Computational Structure Prediction and Folding Pathway Studies of the Trp-Cage Protein

A Dissertation Presented

By

Bentley Strockbine

To

The Graduate School

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

In

Molecular and Cellular Pharmacology

Stony Brook University

August 2005

Stony Brook University
The Graduate School

Bentley Andrew Strockbine

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

Carlos Simmerling, Ph.D., Associate Professor, Advisor
Department of Chemistry

Moises Eisenberg, Ph.D., Professor, Committee Chair
Department of Pharmacological Sciences

Caroline Kisker, Ph.D., Associate Professor, Committee Member
Department of Pharmacological Sciences

Carlos de los Santos, Ph.D., Associate Professor, Committee Member
Department of Pharmacological Sciences

Robert Rizzo, Ph.D., Assistant Professor, Outside Committee Member
Department of Applied Math and Statistics at Stony Brook University

This Dissertation is accepted by the Graduate School.

Dean of the Graduate School

Abstract of the Dissertation

Computational Protein Folding Studies with Implicit and Explicit Solvent Models

by

Bentley Strockbine

in

Molecular and Cellular Pharmacology
Stony Brook University
2005

The simplicity of the composition of proteins belies the complexity of their structure. Intense effort from the scientific community has been spent on improving experimental and theoretical methods to determine the native structure of proteins and model folding pathways to aid in the understanding of how proteins fold to their native state.

For molecular mechanical tools to be useful, they must accurately evaluate the relative energies of different structures. In the first part of this dissertation we present a modified parameter set for the AMBER molecular modeling package aimed at improving predictions of the relative energies of alternate protein conformations.

Next we describe results from using these parameters in all-atom fully unrestrained *ab initio* folding simulations for trp-cage, a stable protein with non-trivial secondary structure elements and a hydrophobic core. The first successful prediction of the atomic-resolution structure of a protein (prior to release of experimental data) is presented. The predicted structure displays features that are suggested by experimental data, yet are not evident in NMR derived family of structures.

Last, we will discuss the details of the progression of events during the folding of the trp-cage protein in reproducible unrestrained folding simulations with explicit inclusion of solvent molecules and extend our previous results to the study of the folding pathway. A specific partially folded intermediate is described and the results are compared directly to the available experimental data.

For my loving, patient, family.

Table of Contents

List of Figures

List of Tables and Equations

Acknowledgements

Thank you, thank you, thank you to all those who, despite my best efforts, helped me achieve this goal. I cannot possibly thank all of you here individually, just know that I appreciate all of the help I was given, whether I graciously accepted it or not.

To my advisor, Dr. Carlos Simmerling, who patiently endured my constant interruptions and endless questions, thank you, this work would literally not have happened without your support.

Thanks to each of my committee members for your patience, your advice, and your time. Thank you Dr. Eisenberg for teaching me about the beauty of mathematics and for all of the stories that I certainly cannot repeat here. Thank you Dr. Kisker for introducing me to structural biology. Thank you Dr. de los Santos for your willingness to join my committee. Thank you Dr. Rizzo for both joining my committee and for ensuring my future in science after graduation.

I would also like to thank the past and present members of the Simmerling lab who safeguarded my sanity (almost) and for making the lab a home. In particular I would like to thank Asim Okur who sat through hours of conversation, gallons of coffee and become one of my best friends anyway. Thanks to Dr. Cui who was with me at the beginning and explained what “cd” stands for. Thanks to Kerri Goldgraben for worrying enough for both of us.

Thanks to Dr. Panebianco and the Running Fevers, I forget how many miles we ran but I won’t forget the time we spent together.

Lastly, and most importantly, I would like to thank my loving family for their encouragement and unwavering support. Grandma German, thanks for all the pirogies, I miss you every day.

Chapter 1. Introduction

1.1 The Importance of Protein Structure

Proteins are fundamental biological macromolecules that perform many essential functions. Every protein can be assembled from a limited set of 20 different amino acids. The simplicity of their composition belies the complexity of their structure and function. Understanding the structure of a protein is essential for understanding the mechanics of the functioning of the protein. Understanding the functioning of proteins is a necessary prerequisite to understanding ourselves.

1.1.1 Proteins' Role in Cellular Processes

Proteins play a myriad of roles in cellular processes including enzymatic catalysis, transport, immune recognition, cellular control, mechanical structure, growth, replication, communication, and differentiation. Because proteins play so many roles and because they are so central to so many of the functions of a cell, it is of our utmost interest to understand them.

Understanding proteins is also central to understanding of many disease processes[1]. Essentially every disease, in some way, involves protein function or malfunction.

1.1.2 Elements of Protein Structure

Proteins are made up of one or more polymeric macromolecules consisting of linear assemblies of amino acids. Amino acids are composed of a central carbon atom, called the α-carbon, that is attached to a hydrogen atom, an amino group, a carboxyl group, and a variable side chain that is commonly referred to as an R-group. The carboxyl group of one amino acid is connected to the amino group of the next amino acid by a peptide bond. The numbering of the sequence of amino acids traditionally starts at the free amino terminus.

Because carbon atoms are tetravalent there can be two absolute spatial conformations of the groups attached to the α-carbon. The two isomeric conformations are called the l-isomer and the d-isomer. The "l" and "d" refer to the levorotary and dextrorotary optical activity of the different isomers. With few natural exceptions, only l-isomer amino acids are found in biologically relevant proteins.

Traditionally, protein structure is divided into four levels; primary, secondary, tertiary, and quaternary. The primary structure is the linear order of the amino acids that comprise the polypeptide. The primary structure of a protein contains all the information necessary to determine the other levels of protein structure[2].

The secondary structure describes small repeating elements of structure that are usually held together with hydrogen bonds[3]. The two most common elements of secondary structure are the α helices and β sheets[4]. α Helices are defined by their tight right handed coil-like structure and hydrogen bonds between the carbonyl oxygen of residue i and the amide hydrogen of residue i+4. Thus each main chain carboxyl and amine group of an α helix participates in hydrogen bonding. In β sheets the backbone is in a linear conformation and the hydrogen bonds are made between different strands of the peptide that do not need to be sequentially local to each other. The strands that form a β sheet can be either parallel in sequence or anti-parallel; in both forms each backbone carbonyl and amine group is involved in an interstrand hydrogen bond. There are several other less common forms of secondary structure including 3_{10} helices, π helices, and polyproline helices[4].

Intermediate autonomously folding elements between secondary structure and tertiary structure are called domains. Some domains are capable of a differentiable portion of the activity of a protein.

The tertiary structure of a protein is the global three dimensional structure of an individual polypeptide. For a protein to be functional it is generally required that the protein is folded into at least the tertiary level of structure. The functional tertiary structure of a protein is commonly called the native state. When a protein is not folded to the tertiary level of structure it is usually described as unfolded or denatured. Many proteins are composed of only one polypeptide so it is not uncommon for the tertiary structure to be the ultimate level of structure.

Some proteins are composed of several polypeptides and the quaternary structure represents the unified spatial arrangement off all the polypeptides required to form the complete protein. In the context of a protein that has more than one polypeptide, each peptide is called a subunit.

1.1.3 The Protein Folding Problem

The work of Anfinsen[2] in the 1960s showed that all the information necessary to determine the three-dimensional structure of a protein, in physiological conditions, is contained in the primary sequence. A corollary to this work known as Anfinsen's hypothesis, states that the native conformation of a protein in physiological conditions is the conformation with the lowest global free energy. This suggests the tantalizing possibility that if the relative free energies of all the different molecular conformations can be determined then the native structure of a protein can be identified.

The main conformational degrees of freedom in the backbone of a protein are the rotation around the main chain bonds on either side of the α carbon. These are the ϕ and φ angles and they are defined by the angular difference between the planes created by the first and last three atoms in the backbone series C-N-Cα-C and N-Cα-C-N respectively. Flory suggested that each pair of ϕ and φ angles is independent of the other pairs [5]. This implies that the number of conformations available to a polypeptide chain increases exponentially with the number of amino acids. The number of conformations available for even small proteins quickly becomes too large for a complete search of all the possible structures. This is true for *in vivo* protein folding as well as for *in silico* computational studies; Levinthal's paradox[6] states that there are many more possible structural states than a protein can visit in the time it has to fold. Levinthal concluded that proteins must fold by a sequence of events or pathway[7] that leads to the folded protein. Recent studies suggest that there may be more than one pathway, potentially many[8], which lead from a disordered state to a folded protein.

1.2 Structural Biology

1.2.1 Sources of Structural Information

Information in the field of structural biology can come from several sources. Each of these sources, whether they are experimental or theoretical, aim to add information about the structure and function of biomolecules. X-ray crystallography was the first technique used to determine the structure of a biological macromolecule. In 1957 John Kendrew used the technique to determine the structure of myoglobin. This seminal event can be used to mark the beginning of the field of structural biology[9].

Electromagnetic radiation is a key tool in most experimental methods used to derive structural information at the atomic level. The maximum resolution of an image is restricted by the wavelength of the radiation used to produce the image. The wavelength of visible light is too long to be used to determine the atomic level details of a biomolecule; techniques other than direct imaging must be used to derive biomolecular-structural information.

Two primary sources of structural information are x-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy. In x-ray crystallographic studies, highly pure crystals of a protein, or other molecules of interest, can be used to determine the structure of the molecule[10]. Because electrons scatter x-rays and the scattering can be related to the local density of the electrons, x-rays can be used to calculate a three dimensional map of the distribution of electron density. The density map can then be used to determine the position of the atomic nuclei, and thus the structure of a protein, in the repeating unit cells of the crystal.

NMR spectroscopy measures the energetic difference between atomic spin states in an applied magnetic field to provide information about number and environment of magnetically distinct atoms in the structure of the molecule being studied[11]. Correlation spectroscopy (COSY), which gives information about hydrogen atoms that are covalently connected by one or two atoms, and nuclear Overhauser effect (NOE) experiments, which give information about sequence remote hydrogen distances, can be used together to determine sets of restraints that describe the distances between hydrogen atoms that near each other in space. The Wüthrich sequential assignment technique[12] can then be used to recreate the relative locations of the hydrogens in the protein structure. Unlike x-ray crystallography, NMR studies can be conducted on molecules that are in solution. For biomolecules the solution state is generally a better approximation of the native environment of the molecule.

Further, the local environment around a magnetic nucleus generates very small magnetic fields that oppose the applied magnetic field. The counteractive effect of the local field relative to the applied magnetic field is called shielding, and this shielding shifts the resonance frequency of the absorbed energy. This shift, known as the chemical shift (δ), is measured and used to describe the local environment of the atoms.

There are many other techniques that provide important structural information including atomic force microscopy[13-15], infra-red spectroscopy[16, 17], circular dichroism[18] and fluorescence resonance energy transfer[19-21]. It is worth noting that structural information would be of little value without concomitant molecular and cellular biology techniques that provide the context for structural information.

1.2.2 Rationale for Computational Modeling

The majority of structural information that is gained about proteins from experiments is used to create static models. These models are averaged over many molecules and long time frames. This has the benefit of producing models that most likely represent the thermodynamically relevant structures. The disadvantages of these techniques are that proteins are not static and fast events in the life of a protein, often including folding, are too fast to be investigated with standard experimental techniques. Atoms in proteins are continually moving in small ways relative to each other and are also often moving in large global motions, in coordination with each other, that include folding, breathing motions, and domain motions.

Computational modeling can provide complimentary information that is often inaccessible by experiment alone. Because computational models can simulate the action of every atom in a protein and they necessarily take time steps that are smaller than the timescales of protein motions, they can provide information, at the atomic level, about dynamics of single molecules and about protein motions that would otherwise be too fast to observe.

Because the time steps in the calculations involved in computational modeling are on the order of femtoseconds, even relatively short simulations, on the order of nanoseconds, require a tremendous number of individual calculations. Also, the number of calculations is related to the number of pairs of atoms in the system that is being modeled. These facts limit the timescales available to computational modeling and the size of the systems that are feasible to model. Experimental techniques and computational techniques are well matched to complement each other. Where experimental techniques average over large numbers of molecules and relatively long timeframes, computational techniques generally investigate single molecules over relatively short timeframes.

Chapter 2. Molecular Modeling Overview

2.1 Molecular Dynamics

Molecular modeling is the use of mathematical models to describe and predict the actions of molecules[23, 24]. Molecular dynamics (MD) is a type of molecular modeling where atomic motions are described in the terms of Newtonian mechanics. In other words, molecular dynamics simulates the temporal evolution of a series of interacting atoms by solving the equations of motion. In a system of N particles of known masses m$_i$, where the particles positioned at $r_1, \ldots, r_N$ are affected by inter-particle interactions defined by the energy function *U(r1, ..., rN)* the force F_i on each particle can be determined. See equation 2.1.

Equation 2.1 Equation of Force $F_i = -\frac{\partial U(r_1, \ldots r_n)}{\partial r_i}$

As the force, F_i, acting on the particle and the mass, m_i, of the particle is known, Newton's second law of motion[25] can be used to determine the acceleration, a_i, and thus the velocity of each particle. See equation 2.2.

Equation 2.2 Newton's Second Law $F_i = m_i a_i$

In molecular dynamics it is assumed that if the time step is small enough (finite differences method[26]) the resultant trajectory will accurately represent a true molecular trajectory. Thus MD can be used to calculate a molecular trajectory based on the energy of the system.

2.1.1 The Force Field

To describe the interactions between the atoms, molecular mechanics treats the interactions between atoms with simple mechanical models (*i.e.* springs). Thus simple mathematical models (*i.e.* Hooke's Law) can be used to describe those interactions.

It is clear that for a molecular mechanics model to correctly identify the native conformation of a protein it has to be able to correctly determine the relative free energies of the conformations of the molecule. Toward this aim, molecular mechanics techniques rely on equations that determine the potential energy of a conformation of a molecule by summing the energies of the components and interactions that comprise the molecule. In particular, the Amber force field, a component of the AMBER[27] suite of molecular dynamics simulation tools (by convention "AMBER" is used to refer to the program suite while "Amber" is used to refer to the force field that is part of that suite), is a widely used force field that is composed of a molecular mechanical equation (see equation 2.3) and a set of constants, the parameters set, that are used together to calculate the energies of molecular conformations.

2.1.1.1 The Force Field Equation

The Amber[27] force field is based on a set of classical molecular mechanical functions for modeling the interactions in a molecular system. The force field has four main components: a bond stretching term, an angle bending term, a bond rotation (dihedral) term, and a term for electrostatic and van der Waals non-bonded interactions. See equation 2.3.

Equation 2.3 The Force Field Equation

$$U(r) = \sum_{bonds} K_r\left(r - r_{eq}\right)^2 + \sum_{angles} K_\theta\left(\theta - \theta_{eq}\right)^2 + \sum_{dihedrals} \frac{V_n}{2}\left(1 + \cos\left(n\phi - \gamma\right)\right) + \sum_{i<j} \left\{ 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{R_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{R_{ij}}\right)^{6}\right] + \frac{q_i q_j}{4\pi\varepsilon_0 R_{ij}} \right\}$$

The first term of the force field equation, the bond stretching term, defines the energetic consequence of deviation of the bond length, *r*, from the reference bond length, r_{eq}, with a functional form based on Hooke’s law. The reference bond length is the length at which the energetic consequence is zero. K_r is the force constant.

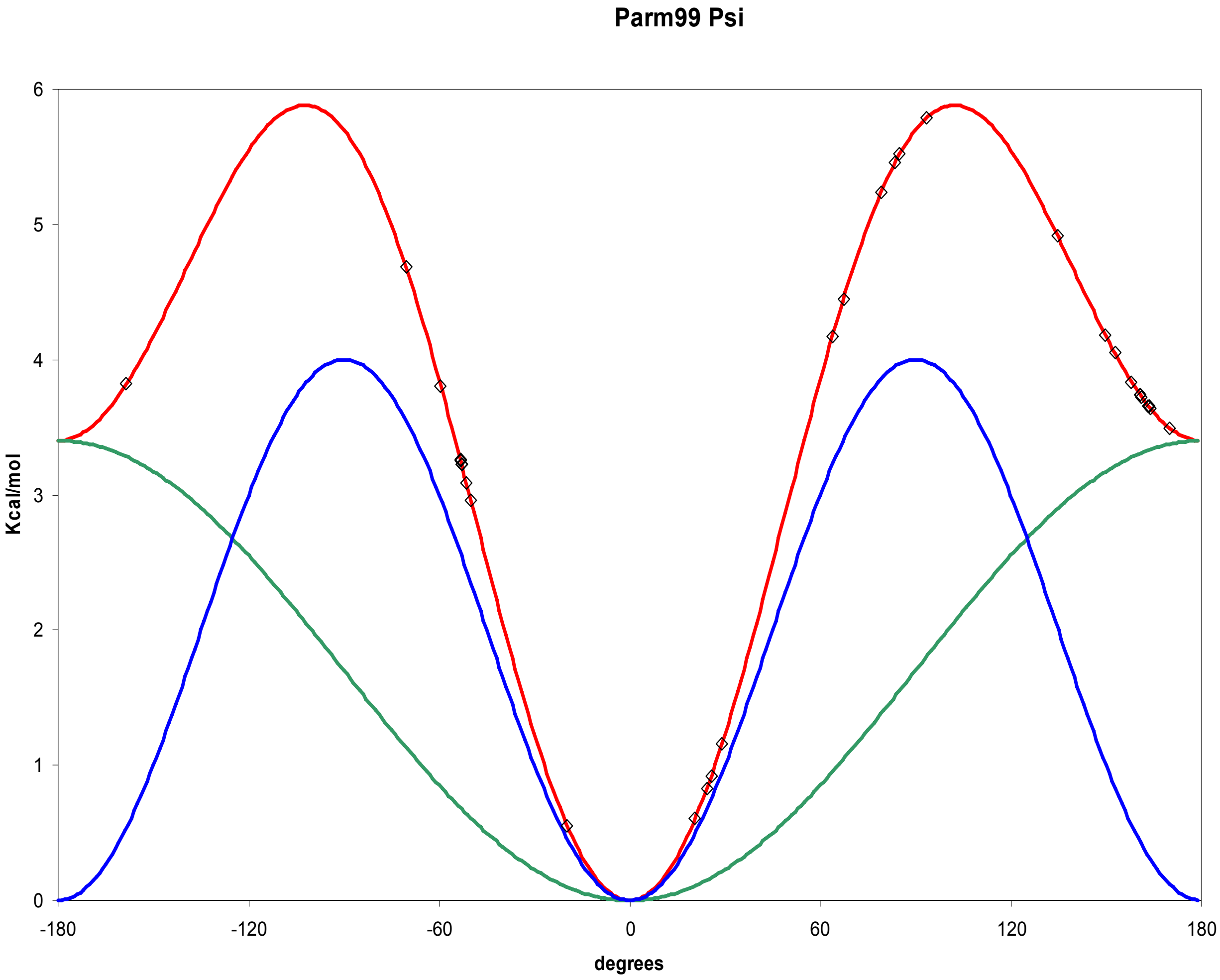
The second term, the angle term, again uses a Hooke’s law formula to determine the energetic consequence of angular deviation of an observed angle, *θ,* of three atoms connected by two bonds from the reference value, θ_{eq}. K_θ is the force constant.

The dihedral term uses a cosine series expansion to determine the energetic consequence of rotation around a bond. Because the atoms are described as single points and the electrons are not explicitly defined, these terms are necessary to replace the interactions between the electrons of neighboring atoms that would affect the rotations around the bonds; a double bond between two carbon atoms would be an example of such an interaction. This rotation is defined by a series of four bonded atoms with the bond being rotated around between the second and third atom. To calculate the energetic barrier to rotation, a dihedral angle, φ, is multiplied by n, the periodicity parameter, which allows for more than one potential barrier per 360 degrees of rotation. Next the phase parameter, γ, is subtracted to shift the relative position of the minima and maxima of the potential. The cosine is taken of the angular measure, one is added and the result is divided by two. This changes the angular measure to a value that can range from zero to one as the dihedral angle rotates from -180° through to 180°. This value is then multiplied by a constant, V_n, which represents the maximum potential barrier for that term. It is important to note that "barrier" represents only the portion of the energy penalty due to the individual term, several terms may be employed for each rotatable bond and other components of force field (i.e. electrostatic interactions) may contribute the total energetic impediment to rotation around a bond.[27] See figure 2.1.

The remaining term in the force field equation is used to calculate the non-bonded interactions for atom pairs *i* and *j*, where atoms *i* and *j* are considered to be "non-bonded" if they are separated by more than three bonds or are not connected by bonding at all. The interactions calculated by these terms include the van der Waals interactions and the electrostatic interactions. The pairwise electrostatic interaction is calculated as a function of the charges, *q*; the charge separation, R; and the dielectric constant, ε_0, using Coulomb's law. The van der Waals interactions are calculated pairwise with a Lennard-Jones 12-6 function where R is the separation, σ is the separation where the energy value is defined to be zero, and ε_{ij} is the well depth for the pair of charges.

Figure 2.1 Parm99 Psi Term Summation.

A set of parameters for a φ or ψ angle defines one or more functions, which in turn determine an energetic penalty for the given dihedral angle. This graph depicts the values for the ψ angle terms from Parm99. As an example, function one, in green, has a periodicity parameter of 1, a phase of 0º, and a barrier height of 3.5 Kcal/mol. The summation of term one (green) and term two (blue) is shown in red. The angles from the tetrapeptide training set used to develop Mod2 are indicated by circles (described in chapter 3).



2.1.1.2 Parameter Sets

The parameter set is the set of constants that are used in the Amber equation. The parameter set for the Amber force field has undergone several generations of improvements and continues to evolve. The charges used in the Amber force field were derived with restrained electrostatic potential (RESP)[28, 29] charge fitting to quantum mechanical calculations. Many of the parameters were derived directly from experimental data, like the reference bond lengths in the bond stretching term, that were derived directly x-ray crystallographic structural data[30]. Other parameters like the ϕ and φ torsional terms were developed to reproduce dipeptide energies derived from high level quantum mechanical calculations[31]. The parameters for the ϕ and φ dihedral terms are especially important for the study of protein folding. As the peptide unit is rigid, the rotations of the peptide units around the α carbon atoms are the main conformational degrees of freedom in a protein backbone. If all of the ϕ and φ angles in a protein are determined, the structure of the backbone is essentially defined. Furthermore, because the parameters for the ϕ and φ angles in the parameter sets were developed to match structural energies they have been used as a corrective term to adjust the performance of the force field. The set of parameters that is currently commonly used with the Amber force field is Parm99[32].

2.1.2 Trajectory Analysis

The calculated trajectory is strictly a series of coordinate sets. The useful meaning from a trajectory comes from the emergent properties that are found in the relationships between the coordinates that were not explicitly defined by the force field. These properties can be investigated in several ways. The coordinates can be compared to each other, they can be compared to another model, or they can be compared to direct experimental observations.

