THE UNIVERSITY OF CHICAGO

DYNAMICS AND THERMODYNAMICS OF PROTEINS: INSIGHTS INTO THE PROTEIN FOLDING PROBLEM

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

DEPARTMENT OF CHEMISTRY

BY

MUHAMMAD HAMID ZAMAN

CHICAGO, ILLINOIS

JUNE 2003

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

To my family

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

It's a divine blessing that I was productive and efficient in my graduate career, however it wouldn't have been possible without the contributions of my advisors, colleagues, friends and family members. I am grateful to all of them for their patience, support and guidance.

Steve, Tobin and Karl made possible what seemed rather impossible four years ago. I still remember my uncertainties, my concerns and my fears at the beginning of my graduate career. My advisers gave me confidence, provided guidance, yet gave me enough independence to pursue all of my crazy ideas. Steve made me think like a scientist and made me bold, independent and innovative. Tobin was always there to challenge my ideas and guide me in the right direction, and Karl initiated my interest in a wide variety of interesting problems. It is due to these people that today I can take pride in my work. I can't thank them enough!

The atmosphere in the three research groups was very different, yet they all complemented each other. Tobin's group was skeptical of simulations and theory, so I always had to prove the accuracy of my methods to a group of skeptical experimentalists. Steve's group was diverse so I had to learn my problems from different angles to satisfy their curiosity. Karl and his group was rigorous in their methods, so I had to make sure I knew enough before I could talk to them about the obstacles in my research. All in all, these three groups made me think and rethink about my research problems and gave me ideas that contributed to my success. I am especially grateful to my colleagues, namely

Nima Panahi in Steve's group, for he was always there as a friend and as a colleague, to Dorel Buta in Karl's group, who taught me some of the basic aspects of simulations, to Bryan, Adarsh, Kevin, Shane and Ali, in the Sosnick group for making my time in the lab so much fun, to Abhishek for his help with Stat-Mech, to Tahlee, for his support in the rough times, and to Min-Yi, for teaching me everything that I know now about MD simulations.

I am also grateful to Professors Norbert Scherer and Steve Kron, for selecting me as a fellow in the Burroughs-Wellcome program and for constantly providing a rigorous atmosphere where every idea was challenged until proven to be true. I am indebted to Steve Kron, for his recommendation played a key role in helping me secure a post-doc position at Whitehead. I also appreciate the comments and suggestions of the Burroughs-Wellcome fellows regarding my research. In this regard, Wendy, Richard and Alexei deserve my special thanks. I also thank Prof. Peter Rossky at the University of Texas for his hospitality and guidance during my stay in Austin in summers of 2001 and 2002.

I am also grateful to the staff in the Chemistry department, especially ZG, for he was extremely patient with me, and taught me a lot about using computer clusters even after I had been guilty of not following the guidelines of the cluster. I thank Mary Giacomoni, Mary Kulberg, Melinda Moore and Joanne Vetroczky, for they are the best secretarial staff any department can have.

Last, but not the least, I am eternally grateful to my family. My mother Bushra, my brother Qasim, my sisters Rabia and Fakiha, and my wife Afreen and her parents Mr and Mrs. Siddiqi, who were always supportive and encouraging of my endeavors. My success would not have been possible without the constant support and prayers of my

family. I would have quit graduate school long time ago if it was not for the encouragement of my wife and the support of my mother, thus it is to my family that I dedicate my thesis.

ABSTRACT

The dynamic and thermodynamic properties of proteins are studied using mathematical models and computer simulations. These models are then tested by comparison with experimental results. First, a method based upon traditional statistical mechanics is developed to study the entropic benefit of cross-linking. This method has applications in protein folding, binding and protein-engineering. Second, a method based upon the transition state theory is used to study macromolecular reactions through multiple pathways. Our method takes into account the intrinsic densities of states of initial and final stages of the reaction, and is able to explain the curvature in the Arrhenius plots in protein folding and unfolding experiments. Third, the dynamics of small peptides are studied as a function of force fields used in computer simulations. Our results show that the force fields show strong biases towards certain conformations and show an ~8-fold dispersion in dynamics. Finally, the effect of nearest neighboring residue's conformation, identity and sequence on a given amino acid is studied using implicit solvent, Langevin Dynamics (LD) simulations. The results show that the Flory Isolated Pair Hypothesis is invalid for small proteins, and that the nearest neighbor interactions play a major role in the overall dynamics and thermodynamics of the system.

1. INTRODUCTION

The Protein-Folding Problem:

Proteins are nature's machinery that perform specific functions. From enzymes that regulate the biological reactions to antibodies that help the immune system to rid the organism of unwanted invaders, proteins play a vital role in biological organisms. In 1961 John Kendrew showed that proteins have a unique three-dimensional structure, and then in early 1970s Chris Anfinsen showed experimentally that the amino acid sequence of a protein contains all the information to form a three-dimensional structure (Anfinsen, 1973). Hence, all the information regarding the final structure of the protein is encoded in its sequence, which in turn determines whether the protein can perform its specific function in the cell. To understand how a protein forms a unique three-dimensional structure from the sequence of its amino acids is one of the most challenging problems in modern biophysics.

This problem has special significance in the post-genomic era, for the genome contains the DNA code that specifies the protein's sequence. In addition, many human proteins are membrane bound and extremely hard to crystallize, therefore the ability to predict the protein structure from its sequence is extremely important. Furthermore, the loss of this three dimensional structure can lead to a protein's inability to perform a specific task, thereby causing diseases such as Alzheimer's, cystic fibrosis etc (Dobson, 2001).

Understanding how a protein folds can also help engineer nanoscale materials that will perform tasks similar to those performed by proteins in their native environment. Finally, the ability of a protein to interact with another protein greatly depends upon its structure, and understanding processes such as binding will require a better understanding of the folding process.

Challenges in studying protein-folding and related problems

Even though the peptide bond was discovered over a hundred years ago, the principles of how a chain of amino acids forms a unique three dimensional structure remains a mystery. Cyrus Levinthal reasoned that a protein molecule can not search through all the possible conformations (Levinthal, 1968), for if each conformation is searched at a rate of conformation per fsec it would require 10^{18} years for a 100 amino acid protein to fold, whereas proteins usually fold in microseconds to seconds. The problem is therefore to find the lowest energy conformation(s) in a very short time.

This time scale of folding demands use of experimental methods with a short time resolution to study the folding problem. Traditional methods such as x-ray crystallography are not useful as they only measure the static structures. In addition, though the proteins typically fold at milliseconds and microseconds, dynamics occurring at nano- and pico-second time scales also have a lot of useful information about the interactions between amino acids. These short timescales therefore require experimental methods which have a short time resolution to probe dynamics at nano and pico-second level.

In addition to the experimental challenges, the protein folding problem is also difficult to study through traditional theoretical methods. This difficulty is due to the complexity arising from a large number of atoms of different types in the system, the complicated many dimensional potential energy surface, and the lack of any obvious symmetry and reaction coordinate to study the protein-folding reaction. In addition, modeling the folding process in aqueous media is challenging due to the complexity of water structure and solute-solvent dynamics.

Current Methods :

This section reviews current methods most often used to study the folding process. The section is divided into two main parts, namely experimental and theoretical methods, and discusses the some of the state-of-the-art techniques to study protein folding.

Experimental Methods

Stopped-Flow spectroscopy

Stopped-Flow spectroscopy is one of the most effective and commonly used technique to study the kinetics of the protein-folding event. Small volumes of solution are driven from syringes through a mixer just before passing through an observation cell, where the measurement is taken. As the solution flows, the reaction is only a few milliseconds old at the time of measurement. The solution then passes into a collection syringe or a stopping syringe, at which point the solution stops flowing. The most common methods of measurement to follow the kinetics of the folding event through measurement of fluorescence, absorbance and circular dichroism.

The technique is attractive as it uses a small volume and can be used to study the folding of small single domain proteins, which often fold in a two-state fashion and have the folding time scales of milli-seconds or more. However, for proteins folding at a much faster-time scale (folding rate ~ 1 microsecond or less), the technique can not be as used at the moment.

Hydrogen Exchange Methods

Hydrogen exchange (HX) is one of the few non-perturbing method to study the folding of a protein. Amide hydrogen atoms in the protein exchange at different rates with the solvent, depending upon the tendency of that atom to ionize, and therefore measurement of the rate of exchange gives information about structural and functional properties of a protein (such as folding, energy transduction etc.

HX experiments can measure the stability and the global unfolding of a protein at native conditions (Englander et al., 1997). As a result, the technique does not suffer from the uncertainties and difficulties often encountered at extreme denaturing conditions. Thus, HX can be used to study the entire folding pathway without any perturbation due to the presence of denaturant. In addition, HX pulse labeling experiments can give valuable information about the kinetic intermediates in a folding reaction. By changing the solvent and the pH of the solution at different mixing stages of the experiment, H-D exchange pattern in the protein can be observed. A variety of such experiments, with pH and solvent changes at different times, can then give information about the time course of structure formation, and hence show the existence or absence of kinetic intermediates (Englander & Mayne, 1992). Finally, the HX experiments have also been used to study

the structure of proteins in their equilibrium molten globule form, and therefore yield information about why certain proteins form molten globules while others don't. Such experiments also outline the presence of secondary structure elements in the molten globule intermediates (Raschke & Marqusee, 1997)

Temperature jump experiments

The temperature at which a protein unfolds is known as melting temperature. Raising the temperature of the solvent above the melting temperature therefore induces unfolding. In a temperature-jump experiment, the solution is kept a few degrees under the melting temperature. With a short laser flash or electrical discharge, the temperature of the solution can be raised to a value above the melting temperature, therefore initiating unfolding of the protein. This technique is used to study the unfolding process at much shorter time scales than the ones available through stopped-flow spectroscopy. The unfolding process induced by the laser flash, can give useful information about the unfolding event as well as some information about the folding of the protein. The detection process is usually carried out by following the fluorescence of Tryptophan excited by UV lasers.

In order to study the folding process, a cold-denatured protein (unfolded protein at low temperature) is subjected to a laser pulse, and the temperature jump results in the folding event. One major drawback in using this technique comes from a short temperature range of the experiment, and hence the reaction is observed under unstable or weakly stable conditions. As a result, the folding or the unfolding reaction can only be observed in the transition region.

Theoretical Methods and Simulations

Theoretical Modeling

Though similar to polymers to a certain extent, proteins pose many problems for traditional mathematical and statistical methods of enumerating conformations, and detailed molecular modeling. Part of this problem arises from the large number of atoms involved, therefore enumeration of all possible conformations is extremely difficult, if not impossible. In order to solve this problem, lattice-models (similar to the ones often used in polymer physics) have been introduced and will be discussed in a later section of this chapter. In addition, the presence of solvent around the protein increases the complexity of the problem.

In spite of these problems, there have been several efforts to use mathematical models to study protein-folding, protein-protein interaction and protein-binding. Many of these studies have focused on calculating the thermodynamic and kinetic properties using Statistical Mechanics (Pande et al., 1997; Zaman et al., 2002) , path integrals (Wang et al., 1996)and Transition State Theory (Zaman et al., 2003b). The presence of solvent has been dealt with implicitly in most of these techniques. Unfortunately, due to the complex nature of the problem, analytic methods are either applicable to very few systems or use approximations that make the system unrealistic. To overcome these obstacles, numerical methods are often used to deal with the complexity and non-linearity of the problem. One such method is the use of computer simulations, which is discussed in the following section.

Molecular Dynamics Simulations and Force Fields

Molecular Dynamics or MD simulations have become a standard computational technique in studying protein-folding. The MD algorithm is based on calculating the "trajectory" (position and velocity) of every single atom in the system (solute and solvent). The calculation of the trajectory is achieved by calculating the force on each atom, which is equal to the negative derivative of the energy with respect to position. The total energy is given by equation 1.1

$$
\begin{aligned}
U &= U_{kinetic} + U_{potential} \\
U_{kinetic} &= \frac{1}{2}\sum_{i}^{N} m_i v_i(t)^2 = \frac{1}{2} n k_B T \\
U_{potential} &= U_{bond-bend} + U_{bond-stretch} + U_{vdw} + U_{torsion} + U_{solv}
\end{aligned}
\tag{1.1}
$$

where k_B is the Boltzmann constant, and n is the number of coordinates. The potential energy can be written as the sum of individual contributions from bond-bending, bond-stretching, van der Waals interactions and interactions between solvent and solute. A central element of MD simulations is a list of parameters known as a "potential" or a Force-Field (FF). This list has all the information about atomic charges, bond bending and stretching, van der Waals radii, penalties for disallowed motions etc. The total energy must stay constant with time, whereas the force between atoms changes over time, since all atoms are moving together. Analytic solution of the position x(t) and velocity v(t) is impossible, so numerical solutions of the position and velocity are calculated.

Though MD simulations are extremely useful in providing pico and nano-second picture of the dynamics of the molecules, there are some fundamental problems with MD simulations. For protein folding simulations, the first and the foremost problem is

computational cost of MD simulations. Since the trajectory for every single atom, including those of the solvent is calculated, a single trajectory of 1 microsecond of a small (~50 amino acids) can take months on even the fastest computers (Duan & Kollman, 1998). This major problem limits the study to only small proteins. Also most proteins fold on the order of milliseconds (three orders of magnitude longer than the longest MD trajectory on a protein), and finally because one trajectory of folding simulation is often insufficient to reach any conclusion.

In order to overcome these problems, several methods have been introduced which include implicit treatment of solvent (later section in this chapter), use of united atom force fields (chapter 4) and distributed computing over the internet (such as Folding@Home; (Zagrovic et al., 2002)) to improve the statistics. One method to overcome the computational costs is to study the unfolding events at higher (> 500 K) temperatures. However, in order to construct a picture of folding events from an unfolding trajectory one has to rely on microscopic reversibility. In addition, these high temperature and pressure conditions are far from being realistic, and can only study the melting of large domains and structures, Finally, MD simulation results are also unreliable due to problems with force fields and their optimization for structure and not dynamics (Chapter 5).

Langevin Dynamics (LD) Simulations and Treatment of Solvent

One of the major obstacles in achieving long time dynamics (> 1 μs) with an all-atom representation of the protein-molecule is the inclusion of the solvent atoms. The

Duan-Kollman trajectory of 1 μs of villin headpiece utilized 256 Cray T3E processors with 255 of them tracking the solvent and one tracking the protein molecule.

One method that has been very successful in reducing the computational costs of all-atom simulations is the use of implicit solvent and Langevin Dynamics to calculate the forces on the protein-molecule. The representation of the implicit solvent has three major requirements (in addition to the usual algorithm for LD/ MD simulations), a microscopic "solvation" potential, a distance dependent dielectric term to screen charge-charge interactions and the calculation of friction-coefficients required for the LD algorithm. The method is discussed in detail in Chapters 4 and 5.

The implicit solvent-LD method has shown to decrease the computational costs of LD simulations by a factor of almost 200. In addition, it has been shown by Shen and Freed (Shen & Freed, 2001) that the difference between explicit solvent and implicit solvent methods is almost an order of magnitude smaller than the difference between the individual force-fields employed to study the dynamics of small proteins.

The implicit solvent-LD simulations are still under development and need improvement on several fronts. One aspect that needs improvement is the use of a better dielectric screening constant. Also, the solvent-solute hydrogen bonds cannot be accurately studied with this method.

MC Simulations and Go Models

Monte Carlo simulations have been used to reduce the computational costs associated with dynamics (MD and LD) simulations. This is achieved by choosing the

configurations in phase space at random by a weighting algorithm rather than integrating the equations of motion. This is usually achieved by using Metropolis sampling, where the configurations are chosen with a probability of exp(-E /k_BT) and the algorithm weighs the configurations evenly. The MC algorithm starts at a configuration of the system A_1, and then the system is perturbed randomly to give a new configuration A_2. In the NVT (constant number of particles, volume and temperature) ensemble the probability of accepting this new configuration is given by:

$$p = \min[1, \frac{\exp(-E_2 / k_B T)}{\exp(-E_1 / k_B T)}]$$

Thus, if the energy of the new configuration A_2 is less than A_1, the new configuration is accepted and recorded, and the process of a random step is repeated with A_2 being the initial configuration. If the energy of A_2 is greater than that of A_1 then, the probability p is compared to a random number z between 0 and 1 and the move is accepted if $p \geq z$, and is rejected otherwise. If a move is rejected, the original configuration, the move is repeated with A_1 being the starting state until the system reaches a new configuration according to the criterion above. The art of running a successful MC calculation lies in choosing a perturbation step of appropriate size, which is not large enough so the rejection rate is high, or not too small such that the volume of the sample phase space increases very slowly with time and makes the process computationally expensive.

MC calculations have been applied to the protein-folding problem in a variety of different flavors. These include MC using torsional dynamics (Fernandez et al., 2001), MC using an all-atom description of the system and MC using a Go model (Shimada &

Shakhnovich, 2002). A Go atom-atom potential (Go, 1983) makes two atoms attract each other if they are neighbors in the native state and repel each otherwise. The energy function strongly favors the native state, making it a global minimum. Though this model is useful in reducing the computational costs, it lacks the ability to accurately search the phase-space and is unable to predict the structure from a sequence of amino-acids for which the native structure is unknown.

One major draw back of using MC simulations is the limited information that can be obtained about the time dependent dynamics of the of individual amino acids Nonetheless MC simulations have had some success in predicting the final structure of the protein from a given sequence.

Answering Fundamental Questions Related to Protein-Folding: Combining Theory, Experiment and Computations.

The current thesis discusses the applications of theory, computer simulations and experiments to understand some of the key processes related to protein-folding. Chapter 2 discusses the application of statistical mechanics to the calculation of entropic benefit in cross-linking. The methodology developed has applications in protein-folding, binding, protein-protein interactions and protein-engineering, where the presence of loops can play a major role in stabilizing a complex.

The next chapter outlines a method rooted in the Transition State Theory to study the kinetics of multiple-path processes. It has been shown that the reaction from the unfolded to the folded state can take place through multiple pathways, depending upon the reaction conditions. There is, however, a surprising dearth of theoretical studies of

such pathways in macromolecular reactions. The chapter discusses a elegant method to study the reactions through multiple pathways as a function of reaction conditions. The method is general enough to be applicable to a wide variety of macromolecular reactions.

The dynamics of small peptides, as a function of various force fields are studied in chapters 4 and 5. Chapter 4 discusses the difference between the united atom and explicit atom force fields, with emphasis on dynamics. The current force fields have been optimized for thermodynamic properties, and their ability to capture the dynamics is unknown. Chapters 4 and 5, study the internal consistencies among the force fields and their ability to reproduce experimental results. The chapters also suggest possible improvements in the current force fields.

Chapter 5 discusses nearest neighbor interactions in peptides. The conformational, geometric and sequence dependence of such interactions are discussed. Such interactions play a vital role in the overall dynamics and thermodynamics of the peptide. The study gives a quantitative measure of these interactions, and the shows that the Flory Isolated Pair Hypothesis is not valid for small peptides.

The final chapter of the current thesis suggests possible future projects and research directions that naturally emerge out of the research presented in earlier chapters.

2. ENTROPIC BENEFIT OF CROSS-LINKING

Introduction

"Thus we are forced to the conclusion that there is no basis for estimating the standard free energy change for the binding of a molecule to a macromolecule from the corresponding energies of binding for molecules representing its component parts without a detailed knowledge of the properties for the system."(Jencks, 1975)

Part of this dilemma posed by Jencks in his classic treatise on enzymology, is the difficulty of calculating from association constant of n component system (K_{ass}, units of M^{-n+1}), the related association constant for the lower order reaction where two of the components are tethered(Jencks, 1981) (K_{ass}, units of M^{-n+2}). This issue can be cast in the specific situation of concentration-dependent bimolecular docking reaction $A + B \leftrightarrow A \bullet B$: Given the free energy of this reaction, ΔG^{o}_{bimol}, can one predict the concentration independent ΔG_{uni} for the corresponding unimolecular reaction where the components are cross-linked with a flexible tether $A \cdots B \leftrightarrow A \bullet B$?

The first important issue to note here is that the loss of entropy upon tethering is completely distinct from the loss of entropy upon binding (Fig. 2.1). Binding of two species of comparable sizes effectively reduces the independent translational and rotational freedom of the two bodies to that of a single body. Upon cross-linking, however, each of the two species retains a considerable amount of translational and rotational freedom, given that the tether is of moderate size and flexibility. The difference between the cross-linking and binding processes is clearly noted by considering the two-step process, $A + B \leftrightarrow A \cdots B \leftrightarrow A \bullet B$, where the tether is introduced, followed by the

two tethered species binding to each other in a complex ($A \bullet B$). Hence, binding and tethering represent fundamentally different processes even though they both reflect a reduction of the dimensionality of the system.

Traditionally, the Sackur-Tetrode equation describing entropy in the gas phase has been used to estimate the entropy of binding, cross-linking, and association of atoms, molecules and macromolecules(Amzel, 1997; Mammen et al., 1998). However, this equation does not consider the molecular volume occupied by the solvent, and hence, probably overestimates the entropy in the liquid phase. In the cell theory of liquids, another approach to the problem, the entire volume of the solvent is divided into fixed cages or cells(Barker, 1963). Each cell contains a single given molecule, an unrealistic assumption that precludes the molecules from fully sampling the entire volume.

To compare the bimolecular to unimolecular system, the center-to-center probability distributions are calculated for the two species before and after their tethering. These distributions, $P_{\mathrm{dimer}}^{helix}(r)$ and $P_{tether}^{helix}(r)$, respectively, are used to calculate the entropy of each state according to

$$S = -R\int_0^\infty 4\pi r^2 P(r)\ln P(r)dr \qquad (2.1)$$

The distribution $P_{\mathrm{dimer}}^{helix}(r)$ inherently depends upon the concentration of reactants, typically chosen to be at 1 M standard state. The more concentrated the reactants, the less entropy is lost upon introduction of the tether. The distribution $P_{tether}^{helix}(r)$ depends upon the length and nature of the tether.

One of the major benefits of using probability distribution functions is that the presence of the solvent does not significantly affect the calculation of the benefit of cross-

linking. Although the volume occupied by the solvent molecules may restrict the available volume of each helix, the distribution functions themselves are not significantly altered by the liquid. Additionally, any reduction should be independent of the presence of the tether. Hence, the entropy of both states should be reduced by the same amount, which cancels out in the calculation of the overall change in entropy.

To calculate the untethered system's distribution, we introduce the "Nearest Neighbor" Method, where $P_{dimer}^{helix}(r)$ is posited to reflect the probability that a partner can travel a given distance from a reference helix while still being its nearest neighbor (NN). The center-to-center and relative angle distribution functions of the tethered system, connected by Gaussian random coils or poly-L-alanine chains, are compared to their untethered counter-parts to estimate the loss of translation and rotational entropy accompanying cross-linking. Here, we illustrate this methodology to the docking of two helices, as well as to the general association of three components where two of them are pre-tethered. The method is extended and compared to experimental results for a variety of proteins.

Methods:

The entropy calculations were performed using a program written in Mathematica® 4.1 developed by Wolfram Research Inc. A chain of n residues (unfolded peptide or the cross-link) was modeled either as a Gaussian random walk with $2n$ segments (length 1.5 Å= ½ (C_α-to-C_α distance)), or as a poly-alanine chain. For the latter, each alanine's conformation was specified by the occupation of three discrete regions in the Ramachandran Φ,Ψ plot (extended, α- and 3_{10} helical regions which are

the upper left, lower left and upper right quadrants, respectively). The regions were approximated as ellipsoids according to Flory's method(Flory, 1953). The Φ,Ψ values were randomly chosen within each region in proportion to the size of each ellipsoid to determine the overall conformation of the chain.

The excluded volume effects were investigated as outlined by Pappu et al(Pappu et al., 2000). Based upon a hard-sphere model, the sterically allowed Φ,Ψ angles for an alanine dipeptide were generated. A steric clash between two atoms exists when their contact distance is less than the hard-sphere contact distance. Equally distributed values of these sterically allowed Φ,Ψ angles were used to generate conformations of chains up to twelve alanine residues. Each chain's conformation was then screened for steric clashes between any two atoms of non-adjacent residues. The conformations without any clashes were tabulated to calculate the percentage of allowed conformations for each chain length.

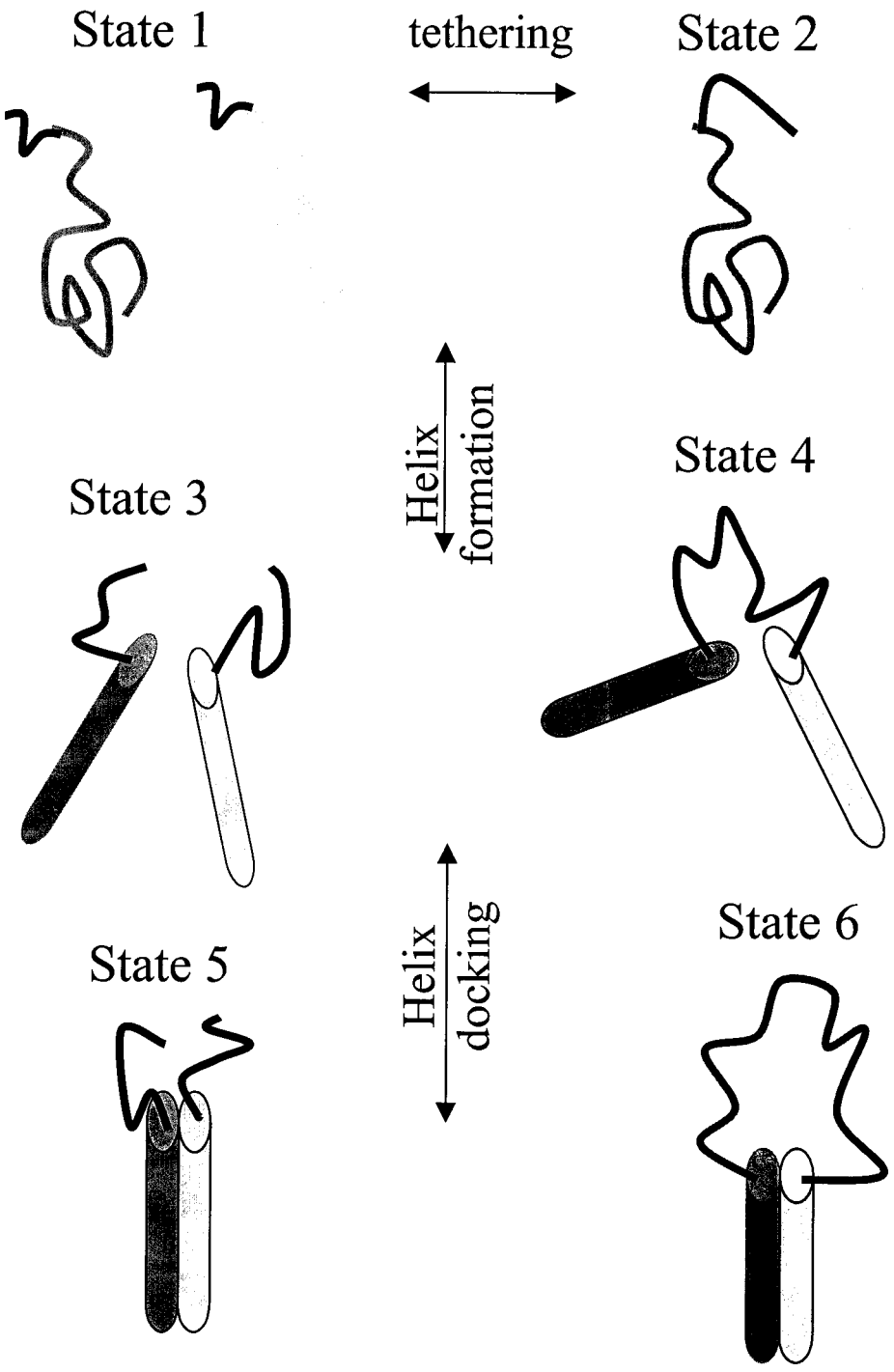
Results and Discussion

Loss of translational and rotational entropy upon cross-linking

The folding and binding of a pair of helices is modeled with six states, which are related by transitions representing either folding, cross-linking, or binding (Fig. 2.1). The horizontal arrows represent the cross-linking process whether the system is in the denatured (1↔2), helical (3↔4), or docked conformations (5↔6). The vertical arms represent the folding from a random coil to helical structure (1↔3 and 2↔4), or the docking of two pre-folded helices (3↔5 and 4↔6). The translational entropy depends

Figure 2.1 Individual steps in binding, folding and cross-linking

The left column represents the process of two untethered, denatured chains that form helices, and then bind to form a helical dimer. The two unstructured monomers are modeled as Gaussian random walks (State 1); the isolated (State 3) and bound (State 5) helices are modeled as thin rods. States 2,4, and 6, are the corresponding states for a system where the two chains have been tethered. The introduction of the tether, a shift from the left to the right column, results in a decrease in the reaction order for the isolated chains ($1\leftrightarrow2$, and $3\leftrightarrow4$), but not for pre-docked helices ($5\leftrightarrow6$). The loss of translational entropy upon introduction of the tether for the upper two transitions is calculated according to the NN method, while for the lower transition, the loss is calculated according to loop closure entropy.


State 1
tethering
State 2
Helix formation
State 3
State 4
Helix docking
State 5
State 6

upon the distribution of center-to-center distances whereas the rotational entropy depends upon the distribution of the relative angles between the two components. The reduction in translational entropy upon cross-linking of pre-folded helices is calculated by comparing the center-to-center probability distribution of the two helices before, $P_{dimer}^{helix}(r)$, and after tethering, $P_{tether}^{helix}(r)$, according to:

$$\Delta S_{trans} = S_{tether}\text{-}S_{dimeric} = -R\int_0^{\infty}\left(4\pi r^2\{(P_{tether}^{helix}(r)\ln P_{tether}^{helix}(r) - P_{dimer}^{helix}(r)\ln P_{dimer}^{helix}(r)\}\right)dr \qquad (2.2)$$

where R is the gas constant, and the distribution has normalization $\int_0^{\infty}4\pi r^2 P_{tether}^{helix}(r)dr = 1$.

Although internal vibrational motions of the docked complex must be accounted for in a calculation of the energetics of a given binding process (ΔG^{o}_{bimol} or ΔG_{uni}) (Brady & Sharp, 1997), these motions are not altered by introduction of an ideal tether. They should contribute equally in the bound state of both the tethered and untethered systems, and do not affect ΔS_{trans} as defined in Eq. 2.2. Thus, internal motions do not need to be considered in the present calculation of the change in entropy upon tethering.

The reduction in rotational entropy is obtained from a comparison of the rotational distribution function for the tethered species, $P_{tether}(\theta,\phi)$, with that of the untethered species, which is a uniform distribution $P_{untether}(\theta,\phi)$:

$$\Delta S_{rot} = -R\int[P_{tether}(\theta,\phi)\ln P_{tether}(\theta,\phi) - P_{uniform}(\theta,\phi)\ln P_{uniform}(\theta,\phi)]d(\theta,\phi) \qquad (2.3)$$

where θ, and ϕ are the angles in spherical coordinates between the axes of the two helices with normalization $\int P(\theta,\phi)d(\theta,\phi)=1$. The unfolded peptide and unstructured tethers are initially approximated as Gaussian random walks (or chains)(Jacobson et al., 1950;

Jacobson & Stockmayer, 1950) and the helices as thin, non-interacting rods. Both the tether and the unfolded polypeptide are assumed to Gaussian random walks. The excluded volume effects of the chain are discussed in a later section.

Translational Entropy

The loss of translational entropy is calculated upon the introduction of the tether for each of the three horizontal transitions shown in Fig. 2.1.

Docked helices:

The process 5↔6 represents the ligation of the tether and the formation of a closed loop. The ligation process is unimolecular (ignoring the covalent peptide bond formation). Hence, the concentration of reactants is irrelevant to the loss of translational and rotational entropy. According to Flory as well as Jacobson-Stockmayer Theory for Gaussian chains, the entropic cost of loop closure of n segments is(Flory, 1953; Jacobson et al., 1950; Jacobson & Stockmayer, 1950)

$$\Delta S_{5\text{-}6} = -3/2\, R \ln (n\, \pi/2) \tag{2.4}$$

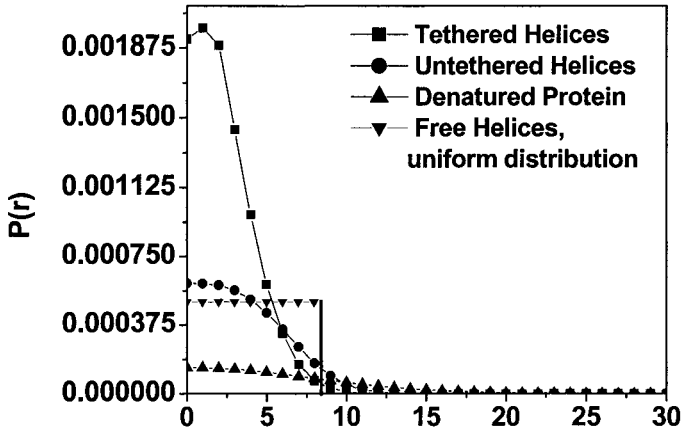
Undocked helices:

The process 3↔4 represents the cross-linking of a system of two pre-folded, but undocked helices. Tethering of the helices results in a change from a bimolecular to a unimolecular system. The loss of entropy for this process is calculated from the change in the probability distribution of center-to-center distances that describes the configuration before and after the introduction of the tether, i.e. $P_{\mathrm{dimer}}^{helix}(r)$ and $P_{tether}^{helix}(r)$ respectively.

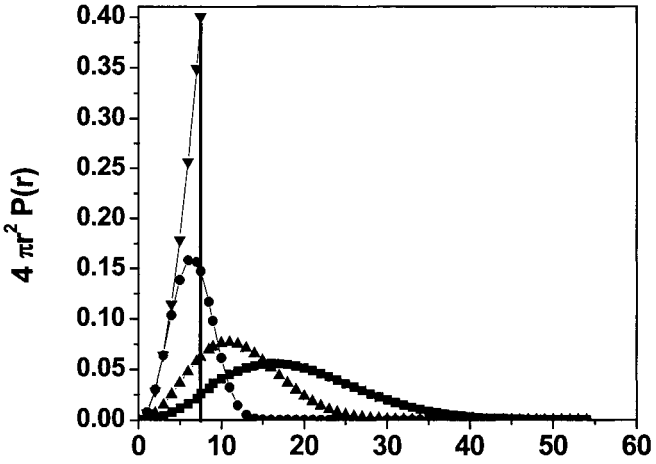
Figure 2.2 Probability Distribution Functions

The center-to-center probability distribution functions for unfolded chains and cross-linked helices, either untethered, (calculated according to the NN Method), or connected by a six residue tether (modeled as a Gaussian random coil). *a)* $P(r)$ distribution. *b)* The unit normalized $4\pi^2 P(r)$ distribution, and *c)* $4\pi r^2 P(r) \ln P(r)$. The area under the curve in *c)* for each distribution (multiplied by -*R*) is the translational entropy. For comparison, the distribution is shown for a molecule uniformly distributed within a spherical volume, but having the same entropy as the NN distribution. Symbols are the same in all three panels.

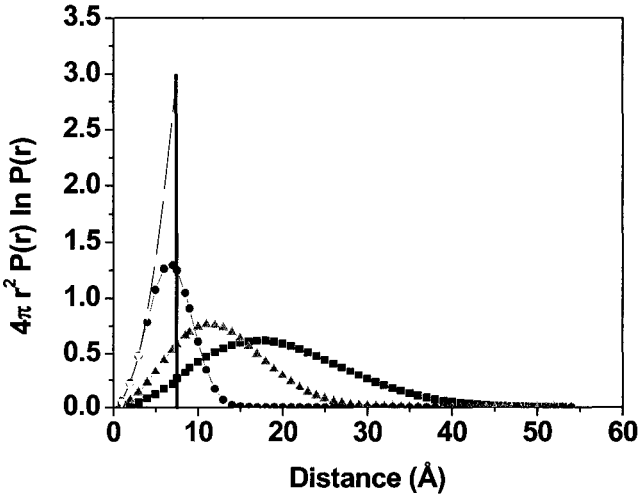
a)


Tethered Helices
Untethered Helices
Denatured Protein
Free Helices, uniform distribution
P(r)
0.001875
0.001500
0.001125
0.000750
0.000375
0.000000
0
5
10
15
20
25
30

b)


4 πr² P(r)
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
10
20
30
40
50
60

c)


4π r² P(r) ln P(r)
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
10
20
30
40
50
60
Distance (Å)

The relative motion of the two cross-linked helices is restricted by the tether which itself has an end-to-end probability distribution (Fig. 2.2). For an n residue tether, this distribution is approximated by a freely-jointed Gaussian random walk of $2n$ segments (each amino acid has two torsion angles):

$$P_{RC}(r) = (\beta/\sqrt{\pi})^3 e^{-\beta^2 r^2} \tag{2.5}$$

where $\beta = \sqrt{3/(2nl^2)}$, and l is the length of each step(Cantor & Schimmel, 1980; Flory, 1953), taken to be 1.5 Å.

