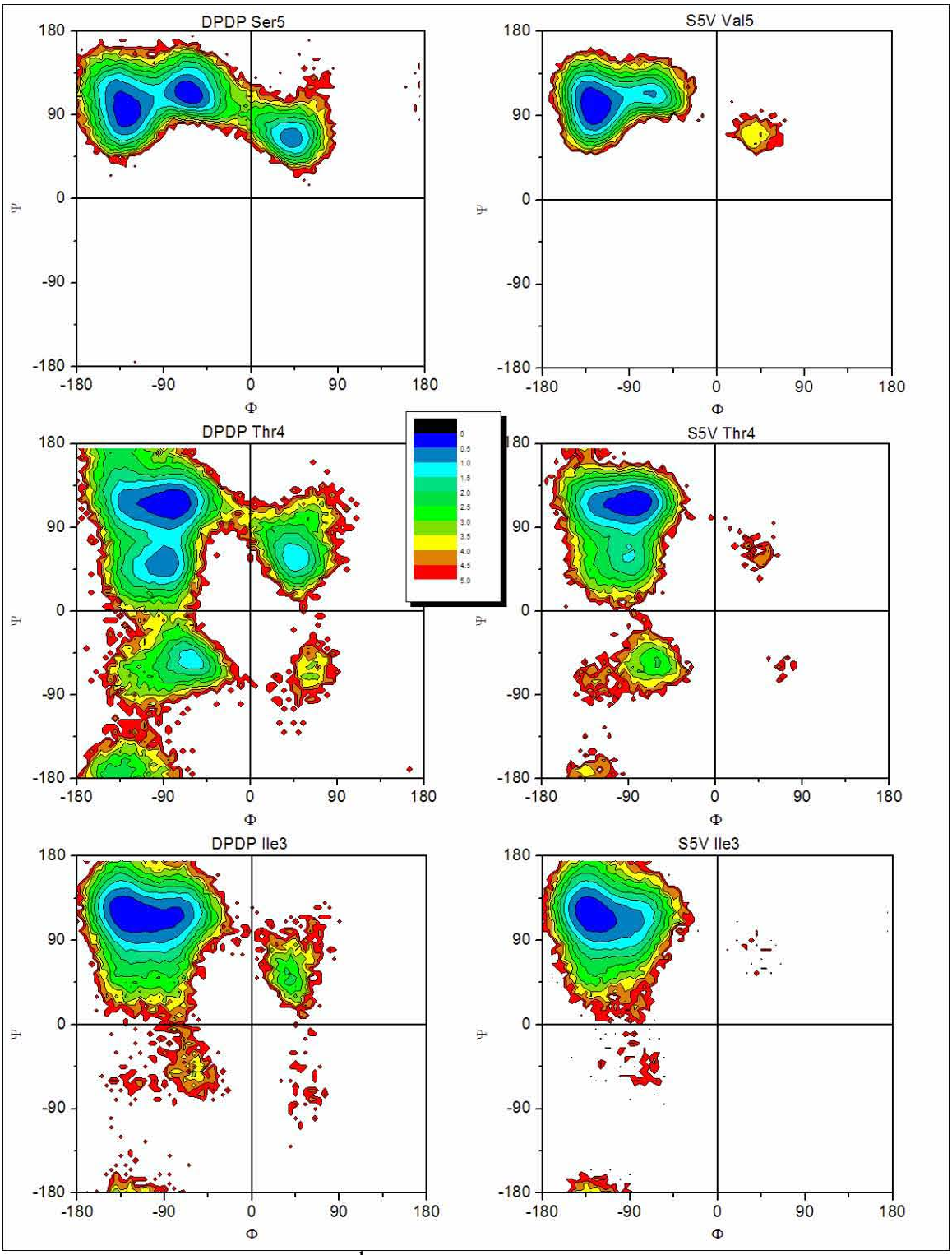

Figure 4-6. Free energy (in kcal mol^{-1}) Ramachandran plots for residues 3, 4, and 5 in DPDP (left column) and S5V (right column). Accessibility to the left-handed helix region of the Ramachandran space is drastically reduced in the S5V mutant.

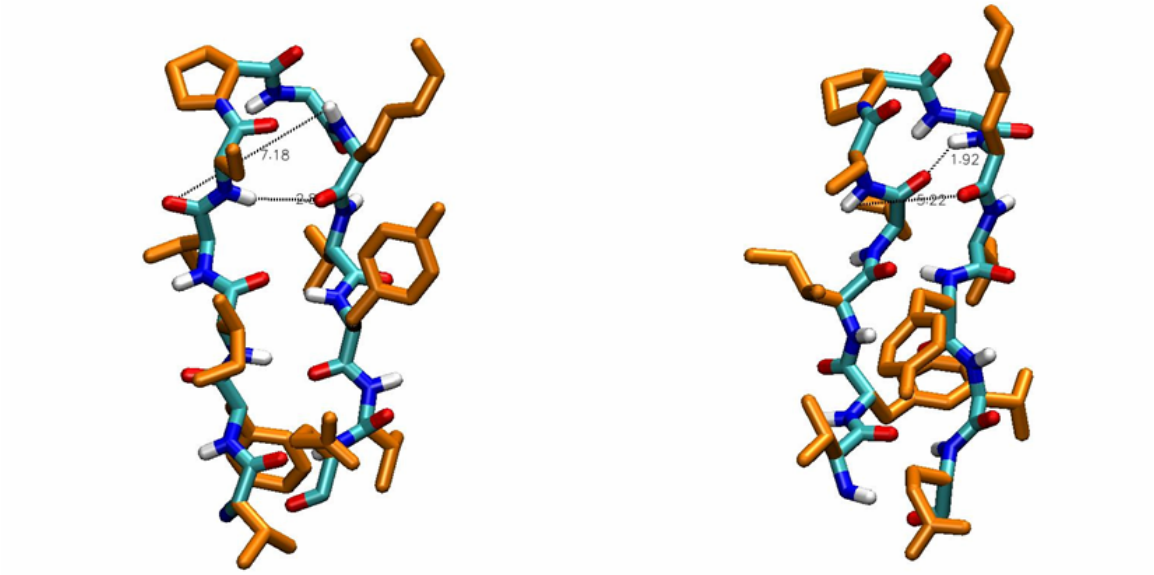

Figure 4-7. Hairpin 1 of DPDP in the normal ß-sheet conformation (left) and the 'kinked' conformation (right). The top dashed line is drawn between the atoms forming the non-native backbone hydrogen bond (T4O-K8H) and the bottom dashed line is drawn between the atoms forming a native backbone hydrogen bond (S5H-K8O).

The S5V mutation, by reducing residue 5 access to the left-handed helical region of the Ramachandran space, prevents formation of this non-native hydrogen bond and the 'kinked' structure. This effect can be seen most clearly in free energy plots using X_{Total} and the non-native hydrogen bond distance (T4O-K8H) as coordinates, shown in Figure 4-8. In DPDP there is a clear thermodynamic pathway in which the non-native hydrogen bond forms at around X_{Total}=0.30; there is only a small barrier to proceed to the unfolded state from this intermediate state. However, in S5V this pathway is all but eliminated. Therefore S5V derives its stability via elimination of an intermediate state that facilitates unfolding of DPDP.

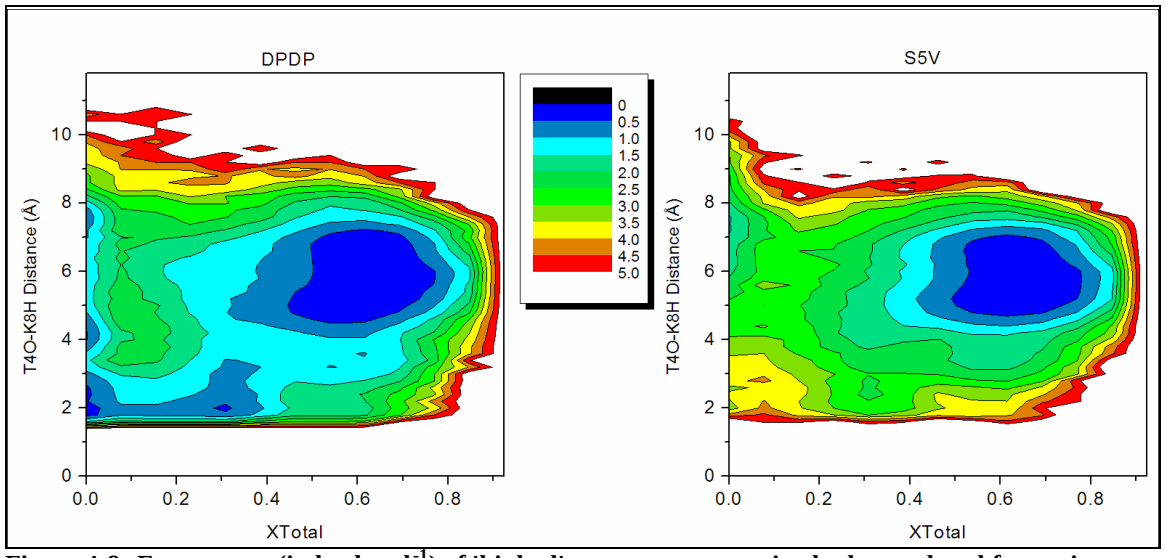

Figure 4-8. Free energy (in kcal mol^{-1}) of 'kinked' structure non-native hydrogen bond formation as a function of XTotal. The intermediate stabilized by formation of this non-native hydrogen bond at XTotal=0.3 in DPDP is eliminated in the S5V mutant.

4.3.3 FTV Mutant

Both the FT (F2T, T11F) and S5V mutations were more stable than the wildtype. A triple mutant was made, FTV (F2T, T11F, S5V) in an attempt to combine the stabilities gained from each set of mutations to produce an even more stable peptide. Melt curves for DPDP, FT, FTV, and S5V calculated from fraction backbone hydrogen bonds (cutoff XH1, XH2 > 0.5 ~ folded sheet) are shown in Figure 4-9. The overall stability of the FTV mutant falls between the FT and S5V mutants, indicating the stability of the 2 mutations does not add and likely compete with each other in some way.

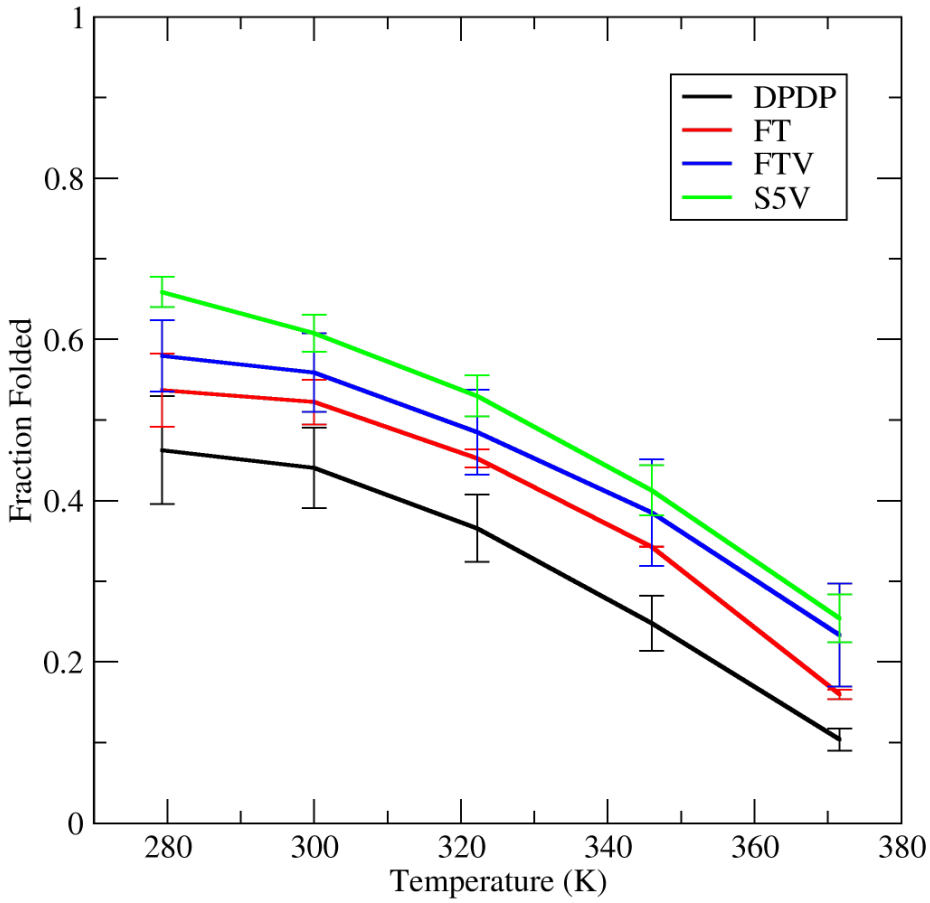

Figure 4-9. Melting curves for wildtype DPDP, the FT mutant, the S5V mutant, and the FTV mutant, which combines the FT and S5V mutations. The stability of FTV falls in-between that of FTV and S5V, suggesting these two mutations somehow compete with each other.

Figure 4-10a and Figure 4-10b show normalized histograms of the radius of gyration for the 'normal' hydrophobic cluster (residues I3, S/V5, Y10, K17, and L19) and 'alternate' hydrophobic cluster (residues F/T2, T4, T/F11, K16, and I18) respectively for wildtype DPDP and the FT, S5V, and FTV mutants. As expected, the alternate hydrophobic cluster is more compact in the FT and FTV mutants than in DPDP or S5V, although this cluster is overall not as compact as the normal hydrophobic cluster in any of the simulations. The position of the normal hydrophobic cluster peak does not change from the wildtype position for any of the mutants.

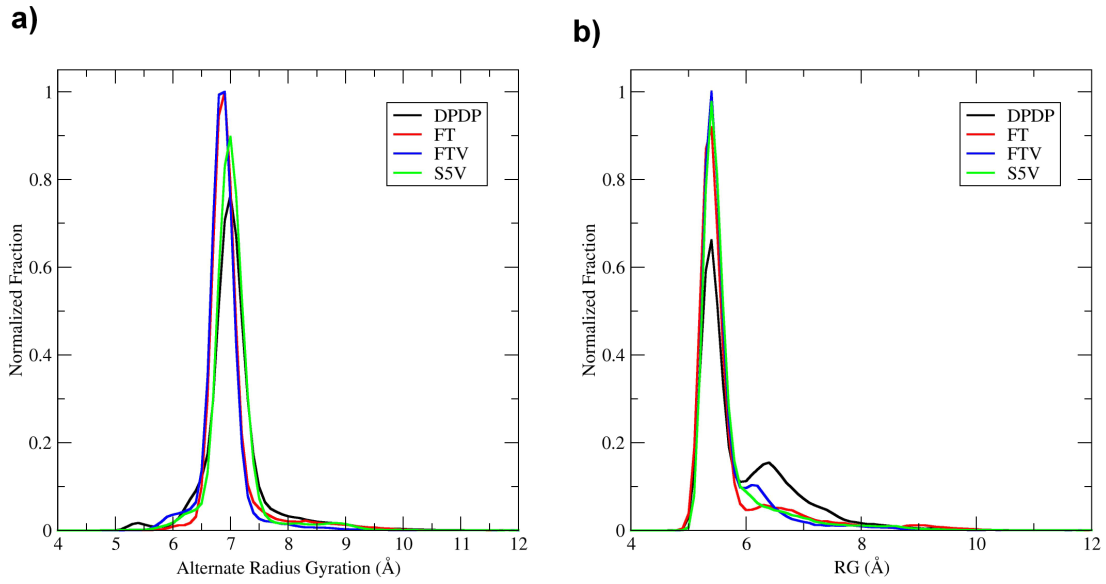

Figure 4-10. a) Normalized histogram of the radius of gyration of the hydrophobic cluster shown in Figure 4-1a for wildtype DPDP and the FT, S5V, and FTV mutants. The cluster in all three mutants is about as compact as it is in wildtype DPDP; it is also present more often in the mutants, reflecting the higher stability of the mutants. b) Normalized histogram of the radius of gyration of the alternate hydrophobic cluster shown in Figure 4-1b for wildtype DPDP and the FT, S5V, and FTV mutants. The FT and FTV mutations result in a slightly more compact cluster than wildtype DPDP or the S5V mutation. The alternate cluster is not as compact as the normal cluster.