To compare two entire models the models must be overlaid, or fit, in such a way as to minimize the differences between the structures. The most common method for fitting two molecules is by minimizing a function that reports the difference between two structures. A common method of making this comparison is to calculate the root mean squared deviation (RMSD) between the structures. If *N* is the number of atoms to be compared, and *d* is the distance between the atoms, the RMSD can be calculated as seen in equation 2.4[33, 34].

Equation 2.4 RMSD $RMSD = \sqrt{\frac{\sum_{i=1}^{N} d_i^2}{N}}$

There are many other possible analyses that can be used to evaluate structures including simple distance measurements, fraction of native contacts (Q), hydrogen bond populations, salt bridge populations, and energy analyses. It is important to note that many of these analyses use a model built from experimental observations. The analyses are not directly compared to the original experimental data. Often a molecular modeling force field is used in the process of model building from experimental data, thus biases from a force field may be present in the reference model. It is also often possible, and usually worthwhile, to compare computational models directly to the primary experimental data.

2.2 Solvent Models

Proteins and other biomacromolecules assume their native state in solution. To correctly model the behavior of proteins we must accurately model the effect of the solvent. To balance the competing computational expense of simulation with the need for realistic replication of solvent effects there are several different solvent models that are used for different purposes[35].

2.2.1 Implicit Solvent Models

Implicit solvent models dramatically reduce the degrees of freedom of a model system at the expense of interactions involving individual solvent molecules. The effect of the solvent is reproduced by representing the solvent as a continuum, presenting an approximation of an averaged solvent effect to the macromolecule. A further advantage of continuum models is the lack of solvent friction; the absence of solvent friction[36] removes an impediment to the search of conformational space and effectively speeds up simulations. Implicit solvent models are reviewed in reference [37].

2.2.1.1 Generalized Born Solvent Model

The generalized Born solvent model (GB)[38] is a continuum model used to approximate the electrostatic contribution to molecular solvation free energies of a high dielectric solvent like water (reviewed here [39].) The effect of the continuum model is to lower the computational cost of simulating molecular motions[40]. Simulations by Tsui and Case of unconstrained A-type DNA helices converge to B-type helices in 20 ps with the GB solvent model as compared to 500 ps for similar simulations with explicitly included solvent. A drawback of GB is the necessary lack of specific water interactions. Small local effects such as structured waters and charge bridging cannot be directly modeled with GB[40-42]. Despite this shortcoming, simulations employing GB have reproduced complicated protein movements[43], including protein folding as this dissertation will describe.

2.2.2 Explicit Solvent Models

Explicit solvent models are exactly as they sound, each atom of the solvent, or an approximation thereof, is modeled along with the solute of interest (reviewed here [35].) Despite advances in implicit solvent modeling, there are still differences in the results of simulations with implicit and explicit solvent. Neymeyer and Garcia were able to show distinct differences in the ensembles of structures sampled with implicit and explicit solvent, both for structures near the native state and for unfolded structures[44]. Furthermore, continuum models are unable to reproduce short range effects such as charge bridging[40, 45, 46]. The most commonly used explicit solvent models are the transferable intermolecular potential (TIP) functions[47]. The simplest models of explicit water use three points (notably TIP3P described below) but there are more accurate, and thus more computationally expensive, models that use more particles to simulate each water.

2.2.2.1 TIP3P Explicit Solvent Model

TIP3P[47] is a three particle water model that has been used successfully to model protein dynamics; reviewed in reference [48]. The bond lengths (0.9572Å) and angles (104.5°) of the described particles are fixed in the TIP3P model but there are other explicit solvent models where these degrees of freedom are flexible[49]. TIP3P was developed to reproduce water propertied and specifically tested for its ability to reproduce free energies of solvation. Relative to both experimental calculations and other more comprehensive solvent models, such as TIP5P, TIP3P underestimates density as a function of temperature for physiologically reasonable temperatures[50] and overestimates the diffusion constant. These approximations limit the properties that can be determined with the TIP3P model but in general the motions of proteins in molecular dynamics are slow enough that these approximations are valid. The computational expense of a more accurate treatment of the solvent must be weighed against the consequential loss of simulation time. In the simulations presented here, the inherent limitations of the TIP3P solvent model were outweighed by the efficiency, and thus it was chosen for these studies.

Chapter 3. Force Field Development

3.1 Identification of the Bias in the Current Amber Force Field

For molecular mechanical tools to be useful, they must accurately evaluate the relative energies of different structures.

The dihedral terms and their associated parameters in the Amber force field are used to calculate energetic penalties for rotation around a bond. Our interest in the dihedral parameters in the Amber force field came from noting that α-helical conformations were over-stabilized in molecular dynamics simulations of proteins using the current set of parameters (Parm99[32]). This suggested that the Parm99 force field was not accurately evaluating the relative energies of the different structures.

In an 8 ns explicit solvent molecular dynamics simulation by Dr. Guanglei Cui, Parm99 turned an unstructured region of the peripheral subunit-binding domain P.S.B.D. peptide[51] to an α-helical conformation. Similarly, during high temperature simulations of a tryptophan zipper fragment trpzip2[52] conducted by Asim Okur with Parm99, an α-helical conformation became stable even though the native structure of the fragment is in a β hairpin conformation. These results suggested that Parm99 energetically favors α-helical conformations. Finally an eleven amino acid fragment of alpha lactalbumin[53], which is only partially 3_{10} helical in the native conformation, was stable in a fully α-helical conformation during simulations with Parm99, further suggesting that the parameter set was favoring the α-helical conformations. See figure 3.1.

To confirm that it was the dihedral parameters that were responsible for the over-stabilization of the α-helices, a parameter set was constructed that was intended to have null values for the dihedral terms (ParmX.) It was later determined that not all of the terms that affect the dihedral terms in ParmX were zeroed. Refining the parameters with truly zeroed parameter terms will be one of the future goals of this project. No stabilization of α-helices was noted after extended molecular dynamics simulations with ParmX. This further suggested that it was the dihedral parameters in Parm99 that were responsible for the over-stabilization of the α-helices. ParmX also afforded the ability to test if the ϕ and φ dihedral parameters were necessary at all. By comparing the energies and RMSD values from simulations with and without the included ϕ and φ dihedral parameters the necessity of the parameters could be tested. Figure 3.2 shows the results of such simulations; the RMSD values were determined from the native structure as described by NMR. The lack of a gap between the energies of low RMSD structures and high RMSD structures in the ParmX graph suggests that the ϕ and φ terms are necessary for the force field to correctly identify the near native structures as the lowest in energy. Further, because ParmX has no ϕ and φ terms, it can be used as a baseline to measure the effect of different added ϕ and φ terms[54].

As a means of clarifying the energetic contributions of terms that go into the energy calculation the individual components were calculated for a complete set of rotations around the ϕ and φ angles of an alanine dipeptide. These values were then plotted on a standard Ramachandran plot. The contribution of the force field without the ϕ and φ dihedral terms was calculated (ParmX), the contribution of only the Parm99 ϕ and φ dihedral term was calculated, and the contribution of a GB implicit solvent model was added to further clarify the energies involved in the dipeptide system. These values can be seen separately and combined in figure 3.3. On the "combined" graph the relative energies of different conformations can be seen.

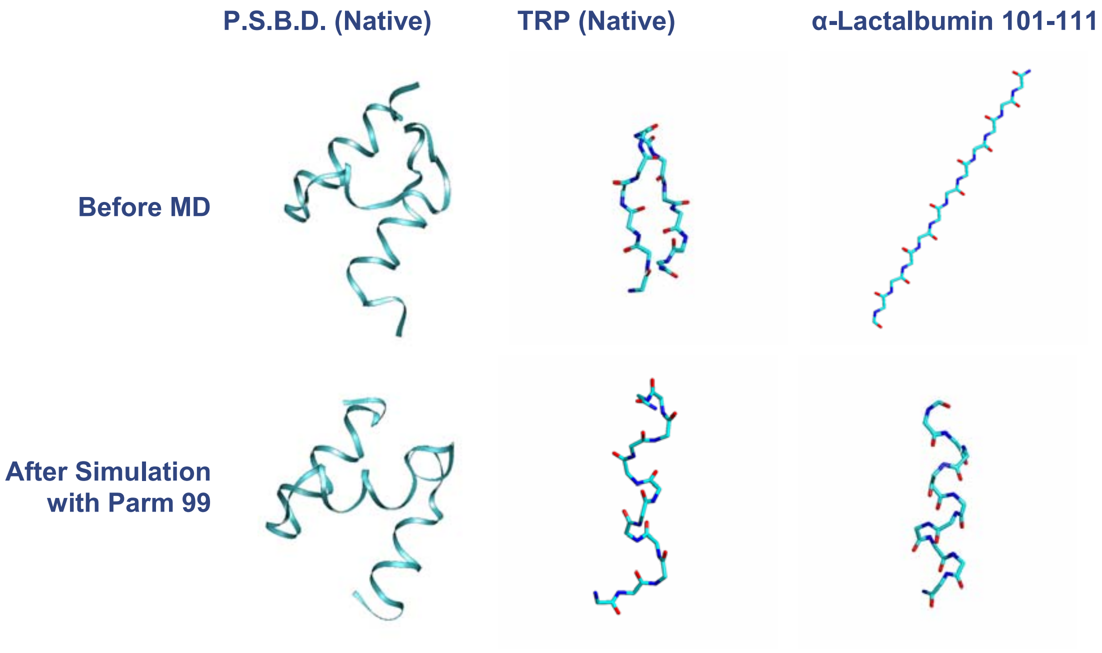


Figure 3.1 Protein Structures Before and After Simulation with Parm99

Backbone traces of P.S.B.D., TrpZip2, and α-Lactalbumin 101-111 before and after simulations with the Parm99 parameter set. All three structures become either partially or wholly α helical in nature. The middle section of P.S.B.D. turns alpha helical after simulation. TrpZip shifts from a β-turn to an alpha helix. α-Lactalbumin 101-111 turns from a linear structure to a stable ordered α helix while the native conformation of the fragment is a partially disordered 3_{10} helix. These changes suggest an α-helical bias in the Parm99 parameter set.

Figure 3.2 ParmX and Parm99 Energy vs. RMSD
These graphs show the energies of 7500 structures of the α-lactalbumin fragment as evaluated by the ParmX and Parm99 parameter sets compared to the RMSD from the native structure as described by NMR. ParmX is a parameter set with no φ or ψ torsional terms. Parm99 shows a large energy gap between the native and non-native structures while ParmX shows no such gap at all. The energy gap provides an explanation for why Parm99 over-samples α-helical structures.

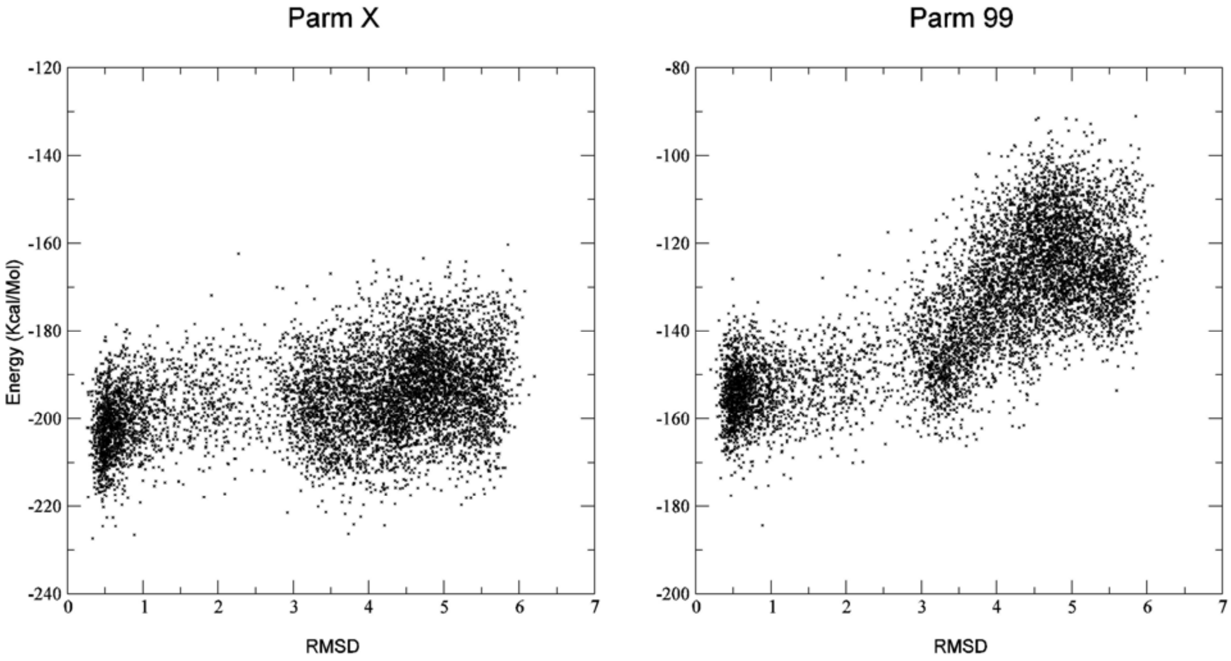
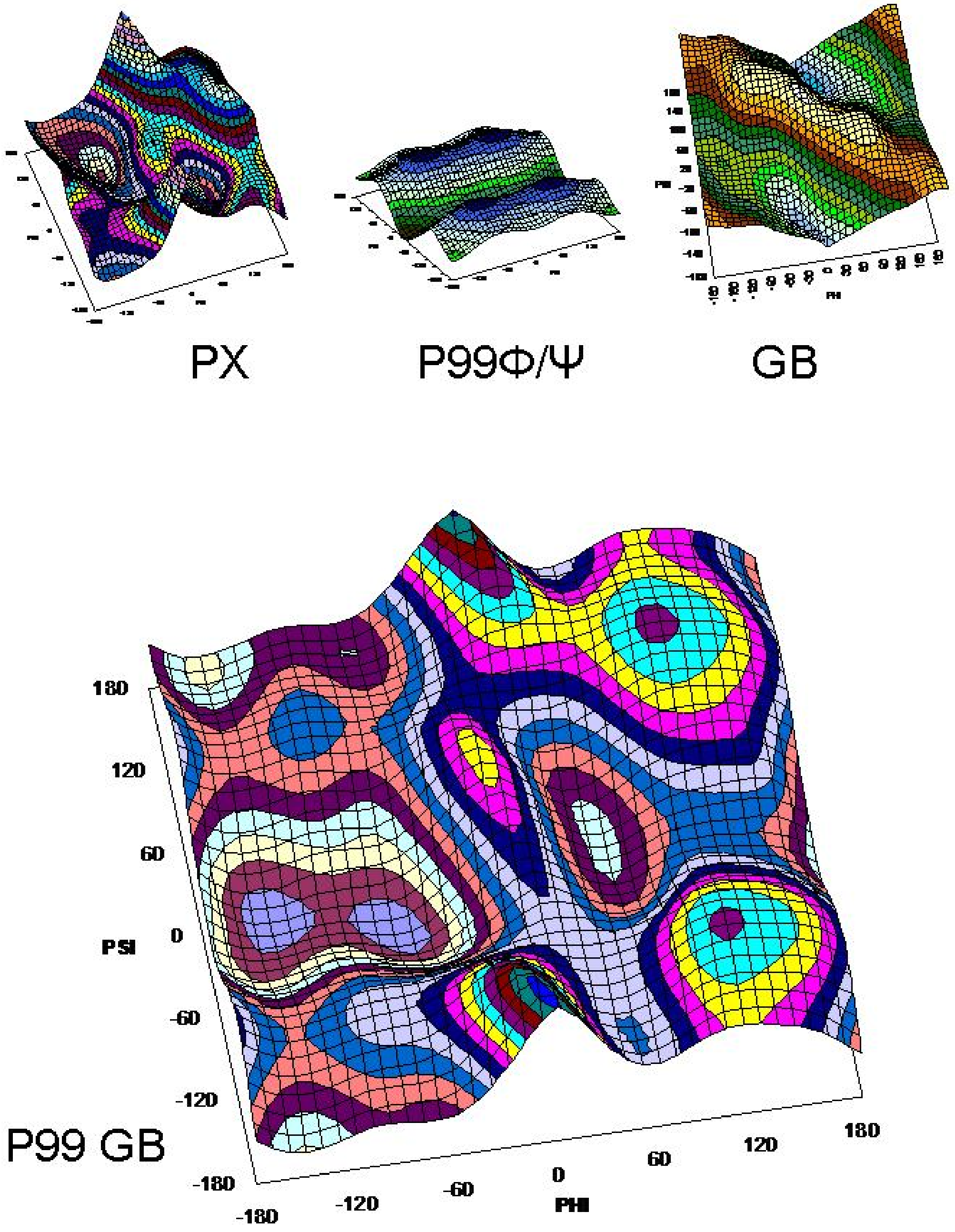


Figure 3.3 Energy Components of an Alanine Dipeptide

The separate components of the ParmX, Parm99 ϕ and φ dihedral terms, and the GB implicit solvent model energy component calculated for all the ϕ and φ rotational values plotted on a Ramachandran plot. The large image is the summation of the smaller graphs. The color gradients represent 1 kcal/mol.



3.2 Training and Testing of Force Fields

Historically, the most used parameter set with the AMBER force field is PARM94. Prior to Parm94, parameter sets, particularly the set of Wiener et al. [30], were developed for gas phase potential functions. Parm94 was developed with empirically derived and quantum mechanically derived parameters to balance solute-solvent and solvent-solvent interactions[31]. The dihedral parameters of Parm94 were optimized to match a set of quantum mechanical energies[55] for a set of simple molecules with the hope that the parameters would be transferable to larger molecules. The Parm99 parameter set was an evolutionary development of the Parm94 force field. Notably, the fourier components of the dihedral parameters for the ϕ and φ were said to be improved.

The dihedral parameters of the Parm99[32] parameter set were developed by a systematic search of parameter space. The parameter sets were evaluated on their ability to match the result of a previous HF/6-31G** quantum mechanical evaluation of the energies[55] of a seven member dipeptide training set. The trained parameter sets were tested based on their ability to similarly predict the quantum mechanically derived conformational energies of an eleven member tetrapeptide test set[56]. The members of the set are numbered 1-10 and alpha; the alpha conformation is so titled because of its ϕ and φ angles are in the alpha helical region of the Ramachandran plot. See Table 3.1. The parameter sets that calculated relative energy values with the smallest average deviation from the values from the quantum mechanical analysis were judged to be the best. The calculation with the lowest overall energy was assumed to have the smallest deviation and was set to be equal to the quantum mechanically derived energy for reference purposes. This introduced the deviation between the reference structure energy and the quantum mechanically derived energy as a systematic error. The effect of this error can be seen by subtracting the energies calculated with Parm99 for the eleven member test set from the energies calculated *ab initio* for the same conformations. See figure 3.4. The alpha helical bias is evidenced by the fact that the alpha helical conformation has the lowest energy difference between the ab initio calculations and the calculation made with Parm99. This suggests that all of the energies were overestimated by the Parm99 calculations but the energy calculations of the alpha helical conformation had the smallest error making in relatively lower in energy, as noted in figure 3.1.

Figure 3.4 Parm99 Energies Minus *ab initio* Energies

The energies of the eleven member test set of structures of the alanine tetrapeptide derived with Parm99 minus the energies calculated *ab initio,* relative to the energy for the alpha helical (alpha) conformation. This graph demonstrates that the alpha helical conformation, as described by Parm99, has smallest difference between the *ab initio* calculations and energies derived with Parm99.

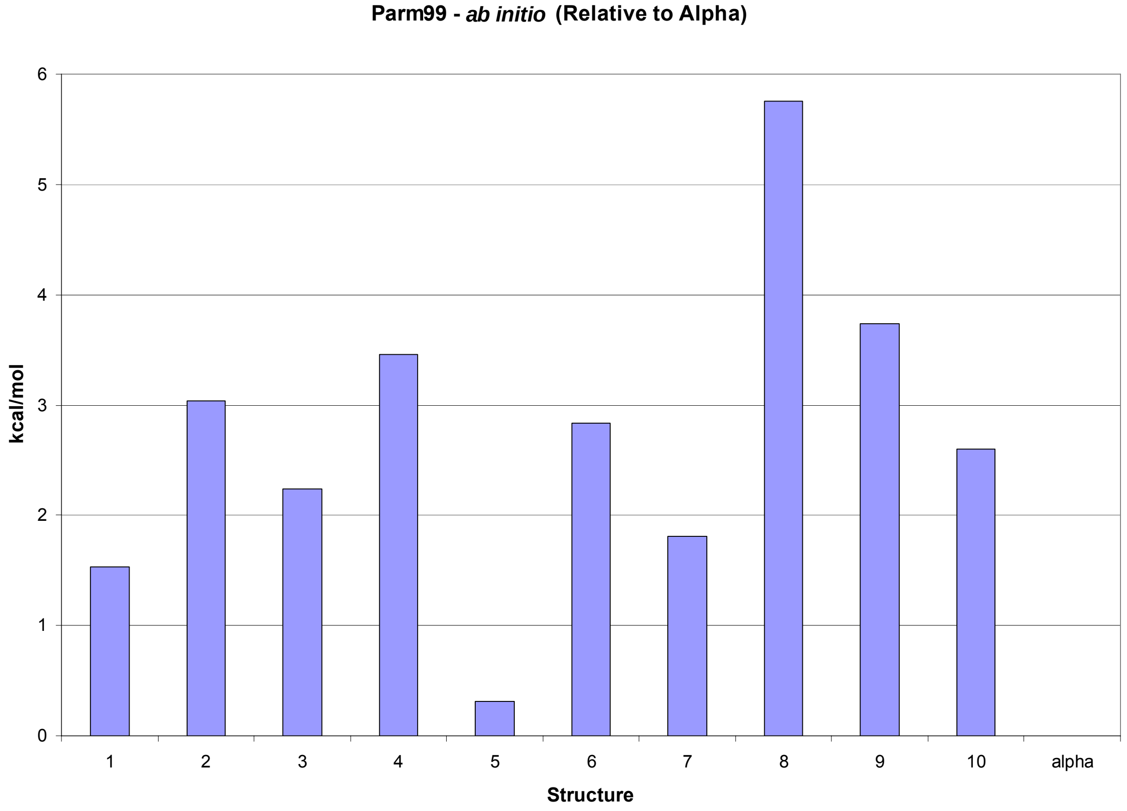


Table 3.1 ϕ and φ Angles of the Eleven Tetrapeptide Set

	Conformation Phi and Psi Angles					
Conformation #	**Phi1**	**Psi1**	**Phi2**	**Psi2**	**Phi3**	**Psi3**
1	-158.50	163.50	-157.80	163.40	-156.20	160.80
2	-158.60	163.90	-154.90	158.10	-86.00	79.20
3	-81.70	93.40	76.30	-53.40	-80.50	85.10
4	-156.90	161.30	-88.80	83.50	-156.00	152.80
5	-157.20	170.00	-76.20	-19.60	-153.80	160.80
6	-89.00	67.30	63.00	24.30	-165.00	149.80
7	56.00	-158.50	-93.00	63.80	-163.30	-50.00
8	72.80	-70.50	-58.10	134.70	62.00	25.70
9	75.70	-59.50	76.10	-53.30	75.50	-53.00
10	62.50	29.00	65.10	20.60	73.80	-51.50
(alpha) 11	-52.00	-53.00	-52.00	-53.00	-52.00	-53.00

3.2.1 Parameter Set Development

Because the dihedral parameters, which act on the torsional angles ϕ and φ and describe the energetic penalties to be attributed to those angles, are so critical for modeling proteins, the aim of this project was to develop an improved set of these parameters. It is specifically the parameters represented by V_n (the barrier height), n (the periodicity factor), and γ (the phase factor) of the dihedral parameters for the ϕ and φ torsional angles that we are investigating.