Rather than the end-to-end distribution of the tether, the relevant quantity for the calculation of the loss of entropy for the helices is their center-to-center distribution. This distribution can be calculated from the tether's end-to-end distribution by adding to the ends, an extra step of half the length of the helix chosen at random exit angles. We obtain $P_{tether}^{helix}(r)$ by simulating such a random walk (Fig. 2.2), and the translation entropy of the tethered helices from Eq. 2.1.

Next the center-to-center distribution function for the unlinked system, $P_{\mathrm{dim}er}^{helix}(r)$ is calculated. In principle, the two unlinked helices are uniformly distributed over the entire volume of the system, V_T. The application of this uniform distribution, however, greatly over-estimates the accessible volume and results in an incorrect, system-size dependence to the entropy.

To circumvent this problem, we introduce the Nearest Neighbor (NN) Method. The relevant volume for a given partner helix is limited to the region around a reference helix where the partner is the nearest neighbor. As the partner diffuses a distance away from the reference helix, another helix is likely to become the closest helix. The distance

the original partner can travel while remaining the nearest neighbor is dependent upon the number density, $\rho=N/V_T\equiv 1/V_{\text{mol}}$.

We posit that the probability of the partner being the closest, $P_{NN}(r)$, is proportional to the center-to-center distance distribution, $P_{\text{dimer}}^{helix}(r)$, between the partner and the reference helix, differing only by a scale factor required to maintain unit normalization. As the initial choice of both the reference helix and partner are arbitrary, and their mutual association is not fixed in time (i.e. a third helix can become the NN), issues related to distinguishability are circumvented.

The NN, and hence, $P_{\text{dimer}}^{helix}(r)$ distribution can be calculated from the probability that no other helix is the NN. For the partner at a distance r from the reference helix, the probability that another particular helix is the NN is $v/V_T = v\rho/N$, where $v = 4\pi r^3/3$. The probability the original partner remains the NN is the probability that none of the other N helices are closer:

$$P_{NN}(r) = (1 - v\rho / N)^N \sim e^{-v\rho} \tag{2.6}$$

After normalization, we obtain $P_{\text{dimer}}^{helix}(r) = e^{-v\rho}/\rho$.

Having determined $P_{\text{dimer}}^{helix}(r)$ in this manner, we can calculate the entropy of the untethered, dimeric helices according to Eq. 1. At 1 M standard state, $1/\rho=V_{\text{mol}}=1661$ Å^3, and S_{dimeric}= -17.02 e.u. Interestingly, this value of entropy is equivalent to that for a helix uniformly distributed within a volume of 1839 Å^3. This volume is very close to V_{mol}, for which S =-16.82 e.u. (Fig. 2.2).

The difference in entropy between the untethered system and the cross-linked helices (process 3↔4) provides the entropic cost of cross-linking in the helical, but undocked state, as per Eq. 2.2. For the introduction of a six residue tether to a bimolecular system at 1 M standard state, the entropy is decreased by 2.15 e.u., or $T\Delta S_{trans}$ = -0.64 kcal mol^{-1} at T=300 K. Values for other length tethers are listed in Table 2.1.

Although the tethering the two undocked helices at 1 M standard state by six residues decreases the entropy of the system, a tether of twelve or more residues results in an increase in the entropy of the system. It may seem paradoxical that the tethering can result in an increase in entropy. The increase indicates the tethered, but undocked helices sample more volume than do the isolated helices at the (high) standard state concentration of 1 M. The entropy of an untethered helix equates to it uniformly sampling a box of dimensions of only (12.2 Å)^3, whereas the mean center-to-center distance of two helices with a twelve residue tether is over two-fold larger. This increase for the tethered helices explains why a twelve-residue tether actually increases the entropy in the cross-linked state *when compared to the free state at 1 M standard state concentration* (see discussion below). Part of the effect comes from the fact that the tethered helix cannot be exchanged with another helix, however far away it moves, whereas the untethered one can when it moves further away than 12.2 Å. For a more realistic standard state concentration of 1 μM, the volume sampled by the untethered system is increased 10^6-fold, and the introduction of the tether does reduce the translational entropy of the helices.

Table 2.1. Change in translational entropy upon cross-linking

Length of tether	$\Delta S_{trans}^{denatured}$ (e.u) [1]	$\Delta S_{trans}^{pre-folded}$ (e.u) [2]
6	3.89	-2.15
9	4.11	-0.838
12	4.32	0.104
15	4.51	0.84
18	4.69	1.45

[1] Tethering of denatured helices, Process 1↔2

[2] Tethering of pre-folded helices, Process 3↔4

The use of the distribution function and Eq. 2.1 to calculate the translational entropy correctly accounts for the concentration dependence, $\Delta S(C) = \Delta S(C_{ref}) - R \ln C_{new}/C_{ref}$, where C_{new} and C_{ref} are the new and reference concentration, respectively. For example, an N-fold increase in solute concentration results in the distribution function being uniformly contracted N-fold along the r-axis (with a commensurate increase in height to maintain unit normalization): $P_{C_{new}}(r) = NP_{c_{ref}}(rN)$. The entropy of the system at higher concentration is

$$\begin{aligned} S(C) &= -R\int 4\pi r^2 P_C(r)\ln P_C(r)dr \\ &= -R\int_0^\infty [4\pi r^2 NP_{C_{ref}}(rN)\ln P_{C_{ref}}(rN) + 4\pi r^2 NP_{C_{ref}}(rN)\ln N]dr \\ &= -R\int_0^\infty 4\pi r^2 P_{C_{ref}}(rN)\ln P_{C_{ref}}(rN)(Ndr) - R\ln N\int_0^\infty 4\pi r^2 P_{C_{ref}}(rN)Ndr \\ &= S(C_{ref}) - R\ln N \end{aligned} \tag{2.7}$$

Hence, the method has the correct dependence on the solute concentration.

Denatured polypeptides.

The analysis for tethering in the denatured state (Process 1↔2 of Fig. 2.1), is similar to that for the undocked helices. Again, the change in translational entropy is calculated from the change in the center-to-center distance distributions, $P_{\mathrm{dim}er}^{denatured}(r)$ and $P_{tether}^{denatured}(r)$. The $P_{\mathrm{dim}er}^{denatured}(r)$ distribution is identical to its helical counterpart, $P_{\mathrm{dim}er}^{helix}(r)$, as the NN Method did not assume any shape for the reactants.

The tether restricts the center-to-center distribution of the two denatured polypeptides. The tether's end-to-end distribution when convoluted with each polymer's

end-to-center distribution is the desired center-to-center distance distribution between the two polypeptides. Alternatively, $P_{tether}^{denatured}(r)$ can be calculated by realizing that it is the end-to-end distribution for the portion of the chain connecting the two centers of the denatured (identical) helices. The number of residues in this portion is $N_{helix}+N_{tether}$.

The calculation of change in entropy upon the introduction of a cross-link in the unfolded state is carried out by using Eq. 2.2. We find ΔS_{trans}=3.9 e.u. ($T\Delta S_{trans}$ = 1.16 kcal mol^{-1} at 300 K) for the introduction of a six amino acid long tether in the denatured state, Process 1↔2. The increase in entropy indicates that the center of tethered polypeptide samples more configurational space than it does when it is untethered at a standard state concentration of 1 M.

Size dependence of cross-linking entropy

The entropic cost of cross-linking depends upon the size of components. An increase from 33 to 66 residues, for example, results in the entropy of tethering increasing by 0.6 and 1.82 kcal M^{-1} (2 e.u. and 6.06 e.u respectively) for denatured and pre-folded helices (Process 1↔2, and 3↔4), respectively, when they are connected by a six-residue helix. The increase in entropy for the denatured helices of 66 residue is because the center-to-center distance distribution is that of a 72 residue (=$N_{helix}+N_{tether}$) random walk rather than the original 39 residue random walk. Likewise, the increase in entropy for the pre-folded helices is because of their increased length, which results in a more extended distribution function. The entropy values listed in the Tables 2.1 and 2.2, and mentioned elsewhere in this paper are for tethers of different lengths. The helix, however, is kept at a constant 33 residues and length of 30 Å.

Table 2.2. Effect of tether on the association of denatured and pre-folded helices

Length of tether	ΔS^o (e.u.)	$\Delta G^o_{bi}-\Delta G_{uni}=-T\Delta S^o$	C_{eff} (M) [1]
6	-10.6 (-4.6)	3.2 (1.4)	0.0049 (0.10)
9	-12.0 (-7.1)	3.6 (2.1)	0.0024 (0.029)
12	-13.1 (-8.9)	3.9 (2.7)	0.0014 (0.012)
15	-14.0 (-10.3)	4.2 (3.1)	0.00092 (0.0056)
18	-14.7 (-11.5)	4.4 (3.4)	0.00063 (0.0032)

Values given in the table are for process 1↔5 relative to process 2↔6, at 1 M standard state concentration.

Values for Process 3↔5 relative to Process 4↔6 is given in parenthesis.

Stability is given in kcal M^{-1}, at T=300 K

[1] Calculated according to $C_{eff}= e^{\Delta S/R}$.

Rotational Entropy.

The analysis of the loss of rotational entropy parallels that for translational entropy, except that the restriction in the relative angle, rather than the distance, between the two helices is the pertinent quantity. Of the three horizontal processes in Fig. 2.1, the relative angle is likely to affect only for the tethering of the two, undocked helices (Process 3↔4). The tether does not change the orientation between two docked helices (Process 5↔6), nor does it significantly restrict the rotational freedom between the two denatured polypeptides (Process 1↔2).

The angular distribution of the undocked helices, $P(\theta,\phi)$, relative to a uniform distribution, $P_{uniform}(\theta,\phi)$, is used to calculate loss of rotational entropy (Eq. 2.3). For arbitrarily shaped objects, a third angular degree of freedom is required. However, for the cylindrically symmetric helices with a freely jointed tether examined here, there is no restriction in this quantity.

A tether composed of amino acids has steric restraints due to restriction of each residue's Φ,Ψ dihedral angles. This restriction may result in a decrease in the angular freedom between the two helices. In order to investigate this effect, simulations are carried out with polyalanine tethers. The dihedral angles for each residue are chosen to reflect the restriction of the polypeptide backbone and side-chain moieties. The tether's conformation is coarsely specified for each residue by the occupation of three discrete regions in the Ramachandran Φ,Ψ plot(Flory, 1953). The regions are approximated as ellipsoids according to Flory's method(Flory, 1953). The Φ,Ψ values are randomly

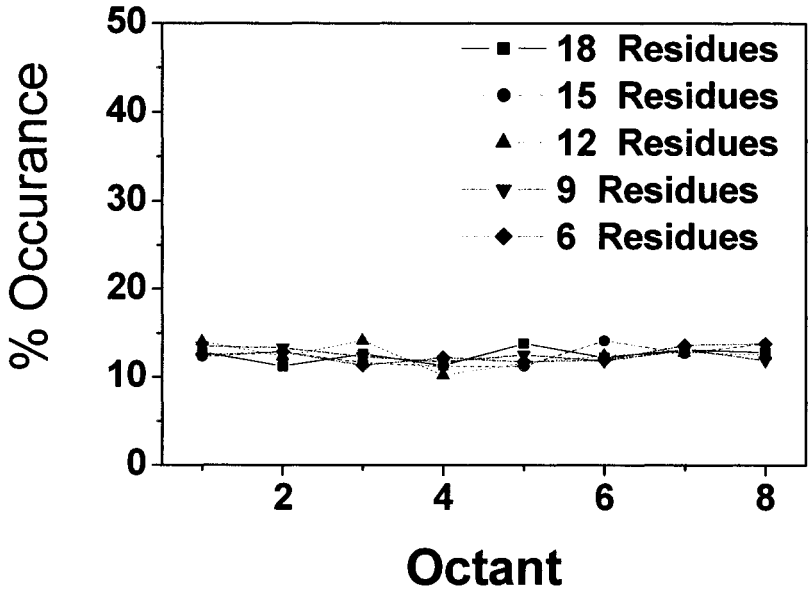


Figure 2.3 Loss of Rotational Entropy.

The angle between the two helices, approximated as the angle between the first segment and the last segment of the polyalanine tether, are shown. The uniform distribution indicates that the loss of rotational entropy upon tethering is minimal at these linker lengths.

chosen within each basin according to their relative areas to determine the overall conformation of the chain.

To assess the loss of rotational entropy, the angle between the long axes of the helices is histogrammed for each configuration. This angle is assumed to be the same as the angle between the first and last segment of the tether (i.e. the helix is fixed in angle relative to the adjoining segment in the tether). The histogram of angles is coarsely divided into octants of a sphere (Fig. 2.3). For a six residue tether, the angular distribution is nearly uniform, indicating that the helices are essentially freely jointed for this length of tether. Hence, the loss of rotational entropy upon tethering is quite small (< 0.5 e.u.), and generally negligible as compared to the loss of translational entropy. Furthermore, the loss of rotational entropy upon tethering is considerably less than 45 e.u. ($T\Delta S$= 13.5 kcal mol^{-1} at T= 300 K), the loss of rotational entropy for the binding of two 4 kD proteins(Mammen et al., 1998).

Entropic benefit of cross-linking.

Having calculated the entropic cost of cross-linking for each of the individual steps, and concluded that the loss of rotational entropy is insignificant, we can estimate the net benefit of cross-linking to the entire docking equilibrium of the unfolded (Process 1↔5 relative to 2↔6) or pre-folded helices (Process 3↔5 relative to 4↔6). Under the assumption that docking is intra-molecular, the introduction of a cross-link results in a loop closure penalty in the docked state. The tether either decreases or increases the entropy of free, undocked state, depending upon its length (Tables 2.1, 2.2). For example,

closing a six residue tether in the docked state (Process 5↔6), results in a loop closure penalty equivalent to a reduction in entropy of 6.7 e.u. At 1 M standard state, the introduction of the tether increases the translational entropy by 3.9 e.u. for the denatured helices (Process 1↔2), but decreases it by 2.15 e.u for the pre-folded, undocked helices (3↔4).

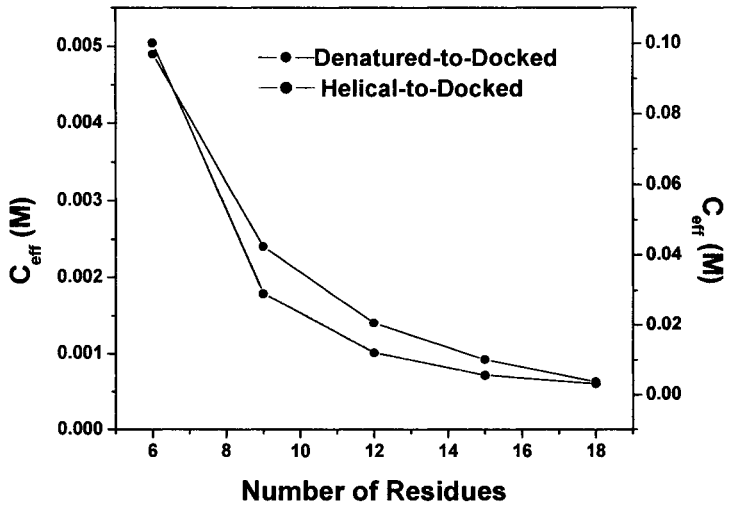
Thus, the introduction of the tether in the denatured state (Process 1↔2) favors the tethered state while in the docked state (Process 5↔6), it favors the untethered state. Hence, for the entire reaction (Process 1↔5 relative to 2↔6), the tether's effect in both the undocked and docked states opposes the formation of the docked state by a total 10.6 e.u. When the undocked state is already helical, the tether's contribution favors the untethered state in both undocked and docked states. The net of these opposing factors still inhibits the formation of the docked state by a total of 4.6 e.u. Therefore, in either situation, the introduction of a tether results in a decrease in the amount of docked species.

However, the conclusion that tethering is entropically destabilizing explicitly depends upon the standard state concentration used in the calculation. The less concentrated the reactants, the more translational entropy each helix has prior to binding. Thus, the loss of entropy upon introduction of the tether in the undocked state is greater when compared to reactants at lower concentrations. The loss of translational entropy generally is calculated relative to the free state at 1 M concentration, the concentration where free energies for bimolecular systems are calculated (i.e. $\Delta G^{o}_{bi} = -R\ T\ \ln K_{ass}$). As noted above, the use of 1 M standard state concentration explains the paradoxical result

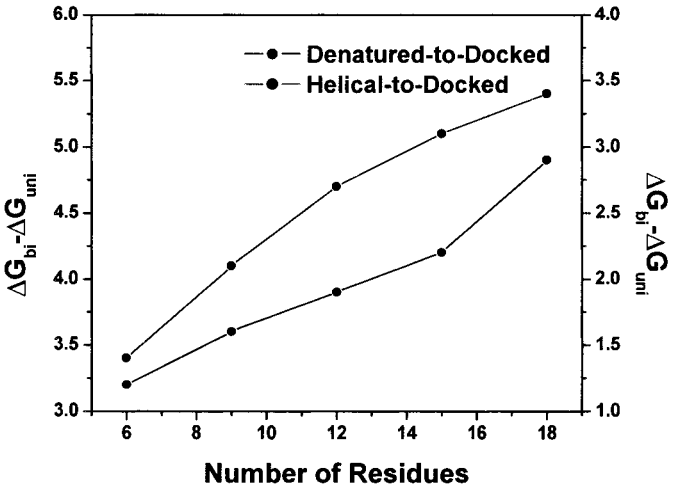
Fig 2.4 Entropic Benefit of Cross-Linking.

a) Decrease in effective concentration as a function of the increase in the number of residues in the cross-link. b) Increase in ΔG_{bi}-ΔG_{uni} as a function of the cross-link length.

a)


Denatured-to-Docked
Helical-to-Docked
C_{eff} (M)
C_{eff} (M)
Number of Residues

b)


Denatured-to-Docked
Helical-to-Docked
ΔG_{bi}-ΔG_{uni}
ΔG_{bi}-ΔG_{uni}
Number of Residues

that a twelve residue tether between two helices, whose average center-to-center distance is 25-30 Å, results in an *increase* in entropy relative to the free species, whose translational entropy is equivalent to the uniform exploration of a cube having a volume of only 1661 $Å^3 \sim (12.2 Å)^3$. In essence, the helices explore more volume when tethered than they do at 1 M standard state concentration according to the NN method.

For macromolecules, the 1 M standard state concentration is unrealistically high. At a more realistic concentration of $C_{ref} = 1$ μM, the entropic penalty of adding the tether is increased by 26.7 e.u. (Eq. 2.7), and the introduction of the tether now is very entropically restrictive, as one expects. Further, at this lower concentration, the assumption that the tethered helices dock intramolecularly is more likely to be valid.

In terms of free energy for the denatured helices at 300 K connected by a six residue tether, ΔG_{uni} is less than ΔG^{o}_{bi} by 3.2 kcal mol^{-1} at 1 M standard state concentration, but is 5 kcal mol^{-1} greater at $C_{ref} = 1$ μM. Hence, the introduction of the tether does increase the population of docked species when concentration of the individual helices is below ~1 mM.

The tether's effect on stability, and its dependence on the choice of C_{ref}, also can be cast in terms of effective concentration of reactants, C_{eff}, upon the introduction of a tether. Once ΔS values are calculated at a given reference concentration, C_{ref}, the effective concentration upon cross-linking, C_{eff}, can be calculated according to:

$$C_{eff} = C_{ref} e^{\Delta S/R} \tag{2.8}$$

For six and eighteen residue tethers, denatured helices have an effective concentration 4.1 and 0.57 mM, respectively (Table 2.2). At a reference concentration equal to C_{eff}, the free

energy of the bimolecular and tethered systems are equal. When the effective concentration induced by the tether exceeds the dissociation constant, the docked complex is stable. $K_{eq}^{complex} = \frac{[docked]}{[undocked]} = \frac{C_{ref}}{K_{diss}} > 1$.

Higher order reactions.

The methodology presented above can be used as a framework for determining the binding constant for a multimeric association after two of the individual components are tethered(Jencks, 1981). The analysis is illustrated with the sequential binding of three individual components, *A*, *B*, and *C* (Fig. 2.5). If the tether does not interfere with either binding step (e.g. $K_{B-C}^{untether} = K_{B-C}^{tether}$), then the difference in the two processes is reduced to the difference in the second binding step where *A* binds to the complex $B \bullet C$. This situation directly corresponds to that for the docking of the pre-formed helix before and after the introduction of the tether (State 3↔State 5 versus State 4↔State 6 in Fig. 2.1). Prior to *A* binding the complex, its effective concentration is altered by the presence of the tether, and consequently its translational entropy is changed. After binding, there is a loop closure penalty associated with the restriction of the ends of the tether. The loss in entropy, $\Delta S_{unbound}$, upon introduction of the tether between *A* and the complex $B \bullet C$ is the same as that calculated using the NN Method for the two helices (although excluded volume issues may become more important). The corresponding loss in entropy in the bound state, ΔS_{bound}, is the loop closure entropy penalty for the tether. This value depends upon the actual end-to-end distance of the tether in the bound state, which may not be

Figure 2.5 Higher order reactions.

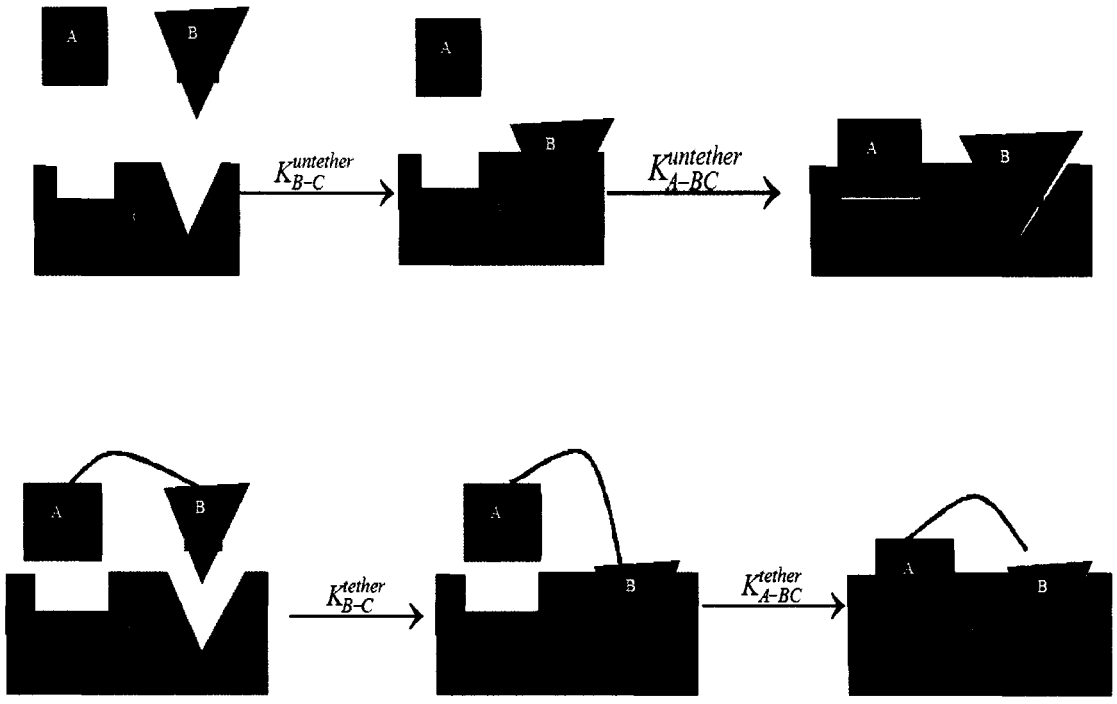
Diagram depicted the sequential binding of three components, A, B and C when A and B are untethered (upper) and pre-tethered (lower). For the untethered system, the association constant is given by

$$K_{A-B-C}^{untether} = K_{B-C}^{untether} K_{A-BC}^{untether} = \frac{[A \bullet B \bullet C]}{[A][B][C]} (K_{B-C}^{untether} ,$$

$K_{A-BC}^{untether}$ in units of M^{-1}, $K_{A-B-C}^{untether}$ in units of M^{-2}), while the association constant for the corresponding reaction where A and B are tethered, is given by:

$$K_{A-B-C}^{tether} = K_{B-C}^{tether} K_{A-BC}^{tether} = \frac{[A \bullet B \bullet C]}{[A \cdots B][C]}$$

(K_{B-C}^{tether} , K_{A-B-C}^{tether} in units of M^{-1}, K_{A-BC}^{tether} is dimensionless).


A
B
A
B
$K_{B-C}^{untether}$
$K_{A-BC}^{untether}$
A
B
A
B
A
B
K_{B-C}^{tether}
K_{A-BC}^{tether}
A
B

zero, and will depend on the details of the system. The loop closure penalty for Gaussian random coil for a given r end-to-end distance (relative to a zero end-to-end distance) is

$$\Delta S_{bound} = S(r) - S(0) = -R\ln[\frac{P_{RC(r)}}{P_{RC(0)}}] = R\beta^2 r^2 \quad (2.9)$$

where β is defined in Eq. 5.

The net effect of the introduction of the tether is $\Delta S_{tether} = \Delta S_{bound} - \Delta S_{unbound}$. This entropy can be converted to C_{eff} according to Eq. 8. We obtain the association constant of the tethered reaction according to

$$K^{tether}_{A-B-C} = C_{eff} K^{untether}_{A-B-C} \quad (2.10)$$

An empirical value of K^{tether}_{A-B-C} which equates to a C_{eff} that is higher than predicted by Eq. 10 generally implies that the tether provides an orientational benefit. Conversely, a lower value of C_{eff} implies that the tether may be directly interfering with binding or that it is insufficiently long or flexible, and is strained in the bound complex.

2.3.4 Gaussian coil approximation.

The simulations of the poly-alanine tether used in the rotational entropy calculation can also be used to test the validity of the Gaussian random coil approximation for real amino acid tethers. The average end-to-end distance distribution for poly-alanine tethers of various lengths is compared to that for an ideal Gaussian random coil in Fig 2.6 *a-c*. A three residue tether behaves significantly differently than

the idealized coil, but a six residue chain already is a reasonable approximation for a Gaussian coil, both in terms of the mean end-to-end distance and the distribution of end-to-end distances.

Excluded volume effects:

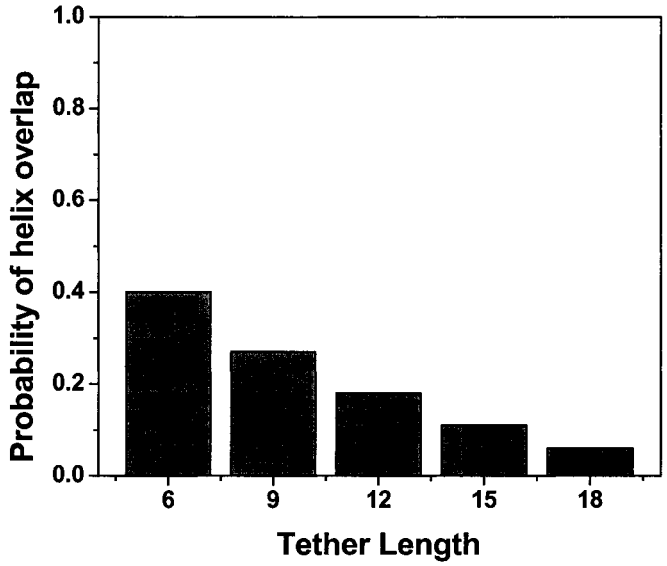
In order to correct for the excluded volume of the helices and steric overlap, simulations are carried with finite size helices. The 30 Å long, 10 Å diameter helices overlap in about 1/3 of the configurations of a six residue tether. For an eighteen residue chain, the occurrence decreases to 5% (Fig. 2.7a). The elimination of these disallowed configurations results in a more extended center-to-center probability distribution. Hence, the entropy of the tethered helices is higher, and the loss of translational entropy is, in fact, reduced. Numerically, this effect results in a decrease of only 0.43 e.u. (2.9 % decrease from the original S_{tether}) and 0.12 e.u (0.7% decrease) in cross-linking entropy for six and eighteen residues respectively.

In order to investigate excluded volume effects due to the steric clashes of non-neighboring residues, we performed hard-sphere simulations similar to Pappu et al(Pappu et al., 2000) (Fig. 2.7b). Polyalanine chains were constructed using dihedral angles randomly chosen from all the sterically allowed possibilities in a dipeptide (without restriction to a given region such as α-helical, as was done by Pappu et al). The number of overlapping chains increases nearly linearly from 7 to 16% going from six to twelve residues. Furthermore, the effect on the end-to-end probability distribution also is very small. As this distribution directly relates to the entropy, excluded volume effects of the tether with itself (and presumably with the helices that extend out in the opposite direction) has a minimal impact on the entropy of tethering.

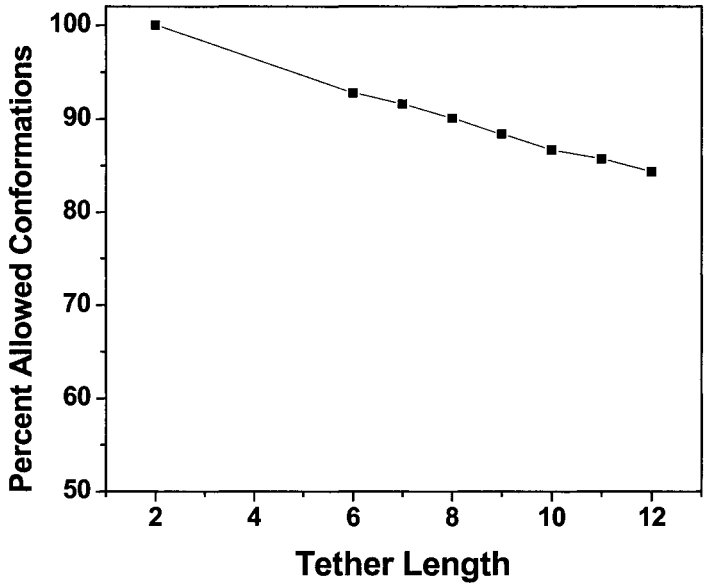
Figure 2.6 Excluded volume effects.

a). The fraction of configurations where the finite sized helices (30 Å long, 10 Å diameter) overlap for different length poly-alanine tethers. b) Number of allowed configurations for different length poly-alanine chains. The steric clashes of non-nearest neighbors were calculated using the hard-sphere model described in the text.

a)


Probability of helix overlap
1.0
0.8
0.6
0.4
0.2
0.0
6
9
12
15
18
Tether Length

b)


Percent Allowed Conformations
100
90
80
70
60
50
2
4
6
8
10
12
Tether Length

Comparison to other methods:

The NN Method provides an explanation as to why the Cell Theory of Liquids provides a reasonable result, even though the basic premise, that the solute particles are localized to cells, is invalid. The distribution function for the untethered system, $P_{\mathrm{dim}er}^{helix}(r)$, has nearly the equivalent entropy to that of a uniform distribution within volume equivalent to the average volume per solute molecule. Hence, the Cell Theory of Liquids provides a reasonable approximation even though the basis of the method is unrealistic.

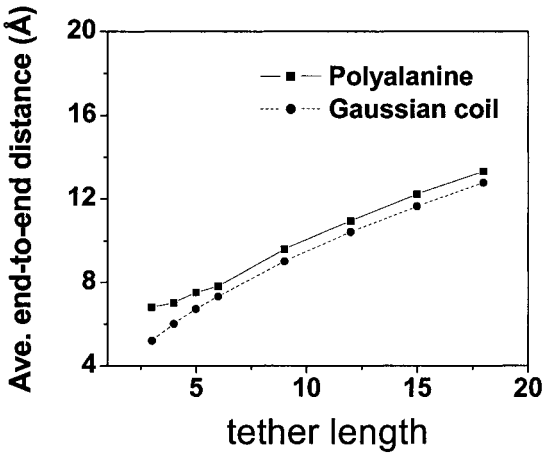
Likewise, the NN Method for calculating the entropy of a free helix (not the loss of entropy for the tethering process) agrees with the gas phase Sackur-Tetrode Eq. in that a solute particle has nearly the same effective volume per particle in both methods. We point out, however, that the Sackur-Tetrode Eq. for classical particles requires the *ad hoc* introduction of an extra factor of N^{-1} in the volume per particle to account for the particles' indistinguishability. No such correction is needed in the NN Method.

Amzel(Amzel, 1997) as well as Whitesides, Shahknovich and co-workers(Mammen et al., 1998) have presented similar corrections to the Sakur-Tetrode Eq. to account for restriction due to the finite volume occupied by solvent molecules. The effective, or free volume of the solute is the volume the center of the solute can sample without bumping into a solvent molecule. Using this method, Amzel accurately estimates the entropy of liquid water. Amzel also predicts the loss of entropy upon binding (Process 4↔6) by comparing the free volume for the solute in solution to the solute's free volume when it is bound to another protein.

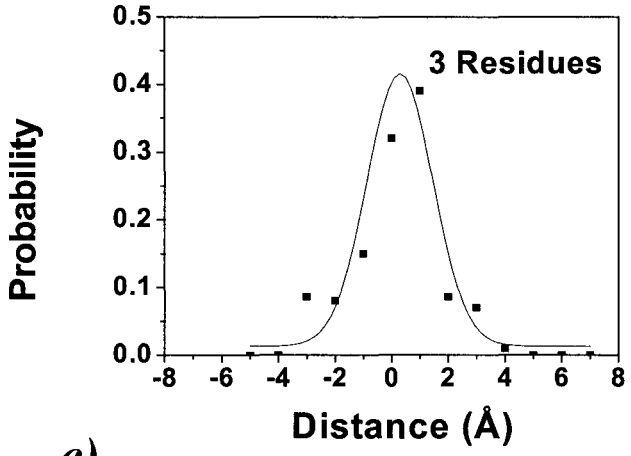
Figure 2.7 Properties of a polyalanine tether.

a) Comparison between end-to-end distances for Gaussian random coils and polyalanine chains. The distributions are significantly different for three residue polyalanine chains, but are similar for six residues (and longer) chains. b) Probability distribution of end-to-end distances for Gaussian random coils and polypeptides. c) The results for six or more residues are well approximated by a Gaussian distribution. The solid line represents a Gaussian fit whose width is within 10% of the corresponding Gaussian random coil distribution of the same length.

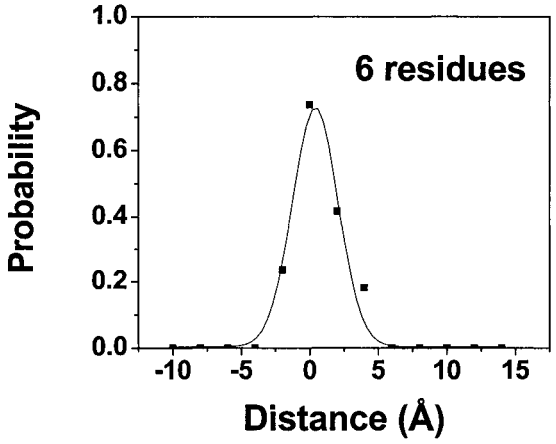
a)


Ave. end-to-end distance (Å)
20
16
12
8
4
5
10
15
20
tether length
Polyalanine
Gaussian coil

b)


Probability
0.5
0.4
0.3
0.2
0.1
0.0
-8
-6
-4
-2
0
2
4
6
8
Distance (Å)
3 Residues

c)


Probability
1.0
0.8
0.6
0.4
0.2
0.0
-10
-5
0
5
10
15
Distance (Å)
6 residues

Mammen et al(Mammen et al., 1998) accurately estimated the standard entropy of condensation of monatomic gases using the concept of free volume. They also estimate the (maximum) loss of entropy for a binding reaction of a tethered system (Process 4↔6), as distinct from our calculation where entropic benefit of introducing the tether is calculated (Process 4↔6 relative to Process 3↔5). Regardless, our calculation is similar in that the loss of entropy is based upon a comparison of the volume accessible prior to docking to that available afterwards. However, certain details are different, for example, the present analysis is based upon probability distribution functions of the tether.