Figure 4-11 shows the normalized distance histograms of the non-native backbone hydrogen bond that is formed in 'kinked' structures which seem to destabilize the native state by providing a favorable unfolding pathway. There is a peak at 2 Å for the DPDP and FT simulations, indicating that this bond is formed. This peak is gone from the S5V and FTV mutations, indicating formation of the kinked structure is no longer as favorable, as seen previously.

While the FTV mutant incorporates both the alternative hydrophobic core and elimination of the non-native backbone hydrogen bond that stabilized the FT and S5V mutants respectively, its overall stability is less than that of the S5V mutation alone. This seems to indicate that the FT mutation competes with the S5V mutation in some as yet unrevealed way. It also suggests that turn optimization is more important to overall stability of the hairpin that optimization of interactions between the strands, as has been noted previously[102].

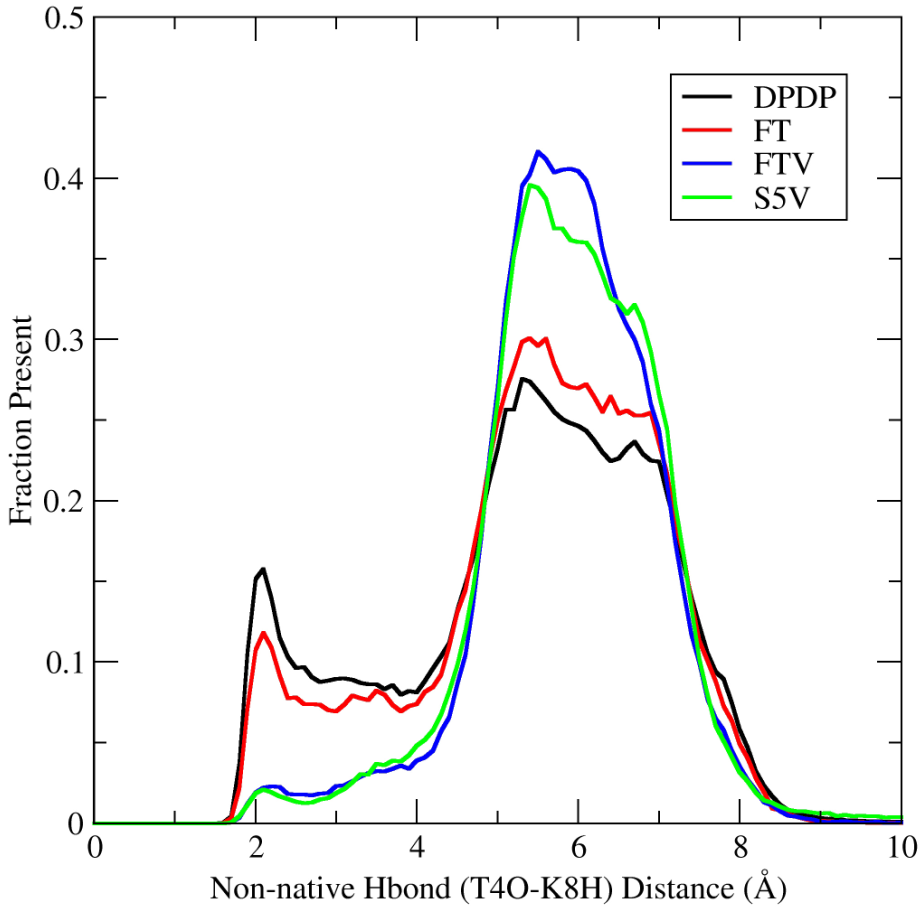

Figure 4-11. Distribution of the distance between the atoms which comprise the non-native hydrogen bond (shown in Figure 4-7). When formed, this non-native hydrogen bond can stabilize an unfolding intermediate and so destabilize the native state of DPDP. The wildtype and FT mutant both show formation of this hydrogen bond, while the S5V and FTV mutants do not, as expected.

4.4 Conclusions

Several mutations were made of various residues in DPDP, revealing underlying roles for certain positions. Mutation of the central residue of a hydrophobic core in DPDP (Tyr10) to a Thr disrupted the core and lowered stability, while a mutation to Val showed little effect, indicating it is important for residues in this position to form hydrophobic contacts, and the importance of a hydrophobic core to overall sheet stability in this model peptide.

A mutation designed to form a complimentary hydrophobic core on the opposite face of DPDP (the FT mutant) succeeded in increasing overall stability by stabilization of Hairpin 1. A second mutation of the *i* residue of turn 1 from Ser to Val increased stability significantly by eliminating an unfolding intermediate structure containing a non-native backbone hydrogen bond. This result indicates that Val may be preferable to Ser in the first position of a type I' or type II' reverse turn, and reinforces the idea that optimization of all residues in a turn is important to the stability of ß-hairpins. Addition of a salt-bridge to hairpin 1 (EVK) was found to have only a marginal effect on stabilizing hairpin 1. Also, the two mutations found to be most stabilizing (FT and S5V into FTV) were found to be non-additive; the FTV mutant actually showed stability in between that of FT and S5V, indicating these mutations compete, and reinforcing the idea that care must be taken when attempting to design in additional stability; optimization of the turn in ß-hairpin structures seems to be dominant over optimization of strand interactions. Intuitively this seems sensible, since a strong turn reduces the entropic cost of bringing two strands together.

It should be noted that although good convergence was achieved for all systems in this study, these results (specifically for the mutants) are likely sensitive to the choice of forcefield and solvent model (see also Conclusions in Chapter 3). For example, since the stability changes in the FT and Y10T mutations seem to result from hydrophobic effects, the accuracy of the solvent model will play a large role in the accuracy of the results. Ideally the mutant peptides in this study would be synthesized and the changes in stability would be studied via experimental methods in order to completely verify the results. Unfortunately, the magnitude of the stability changes shown here may not be detectable by current experimental methods for this specific system. Therefore it is hoped that perhaps the methods of optimization suggested here can be applied to larger systems where the effects might be more amenable to detection by experiment.

Chapter 5
Secondary Structure Bias in Generalized Born Solvent Models: Comparison of Conformational Ensembles and Free Energy of Solvent Polarization from Explicit and Implicit Solvation

5.1 Introduction

To correctly model protein behavior in an aqueous environment it is given that an accurate representation of solvent is necessary. In computational simulations of proteins it is common to either represent the solvent atoms explicitly or to estimate the solute response to bulk solvent using a dielectric continuum model, which is generally referred to as implicit solvation[114]. Although explicit solvent models are more realistic and physically rigorous[115], implicit solvent models have several features that make their use attractive. Not having to include solvent atoms can considerably reduce the size of a system, which can result in a significant decrease in the computational cost of a simulation. In addition, conformational sampling is increased from the lack of explicit solvent molecules in two ways: 1) there is no need to average over the extremely large number of solvent configurations in a simulation, 2) the lack of friction from solvent molecules effectively removes the viscosity of the solvent environment, accelerating molecular motions[61].

In an implicit solvent model, the overall free energy cost of solvating a solute molecule is typically decomposed into a non-polar component (?GNonpol) and a polar component (?GPol)[116]. ?GNonpol is the free energy cost of rearranging the solvent to accommodate an uncharged solute molecule of arbitrary shape, and ?GPol is the free energy cost of solvent polarization due to solute charges. The most accurate method for calculating ?GPol in a continuum dielectric environment (neglecting salt effects) is solving the Poisson Equation (PE)[117]. However, this method is not easily incorporated into molecular dynamics (MD) simulations due to computational expense. Despite the recent advances that have been made in using implicit solvent models based on PE in MD simulations[118-120], this calculation remains highly computationally demanding[121]. In light of this, another method of calculating ?GPol is often used: the generalized Born (GB) implicit solvent model[122]. GB is based on PE but contains several approximations which increase the speed of the calculation. As a result, the GB model has become quite popular in computational simulations[123].

However, this increase in speed comes at the cost of accuracy. Although the GB model has been shown to give solvation free energies in agreement with experiment for small molecules[122, 124], there has been some question as to the performance of this model for simulations of larger biomolecules. Grycuk has shown that significant errors arise in GB calculations due to the Coulomb-field approximation[125]. Several studies[72, 95, 111-113] have also shown that GB models tend to over-stabilize ion pair interactions, which can lead to the trapping of molecules in (and thus over-population of) non-native states. There have been several reports suggesting that certain GB models tend

to over-stabilize a-helical conformations[71-74], although the exact cause for this remains unclear. In addition, it has been shown for several biological macromolecules that accuracy of GB often results from widespread cancellation of errors[126, 127].

Due to these issues it is desirable to quantitatively compare ensemble properties from simulations with implicit and explicit solvent models. However, this kind of direct comparison can be difficult since explicit solvent simulations require a greater length of time to converge than implicit solvent simulations due to considerably slower conformational sampling for flexible solutes. Recently, the development of enhanced sampling techniques such as Parallel Tempering[35] or Replica Exchange Molecular Dynamics (REMD)[36] have provided a means to bridge the sampling gap between implicit and explicit solvent simulations.

In this study we assess the performance of three GB implicit solvent models implemented in Amber[65] as compared to the TIP3P explicit solvent model and the PE implicit solvent model. Our test peptide is alanine decapeptide (Ala10, Ace-A10-NH2). We chose this model system to compare explicit and implicit solvent models as there are no potential salt bridges, eliminating formation of these as an issue. Ala10 is also long enough to form more than one or more repeats of basic secondary structure types found in larger proteins, such as helices and ß-hairpins. We recently reported extensive simulations of this peptide in water[74].

We compare ensembles of structures from well-converged REMD simulations of Ala10 using either the TIP3P explicit solvent or three variations of the GB implicit solvent model implemented in Amber[65]. It is shown that in simulations of Ala10 with the TIP3P solvent model, residues predominantly adopt a polyproline II (PP2) conformation, in agreement with various experimental observations of short Alanine-rich peptides (see discussion in Ref. 31). However, it is then shown that the conformational preferences of Ala10 are altered in simulations with GB solvent models; in particular, certain GB models appear to strongly foster the formation of a-helical conformations. The results suggest that these models may have serious limitations when one wants to quantitatively investigate the conformational preferences of peptides and proteins.

To explain these observed differences between explicit and implicit simulations, we first directly compare explicit solvent ?GPol values obtained from Thermodynamic Integration (TI) calculations to ?GPol values from PE and GB implicit solvent models for four basic secondary structure types: right-handed a-helix, left-handed a-helix, ß-hairpin, and polyproline II helix. In particular, we focus on comparing the difference in the electrostatic component of the solvation free energy between these conformations (??GPol), and how this relates to the ensembles of structures observed in the REMD simulations. In particular, we show that the observed a-helical bias in certain GB models results from overestimation of ??GPol for a-helical structures. We also show that in terms of reproducing TIP3P ??GPol values, the PE implicit solvent model has the best performance overall.

Given that the PE implicit model has the best performance, we then compare effective Born radii calculated with GB to 'perfect' effective Born radii calculated with PE, and show that there are large discrepancies, especially for backbone atoms. It is shown that use of 'Perfect' effect Born radii improves the accuracy of the Self and Interaction terms of the GB energy calculation with respect to PE results (as has been reported previously[126]). However, it is also shown that in terms of reproducing TIP3P

??GPol values, a GB model with 'Perfect' effect Born radii does not approach the performance of the PE model, and indeed does not provide an appreciable improvement over any of the other GB models studied here. This suggests that there is a limit to how far radii optimization alone can improve the GB solvent model.

5.2 Methods

5.2.1 REMD Simulation Details

The peptide simulated is Ala10 (Ace-A10-NH2) in TIP3P[128] and several variations of the GB implicit solvent model; GBHCT[60], GBOBC[129], and GBNeck[130] (igb = 1, 5, and 7 respectively in Amber 9). A variant of GBOBC with different a, ß, and ? parameters (discussed below) was also used (igb = 2 in Amber9). In the text, GBOBC will be used to refer to results with igb = 5, and results from GBOBC with igb = 2 parameters will be specifically noted using the igb value. For TIP3P REMD simulations, Ala10 was solvated in a truncated octahedral box with 983 solvent molecules. Amber 9[65] was used with the ff99SB force field[131] for all REMD simulations. It has been recently shown that this force field performs quite well and is able to reproduce NMR observables for ubiquitin in TIP3P water[132]. For consistency, MBondi2 radii[129] were used in both the GB REMD simulations and subsequent GB and PE energy calculations described below.

For each solvent model, two separate REMD simulations of Ala10 were run starting from different initial conformations: an extended conformation and a collapsed conformation. The distribution of temperatures was chosen to ensure good overlap of potential energy between replicas and to achieve an exchange acceptance ratio of 0.20. The TIP3P REMD simulations involved 40 replicas at temperatures ranging from 266.9 to 571.2 K. Since the GB REMD simulations had far fewer degrees of freedom, only 8 replicas were required at temperatures ranging from 269.5 to 570.9 K. All data analysis was performed on REMD structure ensembles at 300.0 K. The high degree of convergence of these ensembles has been demonstrated in a previously published study[74].

Bonds to hydrogen atoms were constrained with the SHAKE[59] algorithm using a geometrical tolerance of 0.000001 Å. The non-bonded interaction cutoff was 7.0 Å for the TIP3P simulations, and 99.0 Å (effectively infinite) for the GB simulations. The TIP3P simulations were run in the nVT ensemble, long range electrostatic interactions were calculated using periodic boundary conditions via the particle mesh Ewald (PME) summation[133], and the non-bonded list was updated every 20 steps. Simulations were run with a time-step of 2 fs, with exchange attempts occurring every 1 ps. Both explicit and implicit solvent simulations employed a weak temperature coupling algorithm[58] with a time constant of 0.1 ps.

5.2.2 Solvent Model Descriptions

The following are brief descriptions of the GB implicit solvent models used in this work (GBHCT, GBOBC, and GBNeck), and the implicit model based on PE. The

basic premise of these implicit models is that the effect of the solvent surrounding a molecule can be reproduced by a continuum dielectric field.

In the case of the GB solvent models, the idea can be traced back to the original Born equation for calculating ?GPol of a spherical ion of a certain size and internal dielectric value with a point charge at its center in an external dielectric medium (essentially Equation 5-6). This equation states that ?GPol will be proportional to the charge of the ion and inversely proportional to the size of the ion. So the larger the radius of an ion, the more the charge of that ion is blocked from the screening effect of the external dielectric, *i.e.* the solvent. In GB, this idea is extended to the case where there are multiple 'ions' (*i.e.* atoms with point charges) clustered together (such as is the case in a molecule), and attempts to take into account the descreening affects from the additional atoms by assigning them 'effective' radii. For example, the effective radius of an atom completely surrounded by solvent is simply equal to its intrinsic radius, but the effective radius of an atom at the center of a spherical cluster of atoms will be approximately equal to the radius of that cluster of atoms. Calculation of the effective radius is the key to the GB models studied here and is covered in more detail in the next section.

Implicit models based on PE are more rigorous than GB. Instead of calculating ?GPol based solely on the positions of the atoms, a 3-dimensional grid is set up that encompasses not only the atoms of interest but regions of solvent as well. At each grid point charge is defined (by distributing the atomic point charges near the grid points in some fashion) and the electrostatic potential is calculated via a finite difference method. Lines connecting each grid point are associated with a dielectric constant, which is either the solute (internal) or solvent (external) dielectric constant based on the location of the dielectric boundary (normally defined by the molecular surface). Because charges are discretized onto a grid, the calculation of the electrostatic potential and therefore ?GPol can be quite sensitive to the grid spacing, and so care must be taken when choosing the size of the grid.

5.2.3 Solvent Model Details

Each GB model used in this study has the same basic formulation. For a given solute (neglecting salt effects), the GB model calculates the electrostatic contribution to the solvation free energy between all atoms in the solute using Equation 5-1.

$$\Delta G_{Pol} = -\frac{1}{2}\left(e_{in}^{-1} - e_{out}^{-1}\right)\sum_{i,j}\frac{q_i q_j}{f_{GB}}$$

Equation 5-1. Generalized Born Equation.

In Equation 5-1 e_{in} and e_{out} are the dielectric constants inside and outside the solute respectively, q_i and q_j are partial atomic charges on atoms i and j, and $\mathbf{f_{GB}}$ is a function that modifies the strength of the charge interaction based on the screening of the charges by other atoms and the solvent. It is common (although other forms have been used[126, 134]) to calculate $\mathbf{f_{GB}}$ using Equation 5-2.

$$f_{GB} = \sqrt{r_{ij}^{\,2} + R_i R_j \exp\left(\frac{-r_{ij}^2}{4R_i R_j}\right)}$$

Equation 5-2. Form of f_{GB} commonly used in Equation 5-1.