3.2.2 Exhaustive Search Methods

To identify a better set of dihedral parameters we performed a coarse search of a limited section of the possible parameter space. The search included both ϕ and φ parameters with a periodicity parameter of 1 or 2, a phase of 0 or 180, and barrier heights of 0 to 2 in steps of 0.1 kcal/mol. This entailed searching 3.1 million parameter sets ($(21*2)^4$). All calculations were performed on the 92 processor (40 800 mHz Pentium III, 52 1.4 gHz Athlon) Simmerling lab beowulf cluster “Ristra”. Molecular dynamics simulations were carried out using the SANDER module of the Amber 6 program suite[27]. Using the seven[56] dipeptide structures as a training[29] set, we evaluated the ability of each parameter set to match the results of the HF/6-31G** quantum mechanical evaluation of the energies of the same structures.

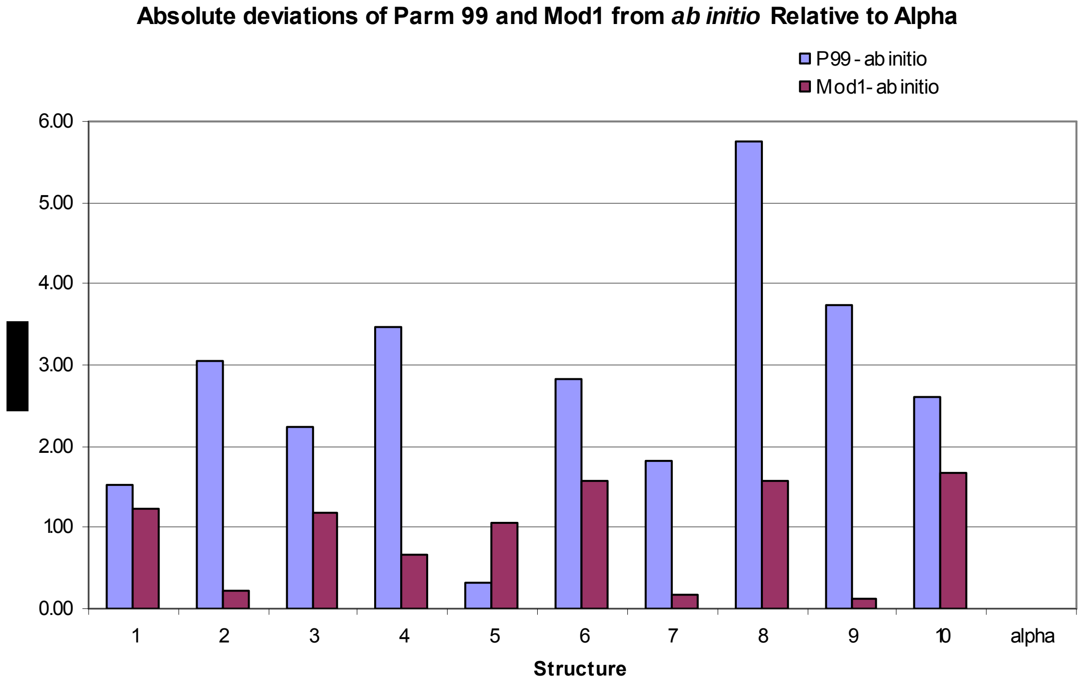
Before the energy evaluations, the test structures were minimized with the Parm94 force field by 10 steps of steepest descent followed by 990 steps of conjugate gradient minimization using a convergence criterion of 0.1 kcal/mol-degree, phi and psi torsional angle restraints to the original angles with a +/- 5 degree, flat bottomed wells with 5 degree parabolic sides, and a 50 kcal/mol-rad^2 force constant.

Each parameter set evaluation was started with a re-minimization using the parameter set to be evaluated to a maximum of 15,000 steps (10 steepest descent steps followed by conjugate gradient). The convergence criterion was 0.1 kcal/mol-degree and most minimizations continued for less than a total of 20 steps. Each minimization was restrained similarly to the original 1000 step minimization. For each set, following minimization, the average absolute energy differences and the maximum absolute energy differences of each of the 21 possible pairs of structures were calculated. The parameter sets that calculated relative energy values with the smallest absolute average deviation and the smallest maximum deviation from all the values from the quantum mechanical analyses were judged to be the best. This approach eliminated the systematic error that was introduced by using one structure as a reference structure.

The set with the lowest average deviation was dubbed Mod1. Mod1 was identified as a parameter set that performed better than the current standard, Parm99. (See figure 3.5.) The small size of the test set limited the number of structure families that could be included and the limits on structure size due to the computational expenses of quantum mechanical evaluations which limit the potential transferability of any identified sets. Limiting the phases to 0 and 180 degrees assured that the energy penalty was symmetrical around 0 degrees.

Because the Mod1 parameter set was trained on the smaller set of dipeptides we were concerned there was insufficient data in the test set to accurately reproduce the correct energies. To rectify the situation we ran a similar search of parameter space, but instead of using the seven member tetrapeptide training set we trained on the eleven member tetrapeptide set that had been originally used as a test set for Mod1. The resultant parameter set was named Mod2. To test the new set we used the set to calculate the energies of a series of dipeptide structures that have all the possible ϕ and φ values in degree increments. These energies were the plotted on a Ramachandran plot to identify the energy basins. This type of plot can also be used to compare different force fields. See figure 3.6.

Figure 3.5 Parm99 and Mod1 Absolute Energies Relative to Conformation α
The energies predicted using the Parm99 and Mod1 force fields for the eleven member tetrapeptide set. On average, the deviations between the Mod1 energies and *ab initio* are lower than the deviations between the Parm99 energies and the *ab initio* energies. This helps to explain why Mod1 favors the alpha helical structures less, relative to Parm99, than the other structures in the test set.



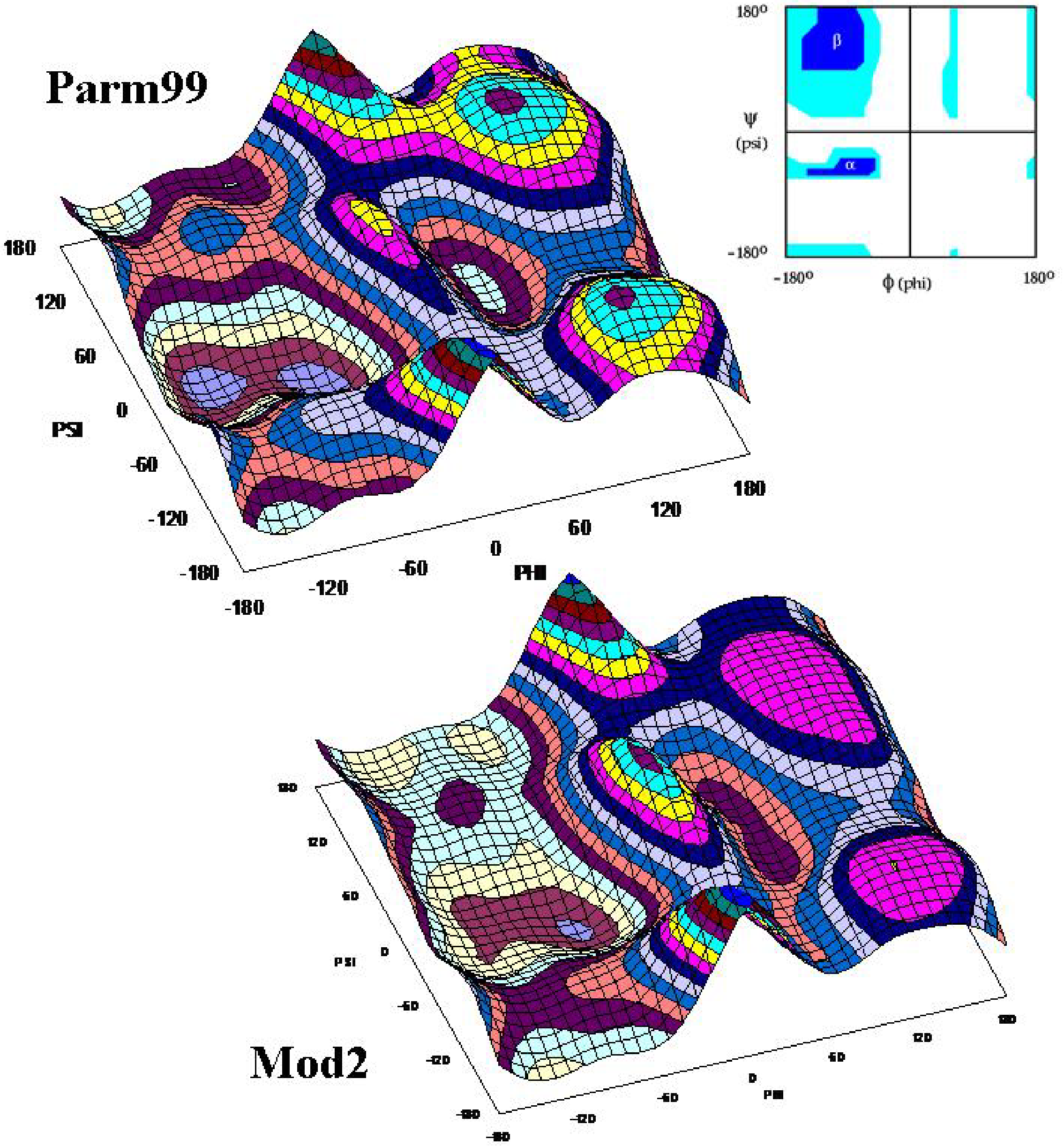
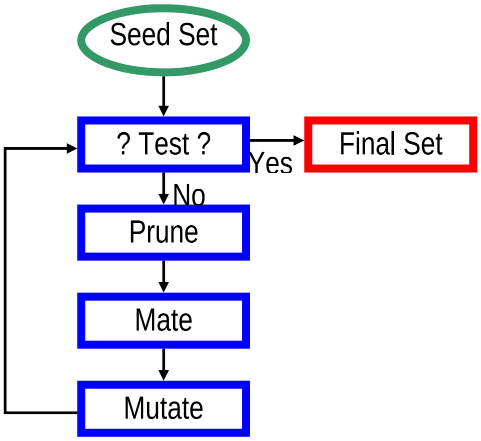


Figure 3.6 Parm99 and Mod2 Energy Ramachandran Plots

The graphs represent the summation of the GB energy, the ParmX energy, and the energies from either the Mod2 or the Parm99 ϕ and φ dihedral parameters for the alanine dipeptide plotted on a Ramachandran plot. The relative difference between the well for the alpha helical portion of the plot and the well for the β sheet portion of the plot is smaller in the Mod2 plot. This helps to explain why α helical structures are less favored with the Mod2 parameter set. The color gradients represent 1kcal/mol.

3.2.3 Genetic Algorithm Based Searches

Another approach to searching parameter space for parameter sets that correctly identify low energy structures is with genetic algorithms[57]. Genetic algorithms are function optimization techniques that actively select functions based on fitness criteria. The fitness function in this search involves the ability to correctly identify relative conformational energies of the test set. A high scoring fraction of the original population (seed set) is carried into the next generation. To introduce diversity, a fraction of the high scoring population are mutated or recombined to fill the next generation. The algorithm is stopped after a designated time, a designated value is reached, or after a designated period in which no improvement is seen in the output from one generation to the next. Genetic algorithms have two large advantages: 1) Instead of searching all possible values of all possible dimensions of parameter space, genetic algorithms make one dimensional moves through the multidimensional parameter space which allows them to be faster and to search the parameter space at a much higher resolution. 2) The conservation of successful sets with the addition of random sets allows for thorough searches of successful areas of parameter space while not getting stuck in only those areas. The disadvantage of genetic algorithms is that they



A simple genetic algorithm.

necessarily do not search every portion of parameter space which leaves open the possibility of not finding the best possible parameter sets.

3.2.3.1 Decoy Set Analysis

The first search of parameter space with a genetic algorithm was trained on the same set of eleven tetrapeptides (see table 3.1) as the previous comprehensive search. The algorithm can be written to output as many of the best parameter sets as desired. Because the training set is limited in its capability to represent secondary and tertiary structure we wanted to further evaluate the sets that were output from the genetic algorithm. To test each of the output sets, we used the parameters to evaluate the energies of a large body of structures over a large range of RMSD values of two peptides that had previously been described experimentally[54]. These structures are called decoy sets. The two peptides were trpzip2[58], a peptide that forms a β hairpin in its native state, and α lactalbumin fragment 101-111, a predominantly helical structure[52]. The sets were derived from simulations that were restrained to the native conformation, unrestrained, and forced to fold and unfold with targeted MD simulations[54]. Together, there were more than 7.5×10^5 structures in the decoy sets. This guaranteed significant sampling of the native state, a large population of unfolded structures, and a significant population of structures from the folding path. Decoy sets can be used to quickly evaluate the ability of the force field to correctly identify native-like structures as those with low energies because the computational expense of evaluating the energy of a structure is trivial as compared to the generation of a trajectory. This process is called decoy analysis.

The parameter set that resulted from the genetic algorithm search with the tetrapeptide set and decoy analysis of the decoy sets was dubbed ParmGA2. See Figure 3.7. ParmGA2 showed a larger helical bias than Parm99. Because the initial search was performed on the eleven tetrapeptides set of structures, it is possible the training process would be irrevocably biased toward the structures in that small set. To address this problem we decided to use the decoy sets as the training sets.

Using the decoy set as a training set requires a different approach to measuring the fitness for the genetic algorithm. A new fitness function using two measurements to evaluate each set was developed by Asim Okur in our lab. The first component is the energy gap between the RMSD of the native structures and the non-native structures. The native structures were defined as those structures with RMSD values below 1.0 angstroms and conformations were defined as non-native if their RMSD values from the native were greater than 1.7 angstroms. To calculate the gap, the 1000 lowest energies of both the native and non-native populations were averaged and the native average was subtracted from the non-native average. Thus positive values indicate that the native structures are lower in energy (on average) than the non-native structures. The energy gap measures how favored the native structures are relative to all the non-native structures. The second component of the fitness function is the Energy vs. RMSD slope. The slope was calculated as a vertical offset least squares linear fit. The slope of the fit line was taken to represent the energetic compulsion toward the native state. The fitness function of the genetic algorithm was the geometric mean of the two values.

The size of the test set, now 7.5 x 10^5 structures, necessitated that the energy calculations be optimized for speed. As only the ϕ and φ terms were changing, only the evaluations that involved those terms were calculated at each step.

The resultant force field was named ParmGA12. Energy/RMSD plots of the decoy sets evaluated with Parm99, and ParmGa12 are shown in Figure 3.8. The energy of high RMSD structures of the trpzip2 fragment as evaluated in Parm99 are lower in energy than the native structures, explaining why simulations using Parm99 did not populate native structures. The native structures of the same decoy set evaluated with ParmGA12 are lower in energy than the non-native structures suggesting the force field would favor the native structures. This effect was evidenced by simulations of trpzip2 using the ParmGA12 parameter set that had native like population fractions similar to that of experimental observations. In contrast, the slope of the energies for the α-lactalbumin decoy set is very steep when evaluated with Parm99 and less severe when evaluated with ParmGA12. This may help to explain why the Parm99 simulations over-stabilize helical conformations. The native structure is favored by both simulations but the non-native structures may be over-penalized in the Parm99 simulations[54].

For the sake of comparison, each of the parameter sets is described in Table 3.2.

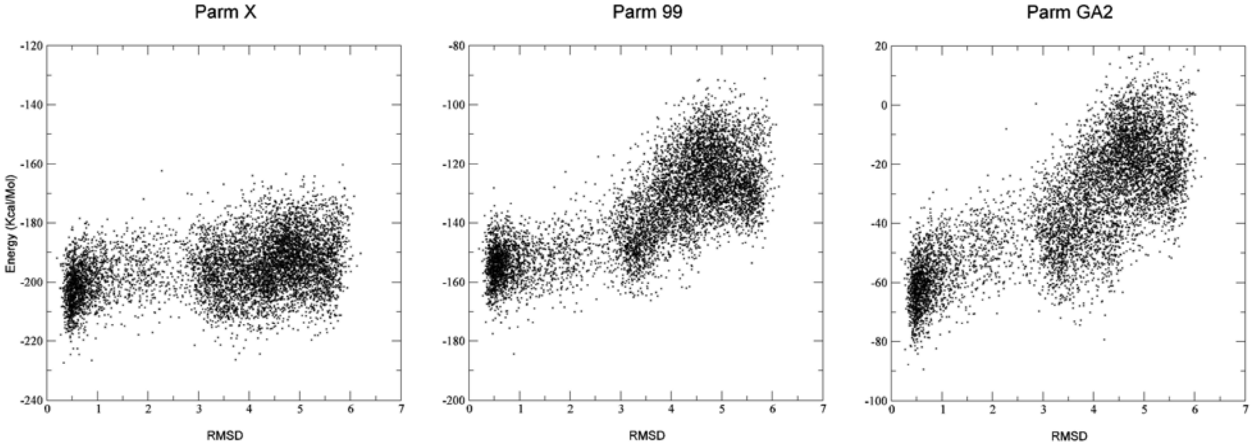


Figure 3.7 ParmGA2 Compared to ParmX and Parm99

These graphs show the energies of 7500 structures of the α-lactalbumin fragment as evaluated by ParmX, Parm99, and ParmGA2. Parm99 demonstrated a significant alpha helical bias (see figure 3.1) and ParmGA2, as shown by the third graph, shows an even larger bias. RMSD values are determined from the native structure as described by NMR.

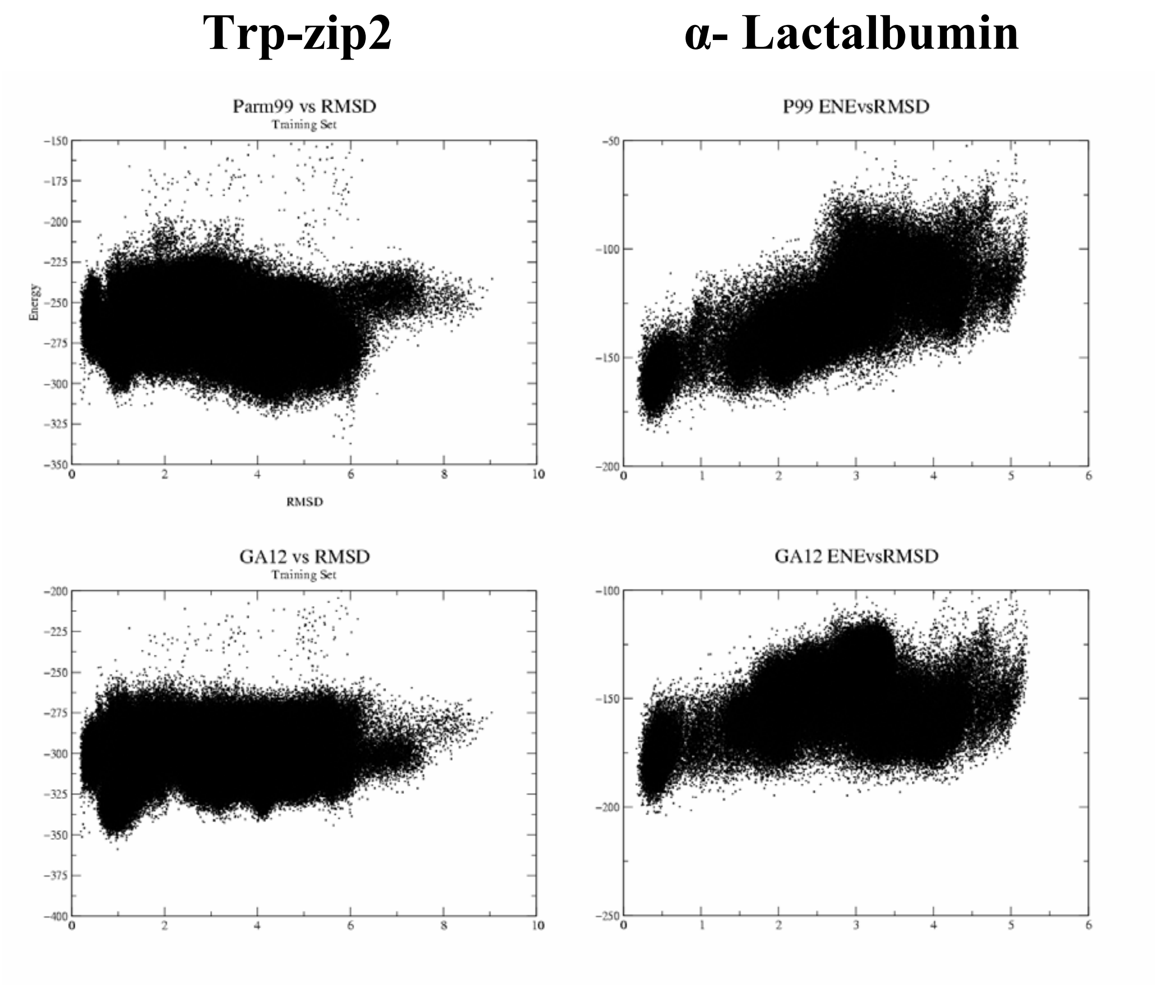


Figure 3.8 Decoy Analyses with Parm99 and Ga12

Each of the graphs represents an energy/RMSD plot of a decoy set. The top graphs are the trpzip2 and α-lactalbumin decoy plots evaluated with the Parm99 parameter set while the bottom plots are the same decoy sets evaluated with the ParmGA12 parameter set. For trpzip2 ParmGA12 produces an energy gap that favors native-like structure while Parm99 favors non-native structures. For α-lactalbumin the energy gap produced by Parm99 suggests that Parm99 will over-stabilize the native structures while ParmGA12 produces a smaller energetic difference between the native and non-native structures.