Another method used to calculate the translational entropy of binding or cross-linking in aqueous media is the empirical quantity "cratic entropy" (mixing entropy) given by:

$$\Delta S = -R \ln 55 \qquad (2.12)$$

The value of 55 reflects the ratio of the concentration of 1 M solute to the molarity of pure water. Cratic entropy was proposed by Gurney(Gurney, 1953) and Kauzman(Kauzmann, 1959) to explain the entropy of mixing of ideal solutes in pure solutions. This correction is not derived from any principles of thermodynamics(Holtzer, 1995) or statistical mechanics(Gilson et al., 1997; Janin, 1996).

Comparison to Experiments:

In this section we compare the NN method to the results for proteins where the entropy or the free energy between the monomeric and dimeric versions has been measured (Table 2.3).

Table 2.3. Comparison between experiment and NN method

Protein (structure)	Experimental ΔS ($\Delta G^{o}_{bi} - \Delta G_{uni}$)	Predicted ΔS ($\Delta G^{o}_{bi} - \Delta G_{uni}$)
GCN4 coiled-coil(Moran et al., 1999) (α-helical)	~0 (~0)	-10.66 (3.2)
Designed coiled-coil(Yu et al., 1999) (α-helical)	5±8 (-1.5±2.4)	-3.4 (1.0)
SSI(Tamura & Privalov, 1997) (α/β)	-5±4 (1.5±1.2)	-7.20 (2.2)
Arc repressor(Robinson & Sauer, 1996) (α/β)	-11.8± 0.5 (3.5± 0.2)	-13.4 (4.0)

Entropy values given in the table are given in e.u., stability in kcal mol^{-1}, calculated according to $\Delta G = -T\Delta S$ at 300 K.

For variants of the dimeric GCN4-p1 coil coiled and analogs cross-linked with a disulfide bridged, Cys-Gly-Gly amino terminal tether, the difference between the dimeric and unimolecular stability was approximately zero(Moran et al., 1999). For this system, however, the NN method predicts the dimeric stability is stronger by 3.2 kcal mol^{-1} for the difference in the stability of the two versions. The excess stabilization of the tethered species may reflect stabilizing interactions between the tether and the helices. The amount of denaturant sensitive surface buried in the native structure (the *m*-value) was increased by about 10% in the tethered version. Potentially, the helix was capped by the glycines in the tether, which stabilized additional helical structure(Krantz et al., 2000).

Privalov and coworkers have performed detailed calorimetric measurements of the stability of homodimeric coiled coil and a version cross-linked with an internal disulfide bond(Yu et al., 1999). They showed that the entire change in free energy upon cross-linking was due to the loss of entropy upon introduction of the tether. The heat capacity of these molecules was not affected by the cross-link, and therefore, the vibrational modes were not perturbed. Also, the $\Delta H_{folding}$ did not change upon introduction of the tether. The pre-docked conformation is essentially fully denatured in this system, thus the calculated loss of entropy is for the difference in ΔS between Processes $1 \leftrightarrow 2$ and $5 \leftrightarrow 6$. For a system with just an internal disulfide bond, the loop closure entropy in the folded complex is minimal (~1 e.u). Therefore, the entire entropy change probably is only due to the tethering process in the denatured state. For such a system, according to the NN Method, the loss of entropy is 3.4 e.u., which is in good agreement with the experimental results (5±8 e.u).

In a very similar study by Tamura and Privalov(Tamura & Privalov, 1997) on Streptomyces subtilisin inhibitor (SSI), the authors have measured the loss of entropy upon cross-linking by calorimetric and magnetic resonance methods. For this homodimeric system, cross-linked by only a single internal disulfide bond, the NN method predicts the loss of entropy to be –7.20 e.u. which is close to the experimentally determined value of –(5±4) e.u.

Robinson and Sauer(Robinson & Sauer, 1996) examined the effect of a 15 residue cross-link on stability and folding kinetics on dimeric arc repressor. The cross-link connects the C-terminus of one Arc subunit to the N-terminus of the second subunit (Arc-L1-Arc). Comparison of the equilibrium stabilities of the linked and unlinked proteins yielded a C_{eff} = 2.7± 0.7 mM. We model the system by taking into account the length of the cross-link and the end-to-end distance between the C-terminus of one Arc subunit and the N-terminus of the second subunit. For the Arc-L1-Arc system, the NN method predicts C_{eff} =1.2 mM, or a difference of 1.6 e.u. from the observed value.

Conclusion:

A clear distinction exists between binding and cross-linking, two processes which are often considered to be equivalent. This chapter presents a method to calculate the entropic benefit of cross-linking. The major underpinning of the method is the realization that the probability distribution for the dimeric system represents the probability that a partner can travel a given distance while still being the closest molecule to a reference helix. Also, by comparing entropies based upon probability distributions, the method is independent of the nature of the solvent. The contribution of rotational entropy is

relatively negligible for a reasonable length tether. The NN method is applicable to a wide variety of protein systems. The methods outlined in this paper can be applied to higher order association processes and nucleic acid hairpins, although the exact results should be sequence dependent. For dimeric proteins, the introduction of a ten-residue tether results in an effective concentration of reactants in the millimolar range. When the concentration of the dimeric system is less than this concentration, the introduction of the tether will increase the fraction of docked species, and generally, result in a net stabilization for systems with mM or stronger dissociation constants.

3. THEORETICAL TREATMENT OF MACROMOLECULAR REACTIONS THROUGH MULTIPLE PATHWAYS

Introduction:

With recent advances in experimental and theoretical methods to study macromolecules, there is a rebirth of interest in the temperature-dependent kinetics of complex reactions(Bryngelson et al., 1995; Chan & Dill, 1998). These theoretical and experimental methods have provided extremely useful information about the potential energy surfaces and the pathways that lead from one state to another. However, the picture is still incomplete.

One uncertainty is the effect of multiple pathways on the temperature dependence of reactions. An analysis of this issue can provide vital information about the shape of the potential energy surface. The problem has been addressed by both experimental (Krantz & Sosnick, 2001; Moran et al., 1999) and theoretical (Chan & Dill, 1998) research groups in recent years. One example is the protein-folding problem where the energy landscape picture has been quite useful. Determining the presence or absence of multiple pathways can provide critical information about the folding landscape and the overall folding behavior. Therefore, a clear experimental signature of multiple pathways can be a very useful tool for the interpretation of complex reactions. In this study, we show that the temperature dependence of rate coefficients k(T), specifically the deviation from the traditional Arrhenius linear dependence of ln k on 1/T, can contain this information.

The analysis of complex kinetics through multiple pathways can be divided into three categories; parallel pathways, sequential pathways and combinations thereof. This chapter, the first case is studied by applying the transition state theory to calculate the rate of a two-state reaction with parallel pathways having different activation energies. The method discussed takes into account the implicit temperature dependence of the reaction rates, and addresses the consequences of the distribution of activation energies of pathways. The results show that the presence of multiple pathways results in curvature in Arrhenius plots. However, the sign of the curvature depends upon the relative rates of increase, with energy, of the densities of states of the saddle and the initial state.

Methods:

According the Transition State Theory (TST),

$$k = \frac{k_B T}{h} \exp(-\frac{\Delta G^{\|}}{RT}) = \frac{k_B T}{h} \exp(-\frac{\Delta H^{\|}}{RT}) \exp(\frac{\Delta S^{\|}}{R}) \qquad (3.1)$$

where the superscript || refers to the activation parameters.

The activation enthalpy of many macromolecular systems (e.g. protein-folding process) is of the order of a few kcals/mol (Scalley & Baker, 1997; Wolynes et al., 1996). Therefore in our model, we use the Arrhenius approximation of $\Delta H^{\|} = E_a - RT$. The entropy of activation of the system is calculated by considering the densities of states, ρ(E), for the initial and the saddle states in the rate-determining step (see equations 3.2 and 3.3 below). The contribution to the entropy of activation from initial and transition (saddle) state densities depends on temperature. For our model, we assume that the

population in the density of states either increases linearly or quadratically with temperature. More complex behavior is plausible, but in the absence of specific information about any particular system, and in the present situation of our examination of general behavior, the linear and quadratic models are sufficient to demonstrate the behavior we seek to interpret.

We also assume that if the density of states of the reactant increases linearly with temperature then the density of states for the saddle state increases quadratically and vice-versa. Once again, this is to ensure that we are using the simplest possible model. For a process such as folding or unfolding, the densities of states of the initial and the saddle states do not increase at the same rate. Therefore we use this assumption to model such processes where the densities of states of the initial and the saddle state increase at different rates.

The change of entropy of between two given states (initial and transition or transition or final) at a given temperature is then calculated by using the densities of states:

$$\Delta S(T) = R(\langle \ln \rho^{TS}(T) \rangle - \langle \ln \rho^{I}(T) \rangle) \tag{3.2}$$

where the superscripts TS and I refer to the densities of states at the saddle state and the initial state respectively. The overall rate coefficient is thus given by:

$$k(T) = \frac{k_B T}{h} \exp(1 - \frac{E_a}{RT}) \exp(\Delta S(T)) \tag{3.3}$$

The density of states at the saddle is a measure of the multiplicity of pathways, but no reference need be made to the extent to which there is mode-coupling among them. In the system with multiple pathways, the pathways differ from each other due to different E_a.

For an ensemble of pathways, each with different activation energies $E_{a,i}$, the observed rate coefficient is the sum of rate coefficients through each individual pathway i is given by:

$$k_{observed} = \sum_i k_i \tag{3.4}$$

Assuming that the activation energies have a Gaussian distribution, we can determine P (E_a), by using the standard equation of a normalized Gaussian distribution with mean activation energy μ and variance σ_{Ea}:

$$P(E_a) = \frac{1}{\sigma_{Ea}\sqrt{2\pi}} \exp\left(-\frac{(E_a - \mu)}{2\sigma_{Ea}^{\ 2}}\right) \tag{3.5}$$

This value of P (E_a) is multiplied to its corresponding E_a in equation 3 to ensure the Gaussian weighting of activation energy. Hence the overall rate depends on the nature of the distribution of activation energies, the mean activation energy and the variance.

Model Systems and Results:

Model 1*:*

Two different model systems are examined in this study. In the first model, there are two routes, a singular low energy pathway with activation energy E_a, and an ensemble of *g*-fold degenerate higher energy pathways having a higher activation energy $E_b = E_a + \delta E_a$ (Fig. 3.1a). The purpose of this model is to examine how the presence of an ensemble of high energy pathways affects the behavior of k_{obs}; explicitly, identifying the temperature range where the flux switches from the lower energy route to the higher energy routes. The net rate of the reaction depends, in addition to the temperature and the

activation energy barriers of the two routes, on the degeneracy of the higher energy pathways.

If the reaction rate through each pathway obeys the Arrhenius approximation, the net rate coefficient of the reaction can be written as the products of the exponentials of entropic and enthalpic terms:

$$k_{obs} = A\exp(\ln[1])\exp(-\frac{E_a}{RT}) + B\sum\exp(\ln[g])\exp(-\frac{E_b}{RT});$$
$$k_{obs} = A\exp(-\frac{E_a}{RT}) + B\sum\exp(\ln[g])\exp(-\frac{E_a + \delta E_a}{RT}) \quad (3.6)$$

where A and B are constants which are obtained from the initial conditions of the reaction. The summation is over all the possible pathways. The degeneracy of the low energy pathway is 1 and hence its entropic contribution is zero. The degeneracy of the higher energy pathway is *g*, and hence its entropic contribution is ln [g].

The rate coefficient of the reaction for a system depicted in Fig 3.1a as a function of temperature is plotted in Fig 3.1b with g = 4000, E_a=10 kcal and δ = 0.51.

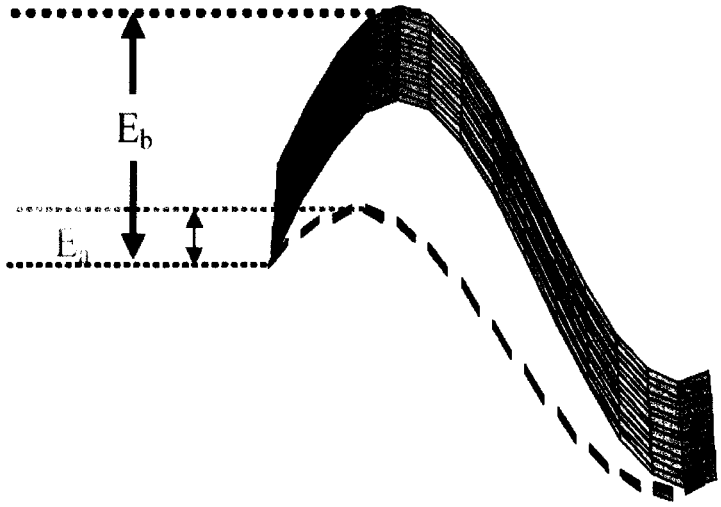
Model 2*:*

This model is in fact an extension of the first model, as this model takes a sharp distribution of energies in Model 1 and spreads them out over a range of energies. The reactant molecules can now go through a wider distribution of singular pathways, i.e. pathways with a Gaussian distribution of activation energies with a variance σ_{Ea}. The rate

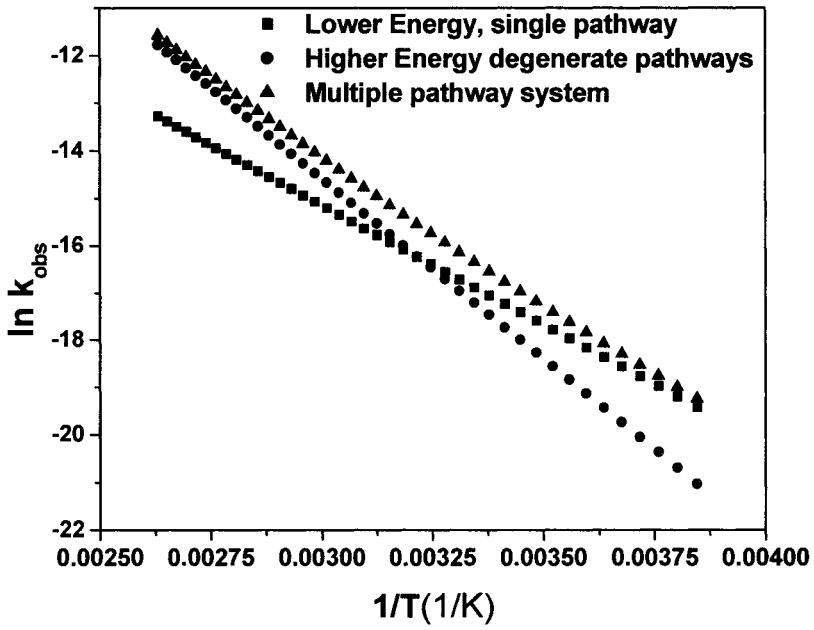
Figure 3.1. Model # 1: The two-pathway reaction

a) The reaction diagram of a two-pathway system. The single lower energy pathway, with a single activation energy E_a, and a higher energy ensemble of pathways with activation energy $E_b = E_a + \delta E_a$ and degeneracy "g", where $g = 4 \times 10^3$.
b) The Arrhenius plot for the system shown in 1a, with $\delta = 0.51E_a$. The observed rate coefficient through only the lower energy pathway, and through the ensemble is plotted to show the "temperature switch" that changes the flux from one pathway to another pathway.

a)


E_b
E_a

b)


Lower Energy, single pathway
Higher Energy degenerate pathways
Multiple pathway system
ln k_{obs}
-12
-14
-16
-18
-20
-22
0.00250
0.00275
0.00300
0.00325
0.00350
0.00375
0.00400
1/T(1/K)

through each pathway then depends upon the density of states in the reactant and the saddle state, whose populations vary with temperature as discussed above in the methods section. The rate of reaction through each individual pathway is calculated according to Eq. 3.3 and the overall observed rate coefficient is given by Eq. 3.4.

The plots of ln k versus 1/T for different mean activation energies are shown in Fig 3.2 (a and c). These plots are obtained by calculating the rate coefficient for N (N=1500) pathways in 5 K temperature intervals. Fig. 3.2 (a) shows the behavior of a two-state system for which the density of states for the reactant increases faster (increases quadratically with T) than the density of states of the in the saddle state (increases linearly as T), i.e. $\rho^{I} > \rho^{TS}$. Fig. 3.2 (c) depicts the opposite behavior ($\rho^{I} < \rho^{TS}$).

The plots in Fig. 3.2 (a and c) show increased curvature at lower and higher temperatures. The curvatures increase several fold at even higher(>373) and lower (< 270) temperatures; however such a temperature range is unlikely to be observed experimentally for biological systems. The results are plotted for various mean activation energies (μ); the variance of the distribution (σ_E) is kept the same for these plots. If the density of states of either the reactant or the saddle state were higher and varied more rapidly, the curvature would of course appear within a narrower range of temperature. For this system, our calculations suggest that dH/ dT is not a constant with temperature, so we have a non-zero value of dC_p/dT. For a process whose the density of states increases faster in the initial state than in the saddle state (i.e. $\rho^{I} > \rho^{TS}$), dC_p/dT is negative, the opposite effect is observed in the case in which the density of states of the saddle state increases faster (Figs. 3.2 b and d respectively).

Figure 3.2. Model # 2: Log Rate Coefficient vs 1/Temperature for different mean activation energies.

a) The log of the observed rate coefficient of the reaction versus 1/T, also known as an Arrhenius plot. The y-axis is $k' = (\frac{kh}{k_B T})$ as often used in macromolecular kinetics plots[9,10]. The origin of curvature is described in detail in the text. The plot represents a system for which the density of states of the initial state increases at a faster rate than that of the saddle state.
b) dH/dT (= C_p) versus temperature for a system for which the density of states of the initial state increases at a faster rate than that of the saddle state. The curvature in the plot suggests that the slope is non-zero ($dC_p/dT < 0$).
c) The y-axis is the same as in Fig 3.2a. The plot represents a system for which the density of states of the initial state increases at a slower rate than that of the saddle state.
d) dH/dT versus temperature for a system for which the density of states of the initial state increases at a slower rate than that of the saddle state. The curvature in the plot suggests that the slope is non-zero ($dC_p/dT > 0$)
The plots are based on different mean activation energies of the Gaussian distribution, however the variances of the distributions are kept the same.

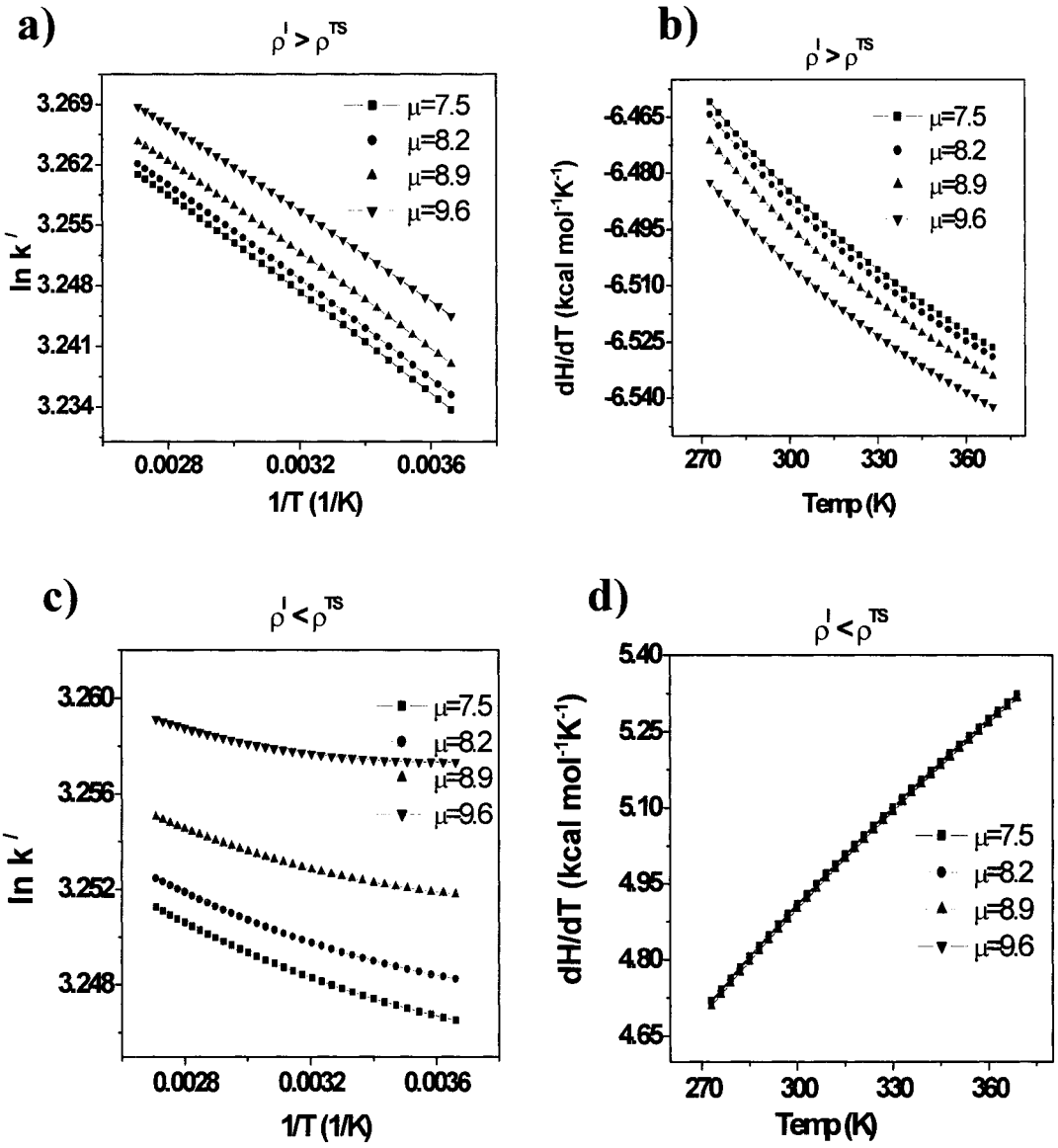

a)
$\rho^I > \rho^{TS}$
3.269
3.262
3.255
3.248
3.241
3.234
ln k'
0.0028
0.0032
0.0036
1/T (1/K)
μ=7.5
μ=8.2
μ=8.9
μ=9.6
b)
$\rho^I > \rho^{TS}$
-6.465
-6.480
-6.495
-6.510
-6.525
-6.540
dH/dT (kcal mol^{-1}K^{-1})
270
300
330
360
Temp (K)
μ=7.5
μ=8.2
μ=8.9
μ=9.6
c)
$\rho^I < \rho^{TS}$
3.260
3.256
3.252
3.248
ln k'
0.0028
0.0032
0.0036
1/T (1/K)
μ=7.5
μ=8.2
μ=8.9
μ=9.6
d)
$\rho^I < \rho^{TS}$
5.40
5.25
5.10
4.95
4.80
4.65
dH/dT (kcal mol^{-1}K^{-1})
270
300
330
360
Temp (K)
μ=7.5
μ=8.2
μ=8.9
μ=9.6

Fig 3.3 shows the behavior of the natural log of observed rate coefficient versus the variance of the distribution at different temperatures. The plots show the behavior of the system at different mean activation energies (μ). Fig 3.3(a) show the behavior of the model two-state system whose density of states of the initial state increases at a faster rate (increases quadratically with temperature) than that of the saddle state (increases linearly with temperature), whereas the opposite effect is depicted in Fig 3.3(c).

Discussion:

The first model involving just two pathways with one lower-energy pathway and an ensemble of degenerate higher energy pathways, though very simple, is quite useful. It shows that the presence of high energy paths can influence the rate of the reaction significantly at physiological temperatures (Fig 3.1b), under certain conditions, e.g. when the degeneracy of higher energy paths is high ($g > 10^3$) and the activation energy gap between the lower and higher energy paths is small ($E_b - E_a \leq 0.5\ E_a$). This model also shows the presence of a temperature "switch", at which the flux from one pathway is greater than that of the other. Such information is going to be very useful to design, conduct and interpret macromolecular kinetic experiments at different temperatures.

The results using model # 2 also show curvature in the Arrhenius plots. As stated earlier, the plot shows more curvature at higher (> 373 K) temperatures than in the intermediate temperature range ($273 < T < 373$). The curvature in our model is due to the presence of multiple pathways and the individual rates of reaction through those pathways. Since we are working with a Gaussian distribution, at lower temperatures

Figure 3.3. Model # 2: Log Rate Coefficient vs Variance of Gaussian distribution plots for different mean activation energies and different temperatures.

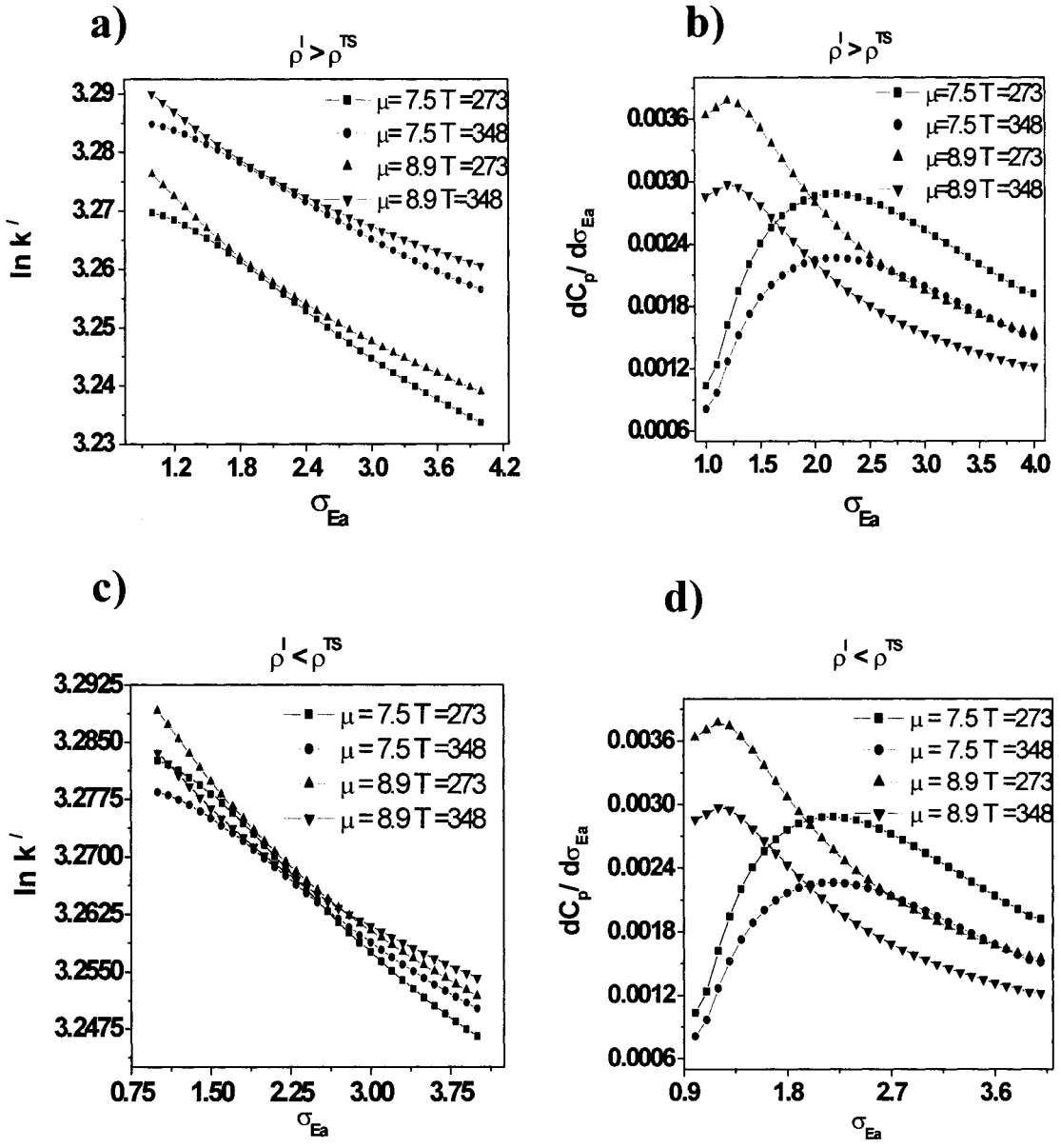
a) The log of the observed rate of the reaction as a function of the variance of the Gaussian distribution. The y-axis shown is $k' = (\frac{kh}{k_B T})$. The plot represents a system for which the density of states of the initial state increases at a faster rate than that of the saddle form.

b) The plot of change of C_p with variance magnifies the fact that these variance plots are very sensitive to changes in activation energy and insensitive to changes in temperature.

c) The plot is similar to Fig. 3.3a and represents a system for which the density of states of the initial state increases at a slower rate than that of the saddle state.

d) The plot is similar to plot 3.3b, except the system plotted is the one for which the density of states of the initial state increases at a slower rate than that of the saddle state

The plots have two different mean activation energies and are plotted at two different temperatures.


a)
$\rho^I > \rho^{TS}$
ln k′
σ_{Ea}
μ= 7.5 T =273
μ= 7.5 T =348
μ= 8.9 T =273
μ= 8.9 T=348
b)
$\rho^I > \rho^{TS}$
$dC_p / d\sigma_{Ea}$
σ_{Ea}
μ=7.5 T=273
μ=7.5 T=348
μ=8.9 T=273
μ=8.9 T=348
c)
$\rho^I < \rho^{TS}$
ln k′
σ_{Ea}
μ = 7.5 T =273
μ = 7.5 T =348
μ = 8.9 T =273
μ = 8.9 T =348
d)
$\rho^I < \rho^{TS}$
$dC_p / d\sigma_{Ea}$
σ_{Ea}
μ = 7.5 T =273
μ = 7.5 T =348
μ = 8.9 T =273
μ = 8.9 T =348

pathways with very low activation energies are populated, whereas at T> 310 K the flux goes through all the possible pathways.

The origin of curvature in Arrhenius plots in macromolecular processes such as protein folding has been a subject of debate (Bryngelson et al., 1995; Chan & Dill, 1998; Scalley & Baker, 1997; Wolynes et al., 1996). According to some theories, the origin is due to diffusion on the landscape (Bryngelson et al., 1995; Chan & Dill, 1998; Scalley & Baker, 1997; Wolynes et al., 1996) whereas the experimentally observed curvature in Arrhenius plots is usually attributed to hydrophobic effect and other temperature dependent interactions that stabilize the folding molecule (Scalley & Baker, 1997).

First principle theoretical methods have failed to reproduce the experimentally observed curvature in Arrhenius plots for protein-folding reactions. In fact, these theoretical results predict curvature in the Arrhenius plots with a sign opposite to that of experiments. This is because the entropic contribution is modeled as a term completely independent of temperature, and hence there is an increase in the observed rate at higher temperatures (also shown in model # 1; fig 3.1b). Based on our model (model # 2, where the densities of states of initial and the saddle state, and their implicit dependence on temperature are taken into account) we can correctly predict the sign of the curvature for protein-folding and unfolding reactions. Though there might be a difference in magnitude of the rate (because of complexities that our model does not take into account) our results agree fairly well with the experimental results (Chen & Schellman, 1989; Jackson & Fersht, 1991) in predicting the sign of the curvature in the Arrhenius plots.

For a protein-folding reaction, the density of states of the initial unfolded state increases faster than that of a saddle state, whereas the opposite effect is to be expected

for the unfolding reaction. This is because the random coil is less restrictive than the transition (saddle) state and hence is expected to have a density of states, up to some rather high energy, that increases rapidly. The transition state is presumably more constrained, hence its density of state increases more slowly, and the native state is the most restricted of all, and hence has the slowest-growing density of states. Interpreting our model with this assumption gives the observed curvature for both the unfolding and the folding reactions.

Our results also suggest that for a multiple path process the effective dH/dT is nonzero. Though the change in magnitude is relatively small, the slope of the plot of C_p (= dH/dT) versus temperature is non-zero for both cases discussed above (Fig 3.2 b and 3.2 d).

From a theoretical perspective, reaction along a single pathway will result in a constant C_p if no other interactions (e.g. hydrophobic effect) are taken into account. This is not the case with our model, since the densities of states of the initial and saddle states are populated as functions of temperature, and result in a temperature-dependent heat capacity, i.e. $dC_p/dT \neq 0$.

It is sometimes said that the hydrophobic effect results in the curvature in the Arrhenius plots. The origins of hydrophobic effect (difference in C_p in the unfolded and the folded state) lie in the difference in organization of water around the folded and the unfolded molecule and the difference in accessible energy states. Therefore the hydrophobic effect can be interpreted as a microscopic model that carries with it a difference in densities of conformational states between the unfolded and transition (or folded) states. This, in turn, results in curvature in the Arrhenius plots. Therefore our

analysis is not contradictory, but complimentary to other work showing the curvature in Arrhenius plots due to hydrophobic effect. Rather, our model shows that *any* effect associated with the differing patterns of densities of state in initial and transition states will yield such curvature.

Another interesting feature of the presence of multiple pathways is the compensation between enthalpic and entropic factors. The enthalpic term in the free energy for a folding reaction is fairly small, and is the only term depending explicitly on the activation energy. However, this might not be the case for other macromolecular reactions, where the enthalpy plays a more significant role. Eq. 3.1 indicates that enthalpy and entropy act in opposite directions. In our model, for a system whose density of states of the reactant increases faster than the saddle state in the rate-determining step (i.e. $\rho^{I} > \rho^{TS}$), the entropic contribution is positive, therefore entropy and enthalpy work in opposite directions (see eq. 3.1, 3.2 and 3.3). Thus for a system with a small enthalpic contribution, or small E_a (e.g. folding reaction) the curvature is primarily due to the entropic contribution. On the other hand, the enthalpy and entropy terms in the rate of the reaction act in the same direction for a system whose density of states of the saddle state increases faster than that of the initial state (i.e. $\rho^{I} < \rho^{TS}$). This is because the entropic contribution in eq. 3.1 and 3.3 is negative, and thus both entropy and enthalpy have the same sign. This is observed from Fig 3.2c (blue curve, μ =9.6), which has the maximum curvature, due to higher activation energy and hence greater enthalpy. Because the enthalpic component of the overall reaction rate depends only upon the activation energy and is independent of any variation in the density of states, the enthalpic component shows the same behavior for the forward and the backward reaction. Consequently, for

macromolecular processes which are governed overwhelmingly by enthalpic forces, we predict the shape of the ln k vs. 1/T plot would be the same for forward and backward reactions.

We can also predict the rates of reactions through parallel pathways as a function of the variance of the distribution of activation energies (Fig. 3.3). Though it is not yet possible to observe the rate of the reaction experimentally as a function of this parameter, nonetheless these plots can give us valuable information about the behavior of the system as a function of the nature of the distribution of activation energies. One of the most striking features of these plots is the change in the shape of the curve as a function of the mean activation energy μ (Figs 3.3a). At the same temperatures, the two curves (Fig 3.3a. black (μ =7.5) and green curves (μ =8.9)) have different shapes, whereas the curves with the same activation energy (Fig 3.3a) have similar shapes even though the systems have different temperatures.

A similar observation can be made about Figs 3.3c for which the density of states of the initial state is higher than that of the saddle state in the rate-determining step (i.e. $\rho^{I} > \rho^{TS}$). The only source of difference in these curves is the difference between the activation energies. Similarly the variations in temperature do not seem to have any significant effect on the shape of the curves. These results show the effect of the change in E_a (i.e. enthalpic contribution) on the overall rate coefficient which is not obvious from the Arrhenius (ln k vs 1/T) plots. It is interesting to note that whereas variations in the entropic term in the rate equation result in changes in the shape of Arrhenius plots, the variance plots show the dependence of the rate coefficient on the changes in activation energy and are insensitive to temperature changes. This observation is magnified in the

corresponding plots in which the derivative of C_p with variance is plotted against variance (Fig 3.3 b and d).