In Equation 5-2 r_{ij} is the distance between atoms *i* and *j*, and R_i and R_j are the effective Born radii of atoms *i* and *j*[122]. The effective Born radius (hereafter referred to as RGB) of an atom reflects the effect of solvent dielectric on the atom charge; the more surrounded an atom is by high-dielectric solvent, the more its charge is screened and the smaller its RGB becomes.

The main difference in the three GB models studied here is in the calculation of RGB. The GBHCT model calculates RGB for each atom using Equation 5-3.

$$R_i^{-1} = r_i^{-1} - I$$

Equation 5-3. Effective Born radius calculation.

In Equation 5-3 $?_i$ is the intrinsic Born radius of atom *i*, and ***I*** is calculated using Equation 5-4.

$$I = \frac{1}{4p} \int_{VDW} q(|\vec{r}| - r_i) \frac{1}{r^4} d^3\vec{r}$$

Equation 5-4. Effective Born radius integral.

Equation 5-4 modifies the intrinsic radius of the atom based on the amount of screening from all other atoms[129]; for a single ion RGB is equal to the intrinsic radius. The integral is calculated over the van der Waals (VDW) radii of those atoms, essentially defining the dielectric boundary as a VDW surface (as opposed to the molecular surface used in solutions to PE[135]). As it is implemented in Amber, the above integral is solved in an analytical and pair-wise way, the exact form of which is given by Hawkins *et al.*[60]. Another functionally identical solution to this integral has been given by Schaeffer & Froemmel[136].

It was shown that the above formulation would give RGB values that were too small for deeply buried atoms[127, 137] due to regions of high dielectric created when the VDW radii of spheres do not overlap inside a molecule, even if the region is inaccessible to solvent. To compensate for this, the GBOBC model introduced a correction to the RGB calculation, Equation 5-5.

$$R_i^{-1} = r_i^{-1} - r_i^{-1} \tanh(a\Psi - b\Psi^2 + g\Psi^3)$$

Equation 5-5. GBOBC effective Born radius adjustment.

In Equation 5-5 ? = I?$_i$, and a, ß, and ? are adjustable empirical parameters[129]. This was designed to increase RGB for buried atoms, while leaving RGB for atoms near the surface relatively unchanged.

Although the GBOBC model compensated for the underestimation of RGB for buried atoms, there remained the possibility that because of the VDW surface representation, regions of high dielectric (or 'Neck' regions) that should be inaccessible

to water could develop between surface atoms, such as atoms in a hydrogen-bonding pair. The GBNeck model was designed to correct for these 'Neck' regions, and in doing so bring the VDW surface calculated in Equation 5-4 more in line with the molecular surface used in PE calculations. This correction is in addition to the one in Equation 5-5, and is applied during the calculation of the integral in Equation 5-4[130].

In order to obtain effective Born radii from the PE model, a method similar to one used by Onufriev *et al.*[126] is used. Equation 5-1, the generalized Born equation, can be separated into Self (i=j) and Interaction (i?j) terms. From Equation 5-1 and Equation 5-2 the Self solvation free energy for atom i, ?Gi, becomes Equation 5-6.

$$\Delta G_i = -\frac{1}{2}\left(e_{in}^{-1} - e_{out}^{-1}\right)\frac{q_i^2}{R_i}$$

Equation 5-6. Atomic self-solvation free energy as related to effective Born radius.

By setting all atomic charges to zero except the charge on atom *i*, $?G_i$ can be solved using PE, from which R_i is easily obtained. Effective Born radii obtained in this fashion will be referred to hereafter as RPE.

All PE calculations were performed with DelPhi version 2.0[135] using a grid spacing of 0.25 Å and an internal relative dielectric of 1.0. The grid spacing of 0.25 Å was found to provide the best balance of speed and accuracy, as smaller grid spacing did not result in significant improvement in calculated energies. Calculations of structures used an external relative dielectric of 78.5 to be consistent with Amber GB models. Calculations of effective Born radii with PE used an external relative dielectric of 1000.0 (effectively infinite) for consistency with standard GB effective radii calculations, as suggested by Sigalov *et al.*[138]. A percent fill value of 80% was used.

5.2.4 Thermodynamic Integration Calculations

Thermodynamic Integration (TI) calculations were performed with Amber in order to obtain ?GPol values for Ala10 in explicit TIP3P solvent. State 0 had all solute atomic charges off, and state 1 had all solute atomic charges on. Calculations were performed on four different conformations of Ala10: a-helix (Alpha), left-handed a-helix (Left), polyproline II helix (PP2), and ß-hairpin (Hairpin). The Alpha, Left, and PP2 conformations were generated with the Leap module of Amber. All f/? dihedrals in these conformations were set to 'idealized' values: Alpha = - 57.8°/-47.0°, Left = 57.8°/47.0°, PP2 = - 75.0°/145°. The Hairpin conformation was generated from the backbone of the ß-hairpin peptide Trpzip2[56] (PDB ID 1LE1). Figure 5-1 shows cartoon representations of these four conformations.

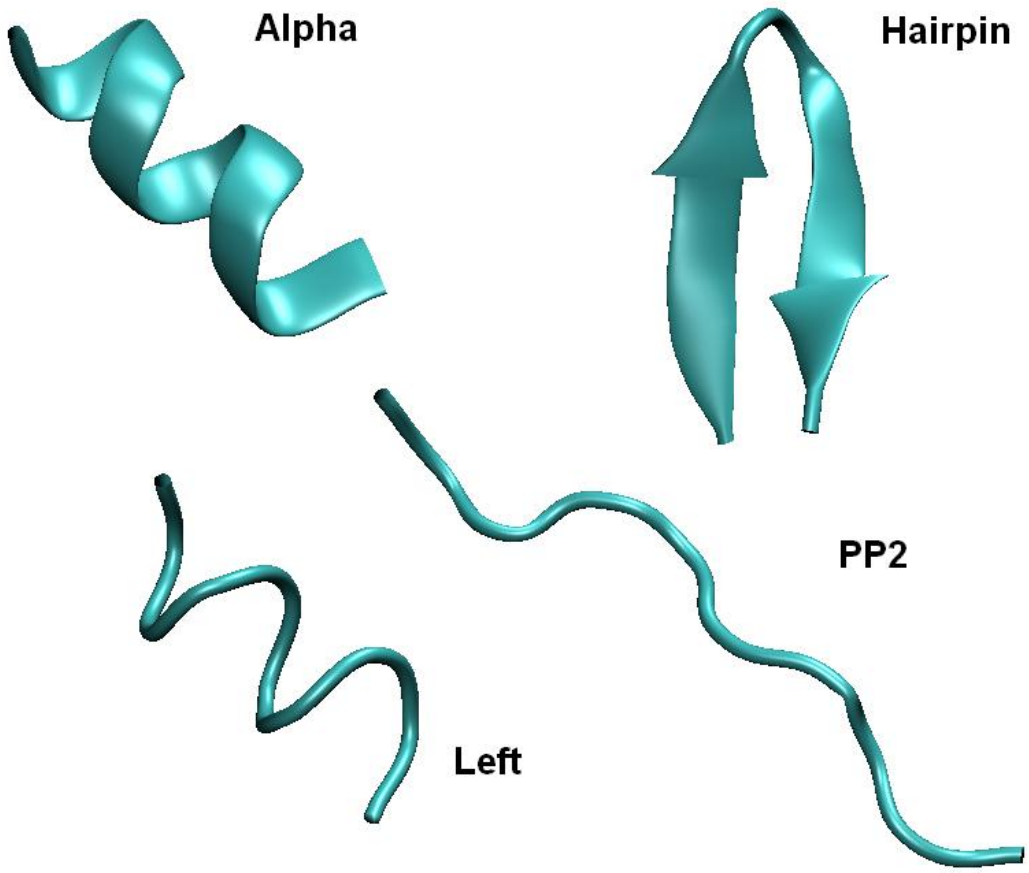

Figure 5-1. Four representative conformations of Ala10 used for TI calculations, shown in a 'Cartoon' style. Picture generated with VMD 1.8.4 [3].

There are two main considerations in these calculations. One is that over the course of the TI calculation the solute may change conformation, which is not desirable since only ?GPol values for specific conformations are desired. This was dealt with by applying simple positional restraints on all atoms to hold the molecule in the desired conformation. Another consideration is that when the charges in the solute are switched on, there are not only solvent-solute charge interactions but intra-solute charge interactions. This requires that two separate TI calculations be done; one in which the molecule is solvated, and one in which the molecule is in the gas phase. Subtracting the free energy values then not only cancels out the intra-solute charge interactions, but the restraint energies as well.

All conformations were solvated with the same number of TIP3P waters as in the REMD simulations, energy minimized, and TI calculations were run for 0.2, 1.0, or 2.0 ns with 5 or 7 ? values in order to test the sensitivity of the results to TI parameters. Conformations were preserved in TI calculations by use of 10 kcal mol^{-1} harmonical Cartesian coordinate restraints on all atoms. Final TI values were obtained from Gaussian integration over all ? values, excluding the first 50 ps of data from each ? value as equilibration.

5.2.5 Secondary Structure and Conformational Analysis

Secondary structure values were calculated using DSSP[139] as implemented in the Ptraj module of Amber, which uses patterns of hydrogen bonding to differentiate between different types of secondary structure. In addition, residues were assigned local conformational preferences (Alpha, Left, PP2, Extended) based on their f /? dihedral angle statistics calculated over the REMD trajectories. A residue is considered in the given conformation if it falls within ±30° of the following f /? values, chosen based on approximate boundaries of the free energy basins sampled in the explicit solvent REMD simulation of Ala10: Alpha (- 70°/-25°), Left (50°/30°), PP2 (- 70°/150°), or Extended (- 150°/155°).

5.3 Results

5.3.1 Secondary Structure and Local Conformational Propensities

Figure 5-2 shows secondary structure and local backbone conformational propensities calculated from backbone dihedral angles (see Methods for details) at 300.0 K for all residues of Ala10 calculated from unrestrained REMD simulations conducted using either the TIP3P, GBHCT, GBOBC, or GBNeck solvent model. Local conformational propensity differs from secondary structure propensity in that it is not dependent on the conformation of neighboring residues; for example a particular residue may be in a helical conformation and yet not be part of any regular helical structure (perhaps its neighbors are in a PP2 conformation). The average secondary structure propensities and local conformational preferences of all residues in each simulation are given in Table 5-1. The overall agreement between independent simulations for each solvent model (as indicated by the small error values) shows that good convergence was achieved for all simulations; excellent convergence for these ensembles has been reported previously[74].

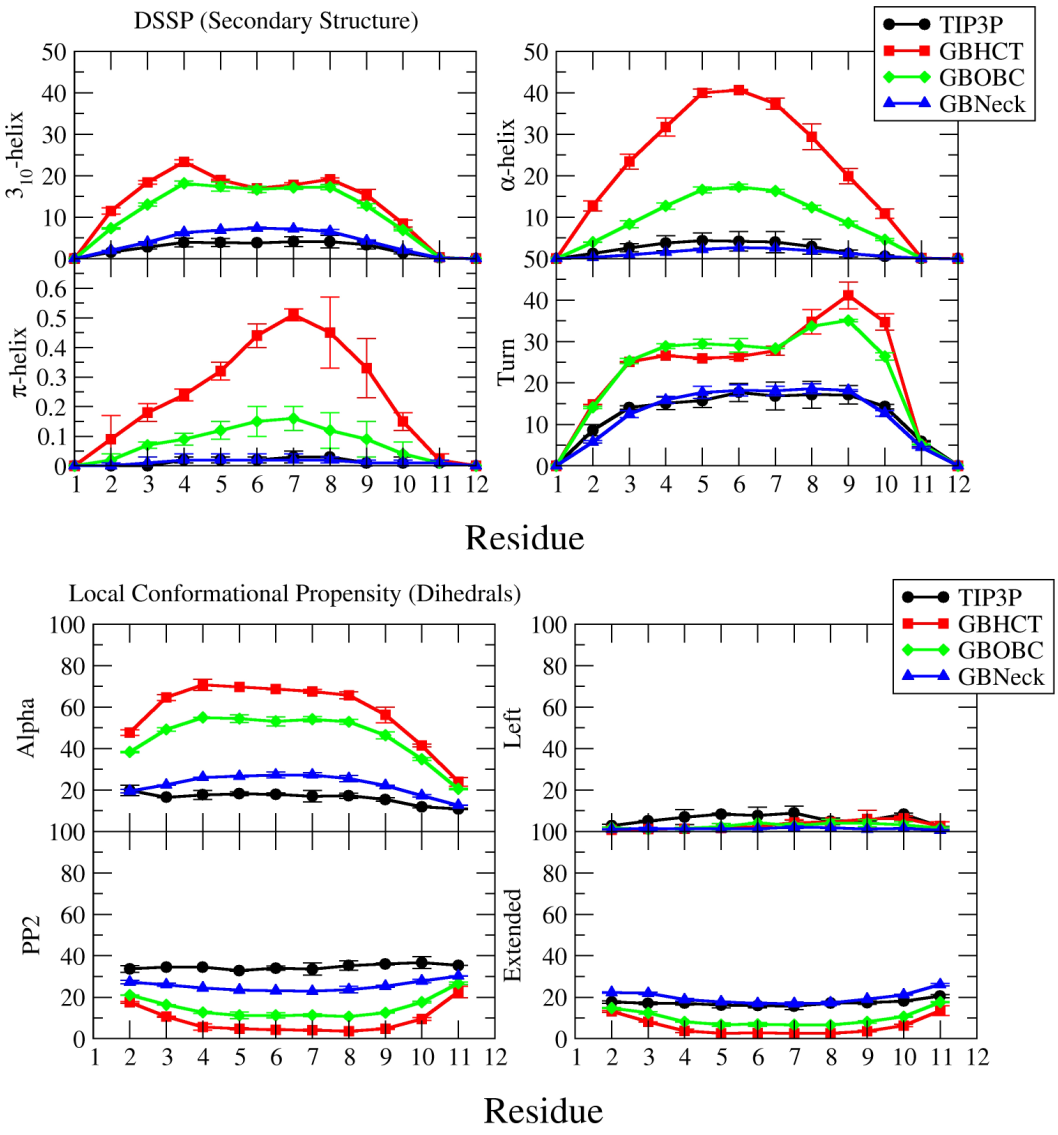

Figure 5-2. Secondary structure and local conformational propensities for each residue of Ala10 from unrestrained REMD simulations using various solvent models at 300.0 K. Residues 1 and 12 are the acetyl and amide N- and C-caps respectively. Error bars are calculated as half the difference of values reported from two independent simulations with the given solvent model, using different initial coordinates.

A) DSSP (Secondary Structure)				
	TIP3P	GBHCT	GBOBC	GBNeck
310-Helix	2.89 ± 0.06	15.01 ± 0.08	12.66 ± 0.07	4.64 ± 0.09
a-Helix	2.45 ± 0.63	24.60 ± 0.06	10.06 ± 0.08	1.37 ± 0.01
p-Helix	0.01 ± 0.01	0.27 ± 0.02	0.09 ± 0.02	0.01 ± 0.01
Turn	14.26 ± 0.18	26.19 ± 0.29	25.54 ± 0.09	14.21 ± 0.30
B) Local Conformational Propensity (Backbone Dihedrals)				
	TIP3P	GBHCT	GBOBC	GBNeck
Alpha	16.20 ± 0.33	57.57 ± 0.20	45.85 ± 0.20	22.63 ± 0.15
Left	6.00 ± 0.28	3.06 ± 0.16	2.58 ± 0.03	1.29 ± 0.04
PP2	34.65 ± 0.29	8.73 ± 0.01	15.14 ± 0.09	25.45 ± 0.04
Extended	17.61 ± 0.38	5.91 ± 0.08	9.87 ± 0.10	19.83 ± 0.15

Table 5-1. A) Average percent secondary structure and B) local conformational propensities from Ala10 REMD simulations. Secondary structure was calculated using DSSP[139] as implemented in Ptraj, and local conformational propensity was calculated based on dihedral angle cutoffs.