		ParameterSet Name					
Periodicity		**P94**	**P99**	**PX**	**Mod1**	**Mod2**	**GA12**
PHI 1	Barrier		0.80		0.20	1.00	0.40
	Phase		0.00		0.00	0.00	262.00
PHI 2	Barrier	0.20	0.85		0.70		0.41
	Phase	180.00	180.00		0.00		303.00
PHI 3	Barrier						0.02
	Phase						287.00
PHI 4	Barrier						0.02
	Phase						333.00
PSI 1	Barrier	0.75	1.70		0.30	0.70	0.48
	Phase	180.00	180.00		0.00	180.00	274.00
PSI 2	Barrier	1.35	2.00		2.00	1.10	0.45
	Phase	180.00	180.00		180.00	180.00	309.00
PSI 3	Barrier						0.12
	Phase						330.00
PSI 4	Barrier	0.40					0.45
	Phase	180.00					316.00

Table 3.2 Parameter Sets
The different parameter sets described in this document are presented with the values for the barrier, phase, and periodicity terms. The periodicity must be a positive integer and was limited to a maximum periodicity of four for each ϕ and φ.

3.3 Discussion

Ultimately, at the time these simulations were run, the ParmMod2 parameter set most accurately reproduced the energies of the structures in our test sets. The work of developing parameter sets is by no means complete and ParmMod2 represents and incremental improvement on its predecessors. New force field parameter sets continue to be developed in the Simmerling lab and elsewhere. In particular, ParmMod2 doesn't discriminate between glycine and other amino acids and newer parameter sets will make treat glycine separately because glycine can reach a larger area of phi/psi space.

Chapter 4. Structure Prediction and Implicit Solvent Simulations of Trp-Cage

4.1 Exendin-4

Exendin-4 is a protein that was originally isolated in 1992[59] from the saliva of *Heloderma Horridum*, more commonly known as the Gila Monster[60]. In 2001 the structure of Exendin-4 was solved by the Anderson laboratory at the University of Washington[61]. The structure revealed a novel protein motif which they called the tryptophan cage. The cage involves the side chain of a tryptophan pi stacked between two residues from other parts of the molecule with a Trp-εNH hydrogen bond to a backbone carbonyl[62]. A small number of other instances of the motif were identified, including instances where the tryptophan belongs to a separate peptide leading to the suggestion that the motif might be important for protein-protein interactions.

4.2 Trp-Cage

With the specific intention of designing a small, ultra-fast folding protein to aid in structural and computational studies, the Anderson group truncated and mutated Exendin-4 to produce the series of trp-cage cage proteins[63]. See table 6.1. In the scientific literature the name trp-cage has become synonymous with TC5b, a particularly fast folding mutant with the amino acid sequence NLYIQ^{5}WLKDG^{10}GPSSG^{15}RPPPS^{20}. The rest of this document will follow that convention. This peptide is the smallest peptide that displays two-state folding kinetics and has significant secondary and tertiary structure. At the time of its introduction trp-cage was the fastest known folding protein, folding in 4 μs[64]. The size of this construct, the rapidity of its folding, and presence of protein-like features mark the design of this mini-protein as a significant milestone[65].

4.3 Predicting the Structure of Trp-Cage

The ultimate goal of these studies is to be able to use force fields in simulations to provide new information. In particular, we would like to address two questions 1) can we find the native state, and 2) can we help explain the folding process. To address the first question we used the ParmMod2 force field to try and predict the structure of trp-cage. Prior to the release of the experimentally derived structure coordinates of trp-cage we ran simulations from a fully extended non-native state. We intended to test both the ParmMod2 force field and ourselves to see if we could identify the native state without the advantage of the experimentally derived molecular coordinates.

4.3.1 Simulation Details

The only trp-cage information that was used in each simulation was the amino acid sequence. With that information we used the LEaP module of the AMBER program suite to build a zwitterionic structure of the molecule. The starting structure was fully extended and primarily linear; the ϕ and φ angles were each set to 180° with the exception of the proline ϕ angle which was set to -61.5° and the φ angle which was set to -176.6° because of the constraints of the pyrrolidine ring. The trajectories were calculated with the ParmMod2 parameter set (see chapter 3) in the Sander module of AMBER 6. The simulations were unrestrained and the effects of the solvent were calculated with the GB solvent model[38]. Simulations were carried out at 300K, 325K and at 400K. The calculations were carried out on the local Simmerling Lab beowulf cluster "Ristra".

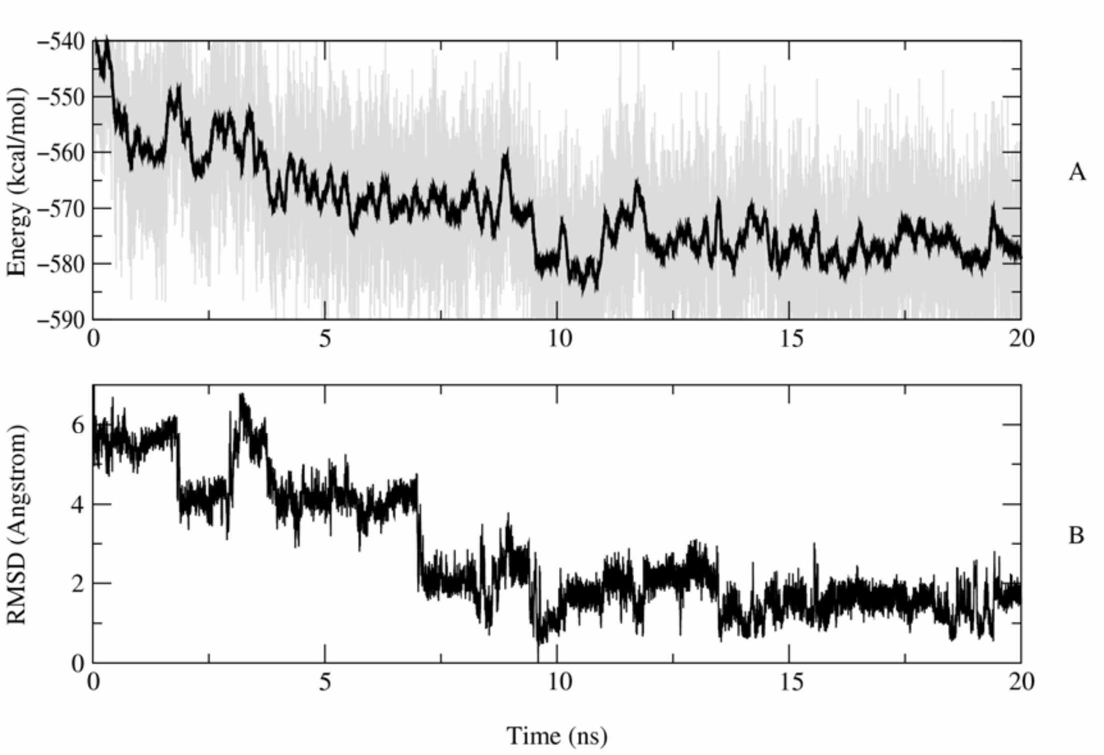
4.3.2 Identifying the Native State

When an experimentally determined structure is not available, it is difficult to evaluate the conformations sampled during a simulation. The convergence of predictions from multiple simulations is a reasonable approach to identify a "folded" state, but this can be misleading if the protein is not completely structured at the temperature of interest (generally physiologically relevant temperatures). We decided to also monitor potential energy (including solvation free energy) during simulations of this peptide. MD simulations of 100ns were performed at 300K, but all were kinetically trapped on this timescale, showing strong dependence on initial conditions and failing to converge to similar conformational ensembles. We therefore increased the temperature to 325K. The potential energy as a function of time during this simulation is shown in Figure 1a. A decrease of approximately 40 kcal/mol is seen over the course of ~10ns, after which no further improvement is noted. Two independent simulations converged to essentially identical families of structures after 5ns and 20ns.

We assigned this family as our "folded" state, and selected the snapshot with the lowest potential energy across the simulation as our representative structure. In Figure 4.1b, we show the backbone RMSD *relative to this structure* during the course of the same simulation from Figure 4.1a. A clear correlation between energy and RMSD is present; the energy plateau is reached at the same time as the convergence to the final structure family with RMSD values of ~1-2Å. The simulation was extended to 50ns, and no significant change in energy or RMSD profiles was observed.

Since folding was not reversible during these simulations, we performed a 20ns simulation at 400K which showed extensive sampling of conformations with RMSD values from <1.0Å to 7Å; even under these conditions the "native" family was transiently located on 6 different occasions and was the lowest energy sampled, although it comprised only 3% of all structures at this elevated T. These data provide additional evidence that the 325K simulations are not trapped in high-energy basins.

Figure 4.1 Evaluations of the trp-cage during simulation
(A) Potential energy of the trpcage as a function of time during MD. The solid line is a running average over 10ps. (B) Backbone RMSD during the same MD, compared to the lowest energy conformation. These graphs together suggest that as the simulation was arriving at lower energy structures it was populating only one structure family.



4.3.3 Comparisons of the Theoretical Prediction to the Experimental Data

Based on this analysis, the low-energy structure was given to the experimental group as our prediction prior to the release of the coordinates of their family of 38 solution state NMR models. The NMR-based coordinates are now available (PDB code 1L2Y), and the similarity of the NMR models to our low-energy snapshot is quite remarkable (Figure 4.2), the Cα RMSD between the structures was 0.97Å. NMR and theoretical structures share all of the following characteristics: residues 2-8 form a short α helix, a single turn of 3_{10} helix is present at residues 11-14, and the rest of the chain wraps back along the helical axis toward the N-terminus of the chain. The indole ring of Trp6 forms the center of a hydrophobic core, flanked by the side chains of Tyr3, Leu7 and 2 non-neighboring prolines (12 and 18). The Pro_3 triplet exhibits a polyproline II helix (which is the first native-like element established during the simulations, reducing the entropic penalty for formation of the cage), with the central Pro18 forming part of the cage. Two unusual intramolecular hydrogen bonds that are present in the NMR structure, between the Trp6 indole NHε1 and the backbone carbonyl of residue i+10 (Arg16) and between Gly11 HN and the carbonyl of residue i-5 (Trp6), are highly populated in the MD structures after folding (92% and 75%, respectively).

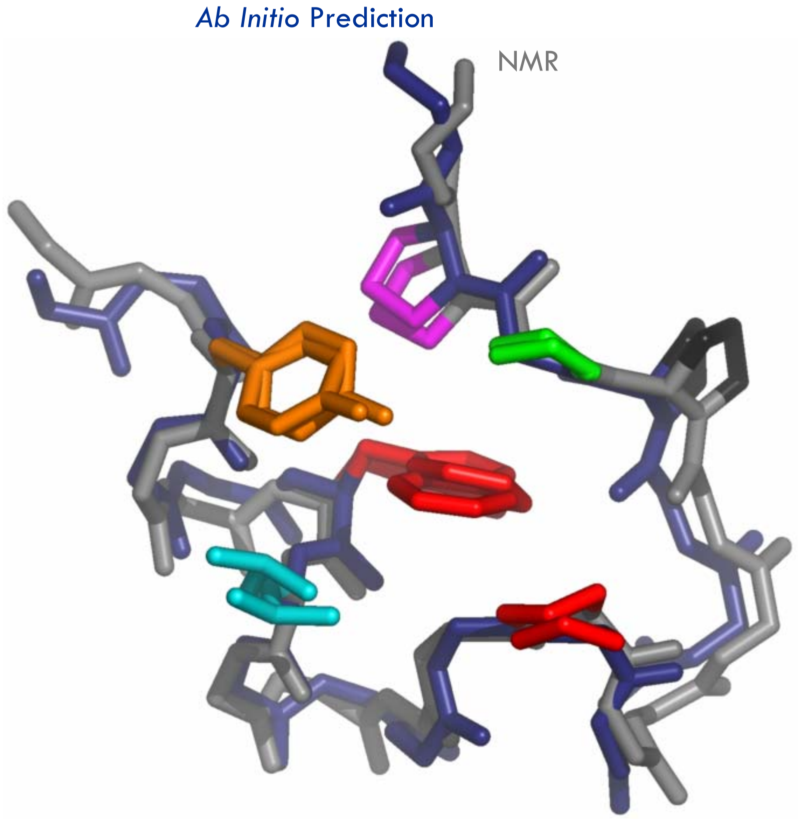


Figure 4.2 Trp-Cage Experimental Structure and Theoretical Prediction
The low energy MD (blue backbone) structure and NMR (gray backbone) structure shown after a best fit overlap. Only key side chains for the trp-cage are shown.

4.3.3.1 Root Mean Square Deviation

Neglecting the first and last residues and 3 side chains (all poorly defined in the NMR models, discussed below), the heavy atom RMSD between the experimental model and our low-energy structure is 1.4Å. Due to the large fluctuations observed using the continuum solvent, we carried out refinement of our model using 2ns 300K MD in explicit water. This resulted in further improvement, and the average over the final 500ps has a heavy atom RMSD of only 1.1Å compared to the NMR model**.**

The RMSD values were calculated between the model structure and the first structure of the family of NMR structures. The first structure of the NMR family was chosen because that structure is traditionally the best representative of the ensemble. The final RMSD values of the simulation are particularly remarkable because pair-wise RMSD values of the NMR ensemble range from 1.6Å to 2.8Å with a standard deviation of 0.3Å[66]. This suggests that our low-energy structure would be essentially indistinguishable from the NMR ensemble.

4.3.3.1 SHIFTS Calculations

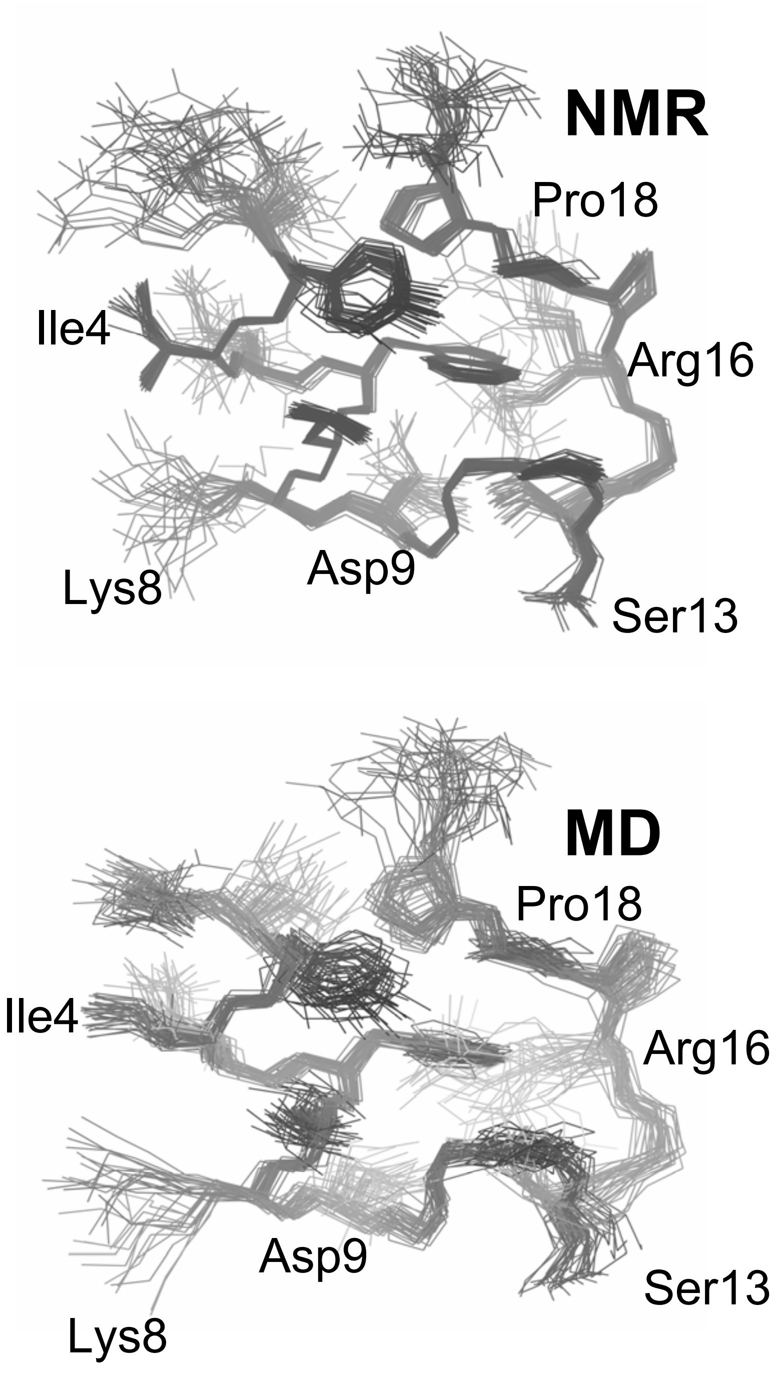
In analogy to calculations reported[67] for the experimental model, collaborators from the Roitberg Lab at the Univeristy of Florida performed ring current shift calculations for our structure using SHIFTS[68] 4.1 (http://www.scripps.edu/case). The correspondence between the chemical shift deviations (CSDs) for theoretical and NMR-based models is excellent, with root mean square error of 0.22 ppm and correlation coefficient of 0.99 for the 2 data sets.

These include the highly stereospecific CSDs for Gly30 Hα (-3.43/-0.96 for model 1 and -3.00/-0.54 for our structure), but we have excluded the outlier Pro18, which is in close contact with Trp6. The experimental structures assigned these prolines in the down pucker [69], and result in a Hβ3 shift of 0.34 ppm. During our simulations however, the down and up puckers are nearly equally populated with rapid sub-ns exchange, and representative structures give Hβ3 shifts of –0.22 and 1.22 ppm, respectively. Because the NMR structure data is necessarily derived from a large number of structures the experimental data may reflect averaging from two proline pucker populations that the simulations are able to distinguish.

4.3.3.2 Structure Ensembles

While it is important for an accurate structure prediction to correctly locate structured atoms, it is also important to predict the available conformational flexibility of a molecule. Consistent with the NMR-based models, the charged terminal residues, the sidechain of Leu2, and the sidechain of Lys8 sample multiple conformations during the simulations (see figure 4.3). In contrast, flexibility of the Arg16 sidechain is markedly reduced in the simulation compared to the NMR models. A potential explanation for this incongruence is described in the following section.

Figure 4.3 Structure Families from NMR and Native-like from MD
The structure families of NMR and the MD simulation show similar stability, notably in the backbone, and similar disorder, notably in the sidechain atoms of the terminal residues.



4.3.3.3 Salt Bridge

While Arg16 shows large variation for nearly all χ dihedral angles in the NMR models, this region exhibits relatively small fluctuations during the simulation. Closer examination of the MD data revealed that Arg16 participates in a solvent exposed salt bridge with the γ-carboxyl group of Asp9. The pairing was stable but transiently lost on multiple occasions. In this case the simulations likely provide the more reliable picture; a lack of NOEs and absence of prochiral assignments for Arg16 β and γ protons may have led to the poor convergence [69] of the NMR-based models. In fact, creation of this salt bridge was the motivation for mutating these residues during trpcage design. Further, our collaborators in the Anderson group at the University of Washington (who are the experimentalists who designed trp-cage and the authors of the original trp-cage paper) suggest that while there have never been enough diagnostic NOEs to produce this as a consistent feature in NMR structure ensembles, there is significant evidence to suggest the presence of the Asp9-Arg16 salt bridge. This includes pH titration experiments of trp-cage that show a large stability dependence coupled to Asp9 protonation[67] and mutants of trp-cage and other similar proteins that lack Asp9 or Arg16 do not display similar pH sensitivity in their melting temperatures. Also, TC10b (amino acid sequence: ^{1}DAYAQ^{5}WLKDG^{10}GPSSG^{15}RPPPS), an exceptionally stable variant of trp-cage melts at 57°C at pH 7 and approximately 18°C lower at pH 2.5. In contrast, the TC10b:Arg16Nva mutant (where Nva is norvaline), that cannot form the salt bridge, melts at 32°C at pH 7 and displayed essentially the same extent of folding throughout the pH range 2.5 – 7. This provides strong suggestive evidence for a stabilizing Asp9-Arg16 salt bridge.

After the simulations were analyzed it was suggested that salt bridges in general are over-stabilized in the Amber force fields. The experimental evidence for the presence of the salt bridge and the concordance the computationally and experimentally derived structures suggests the validity of the hypothesis that the salt bridge is important for the structure but discovering how important remains an area for future studies of this molecule. Accurately representing the stability of salt bridges continues to be an area of difficulty for computational simulations as evidenced by the results presented in Chapter 6 of this dissertation.

The only remaining significant difference between our model and that determined by NMR is the orientation of the side chain of Leu7; which may reveal limitations of our model.

One native simulation unfolded, resulting in loss of all elements of the hydrophobic core except a Trp6-Pro12 pair. A reduction in distances between the indole ring and Gly11/Pro12 was observed, consistent with experimental evidence for more negative chemical shift deviations (CSDs) at this T. Due to the complex nature of the unfolded ensemble, further simulations and analyses were warranted and are discussed in the following chapter.

Experimental data also suggests that a 16-residue sequence obtained from truncation of the C-terminal PPPS in trpcage does not significantly populate a single fold[67]; a 40ns simulation of this construct did not converge to any single structure, further strengthening the hypothesis that the cage motif contributes to the stability of this protein.

4.4 Discussion

While the CASP competitions[70] offer the opportunity for verifiable blind predictions of protein structure, we undertook this study due to the creation of the small and unusually stable mini-protein. The simulations we have described did not include any structural or other experimental data for the trp-cage but still converged to a highly similar family of conformations. In addition, our simulations suggest plausible structural details beyond those available from NMR models, such as the Asp9-Arg16 salt bridge. This demonstrates that MD simulations have reached the point where accurate structure refinement and prediction through direct simulation are not only becoming possible, but may soon be routine enough to contribute significantly to our understanding of the factors that determine folding.

These simulations were described in the first publication of Trp-cage simulations[43]. Because of its folding speed and stability Trp-cage has become an important tool for both enhancing our understanding of protein folding and for evaluating simulation tools[66, 71, 72]. While this work highlights the abilities of current force fields to accurately fold the trp-cage peptide, it does not suggest that the force field will be able to fold all proteins. (Specific limitations of the force field are discussed further in chapter 6.)

Chapter 5. Explicit Solvent Simulations of Trp-Cage

5.1 Folding Proteins with an Explicit Solvent Model

As discussed previously, in these studies we would like to address two questions 1) can we find the native state (structure prediction)? and 2) can we help explain the folding process (protein folding)? Studies of trp-cage with implicit solvation suggested that the force fields we used were indeed able to find the native state. It is the aim of this section to address the second, more difficult question. In particular, this section will discuss the details of folding simulations of trp-cage with explicit inclusion of solvent molecules and extend our previous results to the study of the folding pathway. Further, here we compare our simulations directly to the available experimental data.