From a theoretical perspective, the Arrhenius plots represent only part of the overall picture. We believe that these rate-variance plots, in combination with rate-temperature plots will give an improved picture of the contribution of various thermodynamic terms to the overall rate.

Conclusion:

In this chapter we present a simple method to address the issue of reactions across a single barrier per pathway in a system of multiple pathways. The presence or absence of multiple pathways can play a very important role in visualizing the potential energy surface of a reaction. This chapter also addresses the issue of the effect of multiple parallel pathways on the overall kinetics of complex reactions. Though initial theoretical studies of macromolecular kinetics failed to agree with experiments, our study shows that the inclusion of the density of states in the initial and the saddle state corrects for the previous discrepancy between theoretical and experimental results. Our study indicates that presence of multiple parallel pathways and the density of states assignment to the initial and the saddle state plays a key role in providing this curvature observed in the Arrhenius plots in protein folding reactions although other factors may also contribute.

The Arrhenius plots show high sensitivity to the density of states, as well as to the variance of the distribution of energies. The rates at which actual densities of states vary with temperature are presumably more complicated than the one discussed here; however

the good agreement with experiments suggests that more sophisticated modifications of the dependence of densities of states on temperature will only improve the quality of the results. The success of our model in predicting the correct sign of curvature in Arrhenius plots is intended to stimulate experiments designed and conducted to observe the presence of multiple pathways, since initial comparisons to experimental results are very encouraging (Chen & Schellman, 1989; Jackson & Fersht, 1991; Scalley & Baker, 1997). More sophisticated calculations taking into account the densities of states of the initial and saddle states, as well as different rates at which these two densities of states increase will provide a more rigorous test of our theory.

4. COMPARISON BETWEEN UNITED ATOM AND EXPLICIT ATOM FORCE FIELDS

Introduction

The accuracy, speed, and reliability of computer simulations depend crucially upon the force fields employed. The commonly used "atomistic" force fields can be divided into two broad categories, namely all-atom (also called explicit atom) and united atom force fields. The explicit atom force fields treat each atom in the molecule as an interaction site (Williams, 1967), whereas the united atom force fields unite the carbon atoms and their directly bonded hydrogen atoms into single, often spherically symmetric interaction sites(Ryckaert & Bellemans, 1978). In other words, the explicit atom force fields represent CH_3, CH_2 and CH in terms of four, three and two interaction sites respectively, whereas the united atom force fields employ single pseudo-atom representations for each CH_3, CH_2 and CH group. The explicit atom force fields appear to be more realistic and are believed to be more appropriate at higher densities. However, the use of united-atom representations is quite desirable because the reduction of aliphatic groups to a single pseudo-atom can increase the simulation speed by as much as an order of magnitude (Martin & Siepmann, 1999) and can thereby render computationally expensive simulations quite inexpensive and tractable on a personal computer. For example, the ambitious "Folding@Home" program currently employs the united atom OPLS force field as part of its efforts at simulating the folding rates of small proteins(Shirts & Pande, 2001; Smith et al., 1993).

Equilibrium and dynamic properties of alkanes (Karayiannis et al., 2002; Martin & Siepmann, 1998; Martin & Siepmann, 1999; Smith & Yoon, 1994; Smith et al., 1993) silicone polymers (Sides et al., 2002) , as well as tri-alanine in water (Mu et al., (in press); Stock & Mu, 2002), have been simulated extensively using both explicit atom and united atom models, and the corresponding structural, equilibrium, and dynamic properties have been compared to experimental data. These studies delineate the regimes where the united atom models agree well with experiments. In contrast, however, very little analysis exists comparing the predictions of these two types of force fields for the conformational dynamics of flexible peptides, i.e., for systems that access a wide range of conformations that depart considerably from those in the neighborhood of the native structures of folded proteins.

As in their application to problems in polymer physics, computer simulations for biopolymers have contributed to the understanding of equilibrium and dynamical properties of these large and complex systems at short and long time scales. This understanding of the dynamics of these systems is important in identifying potential candidates for better and more effective drug design, for the influence of mutations on different proteins, and for the interaction of proteins with ligands, membranes, and solvents. In order to perform these simulations with reasonable computational times, united atom models are often used, for example in studies of protein folding (Bryant et al., 2000), but no direct comparison has been made of the dynamical properties emerging from simulations using explicit atom and united atom models. Moreover, force fields for proteins have generally been devised by comparison with experimental thermodynamics and structural data (perhaps with some *ab initio* information), but the peptide dynamics

samples a wide range of non-native conformations and thereby provides a far more stringent test of the quality of the potential functions. Thus, force fields that predict similar native structures may deviate considerably in their description of the dynamics of very flexible peptides such as the one studied herein.

We study the long time dynamical behavior of a penta-peptide, Met-enkephalin using an implicit solvent method that has been tested by comparison with explicit solvent MD simulations for this same peptide (Shen & Freed, 2002b) and for the initial stages of folding of the villin headpiece (Shen & Freed, 2002a). Shen and Freed have shown that implicit water LD simulations are 200 times faster than explicit solvent MD simulations(Shen & Freed, 2002b), therefore making the study of peptide dynamics more tractable on personal computers. The additional speed enhancement from using united atom models scales roughly as the square of the ratio of numbers of united atom groups to explicit atoms, which for Met-enkephalin is a factor of $(75/57)^2 = 1.73$.

Met-enkephalin (Tyr-Gly-Gly-Phe-Met) is one of the smallest neurotransmitter peptides and was first isolated from pig brains (Hughes et al., 1975). This peptide has been studied extensively using X-ray crystallography (Hughes et al., 1975), NMR (Smith & Griffin, 1978) and computer simulations (Deber & Behnam, 1984; Shen & Freed, 2002b; Wang & Kuczera, 1996). This large body of research has established that similar to other short peptides Met-enkephalin does not exhibit a single native conformation. Rather, Met-enkephalin rapidly traverses a wide range of different conformations in aqueous solution (Graham et al., 1992; Shen & Freed, 2002b). The implicit solvent-LD method has been tested against explicit solvent MD simulations, and has been shown to work better than nearly two dozen other implicit solvent methods screened (Shen &

Freed, 2002b). In order to compare and contrast the predicted dynamical behavior of Met-enkephalin by united atom and explicit atom models, we study the long time behavior (~130 ns) using six different commonly used force fields. Four of these force-fields (AMBER 94, AMBER 96, CHARMM-27 and OPLS all-atom) are explicit atom force fields, whereas two of them (OPLS and CHARMM-19(Neria et al., 1996)) employ the united atom method. Three pairs are matched sets that have been developed by the same group and therefore provide checks on their internal consistency. The main aim of our study is not to show the superiority or inferiority of any of the force fields; rather it is aimed at suggesting criteria that can be used to test and improve existing force fields.

Computational Details

The implicit water Langevin Dynamics (LD) simulations follow the procedures discussed in detail by Shen and Freed. Thus, the method is reviewed only briefly in this section. Within the implicit solvent model, the total system energy is given by

$$U_{tot} = U_b + U_{bend} + U_{imp\text{-}tors} + U_{tors} + U_{ch}(\varepsilon) + U_{vdw} + U_{solv}, \qquad (4.1)$$

where the subscripts *b*, *bend*, *imp-tors*, *tors* and *vdw* denote the bonding, bond-bending, improper torsions, torsions, and van der Waals interaction terms in the overall system energy. The subscript *ch* designates the contribution involving the dielectric screening of electrostatic interactions, and the subscript *solv* denotes the solvation potential portion of the overall system energy. The implicit water expression for the overall system energies differs from its explicit water counterpart in presence of the dielectric screening in $U_{ch}(\varepsilon)$ and in the solvation term (U_{solv}) that replace the Coulomb interactions and the protein-solvent and solvent-solvent interactions, respectively, in the explicit solvent treatments.

We utilize the macroscopic solvation potentials U_{sol}, given by the Ooi-Scheraga solvent-accessible surface area (SASA)(Ooi et al., 1987) method because comparisons between implicit and explicit solvent simulations demonstrate the superiority of this potential in more faithfully approximating the results of explicit solvent simulations. The potential contains a contact free energy term that is evaluated in terms of the accessible surface area (σ_i) of all atoms i in the peptide. The SASA accessible surface area is computed from a hypersurface bisecting the first solvent shell using a water (probe) radius of 1.4 Å. Therefore, the overall solvation free energy can be written as a sum of free energy contribution from all atoms,

$$U_{solv} = \sum_{i=1}^{i=N} g_i \sigma_i , \tag{4.2}$$

where g_i denotes the empirical atom solvation energy parameters[28] determined by fitting experimental aqueous solvation free energies of amino acids and selected organic compounds to equation 2.

The LD simulations are based on non-linear generalized Langevin equations (GLE) and a procedure similar to MD simulations, apart for the need of an additional algorithm for computing the frictional forces and the corresponding random forces that represent the frictional forces due to implicitly treated water. These friction coefficients are computed by the method of Pastor and Karplus(Pastor & Karplus, 1988) and are updated every 100 integration steps. More explicitly, the LD simulations are generated by integrating the atom positions and velocities by using the standard velocity Verlet algorithm,(Allen, 1987)

$$r_i(t+\Delta t) = r_i(t) + c_{1i}v_i(t)\Delta t + \frac{1}{2}c_{2i}a_i(t)\Delta t^2 + r_{gi} \quad (4.3)$$

$$v_i(t+\Delta t) = c_{0i}v_i(t) + c_{1i}a_i(t)\Delta t + v_{gi} \quad (4.4)$$

The coefficients c_{0i}, c_{1i} and c_{2i} are given by

$$c_{0i} = \exp(-\zeta_i \frac{\Delta t}{m_i}) \quad (4.5)$$

$$c_{1i} = (-\zeta_i \frac{\Delta t}{m_i})^{-1}(1 - c_{0i}) \quad (4.6)$$

,

$$c_{2i} = (-\zeta_i \frac{\Delta t}{m_i})^{-1}(1 - c_{1i}) \quad (4.7)$$

where m_i is the mass of i^{th} atom. The r_{gi} and v_{gi} are Gaussian random variables with variances depending upon the friction coefficients in the usual manner. The friction coefficients ζ_i are determined from the solvent accessible surface area (σ_i') with zero probe radius using stick boundary conditions,

$$\zeta_i = 6\pi\eta r_{eff:i} \quad (4.8)$$

where η is the solvent viscosity and $r_{eff:i}$ is the effective hydrodynamic radius of atom i that is computed from the solvent accessible area by

$$r_{eff:i} = \sqrt{\frac{\sigma_i'}{4\pi}} \quad (4.9)$$

As mentioned above, we compare and contrast the results of two united atom and four all atom force fields that are present in commercially available packages. The LD simulations have been performed using a modified version (by Shen and Freed) of the

TINKER software package(Ponder, 1999) for protein molecular dynamics simulations. The parameters of the CHARMM-19 force field were incorporated into the TINKER package using the published parameters by Karplus and co-workers (Neria et al., 1996). Both the explicit atom and united atom force fields have been used to generate 130 ns trajectory of Met-enkephalin. The simulations use a 1.5 fs time step and are run on 1.4 GHz Pentium IV and 1.2 GHz AMD machines.

Results and Discussion

Met-enkephalin does not occupy a unique native state in aqueous solution; rather, its dynamics are highly flexible. Previous studies show that the molecule jumps between extended, semi-packed, and packed states(Shen & Freed, 2002b) (Fig 4.1 a, b, and c, respectively), that may be classified in terms of the distribution for the square of the radius of gyration, which is defined by

$$R_g^2 = \frac{1}{N}\sum_{i=1}^{i=N}(\mathbf{r}_i - \mathbf{r}_g)^2 \qquad \textbf{(4.10)}$$

where $\mathbf{r}_g$ denotes the position of the center of gravity of the molecule and $\mathbf{r}_i$ denotes the position of the ith atom. The extended state is classified as the range of conformations where $R_g^2 > 40$ Å^2, the semi-packed state for conformations having R_g^2 between 20 Å^2 and 40 Å^2, and the packed state for conformations with R_g^2 less than 20 Å^2. The normalized distributions for R_g^2 are computed by binning the values of R_g^2 into 100 discrete regions with values between 0 and 60 Å^2.

Figure 4.2 exhibit the probability distributions of R_g^2 for the six different force fields. The first interesting feature of Fig. 4.2 is the contrast between the explicit atom

a)

b)

c)

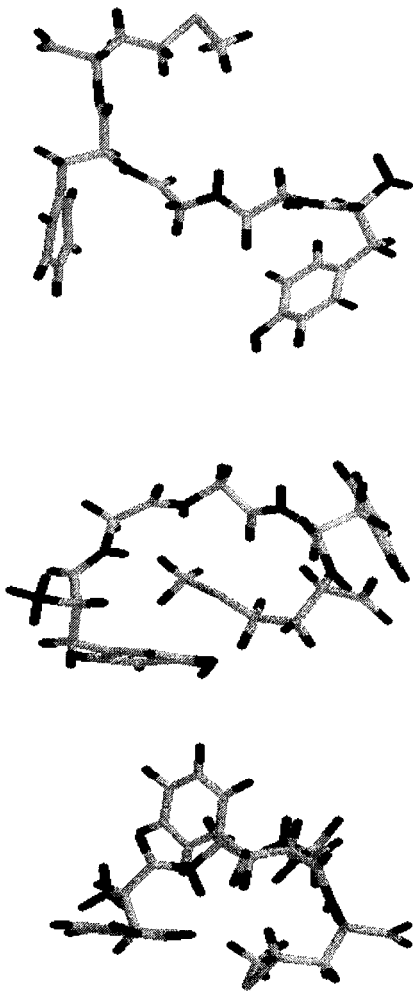


Figure 4.1. Three classes of conformation accessed during the dynamics of Met-enkephalin.

a) Typical examples of extended (R_g^2 ~ 42 A^2), b) semi-compact (R_g^2 ~33 A^2), and c) compact (R_g^2 ~18 A^2) conformations. Carbon atoms are shown in gray, nitrogen in blue, oxygen in red and hydrogen atoms in green.

force fields. Whereas the all-atom AMBER 94 and CHARMM-27 have fairly similar behaviors, the OPLS all-atom force field displays different dynamics in which the Met-enkephalin molecule spends most of its time on the border between semi-compact and extended conformations and almost never samples the compact state. AMBER 96 shows a behavior similar to that of OPLS all atom force field. In contrast, the CHARMM-27 and AMBER 94 all-atom simulations describe Met-enkephalin as remaining mostly on the border between compact and semi-compact states and spending very little time in the extended state. These differences are, however, rather small compared to the departures between the explicit atom and united atom distributions. The LD simulations with the CHARMM-19 united atom force field suggest that the molecule remains in the semi-compact state (very close to the compact state) essentially all of the time, in sharp contrast with the other force-field predictions for which Met-enkephalin prefers one state but still significantly samples all the other conformations. The OPLS united atom distribution departs considerably from the OPLS all-atom distribution but is more similar to the other all-atom distributions for the radius of gyration. The OPLS united atom simulations thus do not overly constrict the peptide to one conformation as is found for the CHARMM-19 united atom simulations, but the OPLS united atom distribution under samples the compact state.

The comparison of the united-atom and explicit atom distributions in Fig. 2 raises two important questions, namely the origin of the differences between the two kinds of force fields and the observation that the CHARMM-19 force field produces a distribution significantly different than those of the other force fields. The answers to both these questions require a careful analysis of the parameters of the force fields. The primary

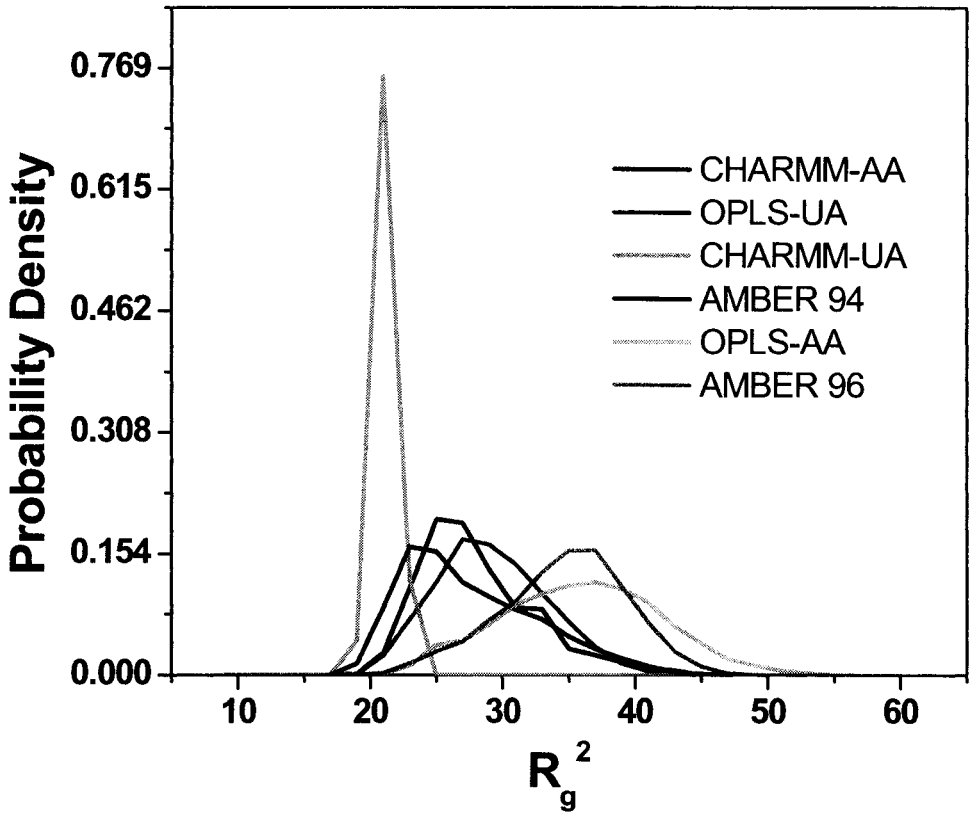


Figure 4.2. The probability distribution for the radius of gyration R_g^2 for Met-enkephalin as computed with the six different force fields.

The radius of gyration is computed using trajectories of 130 ns by employing the implicit solvent LD simulations using the methods of Shen and Freed (Shen & Freed, 2002b). The plot displays the R_g^2 distribution for the six force fields discussed in the text.

difference between united atom and explicit atom force fields (apart from the numbers of interaction sites) is the presence of partial charges on aliphatic groups. Both CHARMM-19 and OPLS united atom force fields assign a vanishing partial charge to all CH_3, CH_2 and CH groups that are connected to other aliphatic groups. This difference can then lead to energetics of molecular conformations that significantly differ from the ones that are observed for the explicit atom force fields where these groups have partial charges on the carbon and hydrogen atoms. Consequently, both the united atom force fields lead to dynamics that sample only parts of the overall conformational space available and, more explicitly, constrict the molecule to only one region of conformational space. The resolution of the second question regarding the differences between the simulations with CHARMM-19 and with the other force fields also requires a careful analysis of the parameters. The CHARMM-19 force field uses a wide variety of "wild-card" parameters; in other words, there are several torsional and improper-torsional parameters that have been assigned the same *ad hoc* value regardless of the types of atoms present. This departs from the procedure used in developing the other force fields, although they too contain a few torsional parameters that are assigned the same *ad hoc* values. This constancy of torsional parameters may not cause a problem for modeling the dynamics of certain alkanes, but proteins have a highly diverse group of bonds and torsions. A proper description of protein dynamics may require more careful parameterization of the torsions associated with such bonds. The difference between the results computed with the AMBER 94 and AMBER 96 force fields is probably due to modified torsional parameters in AMBER 96, which are based upon empirical data and which have been adjusted to reproduce the energy difference between extended and constrained alpha

helical energies for the alanine tetrapeptide. The similarity between the results of AMBER 96 and OPLS all atom force field simulation has also been observed by other groups (Mu et al., (in press)). Likewise, the results produced by AMBER 94 and CHARMM-27 have also been found to be quite similar (Mu et al., (in press)). The two united atom force fields discussed here also differ in their description of the van der Waals radius of the hydrogen atoms attached to nitrogen or oxygen atoms. While the OPLS-united atom force field assigns a value of zero to the van der Waals radius of hydrogen atoms attached to oxygen or nitrogen atoms, CHARMM-19 assigns a non-zero radius. This difference can also lead to divergent results between the two united atom force fields. Similarly, slight differences among the explicit atom force fields also arise because the OPLS all atom force field specifies a vanishing van der Waals radius for hydrogen atoms bonded to oxygen or nitrogen atoms, whereas AMBER (94 and 96) assigns a non-zero radius. CHARMM-27, on the other hand specifies a non-zero van der Waals radius for hydrogen atoms bonded to either oxygen or nitrogen atoms.

We also compare and contrast a variety of time-correlation functions (TCF) for Met-enkephalin as predicted by the four explicit atom and two united atom force fields. The TCFs compared here are P_1 dipole autocorrelation functions of the interatomic position vectors, which are defined as,

$$C_{ij}(t) = \frac{< \boldsymbol{l}_{ij}(0) \bullet \boldsymbol{l}_{ij}(t) >}{l_{ij}^2} \tag{4.11}$$

where the interatom vectors $\boldsymbol{l}_{ij}$ are $\boldsymbol{l}_{ij} = \boldsymbol{r}_i - \boldsymbol{r}_j$. The angular brackets in Eq. 11 denote the equilibrium average. The P_1 correlation function depicts the local or global flexibility of

the molecule depending upon whether the atoms i and j are distant or proximate. We compare three TCFs for the united and explicit atom force fields to sample some interesting local and global motions. Explicitly, these TCFs are those for the end-to-end vector, the C_γ-C_γ vector, and the central backbone C-C vector.

The statistical error in the correlation function C(t) due to the finite trajectory is estimated by the method of Zwanzig and Ailawadi (Zwanzig & Aliwadi, 1969) as:

$$\sigma = \sqrt{\frac{2\tau'}{T}}[(1-C(t)] \tag{4.12}$$

where $T >> \tau'$ is the duration of the trajectory and τ' is the correlation time defined by

$$\tau' = \int_0^\infty [C(t')]^2 dt' \tag{4.13}$$

Figures 4.3-4.5 present several computed TCFs and contain error bars placed at 500 ps, 1000 ps and 2000 ps with the extremum values $C_\pm(t) = C(t) \pm \sqrt{\frac{2\tau'}{T}}[(1-C(t)]$.

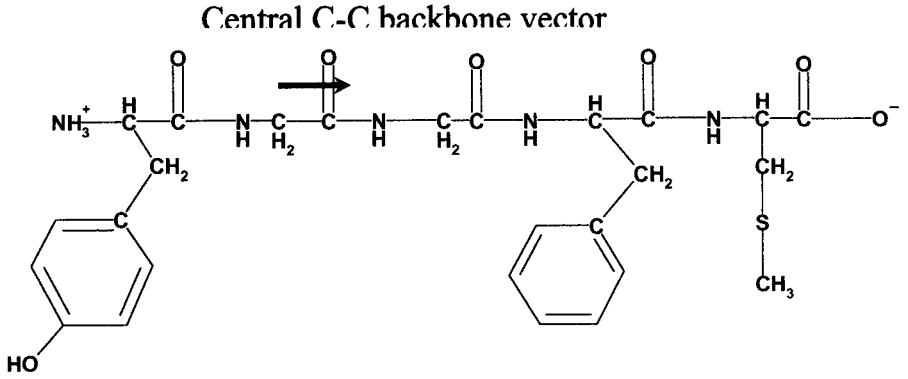
Figure 4.3 depicts the dipole correlation function for the central backbone C-C bond that is exhibited over the curves. The short time dynamics produced with the united atom force fields are fairly similar, but this correspondence is lost at longer times where the CHARMM-UA (i.e. CHARMM-19) force field curve decays faster than its OPLS counterpart. The explicit atom C-C TCFs (except for that from AMBER 96) are fairly similar, and partially mirror the similar R_g^2 distributions from the explicit atom AMBER 94 and CHARMM-27. A larger difference between the predictions from the OPLS-UA

and CHARMM-UA force fields at longer time scales is also obvious from Fig. 2 for the R_g^2 from the two force fields. The TCFs for the C-C backbone vector (Fig. 4.3) from the two united atom force fields are more similar at short (~1 ns) time scales; however, the R_g^2 distributions are more sensitive to the dynamics over much larger (~ 120 ns) time ranges. The TCFs for the two united atom force fields differ considerably and probably reflect the differences in the corresponding Rg^2 distributions. Interestingly, the TCFs from both OPLS force fields become more similar for longer times.

Figure 4.4 presents the TCFs for the dynamics of the end-to-end (N(Tyr1) to O(Met5)) vector. (The TCF for the backbone end-to-end ($C_{1\alpha}$–$C_{5\alpha}$) vector behaves very much the same.) The OPLS-UA and OPLS-AA TCFs have quite similar shapes but yield very different R_g^2 distributions. Once again, this difference probably arises from the difference in time scales that are relevant to the two sets of properties. The OPLS-UA and OPLS-AA TCFs agree better at shorter time scales but begin to depart for longer times. The AMBER 94 and CHARMM-AA (CHARMM-27) force fields again produce fairly similar TCFs, but the dynamics from these two force fields differ considerably from those calculated using the OPLS-AA force field. Once again, the AMBER 96 force field correlation function decays the slowest.

Figure 4.5 displays the TCFs for of the phenyl-phenyl C_γ-C_γ vector. Once again, the CHARMM-UA TCF decays the fastest, the AMBER 96 decays the slowest, the TCFs from the OPLS-UA and OPLS-AA force fields are quite similar, as are the AMBER 94 and CHARMM-AA TCFs.

A further analysis of the TCFs reveals several important features about the nature of the force fields. The OPLS-UA and OPLS-AA force fields yield similar TCFs at



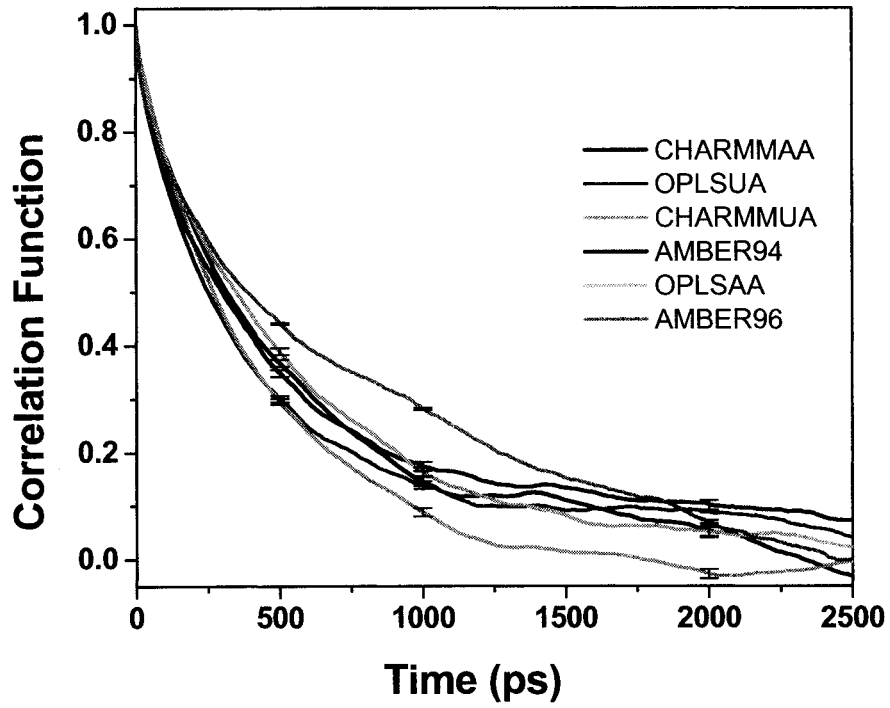


Figure 4.3. The central backbone C-C bond TCF from different force-fields.

The backbone vector is shown by an arrow in the structure of Met-enkephalin. The two united atom force fields exhibit similar behavior at very short time scales (< 1000 ps), whereas the explicit atom force fields display similar dynamics.

shorter time scales for the motions that vary more slowly (such as the end to end vectors and the phe-phe C_γ-C_γ vector), whereas they depart considerably at shorter time scales for relatively rapid motions (such as the backbone C-C vector). The CHARMM-UA yields TCFs that depart substantially from those from all the other force fields, due to the reasons discussed above. The TCFs computed from the AMBER 94 and the CHARMM-AA force fields exhibit the most consistent agreement over all time scales and reflect the similarities in their R_g^2 distributions. While their shapes differ slightly, both sample almost equally from the three sections of the peptide conformational space. The differences between the distributions of R_g^2 from the OPLS-AA force field, on one hand, and from AMBER 94 and CHARMM-27, on the other hand, is also reflected in their respective TCF plots. Except for the C-C backbone dynamics, the TCFs from OPLS-AA deviate from those computed with the other three explicit atom force fields. The similarities between the OPLS-AA and OPLS-UA force fields TCFs at shorter time scales are lost in the longer time behavior of the peptide. The AMBER 96 force field shows R_g^2 distributions that are similar to OPLS-AA but depart considerably from all the force fields for all the three TCFs. Once again, we believe the similarity in the results between AMBER 96 and OPLS-AA becomes more pronounced at longer time scales, whereas for shorter time scales the two force fields produce significantly different results. The R_g^2 distributions and the TCFs demonstrate that the explicit and united atom force fields predict quite disparate dynamics, probably due to the absence of partial charges on aliphatic groups and because all of these force fields have been optimized for equilibrium structures and not for large scale conformational dynamics.

Conformational Dynamics:

We have also studied the time-dependent conformational dynamics of Met-enkephalin. The study is carried out by inspecting Ramachandran maps for the various torsional angles (Φ,Ψ) of the peptide (Fig. 4.6a). The phi-psi map shown is similar to the one used by Pappu *et al* (Pappu & Rose, 2002). The map is divided into eight basins as indicated by the labels. Figure 4.6(b-g) displays the time-dependent dynamics for the central glycine residue (Gly-3) of the peptide. The other residues in the peptide exhibit very similar patterns of behavior as functions of the force field, so the Gly-3 serves to illustrate the general trends. Figure 4.6 depicts the Ramachandran basin occupied by Gly-3 as a function of time. Table 4.1 summarizes the same information by presenting the time-averaged occupancy of the eight basins for Gly-3.

The figures and Table demonstrate that, apart from the CHARMM-19 force field simulations, the Gly-3 residue executes many hops between basins for all the other force fields. The CHARMM-27, AMBER –94 and CHARMM-19 trajectories show the peptide spending maximal time in basin 2, though for CHARMM-27, basin 1 is almost equally populated. The AMBER-94 and CHARMM-27 cases exhibit very similar basin populations for all basins except for basins 1 and 6. The OPLS-UA dynamics behaves very similar to CHARMM-27 apart from a much smaller presence in basin 2. The dynamical pattern produced by the AMBER-94 trajectory is strikingly different from that of AMBER –96, probably due to the different torsional parameters as discussed above. The AMBER-96 trajectory, in sharp contrast to AMBER 94, shows very little preference for any basins other than 1, 2 and 3. The Ramachandran populations from the OPLS-UA and OPLS-AA force fields exhibit similar behaviors just as observed for the TCFs. The

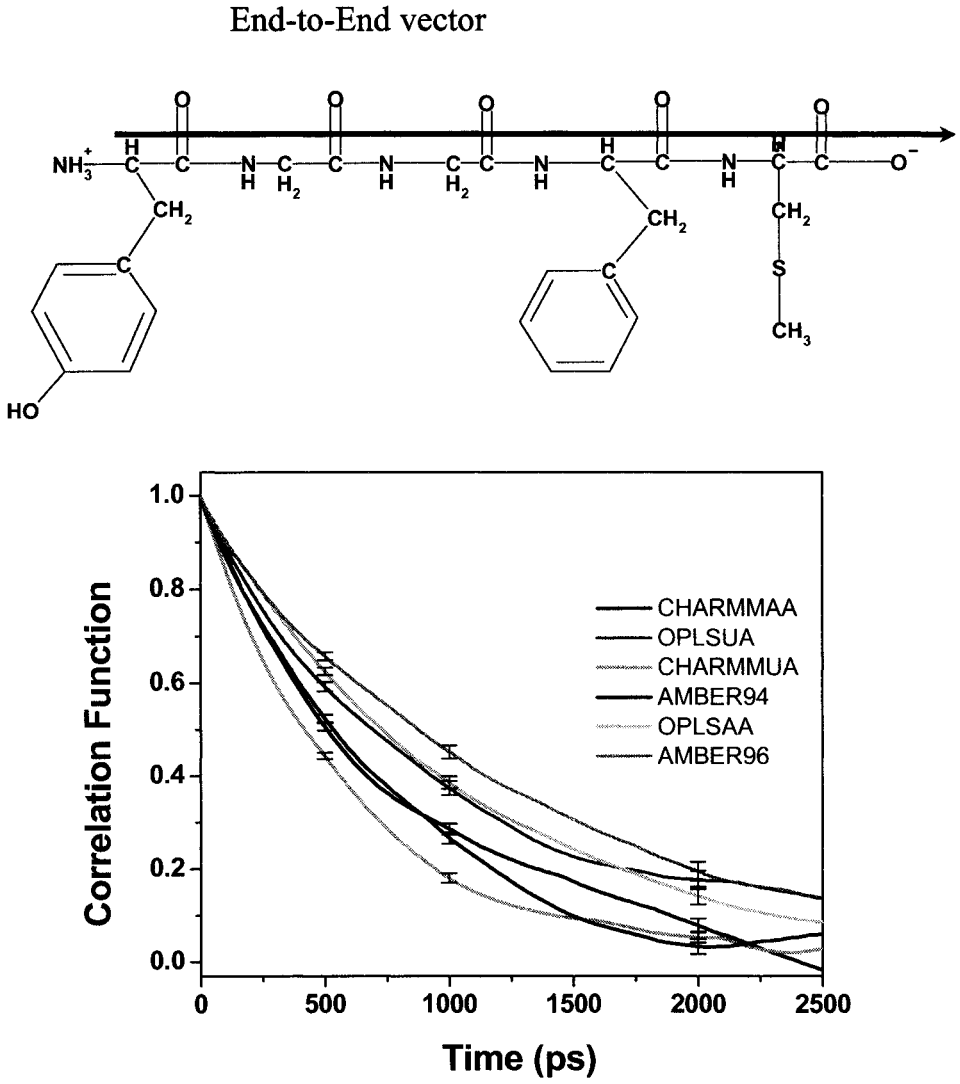


Figure 4.4. The peptide end-to-end [N(Tyr 1) – O(Met 5)] TCF from different force-fields.

The dynamics are computed using the two united atom and four explicit atom force fields. The end-to-end vector is depicted by an arrow in the structural diagram for Met-enkephalin. The TCFs from the OPLS-UA and OPLS-AA force fields are similar on this short time scale. The CHARMM-UA force field TCF decays faster than that of any other force field, the AMBER 96 force field TCF decays the slowest, whereas the AMBER 94 and CHARMM-AA TCFs are again quite similar.

OPLS-AA basin populations display a preference for basins 1, 5 and 7, and the central Gly residue almost never samples basin 2, which is the predominant basin for both the AMBER-94 and CHARMM-27 trajectories. Our conformational dynamics results for AMBER 94 and AMBER 96 show good agreement with Sanbonmatsu et. al's results using explicit solvent replica-exchange MD simulations on Met-enkephalin using PARM 94 and PARM 96. The phi-psi basin populations indicate similar dynamics from the AMBER-94, CHARMM-27 and OPLS-UA force fields, and sharp differences between AMBER-96 and AMBER-94, in accord with our observations from the R_g^2 and TCF plots. The conformational dynamics exhibit common characteristics for the OPLS-UA and OPLS-AA force fields as is also evident from TCF plots as well. Hence, our analysis demonstrates that a careful study of the conformational dynamics of small, very flexible peptides should provide additional information for improving the representation of current force fields in certain areas of conformational space that are not well sampled by folded protein structures. Further studies of additional dynamical properties that probe the dynamics at different time scales should provide a better overall picture the regimes where the force fields need further improvement.