5.3.1.1 Explicit Solvent Simulations

The unrestrained REMD simulations of Ala10 with the TIP3P solvent model give results that are consistent with several recent theoretical and experimental studies of related polyalanine peptides. On average, Ala10 residues in the TIP3P simulation are predominantly in the PP2 conformation, consistent with free energy calculations done by Mezei *et al.*[140]. The average amount of PP2 observed (34.65±0.29%) is in reasonable agreement with values obtained for a similar polyalanine peptide XAO (Ace-X2A7O2-NH2, X=diaminobutyrate, O=ornithine), from both a previous explicit solvent computational study (42-47%[141]) and from experiment (40±8%[142]). Amide hydrogen atoms are involved in intramolecular hydrogen bonds for about 10% of the TIP3P simulation (data not shown), in close agreement with the value obtained from NMR data by Scheraga *et al.* (9%[143]) for XAO.

The predominant secondary structure type identified by DSSP for the TIP3P simulation is Turn, indicating that any inter-residue hydrogen bonds that form tend to be in no specific pattern. Although there is a tendency for residues to adopt an Alpha conformation locally (16.20±0.33%), there is almost no a-helical or 310-helical structure (5.34±0.63% total). There is a similar tendency for residues to adopt Extended conformations locally (17.61±0.38%), but little parallel or anti-parallel ß-sheet structure formation (1.54±0.44% total). Residues very rarely adopt the Left conformation locally, consistent with the fact that this conformation is sterically hindered.

5.3.1.2 Implicit Solvent Simulations

In the unrestrained REMD simulations with the GBHCT and GBOBC solvent models there is clearly much greater preference for residues to be in the Alpha conformation locally compared to the TIP3P simulation; the GBHCT simulation in particular contains about 10 times the amount of average a-helical structure compared to the TIP3P simulation, and the GBOBC simulation contains about 4 times as much. A qualitative tendency for the GBHCT model to favor helix formation has been reported previously[73]. Similarly, there are greater amounts of 310-helical, a-helical, and even p-

helical structure present in these simulations. There is also a greater amount of Turn structure in both GB simulations than in the TIP3P simulations, reflecting an increased amount of localized inter-solute interaction. This is consistent with the increased helical populations observed in the GB simulations. In both the GBHCT and GBOBC simulations there is much less tendency to adopt the PP2, Extended, and Left local conformations.

Compared to the other GB models, the GBNeck simulation shows overall better agreement with the TIP3P simulation results. In particular, the amount of Extended local conformational propensity and percent Turn structure agree quite well with the TIP3P values. However, there is still a slightly larger preference for residues to be in the Alpha conformation locally (22.63±0.15% vs. 16.20±0.33% TIP3P). Also, while the GBNeck simulation contains about twice the amount of 310-helical structure as the TIP3P simulation, it contains only about half the amount of a-helical structure. As with GBHCT and GBOBC there is much less of a tendency to adopt the PP2 and Left local conformations than in the TIP3P simulations.

These results show that even for a simple system such as Ala10 which has no problematic salt bridges, the choice of solvent model has a large impact on secondary structural propensities and the local backbone dihedral conformation of residues. In particular, the GBHCT and GBOBC solvent models appear to foster the formation of a-helical structure when compared to the TIP3P solvent model, and although the GBNeck model appears to give better agreement with TIP3P solvent, there are still significant deviations.

There are two questions that should be addressed at this point: 1) Are implicit models simply unable to reproduce explicit solvent results, or 2) is the specific form of the implicit model the cause of the bias? Answering yes to the first question implies that fundamental assumption of implicit models – that is, that the bulk properties of water can be represented as a continuum dielectric – is incorrect, at least for Ala10. Studies have shown that the behavior of water near the water-peptide interface can deviate significantly from that of bulk water[144, 145]. Answering yes to the second question implies that the problem is in the GB model itself, perhaps arising from its approximate nature with respect to PE. We address the first question by comparing the GBHCT, GBOBC, GBNeck, and PE models directly to the TIP3P explicit water model, and the second question by comparing the GB models directly to PE calculations.

5.3.2 Comparison of Free Energies of Solvent Polarization from Explicit and Implicit Solvents

Since the electrostatic component of the solvation free energy (?GPol) is expected to be dominant, it is desirable to directly compare ?GPol obtained from both implicit and explicit solvent simulations. Since there is no direct calculation of ?GPol in explicit solvent models, other methods must be employed. Thermodynamic Integration (TI) is a method by which the free energy is calculated as the work done in changing a system from one state to another (State 0 ? State 1) by way of a switching function, usually represented by f(?), ? ranges from 0? 1[146]. Since ?GPol can be interpreted as the free energy cost associated with perturbing the solvent when the solute goes from an uncharged to a charged state, it can be calculated for a molecule in explicit water via TI

by making state 0 and state 1 the uncharged and charged states respectively, as has been done previously[147].

TI calculations were performed to obtain ?GPol values for four conformations of Ala10; three idealized conformations in which all backbone dihedral angles were approximately equal across all residues (Alpha, Left, and PP2), and an additional conformation generated from the backbone of a model ß-hairpin (Hairpin, see Methods for complete details). TI calculations were run with either 5 or 7 ? values and for different lengths of time to test the accuracy and sensitivity of the results, which are given in Table 5-2.

	Alpha	PP2	Left	Hairpin
0.2 ns 5 ?	-44.23	-75.62	-51.49	-55.09
1.0 ns 5 ?	-44.10	-76.51	-51.29	-53.87
1.0 ns 7 ?	-44.10	-76.43	-51.19	-54.36
2.0 ns 5 ?	-44.04	-76.22	-51.42	-54.25

Table 5-2. ?GPol (in kcal mol^{-1}) for four representative conformations of Ala10 in explicit solvent calculated with TI using varying lengths of time and ? values. A TI simulation time of 1.0 ns or greater appears to give the best results; only these TI values are considered for comparison with implicit solvent. Varying the number of ? values from 5 to 7 has comparatively little effect.

The ? GPol values generated from the TI calculations appear well converged; the difference between values is less than 1.0 kcal mol^{-1} over all variable changes. Increasing the simulation length from 0.2 ns to 1.0 ns has the largest effect, most likely from allowing the system more time to equilibrate. Because of this, only values from TI simulations 1.0 ns or greater in length are considered in the analysis. Increasing the number of ? values from 5 to 7 has little effect on final results, indicating that for this system 5 ? values is adequate.

Table 5-3A shows the comparison of ?GPol values from explicit solvent to implicit solvent models for the four conformations of Ala10. The implicit solvent model values were obtained by averaging ?GPol from the set of structures (1000 for each conformation) generated during the 1.0 ns TI calculations. Each solvent model has the same overall trend in terms of which conformation has the most favorable (lowest) solvation free energy; PP2<<Hairpin<Left<Alpha. It is interesting to note that the less solvent exposed the conformation, the more ?GPol values from the various solvent models deviate from each other, as shown in the last column of Table 5-3A (labeled Stdev). For example, the ?GPol values from both explicit and implicit solvent models are very similar the well-solvated PP2 conformation, as shown by the small standard deviation of ?GPol across all models (0.69 kcal mol^{-1}). The differences between the explicit and implicit solvent models show up more clearly in the less solvent-exposed Hairpin, Left, and Alpha conformations, with larger standard deviations of 2.02, 2.83, and 3.56 kcal mol^{-1} respectively.

A) ?GPol	TIP3P	PE	GBHCT	GBOBC	GBNeck	Stdev
Alpha	-44.08 ± 0.04	-47.96 ± 0.77	-51.69 ± 1.21	-49.38 ± 1.21	-43.26 ± 0.90	3.56
PP2	-76.39 ± 0.15	-78.04 ± 0.91	-77.35 ± 1.05	-78.07 ± 1.09	-77.59 ± 1.02	0.69
Left	-51.30 ± 0.12	-54.85 ± 0.90	-55.05 ± 1.08	-52.67 ± 1.10	-48.19 ± 0.91	2.83
Hairpin	-54.16 ± 0.25	-57.27 ± 1.13	-57.48 ± 1.45	-56.03 ± 1.47	-52.85 ± 1.29	2.01

B) ??GPol	TIP3P	PE	GBHCT	GBOBC	GBNeck
PP2-Alpha	-32.31	-30.07	-25.67	-28.69	-34.33
PP2-Left	-25.09	-23.19	-22.31	-25.40	-29.40
PP2-Hairpin	-22.23	-20.77	-19.87	-22.03	-24.73
Alpha-Left	7.22	6.88	3.36	3.29	4.93
Alpha-Hairpin	10.08	9.31	5.80	6.66	9.60
Left-Hairpin	2.86	2.43	2.43	3.37	4.67

C) ??GPol RMSD†	PE	GBHCT	GBOBC	GBNeck
Overall	1.39	3.89	2.60	2.51
PP2	1.89	4.37	2.10	3.11
Non-PP2	0.55	3.34	3.02	1.71

†RMSD from TIP3P ??GPol values.

Table 5-3. A) ?GPol (in kcal mol^{-1}) calculated for four representative conformations from the TIP3P explicit solvent and various implicit solvent models. The last column, labeled Stdev, gives the standard deviation of all implicit models from the TIP3P value. B) ??GPol between all four conformations for all solvent models. C) RMSD of implicit solvent model ??GPol values from the TIP3P values. PP2 refers to the RMSD between PP2 and the compact structures (Alpha, Left, and Hairpin), and Non-PP2 refers to the RMSD between the compact structures themselves.

It is not expected that the results from implicit solvent models will agree directly with the TI results from the TIP3P model since the intrinsic Born radii set used (Mbondi2) has not been optimized to reproduce explicit solvent values for some of these implicit models. It is still useful, however, to compare the differences in ?GPol between different conformations (??GPol), as this has a direct affect on the thermodynamics of the system, and so provides a way to relate individual ?GPol values from various solvent models to the ensembles of structures generated in the REMD runs. The ??GPol values between all conformations are given in Table 5-3B.

The first three sets of ??GPol values considered are those between the PP2 conformation and all other conformations. As the PP2 conformation is much more highly solvated and extended compared to the other conformations, these comparisons give insight into the changes in solvation that accompany peptide or protein folding. It is shown in Table 5-3B that compared to TIP3P, ??GPol between the PP2 and Alpha conformations is underestimated by PE, GBOBC, and GBHCT models by - 2.23, - 3.62, and - 6.64 kcal mol^{-1} respectively. This indicates an insufficient desolvation penalty upon the transition to the Alpha conformation. In contrast, the GBNeck model overestimates ??GPol by 2.02 kcal mol^{-1}, indicating there is too much of a desolvation penalty upon the transition to Alpha.

It is interesting to note that the PP2 and Alpha ??GPol values from both explicit and implicit solvent models correlate well (natural log fit, R2 = 0.9946) with the fractional a-helical structure (%a /[100-%a]) obtained from DSSP analysis of the corresponding REMD simulations (Figure 5-3). This shows a direct relationship between the change in free energy of solvation of a structure, and how much of that structure is observed in simulation. Based on the fit, the PE ??GPol value of –30.07 kcal mol^{-1} would translate into ~6% a-helical structure for an ensemble sampled using PE (which

was not computationally feasible for this study). This suggests that even a model based on PE may be slightly too helical compared to TIP3P, although its performance is still much better than GBHCT or GBOBC. Of course this value is simply an extrapolation, and ideally simulations using implicit solvent based on PE will be used in the future to generate well-converged ensembles.

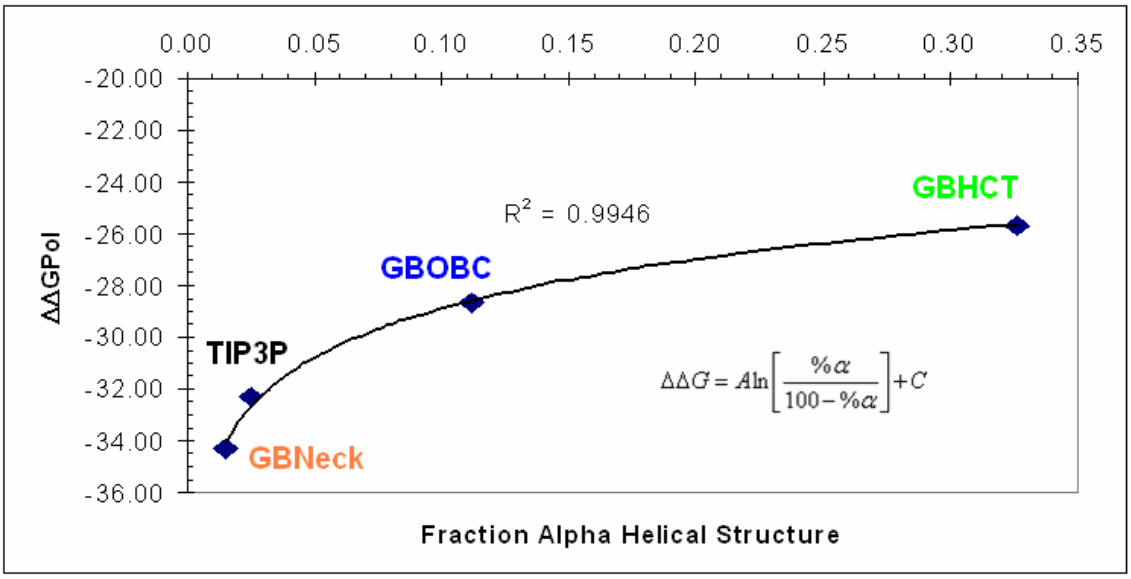

Figure 5-3. Plot of fractional a-helical structure (%a/[100-%a]) obtained from DSSP analysis of REMD simulations with various solvent models versus the corresponding ??GPol value between the PP2 and Alpha conformations. The data points from right to left are for the GBNeck, TIP3P, GBOBC, and GBHCT solvent models. As the solvation free energy gap in the given solvent model between the PP2 and Alpha structures decreases, the amount of a-helical structure in simulations with that model increases.

The ??GPol values between PP2 and Left follow a slightly different trend. Compared to TIP3P values, the PE and GBHCT models underestimate ??GPol by - 1.89 and - 2.78 kcal mol^{-1} respectively, which is consistent with the smaller ratio of PP2 to Left conformation (as determined from the values in Table 5-1) observed in the GBHCT REMD simulation (2.9) compared to the TIP3P simulation (5.8). The GBOBC model is almost an exact match, only overestimating ??GPol by 0.31 kcal mol^{-1}, consistent with the fact that the ratio of PP2 to Left in the GBOBC REMD simulation (5.9) is quite similar to the TIP3P value. The GBNeck model greatly overestimates ??GPol in this case by 4.31 kcal mol^{-1}, consistent with the greatly increased ratio of PP2 to Left conformation found in the GBNeck REMD simulation (19.7).

It is noted that while a repeating Left conformation itself is a high energy and not very realistic conformation, adopting a local left-helical conformation is important for residues in structures incorporating reverse-turns, such as ß-hairpins. It is perhaps unsurprising then that the ??GPol values between PP2 and Hairpin follow a similar trend to those between PP2 and Left. The PE and GBHCT models underestimate ??GPol by - 1.51 and - 2.40 kcal mol^{-1} respectively. The GBOBC model is again almost exact, underestimating by only - 0.24 kcal mol^{-1}. The GBNeck model overestimates ??GPol by 2.46 kcal mol^{-1}.

The last three sets of ??GPol values considered are between the Alpha, Hairpin, and Left conformations, which are less solvated and have more favorable internal contacts compared to the PP2 conformation. The performance of PE in all three cases is superb; the largest deviation from TIP3P is ??GPol between Alpha and Hairpin, which is only 0.73 kcal mol^{-1}.

The overall performance for all three GB models for these compact structures is markedly worse than PE. All three GB models overestimate ??GPol between Alpha and Left; GBHCT and GBOBC by about 3.9 kcal mol^{-1}, and GBNeck by about 2.3 kcal mol^{-1}. The desolvation penalty between these two conformations being too large is consistent with the increased ratio of Alpha to Left conformational propensity observed in the GBHCT, GBOBC, and GBNeck REMD simulations (~18) compared to the ratio from the TIP3P REMD simulation (~3).

The remaining comparisons show no consistent pattern and serve only to highlight how the performance of each GB model depends on conformation. The GBOBC and GBHCT models overestimate ??GPol between Alpha and Hairpin by 4.28 and 3.42 kcal mol^{-1} respectively, while GBNeck only overestimates by 0.44 kcal mol^{-1}. In contrast, the GBNeck model underestimates ??GPol between Left and Hairpin by 1.79 kcal mol^{-1}, while the GBOBC and GBHCT models are within 0.5 kcal mol^{-1} of the TIP3P value.