In the past several years, the study of peptide and protein folding has advanced considerably[9, 23]. Recent complementary advances in experimental techniques and computational simulations are enabling both tools to be applied to the same model systems. Experimentally, single molecule techniques such as atomic force microscopy[13, 14] and single-pair fluorescence resonance energy transfer[19] allow direct measurements of single molecules, while ultra fast spectroscopic techniques, such as two dimensional infrared spectroscopy, measure protein conformational fluctuations in time scales of femtoseconds[16, 17].

Simulations that reach timescales that permit direct observation of folding events of even the fastest folding proteins have only recently become computationally feasible [73]. These simulations can supplement experimental observations by describing the nature and distribution of barriers and intermediates encountered during folding. Early reports focused on unfolding events and assumed reversibility of paths[74] or incompletely sampled the folding pathway[75]. Duan et al. published the first, and so far only, simulation to reach the microsecond timescale with a single continuous simulation[75]. This simulation of the Villain Headpiece sub-domain, published in 1998, started from the extended state but never found as many as 50% of the native contacts. Recent reports of extensive sampling of trp-cage using replica exchange molecular dynamics (REMD)[76] have described thermodynamically relevant structural ensembles [72, 77]. REMD simulations sample various temperatures to enhance the sampling of free energy basins but they cannot, by construction, study issues related to how a protein folds because the method breaks time-continuity at single temperatures. Pitera and Swope[72] described the folding trp-cage with REMD to within 2Å of the native state

Enhanced sampling techniques like REMD and increases in computer power have made possible all-atom simulations that are long enough to encompass full protein folding events; chapter 4 in this document provides one example. This success is also due to small proteins purposefully designed to fold extremely fast that effectively reduce the computational expense of folding simulations. Among these, trp-cage has become an important model system for protein folding studies.

5.1.1 Methods

With the intent of simulating the folding process of trp-cage we started several simulations at various temperatures with explicitly included solvent. These simulations were started from both fully extended and collapsed non-native structures. We simulated the folding of the trp-cage TC5b zwitterionic sequence NLYIQ^{5}WLKDG^{10}GPSSG^{15}RPPPS^{20} as described by Neidigh et al.[67]. The LEaP module of AMBER was used to generate an extended conformation for TC5b using only the amino acid sequence. Backbone φ and ψ angles were set to 180° for all residues except proline, which were initially set to -61.5° and -176.6° respectively. This extended structure was solvated in a 64.7 x 35.7 x 35.7Å periodic box with at least a 2Å water buffer and 1744 water molecules. Because the protein molecule was oriented along the longest diagonal in the box, with the intent of keeping the solvation box and thus the number of waters as small as possible, only a small fraction of the surface of the protein is within 2Å of the edge of the box; for the majority of the structure there was a considerably larger water buffer between the repeating copies (see figure 5.1). All simulations were carried out using AMBER 7.0[78] and the TIP3P[48] explicit water model, with the Particle Mesh Ewald[79] approach for calculating long-range electrostatics**.** The ParmMod2 parameter set was used in place of the AMBER 7 default force field (Parm94).

An MD simulation with constant pressure, and thus a variable box size, was performed at 350K, during which the extended conformation (see conformation B, figure 5.3) collapsed to a compact structure. After collapse, a smaller periodic box with 941 water molecules was used to reduce computational requirements. This structure was used as the initial structure for a simulation at 350K and 3 simulations around 325K (324K, 325K, and 326K). Two other simulations were started from two independent initial conformations (conformation A, 3.3Å and conformation C, 4.7Å RMSD from the native, see figure 5.3) that were generated through constant volume simulations of the native conformation at 600K and 800K, each solvated with 947 waters. Each of these conformations was subjected to further simulation at 350K. Backbone RMSDs are evaluated relative to the first structure of the published NMR ensemble for TC5b (PDB ID 1L2Y). The simulation temperatures were chosen to be above the experimentally determined melting transition state (315K) to encourage increased conformation sampling.

Zhou et al. [77] defined a native contact as any pairwise distance of sequentially non-adjacent α-carbons that is less than 6.5 Å in the first NMR structure. To directly compare our work with theirs we kept that definition. With this definition there are 34 native contacts in the native conformation (see table 5.1). Q is defined as the fraction of native contacts satisfied in an individual structure.

With the intent of using the same NOE-based distance restraints that were used to construct the NMR structures[67] to test our simulations, we used a set of 29 key restraints that our experimental collaborators determined to be necessary and sufficient to define the native structure (vide infra). See table 5.2. The fraction of restraints satisfied was used as a measure of the accuracy of the folded structure.

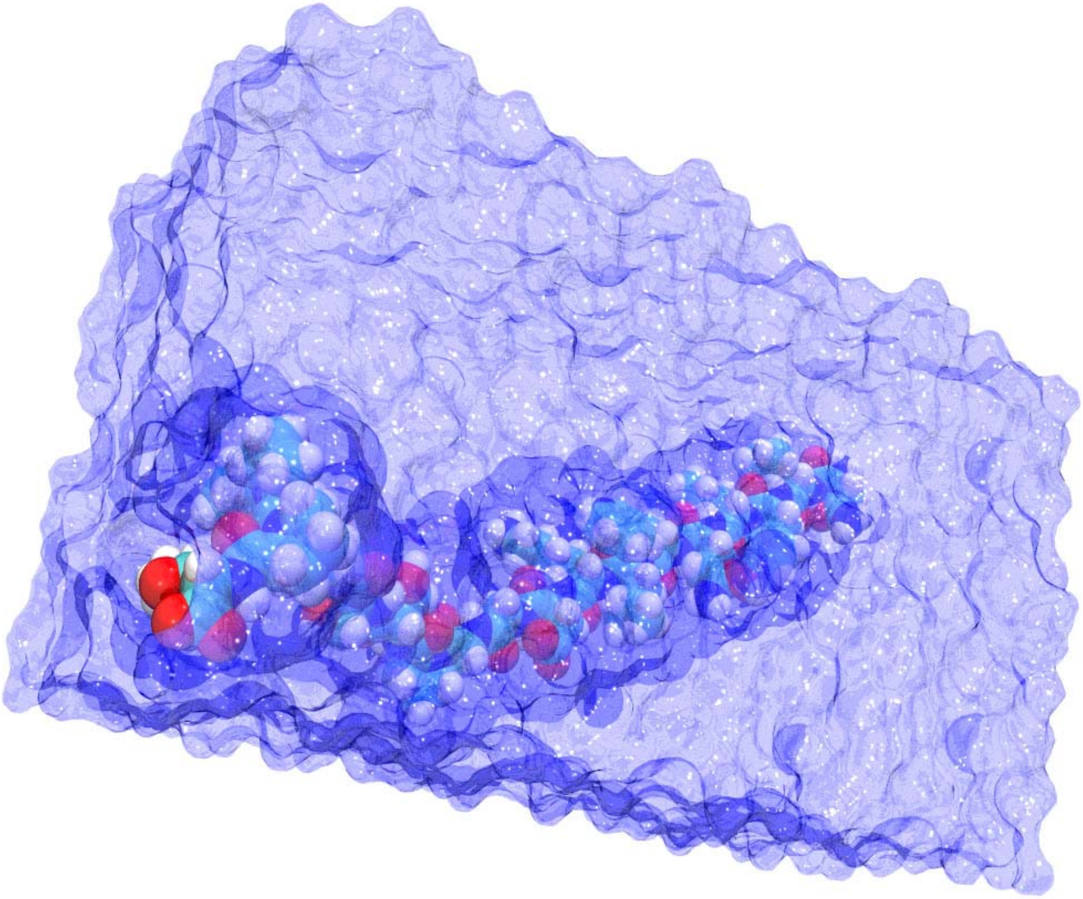


Figure 5.1 Solvated Trp-cage

The linear structure of trp-cage that was used to start the simulations. The protein is shown in standard molecular colors; the water filling the periodic box is presented as a transparent blue surface. This image is of the extended non-native structure before minimization and the uncollapsed water surface can be seen surrounding the protein structure.

Native Contacts in the First Trp-Cage NMR Structure

Residue i

Residue i+x, x>1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
3	5.6																	
4	5	5.5																
5	6.3	5.2	5.4															
6		6.4	4.9	5.4														
7			6.1	5.1	5.6													
8				6.1	5.4	5.7												
9						5.6	5.4											
10							5.2	5.5										
11						5.5	4.3		5.9									
12										6.3								
13										5.3	5.5							
14									6.0	5.2	5.8	5.6						
15													6					
16												6.5		5.6				
17																		
18																		
19			6.3															
20																		6.5

Table 5.1 Native Contacts
Pairwise distances in the first NMR structure in Å; a native contact is defined as a distance of less than 6.5Å between non-sequential α carbons in the first NMR structure. Empty boxes reflect distances between α carbons that are greater than 6.5 Å.

Atom1	Atom2	d	d-	d+
N-terminal helix				
residue 4 atom hn	residue 5 atom hn	3.00	0.50	0.50
residue 5 atom hn	residue 7 atom hn	4.00	0.70	1.00
residue 3 atom hn	residue 4 atom hn	2.50	0.50	0.50
residue 4 atom ha	residue 5 atom hn	3.50	0.60	0.50
residue 5 atom ha	residue 6 atom hn	3.50	0.60	0.50
residue 6 atom hn	residue 7 atom hn	3.00	0.50	0.50
residue 6 atom ha	residue 7 atom hn	3.50	0.60	0.50
residue 7 atom hn	residue 8 atom hn	3.00	0.50	0.50
residue 7 atom ha	residue 8 atom hn	3.50	0.60	0.50
residue 2 atom ha	residue 5 atom hb*	3.50	0.60	0.70
residue 3 atom ha	residue 6 atom hn	3.50	0.60	0.50
residue 3 atom ha	residue 6 atom hb2	3.50	0.60	0.50
residue 4 atom ha	residue 7 atom hn	3.00	0.50	0.50
residue 5 atom ha	residue 8 atom hn	3.00	0.50	0.50
residue 5 atom ha	residue 8 atom hb*	3.50	0.60	0.70
Local Structuring				
residue 7 atom ha	residue 10 atom hn	4.00	0.70	1.00
residue 7 atom ha	residue 11 atom hn	2.50	0.50	0.50
residue 12 atom ha	residue 15 atom hn	4.00	0.70	1.00
Core Packing				
residue 3 atom ha	residue 19 atom hg*	3.50	0.60	0.50
residue 6 atom hz2	residue 12 atom ha	2.50	0.50	0.60
residue 6 atom he1	residue 17 atom ha	3.50	0.60	0.50
residue 6 atom hz2	residue 18 atom hd1	3.50	0.60	0.50
residue 6 atom he1	residue 18 atom ha	3.50	0.60	0.50
residue 7 atom hd2*	residue 12 atom hd1	3.50	0.60	0.70
residue 6 atom hh2	residue 12 atom hd1	3.00	0.50	0.50
residue 6 atom hz2	residue 18 atom hb1	4.00	0.70	1.00
residue 6 atom hh2	residue 18 atom hb1	4.00	0.70	1.00
residue 6 atom hd1	residue 16 atom hb*	3.50	0.60	0.70
residue 6 atom hz2	residue 18 atom hd2	4.00	0.70	1.00

Table 5.2 Key NMR Restraints

29 Key restraints deemed necessary and sufficient for the determination of the NMR derived native structure by the Anderson Group at the University of Washington. “d” is the distance between the atoms, “d-“ is the amount subtracted from “d “ to arrive at the lower bound, “d+” is the distance added to “d” to arrive at the upper bound for the distance range. * Signifies that the protons attached to the carbon were indistinguishable.

5.2 Comparing Simulation to Experimental Data

We compare here our theoretically derived simulation data to the experimental data from our collaborators in the Anderson Group at the University of Washington that define the structural elements of the trp-cage fold. The NMR data of the protein define the key diagnostics for judging the extent to which our unrestrained dynamics trajectories can reproduce the trp-cage fold.

5.2.1 Defining the Folded State

Any discussion of protein folding must be carried out in the context of a definition of the folded state. With such a reference condition one can then study how this state is located and what types of structures are populated that do not fall under this definition. Direct comparisons of experimental data with the spectroscopic observations predicted for a dynamics-derived fold represents the most direct method for verifying the validity of a folding simulation.

The three NMR observations discussed herein that fall in this category are: chemical shifts deviations (CSD's, particularly those that reflect ring-current effects on sequence-remote hydrogens), NH exchange protection factors (which reflects sequestration from bulk water interactions due to persistent H-bonding), and interatomic distances obtained from NOE observations. These experiments observe folded (or unfolded) ensembles of structures, not on any one structure. This means the data can be used to help define the folded state but it cannot be directly used to describe the folding process.

To compare our results to the experimental data that represents a folded ensemble, we needed to select a group of structures from the simulation trajectories to act as a folded ensemble. We selected the first one thousand structures (1ns) from the 326K simulation starting with the structure with the lowest 3-19 cα RMSD (0.49Å) from the first NMR model (native). The ensemble has an average 3-19 cα RMSD to the native of 0.76Å with a standard deviation of .15Å.

5.2.1.1 SHIFTS Analyses

To address the question of whether the structures that we are considering folded are consistent with the NMR observations, we predicted the chemical shift deviations (CSD), using Shifts 4.1 [68], for the representative folded ensembles derived from the contiguous portions of the 326K dynamics trajectories maintaining RMSDs similar to the first NMR structure (see table 5.3). This also allows us to compare these results to the shifts calculated for the NMR ensemble with the same program as the Anderson group. Both sets of values were then compared to the experimentally calculated values. The experimental values were observed at pH 7.0 in water at 280K.

There are two significant disagreements from the experimental values in the shift predictions from the NMR structure ensemble. The first was a slight over-estimation of the upfield shifts of sites in Pro18 and Pro19, which may indicate the NMR structure places these two core units too close to the indole ring. The simulation is in close agreement with the experimental Pro18 and Pro19 CSD data. Additionally, the simulation shift calculations estimate the Pro12Hδ3 shift to be upfield while the shift is experimentally measured to be downfield. This shift was previously identified to be particularly sensitive to temperature change[63] so the reported difference may be due to the difference in the temperatures between the experimental system (280K) and the simulation temperature (326K.)

The remaining shifts were accurately reproduced by both the NMR ensemble and the simulation ensemble. Notably, both ensembles predict the large diastereotopic difference in ring current effects at the two α hydrogens of Gly11.

Table 5.3 NMR Chemical Shift Deviations
Chemical shift deviations experimentally observed, calculated for the NMR derived native ensemble, and calculated for the molecular dynamics simulation derived native ensemble. Values are expressed in parts per million. The NMR and MD values are averaged over their respective entire ensembles and are followed by the standard deviations of that ensemble. The CSDs are calculated for the ring current effects alone. *The 19Hδ CSDs are reported as an average as the protons displayed very similar CSDs both experimentally and in the theoretical models.

Atom	Exp. Observed	NMR	s.d.	MD	s.d.
11Hα2	**-2.97**	**-2.70**	0.29	**-2.45**	0.56
11Hα3	**-0.88**	**-0.92**	0.20	**-0.84**	0.24
12Hβ3	**0.23**	**0.12**	0.02	**0.11**	0.08
12Hδ2	**-0.31**	**-0.14**	0.05	**-0.23**	0.08
12Hδ3	**0.19**	**-0.15**	0.12	**-0.25**	0.20
18Hα	**-2.04**	**-2.58**	0.18	**-2.09**	0.41
18Hβ3	**-1.82**	**-2.07**	0.43	**-1.62**	0.71
19Hδ*	**-0.58**	**-0.92**	0.32	**-0.70**	0.42

5.2.1.2 NH Protection Factors

Amide proton (NH) exchange rates can be used as an experimental measure of the fraction of the folded ensemble that affords the proton protection [80]. In trp-cage two NH groups were measurably protected from exchange in the folded ensemble. In the folded state the Gly11 amide hydrogen and Trp6-NHε1 were protected suggesting hydrogen bonding to the backbone carbonyl of Trp6 and Arg16 respectively. Because the amide protons were not protected from exchange in the unfolded states, the protection can be used to define the fraction of the folded ensemble that has the presumed hydrogen bond. The Anderson group found that the interactions that afford protection in the experimental trp-cage structures are present in 98.4% of the structures. We would expect that if our selection of native structure was accurately representing this folded ensemble we would find a similar percent of structures that have the same hydrogen bonds. In the NMR ensembles, the two key hydrogen bonds were observed by our collaborators to be present in greater than 80% of the structures even though these structures were defined from the NOE distances and without the exchange information.

To measure the hydrogen bond populations in the simulation structures, we counted a hydrogen bond as being present in all cases where the distance between the hydrogen and heavy atom in question was less than or equal to 2.8Å; no angular parameter was used. 2.8Å was chosen because it is near the end of the range of hydrogen bond and it separated the first peak in the distribution of distances between the atoms over the course of the entire simulation. In the ensemble of folded structures from the simulation we observed the Gly11-Trp6O hydrogen bond to be present in 98.8% of the structures with a mean distance of 2.03Å and a standard deviation of 0.18Å. The Trp6NHε1-Arg16O hydrogen bond was found to be present in 94.5% of the structures with a mean distance of 2.11Å and a standard deviation of 0.24. The standard deviation is included as a rough measure of the spread of the distances despite the fact that data is biased towards longer distances and thus does not present in a standard normal distribution.

5.2.1.3 Violation of NOE-Based Distance Restraints

Because the NOE-based distance restraints were directly used to build the experimental model, we were interested in comparing our model directly to that data. This technique has been used by Zhou for both simulations of trp-cage[77] and other proteins (protein G) [45]. This technique also affords the opportunity to evaluate the utility of RMSD as a measure of folding. Published accounts of trp-cage folding simulations defined the cutoff for the folded state by Cα RMSD in the 2.0 [72] to 2.5 [71] Å or violations around 20% [66] of the 169 NOE distance restraints. Our collaborators from the University of Washington note that of the NOEs, only 64 (38%) are for i/i+n distances with n > 1 and that 50 – 75% of random unfolded structures yield folded structures that agree with all of these NOE distances (maximum restraint violation < 0.12 Å). To evaluate the folding simulations, they were able to reduce the set of 169 NOE distances to a set of 29 (see table 5.2) that were necessary and sufficient to reproduce NMR structures with a pairwise backbone RMSD of 1.34 Å versus the first published NMR structure.

To evaluate the folding simulations we evaluated the 326K trajectory of structures with the distance restraints. Each of the restraints was satisfied over time but no one structure satisfied every restraint. This is not entirely unexpected because the NOE restraint data reflects a time average over a number of protein molecules on the order of Avagadro's number while our data is for an individual structure with no range of structural fluctuations. To account for this we extended the NOE restraints by .5Å and compared the 326K simulation running fraction of NOE violations to the running 3-19 backbone heavy atom RMSD from the first NMR structure. (See figure 5.2) With the extended NOE range, the trajectory structures satisfy all of the NOE restraints on several occasions.

In less than 10ns the simulation reaches structures that satisfy more than half of the experimental data, followed by a period of fluctuation at ~75% and eventually a very rapid transition into a native-like ensemble shortly after 50ns. This transition into the native conformation is more apparent using the NOE violations than any of the other folding measures that we employed. The transition to the folded state as measured by NOE restraint violations also occurred at essentially the exact time the transition occurred as described by RMSD measurements.

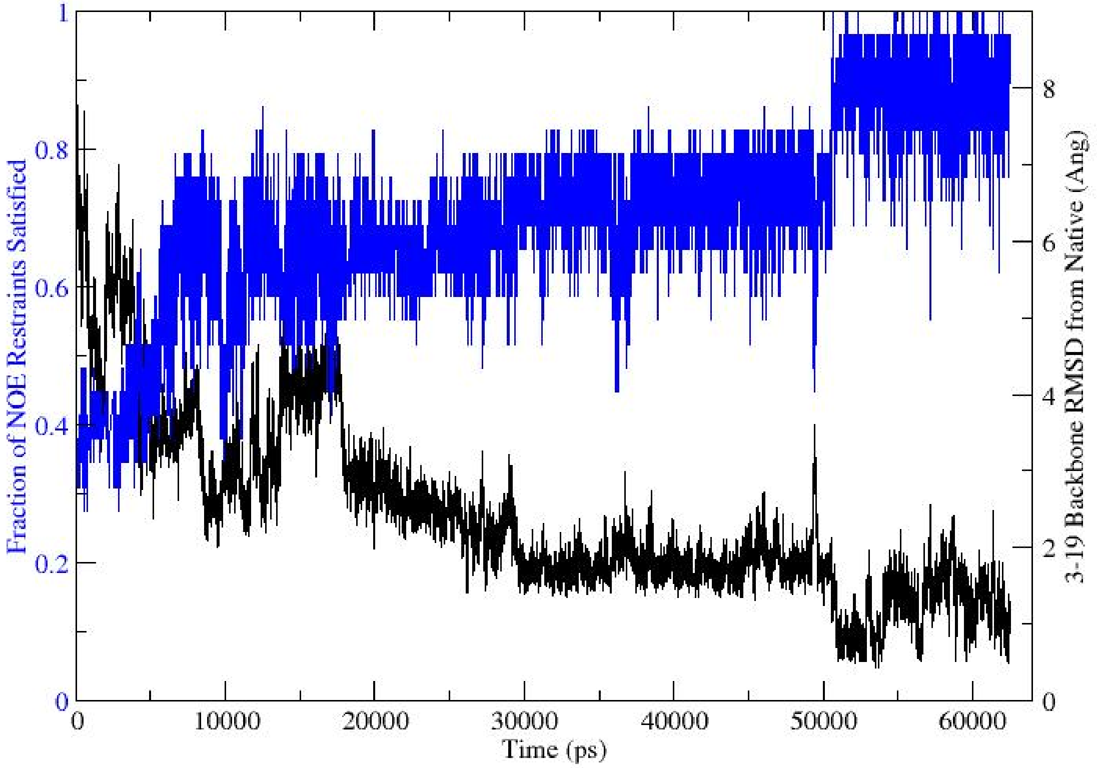


Figure 5.2 Fraction of Key NOE Restraints Satisfied
The fraction of NOE restraints satisfied for each structure during the 326K simulation is plotted in blue and described on the left y-axis while the 3-29 backbone RMSD from the NMR derived native structure is plotted in black and described on the right y-axis. Both data sets are plotted vs. time in ps. As the structure folds to more native-like conformations the fraction of key NOE restraints that is satisfied increases until structures that satisfy all the they key NOE restraints are reached.