Conclusion:

We present the first comparison of its kind for the dynamical properties of peptides that are predicted by united atom and explicit atom force fields. Our study considers six commonly used force fields and subjects them to rigorous tests for the dynamics of a short and highly flexible peptide. The united atom and explicit atom

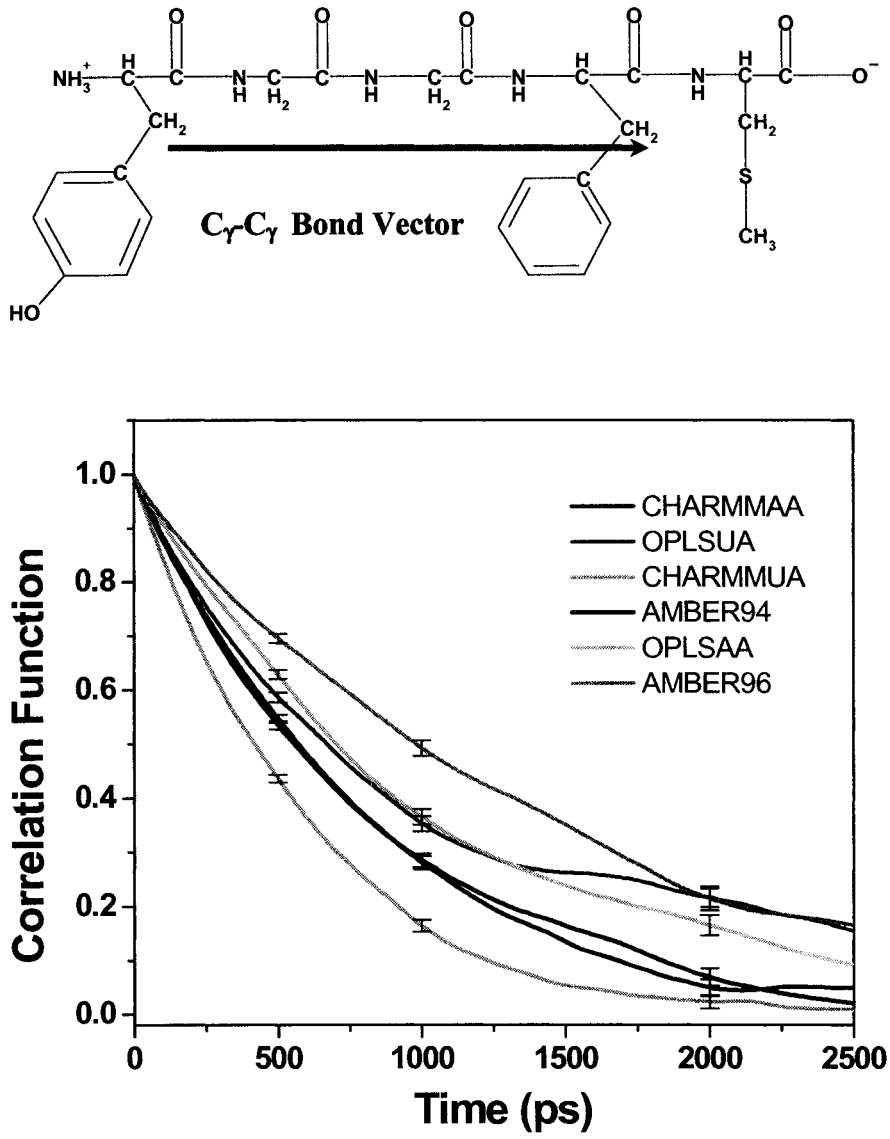


Figure 4.5. The Phe-Phe ($C_\gamma - C_\gamma$) vector

This vector represents another slowly varying variable. The TCF form the CHARMM-UA force field decays the fastest, the AMBER 94 and CHARMM-AA force field TCFs again are very similar, while at shorter times (~ 1200 ps) the OPLS-AA and OPLS-UA TCFs are close to each other.

Table 4.1 Percentage occurrences of Gly-3 individual Ramachandran basins

	BASIN 1	BASIN 2	BASIN 3	BASIN 4	BASIN 5	BASIN 6	BASIN 7	BASIN 8
AMBER-96	47.0	47.0	47.0	47.0	47.0	47.0	47.0	47.0
OPLS-AA	51.0	51.0	51.0	51.0	51.0	51.0	51.0	51.0
AMBER-94	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0
CHARMM-UA(19)	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
OPLS-UA	47.0	47.0	47.0	47.0	47.0	47.0	47.0	47.0
CHARMM-AA(27)	29.0	29.0	29.0	29.0	29.0	29.0	29.0	29.0

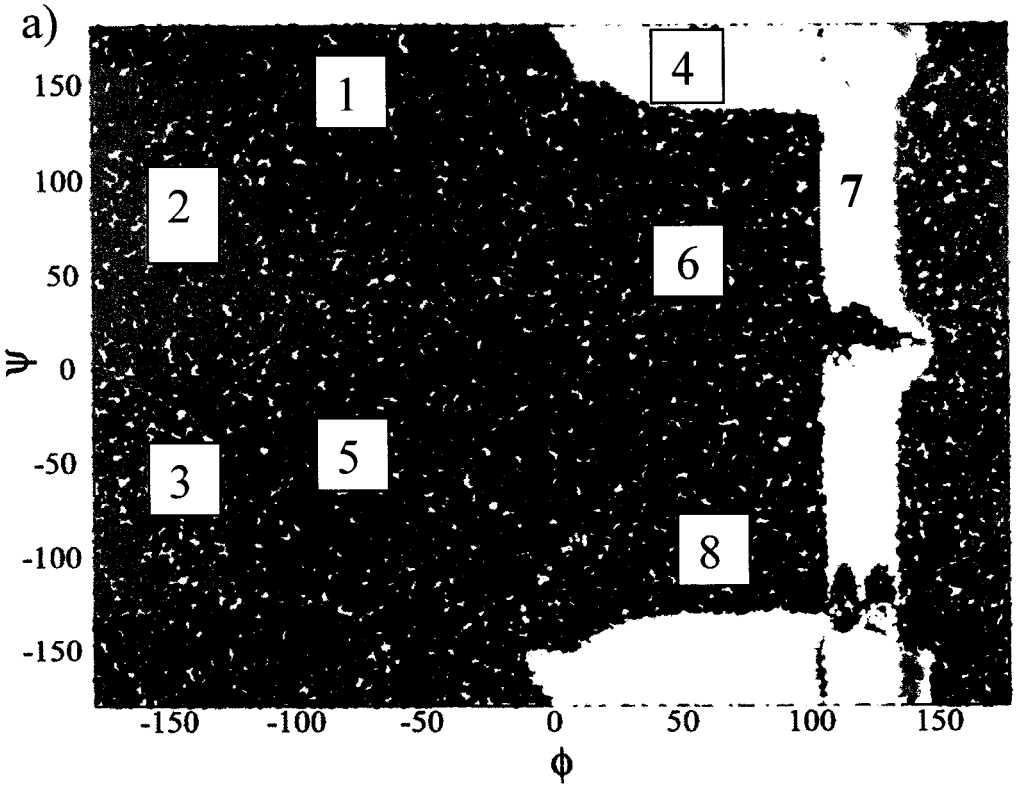
models exhibit different dynamical descriptions for the motion of the penta-peptide, Met-enkephalin. The all atom CHARMM-27 and AMBER 94 force fields produce rather similar results, whereas the OPLS explicit atom force field (OPLS-AA) produces slightly different dynamics which are similar to those found when using the AMBER 96 force field. There are however, significant differences in the results produced by AMBER-94 and AMBER-96.

The united atom CHARMM-19 force field suffers from limitations in not having partial charges on certain united atom groups and in using *ad hoc* "wild card" parameters for certain torsions, and these deficiencies are manifest in the simulations that constrict the peptide to remain close to a given conformation. These problems are not experienced by the explicit atom force fields whose simulations display a preference to one conformation but indeed sample the whole range of conformations. The OPLS-UA case is somewhat similar to that for the explicit atom force fields in this regard.

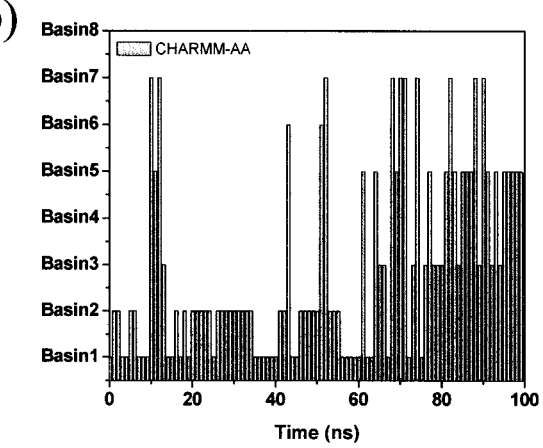
These comparisons are meant to identify problems and develop methods for improving the current force fields for studying biological molecules since both the united atom and explicit atom force fields have already shown great promise in treating alkanes (Sides et al., 2002). The approach presented in this chapter represents a point of departure for further comparisons between the use of united atom and explicit atom force fields for describing the dynamics of biological molecules. With the increasing power of MD simulations for biological systems, tests such as ours will be helpful in achieving a more accurate representation of both the short and long scale dynamics of proteins. However, further checks for larger and structurally different peptides will be extremely useful to aid

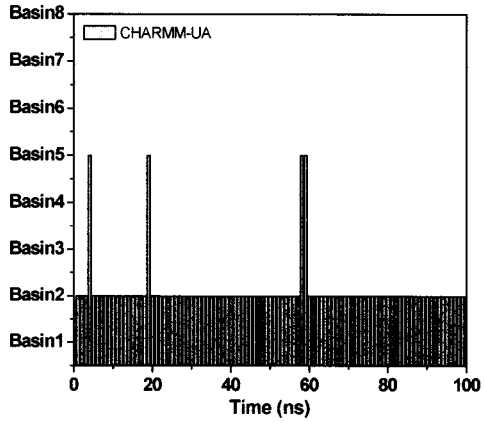
Figure 4.6. Time Dependent Conformational Dynamics of the central Gly-3 residue.

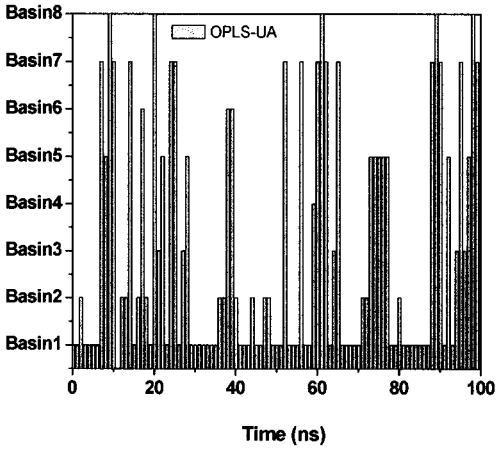
a) Phi-Psi plot of torsional angles is divided into eight basins labeled from 1 through 8. The plot is similar to the one reported by Pappu *et al.* and is reproduced with the author's permission. Basin occupations of Gly-3 are shown as functions of time for b) CHARMM-27 (explicit atom), OPLS-UA, CHARMM-19 (united atom), AMBER-94 , OPLS-AA and AMBER-96.

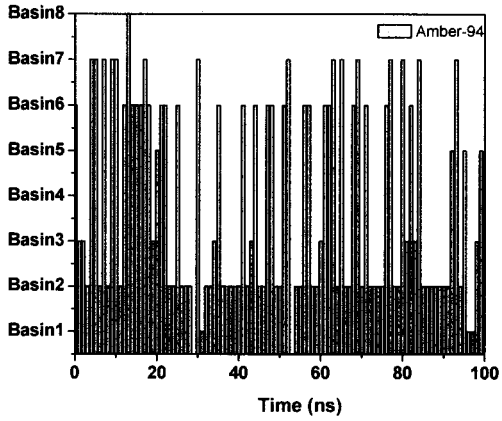

a)
150
100
50
0
-50
-100
-150
ψ
-150
-100
-50
0
50
100
150
ϕ
1
2
3
4
5
6
7
8

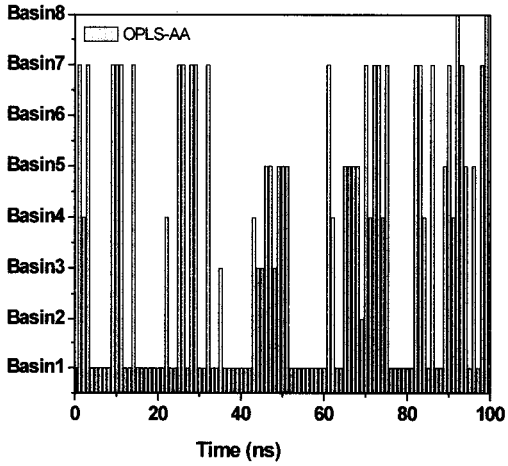
b)

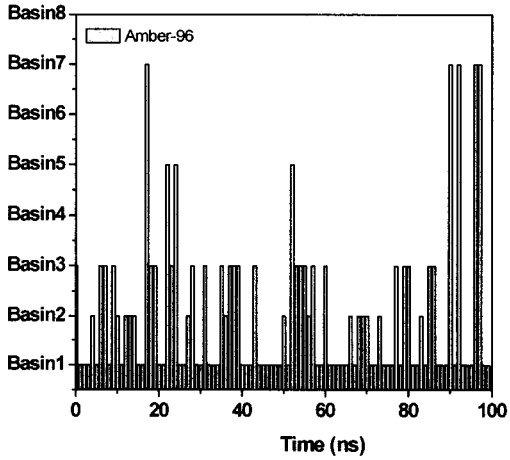

CHARMM-AA
Basin8
Basin7
Basin6
Basin5
Basin4
Basin3
Basin2
Basin1
0
20
40
60
80
100
Time (ns)


CHARMM-UA
Basin8
Basin7
Basin6
Basin5
Basin4
Basin3
Basin2
Basin1
0
20
40
60
80
100
Time (ns)


OPLS-UA
Basin8
Basin7
Basin6
Basin5
Basin4
Basin3
Basin2
Basin1
0
20
40
60
80
100
Time (ns)


Amber-94
Basin8
Basin7
Basin6
Basin5
Basin4
Basin3
Basin2
Basin1
0
20
40
60
80
100
Time (ns)


OPLS-AA
Basin8
Basin7
Basin6
Basin5
Basin4
Basin3
Basin2
Basin1
0
20
40
60
80
100
Time (ns)


Amber-96
Basin8
Basin7
Basin6
Basin5
Basin4
Basin3
Basin2
Basin1
0
20
40
60
80
100
Time (ns)

in overcoming the shortcomings of these force fields and thereby in developing faster, more accurate and realistic potentials.

5. BACKBONE DYNAMICS, FLORY ISOLATED PAIR HYPOTHESIS AND INTER-BASIN DYNAMICS OF AMINO ACIDS

Introduction

A fundamental descriptor of a polypeptide's conformation is the set of its backbone dihedral or torsional angles. For each residue, these angles specify a location in the Ramachandran plot of Φ, Ψ angles (Ramachandran et al., 1963; Ramachandran & Sasisekharan, 1968). The intrinsic preference for each peptide unit to be in one Ramachandran basin or another and the inter-basin hopping rates directly affect secondary structure preferences and residual structure in the denatured state, as well as the overall thermodynamics and kinetics of protein folding. In spite of this significance, only a few studies focus on the peptide backbone dynamics using atomic-level force fields (FFs) in an aqueous environment (Bolhuis et al., 2000; Hu et al., 2003; Mu et al., (in press)). Furthermore, an analysis of these backbone dynamics and structure is useful to reveal any dependence on context, including the conformation and chemical identity of the nearest neighbor (NN) residues. In this study we present such a study for amino-acetylated (Ace) and carboxy-amidated (Nme) versions of a mono-alanine "dipeptide" (i.e. Ace-Ala-Nme) and for di-and tri-amino acids (Fig. 5.1) with one, two and three pairs of Φ,Ψ dihedral angles, respectively.

Our analysis tests the applicability of the Flory isolated-pair hypothesis (IPH) (Flory, 1969), which is implicitly invoked in many equilibrium and kinetic treatments of protein folding, including helix-coil theories. According to the IPH, the Ramachandran basin populations of one residue are independent of its neighbors' conformations (except

for proline, and residues preceding prolines): " the interactions associated with rotations of one such independent pair are quite independent of the angles assumed by neighboring pairs" (Flory, 1969). When this pivotal isolated-pair assumption is valid, the backbone entropy of the system can be expressed as the sum of individual residues' entropies. Within the IPH, a single helix-coil equilibrium constant can be assigned to each amino acid species without qualification to either its neighbors configuration or identity, as is done in nearly all analyses of helix-coil transitions.

Pappu *et al.*(Pappu et al., 2000) consider the reduction in sampling due to nearest neighbor's configuration in polyalanine. In contradiction to the IPH, they find that the central residue, located between two residues with helical geometries, is sterically hindered by these neighbors. However, when the dihedral angles in a polypeptide are chosen according to their relative basin probabilities without restriction to the helical basin, the number of overlapping conformations is minor, for example, only 16% for a twelve residue chain (Zaman et al., 2002). Hence, steric hard-core type overlap provides only a minor reduction in the total conformational entropy of the unfolded state (in the absence of extensive helical configurations).

Molecular dynamic (MD) simulations have recently demonstrated that different force fields (FFs) can produce rather large differences in basin populations (Garcia & Sanbonmatsu, 2002; Hu et al., 2003; Mu et al., (in press)) and references therein]. Garcia and coworkers find that the AMBER 96 FF must be altered so that a largely alanine-containing peptide is predicted to undergo helix-coil transitions at the experimentally observed temperatures (Garcia & Sanbonmatsu, 2002). Their alteration involves the elimination of an additional, backbone dihedral, or torsional potential, which is present

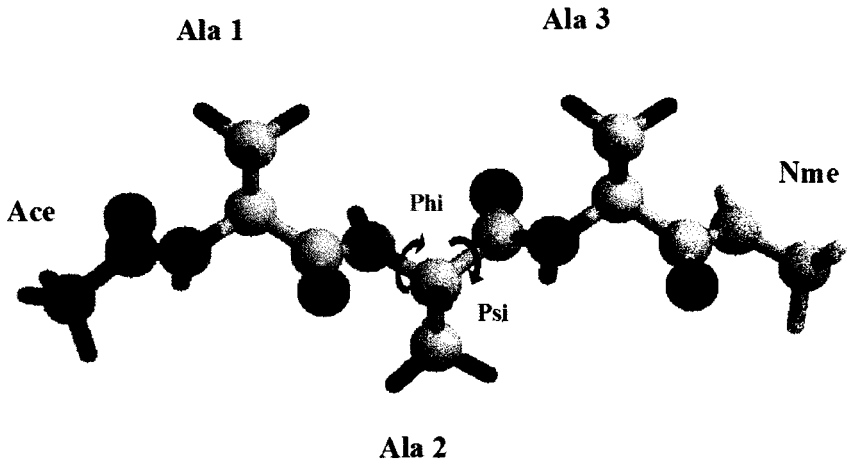


Figure 5.1. Tri-alanine

Ace-$(Ala)_3$-Nme peptide with center of the three pairs of backbone dihedral angles highlighted. The hydrogen atoms are shown in stick representation (black), whereas oxygen (red), nitrogen (blue) and carbon (grey) are depicted in ball-and-stick representation.

with varying topographies in most commonly used FFs (Hu et al., 2003). Upon elimination of this added potential, the basin preference in Garcia's FF is entirely determined by backbone, side-chain, electrostatic, and solvent interactions and geometries. Similarly, explicit solvent MD simulations by Hu *et al.* and Mu *et al.* show that the preference for the helical basin ranges from ~10-90% and that individual inter-basin hopping rates can vary up to 10-fold when computed from different FFs for the simple examples of alanine and glycine dipeptides (Hu et al., 2003) and tri-alanine (Mu et al., (in press)).

Because of the considerable influence a FF can exert on predicted inter-basin hopping frequencies, we test the reliability of our conclusions by performing independent calculations employing seven commonly used FFs, namely AMBER 94 (Pearlman, 1995), AMBER 96 (Kollman, 1997), Garcia's modified AMBER 96 (Garcia & Sanbonmatsu, 2002) (referred to as G-A-96 in this paper), CHARMM-27 (MacKerell et al., 1998), OPLS-united atom (Jorgensen, 1988), OPLS-AA-97, and the latest OPLS-AA-01 (Kaminski et al., 2001). The comparison of predictions obtained from the different FFs is also motivated by the knowledge that they have been optimized to reproduce thermodynamic data (and, in some cases, *ab initio* quantum calculations) and are generally validated by their ability to describe protein structures. Consequently, their suitability for dynamical calculations is unclear because the dynamics is sensitive to the heights of kinetic barriers, whereas thermodynamics and native structures are not.

Our Langevin dynamics (LD) simulations with a implicit solvent model (Shen & Freed, 2002b) produce nearly the same, strong FF dependence of basin populations and dynamics obtained from MD calculations with explicit solvent (Mu et al., (in press)).

Moreover, where the same FFs are used for explicit and implicit solvent simulations, good agreement is found, thereby supporting the validity of our computationally far less expensive approach.

In the present extensive study at 300 K, we examine the validity of the IPH using molecular mechanics potentials to construct and analyze the conformational and dynamical properties of peptides composed of many different amino acid combinations (60 different species in all). For all seven FFs considered, the time course of the LD trajectories reveals that a residue's

basin population and dynamics may be strongly influenced by the nearest neighbor amino acid's conformation and chemical identity. We calculate the backbone conformational entropy in the unfolded state for each residue according to its sampling of the phi-psi plot. This calculation is conducted separately by assuming that the samplings for each residue are independent (the IPH assumption) and by considering the correlated motions in order to quantify the error in IPH. We also discuss the implications of the different thermodynamics and dynamics produced by the various widely used FFs upon the ability of all-atom simulations to describe the free energies, folding pathways and time-scales in protein folding.

Results

The Ramachandran basin assignments are derived from the observed time course of the population distributions (Fig 5.2; see methods section). A common definition is suitable for all seven FFs. The most populated basins are the polyproline II (basin 1, B1), extended β (basin 2, B2), and α-helical (basin 3, B3) conformations (see Table 5.1). The

Table 4.1. Basin populations and configurational entropy for different force fields

Force Field	PP-II (%)	extended β (%)	α-helical (%)	$T\Delta S$ [1] (kcal mol^{-1} K^{-1})
AMBER 94	1.08 (13)	1.5 (3)	96.86 (80)	0
AMBER 96	14.15 (41)	76.27 (44)	5.02 (14)	-0.187
Garcia-A96	30.24	17.8	45.31	-0.358
CHARMM27	24.20 (55)[2]	18.33	47.62(45)	-0.365
OPLS-AA-97	82.97[3] (88)[2]		12.57 (12)	-0.355
OPLS-AA-01	31.02	41.17	20.75	-0.372
OPLS-UA	59.31[3]		33.93	-0.427

Values given in the table are for Ala^2 in Ala^1-Ala^2-Ala^3 at T = 300 K and values in parenthesis are from an explicit solvent MD calculation for trialanine (Mu et al., (in press))

[1] Calculated using Eq. 1 and referenced to value for AMBER 94.

[2] Combined values for PP-II and extended β.

[3] PP-II and extended β basins are not distinguished in this FF.

polyproline II (PP-II) and extended β basins are separated by a free energy barrier for all the FFs except the OPLS-UA and OPLS-AA-97 FFs where only a single basin is present in this region of the Ramachandran plot. The existence of a distinct PP-II basin is well established both experimentally (Shi et al., 2002a; Shi et al., 2002b; Woutersen et al., 2002) and references therein) and in MD simulations (Mu et al., (in press); Pappu & Rose, 2002)).

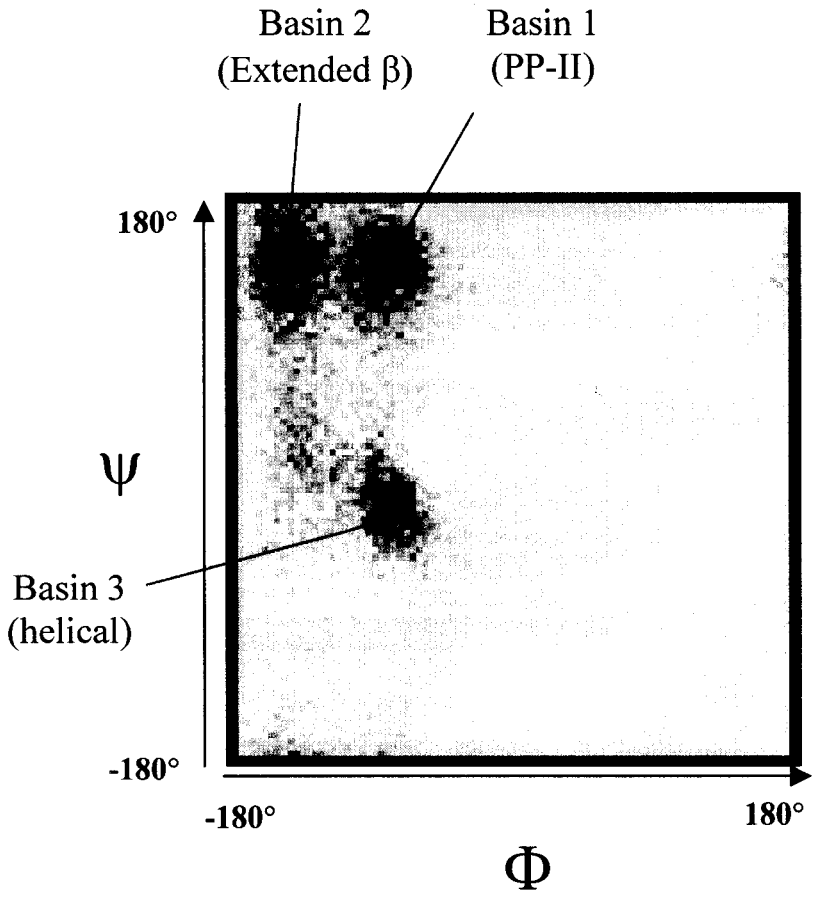
Figure 5.3 presents the time course of basin occupancies for the central Ala in the Ala-Ala-Ala peptide as calculated with the three different FFs using the color code at the top of the figure. The color variations between the trajectories from the different FFs strikingly expose the qualitatively different dynamics predicted by the various FFs. AMBER 94 predominantly populates the helical basin, whereas the distribution among the three dominant wells is more uniform for G-A-96 and OPLS-AA-01, though G-A-96 yields significantly more helical population that OPLS-AA-01.

Sequence dependence of NN effects

Underlying the IPH is the assumption of a lack of correlations between the (Φ,Ψ) dihedral angles of neighboring residues due to the rigidity of the peptide bond. Our first investigation focuses on the importance of the flanking moieties. A series of simulations is performed contrasting the behavior of a single alanine capped with acetyl and amide groups (Fig. 5.4) with that of an alanine flanked on both sides with alanines (Fig. 5.5). The presence of less bulky neighbors in the single alanine molecule increases the fraction of time the alanine spends in the extended β and PPII conformations (basins 1 and 2). For

Figure 5.2. Ramachandran plot of Ala^2 in Ala^1-Ala^2-Ala^3 .

Computed using the OPLS-AA-01 FF shows the presence of three distinct basins.



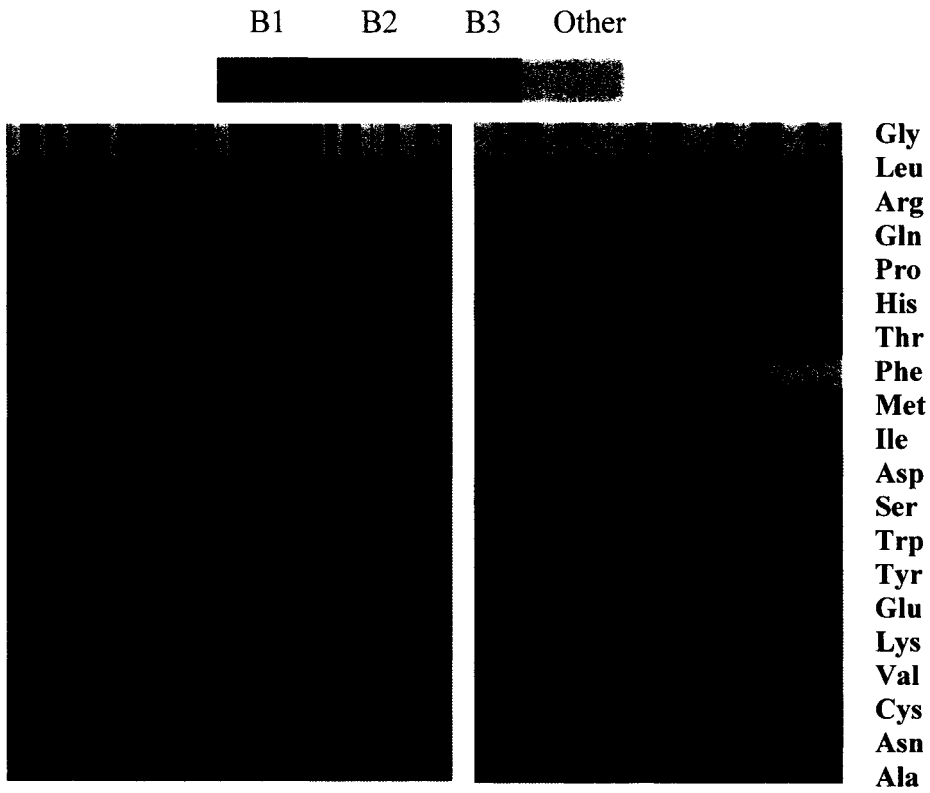
example, using the AMBER 94 FF, essentially the entire population is in the helical basin 3 for the (capped) tri-alanine molecule, whereas ~ 20% populates the other two basins in the (capped) mono-alanine molecule. These results are similar to those of Hu *et al.* who observe that mono-alanine populates the helical basin 84% of the time (Hu et al., 2003; Mu et al., (in press); Zaman et al., 2003a). The difference between mono- and tri-alanine already demonstrates that the rigidity of the peptide backbone does not prevent the neighbor moieties from influencing the backbone configuration even of a small amino acid such as alanine.

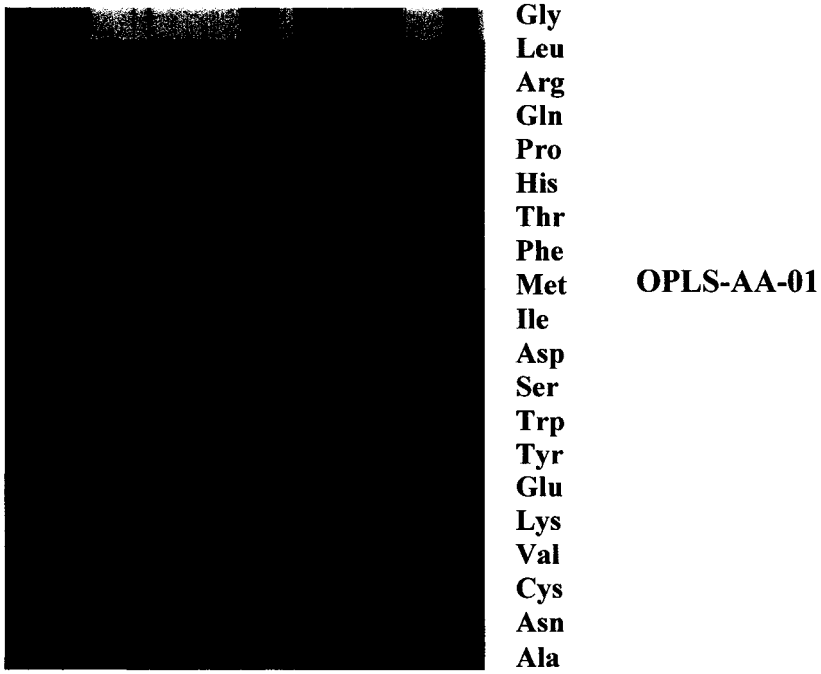
The implicit solvent LD simulations for the tri-alanine basin populations accord reasonably with the explicit solvent MD simulations for the same system by Mu *et al.* (Mu et al., (in press)).The LD simulations with the AMBER 94 FF differ for the α-helix population by ~15 %. With the AMBER-96 and CHARMM 27 FFs, the LD simulations under- and over-estimate the extended β population by 30 and 12% respectively, while very close agreement with the MD simulations is found for the OPLS-AA-97 FF (within 6%). This general agreement provides additional strong justification for use of the implicit solvent model in the more extensive study of NN neighbor effects that follow.

The three FFs (AMBER 94, OPLS-AA-01 and G-A-96) generate different basin preferences for the tri-alanine molecule (Fig. 5.3 and first row in Table 5.2). The AMBER 94 FF predicts a predominant helix basin population, while the G-A-96 and OPLS-01 yield helix, extended, and PPII basin populations in the ratios roughly of 3:2:1 and 2:4:3, respectively. Table 5.2 also illustrates the NN effect on the central Ala residue in the peptides Ala-Ala-X and X-Ala-Ala for seven different residues X (of varying character), while Table 5.2 and Figs. 5.6 and 5.7 display the NN effect for X in the seven

Figure 5.3. Backbone dynamics of different center residues in Ala-X-Ala.

The 15 ns time course is presented for the basin populations, colored according to the legend given at the top of the figure. Simulations for three representative FFs are provided to demonstrate the wide variation in populations between the FFs. Residues spend considerably more time in basin 3 (helical) when the AMBER 94 FF is used compared to the G-A-96 and OPLS-AA-01 FFs, where there is higher probability for extended β structures (basin 2) and PP-II (basin 1).


B1
B2
B3
Other
Gly
Leu
Arg
Gln
Pro
His
Thr
Phe
Met
Ile
Asp
Ser
Trp
Tyr
Glu
Lys
Val
Cys
Asn
Ala
AMBER 94
G-A-96


Gly
Leu
Arg
Gln
Pro
His
Thr
Phe
Met
Ile
Asp
Ser
Trp
Tyr
Glu
Lys
Val
Cys
Asn
Ala
OPLS-AA-01

pairs of di-peptides Ala-X and X-Ala. The G-A-96 and OPLS-AA-01 FFs produce an appreciable NN effect, with the alanine basin populations sometimes changing by a factor of three as the neighboring side-chains are varied. For example, the helix basin population from the G-A-96 FF ranges from the low of 16.4% when the C-terminal NN is X^3 = Gly to a high of 48.4% for X^3=Trp. For N-terminal NNs, the center alanine's helix populations are 40.5% with X^1 = Gly and 4.5% for X^1 = Trp. Similarly, large NN effects are evident for the G-A-96 FF in Table 5.2.

The AMBER 94 FF only yields a marginal NN effect and only in the di-peptides (Fig 5.8). This difference arises because the AMBER 94 FF predicts that the alanine backbone almost always remains in the helical basin, regardless of the NN, whereas the helical basin population varies between 5 and 75% for the other two FFs. Hence, much of our analysis focuses on the two more realistic FFs, G-A-96 and OPLS-AA-01.

The NN effects computed for the di-amino acids are of similar magnitude to those obtained for the tri-amino acids (data not shown), which confirms that the observations concerning NN effects are not artifacts of longer range I-1,I+1 side-chain interactions.

Backbone entropy

The influence of nearest neighboring residues can be quantified in terms of the change in an alanine's backbone entropy due to presence of different neighbors. Using the basin populations on the phi-psi map, we calculate the backbone conformational entropy according to the relation (see Methods),

$$S = -R \sum_{i=1}^{120} \sum_{j=1}^{120} P_{ij} \ln P_{ij} \qquad (5.1)$$

where P_{ij} is the normalized probability of being in the i,j^{th} 3°x3° mesh element in the phi-psi map, and R is the gas constant. Although this calculation of S depends on the mesh-size (i.e., the volume per configuration in phase space), relative entropy differences between residues, or between those calculated with different FFs, do not. The difference in basin populations for the different FFs is manifest in residue dependent backbone entropies (Table 5.1 and Fig. 5.9). For example, the entropy is the lowest with the AMBER 94 FF where essentially all the population is in the helical basin. For the center alanine in a tri-alanine peptide, the backbone entropy $T\Delta S$ calculated using AMBER 96 is larger than $T\Delta S$ calculated using AMBER 94 by 0.18 kcal mol^{-1}, which reflects the binary basin occupancy between the extended and PP-II basins for AMBER 94. Because the simulations with the other five FFs yield a more uniform distribution of these three basins, the backbone entropy of the center residue in tri-alanine for most FFs exceeds the AMBER 94 entropy by $T\Delta S \sim 0.4$ kcal mol^{-1}.