It is clear that the performance of implicit solvent models is dependent on the conformation of Ala10. As a way to gauge the overall performance of each implicit solvent model with respect to the TIP3P solvent model, the RMSD from TIP3P ?? GPol values for each implicit solvent model was calculated (Table 5-3C). The best overall performance is from PE, with an overall RMSD of 1.39 kcal mol^{-1}. The next best performance is by the GBNeck and GBOBC models, with RMSDs of 2.51 and 2.60 kcal mol^{-1} respectively. The worst performance is from the GBHCT model, with an overall RMSD of 3.89 kcal mol^{-1}. For reproducing the difference between PP2 and more compact states (analogous to folding, PP2 column in Table 5-3C), PE again has the best performance (1.89 kcal mol^{-1}), with GBOBC coming in a close second (2.10 kcal mol^{-1}). GBNeck and GBHCT perform worse, with RMSDs of 3.11 and 4.37 kcal mol^{-1}. For reproducing the differences between compact states themselves (Non-PP2 column in Table 5-3C), PE is clearly superior to all of the GB models, with a RMSD of 0.55 kcal mol^{-1}. GBNeck is a distant second with a RMSD of 1.71 kcal mol^{-1}, while GBOBC and GBHCT have RMSDs of 3.02 and 3.34 respectively.

The overabundance of helical structure in the REMD ensembles obtained with the GBHCT and GBOBC solvent models can now be rationalized. Essentially, these models over-stabilize a-helices because not enough of a desolvation penalty is paid for forming the a-helical structure; the already favorable internal energy of the a-helix is accompanied by an overly favorable solvation free energy. In contrast, the desolvation penalty for formation of a-helical structure with the GBNeck model is comparable to PE and TIP3P, and a-helical structure is not overly-abundant in the REMD simulations with this model.

Overall, PE is the best of the implicit models at reproducing the differences in ?GPol between different conformations of Ala10, while all GB models perform considerably worse. PE and GBOBC are both good at reproducing the differences between PP2 and the more compact conformations. Although PE is clearly superior to all

GB models at reproducing the differences between the compact conformations, it should be noted that GBNeck is still much better at this than GBHCT or GBOBC. It is interesting to point out that in particular all GB models have difficulty reproducing the difference between the right-handed and left-handed alpha helix.

The reason for the relatively poor performance of these GB models compared to the PE model, especially for reproducing ??GPol between the more compact structures (Alpha, Left, and Hairpin) is not clear at this point. In the next section, this problem is explored by comparing the effective radii and energy calculations of these three GB models to effective Born radii and energy calculated with PE.

5.3.3 Direct Comparison of GB to PE

5.3.3.1 Effective Radii

All implicit models rely on an accurate description of the dielectric boundary for good performance[116]. In this study, the model based on PE (DelPhi 2.0) calculates this boundary based on the molecular surface accessible to a probe with a radius comparable to that of a water molecule (1.4 Å), which is then mapped onto a cubic lattice. In these GB models, instead of a specific dielectric boundary, each atom is assigned an effective Born radius (RGB), which is essentially a measure of how well solvated the atom is. For atoms that are well-solvated (*i.e.* atoms that have a more favorable solvation free energy) this radius is small, reflecting the damping effect that a solvent with high dielectric has on atomic charge. The relationship between RGB and atomic solvation free energy (Self Energy) can be seen clearly from Equation 5-6.

The fundamental difference between the GBHCT, GBOBC, and GBNeck models discussed here is in the calculation of RGB (see Methods for more details). Onufriev *et al.* showed that when RGB is calculated from atomic ?GPol obtained using PE, the resulting 'perfect' Born radii (RPE) improve the accuracy of both GB Self and Interaction energy terms, and improve overall agreement with PE[126]. Since out of all the implicit models, PE had the best performance in reproducing explicit solvent ??GPol values, examining the deviation between 'perfect' radii obtained via PE and those calculated with the various GB models may provide insight into areas where GB is deficient, and reveal specific areas to improve.

Effective Born radii were calculated with PE (RPE), and compared to RGB obtained from the GBHCT, GBOBC, and GBNeck implicit solvent models using a subset of the last 500 structures from the Alpha, Left, and Hairpin TI calculation trajectories, and a subset of 100 structures (frames 500-599) from the PP2 TI calculation trajectories. A subset of structures was chosen since derivation of RPE for many structures is particularly time consuming as it requires a PE calculation for every atom in every structure. Fewer structures were used for PP2 as the PE calculations for these structures are particularly time-consuming (because of the large solvent-exposed surface area of this conformation).

Table 5-4A-I shows the RMSD of RGB from RPE for each of the GB solvent models across all residues of Ala10 for the given atom type, averaged over all structures used in the ? GPol analysis shown in Table 5-3. The atom types considered are amide

hydrogen (H), carbonyl oxygen (O), amide nitrogen (N), carbonyl carbon (C), a carbon (CA), ß carbon (CB), a hydrogen (HA), backbone atoms (BB, representing H, O, N, C, and CA), and all atom types. Table 5-4J shows the average RMSD value over all conformations for the given solvent model. Table 5-5 shows the average difference instead of RMSD for each atom type, to convey whether RGB is under-estimated or over-estimated with respect to RPE.

GB Effective Radii Average RMSD from Perfect (PE) Radii (?)

A) All	GBHCT	GBOBC	GBNeck
alpha	0.25 ± 0.01	0.19 ± 0.01	0.22 ± 0.02
hairpin	0.18 ± 0.01	0.16 ± 0.01	0.12 ± 0.01
left	0.20 ± 0.01	0.20 ± 0.01	0.32 ± 0.03
pp2	0.06 ± 0.00	0.11 ± 0.01	0.04 ± 0.00

B) BB	GBHCT	GBOBC	GBNeck
alpha	0.35 ± 0.02	0.26 ± 0.01	0.31 ± 0.03
hairpin	0.20 ± 0.02	0.20 ± 0.01	0.16 ± 0.01
left	0.27 ± 0.02	0.26 ± 0.02	0.45 ± 0.04
pp2	0.06 ± 0.00	0.12 ± 0.01	0.05 ± 0.00

C) H	GBHCT	GBOBC	GBNeck
alpha	0.71 ± 0.04	0.29 ± 0.04	0.19 ± 0.06
hairpin	0.39 ± 0.04	0.19 ± 0.03	0.15 ± 0.03
left	0.50 ± 0.04	0.13 ± 0.02	0.43 ± 0.06
pp2	0.04 ± 0.00	0.10 ± 0.01	0.02 ± 0.00

D) O	GBHCT	GBOBC	GBNeck
alpha	0.16 ± 0.01	0.09 ± 0.01	0.20 ± 0.02
hairpin	0.16 ± 0.01	0.08 ± 0.01	0.11 ± 0.02
left	0.18 ± 0.01	0.07 ± 0.01	0.24 ± 0.02
pp2	0.03 ± 0.00	0.02 ± 0.00	0.04 ± 0.00

E) N	GBHCT	GBOBC	GBNeck
alpha	0.26 ± 0.01	0.30 ± 0.03	0.37 ± 0.04
hairpin	0.16 ± 0.01	0.21 ± 0.02	0.11 ± 0.02
left	0.27 ± 0.02	0.31 ± 0.03	0.52 ± 0.05
pp2	0.08 ± 0.01	0.13 ± 0.01	0.07 ± 0.01

F) C	GBHCT	GBOBC	GBNeck
alpha	0.16 ± 0.01	0.19 ± 0.02	0.42 ± 0.03
hairpin	0.08 ± 0.01	0.20 ± 0.01	0.25 ± 0.02
left	0.12 ± 0.01	0.29 ± 0.03	0.58 ± 0.04
pp2	0.07 ± 0.01	0.09 ± 0.01	0.05 ± 0.01

G) CA	GBHCT	GBOBC	GBNeck
alpha	0.05 ± 0.01	0.34 ± 0.02	0.26 ± 0.02
hairpin	0.09 ± 0.01	0.26 ± 0.01	0.12 ± 0.01
left	0.07 ± 0.01	0.36 ± 0.02	0.37 ± 0.03
pp2	0.03 ± 0.01	0.18 ± 0.01	0.04 ± 0.01

H) CB	GBHCT	GBOBC	GBNeck
alpha	0.03 ± 0.00	0.03 ± 0.00	0.12 ± 0.01
hairpin	0.04 ± 0.00	0.04 ± 0.00	0.08 ± 0.00
left	0.01 ± 0.00	0.02 ± 0.00	0.10 ± 0.00
pp2	0.03 ± 0.00	0.04 ± 0.00	0.05 ± 0.00

I) HA	GBHCT	GBOBC	GBNeck
alpha	0.07 ± 0.00	0.19 ± 0.01	0.04 ± 0.01
hairpin	0.34 ± 0.03	0.23 ± 0.02	0.12 ± 0.03
left	0.10 ± 0.01	0.21 ± 0.01	0.11 ± 0.02
pp2	0.08 ± 0.00	0.18 ± 0.01	0.02 ± 0.00

J) Overall Averages

	GBHCT	GBOBC	GBNeck
All	0.17	0.16	0.17
BB	0.22	0.21	0.24
H	0.41	0.18	0.20
O	0.13	0.06	0.15
N	0.19	0.24	0.27

	GBHCT	GBOBC	GBNeck
C	0.11	0.19	0.33
CA	0.06	0.29	0.20
CB	0.03	0.03	0.09
HA	0.15	0.20	0.08

Table 5-4. A-I) RMSD of effective GB radii calculated with various GB models from effective radii calculated with PE (perfect radii) for four conformations of Ala10, shown for various atom types: All, BB (backbone atoms, H, O, N, C, CA), H (amide hydrogen), O (carbonyl oxygen), N (amide nitrogen), C (carbonyl carbon), CA (a carbon), CB (ß carbon), and HA (a hydrogen). J) Overall average RMSD over the four conformations of Ala10 for each GB solvent model.

Two trends are readily apparent from the effective radii RMSDs shown in Table 5-4A for all atom types and Table 5-4B for all backbone atom types: 1) The largest deviations of RGB from RPE are in backbone atoms, and 2) the deviation of RGB from RPE in PP2 conformations is significantly smaller than for the more compact Alpha, Left, and Hairpin conformations across all GB models. These two observations are consistent with the idea that the performance of GB models decreases the more buried an atom is, and also consistent with previously published comparisons of RGB with RPE[73,

126]. The corresponding average differences in Table 5-5A and Table 5-5B show that in general the GBOBC and GBNeck models tend to overestimate RGB (and thus underestimate solvation), while the GBHCT model underestimates RGB.

GB Effective Radii Average Difference from Perfect (PE) Radii (?)

A) H	GBHCT	GBOBC	GBNeck
alpha	0.59 ± 0.38	0.18 ± 0.19	-0.11 ± 0.12
hairpin	0.18 ± 0.34	-0.02 ± 0.18	0.03 ± 0.13
left	0.39 ± 0.30	-0.01 ± 0.10	-0.37 ± 0.20
pp2	-0.04 ± 0.02	-0.10 ± 0.02	0.00 ± 0.02

B) O	GBHCT	GBOBC	GBNeck
alpha	0.12 ± 0.10	0.06 ± 0.06	-0.17 ± 0.10
hairpin	0.08 ± 0.13	0.03 ± 0.07	-0.09 ± 0.06
left	0.15 ± 0.11	0.02 ± 0.05	-0.20 ± 0.12
pp2	-0.01 ± 0.03	-0.01 ± 0.02	-0.04 ± 0.01

C) N	GBHCT	GBOBC	GBNeck
alpha	0.24 ± 0.10	-0.26 ± 0.13	-0.30 ± 0.21
hairpin	0.11 ± 0.11	-0.16 ± 0.12	-0.06 ± 0.09
left	0.24 ± 0.11	-0.27 ± 0.14	-0.43 ± 0.29
pp2	0.07 ± 0.04	-0.11 ± 0.07	0.05 ± 0.04

D) C	GBHCT	GBOBC	GBNeck
alpha	0.14 ± 0.07	-0.17 ± 0.08	-0.37 ± 0.21
hairpin	0.03 ± 0.07	-0.17 ± 0.10	-0.21 ± 0.13
left	0.10 ± 0.06	-0.26 ± 0.11	-0.50 ± 0.28
pp2	0.04 ± 0.05	-0.07 ± 0.04	-0.02 ± 0.04

E) CA	GBHCT	GBOBC	GBNeck
alpha	0.03 ± 0.03	-0.33 ± 0.07	-0.23 ± 0.11
hairpin	0.05 ± 0.07	-0.25 ± 0.07	-0.09 ± 0.07
left	0.05 ± 0.05	-0.35 ± 0.09	-0.32 ± 0.18
pp2	0.02 ± 0.02	-0.18 ± 0.03	0.03 ± 0.02

F) CB	GBHCT	GBOBC	GBNeck
alpha	0.00 ± 0.03	0.00 ± 0.03	-0.11 ± 0.03
hairpin	0.02 ± 0.03	0.03 ± 0.03	-0.08 ± 0.02
left	0.00 ± 0.01	0.01 ± 0.01	-0.10 ± 0.03
pp2	0.02 ± 0.02	0.03 ± 0.02	-0.05 ± 0.01

G) HA	GBHCT	GBOBC	GBNeck
alpha	-0.05 ± 0.04	-0.18 ± 0.03	-0.03 ± 0.03
hairpin	0.13 ± 0.31	-0.07 ± 0.21	0.04 ± 0.10
left	-0.04 ± 0.09	-0.20 ± 0.05	-0.09 ± 0.05
pp2	-0.08 ± 0.01	-0.18 ± 0.04	0.02 ± 0.01

H) Overall Averages

	GBHCT	GBOBC	GBNeck		GBHCT	GBOBC	GBNeck
H	0.28	0.01	-0.11	*CA*	0.04	-0.28	-0.15
O	0.08	0.02	-0.13	*CB*	0.01	0.02	-0.08
N	0.17	-0.20	-0.18	*HA*	-0.01	-0.16	-0.02
C	0.08	-0.17	-0.27				

Table 5-5. A-I) Average deviation of effective GB radii calculated with various GB models from effective radii calculated with PE (perfect radii) across all residues of four conformations of Ala10, shown for various atom types: H (amide hydrogen), O (carbonyl oxygen), N (amide nitrogen), C (carbonyl carbon), CA (a carbon), CB (ß carbon), and HA (a hydrogen). J) Overall average deviation over the four conformations of Ala10 for each GB solvent model.

Each GB model shows different behavior across different atom types and conformations (Table 5-4C-I and Table 5-5A-G). The largest deviation in the GBHCT model is from the amide hydrogens (H), which has an average RMSD across all residues of 0.41 Å; this is the worst of all three GB models. A detailed look at the H atoms confirms that the deviation is greatest when the atoms are buried, such as when involved in hydrogen bonding. For example, the H atom of residue A1 in the hairpin structure (which is solvent exposed) shows almost no deviation, while RGB for the H atom of the very next residue (which is involved in a hydrogen bond) is underestimated by 0.70 Å (Data not shown).

The average deviations across the Alpha, Hairpin, and Left structures seen in Table 5-5A indicate that in the GBHCT model RGB is always underestimated for H atoms, meaning that they are considered more solvent exposed than they should be according to PE. In addition, RGB is also underestimated for carbonyl oxygen (O) atoms in these conformations. This leads to the conclusion that in this model, backbone hydrogen bonding between H and O atoms will be over-stabilized due to an insufficient

desolvation penalty, consistent with the overabundance of helical structures observed in the unrestrained REMD structural ensembles.

RGB is underestimated in general for all other atom types in GBHCT, particularly the amide nitrogen (N) atoms (average RMSD of 0.19 Å). However, the performance for carbonyl carbon (C) and a carbon (CA) atoms is the best of all the GB models (average RMSDs of 0.11 and 0.06 Å respectively). Overall, the performance of this model for Ala10 becomes progressively worse the less solvated the structure becomes. This behavior is consistent with previous observations of this GB model[127, 137].

The behavior of the GBOBC model is slightly more varied. The RGB for H and O atoms is still underestimated, particularly when these atoms are buried, but to a much lesser extent than in GBHCT (average RMSDs of 0.18 and 0.06 Å respectively). In fact, the GBOBC model has the best performance for O atoms out of any of the GB models. This indicates that backbone hydrogen bonds between H and O atoms may still be over-stabilized, but to a lesser extent than in GBHCT. It is also interesting to note that the deviation of RGB for H atoms in the Left conformation is quite small compared to the other two GB models. However, RGB is overestimated for N, HA, C, and CA atoms (average RMSDs of 0.24, 0.20 Å, 0.19, 0.29, and respectively). The deviation for CA atoms is particularly large compared to that for GBHCT; in fact GBOBC has the worst performance for CA atoms out of the three GB models. As with the GBHCT model, the performance for the GBOBC model is worse for less well-solvated structures.