5.3 Examining the Folding Process

Using the criteria described above, we saw complete folding events in three independent simulations. With simulations of the entire folding process we can use the advantages of molecular dynamics and investigate the fine details of the folding process. They also allow us to make many measurements beyond those we are capable of in experiments. While the limited simulation time and number of folding events precludes any discussion of thermodynamics or kinetics, they do allow for a discussion about the temporal order and commonality of proceedings during the observed folding events.

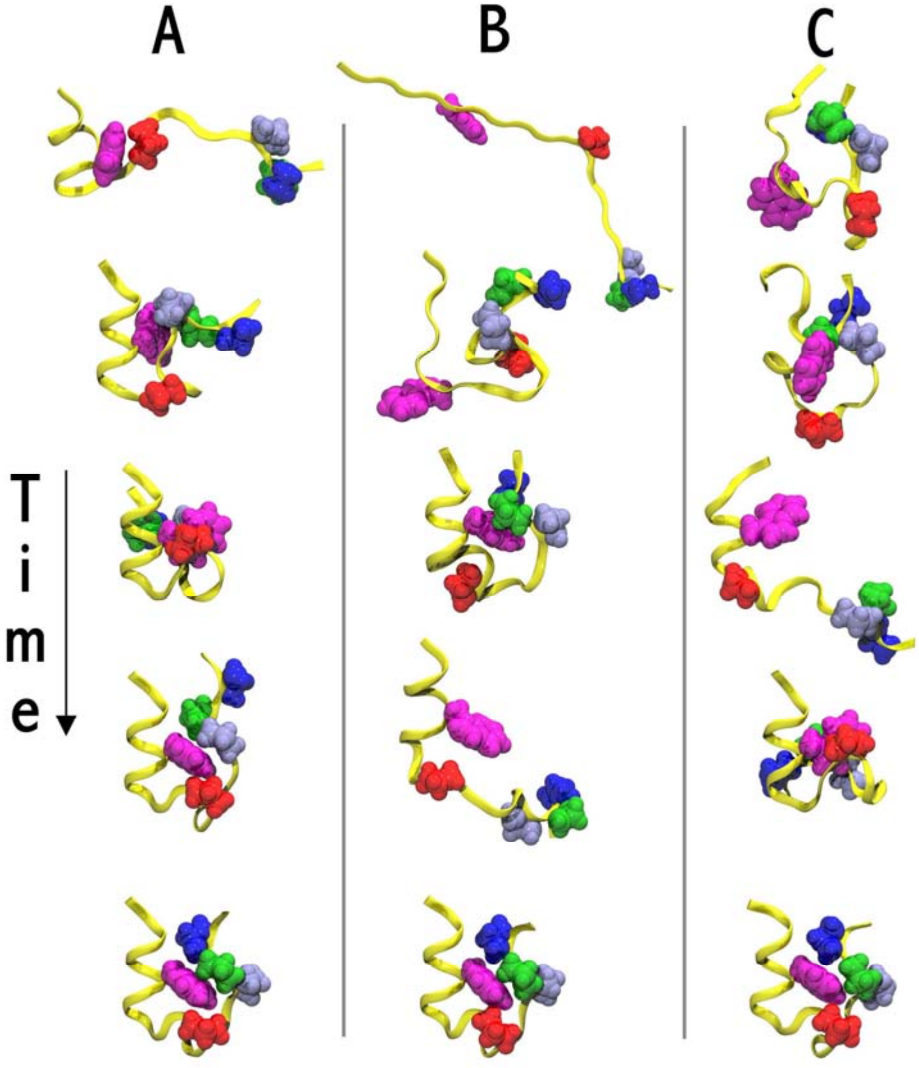
5.3.1 Which Simulations Folded

A total of seven simulations of trp-cage in explicit solvent were run. One simulation at 350K started from the fully extended state, collapsed within 10ns to a compact state with ~5Å Cα RMSD from the native. The size of the water box was reduced after collapse to reduce the number of waters used in the calculations that were no longer necessary to surround the compact structure. The simulation was continued and the structure folded to the native state, unfolded, and subsequently refolded on two occasions. A simulation at 326K, started from the same collapsed state, also folded to the native state. Another simulation at 350K, in a different water box, started from a different collapsed conformation, folded to the native state. Again, unfolding and refolding was observed, suggesting the simulations were not stuck in a local energy minimum. Four other simulations did not reach the native state. Figure

5.3 shows structures from the trajectories of each of the folding simulations. Each starts with the initial conformation and ends with the folded structure.

In total, three of the simulations folded to native like structures. Because our simulations were limited to relatively short time periods it is not unexpected more than half of the simulations did not fold. Each of the non-folding simulations explored a variety of structural conformations and non appeared to have arrived at a structure more stable than the native. In experimental conditions Trp-cage has been shown to fold in 4μs. This suggests (as is conventionally known) that because 3 of the 7 simulations folded in less than 300ns that the MD simulations folded significantly faster than experiment. This is a benefit in terms of computational expense but it does suggest limitation of the model.

Figure 5.3 Folding Pathways
Snapshots along the folding pathways in the 3 simulations, showing the initial conformation at the top and essentially indistinguishable native conformations at the bottom. All rapidly adopt compact structures, with simulations B and C sampling less compact structures before folding. Significant (but transient) α-helical content is present in early stages of folding. The final step in folding in all simulations is the docking of the pre-formed PPII helix (gray, green and blue) onto the "1/2 cage" structure stabilized by contact between W21 (purple) and P31 (red).



5.3.2 Measuring the Procession of Folding

We now describe the sequence of formation of specific features of the trp-cage structure during folding. Having full structures saved from every picosecond of simulation allows us to track even small movements at high resolution. This allows us to investigate minute details of each simulation and also to compare independent simulations. There were both notable similarities and notable differences in the folding simulations. We find that the simulations are consistent with the diffusion-collision model of folding[81-83]. This model suggests that elements smaller than the overall tertiary structure, which may or may not be secondary structure elements, can fold independently of the overall tertiary structure; these elements are called microdomains. Folding would then follow a series of coalescence steps whereby the microdomains together form the tertiary structure.

In our simulations, as would be expected from a diffusion-collision process, semi-stable secondary structure elements appeared before the native tertiary contacts were formed. Snapshots along the folding pathway for each of the simulations are shown in figure 5.3, with emphasis given to secondary structure and the side chains of Trp6 and the four prolines.

Figure 5.4 shows several of the folding measures described above as a function of time for the 326K folding simulation. In all 3 successful folding simulations, structures with the 3-18 backbone RMSD values over 6Å are sampled before reaching native conformations (figure 5.4-A) with RMSDs below 1Å from the native, corresponding to structures with 100% of the native contacts (figure 5.4-C). The folded structures of each of the folding simulations achieved 100% of the native contacts formed and satisfied all of the set of 29 NOE restraints deemed necessary and sufficient to define the fold.

The radius of gyration (Rgyr) was used to discriminate the collapsed states from the extended states. It was sufficient for this purpose but the measurements were not useful for discerning folded structures from collapsed structures (figure 5.4-B). While the large radius of gyration values corresponded to high RMSD values, many non-native compact conformations have radius of gyration values that are similar to the native state value. In the 326K simulation the radius of gyration dropped to values below 8Å, followed by a rise above 11Å before a final collapse to values that were comparable to radius of gyration values of the folded structures.

The first secondary structure element formed was the polyproline II helix (PPII) for prolines 17-19 (figure 5.4D). Because of the backbone structural limitations required by proline the polyproline helix was essentially formed in the linear structure (.85 Å RMSD from native). Thus, in every simulation the polyproline helix was formed immediately and essentially never unfolded (figure 5.4-D); at no time did the heavy atom RMSD of the polyproline helix ever exceed 1.1 Å from the native. Although the initial extended conformation was very close to the PPII conformation, the backbone RMSD for this region also never exceeded 1.1 Å during a 600K unfolding simulation (data not shown), confirming the stability of this secondary structure element. The stability of the polyproline helix lowers the entropic penalty of forming the hydrophobic core which helps to explain the folding speed and stability of the trp-cage motif.

Next, following the collapse to the compact state, the alpha and 3_{10} helices formed (Figure 5.4 graphs B, F, and E). Unlike the PPII helix, these structures were not stable and were transiently lost on several occasions before location of the native fold.

The initial fraction of native contacts (Q) (figure 5.3 C) ranges from 40-50% for the compact initial structures to 20% for the extended conformation. In each simulation, we observe a gradual increase in Q to values above 80% within the first 10 ns. Q fluctuates around a plateau at 80% until the 3_{10} helix and α helix form at which point it rises rapidly to nearly 100%. The specific contacts formed in this final stage of folding are discussed below. It is interesting to note that Q has already reached 90% by 30ns, and little change is seen during the sudden transition in NOE violations at 50 ns. Corresponding to Q approaching 90% is the arrival of a stable α helical structure (figure 5.4-E). Because the folding of the α helix satisfies the majority of the native contacts and because there were only two native contacts between residues that were separated by more than three residues (Tyr3-Pro19 and Asp9-Serine14) this native contact measure does not discern the folded from unfolded structures. Also, because there are no interstrand contacts between the residues that make up the eponymous tryptophan cage (Trp6, Pro12, and Pro18) this native contact definition cannot evaluate the progression of the folding of this key structure element.

The α helix folded and partially unfolded on at least three occasions before reaching a stable structure. The folding of the α helix is the second folding event of a piece of secondary structure event in all the simulations.

The next native-like feature adopted in the simulations is the contact between the side chains of Trp6 and that of Pro12, which forms the bottom of the “cage”. We called this sub-structure the “half-cage”. It is attained when the indole ring of the tryptophan, in the local context of the α helix, stacks on the ring of the first proline in the sequence. This sub-structure, therefore, forms after the alpha helix, but before the polyproline segment docks on to the top of the W21 side chain. The half-cage can be seen in each of the folding trajectory paths in figure 5.3. The third structure in column A and the fourth structures in columns B and C all show the half-cage. The early presence of this folding intermediate during the folding process is demonstrated in figure 5.5, which shows the RMSD for the half-cage formed by residues 3-12 against that of the full trp-cage fold (residues 3-18). Each simulation shows native-like structures for the half-cage segment (RMSD values < 1Å) in structures that have overall RMSD values up to 3Å. This demonstrates that there is a significant population of structures that have attained the half-cage but not the full cage.

After evaluating the results of a 5ns REMD simulation of trp-cage with explicit solvent Zhou proposed a semi-stable folding intermediate with a significantly different structure than the half-cage that we propose[77]. Zhou’s intermediate has two hydrophobic cores separated by an Asp9-Arg16 salt bridge. He suggests that the structure is stable until the breaking of the salt bridge allows for the rearrangement of the two hydrophobic cores into one. This model would allow for the formation of the

α helix and polyproline helix before formation of the tryptophan cage. This model does not allow for interactions between the residues in the upper hydrophobic core (residues 1-8 and 17-20) and the lower hydrophobic core (residues 10-15.)

Experimental chemical shift data support the suggestion that there is a stable intermediate in the folding path. If folding is truly two state (no stable intermediates) folding should be fully cooperative. Any significant local minima (intermediates) would be detectable as being non-cooperative. Indeed, while all of the other proline resonances move *toward* random coil values during thermal denaturation, the Pro12δ3 resonance moved upfield, *away* from random coil values[63]. This suggests non-cooperative structuring of that residue during the melting process.

While the experimental data support Zhou's contention that there is a folding intermediate, the data contradict the contention that there are two hydrophobic cores separated by an Asp9-Arg16 salt bridge. The data suggest that Pro12δ3 moves closer to Trp6 in the folding intermediate, not further from it as would be required by the two hydrophobic cores.

Examination of the interaction of the side chains of Trp6 and Pro12 in the half cage during the course of our folding simulations, whose contact stabilizes the half-cage motif, can explain this result. The distance between these side chains is shown for structures of different overall RMSD values in figure 5.6. While the relative sampling of alternate conformations differs among these non-equilibrium simulations, the patterns are strikingly similar in each case. Native conformations with RMSD values near 1.0Å all have Trp6:Pro12 contact distances above 6Å. However, a shift to shorter distances (4-5Å) between these side chains is seen in non-native structures with RMSD values above 2Å. Thus the formation of the full cage fold results in an increase in the distance between the side chains that comprise the bottom of the cage (as compared to structures with only the half-cage motif). This data suggests that the stable intermediate observed by experiment can be accounted for by the half-cage intermediate.

While each of the simulations shared a common order of events at the beginning of the simulations, the order of the final events in the folding paths diverged. The events that occurred at the end of the folding process include the arrangement of the 3_{10} helix, the formation of the Trp6hε1-Arg16O hydrogen bond, the formation of the Asp9-Arg16 salt bridge, and the docking of the polyproline helix on the top of the cage. In the 326K simulation the 3_{10} helix (residues 11-14) backbone RMSD from the native and the 3-12 backbone (shown as RMSD from the native respectively in figures 5.4-F and G) arrived at native like conformations well before the final folding event. The polyproline helix also docked early, but it docked incorrectly on top of the cage (figure 5.4-J). The rearrangement of the top of the cage allowed for the formation of the salt bridge (figure 5.4-I) and the Trp6-R16 hydrogen bond (figure 5.4-H) which completed the folding process. In the two other folding simulations the folding was completed by the docking of the polyproline helix (data not shown.)

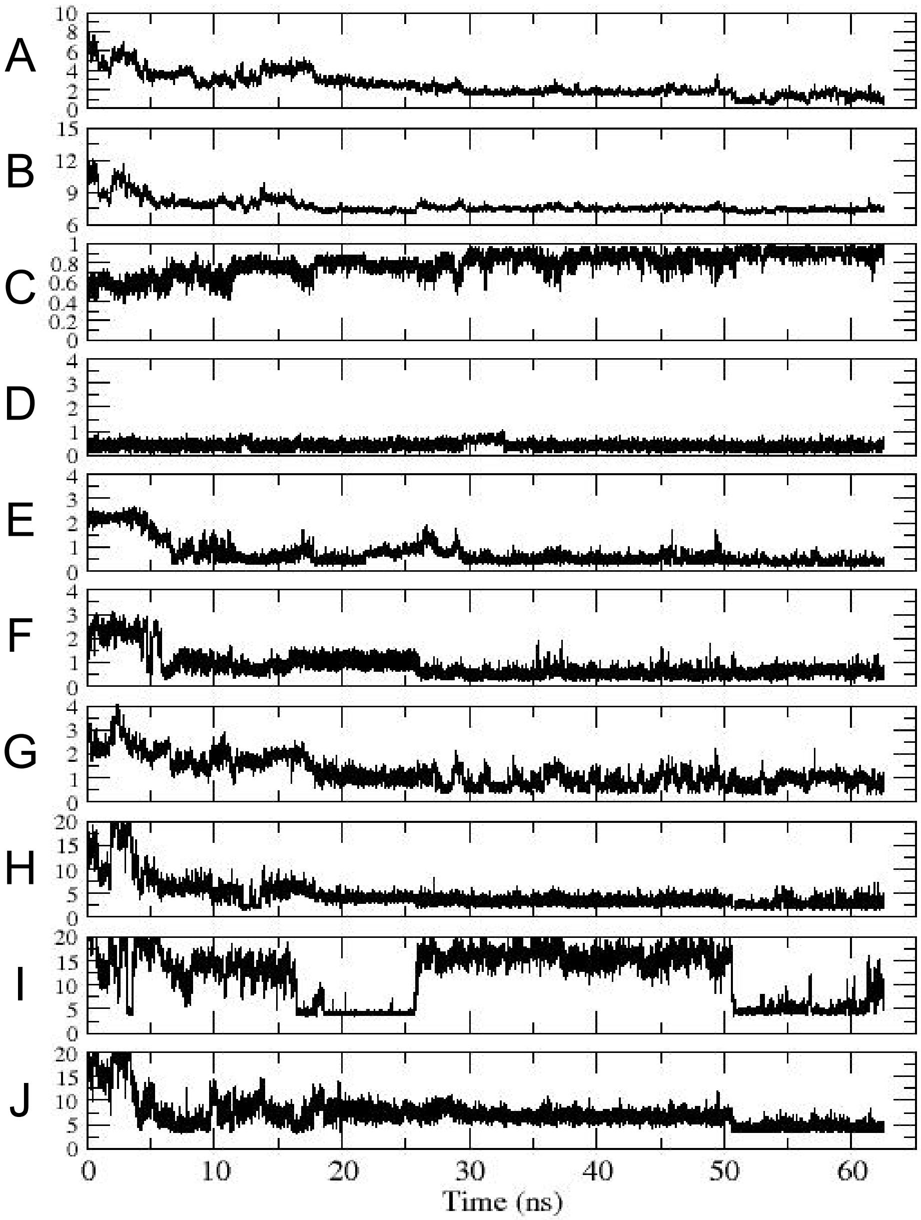


Figure 5.4 Folding Measures
Time evolution of various folding measures during folding simulations. A) 3-18 backbone RMSD from native, B) Radius of gyration in Å, C) Fraction of native contacts (Q) (Cα pairs, i to $i_{>1}$, < 6.5Å), D) Polyproline helix backbone RMSD, E) α Helix backbone RMSD, F) 3_{10} Helix backbone RMSD, G) 3-12 backbone RMSD, H) W6hε1-R16O distance, I) D9γc-R16ζc distance (salt bridge), J) W6-P18 Ring centroid distance. (This figure is discussed in the text.)

Figure 5.5 Presence of the Half-Cage Structure

These figures show the 3-19 backbone RMSDs vs. the 3-12 backbone RMSDs from the nartive. Each of the three simulations has a population of structures where the 3-12 backbone RMSD is less than 1 Å while the full RMSD is greater than 2 Å. These populations demonstrate the presence of the half-cage structure in the absence of the overall folded structure.

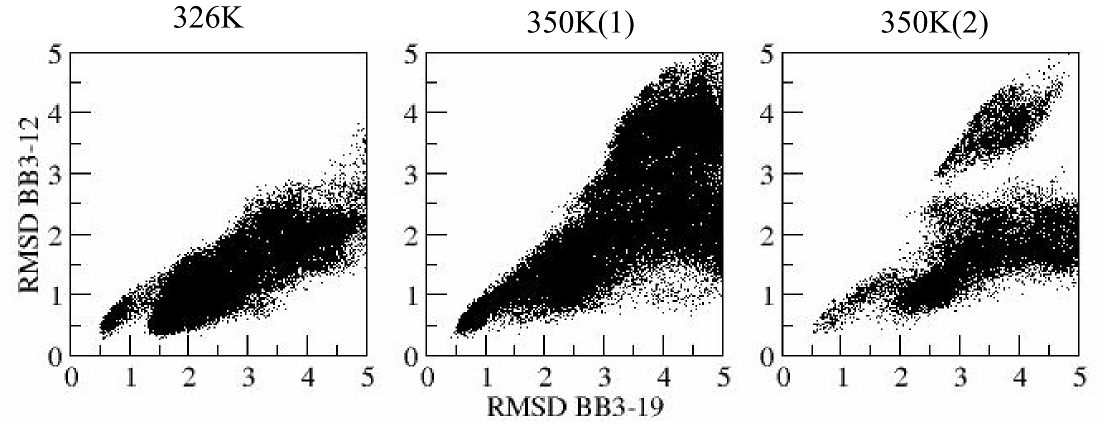
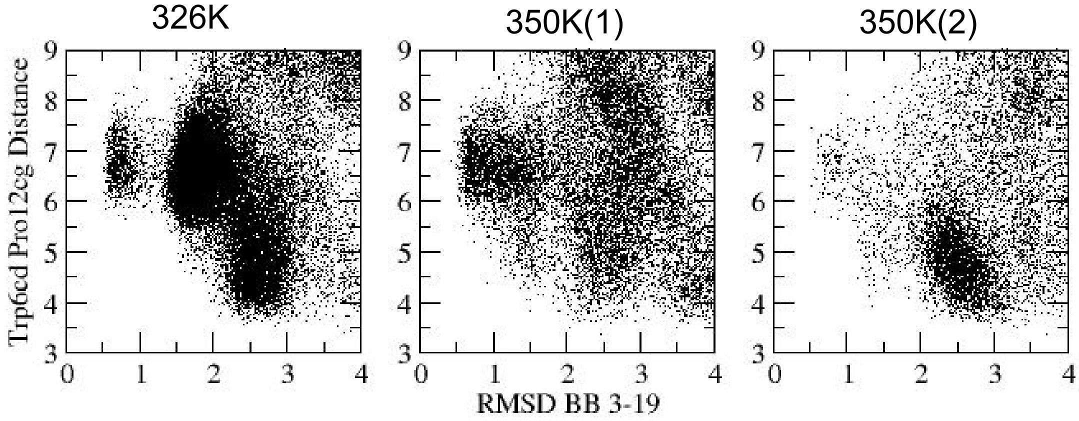


Figure 5.6 Tryptophan Proline 12 Distance

Distances between the side chains of Trp6 and Pro12 shown vs. backbone RMSD value. Each of the 3 folding simulations is shown in a different plot, with each data point corresponding to a single snapshot during that simulation. Well-folded conformations (RMSD ~1Å) have contact distances near 6Å, while structures with larger RMSD values sample significantly shorter packing distances (~4Å) for these side chains that stabilize the half-cage motif.



5.3.3 Conformation of the Trp6 Indole Ring

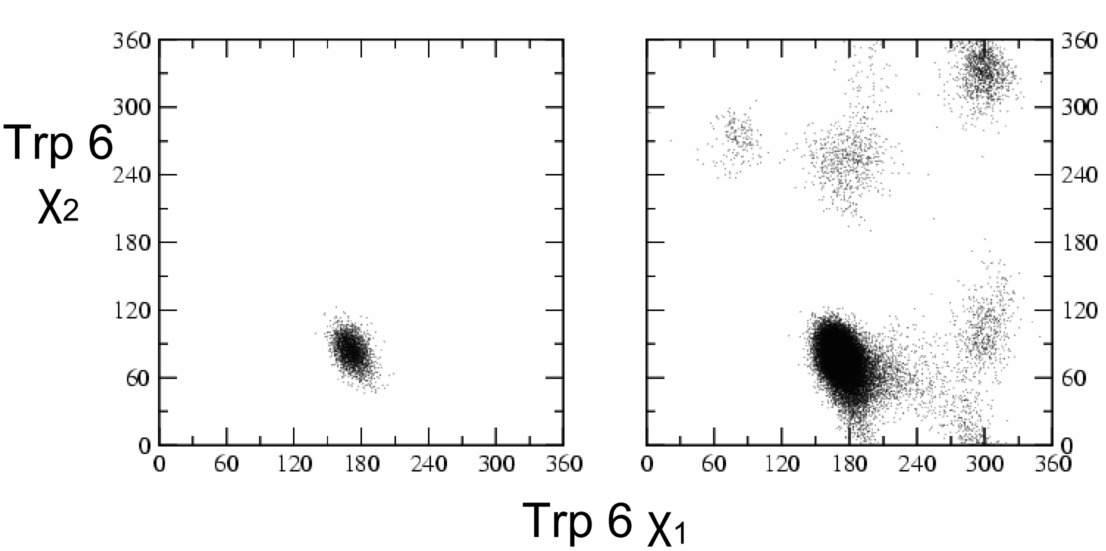
The orientation of the indole ring of the Trp6 sidechain is important for the stability of the folded structure. Before arriving at the native state, simulations on several occasions adopted structures with native-like topology but incorrect orientation of the indole ring in the proline cage. These conformations were only transiently stable and gave way to less compact states before the true native state could subsequently be located.