The change in an alanine's backbone entropy with different neighbors is on the same order of magnitude as the difference in backbone entropy between different residues (Table 5.4). On average, the change in backbone entropy of Ala with different neighbors is $T\Delta S \sim 0.1$ kcal mol^{-1}, which is approximately the average difference between entropy of individual residues. This difference in entropy between individual residues is illustrated in Fig. 5.7 where the backbone entropy is presented for each of the three residues in Ala-X-Ala where X ranges over the 20 naturally occurring amino acids. Although the backbone entropies for the G-A96 and OPLS-AA-01 FFs often differ for individual amino acids, values for the flexible glycine and the highly restricted proline lie

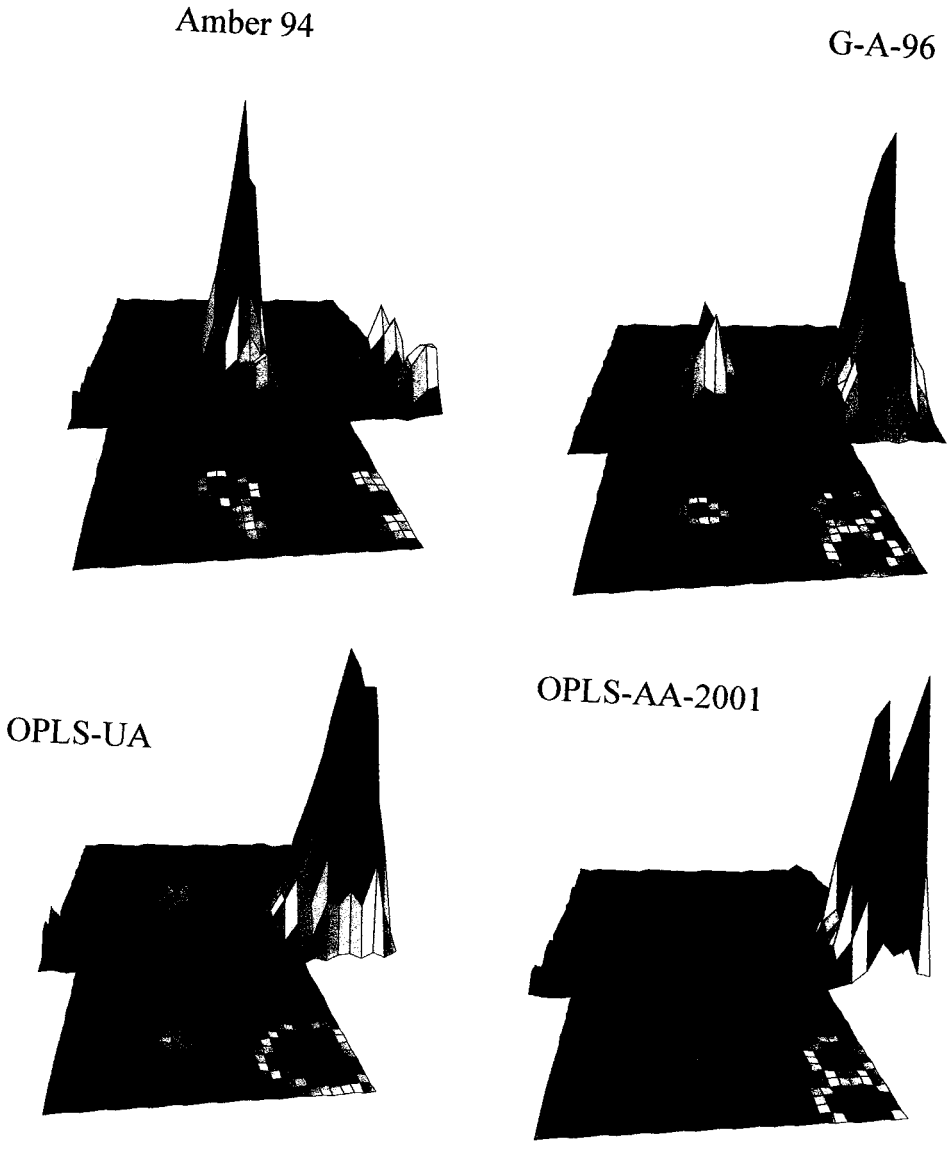


Figure 5.4. Phi-Psi basin populations for Ace-Ala-Nme for different force fields.

Populations are obtained from 45 ns LD trajectories.

Table 5.1. Alanine conformational preference as a function of its NN chemical identity

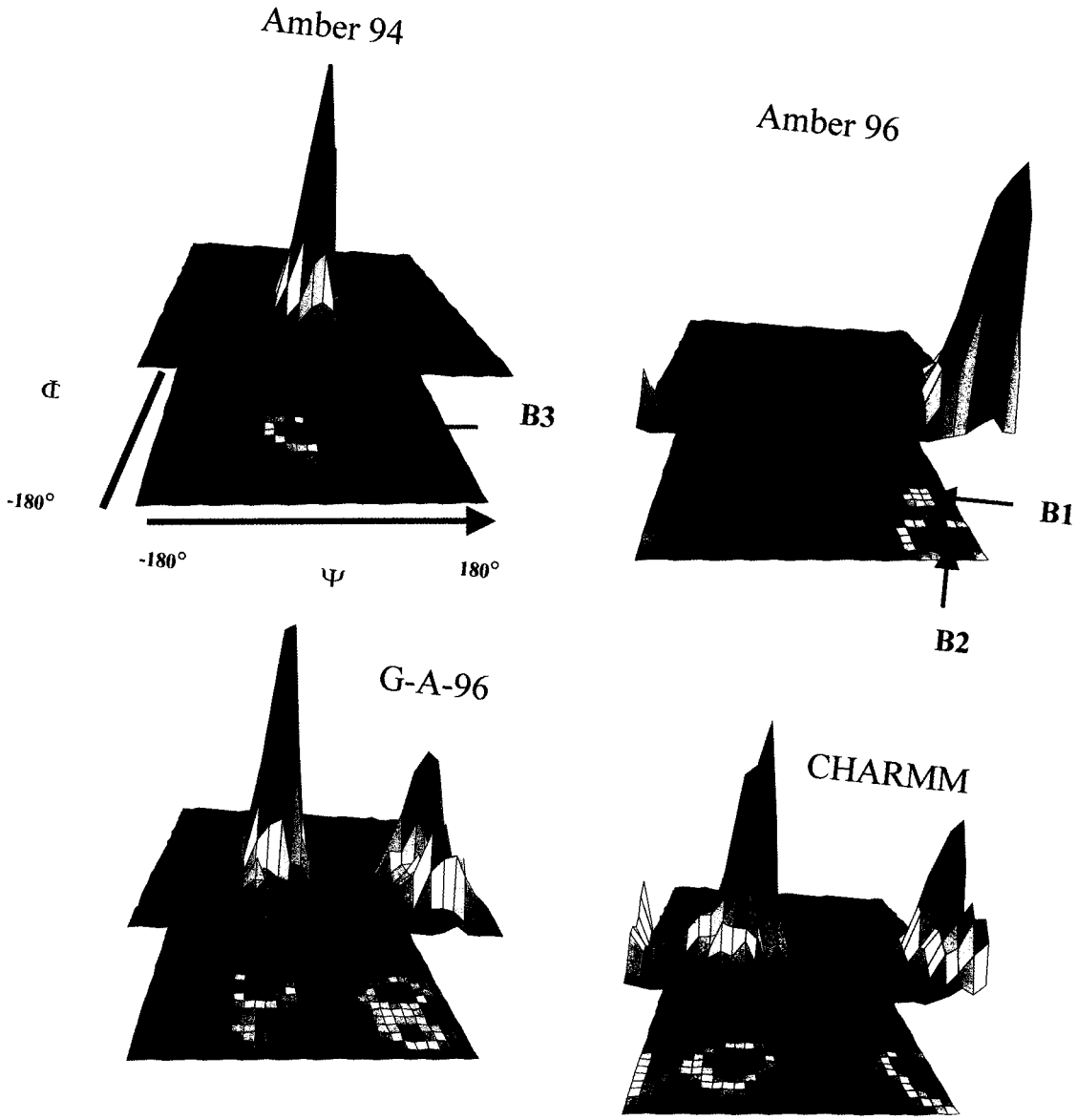
X	Ala-Ala-X		
	Basin 1	**Basin 2**	**Basin 3**
Ala	1.1/30.02/31.02	1.5 /17.9/41.02	96.8/45.75/20.2
Trp	6.6/27.25/19.4	11.6 /17.83/31.2	80.26/48.41/43.6
Met	1.0/34.08/24.5	1.42 /20.22/23.25	97.13/39.91/45.0
Asp	0.4/23.25/12.2	0.4 /45.08/40.83	99.0/25.75/44.12
Asn	3.87/39.8/10.0	7.9 /22/14.5	87.0/32.58/17.81
Leu	0.6/51.51/23.65	0.1 /25.7/21.81	99.4/14.33/46.33
Gly	6.8/44.75/35.12	12.33/28.25/41.91	79.13/16.41/15.41

X	X-Ala-Ala		
	Basin 1	**Basin 2**	**Basin 3**
Ala	1.1/30.02/31.02	1.5 /17.9/41.02	96.8/45.75/20.2
Trp	2.8/48.0/21.66	9.06/41.16/50.41	87.66/4.50/24.33
Met	2.0/19.33/23.16	4.2/13.58/27.13	93.26/63.68/41.46
Asp	0.2/26.25/28.6	1.04/18.25/23.16	98.53/48.41/41.7
Asn	3.16/31.41/36.16	5.00/22.5/35.17	91.20/41.16/20.53
Leu	3.34/22.3/27.83	4.51/28.3/34.25	92.34/39.1/28.50
Gly	4.80/32.0/38.68	4.00/21.75/36.58	90.2/40.48/18.25

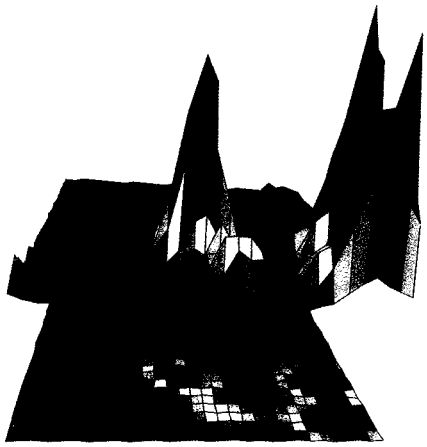
The table gives values for the influence on the center alanine's basin populations. Values are given for the AMBER 94/G-A96/OPLS-AA-01 FFs, respectively.

Figure 5.5. Basin Populations of Tri-peptides.

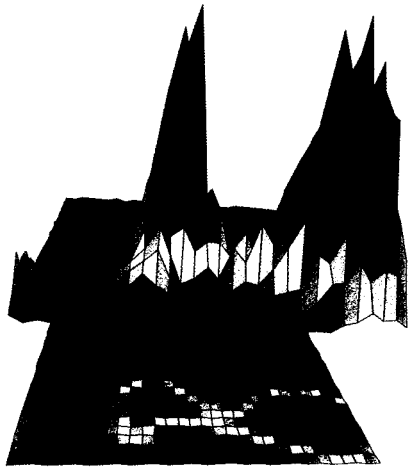
Basin populations for Ala^2 for the seven different FFs, calculated from averages along the time trajectories such as those illustrated in Fig. 1B. The most populated basins are PP-II (basin 1), extended β (basin 2) and α-helical (basin 3).


Amber 94
Φ
-180°
B3
-180°
Ψ
180°
Amber 96
B1
B2
G-A-96
CHARMM

OPLS-UA



OPLS-AA-97



OPLS-AA-01

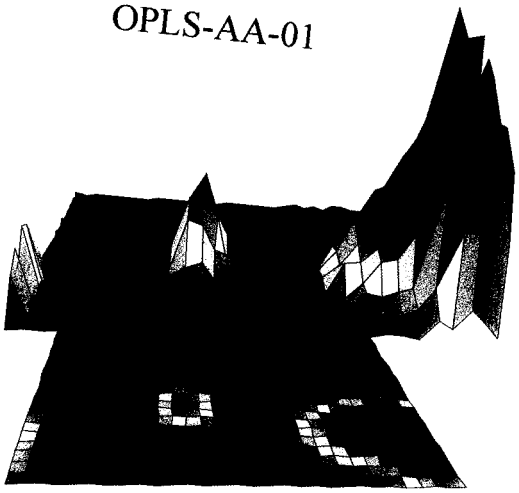


Fig 5.5 (contd.)

near the extrema in both FFs. The calculations also reproduce the known feature that residues preceding *trans*-prolines are conformationally restricted. This effect is illustrated in Fig 5.9 where both G-A-96 and OPLS-AA-01 depict alow entropy for Ala^1 when it precedes proline.

Geometric Dependence of NN effect:

In addition to being sensitive to its NN side-chain *identity*, a residue's conformation is influenced by its NN's *conformation*. The helical basin population of residue X in Ala-X-Ala often changes by two-fold or more when both flanking alanines are in the helical basin. Figure 5.10 illustrates the influence of the NN conformation by presenting the difference in the backbone entropy ($S_{NN\ free} - S_{NN\ constrained}$) = $T\Delta S$ for each of the 20 amino acids as computed when both the flanking alanines are free to occupy all basins according to the equilibrium populations and when they are constrained to be in the helical basin. This entropy difference nearly vanishes for 6-8 of the residues, depending upon the force field. However, $T\Delta S$ lies in the range of -0.5 to 0.12 kcal mol^{-1} for the majority of residues using either the G-A-96 and OPLS-AA-01 FFs. Thus, a residue's configuration can be significantly affected by its NN conformations.

Because a residue's entropy depends upon its neighbors' conformation, the backbone entropy of the system is not the sum of the individual residues' entropies. To estimate the magnitude of the non-additivity, the entropy of pairs of residues in a tri-amino acid molecule are calculated from the location of the pair's configuration in a four dimensional phi-psi plot ($(\Phi,\Psi)_{i=1,2}$). This behavior is illustrated for the peptides AAA, LLL, VVV and a pseudo-random sequence Ala-Glu-Thr-Asn. The difference in the

Figure 5.6. Sequence dependence of Nearest Neighbor effects (AAX).

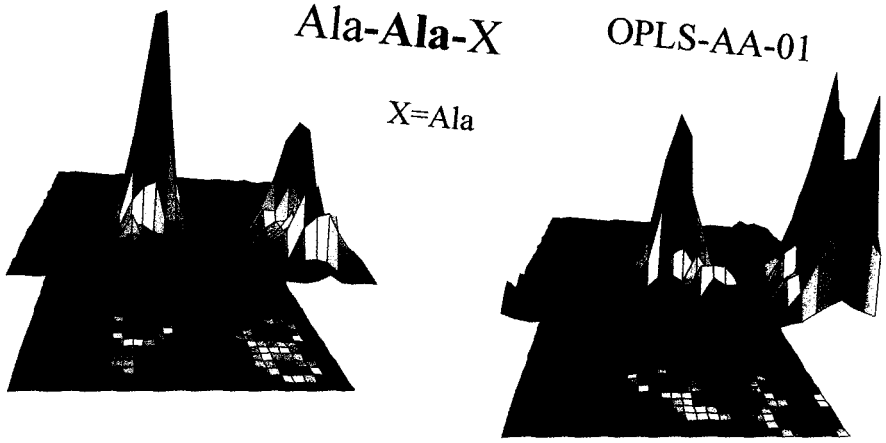
The population distribution for the center Ala is presented for Ala-Ala-X. for the G-A-96 and OPLS-AA-01 FFs for X=[ala, gly, leu, trp, met, asn, asp]. The fractional population of the Ala varies significantly with the neighboring residue type and also whether the other Ala is N- or C-terminal to the neighbor.

G-A-96

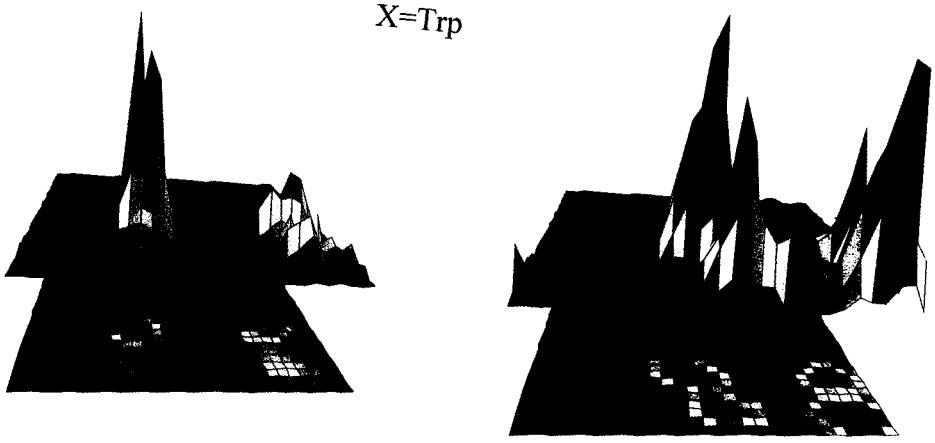
Ala-**Ala**-*X*

OPLS-AA-01

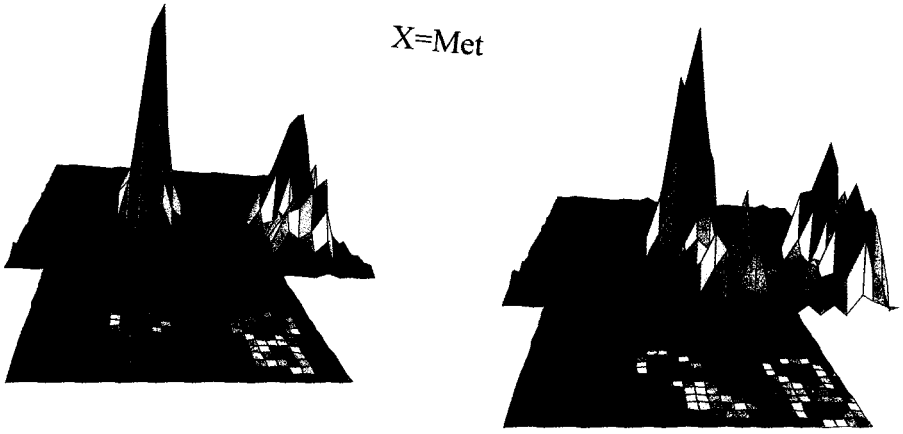
X=Ala



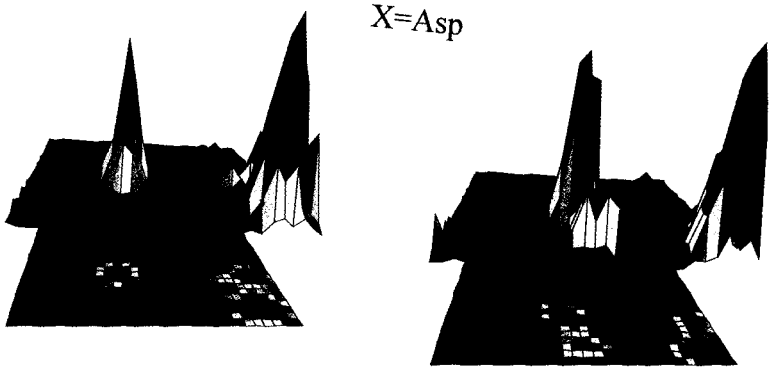
X=Trp



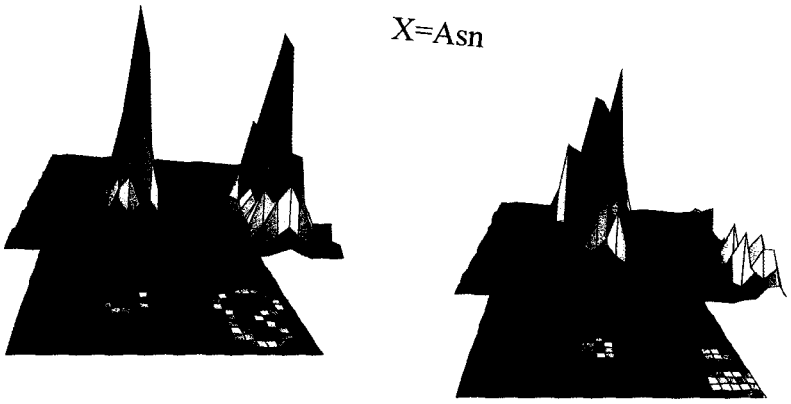
X=Met



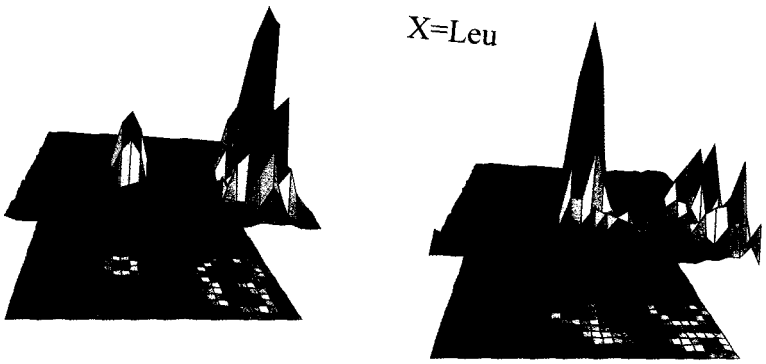
X=Asp



X=Asn



X=Leu



X=Gly

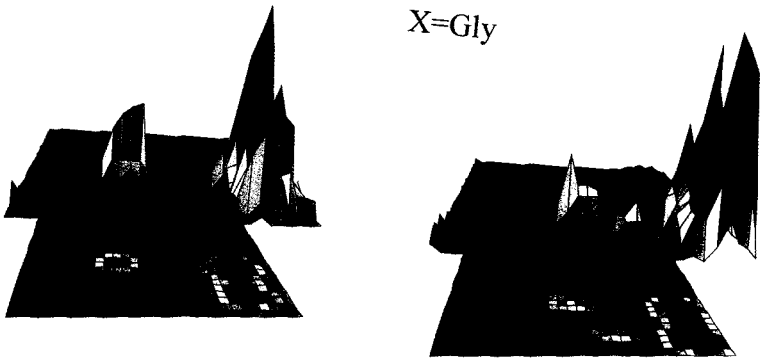


Fig. 5.6 (contd).

Figure 5.7. Sequence dependence of Nearest Neighbor effects (XAA).

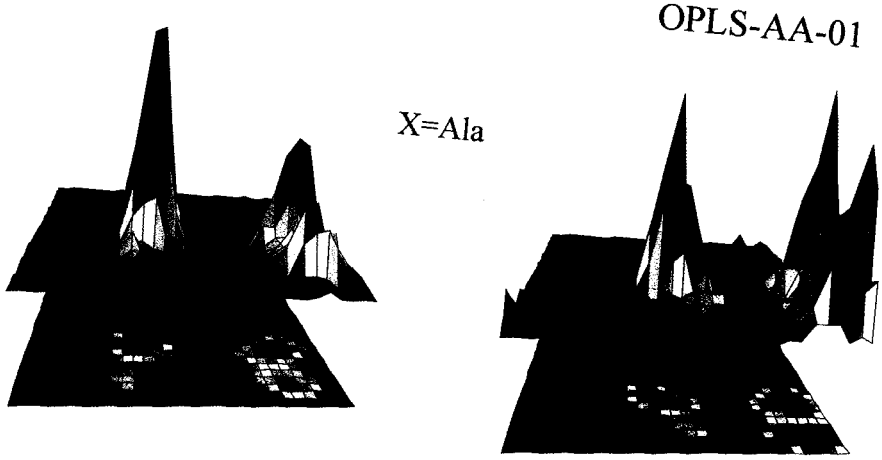
The population distribution for the center Ala is presented for X-Ala-Ala for the G-A-96 and OPLS-AA-01 FFs for X=[ala, gly, leu, trp, met, asn, asp]. The fractional population of the Ala varies significantly with the neighboring residue type and also whether the other Ala is N- or C-terminal to the neighbor.

X-**Ala**-Ala

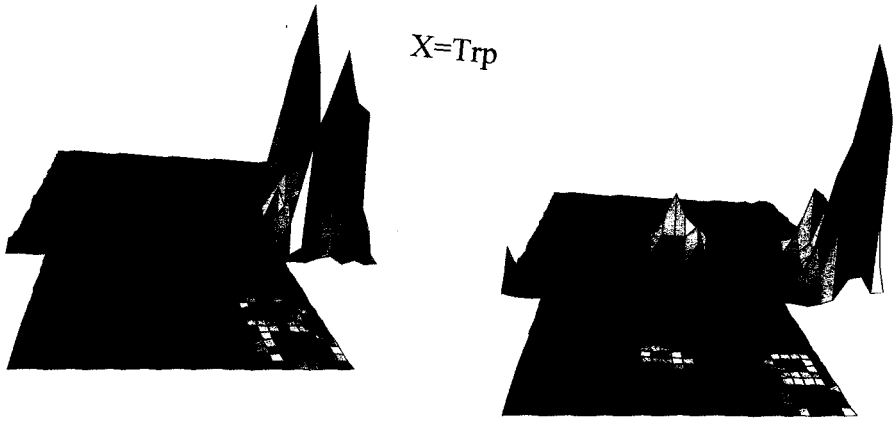
G-A-96

OPLS-AA-01

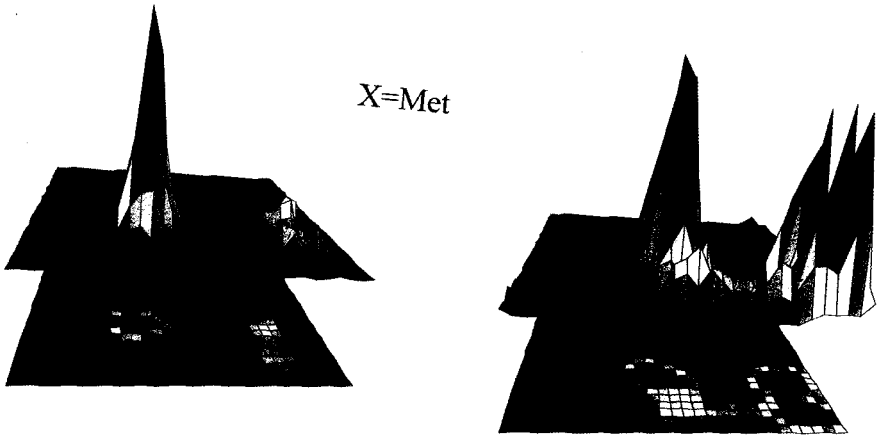
X=Ala



X=Trp



X=Met



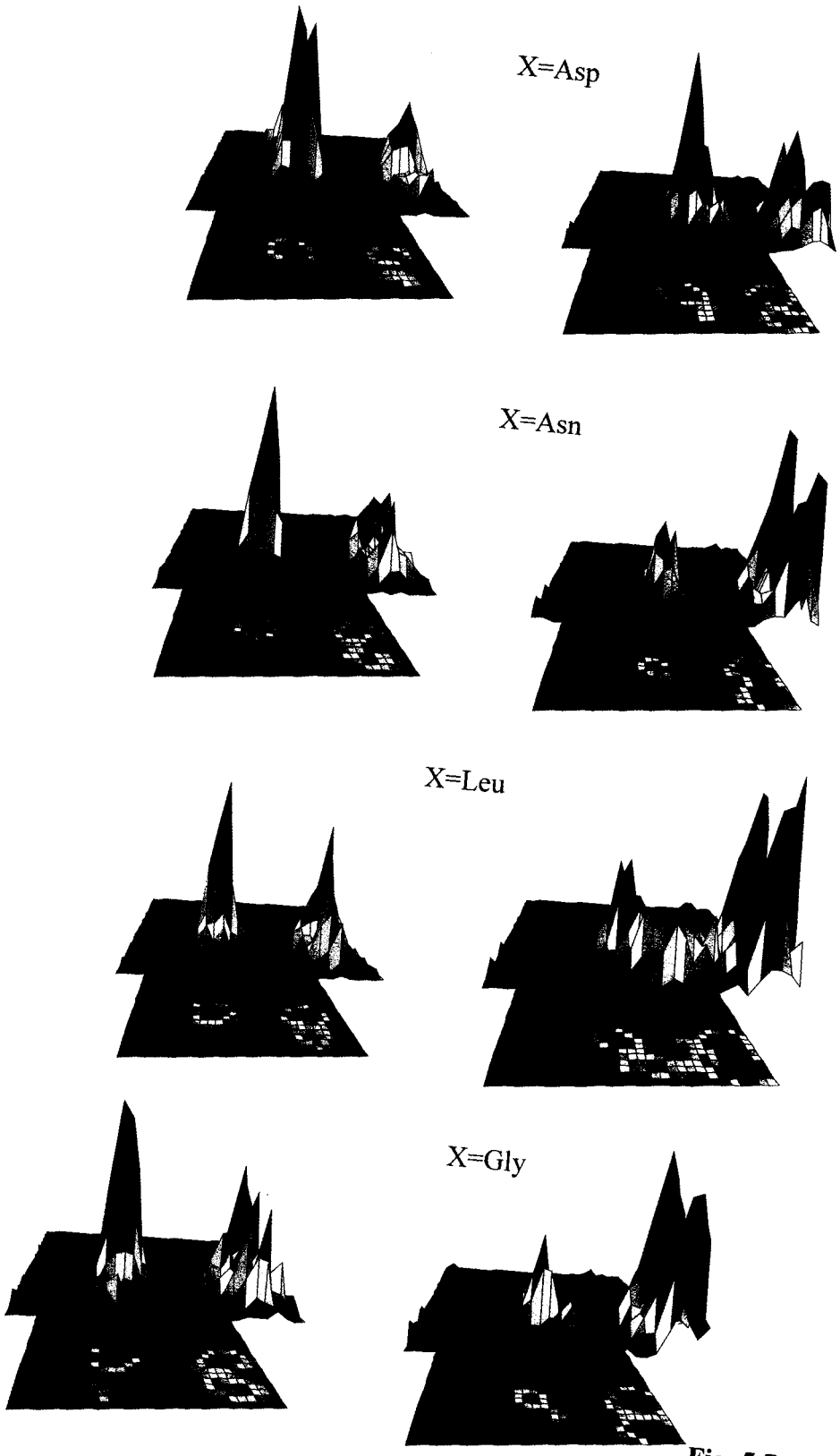


Fig. 5.7 (contd).

correlated entropy and the sum of the entropies of the individual residues, calculated assuming that they are independent of their NNs' conformation, is in the range of $T\Delta S \sim$ 0.3-0.7 kcal mol^{-1} $residue^{-1}$ depending upon the FF employed (Table 5.5). This range of non-additive contributions is about half the estimated loss of backbone entropy per residue upon unfolding based on experimental data ((D'Aquino et al., 1996; Thompson et al., 2002) and references therein). Hence, the non-additive correction is quite significant, and the IPH is inadequate to describe the backbone entropy of short peptides. Therefore, an accurate calculation of the unfolded state entropy must include correlations due to the neighboring residues.

Backbone dynamics:

The rates of transitions between basins (or basin escape rates) are studied for each of the seven force fields using the basin auto-correlation function,

$$C_i(t) = < P_i(t) \bullet P_i(0) > \qquad (5.2)$$

where $P_i(t)$ is the probability of being in the i^{th} basin at time t. $P_i(t)$ is defined as unity if the residue is in basin *i* at time t and is zero otherwise. The long time limit of the correlation function $C_i(t)$ approaches a constant that equals the equilibrium population of basin *i* for the FF. The correlation functions for the helical basin 3 are nearly exponential for the different FFs (Fig. 5.11a), a behavior consistent with first order kinetics for the escape from the basins. Poor fits to an exponential arise for transitions out of basins with very low populations because of meager statistics in these cases. This trend of exponential decay kinetics is also observed in the basin escape rates for basins 1 and 2 (data not shown). Inter-basin transition rates k_{ij} are obtained from fitting the correlation functions with an exponential decay towards the constant long time limit as described in

Table 5.2. Influence of NN sequence on Alanine's basin population fractions

X in Ala-X	AMBER 94 (kcal mol^{-1})			G-A-96 (kcal mol^{-1})			OPLSAA-2001 (kcal mol^{-1})		
	B1	B2	B3	B1	B2	B3	B1	B2	B3
Ala-Ala	0	0	0	0	0	0	0	0	0
Ala-Trp	0.32187	0.38131	-0.08142	0.2567	0.57102	-0.95704	0.28123	0.16389	-0.46162
Ala-Met	0.14866	0.38131	-0.02083	-0.07634	0.36265	-0.41424	0.14119	0.34036	-0.48059
Ala-Asp	-0.32541	-0.80858	0.49557	0.36313	-0.36086	1.11653	0.55953	0.00249	-0.46874
Ala-Asn	-0.12658	0.12945	-0.00733	-0.01569	0.32163	-0.52729	0.67884	0.62365	0.07555
Ala-Leu	0.05046	-0.02	0.00648	-0.16657	0.35422	-0.22375	0.16237	0.37872	-0.49806
Ala-Gly	-0.27493	0.1358	0.00382	-0.07877	0.12601	-0.13025	-0.07487	-0.01317	0.1624
Ala-**Ala**	0	0	0	0	0	0	0	0	0
Trp-**Ala**	-0.10725	-0.24666	0.05502	0.15756	-0.00262	-0.41616	0.21511	-0.12397	-0.11162
Met-**Ala**	0.10302	0.25847	-0.03675	-0.05301	0.11203	-0.09476	0.17494	0.24776	-0.43143
Asp-**Ala**	0.57189	0.16253	-0.05805	0.13285	0.08293	-0.48917	0.04835	0.34269	-0.43489
Asn-**Ala**	0.05046	-0.0249	5.432 E-4	0.08533	0.04877	-0.34178	-0.09238	0.09203	-0.00972
Leu-**Ala**	0.32554	0.24839	-0.06701	0.01402	0.00481	-0.14293	0.06472	0.10793	-0.20653
Gly-**Ala**	-0.82245	-0.23302	0.11512	0.0705	0.20779	-0.63661	-0.1328	0.06844	0.06091

For the alanine in bold face, the values are calculated according to –RT [ln (fractional population in Basin 1 for alanine when NN is residue X) – ln (fractional population in basin y for alanine when NN is alanine)]. Only the fractional populations in the Basins 1, 2 and 3 are presented. The NN conformation is unconstrained. The negative values indicate that the NN effect due to residue X on Ala is less than the NN effect of Ala on Ala.