The performance of GBNeck for H atoms is comparable to that of GBOBC (overall RMSD of 0.20), except for the Left conformation, where it has deviations as large as those of GBHCT. The performance of GBNeck for O atoms is also about as poor as GBHCT (overall RMSD of 0.15). In contrast to GBHCT and GBOBC however, GBNeck overestimates RGB for H and O atoms, the net result of which is a destabilization of hydrogen bonds between these two atoms due to an increased desolvation penalty. In fact, the GBNeck model in general overestimates RGB for all atom types. The performance of GBNeck for C atoms is particularly bad compared to the other two GB models (overall RMSD of 0.33), as is its performance for ß carbon (CB) atoms. The only atom type for which GBNeck performs well compared to the other GB models is a hydrogen (HA) atoms (overall RMSD of 0.08). Like the GBHCT and GBOBC models, the performance of the GBNeck model is worse for less well-solvated structures, except it has more deviation for the Left conformation than the Alpha conformation; the reason for this is not clear.

It is seen here that each GB model has significant deviations in calculation of RGB for various atom types, and the differences are in general not consistent between the GB models. The only real consistency is that RGB approaches RPE for well-solvated structures. In terms of overall RGB RMSD from RPE, each model performs about equally, except for the GBNeck model and the Left conformation as noted above. The differences between the GB models will be further examined by translating the effective radii into actual solvation free energies.

5.3.3.2 Solvation Free Energy

Equation 5-6 shows that the effective Born radius of an atom is directly related to its solvation free energy; this is the Self energy portion of the GB equation (sum of terms

in Equation 5-1 when i=j). However, it is important to note that this energy is also highly dependent on the charge of the atom. The magnitude of the differences between the GB and PE effective Born radii shown in Table 5-4 and Table 5-5 will be strongly modified by the charges on the atoms. For each of the three GB solvent models, the average RMSD of PE self energies from GB self energies across all residues of Ala10 for various atom types are shown in Table 5-6.

GB Atomic Self Energy RMSD from Perfect (PE) Atomic Self Energy (kcal/mol)

A) All	GBHCT	GBOBC	GBNeck
alpha	0.90 ± 0.03	0.65 ± 0.03	1.11 ± 0.08
hairpin	0.70 ± 0.04	0.73 ± 0.04	0.79 ± 0.06
left	0.83 ± 0.04	0.75 ± 0.05	1.37 ± 0.08
pp2	0.32 ± 0.02	0.39 ± 0.03	0.30 ± 0.02

B) BB	GBHCT	GBOBC	GBNeck
alpha	1.28 ± 0.05	0.93 ± 0.05	1.59 ± 0.11
hairpin	1.00 ± 0.05	1.04 ± 0.05	1.13 ± 0.08
left	1.19 ± 0.05	1.07 ± 0.07	1.96 ± 0.11
pp2	0.45 ± 0.03	0.56 ± 0.04	0.43 ± 0.03

C) H	GBHCT	GBOBC	GBNeck
alpha	1.35 ± 0.06	0.49 ± 0.06	0.24 ± 0.06
hairpin	0.82 ± 0.06	0.47 ± 0.04	0.30 ± 0.04
left	1.03 ± 0.06	0.31 ± 0.04	0.64 ± 0.07
pp2	0.18 ± 0.02	0.40 ± 0.03	0.10 ± 0.02

D) O	GBHCT	GBOBC	GBNeck
alpha	1.86 ± 0.08	1.00 ± 0.11	1.92 ± 0.15
hairpin	1.73 ± 0.10	0.94 ± 0.10	1.16 ± 0.12
left	1.94 ± 0.09	0.72 ± 0.09	2.00 ± 0.16
pp2	0.50 ± 0.05	0.38 ± 0.06	0.68 ± 0.05

E) N	GBHCT	GBOBC	GBNeck
alpha	1.07 ± 0.06	1.00 ± 0.09	1.10 ± 0.10
hairpin	0.76 ± 0.07	0.98 ± 0.08	0.53 ± 0.07
left	1.07 ± 0.07	1.03 ± 0.09	1.46 ± 0.11
pp2	0.45 ± 0.04	0.69 ± 0.05	0.38 ± 0.04

F) C	GBHCT	GBOBC	GBNeck
alpha	1.29 ± 0.09	1.39 ± 0.14	2.69 ± 0.18
hairpin	0.81 ± 0.08	1.78 ± 0.12	2.10 ± 0.14
left	0.98 ± 0.09	1.97 ± 0.16	3.47 ± 0.18
pp2	0.71 ± 0.08	0.87 ± 0.08	0.52 ± 0.06

G) CA	GBHCT	GBOBC	GBNeck
alpha	0.00 ± 0.00	0.01 ± 0.00	0.01 ± 0.00
hairpin	0.00 ± 0.00	0.01 ± 0.00	0.00 ± 0.00
left	0.00 ± 0.00	0.01 ± 0.00	0.01 ± 0.00
pp2	0.00 ± 0.00	0.01 ± 0.00	0.00 ± 0.00

H) CB	GBHCT	GBOBC	GBNeck
alpha	0.03 ± 0.00	0.03 ± 0.01	0.14 ± 0.01
hairpin	0.05 ± 0.01	0.05 ± 0.01	0.09 ± 0.01
left	0.02 ± 0.00	0.02 ± 0.00	0.12 ± 0.01
pp2	0.03 ± 0.00	0.05 ± 0.00	0.07 ± 0.00

I) HA	GBHCT	GBOBC	GBNeck
alpha	0.03 ± 0.00	0.07 ± 0.00	0.02 ± 0.00
hairpin	0.09 ± 0.01	0.07 ± 0.00	0.02 ± 0.01
left	0.04 ± 0.00	0.08 ± 0.00	0.04 ± 0.00
pp2	0.03 ± 0.00	0.07 ± 0.00	0.01 ± 0.00

J) Overall Averages

	GBHCT	GBOBC	GBNeck
All	0.69	0.63	0.89
BB	0.98	0.90	1.28
H	0.84	0.42	0.32
O	1.51	0.76	1.44
N	0.84	0.92	0.87

	GBHCT	GBOBC	GBNeck
C	0.95	1.50	2.19
CA	0.00	0.01	0.00
CB	0.03	0.04	0.10
HA	0.05	0.07	0.02

Table 5-6. RMSD of the polar component of atomic self solvation free energy calculated with effective radii obtained using various GB models from atomic self solvation free energy calculated with perfect radii for four conformations of Ala10, shown for various atom types: All, BB (backbone atoms, H, O, N, C, CA), H (amide hydrogen), O (carbonyl oxygen), N (amide nitrogen), C (carbonyl carbon), CA (a carbon), CB (ß carbon), and HA (a hydrogen). J) Overall average RMSD of GB self solvation free energy from PE self solvation free energy over the four conformations of Ala10 for each GB solvent model.

There is of course a direct relationship between deviations in effective radii and deviations in Self solvation free energy; an atom whose effective radius has been underestimated will have an overestimated solvation free energy, and vice versa. What is less clear is the relationship between the magnitude of deviation of effective radii and magnitude of deviation of self solvation free energy. It is apparent that the relatively small (for the most part < 0.5 Å) deviations in effective radii in Table 5-4 and Table 5-5

can translate into significant differences in Self energy on the order of ~1.0 kcal/mol, but this is, of course, highly dependent on the charge of the atom. For example, in the GBHCT model even though the average radii RMSD for H atoms was about three times as large as the average radii RMSD for O atoms, the average self energy RMSD for H atoms is only about half as large. As expected, radii deviations for atoms with small charges become almost insignificant in terms of energy. For example, although large deviations in the effective radius were observed for CA atoms in the GBOBC model, the energy differences are negligible (< 0.01 kcal/mol).

Of course, the Self energy is just part of the GB model; only the Total GB energy can be directly related to observed structural ensembles, so it is important to calculate the Interaction energy as well (sum of terms in Equation 5-1 when ?j). Table 5-7 shows the Self, Interaction, and Total GB energies computed with effective radii obtained with the GBHCT, GBOBC, and GBNeck models (RGB), and PE derived effective radii (Perfect radii, RPE) for the structures used in the analysis shown in Table 5-4, Table 5-5, and Table 5-6. Note the excellent agreement of the Total ?GPol values in Table 5-7 with ?GPol values in Table 5-3A, showing that choosing a subset of structures for the effective radii analysis has not adversely affected the results.

Total	PE	Perfect	GBHCT	GBOBC	GBNeck
alpha	-47.96 ± 0.77	-47.42 ± 0.77	-51.64 ± 0.94	-49.38 ± 0.98	-43.27 ± 0.82
hairpin	-57.28 ± 1.15	-57.70 ± 1.04	-57.45 ± 1.17	-56.02 ± 1.19	-52.83 ± 1.05
left	-54.85 ± 0.90	-52.45 ± 0.82	-55.05 ± 0.90	-52.70 ± 0.93	-48.24 ± 0.08
pp2	-78.00 ± 0.92	-81.22 ± 0.97	-77.26 ± 0.94	-77.99 ± 0.95	-77.48 ± 0.92

Self	PE	Perfect	GBHCT	GBOBC	GBNeck
alpha	-763.77 ± 1.68	-763.77 ± 1.68	-813.34 ± 2.91	-748.18 ± 4.85	-703.92 ± 5.51
hairpin	-822.96 ± 2.96	-822.96 ± 2.96	-843.65 ± 2.23	-798.59 ± 3.55	-787.25 ± 3.58
left	-754.11 ± 1.86	-754.11 ± 1.86	-798.61 ± 3.00	-724.13 ± 5.15	-676.81 ± 5.54
pp2	-882.27 ± 1.45	-882.27 ± 1.45	-885.93 ± 1.47	-862.55 ± 2.12	-875.09 ± 1.52

Interaction	PE†	Perfect	GBHCT	GBOBC	GBNeck
alpha	715.81	716.36 ± 1.75	761.70 ± 2.48	698.80 ± 4.34	660.65 ± 5.17
hairpin	765.68	765.26 ± 2.94	786.20 ± 1.89	742.57 ± 3.07	734.42 ± 3.22
left	699.26	701.66 ± 2.03	743.56 ± 2.75	671.43 ± 4.85	628.57 ± 5.35
pp2	804.27	801.05 ± 1.44	808.67 ± 1.32	784.56 ± 1.86	797.61 ± 1.36

†Calculated from PE(Total) - PE(Self)

Table 5-7. Total, Self, and Interaction components of ?GPol (in kcal mol^{-1}) calculated with either the PE or one of the GB implicit solvent models.

In Table 5-7 it is apparent that although the deviations in the Total energy between PE and each GB model are on the order of a few kcal/mol, there are significant differences in the Self and Interaction GB energies which end up cancelling in the Total solvation free energy. This behavior for GB models has been observed previously[126, 127].

As was noted by Onufriev *et al.*[126], use of effective Born radii calculated via PE improves the quality of interaction energies as well as self energies; surprisingly, this improvement is not always reflected in the Total energy, where other GB models may happen to have better agreement with PE results due to fortuitous cancellation of error.

For example, although perfect radii give the lowest Total energy deviation for the Alpha conformation (- 0.55 kcal/mol), it does not for the Left conformation (- 2.40 kcal/mol); in that case the lowest deviation is from the GBHCT model (0.20 kcal/mol).

As in the previous section, the differences in Total, Self, and Interaction energies shown in Table 5-7 between different conformations are considered (Table 5-8) in order to better compare the performance of each implicit model. Here it is seen that despite the fact that using perfect effective radii brings the Self and Interaction GB energies much closer to those calculated with PE, the use of perfect radii shows no improvement over other GB effective radii calculations in terms of reproducing the solvation free energy differences between different conformations of Ala10. This finding is consistent with that from a study by Stultz, who suggested that agreement with PE alone may be an inadequate way to parameterize GB models for the purpose of calculating free energy differences[148].

A) ??GPol Total

	P-A	P-L	A-L	P-H	A-H	L-H
PE	-30.03	-23.15	6.89	-20.72	9.32	2.43
Perfect	-33.80	-28.77	5.03	-23.52	10.28	5.25
GBHCT	-25.62	-22.21	3.41	-19.81	5.80	2.40
GBOBC	-28.61	-25.29	3.32	-21.97	6.64	3.32
GBNeck	-34.21	-29.25	4.97	-24.66	9.56	4.59

B) ??GPol Self

	P-A	P-L	A-L	P-H	A-H	L-H
PE	-118.49	-128.15	-9.66	-59.31	59.18	68.84
Perfect	-118.49	-128.15	-9.66	-59.31	59.18	68.84
GBHCT	-72.58	-87.32	-14.74	-42.28	30.30	45.04
GBOBC	-114.37	-138.42	-24.05	-63.96	50.41	74.45
GBNeck	-171.17	-198.28	-27.11	-87.85	83.33	110.43

C) ??GPol Interaction

	P-A	P-L	A-L	P-H	A-H	L-H
PE	88.46	105.01	16.55	38.59	-49.87	-66.42
Perfect	84.69	99.39	14.69	35.79	-48.90	-63.59
GBHCT	46.97	65.11	18.14	22.47	-24.50	-42.64
GBOBC	85.76	113.13	27.37	42.00	-43.76	-71.14
GBNeck	136.96	169.04	32.08	63.19	-73.77	-105.84

D) RMSD from PE

	Total	Self	Interaction
Perfect	3.32	0.00	3.32
GBHCT	2.76	30.25	28.24
GBOBC	2.19	8.75	6.60
GBNeck	3.62	43.06	39.63

A=Alpha, P=PP2, L=Left, H=Hairpin

Table 5-8. A-C) Differences in components of ?GPol (from Table 5-7) between conformations of Ala10 (kcal mol-1). D) RMSD of ??GPol calculated with GB models from PE ??GPol for the Total, Self, and Interaction components of solvation free energy.

5.4 Conclusions

In this study, we directly compared the TIP3P explicit solvent model to results from PE and three GB solvent models. Well-converged REMD simulations using either the TIP3P solvent model or each of the three GB solvent models revealed that simulations with GB models show markedly different conformational and structural preferences. In particular, the GBHCT and GBOBC models contained an overabundance of helical structure compared to explicit solvent results and experiment. Thus the different solvent models not only provide ensembles with different secondary structure populations, but the "native" structure in each solvent model (as defined by the dominant conformation in the ensemble) differs depending on the solvent model used for the simulation. This result has significant implications for the use of these GB models for structure prediction or characterization of folding landscapes.

Using the TIP3P model as the standard, we directly compared free energies of solvent polarization from each model for four different conformations of Ala10; right-handed a-helix (Alpha), left-handed a-helix (Left), ß-hairpin (Hairpin), and polyproline II helix (PP2). The performance of implicit models was found to be dependent on conformation; in general, agreement with TIP3P results was best for the well-solvated PP2 conformation, growing progressively worse for more compact conformations (Hairpin, Left, and Alpha). PE was found to have the best overall performance in terms of reproducing differences in solvation free energy between the different conformations. It was also found that the amount of a-helical structure in the unrestrained REMD simulations is correlated to the solvation free energy gap between the PP2 (unfolded model) and Alpha conformations; in the GBHCT and GBOBC solvent models this gap was too small, which is related to the observed overabundance of helical structure in the REMD simulations.

One difference between the TIP3P and GB REMD simulations reported here is the lack of a specific term for ?GNonpol in the GB simulations. The absence of this term could potentially further affect the thermodynamics in these simulations. However, it has been shown that the errors in ?GPol from the various GB models correlate well with structure populations observed in the REMD simulations, and so it appears ?GPol dominates the solvation free energy. Therefore it is likely that inclusion of a term for ?GNonpol would not change the results significantly.

The effective Born radius calculation of each GB model (RGB) was compared to effective Born radii calculated with PE (RPE). While small deviations in effective radii were found for PP2, significant deviations were found for the more compact conformations. It is likely that backbone hydrogen bonds are too stable in the GBHCT and GBOBC models because RGB is underestimated for amide hydrogen (H) and carbonyl oxygen (O) atoms, leading to an insufficient desolvation penalty for hydrogen bonds. Likewise, the GBNeck model overestimates RGB for these atoms, leading to unstable hydrogen bonds and a lower helical population.