The importance of the orientation of the indole ring highlights the need for judicious use of RMSD as a folding measure. As discussed previously, other published reports used Cα RMSD values of 2Å [72] and 2.4Å [71] as the cutoff for defining the native state. We show in Figure 5.7 the conformation of the Trp6 side chain in two sets of structures sampled during simulations. While structures with 3-19 Cα RMSD values below 1Å all have correct indole ring conformations, a large fraction of the population with 3-19 Cα RMSD values below 2Å have the indole ring oriented incorrectly.

Chowdhury et al. described a similar folding path and suggested that the repositioning of Trp6, as measured by the fraction of Trp6 native contacts, is associated with three out of the four main folding transitions that they describe including the final folding event, the repacking of Trp6, that they call the rate limiting step [66]. While we do not find similar Trp6 interactions involved in the earlier folding events we do find that the compact nature of the core prevents rotation of the bulky indole in these structures and unfolding is required before alternate rotamers can be sampled. These events may therefore have a role in determining the folding rate, but it is important to note that because our data does not represent converged sampling of the folding path this observation does not demonstrate that this repacking is the rate-limiting step in folding.

Figure 5.7 Trp6 Indole Ring Conformation

Trp6 side-chain conformations in low RMSD structures. A single conformation is seen in well-folded structures (left, all structures with backbone RMSD<1Å) while near-native backbone conformations often trap the side-chain in incorrect rotamers (right, all structures with backbone RMSD<2Å).



5.3.4 Water Molecule Participation in the Trp-Cage Structure

As a measure of the formation of the hydrophobic core we measured the number of waters around the Trp6 sidechain during the course of the 326K simulation (see figure 5.8). The Trp6 was chosen because of its position at the center of the hydrophobic core and a cutoff of 5Å was chosen to allow measurement of several waters without the undue counting of the bulk solvent. Over the course of the simulation the number of waters around the tryptophan drops from greater than 30 to less than 14. The striking feature of the process is that the expulsion of water closely parallels the changes in RMSD. Every significant step toward the native structure occurs with a reduction in the number of waters near the tryptophan.

As discussed by Rhee et al. [84], two notably different theories about the role of the water in the hydrophobic core during the folding process have been proposed. ten Wolde and Chandler used a coarse grained model to investigate the hydrophobic collapse of a non-protein polymer and concluded that the evaporation of the waters around the hydrophobic core drives the hydrophobic collapse[85]. In contrast, Sheinerman and Brooks used molecular dynamics simulations to investigate the hydrophobic collapse of the B1 segment of the streptococcal protein G with explicit solvent and found that the collapse of the core expelled the waters[86]. Our system, interestingly, suggested neither of these mechanisms. If the leaving of the waters was causing the hydrophobic collapse we would expect a decrease in the numbers of waters around the core residues prior to the structural collapse and if the collapse were expelling the waters we would expect that the initiation of the collapse would consistently occur prior to a reduction in the number of waters local to the core. After looking at each major collapse of the hydrophobic cores in two of the folding simulations (326K and 350K from the native) we find that neither event consistently precedes the other. In figure 5.8-B we show one such collapse that takes part in two phases. In the first phase the change in the structure precedes the expulsion of water from around Trp6 and in the second phase the leaving of the waters precedes the structural change.

The arrival of the simulation at a stable structure allows the investigation of structured waters. To see if any waters were structured in the folded simulation we used the Ptraj module from the Amber 8 suite of programs to count the number of waters (counted by the location of the center of the oxygen in each water) per box in a 0.5 Å grid that covered the volume accessible to the simulation. The waters were counted for a 1ns section of the trajectory that started at the structure with the lowest RMSD from the native. During this trajectory the RMSD from the native never rose above 1Å. The fraction of the frames where a box contained a water molecule was used to calculate a "density" of the water through time. The density was used to create a map of the areas where the water density was higher than bulk solvent and these areas are displayed around a surface map of trp-cage in figure 5.9. Because the core of the protein is relatively stable there is not a clearly available path for water to enter the core, we would expect that a water structured inside the protein would appear as a region of density higher than that of bulk solvent. No such density was found inside the structure suggesting that the structure of the core of the protein doesn't rely on structured waters.

Figure 5.8 Desolvation of the Tryptophan

A) The blue line below shows the number of waters within 5Å of Trp6 during the 326K folding simulation shown as a running average over 100 structures. As a reference the 3-19 Cα backbone RMSD is shown in black. B) The blue line is the number of waters with 5Å from Trp6, without a running average, for the folding event after 50ns in the same simulation as in figure A. Again the black line is the 3-19 Cα backbone RMSD. As the trp-cage structure approaches more native-like conformations the number of waters near Trp6 declines suggesting the desolvation of the hydrophobic core.

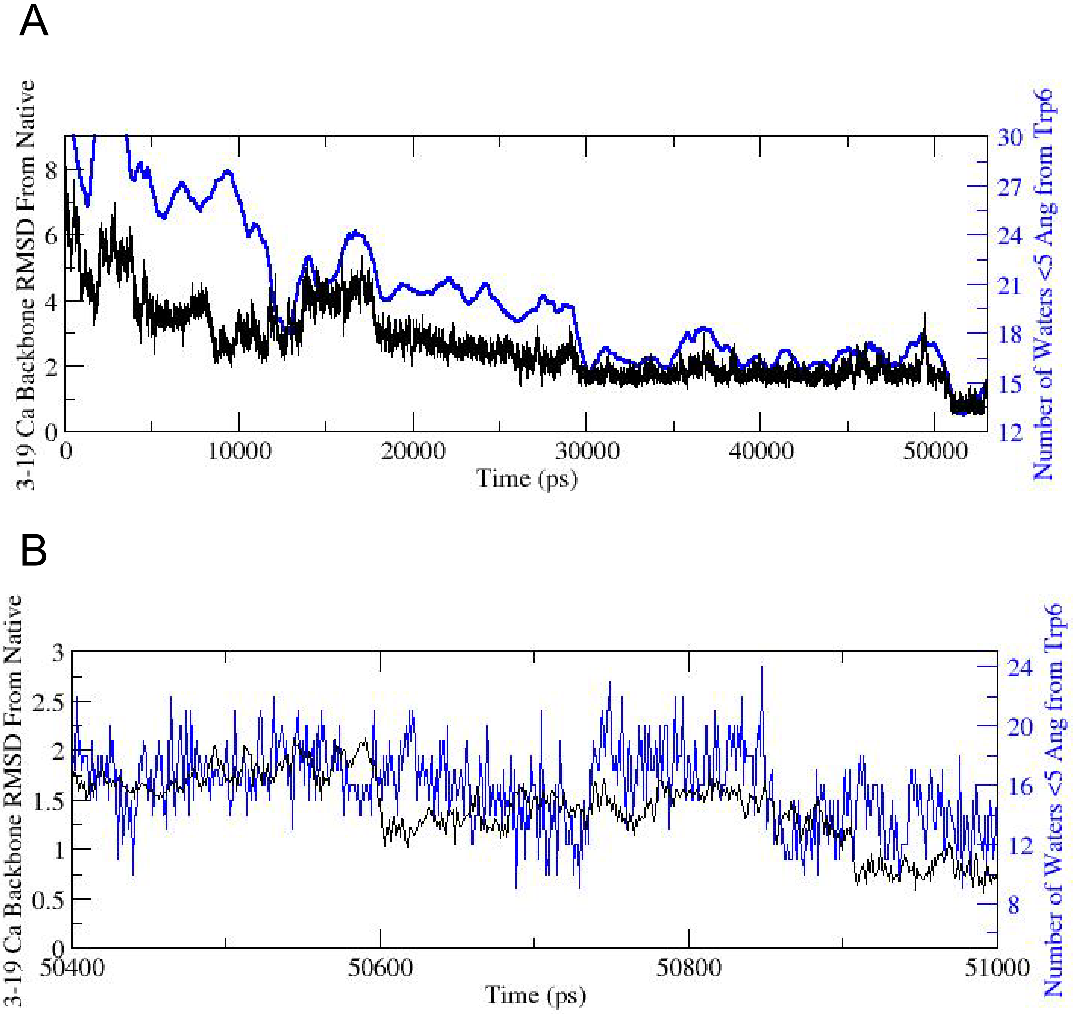
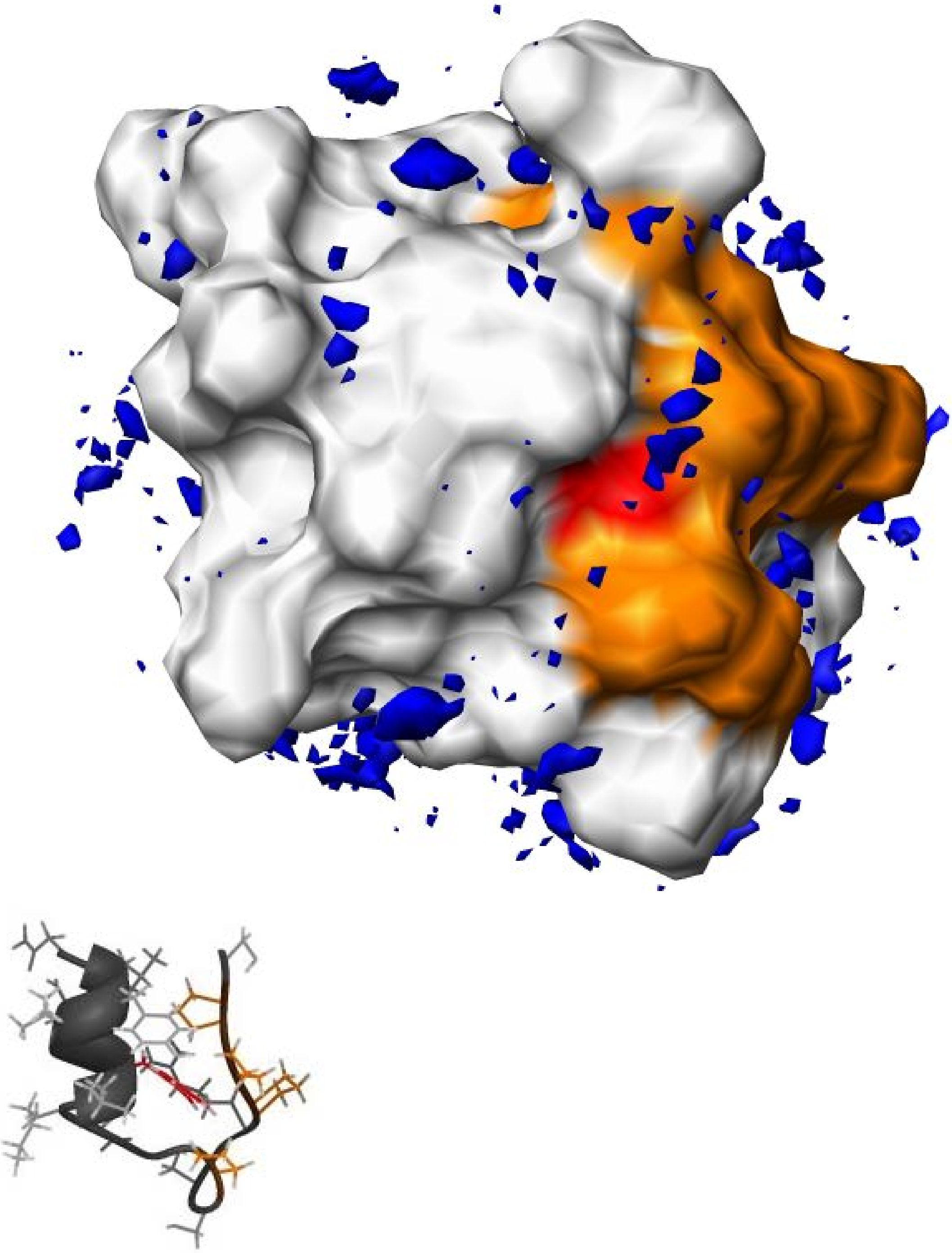


Figure 5.9 Water Density Through Time
This figure shows the solvent accessible surface area (with a probe radius of 1.4Å) of the protein surrounded by the locations of high water densities. The protein surface is white, the proline surfaces is colored orange, the tryptophan surface is red, the water density is colored blue, and the amino terminus is in the upper left hand corner. A reference structure is shown on the lower left.



5.3.5 Comparing Folding in Explicit and Implicit Solvent Models

Because these simulations were run in a similar manner to the implicit solvent simulations we can examine the effects of the different solvent models on the folding process. Although friction isn't entirely due to the solvent, Qui et al. demonstrated that internal friction has a meaningful effect on the folding rate [87], the increased friction expected in the explicit model increased the time it took for the linear structure to collapse to a compact state. With Generalized Born (GB), the radius of gyration reached a native-like value of ~8Å after only 25ps, while this process took over 4ns in explicit solvent (experimentally the folding time has been measured at 4μs [64].) The time it took for the structures to fold from the compact state to the native state was also significantly longer in simulations with the explicit solvent model. In the GB simulations the time from collapse to finding a structure under 1 Å RMSD from the native state took 9 ns while the explicit solvent simulations took greater than 52 ns. Also, the salt bridge population for the Asp9-Arg16 ion pair was significantly higher in the GB simulations (75%) than the explicit solvent simulations (25%). This result could suggest inappropriate treatment of salt bridges in the GB solvent model. This result would be consistent with a recent study by Zhou and Berne on the effect of solvent models on free energy landscapes that suggested that GB models do overstabilize salt bridges and has overstabilized them enough to change the global free energy minimum of the C-terminal β hairpin fragment of protein G[88].

5.4 Discussion

The simulation and analyses presented here show that it is possible to directly simulate the folding process at relevant temperatures with full atomic detail, for both protein and solvent, without the aid of enhanced-sampling techniques. Starting from a number of different initial conditions, we have been able to observe, in a direct dynamical sense, the complete pathway from unfolded to folded conformations, including some transiently populated, non-thermodynamically populated misfolded structures. Such simulations have the potential to provide unique insights into complex process of protein folding. In the present case, we observed that the adoption of native packing in the core of trp-cage was preceded by formation of the α helical secondary structure and the adoption of a semi-stable "half-cage" intermediate.

Older theories on protein folding suggested that proteins folded by a specific step by step process. Newer views suggest that that proteins fold via many different paths with, potentially, common intermediates. Interestingly, this protein folds according to the old *and* the new views of protein folding: the sequence of events is almost unique in the early stages, but the final steps show significant variation among trajectories. This is slightly contrary to expectations. Because of the large number of possible unfolded structures and the smaller number of folded structures one might expect the variations to be found earlier rather than later in the folding paths. In this instance, we observed that the conformational flexibility of the system allowed local energetically favorable events to happen early in the trp-cage folding sequence.

Chapter 6. Studies of the Exendin Protein and its Derivatives

6.1 Biological Significance of Exendin4 and Exenetin

Exendin-4 is a peptide identified from Gila Monster Heloderma Suspectum saliva[59] that has several biological activities that suggest potential use as a therapeutic agent for type II diabetes[89]. Greater than 6% of the United States population displays the clinical manifestation of Diabetes[90]. The clinical manifestation of diabetes is abnormally high blood glucose levels. The body normally keeps the circulating sugar levels in check by secreting insulin. In the diabetic state the body either produces insufficient levels of insulin, no insulin, or has become resistant to the secreted insulin[91].

The biological activites of exendin-4 include the promotion of β-cell neogenisis[92, 93], glucose-dependent stimulation of insulin secretion[94] with a concomitant reduction in blood glucose levels, inhibition of gastric emptying, and an inhibition of food intake[95]. Synthetic exendin-4, called Exenatide, from Amylin Pharmaceuticals completed phase III clinical trails as an anti-diabetes therapeutic in 2004[96] and reached the market in 2005 as the drug Byetta[97].

6.1.1 GLP-1

Glucagon-like polypeptide 1 (GLP-1) is a regulatory peptide in the family of incretins processed from the proglucagon gene and released from the intestinal L-cells in the gut in response to food intake[98]. GLP-1 causes the release of insulin from the pancreatic β-cells and inhibits the release of glucagons. GLP-1 acts by binding to its receptor, GLP-1r, and is degraded by the dipeptidyl-peptidase-IV (DPP-IV)[99] with a circulating half life of two minutes[100]. Exendin-4 acts as a structural mimetic of GLP-1 [101], binding to the same receptor, and was chosen specifically for its improved physiological stability relative to GLP-1[100]. The N-terminal alpha helical regions of GLP-1 and exendin-4 are important for binding to the receptor GLP-1r[102]. The N-terminal region of exendin-4 doesn’t bind to the receptor as strongly as the N-terminal region of GLP-1; the C-terminal region of exendin-4 binds as a compensating mechanism for the weaker binding of the N-terminal region[103]. This highlights the importance of the C-terminal region of exendin-4.

6.2 Simulation and Construct Rationale

Because the 20 amino acids of trp-cage are 80% identical (see table 6.1) to the last 20 amino acids of exendin-4 we hypothesized that a protein composite made from the N-terminal extended helix from exendin-4 and the tryptophan cage from trp-cage would have a tertiary structure similar to that of exendin-4 but would exhibit increased stability and faster folding than exendin-4. To test this we analyzed folding simulations of both exendin-4 and the composite protein.

6.3 Simulation Details

We started 16 simulations of each of the exendin-4 and the exendin-4/trp-cage 5b chimera (ex4-5b). The simulations were run each started from the extended state with the GB solvent model and the Mod2 force field with the Amber 7 [78] molecular modeling suite of programs. The simulations were minimized with 10 steps of steepest descent and 490 steps of conjugate gradient energy minimization to eliminate initial strained structures. Following minimization the simulations were run at 350K for approximately 100ns each on average with individual times ranging from 50ns to 315ns. The simulations often folded to RMSD values below 4Å but the majority of their structures were above 9Å RMSD from the native exendin-4 (79.4% for ex4-5b, 67.83% for exendin-4.) See figure 6.1.

6.4 Simulation Analyses

Because we wanted to evaluate the propensity of the chimeric protein to adopt structures similar to that of exendin-4 and because there is no known experimental structure of the chimeric protein, we used the Cα RMSD from the exendin-4 structure as the measure of the ex4-5b folding. The exendin-4 simulations were analyzed in the same manner to allow a comparison between the simulations.

6.5 Discussion

Detailed investigations of the salt bridge stability by Raphael Geney, of our lab, convinced us that salt bridges are over-stabilized by the particular GB solvation model we used. As the structures of ex4-5b can have 5 or more salt bridges in the disordered state (while the predicted conformation is expected to have only one) we suspected that the non-exendin-4 like families that were heavily populated in our simulations were being favored in response to the excessive salt bridge stability. Indeed, upon visual examination many of the non-native-like structures had several salt bridges.

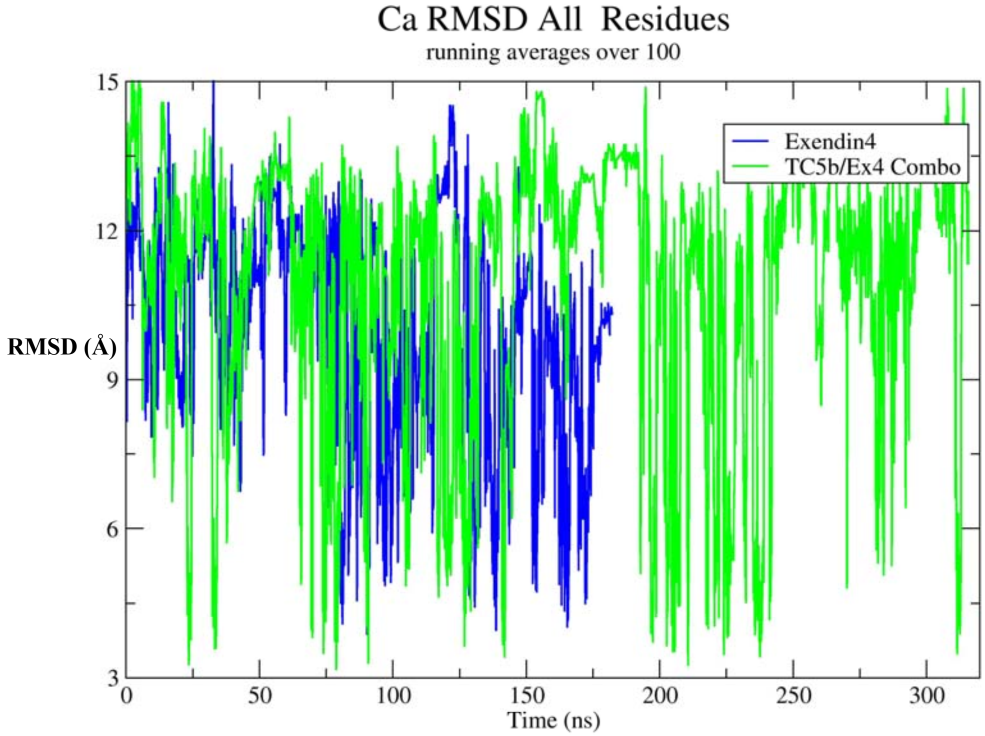
Furthermore, a paper by Al-Sabah and Donnely[104] and personal communications with Dr. Dan Donnely of the School of Biomedical Sciences at the University of Leeds, United Kingdom, demonstrated experimentally that a chimeric protein very similar to the one we proposed EX-4:Ala-2:TC5a (different only at residue 20, see table 6.1) bound less stably to the GLP1r (exendin-4 pIC_{50} 9.1 ± 0.1 M^{-1}, ex4-5b pIC_{50} 8.4 ± 0.4 M^{-1} (mean ± S.E. of three experiments) than the wild type exendin-4 in competition assays with the GLP1r antagonist ^{125}I-exendin(9-39). With the likelihood of improving the compound diminished and with the treatment of the salt bridges cast in doubt we discontinued the simulations.