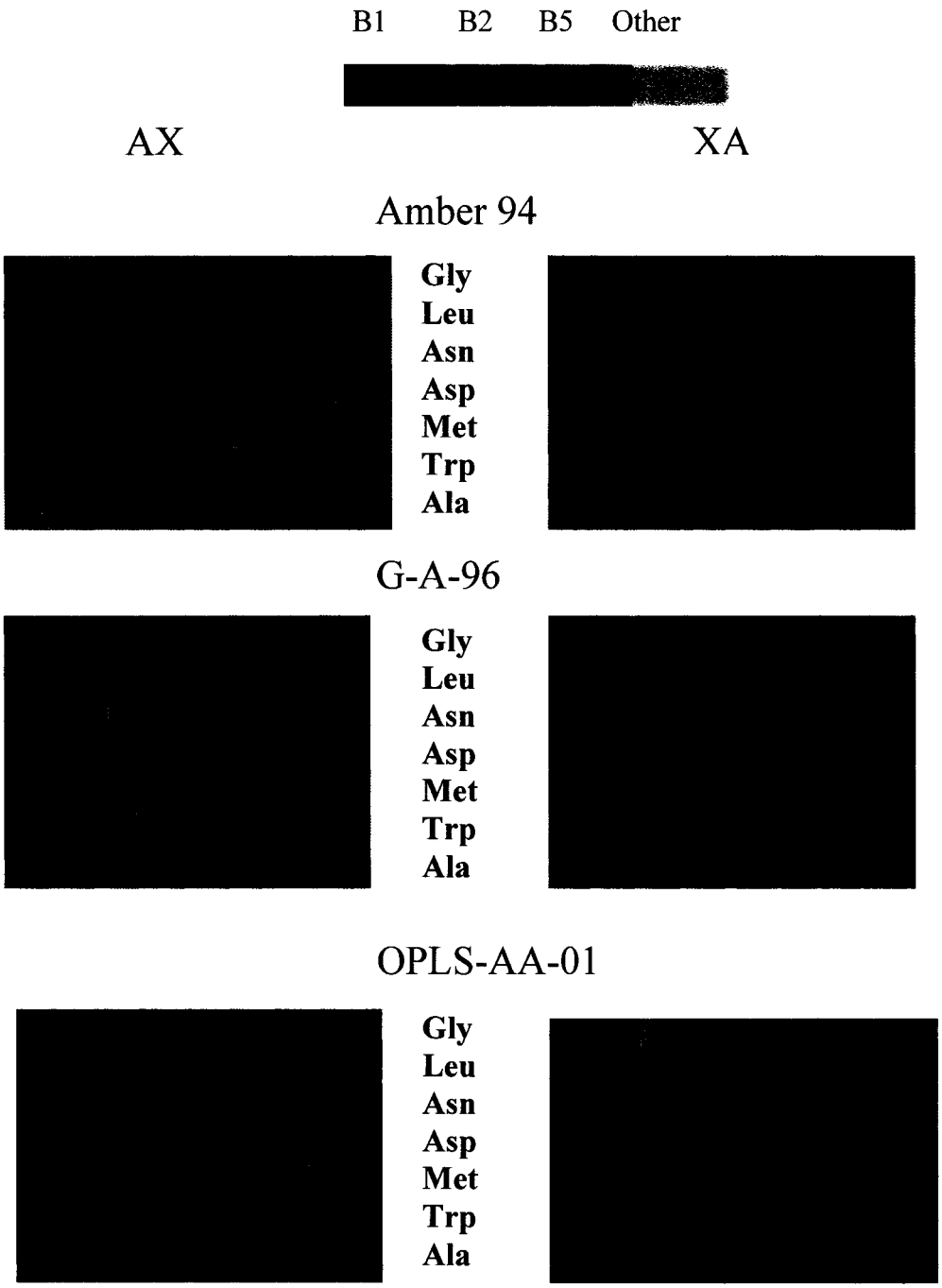


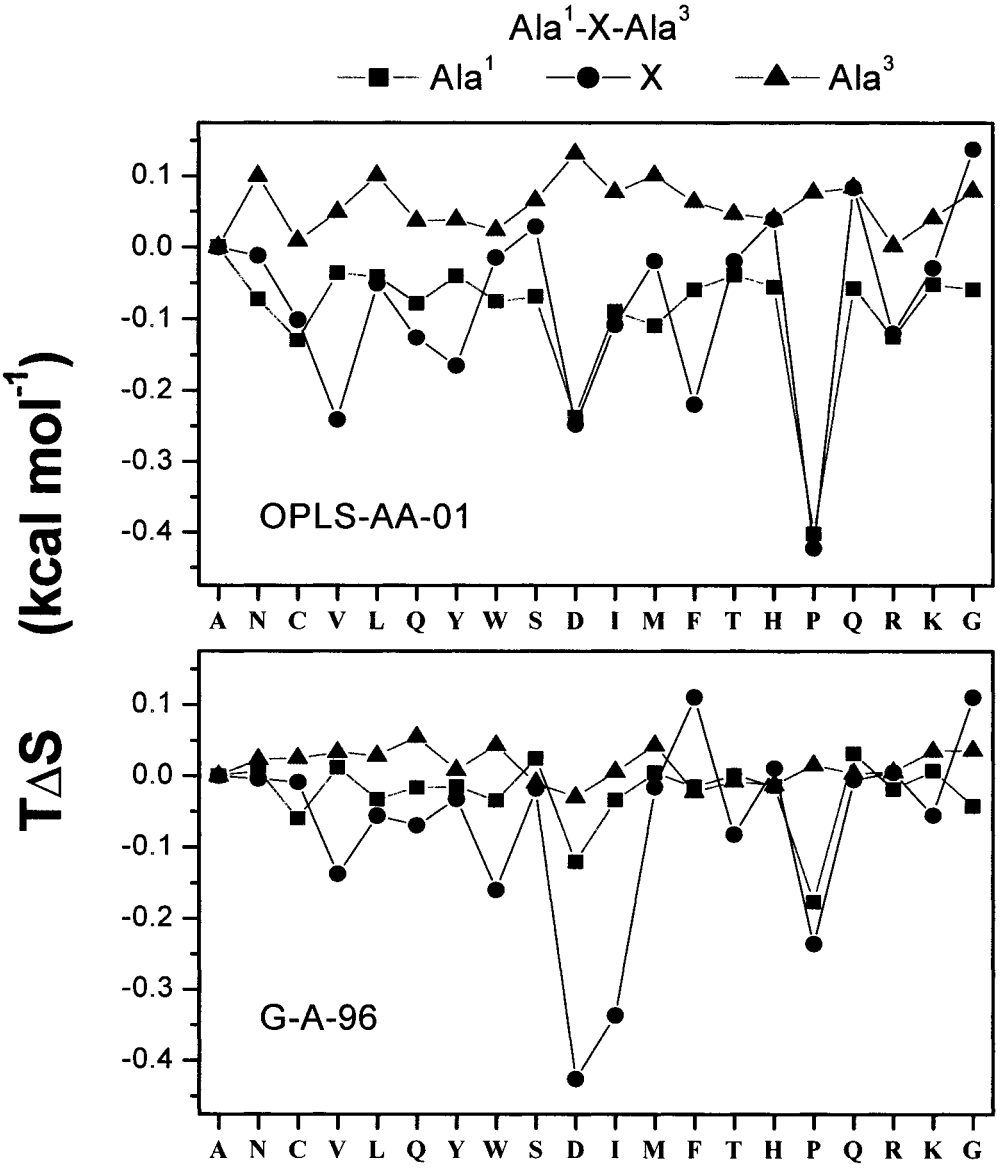
Figure 5.8. Backbone dynamics in di-peptides.

The left panel shows the backbone dynamics of Ala in Ala-X and while the right column shows the backbone dynamics of Ala in X-Ala. The 15ns time trajectory is presented for basin occupancies for three FFs, namely AMBER 94, G-A-96 and OPLS-AA-01.

Methods. While the AMBER 96 and OPLS-AA-97 FF produces the fastest rates due to their negligible populations in the helical basin. The AMBER 96 and G-A-96 rates differ by a factor of five, which arises solely from the flattening of the added torsional potential for the G-A-96 FF (Fig. 5.12). The correlation functions for the other basins exhibit a very similar dispersion in rates, as do those for the alanine in an Ala-Ala di-peptide. A similar dispersion in rates appears in the explicit solvent calculations of Mu *et al* (Mu et al., (in press)) for the tri-alanine peptide, where the authors suggest that the hopping rates vary by almost an order of magnitude for different FFs. An interesting aspect of the dynamics is the directional sampling of the basins in the phi-psi map i.e., the existence of preferential transitions between certain basins. An analysis of the inter-conversions among the three major basins indicates that transitions are predominantly between basin 2 and either basin 1 or basin 3 (Fig. 5.11 b) The time constant for escape from the helical basin of an Ala residue exhibits an eight-fold dispersion as the FF is varied (Table 5.1 and Fig. 5.11). As expected, the AMBER 94 FF yields the slowest rate due to its overwhelming population in basin 3, but not between basins 1 and 3. This behavior is common for all FFs (except the OPLS-UA and OPLS-AA-97 FF where basins 1 and 2 coalesce into a single basin), indicating that directional basin sampling is general. The origin of the directional sampling can be viewed, for example, as the requirement that the left-handed PP-II conformation (basin 1) tends first to untwist (basin 2) before it can re-twist into the right-handed α-helical conformation (basin 3). The basin hopping rates also depend on the NN identity. The hopping rate of Ala^2 in AAX and XAA changes by

Figure 5.9. Backbone entropy and sequence dependence of NN effect.

Entropy for each residue in Ala^1-X-Ala^3, referenced to that of a tri-alanine peptide as calculated for the OPLS-AA-01 (top plot) and G-A-96 (bottom plot) FFs. The value depicted for residue X represents the variation in backbone entropy with amino acid type. Changes in the entropy of Ala^1 or Ala^2 reflects their dependence on residue X, while their difference is due to being N- or C-terminal to the center residue, as well as being at either end. The abscissa is the one-letter code for the amino acids.


Ala1-X-Ala3
Ala1
X
Ala3
TΔS (kcal mol-1)
OPLS-AA-01
G-A-96
0.1
0.0
-0.1
-0.2
-0.3
-0.4
A N C V L Q Y W S D I M F T H P Q R K G

almost 50% between X=Ala and X=Gly. Similarly, X=Asn and X=Ala also display a difference of about 50% in hopping rates.

Discussion

We have investigated the backbone dynamics of different peptides using Langevin dynamics simulations with a validated implicit solvent model and employing a variety of commonly used FFs. A residue's conformation, as well as its location in the peptide sequence, can significantly affect its neighbor's Ramachandran basin populations and basin inter-conversion rates (except with the AMBER 94 FF). For example, when the two flanking residues in a tri-amino acid are restricted to the helical basin, the residue's backbone entropy may change by the same order of magnitude as the difference in backbone entropy between different amino acids. These results are similar to the ones reported by Pappu *et al.*, though quantitative differences exist due to their use of hard-sphere potentials. The influence of either neighboring residues' identity on the backbone entropy of a residue is of the same magnitude. Additionally, the identity of the NN can alter the rate at which an alanine leaves, for example, the helical basin by nearly 50%.

Decrease in Backbone entropy due to correlated motions.

The influence of the NN's conformation on the torsional populations and kinetics of a residue demonstrates the invalidity of the Flory IPH. A similar conclusion is reached by Pappu *et al.*, who also observe a reduction in available conformations for the terminal alanine of a helical segment. We quantify the extent to which backbone conformations are coupled by calculating the difference between the sum of the independent entropies of each residue for a bonded pair of amino acids and that for the correlated pair. This

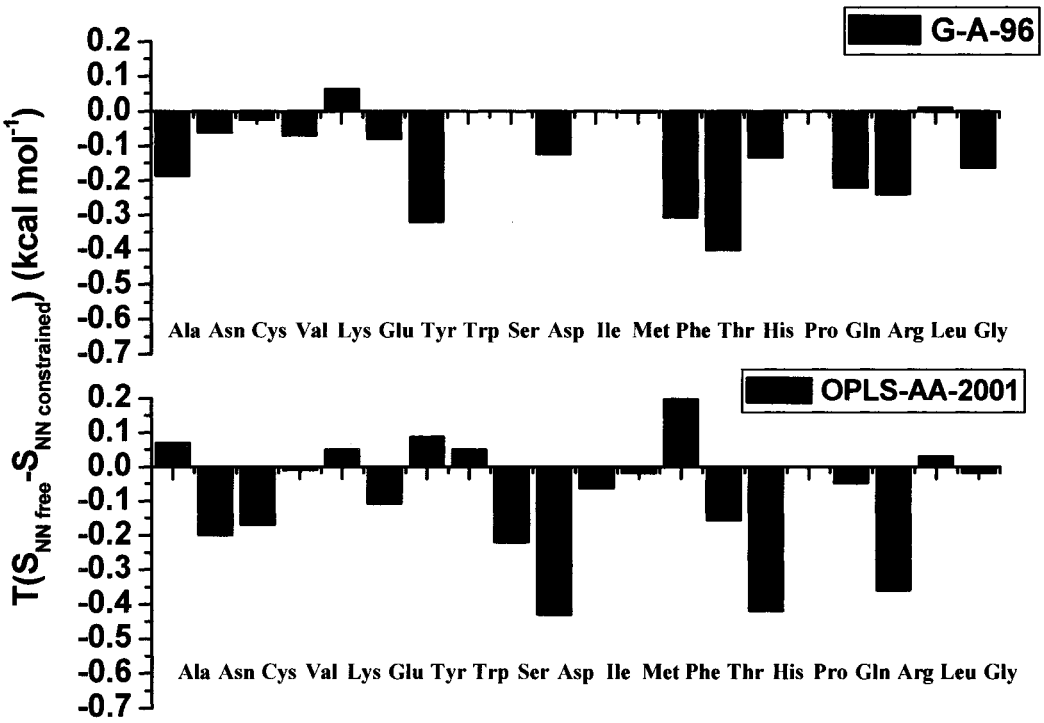


Figure 5.10. Backbone entropy and conformational dependence of NN effect.

The difference in the backbone entropy for residue X in the tripeptide Ala-X-Ala when the flanking alanines are free to be in any basin versus when the flanking residues are in the helical basin (B3).

Table 5.3. Reduction in backbone entropy due to NN correlations

	$TS_{1,2}$	$T(S_1+S_2)$	$T\Delta S_{1+2}$	$TS_{2,3}$	$T(S_2+S_3)$	$T\Delta S_{2+3}$
G-A-96	-3.66	-2.40	1.26	-3.69	-2.57	1.12
CHARMM	-3.63	-2.30	1.33	-3.77	-2.68	1.09
OPLS-UA	-3.61	-2.18	1.42	-3.63	-2.34	1.29
OPLS-AA-01	-3.66	-2.47	1.19	-3.66	-2.45	1.21

Ala-Ala-Ala

	$TS_{1,2}$	$T(S_1+S_2)$	$T\Delta S_{1+2}$	$TS_{2,3}$	$T(S_2+S_3)$	$T\Delta S_{2+3}$
G-A-96	-3.77	-2.88	0.88	-3.85	-3.16	0.69
CHARMM	-3.83	-3.13	0.70	-4.04	-3.41	0.63
OPLS-UA	-4.01	-3.49	0.52	-3.88	-3.16	0.72
OPLS-AA-01	-4.00	-3.36	0.64	-4.01	-3.51	0.50

Val-Val-Val

	$TS_{1,2}$	$T(S_1+S_2)$	$T\Delta S_{1+2}$	$TS_{2,3}$	$T(S_2+S_3)$	$T\Delta S_{2+3}$
G-A-96	-3.91	-3.34	0.55	-3.76	-2.97	0.78
CHARMM	-3.84	-3.02	0.82	-4.21	-3.58	0.63
OPLS-UA	-3.69	-2.57	1.12	-3.71	-2.68	1.02
OPLS-AA-01	-3.78	-2.95	0.83	-3.79	-2.96	0.92

Leu-Leu-Leu

Values in the table are given in kcal mol^{-1}. Reduction in backbone entropy due to NN correlations is obtained according to ΔS_{i+j} = (Entropy of $Residue^i$ + Entropy of $Residue^{i+1}$) – (The entropy of the system composed of $Residue^i$ and $Residue^{i+1}$, calculated using 4-dimensional phi-psi map) resolved in 10°x 10 ° grid elements. As with all calculations of entropy, the value for S depends on the mesh-size, and numbers listed are relative (see text). However, entropy differences (ΔS) do not depend on mesh size, and
are in absolute terms.

Ala-Glu-Thr-Asn

	$TS_{1,2}$	$T(S_1+S_2)$	$T\Delta S_{1+2}$	$TS_{2,3}$	$T(S_2+S_3)$	$T\Delta S_{2+3}$	$TS_{3,4}$	$T(S_3+S_4)$	$T\Delta S_{3+4}$
G-A-96	-3.68	-2.42	1.25	-3.77	-2.64	1.13	-3.74	-2.68	1.05
CHARMM	-3.68	-2.59	1.08	-3.86	-3.13	0.73	-3.93	-3.36	0.57
OPLS-UA	-3.62	-2.26	1.35	-3.68	-2.41	1.27	-3.68	-2.61	1.07
OPLS-AA-01	-3.78	-3.01	0.77	-4.04	-3.54	0.50	-4.13	-3.63	0.50

Table 5.4 (contd.)

difference is sizable, $T\Delta S \sim 0.3\text{-}0.7$ kcal mol^{-1} $residue^{-1}$ (Table 5.4). Fortunately, most experiments measure the entropy of the system as a whole and therefore automatically include all contributions from the correlations. Future studies will investigate the magnitude of the neighbor effects in an entire protein sequence and whether the entropy of the system is strongly dependent upon sequence order, rather than just composition.

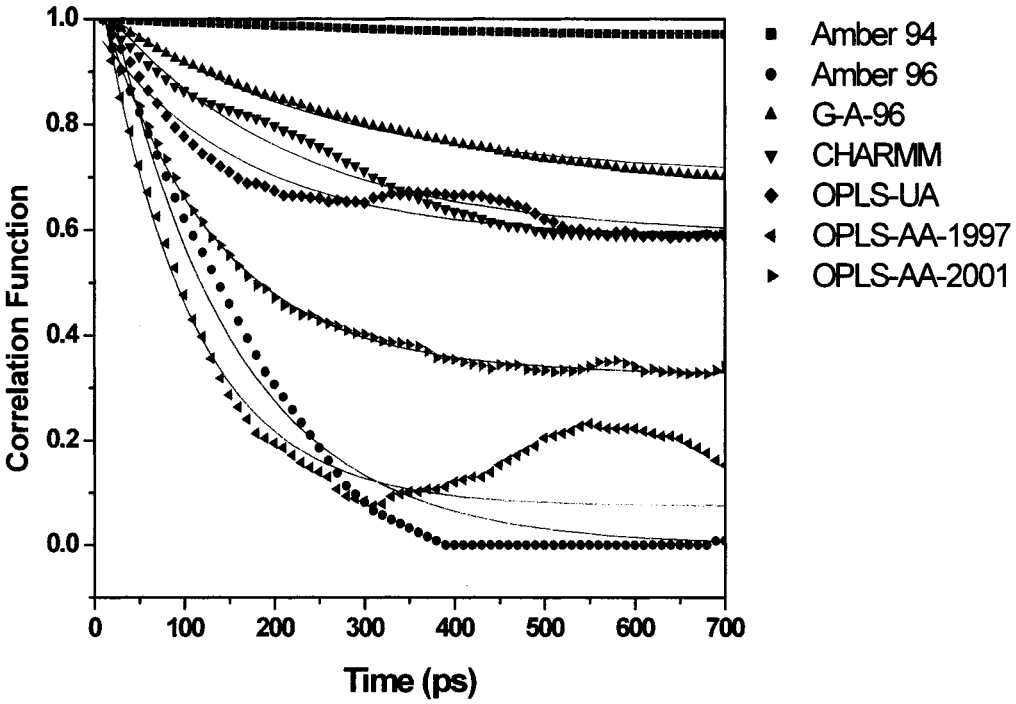
Differences and accuracy of FFs.

We have studied the equilibrium populations and inter-basin hopping kinetic with seven widely used FFs to examine the robustness of our conclusions, as well as to address questions concerning the consistency and reliability of the FFs for treating protein dynamics. As noted by Hu *et al.* (Hu et al., 2003)and Mu *et al.*(Mu et al., (in press)), an important difference among the force fields is the bias towards certain basins (as exhibited in Figs. 5.5 and 5.12). The AMBER 94 FF describes alanine-like residues as largely populating only the helical basin, while the AMBER 96 FF avoids this basin completely. The remaining five FFs lead to the helical, extended and PP-II basins as being populated more equally, although non-helical basins are not distinct in the OPLS-UA and OPLS-AA-97 FF. Garcia and coworkers correct for the "helophobicity" of the AMBER 96 FF by completely flattening the added AMBER 96 torsional potential, which is 1.5 kcal mol^{-1} unfavorable at the helical basin (Fig. 5.12). The OPLS-UA FF also has a flat added torsional potential, while the added potential varies by 0.5 kcal mol^{-1} for CHARMM and by as much as 2 kcal mol^{-1} for the OPLS-AA-97 and OPLS-AA-01 FFs. However, the added torsional potential only determines a portion of the backbone distribution because other interactions, such as partial charges, side-chain dihedral

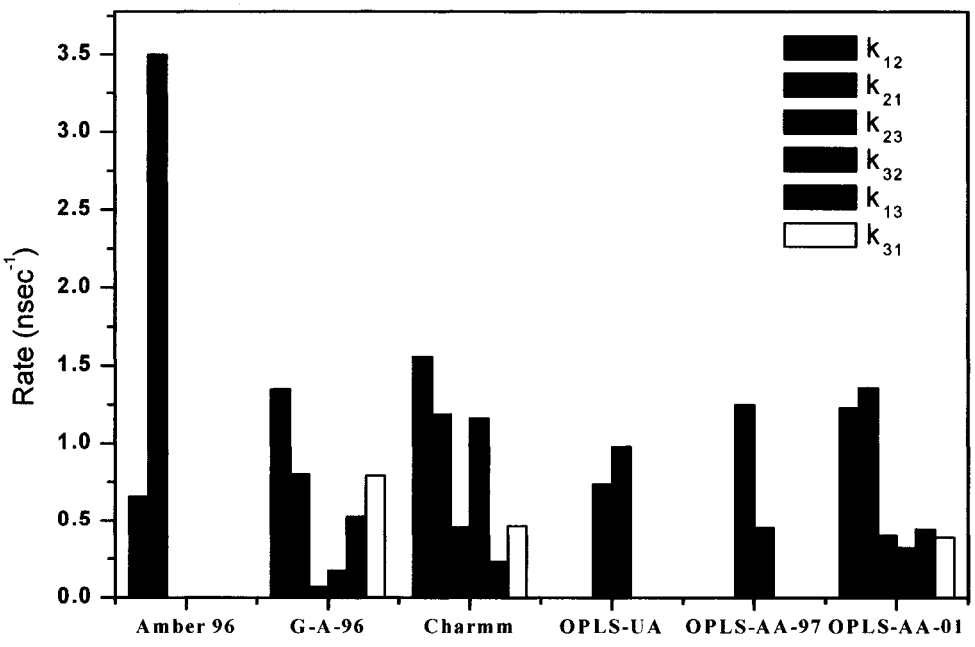
Figure 5.11. Basin Hopping Rates and directional sampling.

a) Correlation function for the basin 3 population for the center alanine in a tri-alanine peptide. The long time limit of the correlation function is the equilibrium population of basin 3, which strongly varies with the FF. Escape rates are obtained from single exponential fits to the correlation functions (red lines). The finite duration of the simulations is responsible for some of the noise in the correlation functions. The poorer exponential fits for the AMBER 96 and OPLS-AA-97 FFs probably arise because of the small basin 3 populations and because a limited number of transitions occur during a 15 ns trajectory for these two FFs. **b)** Inter-basin hopping rates for tri-alanine as calculated with several FFs. Rates for the AMBER 94 FF are not presented because essentially only one basin is populated and there are very few transitions. Similarly, the rates between basin 2 and basin 3 for AMBER 96 are omitted because basin 3 is rarely occupied. For the OPLS-UA and OPLS-AA-97 FF, the rates are between basin 3 and the combined basins 1 and 2, which are not distinct in these FFs.

a)


Correlation Function
Time (ps)
Amber 94
Amber 96
G-A-96
CHARMM
OPLS-UA
OPLS-AA-1997
OPLS-AA-2001

b)


Rate (nsec^{-1})
k_{12}
k_{21}
k_{23}
k_{32}
k_{13}
k_{31}
Amber 96
G-A-96
Charmm
OPLS-UA
OPLS-AA-97
OPLS-AA-01

potentials, and van der Waals interactions, contribute as well (Hu et al., 2003; Mu et al., (in press); Zaman et al., 2003a). Recent experimental studies have shown that alanine-rich unfolded peptides predominately populate the PP-II basins (Dukor & Keiderling, 1991; Schweitzer-Stenner, 2002; Shi et al., 2002a; Woutersen & Hamm, 2001a; Woutersen & Hamm, 2001b; Woutersen et al., 2001; Woutersen et al., 2002). Except for the AMBER 94 FF and the OPLS-UA FF, all FFs predict significant sampling of PP-II conformations. However, the PP-II basin still is not the most populated for any of the force fields. Thus, there is a disparity between the predictions of the FFs and experimental observations for very small peptides.

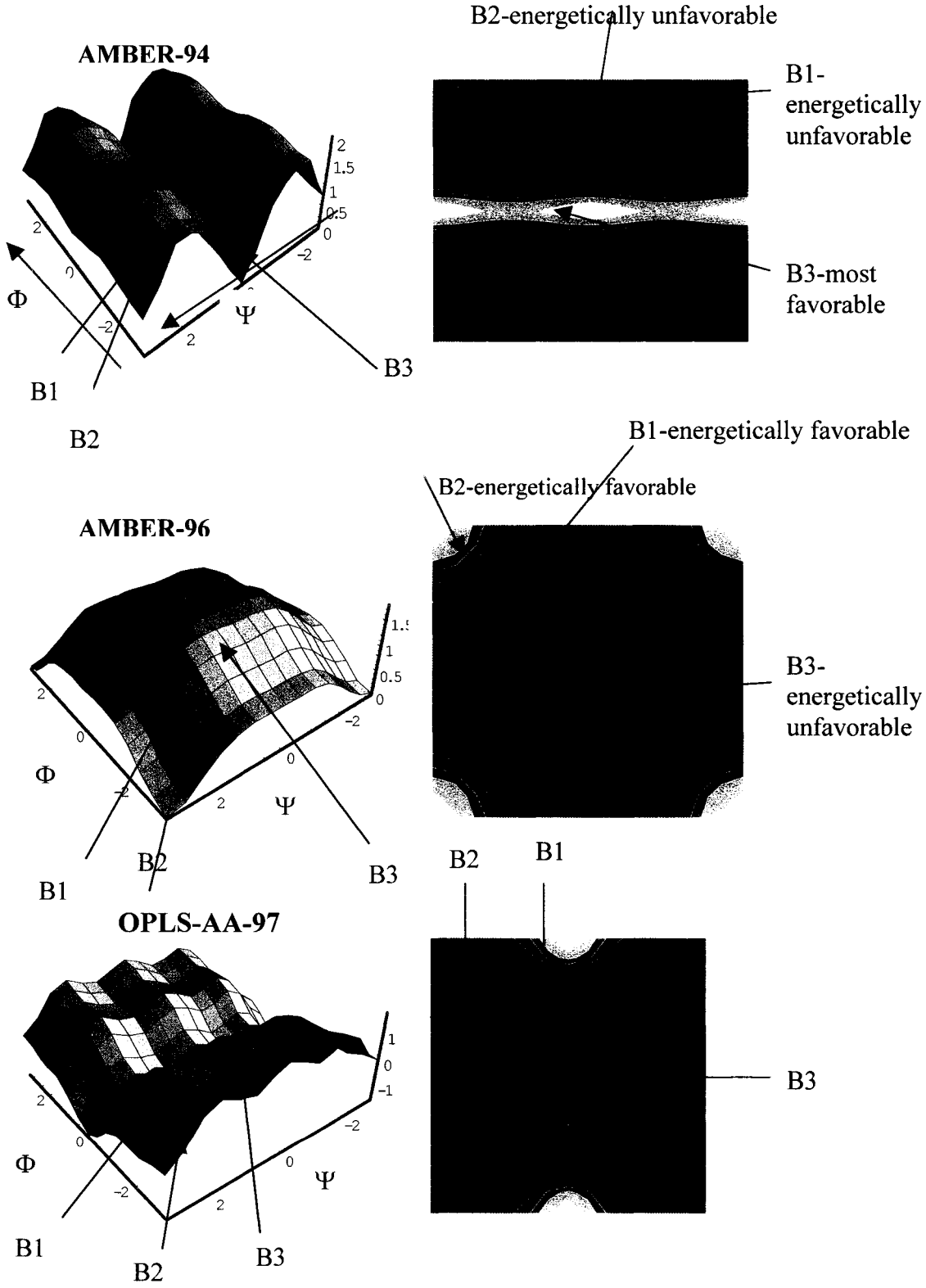
We suggest that this discrepancy may be a reflection of the fact that the FFs have been designed based on thermodynamic data (perhaps with some *ab initio* computations centered near potential minima). The protein FFs have generally been validated by the degree to which they can reproduce the structures of folded proteins. However, folded proteins tend to have much less PP-II structures than either helical or β-sheet structures, so the under weighting of the PP-II basin by the FFs is, perhaps, not too surprising. Additionally, the dynamics sensitively reflects the heights of the saddle-points connecting the basins, while the thermodynamic and quantum data used to parameterize the FFs are insensitive to these kinetic barriers.

Glycine flexibility and helical propensity.

Compared to alanine, the backbone of glycine is more flexible as it can traverse a larger range of the phi-psi map (Fig. 5.13). However, glycine still exhibits strongly preferred regions. This preference reduces the over-all sampling of configurations, and the

Figure 5.12. Torsional Biases in the Force Fields.

The added backbone torsional potential of the different force fields. The right column shows a contour plot of the surface in the left column. The axes are labeled in radians.


AMBER-94
B2-energetically unfavorable
B1-energetically unfavorable
B3-most favorable
Φ
Ψ
B1
B2
B3
B1-energetically favorable
B2-energetically favorable
AMBER-96
B3-energetically unfavorable
B1
B2
B3
OPLS-AA-97
B3

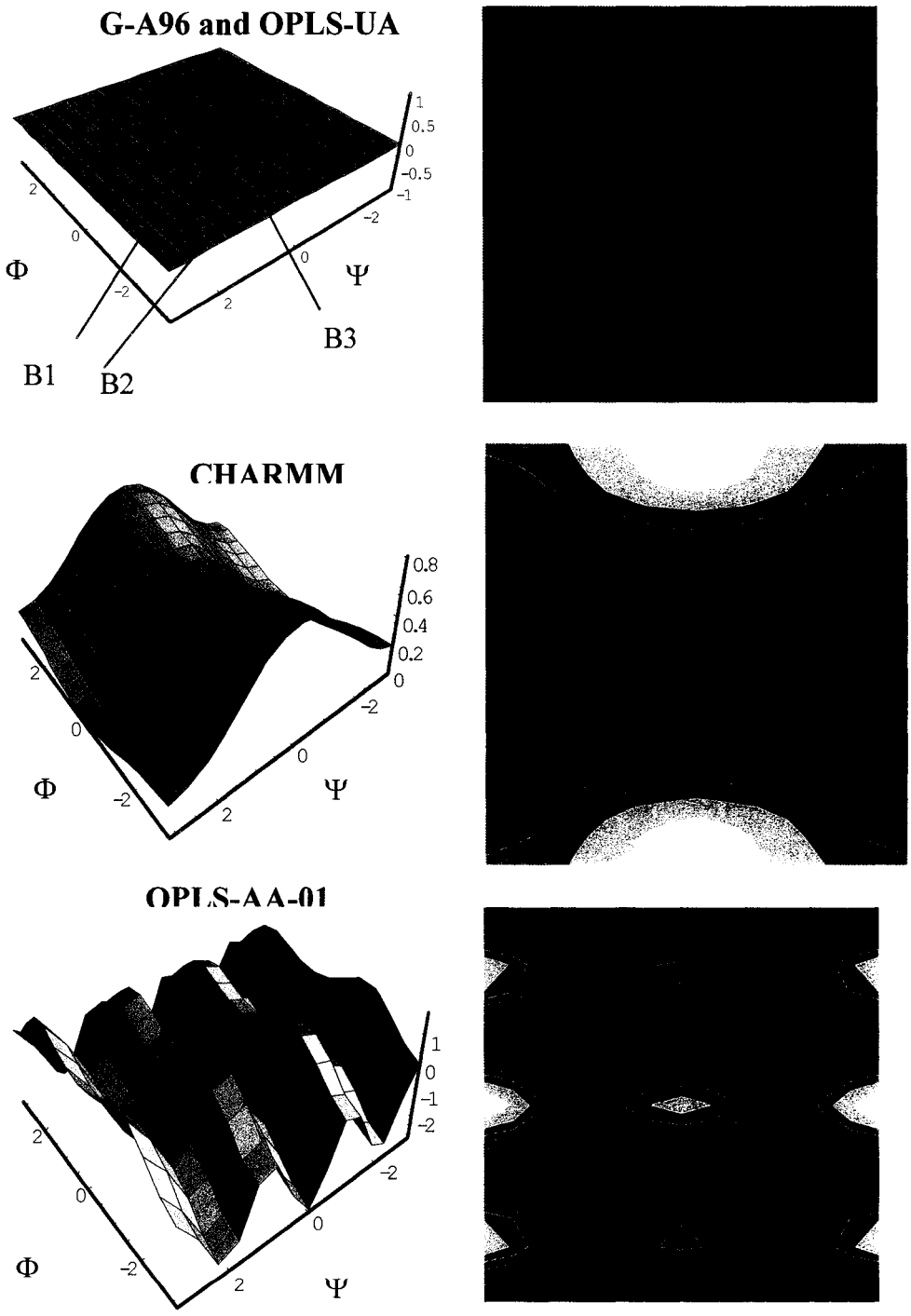


Fig. 5.12 (contd.)

Table 5.4. Sequence dependence of backbone entropy in Ala-X-Ala with unconstrained neighbors

X	$T(S_X - S_{Ala})$ (kcal mol^{-1})			
	AMBER 94	G-A-96	OPLS-AA-01	OPLS-UA
Ala	0	0	0	0
Asn	0.0045	-0.053	-0.008	-0.0515
Cys	0.06	0.0245	-0.017	-0.049
Val	-0.076	0.067	-0.276	-0.1495
Lys	-0.0815	0.033	-0.111	-0.0455
Glu	-0.041	4E-4	-0.139	-0.1135
Tyr	-0.06765	-0.125	-0.065	-0.02
Trp	0.1015	-0.0945	-0.32	-0.0095
Ser	-0.024	0.025	-0.036	-0.0235
Asp	-0.251	-0.1485	-0.852	-0.12495
Ile	-0.0425	-0.181	-0.674	-0.1565
Met	0.0255	-0.09995	-0.033	-0.0515
Phe	0.3465	-0.2061	0.22	-0.0645
Thr	-0.024	-0.022	-0.164	-0.1005
His	0.1795	0.0105	0.021	-0.0955
Pro	-0.2675	-0.398	-0.473	-0.2505
Gln	-0.02	-0.0385	-0.01	-0.0383
Arg	-0.0415	0.006	0.011	-0.0635
Leu	0.09	-0.032	-0.11	-0.04995
Gly	0.498	0.045	0.22	0.0455

backbone entropy is only modestly increased (using realistic FFs), $T(S_{Ala\text{-}Gly\text{-}Ala} - S_{Ala\text{-}Ala\text{-}Ala}) \leq 0.11$ kcal mol^{-1} in an Ala-Gly-Ala tripeptide (Table 5.5). It is generally believed that the difference in the helical propensities between Ala and Gly at a solvent exposed position is entirely attributed to differences in backbone entropy of the unfolded state, because the folded state has the same entropy and interactions (Creamer & Rose, 1994; D'Aquino et al., 1996). The difference in helical propensity between glycine and alanine in the folded state is greater than 0.7 kcal mol^{-1} (Creamer & Rose, 1994; D'Aquino et al., 1996), far larger than their difference in backbone entropy for the unfolded state (~0.11 kcal mol^{-1}; Table 5.5). This discrepancy between the known helical propensity and our calculation of backbone entropy implies either that 1) the assumption is incorrect and the difference in backbone entropy in the unfolded state is the primary factor determining the difference in helical propensity for these two residues, or 2) the FFs do not accurately reproduce the sampling of the unfolded state for alanine and/or glycine.

Time Scales and Comparisons with Experiments.

Our results indicate that simulations employing different commonly used FFs can produce basin hopping rates differing by ~5-fold as well as differentially populate the major basins. These findings agree with recent tri-alanine simulations obtained by Hu *et al.* (Hu et al., 2003)and Mu *et al.*(Mu et al., (in press)) using explicit solvent MD simulations and several different FFs. Inter-basin hopping rates and basin sampling affect the folding pathways, and, hence, the overall dynamics that are predicted by simulations. Consequently, the folding rate determined from folding simulations may contain further uncertainties. Given these issues, the uncertainties in FFs impart at least a factor of 2-3

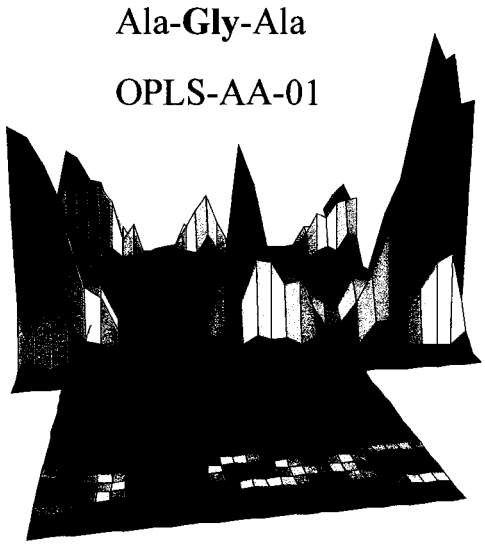


Figure 5.13. Basin population for Gly in Ala-Gly-Ala using the OPLS-AA-01 FF.