As has been reported by others, we note that substantial errors in the Self and Interaction GB energies tend to cancel in the net Total energies. The significant cancellation of error that we observe supports the idea that individual GB energy components should be considered when comparing total GB energies to results from PE, as is often done during development or validation of GB models.

As has been seen before, using RPE in the GB function improves the agreement between Self and Interaction energies compared to PE. However, this improvement does not translate into overall better performance; so-called ‘perfect’ radii are no better at capturing the difference between the conformations here than any other GB model that we tested. This may suggest a limit to how much GB models can be improved solely through optimization of the effective Born radius calculation.

Chapter 6
Summary

The advance of computational methods used to study various problems in structural biology has been nothing short of astounding; consider that within the past few decades the application of these methods has gone from simple sub-picosecond simulations of a protein to multi-microsecond simulations and accurate all atom prediction of the three-dimensional structure of a protein solely from its sequence. There have been advances not only in computational resources (processor speed, storage capacity, parallel computation, *etc.*), but also in the techniques and algorithms used. All of these advances have enabled computational techniques to be applied to an ever-widening range of problems.

The work presented in this thesis is by no means a complete illustration of all the various applications of computational methods. Nevertheless, this work has shown how several of these methods (molecular dynamics, enhanced sampling, free energy calculations, *etc.*) can be used to describe protein folding pathways, examine cooperativity in protein folding, and analyze protein stability. This work has also emphasized that the accuracy of results from computational methods should always be verified against experimental data if possible, since computational models often rely on assumptions and simplifications (such as in the case of the GB implicit solvent model) that can result in significant errors in some cases.

This work has focused largely on the ß-sheet secondary structure type; in particular the ß-hairpin model Trpzip2, and the 3-stranded ß-sheet model DPDP. Since a ß-hairpin represents the simplest form of ß-sheet structure, study of the folding mechanism of the hairpin can provide insight into the earliest stages of folding of larger ß-sheet structures (and therefore the earliest stages of protein folding in general). In the study of Trpzip2 presented here, the computational model used was verified via comparison of calculated thermodynamic parameters from well-converged data with experimental thermodynamic parameters. Although multiple pathways were observed, the key feature of hairpin formation was cross-strand hydrophobic pairing. This type of contact not only stabilized the folded hairpin, but helped keep the individual strands of the hairpin in relatively close contact in the unfolded state. It seems that in Trpzip2, backbone hydrogen bond formation appears largely dependent on formation of the nearest hydrophobic contact pair. Also important was the turn region, which appeared turn-like even in the unfolded state and likely serves in conjunction with cross-strand hydrophobic pairing to reduce the conformational search of bringing the two strands of the hairpin together.

The next simplest ß-sheet system after the ß-hairpin is the three stranded ß-sheet; in this case DPDP was studied. In studying ß-sheets one is able to examine the concept of folding cooperativity parallel to strand direction; *i.e.* how easy is it to add strands to an extant ß-hairpin. Just as the formation of the first helical hydrogen bond is the nucleation step for formation of an a-helix, the formation of a ß-hairpin may be the nucleation step for ß-sheet formation. Since the formation of extended ß-sheet-like structures (such as in

amyloid fibrils) is a hallmark of protein misfolding diseases (such as Alzheimer's), study of cooperativity in ß-sheet formation may give insight into the earliest phases of formation of these structures. Again, the computational model was verified against experimental data; in this case the lower limit of folding cooperativity was reproduced when studied in a manner analogous to the experiment. However, after verifying that the model could reproduce experimental observables a more in-depth analysis was performed that was able to calculate the actual cooperativity of strand formation, which was found to be significantly higher than the experimental estimate. This indicates that once two strands of a ß-sheet form a hairpin, the addition of further strands is essentially a downhill process. This is in agreement with observations from experiments on both DPDP and a related 4-stranded ß-sheet, and has important implications for the formation of amyloid fibrils and other extended ß-sheet structures. The exact cause of this cooperativity is unclear, but may arise from the fact that once one hairpin is formed, the entropy cost of fixing the central strand has already been paid, so all that remains is to add the remaining strand.

In addition to cooperativity, DPDP provided an opportunity to study the underlying forces of hairpin stability. In Trpzip2, both the turn and cross-strand hydrophobic contacts were observed to contribute to stability. DPDP has two component hairpins, of which the C-terminal hairpin was observed to be the more stable of the two. Several mutations of DPDP were made to assess both turn and side chain contributions to individual hairpin (as well as overall) stability. The two most stabilizing mutations involved moving a hydrophobic residue to the center strand, improving hydrophobic contacts, and optimization of the N-terminal turn region. Addition of a salt bridge to the N-terminal hairpin was not found to have any affect on stability. The dual studies of Trpzip2 and DPDP ultimately give a larger picture of ß-sheet systems; the turn region and hydrophobic contacts are important for both folding and stability, with backbone hydrogen bonds serving largely to stabilize folded structures rather than driving anything. In particular, persistent hydrogen bond formation seems dependent upon adjacent hydrophobic pairing, although more study needs to be done to confirm this general result.

Since the computational studies of Trpzip2 and DPDP were done using a generalized Born (GB) implicit solvent model (GBHCT, which was known to have certain deficiencies), it then made sense to perform an in-depth analysis of the accuracy and precision of the GB model. Although these deficiencies were largely accounted for by using a force field specifically designed for and tested on ß-type systems in implicit solvent, it was desirable to quantify errors in the solvent model for future studies. In this way it may be possible to avoid adjusting force field parameters to correct for solvent model inadequacies, as this approach may not be extensible to other systems (for example the correction force field used for Trpzip2 and DPDP is likely not transferable to a-helical systems).

The performance of the GB model and two others was assessed by comparison to another implicit model based on the Poisson equation (PE) and the TIP3P explicit solvent model (via use of thermodynamic integration calculations). In terms of reproducing explicit solvent results, the implicit model based on PE was found to perform the best, whereas the GB models all showed deficiencies. In particular, certain GB models were found to have a bias towards forming a-helical structures. It should be noted here that the errors of the GBHCT model (used in the Trpzip2 and DPDP studies) for ß-hairpins were

significantly less than those for a-helices. Attempts to correct the GB model by fitting to results from PE showed little overall improvement. The results suggest that the ability of GB implicit models to provide quantitative data may be limited, and that uncorrected simulations will only provide qualitative results.

Though the work that has been presented here encompasses several years of study, it exposes but a mere fraction of the secrets that the field of structural biology has to yield. Advances in computational power and techniques will no doubt increase our knowledge by leaps and bounds in the coming years. Just as the work done by countless others has made this work possible, it is hoped that this work will someday provide the foundation for a greater understanding of protein folding and solvation.

References

1. Voet, D. and J.G. Voet, *Biochemistry*. 2nd ed. 1995, New York: J. Wiley & Sons. xvii, 1361.

2. Southall, N.T., K.A. Dill, and A.D.J. Haymet, *A view of the hydrophobic effect.* Journal of Physical Chemistry B, 2002. **106**(3): p. 521-533.

3. Humphrey, W., A. Dalke, and K. Schulten, *VMD: Visual molecular dynamics.* Journal of Molecular Graphics, 1996. **14**(1): p. 33-&.

4. Pauling, L. and R.B. Corey, *2 Hydrogen-Bonded Spiral Configurations of the Polypeptide Chain.* Journal of the American Chemical Society, 1950. **72**(11): p. 5349-5349.

5. Pauling, L. and R.B. Corey, *The Pleated Sheet, a New Layer Configuration of Polypeptide Chains.* Proceedings of the National Academy of Sciences of the United States of America, 1951. **37**(5): p. 251-256.

6. Kendrew, J.C., G. Bodo, H.M. Dintzis, R.G. Parrish, H. Wyckoff, and D.C. Phillips, *3-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis.* Nature, 1958. **181**(4610): p. 662-666.

7. Anfinsen, C.B., *Principles That Govern the Folding of Protein Chains.* Science, 1973. **181**(96): p. 223-230.

8. Stefani, M., *Protein misfolding and aggregation: new examples in medicine and biology of the dark side of the protein world.* Biochimica Et Biophysica Acta-Molecular Basis of Disease, 2004. **1739**(1): p. 5-25.

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

10. Flamant, F., K. Gauthier, and J. Samarut, *Thyroid hormones signaling is getting more complex: STORMs are coming.* Molecular Endocrinology, 2007. **21**(2): p. 321-333.

11. Gonzalez-Ruiz, D. and H. Gohlke, *Targeting protein-protein interactions with small molecules: Challenges and perspectives for computational binding epitope detection and ligand finding.* Current Medicinal Chemistry, 2006. **13**(22): p. 2607-2625.

12. Levinthal, C., *Are There Pathways for Protein Folding.* Journal De Chimie Physique Et De Physico-Chimie Biologique, 1968. **65**(1): p. 44-&.

13. Kubelka, J., J. Hofrichter, and W.A. Eaton, *The protein folding 'speed limit'.* Current Opinion in Structural Biology, 2004. **14**(1): p. 76-88.

14. Leopold, P.E., M. Montal, and J.N. Onuchic, *Protein Folding Funnels - a Kinetic Approach to the Sequence Structure Relationship.* Proceedings of the National Academy of Sciences of the United States of America, 1992. **89**(18): p. 8721-8725.

15. Creighton, T.E., *Proteins : structures and molecular properties*. 2nd ed. 1993, New York: W.H. Freeman. xiii, 507 p.

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

17. Campbell, I.D., *The march of structural biology.* Nature Reviews Molecular Cell Biology, 2002. **3**(5): p. 377-381.

18. Berman, H.M., J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne, *The Protein Data Bank.* Nucleic Acids Research, 2000. **28**(1): p. 235-242.

19. Eyal, E., S. Gerzon, V. Potapov, M. Edelman, and V. Sobolev, *The limit of accuracy of protein modeling: Influence of crystal packing on protein structure.* Journal of Molecular Biology, 2005. **351**(2): p. 431-442.

20. Rahman, A. and Stilling.Fh, *Molecular Dynamics Study of Liquid Water.* Journal of Chemical Physics, 1971. **55**(7): p. 3336-&.

21. Mccammon, J.A., B.R. Gelin, and M. Karplus, *Dynamics of Folded Proteins.* Nature, 1977. **267**(5612): p. 585-590.

22. Daura, X., *Molecular dynamics simulation of peptide folding.* Theoretical Chemistry Accounts, 2006. **116**(1-3): p. 297-306.

23. Leach, A.R., *Molecular modelling : principles and applications.* 2nd ed. 2001, Harlow, England ; New York: Prentice Hall. xxiv, 744 p., [16] p. of plates.

24. Raugei, S., F.L. Gervasio, and P. Carloni, *DFT modeling of biological systems.* Physica Status Solidi B-Basic Solid State Physics, 2006. **243**(11): p. 2500-2515.

25. Bayly, C.I., P. Cieplak, W.D. Cornell, and P.A. Kollman, *A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges - the Resp Model.* Journal of Physical Chemistry, 1993. **97**(40): p. 10269-10280.

26. White, S.H., *The progress of membrane protein structure determination.* Protein Science, 2004. **13**(7): p. 1948-1949.

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

28. Greulich, K.O., *Single molecule techniques for biomedicine and pharmacology.* Current Pharmaceutical Biotechnology, 2004. **5**(3): p. 243-259.

29. Brunger, A.T. and P.D. Adams, *Molecular dynamics applied to X-ray structure refinement*. Accounts of Chemical Research, 2002. **35**(6): p. 404-412.

30. Wickstrom, L., Y. Bi, V. Hornak, D.P. Raleigh, and C. Simmerling, *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.

31. Alonso, H., A.A. Bliznyuk, and J.E. Gready, *Combining docking and molecular dynamic simulations in drug design.* Medicinal Research Reviews, 2006. **26**(5): p. 531-568.

32. Allen, T.W., O.S. Andersen, and B. Roux, *Molecular dynamics - potential of mean force calculations as a tool for understanding ion permeation and selectivity in narrow channels.* Biophysical Chemistry, 2006. **124**(3): p. 251-267.

33. Sorin, E.J., Y.M. Rhee, M.R. Shirts, and V.S. Pande, *The solvation interface is a determining factor in peptide conformational preferences.* J Mol Biol, 2006. **356**(1): p. 248-56.

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

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

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

37. Rathore, N., M. Chopra, and J.J. de Pablo, *Optimal allocation of replicas in parallel tempering simulations.* Journal of Chemical Physics, 2005. **122**(2): p. -.

38. Kone, A. and D.A. Kofke, *Selection of temperature intervals for parallel-tempering simulations.* Journal of Chemical Physics, 2005. **122**(20): p. -.

39. Zhang, W., C. Wu, and Y. Duan, *Convergence of replica exchange molecular dynamics.* Journal of Chemical Physics, 2005. **123**(15): p. -.

40. Kim, P.S. and R.L. Baldwin, *Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding.* Annu Rev Biochem, 1982. **51**: p. 459-89.

41. Karplus, M. and D.L. Weaver, *Protein-folding dynamics.* Nature, 1976. **260**(5550): p. 404-6.

42. Fersht, A.R., *Transition-state structure as a unifying basis in protein-folding mechanisms: Contact order, chain topology, stability, and the extended nucleus mechanism.* Proceedings of the National Academy of Sciences of the United States of America, 2000. **97**(4): p. 1525-1529.

43. Osterhout, J.J., *Understanding protein folding through peptide models.* Protein Pept Lett, 2005. **12**(2): p. 159-64.

44. Stotz, C.E. and E.M. Topp, *Applications of model beta-hairpin peptides.* Journal of Pharmaceutical Sciences, 2004. **93**(12): p. 2881-2894.

45. Cho, J.H., S. Sato, and D.P. Raleigh, *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.

46. Neidigh, J.W., R.M. Fesinmeyer, and N.H. Andersen, *Designing a 20-residue protein.* Nature Structural Biology, 2002. **9**(6): p. 425-430.

47. Cochran, A.G., N.J. Skelton, and M.A. Starovasnik, *Tryptophan zippers: Stable, monomeric beta-hairpins.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **98**(13): p. 9081-9081.

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

49. Roe, D.R., A. Okur, L. Wickstrom, V. Hornak, and C. Simmerling, *Secondary Structure Bias in Generalized Born Solvent Models: Comparison of Conformational Ensembles and Free Energy of Solvent Polarization from Explicit and Implicit Solvation.* J Phys Chem B Condens Matter Mater Surf Interfaces Biophys, 2007. **111**(7): p. 1846-1857.

50. Bonvin, A.M.J.J. and W.F. van Gunsteren, *beta-Hairpin stability and folding: Molecular dynamics studies of the first beta-hairpin of tendamistat.* Journal of Molecular Biology, 2000. **296**(1): p. 255-268.

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

52. Pande, V.S. and D.S. Rokhsar, *Molecular dynamics simulations of unfolding and refolding of a beta-hairpin fragment of protein G.* Proceedings of the National Academy of Sciences of the United States of America, 1999. **96**(16): p. 9062-9067.

53. Dinner, A.R., T. Lazaridis, and M. Karplus, *Understanding beta-hairpin formation.* Proceedings of the National Academy of Sciences of the United States of America, 1999. **96**(16): p. 9068-9073.

54. Klimov, D.K. and D. Thirumalai, *Mechanisms and kinetics of beta-hairpin formation.* Proceedings of the National Academy of Sciences of the United States of America, 2000. **97**(6): p. 2544-2549.

55. Zhou, R.H., B.J. Berne, and R. Germain, *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.

56. Cochran, A.G., N.J. Skelton, and M.A. Starovasnik, *Tryptophan zippers: Stable, monomeric beta-hairpins.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(10): p. 5578-5583.

57. Pearlman, D.A., D.A. Case, J.W. Caldwell, W.S. Ross, T.E. Cheatham, S. Debolt, D. Ferguson, G. Seibel, and P. Kollman, *Amber, a Package of Computer-Programs for Applying Molecular Mechanics, Normal-Mode Analysis, Molecular-Dynamics and Free-Energy Calculations to Simulate the Structural and Energetic Properties of Molecules.* Computer Physics Communications, 1995. **91**(1-3): p. 1-41.

58. Berendsen, H.J.C., J.P.M. Postma, W.F. Vangunsteren, A. Dinola, and J.R. Haak, *Molecular-Dynamics with Coupling to an External Bath.* Journal of Chemical Physics, 1984. **81**(8): p. 3684-3690.

59. Ryckaert, J.P., G. Ciccotti, and H.J.C. Berendsen, *Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes.* Journal of Computational Physics, 1977. **23**(3): p. 327-341.