Table 6.1 Amino Acid Sequences of GLP-1, EX-4, Trp-Cage, EX-4-5b, and EX-4:Ala-2:TC5a

GLP-1			
HAEGTFTSDV^{10}	**SSYLEGQAAK**^{20}	**EFIAWLVKGR**^{30}	
EXENDIN-4			
HGEGTFTSDL^{10}	**SKQMEEEAVR**^{20}	**LFIEWLKNGG**^{30}	**PSSGAPPPS**^{39}
TRPCAGE 5B			
	N^{1}	**LYIQWLKDGG**^{11}	**PSSGRPPPS**^{20}
EXENDIN 4 – TCAGE 5B CHIMERA			
HGEGTFTSDL^{10}	**SKQMEEEAVN**^{20}	**LYIQWLKDGG**^{30}	**PSSGRPPPS**^{39}
EX-4:Ala2:TC5a			
HAEGTFTSDL^{10}	**SKQMEEEAVR**^{20}	**LFIQWLKDGG**^{30}	**PSSGRPPPS**^{39}

Figure 6.1Exendin-4 and Ex4-5b Simulation

The RMSD of the Exendin-4 simulation and the Ex4-5b simulation from the native exendin-4 structure. Neither simulation sampled structures within 3Å of the native Exendin4 structure.



Concluding Remarks

The first study presented in this dissertation discussed the problem of an α-helical bias in the then current parameter set of the Amber forcefield. The root of that problem was identified and methods for developing a new parameter set were presented. A new force field parameter set developed with these tools, Mod2, was used successfully in much of this work, but limitations of this set were discovered. Other work confirmed our inappropriate treatment of the glycine parameters and suggests a future means from improving the parameter sets. Newer parameter sets, currently being developed in the Simmerling lab, will resolve this problem by treating the glycine parameters separately from the parameters for the other amino acids.

In the second project presented here, the Mod2 parameter set was used to predict the three dimensional structure of the mini-protein trp-cage before the release of the experimentally derived structure coordinates. To our knowledge this is the first time that the structure of such a complicated molecule has been predicted to such a high resolution without previous knowledge of the structure. Trp-cage is a very small and fast folding protein, these attributes made it an attractive target for folding simulations but significantly improved computer performance will be necessary before folding of larger more typical proteins can be simulated.

To better understand the process of trp-cage folding the third project presented here investigated folding events of trp-cage with the use of the TIP3P explicit solvent model. These simulations demonstrated that simulations of the full folding process from fully unfolded structures to structures that were essentially indistinguishable from the experimentally derived ensemble of folded structures are possible. Further we observed that the adoption of native packing in the core of trp-cage was preceded by formation of the α helical secondary structure and the adoption of a semi-stable "half-cage" intermediate. Future studies of this system will require an improved understanding of the folding energy surface; improved sampling techniques such as Hybrid Implicit/Explicit Solvent Replica Exchange[105] that are currently being developed will allow exactly these explorations.

Finally the last study presented here heralds both the potential of these tools and their limitations. This study intended to explore the relationship between the stability of the N-terminal extended helix and the C-terminal region of the biologically important exendin-4 protein by incorporating the highly stable tryptophan cage from trp-cage into the exendin-4 C-terminal region. Ultimately it became apparent that the solvent model we were using over-stabilized salt bridges in the structures we were sampling and that our results would not be useful for the intended comparison. Dr. Raphael Geney in the Simmerling Lab is currently working towards an improved treatment of salt bridges in Generalized Born solvent models and with those tools this project might well be reconsidered.

References

1. Dobson CM: **Protein-misfolding diseases: Getting out of shape**. *Nature* 2002, **418**(6899):729-730.

2. Goldberger RF, Epstein CJ, Anfinsen CB: **Acceleration of Reactivation of Reduced Bovine Pancreatic Ribonuclease by a Microsomal System from Rat Liver**. *Journal of Biological Chemistry* 1963, **238**(2):628-&.

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

4. Chothia C, Finkelstein AV: **The Classification and Origins of Protein Folding Patterns**. *Annu Rev Biochem* 1990, **59**:1007-1039.

5. Flory PJ: **Statistical Mechanics of Chain Molecules**: Wiley; 1969.

6. Levinthal C: **How to Fold Graciously**. In: *Mossbauer Spectroscopy in Biological Systems: 1969; Allerton House, Illinois*: University of Illinois Press; 1969.

7. Dill KA, Chan HS: **From Levinthal to pathways to funnels**. *Nature Structural Biology* 1997, **4**(1):10-19.

8. Levinthal C: **Are There Pathways for Protein Folding**. *J Chim Phys Pcb* 1968, **65**(1):44-&.

9. Campbell ID: **Timeline: the march of structural biology**. *Nat Rev Mol Cell Biol* 2002, **3**(5):377-381.

10. Rhodes G: **Crystallography Made Crystal Clear**: Academic Press; 2000.

11. Brandon C, Tooze J: **Introduction to Protein Structure**: Garland Science; 1998.

12. Wuthrich K: **NMR of Proteins and Nucleic Acids**. New York: Wiley; 1986.

13. Fisher TE, Marszalek PE, Fernandez JM: **Stretching single molecules into novel conformations using the atomic force microscope**. *Nat Struct Biol* 2000, **7**(9):719-724.

14. Marszalek PE, Lu H, Li HB, Carrion-Vazquez M, Oberhauser AF, Schulten K, Fernandez JM: **Mechanical unfolding intermediates in titin modules**. *Nature* 1999, **402**(6757):100-103.

15. Engel A, Muller DJ: **Observing single biomolecules at work with the atomic force microscope**. *Nat Struct Biol* 2000, **7**(9):715-718.

16. Sporlein S, Carstens H, Satzger H, Renner C, Behrendt R, Moroder L, Tavan P, Zinth W, Wachtveitl J: **Ultrafast spectroscopy reveals subnanosecond peptide conformational dynamics and validates molecular dynamics simulation**. *Proc Natl Acad Sci U S A* 2002, **99**(12):7998-8002.

17. Woutersen S, Mu Y, Stock G, Hamm P: **Subpicosecond conformational dynamics of small peptides probed by two-dimensional vibrational spectroscopy**. *Proc Natl Acad Sci U S A* 2001, **98**(20):11254-11258.

18. Greenfie.N, Davidson B, Fasman GD: **Use of Computed Optical Rotatory Dispersion Curves for Evaluation of Protein Conformation**. *Biochemistry* 1967, **6**(6):1630-&.

19. Ha T, Enderle T, Ogletree DF, Chemla DS, Selvin PR, Weiss S: **Probing the interaction between two single molecules: Fluorescence resonance energy transfer between a single donor and a single acceptor**. *P Natl Acad Sci USA* 1996, **93**(13):6264-6268.

20. Haustein E, Schwille P: **Single-molecule spectroscopic methods**. *Curr Opin Struct Biol* 2004, **14**(5):531-540.

21. Zhuang XW, Rief M: **Single-molecule folding**. *Curr Opin Struc Biol* 2003, **13**(1):88-97.

22. Holzwarth JF, Schmidt A, Wolff H, Volk R: **Nanosecond Temperature-Jump Technique with an Iodine Laser**. *Journal of Physical Chemistry* 1977, **81**(24):2300-2301.

23. Shea JE, Brooks CL, 3rd: **From folding theories to folding proteins: a review and assessment of simulation studies of protein folding and unfolding**. *Annu Rev Phys Chem* 2001, **52**:499-535.

24. Daggett V, Fersht A: **The present view of the mechanism of protein folding**. *Nat Rev Mol Cell Biol* 2003, **4**(6):497-502.

25. Newton SI: **Philosophiae Naturalis Principia Mathematica**. 1687.

26. Verlet L: **Computer Experiments on Classical Fluids .I. Thermodynamical Properties of Lennard-Jones Molecules**. *Phys Rev* 1967, **159**(1):98-&.

27. D.A. Case DAP, J.W. Caldwell, T.E. Cheatham III, W.S. Ross, C.L. Simmerling, T.A. Darden, K.M. Merz, R.V. Stanton, A.L. Cheng, J.J. Vincent, M. Crowley, V. Tsui, R.J. Radmer, Y. Duan, J. Pitera, I. Massova, G.L. Seibel, U.C. Singh, P.K. Weiner and P.A. Kollman: **AMBER 6**: University of California, San Francisco; 1999.

28. Bayly CI, Cieplak P, Cornell WD, Kollman PA: **A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges - the Resp Model**. *Journal of Physical Chemistry* 1993, **97**(40):10269-10280.

29. Wang JM, Cieplak P, Kollman PA: **How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?** *J Comput Chem* 2000, **21**(12):1049-1074.

30. Weiner SJ, Kollman PA, Case DA, Singh UC, Ghio C, Alagona G, Profeta S, Weiner P: **A New Force-Field for Molecular Mechanical Simulation of Nucleic-Acids and Proteins**. *J Am Chem Soc* 1984, **106**(3):765-784.

31. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA: **A Second Generation Force Field For the Simulation of Proteins, Nucleic Acids, and Organic Molecules**. *J Am Chem Soc* 1995, **117**(19):5179-5197.

32. Cheatham TE, Cieplak P, Kollman PA: **A modified version of the Cornell et al. force field with improved sugar pucker phases and helical repeat**. *J Biomol Struct Dyn* 1999, **16**(4):845-862.

33. Leach AR: **Molecular Modelling: Principles and Applications**, 2nd Edition edn. Essex: Pearson Education Limited; 2001.

34. Kabsch W: **Discussion of Solution for Best Rotation to Relate 2 Sets of Vectors**. *Acta Crystallogr A* 1978, **34**(SEP):827-828.

35. Orozco M, Luque FJ: **Theoretical methods for the description of the solvent effect in biomolecular systems**. *Chem Rev* 2000, **100**(11):4187-4225.

36. Zagrovic B, Pande V: **Solvent viscosity dependence of the folding rate of a small protein: distributed computing study**. *J Comput Chem* 2003, **24**(12):1432-1436.

37. Cramer CJ, Truhlar DG: **Implicit solvation models: Equilibria, structure, spectra, and dynamics**. *Chem Rev* 1999, **99**(8):2161-2200.

38. Still WC, Tempczyk A, Hawley RC, Hendrickson T: **Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics**. *J Am Chem Soc* 1990, **112**(16):6127-6129.

39. Tsui V, Case DA: **Theory and applications of the generalized Born solvation model in macromolecular simulations**. *Biopolymers* 2000, **56**(4):275-291.

40. Simonson T: **Macromolecular electrostatics: continuum models and their growing pains**. *Curr Opin Struc Biol* 2001, **11**(2):243-252.

41. Bashford D, Case DA: **Generalized born models of macromolecular solvation effects**. *Annual Review of Physical Chemistry* 2000, **51**:129-152.

42. Onufriev A, Bashford D, Case DA: **Modification of the generalized Born model suitable for macromolecules**. *J Phys Chem B* 2000, **104**(15):3712-3720.

43. Simmerling C, Strockbine B, Roitberg AE: **All-atom structure prediction and folding simulations of a stable protein**. *J Am Chem Soc* 2002, **124**(38):11258-11259.

44. Nymeyer H, Garcia AE: **Simulation of the folding equilibrium of alpha-helical peptides: A comparison of the generalized born approximation with explicit solvent**. *P Natl Acad Sci USA* 2003, **100**(24):13934-13939.

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

46. Shen MY, Freed KF: **Long time dynamics of met-enkephalin: Comparison of explicit and implicit solvent models**. *Biophys J* 2002, **82**(4):1791-1808.

47. Jorgensen WL: **Transferable Intermolecular Potential Functions for Water, Alcohols, and Ethers - Application to Liquid Water**. *J Am Chem Soc* 1981, **103**(2):335-340.

48. Jorgensen WL, Chandrasekhar J, Madura JD: **Comparison of simple potential functions for simulating liquid water**. *J Chem Phys* 1983, **79**(2):926-935.

49. Ferguson DM: **Parameterization and Evaluation of a Flexible Water Model**. *J Comput Chem* 1995, **16**(4):501-511.

50. Paschek D: **Temperature dependence of the hydrophobic hydration and interaction of simple solutes: An examination of five popular water models**. *J Chem Phys* 2004, **120**(14):6674-6690.

51. Spector S, Raleigh DP: **Submillisecond folding of the peripheral subunit-binding domain**. *J Mol Biol* 1999, **293**(4):763-768.

52. Cochran AG, Skelton NJ, Starovasnik MA: **Tryptophan zippers: Stable, monomeric beta-hairpins**. *P Natl Acad Sci USA* 2001, **98**(10):5578-5583.

53. Demarest SJ, Hua YX, Raleigh DP: **Local interactions drive the formation of nonnative structure in the denatured state of human alpha-lactalbumin: A high resolution structural characterization of a peptide model in aqueous solution**. *Biochemistry* 1999, **38**(22):7380-7387.

54. Okur A, Strockbine B, Hornak V, Simmerling C: **Using PC clusters to evaluate the transferability of molecular mechanics force fields for proteins**. *J Comput Chem* 2003, **24**(1):21-31.

55. Gould IR, Cornell WD, Hillier IH: **A Quantum-Mechanical Investigation of the Conformational Energetics of the Alanine and Glycine Dipeptides in the Gas-Phase and in Aqueous-Solution**. *J Am Chem Soc* 1994, **116**(20):9250-9256.

56. Beachy MD, Chasman D, Murphy RB, Halgren TA, Friesner RA: **Accurate ab initio quantum chemical determination of the relative energetics of peptide conformations and assessment of empirical force fields**. *J Am Chem Soc* 1997, **119**(25):5908-5920.

57. Wang JM, Kollman PA: **Automatic parameterization of force field by systematic search and genetic algorithms**. *J Comput Chem* 2001, **22**(12):1219-1228.

58. Garcia AE, Sanbonmatsu KY: **Exploring the energy landscape of a beta hairpin in explicit solvent**. *Proteins* 2001, **42**(3):345-354.

59. Eng J, Kleinman WA, Singh L, Singh G, Raufman JP: **Isolation and characterization of exendin-4, an exendin-3 analogue, from Heloderma suspectum venom. Further evidence for an exendin receptor on dispersed acini from guinea pig pancreas**. *J Biol Chem* 1992, **267**(11):7402-7405.

60. Kellog R: **The Giant Gila Monster**. In. U.S.A.; 1959.

61. Neidigh JW, Fesinmeyer RM, Prickett KS, Andersen NH: **Exendin-4 and glucagon-like-peptide-1: NMR structural comparisons in the solution and micelle-associated states**. *Biochemistry* 2001, **40**(44):13188-13200.

62. Barua B, Andersen NH: **Determinants of miniprotein stability: can anything replace a buried H-bonded Trp sidechain?** *Letters in Peptide Science* 2001, **8**(3-5):221-226.

63. Neidigh JW, Fesinmeyer RM, Andersen NH: **Designing a 20-residue protein**. *Nat Struct Biol* 2002, **9**(6):425-430.

64. Qiu L, Pabit SA, Roitberg AE, Hagen SJ: **Smaller and faster: the 20-residue Trp-cage protein folds in 4 micros**. *J Am Chem Soc* 2002, **124**(44):12952-12953.

65. Gellman SH, Woolfson DN: **Mini-proteins Trp the light fantastic**. *Nat Struct Biol* 2002, **9**(6):408-410.

66. Chowdhury S, Lee MC, Xiong G, Duan Y: **Ab initio folding simulation of the Trp-cage mini-protein approaches NMR resolution**. *J Mol Biol* 2003, **327**(3):711-717.

67. Neidigh JW, Fesinmeyer RM, Andersen NH: **Designing a 20-residue protein**. *Nat Struct Biol* 2002, **9**(6):425-430.

68. Osapay K, Case DA: **A New Analysis of Proton Chemical-Shifts in Proteins**. *J Am Chem Soc* 1991, **113**(25):9436-9444.

69. Andersen NH: **Personal Communication**. In.; 2002.

70. Moult J, Fidelis K, Zemla A, Hubbard T: **Critical assessment of methods of protein structure prediction (CASP): Round IV**. *Proteins-Structure Function and Genetics* 2001:2-7.

71. Snow CD, Zagrovic B, Pande VS: **The Trp cage: folding kinetics and unfolded state topology via molecular dynamics simulations**. *J Am Chem Soc* 2002, **124**(49):14548-14549.

72. Pitera JW, Swope W: **Understanding folding and design: replica-exchange simulations of "Trp-cage" miniproteins**. *Proc Natl Acad Sci U S A* 2003, **100**(13):7587-7592.

73. Gnanakaran S, Nymeyer H, Portman J, Sanbonmatsu KY, Garcia AE: **Peptide folding simulations**. *Curr Opin Struct Biol* 2003, **13**(2):168-174.

74. Alonso DO, Daggett V: **Molecular dynamics simulations of protein unfolding and limited refolding: characterization of partially unfolded states of ubiquitin in 60% methanol and in water**. *J Mol Biol* 1995, **247**(3):501-520.

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

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

77. Zhou R: **Trp-cage: folding free energy landscape in explicit water**. *Proc Natl Acad Sci U S A* 2003, **100**(23):13280-13285.

78. Case DA, Pearlman DA, Caldwell JA, Cheatham TE, Wang J, Ross WS, Simmerling CL, Darden TA, Merz KM, Stanton RV *et al*: **AMBER 7**. In.: University of California, San Francisco; 2002.

79. Darden T, York D, Pedersen L: **Particle mesh Ewald: An *N*log(N) method for Ewald sums in large systems**. *J Chem Phys* 1993, **98**(12):10089-10092.

80. Zhang YZ, Paterson Y, Roder H: **Rapid Amide Proton-Exchange Rates in Peptides and Proteins Measured by Solvent Quenching and 2-Dimensional Nmr**. *Protein Science* 1995, **4**(4):804-814.

81. Karplus M, Weaver DL: **Protein folding dynamics: the diffusion-collision model and experimental data**. *Protein Sci* 1994, **3**(4):650-668.

82. Karplus M, Weaver DL: **Protein-Folding Dynamics**. *Nature* 1976, **260**(5550):404-406.

83. Islam SA, Karplus M, Weaver DL: **Application of the diffusion-collision model to the folding of three-helix bundle proteins**. *J Mol Biol* 2002, **318**(1):199-215.

84. Rhee YM, Sorin EJ, Jayachandran G, Lindahl E, Pande VS: **Simulations of the role of water in the protein-folding mechanism**. *P Natl Acad Sci USA* 2004, **101**(17):6456-6461.

85. ten Wolde PR, Chandler D: **Drying-induced hydrophobic polymer collapse**. *P Natl Acad Sci USA* 2002, **99**(10):6539-6543.

86. Sheinerman FB, Brooks CL: **Calculations on folding of segment B1 of streptococcal protein G**. *J Mol Biol* 1998, **278**(2):439-456.

87. Qiu LL, Hagen SJ: **Internal friction in the ultrafast folding of the tryptophan cage**. *Chem Phys* 2004, **307**(2-3):243-249.

88. Zhou R, Berne BJ: **Can a continuum solvent model reproduce the free energy landscape of a beta -hairpin folding in water?** *Proc Natl Acad Sci U S A* 2002, **99**(20):12777-12782.

89. Fineman MS, Bicsak TA, Shen LZ, Taylor K, Gaines E, Varns A, Kim D, Baron AD: **Effect on glycemic control of exenatide (synthetic exendin-4) additive to existing metformin and/or sulfonylurea treatment in patients with type 2 diabetes**. *Diabetes Care* 2003, **26**(8):2370-2377.

90. Services USDoHaH: **National diabetes fact sheet: general information and national estimates on diabetes in the United States**. In. Atlanta, GA: Centers for Disease Control and Prevention; 2003.

91. et.al. AA: **Endotext**. In. www.endotext.com: MDText.com; 2005.

92. Tourrel C, Bailbe D, Lacorne M, Meile MJ, Kergoat M, Portha B: **Persistent improvement of type 2 diabetes in the Goto-Kakizaki rat model by expansion of the beta-cell mass during the prediabetic period with glucagon-like peptide-1 or exendin-4**. *Diabetes* 2002, **51**(5):1443-1452.

93. Brubaker PL, Drucker DJ: **Minireview: Glucagon-like peptides regulate cell proliferation and apoptosis in the pancreas, gut, and central nervous system**. *Endocrinology* 2004, **145**(6):2653-2659.

94. Idris I, Patiag D, Gray S, Donnelly R: **Exendin-4 increases insulin sensitivity via a PI-3-kinase-dependent mechanism: contrasting effects of GLP-1**. *Biochem Pharmacol* 2002, **63**(5):993-996.

95. Edwards CM, Stanley SA, Davis R, Brynes AE, Frost GS, Seal LJ, Ghatei MA, Bloom SR: **Exendin-4 reduces fasting and postprandial glucose and decreases energy intake in healthy volunteers**. *Am J Physiol Endocrinol Metab* 2001, **281**(1):E155-161.

96. Dalton L: **Drugs For Diabetes**. In: *Chemical & Engineering News.* vol. 82; 2004: 59-67.

97. Amylin: **Byetta (Product Information)**. In.; 2005: 1-16.

98. Runge S, Wulff BS, Madsen K, Brauner-Osborne H, Knudsen LB: **Different domains of the glucagon and glucagon-like peptide-1 receptors provide the critical determinants of ligand selectivity**. *Br J Pharmacol* 2003, **138**(5):787-794.

99. Deacon CF, Ahren B, Holst JJ: **Inhibitors of dipeptidyl peptidase IV: a novel approach for the prevention and treatment of Type 2 diabetes?** *Expert Opin Investig Drugs* 2004, **13**(9):1091-1102.

100. Nielsen LL, Young AA, Parkes DG: **Pharmacology of exenatide (synthetic exendin-4): a potential therapeutic for improved glycemic control of type 2 diabetes**. *Regul Pept* 2004, **117**(2):77-88.

101. Goke R, Fehmann HC, Linn T, Schmidt H, Krause M, Eng J, Goke B: **Exendin-4 is a high potency agonist and truncated exendin-(9-39)-amide an antagonist at the glucagon-like peptide 1-(7-36)-amide receptor of insulin-secreting beta-cells**. *J Biol Chem* 1993, **268**(26):19650-19655.

102. Lopez de Maturana R, Donnelly D: **The glucagon-like peptide-1 receptor binding site for the N-terminus of GLP-1 requires polarity at Asp198 rather than negative charge**. *FEBS Lett* 2002, **530**(1-3):244-248.

103. Al-Sabah S, Donnelly D: **A model for receptor-peptide binding at the glucagon-like peptide-1 (GLP-1) receptor through the analysis of truncated ligands and receptors**. *Br J Pharmacol* 2003, **140**(2):339-346.

104. Al-Sabah S, Donnelly D: **The primary ligand-binding interaction at the GLP-1 receptor is via the putative helix of the peptide agonists**. *Protein Pept Lett* 2004, **11**(1):9-14.

105. Okur A, Simmerling C: **Personal Communication**. In.; 2005: Hybrid Explicit/Implicit Replica Exchange.