60. Hawkins, G.D., C.J. Cramer, and D.G. Truhlar, *Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium.* Journal of Physical Chemistry, 1996. **100**(51): p. 19824-19839.

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

62. Cornell, W.D., P. Cieplak, C.I. Bayly, I.R. Gould, K.M. Merz, D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell, and P.A. Kollman, *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.

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

64. Yang, W.Y. and M. Gruebele, *Detection-dependent kinetics as a probe of folding landscape microstructure.* Journal of the American Chemical Society, 2004. **126**(25): p. 7758-7759.

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

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

67. Fersht, A.R., *On the simulation of protein folding by short time scale molecular dynamics and distributed computing.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(22): p. 14122-14125.

68. Zwanzig, R., *Two-state models of protein folding kinetics.* Proceedings of the National Academy of Sciences of the United States of America, 1997. **94**(1): p. 148-150.

69. Thirumalai, D., D.K. Klimov, and S.A. Woodson, *Kinetic partitioning mechanism as a unifying theme in the folding of biomolecules.* Theoretical Chemistry Accounts, 1997. **96**(1): p. 14-22.

70. Matagne, A., S.E. Radford, and C.M. Dobson, *Fast and slow tracks in lysozyme folding: Insight into the role of domains in the folding process.* Journal of Molecular Biology, 1997. **267**(5): p. 1068-1074.

71. Nymeyer, H. and A.E. Garcia, *Simulation of the folding equilibrium of alpha-helical peptides: A comparison of the generalized born approximation with explicit solvent.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(24): p. 13934-13939.

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

73. Zhu, J., E. Alexov, and B. Honig, *Comparative study of generalized Born models: Born radii and peptide folding.* Journal of Physical Chemistry B, 2005. **109**(7): p. 3008-3022.

74. Okur, A., L. Wickstrom, M. Layten, R. Geney, K. Song, V. Hornak, and C. Simmerling, *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.

75. Chen, J.H., W.P. Im, and C.L. Brooks, *Balancing solvation and intramolecular interactions: Toward a consistent generalized born force field.* Journal of the American Chemical Society, 2006. **128**(11): p. 3728-3736.

76. Wouters, M.A. and P.M.G. Curmi, *An Analysis of Side-Chain Interactions and Pair Correlations within Antiparallel Beta-Sheets - the Differences between Backbone Hydrogen-Bonded and Non-Hydrogen-Bonded Residue Pairs.* Proteins-Structure Function and Genetics, 1995. **22**(2): p. 119-131.

77. deAlba, E., M.A. Jimenez, and M. Rico, *Turn residue sequence determines beta-hairpin conformation in designed peptides.* Journal of the American Chemical Society, 1997. **119**(1): p. 175-183.

78. Munoz, V. and L. Serrano, *Elucidating the Folding Problem of Helical Peptides Using Empirical Parameters .2. Helix Macrodipole Effects and Rational Modification of the Helical Content of Natural Peptides.* Journal of Molecular Biology, 1995. **245**(3): p. 275-296.

79. Stapley, B.J. and A.J. Doig, *Hydrogen bonding interactions between glutamine and asparagine in alpha-helical peptides.* Journal of Molecular Biology, 1997. **272**(3): p. 465-473.

80. Wieczorek, R. and J.J. Dannenberg, *H-bonding cooperativity and energetics of alpha-helix formation of five 17-amino acid peptides.* Journal of the American Chemical Society, 2003. **125**(27): p. 8124-8129.

81. Schenck, H.L. and S.H. Gellman, *Use of a designed triple-stranded antiparallel beta-sheet to probe beta-sheet cooperativity in aqueous solution.* Journal of the American Chemical Society, 1998. **120**(19): p. 4869-4870.

82. Stanger, H.E., F.A. Syud, J.F. Espinosa, I. Giriatt, T. Muir, and S.H. Gellman, *Length-dependent stability and strand length limits in antiparallel beta-sheet secondary structure.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(21): p. 12015-12020.

83. Guo, C.L., M.S. Cheung, H. Levine, and D.A. Kessler, *Mechanisms of cooperativity underlying sequence-independent beta-sheet formation.* Journal of Chemical Physics, 2002. **116**(10): p. 4353-4365.

84. De Alba, E., J. Santoro, M. Rico, and M.A. Jimenez, *De novo design of a monomeric three-stranded antiparallel beta-sheet.* Protein Science, 1999. **8**(4): p. 854-865.

85. Sharman, G.J. and M.S. Searle, *Cooperative interaction between the three strands of a designed antiparallel beta-sheet.* Journal of the American Chemical Society, 1998. **120**(21): p. 5291-5300.

86. Griffiths-Jones, S.R. and M.S. Searle, *Structure, folding, and energetics of cooperative interactions between the beta-strands of a de novo designed three-stranded antiparallel beta-sheet peptide.* Journal of the American Chemical Society, 2000. **122**(35): p. 8350-8356.

87. Syud, F.A., H.E. Stanger, H.S. Mortell, J.F. Espinosa, J.D. Fisk, C.G. Fry, and S.H. Gellman, *Influence of strand number on antiparallel beta-sheet stability in designed three- and four-stranded beta-sheets.* Journal of Molecular Biology, 2003. **326**(2): p. 553-568.

88. Haque, T.S., J.C. Little, and S.H. Gellman, *Stereochemical requirements for beta-hairpin formation: Model studies with four-residue peptides and depsipeptides.* Journal of the American Chemical Society, 1996. **118**(29): p. 6975-6985.

89. Haque, T.S., J.C. Little, and S.H. Gellman, *Mirror-Image Reverse Turns Promote Beta-Hairpin Formation.* Journal of the American Chemical Society, 1994. **116**(9): p. 4105-4106.

90. Santiveri, C.M., J. Santoro, M. Rico, and M.A. Jimenez, *Factors involved in the stability of isolated beta-sheets: Turn sequence, beta-sheet twisting, and hydrophobic surface burial.* Protein Science, 2004. **13**(4): p. 1134-1147.

91. Kuznetsov, S.V., J. Hilario, T.A. Keiderling, and A. Ansari, *Spectroscopic studies of structural changes in two beta-sheet-forming peptides show an ensemble of structures that unfold noncooperatively.* Biochemistry, 2003. **42**(15): p. 4321-4332.

92. Tang, Y.F., D.J. Rigotti, R. Fairman, and D.P. Raleigh, *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.

93. Syud, F.A., J.F. Espinosa, and S.H. Gellman, *NMR-based quantification of beta-sheet populations in aqueous solution through use of reference peptides for the folded and unfolded states.* Journal of the American Chemical Society, 1999. **121**(49): p. 11577-11578.

94. Tsui, V. and D.A. Case, *Molecular dynamics simulations of nucleic acids with a generalized born solvation model.* Journal of the American Chemical Society, 2000. **122**(11): p. 2489-2498.

95. Zhou, R.H. and B.J. Berne, *Can a continuum solvent model reproduce the free energy landscape of a beta-hairpin folding in water?* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(20): p. 12777-12782.

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

97. Zagrovic, B., E.J. Sorin, and V. Pande, *beta-hairpin folding simulations in atomistic detail using an implicit solvent model.* Journal of Molecular Biology, 2001. **313**(1): p. 151-169.

98. Paci, E., A. Cavalli, M. Vendruscolo, and A. Caflisch, *Analysis of the distributed computing approach applied to the folding of a small beta peptide.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(14): p. 8217-8222.

99. Snow, C.D., L.L. Qiu, D.G. Du, F. Gai, S.J. Hagen, and V.S. Pande, *Trp zipper folding kinetics by molecular dynamics and temperature-jump spectroscopy.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(12): p. 4077-4082.

100. Bursulaya, B.D. and C.L. Brooks, *Folding free energy surface of a three-stranded beta-sheet protein.* Journal of the American Chemical Society, 1999. **121**(43): p. 9947-9951.

101. Colombo, G., D. Roccatano, and A.E. Mark, *Folding and stability of the three-stranded beta-sheet peptide betanova: Insights from molecular dynamics simulations.* Proteins-Structure Function and Genetics, 2002. **46**(4): p. 380-392.

102. Chen, P.Y., C.K. Lin, C.T. Lee, H. Jan, and S.I. Chan, *Effects of turn residues in directing the formation of the beta-sheet and in the stability of the beta-sheet.* Protein Science, 2001. **10**(9): p. 1794-1800.

103. Ma, B.Y. and R. Nussinov, *Molecular dynamics simulations of a beta-hairpin fragment of protein G: Balance between side-chain and backbone forces.* Journal of Molecular Biology, 2000. **296**(4): p. 1091-1104.

104. Irback, A. and F. Sjunnesson, *Folding thermodynamics of three beta-sheet peptides: A model study.* Proteins-Structure Function and Bioinformatics, 2004. **56**(1): p. 110-116.

105. Wang, H.W. and S.S. Sung, *Molecular dynamics simulations of three-strand beta-sheet folding.* Journal of the American Chemical Society, 2000. **122**(9): p. 1999-2009.

106. Levitt, M., *Conformational Preferences of Amino-Acids in Globular Proteins.* Biochemistry, 1978. **17**(20): p. 4277-4284.

107. Ciani, B., M. Jourdan, and M.S. Searle, *Stabilization of beta-hairpin peptides by salt bridges: Role of preorganization in the energetic contribution of weak interactions.* Journal of the American Chemical Society, 2003. **125**(30): p. 9038-9047.

108. Tsai, J. and M. Levitt, *Evidence of turn and salt bridge contributions to beta-hairpin stability: MD simulations of C-terminal fragment from the B1 domain of protein G.* Biophysical Chemistry, 2002. **101**: p. 187-201.

109. Dill, K.A. and D. Shortle, *Denatured States of Proteins.* Annual Review of Biochemistry, 1991. **60**: p. 795-825.

110. Chou, P.Y. and G.D. Fasman, *Conformational Parameters for Amino-Acids in Helical, Beta-Sheet, and Random Coil Regions Calculated from Proteins.* Biochemistry, 1974. **13**(2): p. 211-222.

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

112. Yu, Z.Y., M.P. Jacobson, J. Josovitz, C.S. Rapp, and R.A. Friesner, *First-shell solvation of ion pairs: Correction of systematic errors in implicit solvent models.* Journal of Physical Chemistry B, 2004. **108**(21): p. 6643-6654.

113. Felts, A.K., Y. Harano, E. Gallicchio, and R.M. Levy, *Free energy surfaces of beta-hairpin and alpha-helical peptides generated by replica exchange molecular dynamics with the AGBNP implicit solvent model.* Proteins-Structure Function and Bioinformatics, 2004. **56**(2): p. 310-321.

114. Feig, M. and C.L. Brooks, *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.

115. Levy, R.M. and E. Gallicchio, *Computer simulations with explicit solvent: Recent progress in the thermodynamic decomposition of free energies and in modeling electrostatic effects.* Annual Review of Physical Chemistry, 1998. **49**: p. 531-567.

116. Roux, B. and T. Simonson, *Implicit solvent models.* Biophysical Chemistry, 1999. **78**(1-2): p. 1-20.

117. Gilson, M.K., M.E. Davis, B.A. Luty, and J.A. Mccammon, *Computation of Electrostatic Forces on Solvated Molecules Using the Poisson-Boltzmann Equation.* Journal of Physical Chemistry, 1993. **97**(14): p. 3591-3600.

118. Baker, N.A., *Improving implicit solvent simulations: a Poisson-centric view.* Current Opinion in Structural Biology, 2005. **15**(2): p. 137-143.

119. Luo, R., L. David, and M.K. Gilson, *Accelerated Poisson-Boltzmann calculations for static and dynamic systems.* Journal of Computational Chemistry, 2002. **23**(13): p. 1244-1253.

120. Prabhu, N.V., P.J. Zhu, and K.A. Sharp, *Implementation and testing of stable, fast implicit solvation in molecular dynamics using the smooth-permittivity finite difference Poisson-Boltzmann method.* Journal of Computational Chemistry, 2004. **25**(16): p. 2049-2064.

121. Honig, B. and A. Nicholls, *Classical Electrostatics in Biology and Chemistry.* Science, 1995. **268**(5214): p. 1144-1149.

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

123. Bashford, D. and D.A. Case, *Generalized born models of macromolecular solvation effects.* Annual Review of Physical Chemistry, 2000. **51**: p. 129-152.

124. Rizzo, R.C., T. Aynechi, D.A. Case, and I.D. Kuntz, *Estimation of absolute free energies of hydration using continuum methods: Accuracy of partial, charge models and optimization of nonpolar contributions.* Journal of Chemical Theory and Computation, 2006. **2**(1): p. 128-139.

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

126. Onufriev, A., D. Case, and D. Bashford, *Effective Born radii in the generalized Born approximation: The importance of being perfect.* J COMPUT CHEM, 2002. **23**(14): p. 1297-1304.

127. Srinivasan, J., M.W. Trevathan, P. Beroza, and D.A. Case, *Application of a pairwise generalized Born model to proteins and nucleic acids: inclusion of salt effects.* Theoretical Chemistry Accounts, 1999. **101**(6): p. 426-434.

128. Jorgensen, W.L., J. Chandrasekhar, J.D. Madura, R.W. Impey, and M.L. Klein, *Comparison of Simple Potential Functions for Simulating Liquid Water.* Journal of Chemical Physics, 1983. **79**: p. 926-935.

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

130. Mongan, J., C. Simmerling, J.A. Mccammon, D.A. Case, and A. Onufriev, *Generalized Born model with a simple, robust molelcular volume correction.* Journal of Chemical Theory and Computation, 2006. **In Press**.

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

132. Showalter, S.A. and R. Brüschweiler, *Validation of Molecular Dynamics Simulations of Biomolecules Using NMR Spin Relaxation as Benchmarks: Application to the AMBER99SB Force Field.* Journal of Chemical Theory and Computation, 2007. **In Press**.

133. Essmann, U., L. Perera, M.L. Berkowitz, T. Darden, H. Lee, and L.G. Pedersen, *A Smooth Particle Mesh Ewald Method.* Journal of Chemical Physics, 1995. **103**(19): p. 8577-8593.

134. Jayaram, B., Y. Liu, and D.L. Beveridge, *A modification of the generalized Born theory for improved estimates of solvation energies and pK shifts.* Journal of Chemical Physics, 1998. **109**(4): p. 1465-1471.

135. Nicholls, A. and B. Honig, *A Rapid Finite-Difference Algorithm, Utilizing Successive over- Relaxation to Solve the Poisson-Boltzmann Equation.* Journal of Computational Chemistry, 1991. **12**(4): p. 435-445.

136. Schaefer, M. and C. Froemmel, *A Precise Analytical Method for Calculating the Electrostatic Energy of Macromolecules in Aqueous-Solution.* Journal of Molecular Biology, 1990. **216**(4): p. 1045-1066.

137. Qiu, D., P.S. Shenkin, F.P. Hollinger, and W.C. Still, *The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii.* Journal of Physical Chemistry A, 1997. **101**(16): p. 3005-3014.

138. Sigalov, G., P. Scheffel, and A. Onufriev, *Incorporating variable dielectric environments into the generalized Born model.* Journal of Chemical Physics, 2005. **122**(9): p. -.

139. Kabsch, W. and C. Sander, *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.* Biopolymers, 1983. **22**(12): p. 2577-637.

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

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

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

143. Makowska, J., S. Rodziewicz-Motowidlo, K. Baginska, J.A. Vila, A. Liwo, L. Chmurzynski, and H.A. Scheraga, *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 Sciences of the United States of America, 2006. **103**(6): p. 1744-1749.

144. Bhattacharyya, S.M., Z.G. Wang, and A.H. Zewail, *Dynamics of water near a protein surface.* Journal of Physical Chemistry B, 2003. **107**(47): p. 13218-13228.

145. Russo, D., G. Hura, and T. Head-Gordon, *Hydration dynamics near a model protein surface.* Biophysical Journal, 2004. **86**(3): p. 1852-1862.

146. Kollman, P., *Free Energy Calculations - Applications to Chemical and Biochemical Phenomena.* Chemical Reviews, 1993. **93**(7): p. 2395-2417.

147. Swanson, J.M.J., S.A. Adcock, and J.A. McCammon, *Optimized radii for Poisson-Boltzmann calculations with the AMBER force field.* Journal of Chemical Theory and Computation, 2005. **1**(3): p. 484-493.

148. Stultz, C.M., *An assessment of potential of mean force calculations with implicit solvent models.* Journal of Physical Chemistry B, 2004. **108**(42): p. 16525-16532.