An unconventional view for the Ramachandran basins is used to enable visualizing both the population (top) plot and the contour (bottom) plot.

uncertainty in simulated rates (Snow et al., 2002). Furthermore, the extreme bias towards the helical basin from the AMBER 94 FF implies that any folding simulation using this FF is unreliable for either dynamics or thermodynamics.

In addition, protein folding simulations with a variety of FFs often tend to exhibit early collapse and the formation of structured intermediates (Alonso & Daggett, 1998; Duan & Kollman, 1998; Linhananta et al., 2002; Shen & Freed, 2002a; Shimada & Shakhnovich, 2002; Zagrovic et al., 2002). In contrast, the folding of small proteins is experimentally observed to be two-state without the accumulation of early intermediates ((Krantz et al., 2002) and references therein). Potentially, the early intermediates observed in the simulations arise due to inherent limitations of the FFs which are primarily designed to describe folded structures and not the dynamics of the folding process.

A possible source of the early collapse found in the simulations may lie in an inadequate treatment of the backbone entropy of the unfolded state. Although the backbone entropies are generally within $T\Delta S \sim 1/2$ kcal mol^{-1} of each other for the seven FFs, these values are for a single residue. Even a 0.1 kcal mol^{-1} error for a small, 100 residue protein could produce a net error of 10 kcal mol^{-1}, or a factor of 10^7 in the equilibrium constant for a fully collapsed species relative to the unfolded state. Thus, small errors in parameters of FFs easily can lead to folding mechanisms that are not observed experimentally.

Conclusions

Our simulations demonstrate that the Flory IPH is invalid because of non-negligible interactions between neighboring amino acids. The basin preference and backbone entropy of a residue depends both on the conformation and the identity of the neighbor. We estimate the magnitude of these effects to be $T\Delta S \leq 0.7$ kcal mol^{-1} $residue^{-1}$. Because Zimm-Bragg (Zimm & Bragg, 1959) and Lifson-Roig (Lifson & Roig, 1961) helix-coil theories do not include either a dependence on the sequence or on the conformation, there is opportunity for improving these theories by correcting for the changes in the entropy of the unfolded state due to NN effects.

Basin populations and inter-conversion rates strongly depend on the choice of force field. This dependence is larger than differences between explicit and implicit solvent calculations using the same force field, suggesting that explicit solvent calculations for the dynamics of small peptides are unnecessary until the FFs are improved.

The information we obtain concerning the basin hopping rates can be used in coarse-grained folding algorithms that are based solely on torsional dynamics (Colubri & Fernandez, 2002). Moreover, the preference of peptides for certain conformations can help characterize the structure and dynamics of the denatured state, and their influence on the folding pathway.

Methods:

The long-time dynamics (15 –45 ns) of the di- and tri-amino acids have been probed using the implicit solvent LD simulation method described by Shen and Freed

(Shen & Freed, 2002b) using seven different FFs at 300 K. The peptides are amino-acetylated and carboxy-amidated in order to model the dynamics of the two or three residues within a larger polypeptide. Similar simulations with uncapped ends lead to very different propensities for individual Ramachandran basins mainly because the charged ends favor elongated configurations more than in capped systems. Average basin populations and dynamics are accumulated after the first 3 ns of the equilibration simulations.

The Langevin dynamics simulations take the total system energy $U_{total} = U_b + U_{bend} + U_{tors} + U_{imp-tors} + U_{ch}(\varepsilon(r)) + U_{vdW} + U_{solv}(\sigma)$ as the sum of the types of interaction potentials between the solute atoms (as explained in chapter 4), while the solvent contributions are modeled using a distance dependent dielectric "constant" to screen charge-charge interactions $U_{ch}(\varepsilon(r))$ and a solvation potential $U_{solv}(\sigma)$. The bonding interactions U_b, bond-bond bending interactions U_{bend}, and improper torsional energies $U_{imp-tors}$ are modeled by harmonic potentials, the regular torsional potentials U_{tors} by standard periodic functions, and the van der Waals interactions by Lennard-Jones 6-12 potentials. The Coulomb interactions $U_{ch}(\varepsilon) = \sum_{i>j} q_i q_j / \varepsilon(r_{ij}) r_{ij}$ are expressed in terms of atomic partial charges q_i and a Ramstein-Lavery style (Ramstein & Lavery, 1988) distance dependent dielectric "constant" $\varepsilon(r)$. The microscopic solvation potential is modeled using the Ooi-Scheraga solvent–accessible surface area (SASA) potential (Ooi et al., 1987), $U_{solv}(\sigma) = \sum_{i=1}^{N} g_i \sigma_i$, where σ_i is the accessible surface area of a

hypersurface bisecting the first solvent shell surrounding protein atom i, and the empirical surface free energy parameters g_i depend on the atom type. Because the g_i are free energy parameters, the U_{total} generates a temperature dependent free energy that contains contributions from solvent reorientation within a mean-field approximation.

The LD simulations employ the velocity Verlet algorithm (Allen, 1987) with a time step of $\Delta t = 2$ fsec for integrating the equations of motion for the protein atom positions and velocities. The length of all X-H type bonds are constrained using the RATTLE algorithm (Anderson, 1983). The computations are performed using a modified version of the TINKER 3.9 molecular design package (Ponder, 1999) with a faster non-bonding force evaluation algorithm FAST-LD (Shen, 2002). The frictional forces and corresponding random forces acting on the protein atoms are computed using the Pastor-Karplus accessible surface area model (Pastor & Karplus, 1988). The solvent accessible surface areas σ_i for the friction coefficients are calculated from the exposed surface area of solute atoms using a probe of zero radius. The smaller probe size for friction coefficients is used to cancel effectively the results of (more expensive to calculate) hydrodynamic interactions. The accessible surface areas, atomic friction coefficients, and solvation potentials are updated every 100 dynamical steps (0.25 ps) since tests show that this approximation incurs negligible error because significant conformational variations occur on a much longer time scale (Shen & Freed, 2002b).

Identification of Basin locations:

The entropy calculations (Eq. 5.1) do not depend upon how each basin is defined, as the probability is calculated for each of the 3°x3° grid elements. However, the

populations shown in Tables 5.1 and 5.2 and the rates depicted in Fig. 5.10 depend upon the definitions of individual basins. Basins 1, 2 and 3 are defined based upon the population of central Ala in tri-alanine (Fig 5.2). Basin 3 is defined with a circle large enough to encompass the population of that basin for all the FFs (Fig 5.3). This definition is used to calculate the rate of escape from basin 3 shown in Fig 5.9. The distinction between basins 1 and 2 is only applicable for G-A-96, OPLS-AA-01 and CHARMM, as other FFs either do not have a clear separation between theses two basins (OPLS-UA and OPLS-AA-97) or have all of its population only in a single basin (AMBER 94 and AMBER 96). For basins 1 and 2, the G-A-96, OPLS-AA-01 and CHARMM FFs are used to define non-overlapping ellipses that are large enough to accommodate >90% of the populations in each of these basins.

Independence of initial conditions and length of simulation.

In order to test the robustness of the computed neighbor effects, simulations have been performed for four different di-peptides with varying initial conditions and variable durations of 45 and 15 ns. The overall difference in basin populations is less than 3% (due to different initial conditions and longer trajectories), indicating that the basins are adequately sampled within 15ns and that the results are not an artifact of the initial conditions or the use of short trajectories.

Calculation of k_{ij} (inter-basin transition rates) from basin auto correlation function:

In order to calculate k_{ij}, the rate of transition from basin i to basin j, the escape rate from each basin is calculated. The population decay rate is obtained from an exponential fit to the autocorrelation function $C_i(t)$ for each basin (after having subtracted the long time basin population). Because transitions from basins 1 and 3 overwhelmingly proceed to basin 2, the decay rates of $C_i(t)$ for the basin 1 and 3 correlation functions equal k_{12} and k_{13}, respectively. The decay rate of $C_2(t)$ is the sum $k_{21} + k_{23}$, which can be separated using the equilibrium basin populations and the detailed balance condition for equilibrium, e.g., [basin 1]/[basin 2]= k_{12}/k_{21}).

Calculation of backbone entropies.

Equation 5.1 is only an approximate relation. The conformational entropy can only be computed rigorously from conformational populations when the latter are obtained from a constant energy simulation. However, both the friction coefficients and the solvation potential are inherently temperature dependent quantities, so constant energy implicit solvent simulations are not possible. A more rigorous approach would be to follow the far more computationally costly simulation methods of Okamoto and coworkers (Mitsutake et al., 2000; Sugita et al., 2000; Sugita & Okamoto, 1999), but this would not be possible for the wide range of tri-peptide systems and FFs studied here. Hence, the approximate for of Eq. 5.1 suffices for our broad study.

6. CONCLUSIONS AND FUTURE WORK

Summary

The work presented in this thesis addresses important and unanswered questions related to protein folding and protein-dynamics. We have addressed the issue of entropic benefit of cross-linking in protein-association with a method that is independent of the shape and size of solute and solvent molecules. Our method has shown good agreement with experimental methods. This method has wide applications in designing macromolecular complexes with tethers and calculating the entropic benefit due to them.

Our work has also addressed the issue of heterogeneity of reaction pathways in protein-folding and other macromolecular reactions. The analysis is conceptually simple and is based on first principle statistical mechanics and the widely used transition state theory. We believe that our theory will improve our understanding of multiple path processes, and will provide as a starting point for more sophisticated calculations that will take into account the density of states of initial and final states.

The computer simulation results discussed in the present thesis have shown some of the inconsistencies of the commonly used force-fields when subjected to rigorous dynamics tests. The results not only demonstrate the efficiency of our implicit solvent-LD algorithm but also suggest that the results of computer simulations should be taken with a grain of salt unless the differences among various force fields and experiments are

resolved. The work done in this regard is not meant to prove which force field is better but is aimed at improving the parameters of force fields in general. It is also aimed at developing parameters which will be optimized for dynamics in addition to structure and thermodynamics.

Finally, the LD simulations on small alanine peptides show that the Flory Isolated-Pair-Hypothesis is invalid for small peptides and that the nearest neighbor amino-acids interact with each other and their motion is correlated. This result has implications for the understanding of protein dynamics at a residue level, and will improve the helix-coil theories by taking into account the correlated motions between amino acids.

The next section discusses some of the avenues for future research that emerge naturally out of the work presented in previous chapters.

Future work

Application of calculation of NN method to RNA:

In spite of its ability to calculate the entropic benefit of cross-linking independent of the nature of the solvent or solute molecules, the NN method discussed in chapter 2 needs more refinement before it can be applied to calculate the entropy costs associated with closing RNA hairpin loops. This is due to sensitive dependence of RNA stability on its sequence. The protein stability is also a function of sequence, however, the loop closure entropy doesn't depend very sensitively on the sequence and therefore our method shows

good agreement with experiments. On the other hand, one base-pair difference in the sequence of RNA hairpins can change the stability significantly. This sensitive dependence requires a more refined NN method that has parameters optimized for the RNA systems. In fact, one can imagine there being two sets of parameters for the NN method, one for the protein loops and one for the RNA hairpins, however more literature search and tests with different parameters are needed for the development of RNA NN method.

As mentioned earlier in chapter 2, the NN method renders itself very well to calculate the entropic benefit of cross-linking in higher order complexes with multiple binding sites where the ligands are linked together. The application of the NN method to higher-order systems will provide a rigorous test of our method and will open new doors for protein engineering. It will also benefit organic chemists interested in designing self assembled structures with tethers.

Application of density of states method to sequential pathways:

The method outlined in chapter 3 is applied to a macromolecular reactions with parallel pathways. The method is optimized to study reactions with heterogeneous pathways, however, a wide variety of biological processes (such as non-two state protein-folding) go through sequential pathways. Our work, will hopefully, be useful in laying the foundations of a method that describes macromolecular reactions proceeding through sequential pathways. Some of the work on sequential pathways has already been done (Despa & Berry, 2001a; Despa & Berry, 2001b; Despa et al., 2003)however, the method

has not been compared to experimental results, and does not take into account the density of states of the initial, intermediate and final states. We hope, that our work, will be useful in fine tuning the present methods and will be helpful in getting better agreement between experiments and theory.

In order to have a complete theoretical understanding of multiple pathway processes in macromolecules it is imperative that a method is developed that takes into account the possibility of a combination of sequential and parallel pathways. It is our hope that such a picture will be useful in elucidating the various aspects of multiple path process and will lay the ground work for determining the fingerprints of a multiple-path folding process from the Arrhenius and other rate-temperature plots.

Improvements in the LD-Implicit solvent algorithm

The LD algorithm discussed in chapters 4 and 5 has shown very promising results. Without compromising on the quality of results, the method has cut down computational costs by a factor of more than 200. In spite of this remarkable achievement of the algorithm, there are still many aspects of the program that need further improvements. The areas that need improvement include a better treatment of the dielectric constant. Another area of improvement is the need for a more accurate method to calculate the solvent-solute hydrogen bonds. The solvation potential, though currently able to reproduce the results of explicit solvent MD, can be further improved. Finally, so far the molecular mechanics simulations have not been able to come up with any reasonable treatment of denaturing solvent. This is important since the experimental unfolded protein is usually in a denaturant and not in conditions similar to those of unfolding simulations

(> 1000 K). A similar starting point for the in-silico and in-vivo folding reactions will also be useful in comparing theoretical and experimental results. We hope that such simulations will be able to describe the events on the folding pathway more accurately. This will also highlight the origin of early collapse seen in simulations and not in experiments, as it is unclear at the moment as to whether the origin of early collapse is the unrealistic starting state or the force-fields.

Combining torsional dynamics and LD simulations

The Folding Machine (FM) based upon torsional dynamics of the residues has proved to be fairly successful in predicting the three dimensional structure of proteins from the amino acid sequence in computationally inexpensive manner. It is one of the only few methods that relies on ab-initio methods (as compared to knowledge based methods) to predict the final structure of the protein from the amino acid sequence. There are, however, many areas of FM that need further improvement. Among these areas are the definition of basins, the basin hopping rates and the directional hopping of residues between different basins. Our work, as outlined in chapter 5, can help in solving these problems by using all atom simulations. The parameters of basin depths and heights, basin hopping rates, and directional hopping of residues between specific basins can be determined by the methods outlined in chapter 5 and sample programs presented in the appendix section. These values, though based upon the unfolded state, will be more accurate than the ad-hoc potential currently used by the FM. The FM currently utilizes the protein database for its many parameters, and therefore lacks information about the

unfolded state. The results from LD simulations will overcome this problem and will give the FM the ability to mimic the protein in the unfolded state.

The combination of coarse grained simulations (FM) and fine grained calculations (LD/ MD) is being carried out at other levels also. It is our hope, that in future, we will be able to combine the two methods, in such a way that the initial stages of the folding event will be simulated by the FM which will identify some of the potential structures and the minima on the potential energy landscape. After this identification has been done, fine tuning of structures, and the determination of the lowest energy minima will be carried out by fine-grained simulations such as LD. Though this seems rather simple and straight-forward, there are practical considerations that must be taken into account. First of all, the potential function for the LD and FM are very different, and as outlined in previous chapters, small perturbations in the potential function can lead to entirely different results. Therefore, a consistent potential function needs to be developed. In addition, the identification of a point, where one can switch from FM to LD is not clear, and requires a lot of tests with different starting points. It is also possible that this "switch point" will lie at different points in the trajectory for different systems, therefore making the transition from FM to LD more complicated. Also, it must be kept in mind that these results will only be valid for structure prediction, and due to the nature of the FM (MC type algorithm) no detailed information about the overall dynamics of the folding pathway can be obtained at the moment.

In spite of all the above mentioned issues, it will be interesting to see whether or not there is any improvement in the results of the FM with the new parameters as

obtained by fine-grained LD simulations. These results, will hopefully, open up new areas of research of combining the two methods, and will enhance our ability to predict 3D structures of proteins from the amino-acid sequence.

Testing FFs for folding

The current thesis presents our analysis of the FFs when they are subjected to rigorous dynamics tests. Since the FFs are optimized for structure and thermodynamics, it is only natural, that we see problems with the FFs when they are tested for dynamics. We plan to further study the problems in the FFs by subjecting them to tests that analyze both the dynamics and the structure prediction simultaneously. This can be done by testing for a given sequence, which FFs fold the protein into a known native structure. The preliminary work for this kind of tests has already been done. We have run simulations (4 simulations with seven different FFs, each of 100 ns) of a 13mer partially helical protein (Glu Asn Glu Val Ala Arg Leu Lys Lys Leu Leu Gly Tyr) and a synthetic 13mer beta-hairpin protein (Ser Trp Thr Trp Glu Gly Asn Lys Trp Thr Trp Lys). The preliminary results show that only two out of the seven (AMBER 94 and OPLS-AA-1997) FFs form a stable helix for the 13mer. The experimental results show about 35-45% helix in 0% TFE, as measured by CD experiments. It will be interesting to see whether the more realistic FFs (G-A-96 and OPLS-AA-2001) are able to predict similar helical content or not. The folding pathway, the dynamics and the loss of entropy upon folding will also provide useful parameters in testing the FFs. Finally, to reach any broad-based conclusion, it will be important to see whether the FFs which produce agreement with experiments for the helical 13mer, are able to provide good results for the beta-hairpin or

not. Our results discussed in previous chapters suggest that the force fields show a strong bias for particular conformations when applied to small peptides (3-5 residues) but it is not clear as to whether these biases play as important role in relatively longer (13mer) peptides as they do in small tri-mers and penta-mers.

Long range effects on Dynamics:

One aspect of dynamics still not clearly understood is its relation with the longer-range effects. As discussed in previous chapters, we have the tools to study the dynamics of short peptides, and it would be only a natural extension to apply our methods to study the basin hopping dynamics of longer peptides. The following strategies are proposed to study the long-range effects and their correlation with dynamics of individual residues:

1. LD simulations of relatively larger peptides. In this regard, we have long (>100 ns) trajectory of different versions of cross-linked GCN4 (72 and 84 residues), villin headpiece (33 residues), helical 13mer and a synthetic 13mer that forms a beta-hairpin (13 residues) and Met-Enkephalin (5 residues). We can study the dynamics of these systems on a residue by residue level and study the basin hopping rates, overall tumbling and other dynamic properties as a function of length of the system.
2. This study will be complemented by LD simulations where all the long-range interactions are turned off, and only i±1 or i±2 interactions are allowed. In other words, only nearest neighbors or next-nearest neighbors can interact, but all the other interactions are turned off. This first of its kind simulation will require significant changes in the algorithm computing the trajectories of individual

atoms, and will also have to include a routine to address the excluded volume issue. Nonetheless, the results from this simulation will complement the ones discussed in point # 1. These results will give us information about how much of the overall dynamics are due to the long-range interactions. In addition, this study will also provide a good model for the unfolded state of proteins, and can be used to compare results with experiments such as small angle x-ray scattering (SAXS) experiments that measure the radius of gyration.

3. Finally, the above mentioned simulations can help us answer the question about the dependence of position in oveall dynamics, i.e. whether the residue at the end of the protein is more dynamic than the residues in the center or not. This question will benefit from trajectories obtained by both using long range interactions and trajectories that ignore long range interactions. Such information will also be useful in fine-tuning programs that are not based on atomistic dynamics such as the FM.

These results will highlight the nature of long-range interactions in protein folding and protein-protein interactions and will help in developing a better model for protein folding.

FF optimization for dynamics

Most the above mentioned goals for future research will utilize the state of the art algorithm to compute the atom trajectories using implicit solvent and a fast LD algorithm. These results, however, will only be meaningful if there is a consistent force field available, since running a simulation with all force fields to see consistency among them

is time consuming, and often not practical. Therefore, there is a need for a consistent FF that is optimized for dynamics. This a multidimensional project that requires collaboration with different groups such as ab-initio quantum theorists, experimentalists carrying out NMR experiments to determine dynamics in solution and many other researchers. Some of this work has already started where the hydrogen exchange rates of a small 18 residue peptide E6ap will be measured experimentally and compared with long time simulations. However, this will only give a part of the picture, as more experiments on proteins with different native structure (and random coils) need to be conducted to determine some of the relevant dynamics quantities. In addition, quantum mechanical calculations on small peptides will be useful in determining the parameters needed for a better force field. Above all, there need to be more tests on dynamics using these FFs (such as the ones outlined in chapters 4 and 5) that will point out the regions where there is a need for improvement.

It is my hope that one day, some graduate student, or a group of graduate students will create or improve one of the current FFs so that it is optimized for dynamics and thus will decrease the gap that exists between theory and experiment today.

APPENDIX 1. PLOTTING RAMA MAPS

```
(*Load the appropriate package*)
<< Graphics`MultipleListPlot`; << Graphics`Graphics`;
<< Statistics`DataManipulation` << Graphics`Graphics3D`;

(*Import the Ramachandran files from the command "property" in TINKER ;
 Here the Rama files happen to be in the directory c:\cap *)

a1 = Import["c:/cap/rama-ala-amber", "Table"];
a2 = Import["c:/cap/rama-ala-96", "Table"];
a3 = Import["c:/cap/rama-ala-garcia", "Table"];
a4 = Import["c:/cap/rama-ala-ch", "Table"];
a5 = Import["c:/cap/rama-ala-opls", "Table"];
a6 = Import["c:/cap/rama-ala-oplsaa", "Table"];
a7 = Import["c:/cap/rama-ala-oplsaal", "Table"];

(*For loop going through all the imported Rama coordinates*)

For[α = 1, α < 8, α++, {
  j = 1;

  (*For loop starting at 2 since we are looking at the populations of
    the center residue; for 15 ns for each residue, we have 4500 points in all*)

  For[i = 2, i < 4501, i++,
   {

    (*The list index d extracts the phi and psi*)

    di = {aα[[i, j]], aα[[i, j + 1]]};
   }];
  t = Table[di, {i, 2, 4500}];
```

```
(*use the command "bincounts" to bin the data,
   here the binsize is 12 degrees each,
   therefore the whole plot will be 30 x 30 bins*)

  y = BinCounts[t, {-180, 180, 12}, {-180, 180, 12}];

  (*The command listshadowplot3D gives a 3 D Rama plot and the viewpoint
    can be altered for the angle of view one is interested in*)

  ListShadowPlot3D[y, ViewPoint→ {-.2, -1.5, .39}, Boxed→ False];}]

(*The plot is plotted using G-A-96*)
```

APPENDIX 2. PLOTTING LTM PLOTS

```
(*importing the rama files for the 20 amino acid combinations*)
a1 = Import["c:/axag/rama-a-ala-ag", "Table"];
a2 = Import["c:/axag/rama-a-asn-ag", "Table"];
a3 = Import["c:/axag/rama-a-cys-ag", "Table"];
a4 = Import["c:/axag/rama-a-val-ag", "Table"];
a5 = Import["c:/axag/rama-a-lys-ag", "Table"];
a6 = Import["c:/axag/rama-a-glu-ag", "Table"];
a7 = Import["c:/axag/rama-a-tyr-ag", "Table"];
a8 = Import["c:/axag/rama-a-trp-ag", "Table"];
a9 = Import["c:/axag/rama-a-ser-ag", "Table"];
a10 = Import["c:/axag/rama-a-asp-ag", "Table"];
a11 = Import["c:/axag/rama-a-ile-ag", "Table"];
a12 = Import["c:/axag/rama-a-met-ag", "Table"];
a13 = Import["c:/axag/rama-a-phe-ag", "Table"];
a14 = Import["c:/axag/rama-a-thr-ag", "Table"];
a15 = Import["c:/axag/rama-a-his-ag", "Table"];
a16 = Import["c:/axag/rama-a-pro-ag", "Table"];
a17 = Import["c:/axag/rama-a-gln-ag", "Table"];
a18 = Import["c:/axag/rama-a-arg-ag", "Table"];
a19 = Import["c:/axag/rama-a-leu-ag", "Table"];
a20 = Import["c:/axag/rama-a-gly-ag", "Table"];
(*loading the graphics packages*)
<< Graphics`MultipleListPlot`; << Graphics`Graphics`;
(*for loop to run through all 20 amino acids*)
For[α = 1, α < 21, α++, {

  (*the index j is used to tell the processor where to look for the phi
    (j=1) or psi (j=1+1) value *)
  j = 1;
  (*For loop to look at the center resdiue, i.e. residue # 2,
     residue 1 corresponds to line 1, residue 2 to line 2, residue 3 to line 3,
     and residue 1 to line 4 etc). The loop runs from line 2 to line
    4500 in multiples of 3, since there are 1500 data points
    per residue*)
  For[i = 2, i < 4501, si = 0;
```

```
(*which loop to define basins;
   s_i is used to say which basin it corresponds to. the value of s_i is
    between 0 and 1, since the command raster for LTM
    can only recognize a value between 0 and 1. The different basins
    correspond to different values,
   thus different colors. The first is helical basin with value 0.3 and
    has green color, the second is PP-II with value 0.7 and has a blue color,
   the third is extended basin with value 1 and has a red color,
   and rest everything is "other basins" with a vlue of 0.1 and a dark yellow color*)

  {Which[

    (-130 ≤ a_α[[i, j+1]] ≤ 40 && -180 ≤ a_α[[i, j]] ≤ -30), s_i = 0.3,

    (-104 < a_α[[i, j]] < -30 && 70 ≤ a_α[[i, j+1]] ≤ 180) ∨
     (-104 < a_α[[i, j]] < -30 && -180 ≤ a_α[[i, j+1]] ≤ -160), s_i = .7,

    (-180 ≤ a_α[[i, j]] ≤ -116 && 70 ≤ a_α[[i, j+1]] ≤ 180) ∨
     (-180 ≤ a_α[[i, j]] ≤ -116 && -180 ≤ a_α[[i, j+1]] ≤ -160),
    s_i = 1,

    (-180 ≤ a_α[[i, j]] ≤ -120 && -140 ≤ a_α[[i, j+1]] ≤ 0) ∨
     (150 ≤ a_α[[i, j]] ≤ 180 && -140 ≤ a_α[[i, j+1]] ≤ -10), s_i = .1,

    (0 ≤ a_α[[i, j]] ≤ 100 && -180 ≤ a_α[[i, j+1]] ≤ -140) ∨
     (0 ≤ a_α[[i, j]] ≤ 100 && 140 ≤ a_α[[i, j+1]] ≤ 180), s_i = 0.1,


    (35.0 ≤ a_α[[i, j]] ≤ 100 && -50 ≤ a_α[[i, j+1]] ≤ 145) ∨
     (0 ≤ a_α[[i, j]] ≤ 35.0 && 0 ≤ a_α[[i, j+1]] ≤ 55), s_i = 0.1,

    (100 ≤ a_α[[i, j]] ≤ 150 && -180 ≤ a_α[[i, j+1]] ≤ 10) ∨
     (100 ≤ a_α[[i, j]] ≤ 150 && 30 ≤ a_α[[i, j+1]] ≤ 180), s_i = 0.1,

    (40 ≤ a_α[[i, j]] ≤ 100 && -140 ≤ a_α[[i, j+1]] ≤ -50) ∨
     (0 ≤ a_α[[i, j]] ≤ 40 && -140 ≤ a_α[[i, j+1]] ≤ -105) ∨
     (35.0 ≤ a_α[[i, j]] ≤ 45.0 && -140 ≤ a_α[[i, j+1]] ≤ -50) ∨
     (15.0 ≤ a_α[[i, j]] ≤ 40 && -110 ≤ a_α[[i, j+1]] ≤ -40), s_i = 0.1];
   i = i+3;
  }];
```

```
(*putting everything in a table form*)
   dα = Table[sj, {j, 2, 4500, 3}];


   }];
(*using raster to display an LTM plot; and colors defined by
   mathematica system hue*)
x1 = Table[dα, {α, 1, 20}];
Show[Graphics[Raster[x1, ColorFunction→ Hue]]];
```

APPENDIX 3. CALCULATION OF CORRELATED ENTROPY

```
(*Load the packages*)
<< Graphics`MultipleListPlot`; << Graphics`Graphics`;
<< Statistics`DataManipulation` << Graphics`Graphics3D`;

(*Import the all the files with Rama coordinates*)
a1 = Import["c:/aetn/rama-aetn-garcia", "Table"];
a2 = Import["c:/aetn/rama-aetn-charmm", "Table"];
a3 = Import["c:/aetn/rama-aetn-op", "Table"];
a4 = Import["c:/aetn/rama-aetn-oplsaal", "Table"] ;

(*ALGORITHM FOR CALCULATION OF CORRELATED ENTROPY*)
(*The for loop goes over all the 4 FF files imported*)
For[α = 1, α < 5, α++, {
  j = 1;

  (*The total points for each residue are 22 ns and since there are
    four residues, the total number of points is 8800*)
  For[i = 1, i < 8800, i = i + 4,
   {
    di = {aα[[i + 2, j]], aα[[i + 2, j + 1]], aα[[i + 3, j]], aα[[i + 3, j + 1]]};
   }];

  t1 = Table[di, {i, 1, 8800, 4}];

  (*2 Dimensional bin counting to make a 4 D Rama plot ;
   each axis has 30 bins, therefore the total size is 30 x 30 x 30 x 30*)

  y = BinCounts[t1, {-180, 180, 12}, {-180, 180, 12}, {-180, 180, 12},
    {-180, 180, 12}];


  For[i = 1, i < 31, i++,
   {For[j = 1, j < 31, j++,
     {For[k = 1, k < 31, k++,
       {For[l = 1, l < 31, l++,
```

```
(*If the bin in the 4D plot has no member, the entropy is
                zero, otherwise entropy is calculated by the formula
                below with normalization due to the number of bins
                and the total number of points collected by LD simulation*)

              {Which[y[[i, j, k, l]] == 0, trop2l = 0, y[[i, j, k, l]] ≠ 0,
                trop2l = N[-2 * (((y[[i, j, k, l]]) / (2200)) *
                     Log[(810000 * y[[i, j, k, l]]) / (2200)])]];

                (*incrementing the values*)
                trop20 = 0;
                trop2l = trop2l + trop2l-1;
                trop3k = trop2l;
              }];
            trop30 = 0;
            trop3k = trop3k + trop3k-1;
            trop4j = trop3k;
            }];
          trop40 = 0;
          trop4j = trop4j + trop4j-1;
          trop5i = trop4j;
          }];}];

  (*total entropy over all the bins in the 4D rama plot*)
  sα = Sum[trop5i, {i, 1, 30}];
  (*Printing the correlated entropy values*)
  Print[sα];
  }];
```

APPENDIX 4. CALCULATION OF BACKBONE ENTROPY

```
(*ALGORITHM TO CALCULATE THE ENTROPY OF A SINGLE RESIDUE;
 SAME ALGORITHM TO CALCULATE ENTROPY OF RESIDUE 1, 2,
 3 OR N IN THE PEPTIDE*)
(*α goes from 1 to 5 to calculate for all the four force fields
  mentioned above in the import command*)
For[α = 1, α < 5, α++, {
   j = 1;
   (*calculating the entropy of the fourth residue in the peptide*)

   For[i = 4, i < 8801, i = i + 4,
    {
     d_i = {a_α[[i, j]], a_α[[i, j + 1]]};
    }];
   t = Table[d_i, {i, 4, 8800, 4}];
   y = BinCounts[t, {-180, 180, 12}, {-180, 180, 12}];
   (*For loops for summation over 2 D Rama map, only phi and psi*)
   For[k = 1, k < 31, k++,

    {For[m = 1, m < 31, m++,
       {
        (*Which command to calculate entropy; if the bin has no member,
         entropy is zero, otherwise its given by the formula*)

        Which[y[[k, m]] == 0, ent_m = 0, y[[k, m]] ≠ 0,
         ent_m = N[-2 * (((y[[k, m]]) / (2200)) * Log[(900 * y[[k, m]]) / (2200)])]];
        ent_0 = 0;
        ent_m = ent_m + ent_(m-1);
        trop_k = ent_m;}];

    }];
   (*summing the entropy from each bin*)
   s_α = Sum[trop_k, {k, 1, 30}];
   (*Printing the value of the entropy*)
   Print[(s_α)];
  }];
```

APPENDIX 5. CALCULATION OF CORRELATION FUNCTION

```
(* Loading the appropriate packages and importing the files*)

<< Graphics`MultipleListPlot`; << Graphics`Graphics`;
<< Statistics`DataManipulation` << Graphics`Graphics3D`;
a1 = Import["c:/aaa/rama-a-ala-a-amber-long", "Table"];
a2 = Import["c:/aaa/rama-a-ala-a-96-long", "Table"];
a3 = Import["c:/aaa/rama-a-ala-a-garcia-long", "Table"];
a4 = Import["c:/aaa/rama-a-ala-a-ch-long", "Table"];
a5 = Import["c:/aaa/rama-a-ala-a-opls-long", "Table"];
a6 = Import["c:/aaa/rama-a-ala-a-oplsaa-long", "Table"];
a7 = Import["c:/aaa/rama-a-ala-a-oplsaal-long", "Table"];
For[α = 1, α < 8, α++, {
  j = 1;

  For[i = 2, i < 13501, si = 0,
   {Which[
     (-130 ≤ aα[[i, j + 1]] ≤ 40 && -180 ≤ aα[[i, j]] ≤ -30), si = 1,

     (-104 < aα[[i, j]] < -30 && 70 ≤ aα[[i, j + 1]] ≤ 180) ∨
      (-104 < aα[[i, j]] < -30 && -180 ≤ aα[[i, j + 1]] ≤ -160), si = 0,

     (-180 ≤ aα[[i, j]] ≤ -116 && 70 ≤ aα[[i, j + 1]] ≤ 180) ∨
      (-180 ≤ aα[[i, j]] ≤ -116 && -180 ≤ aα[[i, j + 1]] ≤ -160),
     si = 0,

     (-180 ≤ aα[[i, j]] ≤ -120 && -140 ≤ aα[[i, j + 1]] ≤ 0) ∨
      (150 ≤ aα[[i, j]] ≤ 180 && -140 ≤ aα[[i, j + 1]] ≤ -10), si = 0,

     (0 ≤ aα[[i, j]] ≤ 100 && -180 ≤ aα[[i, j + 1]] ≤ -140) ∨
      (0 ≤ aα[[i, j]] ≤ 100 && 140 ≤ aα[[i, j + 1]] ≤ 180), si = 0,


     (35.0 ≤ aα[[i, j]] ≤ 100 && -50 ≤ aα[[i, j + 1]] ≤ 145) ∨
      (0 ≤ aα[[i, j]] ≤ 35.0 && 0 ≤ aα[[i, j + 1]] ≤ 55), si = 0,
```

```
(100 ≤ a_α[[i, j]] ≤ 150 && -180 ≤ a_α[[i, j + 1]] ≤ 10) ∨
        (100 ≤ a_α[[i, j]] ≤ 150 && 30 ≤ a_α[[i, j + 1]] ≤ 180), s_i = 0,

       (40 ≤ a_α[[i, j]] ≤ 100 && -140 ≤ a_α[[i, j + 1]] ≤ -50) ∨
        (0 ≤ a_α[[i, j]] ≤ 40 && -140 ≤ a_α[[i, j + 1]] ≤ -105) ∨
        (35.0 ≤ a_α[[i, j]] ≤ 45.0 && -140 ≤ a_α[[i, j + 1]] ≤ -50) ∨
        (15.0 ≤ a_α[[i, j]] ≤ 40 && -110 ≤ a_α[[i, j + 1]] ≤ -40), s_i = 0.];
     i = i + 3;
     }];



  d_α = Table[s_j, {j, 902, 13500, 3}];

  m = Count[d_α, 1]; y_α = N[m / 4500];

  For[q = 500, q < 4000, q++,
    {
     For[t = q, t < (q + 85.0), t++,
      {
       corr_t = N[(d_α[[t]]) * (d_α[[q]])];
      }];
     tt_q = Table[corr_t, {t, q, q + 84.0}];
    }];
  ct_α = N[Sum[tt_q, {q, 500, 3999}]];

  }];
(* exporting the files in .dat format to be used in other graphical software
  such as origin*)
Export["corr-5nb-garcia.dat", ct_3]; Export["corr-5nb-amber.dat", ct_1];
Export["corr-5nb-96.dat", ct_2]; Export["corr-5nb-opls.dat", ct_5];
Export["corr-5nb-charmm.dat", ct_4]; Export["corr-5nb-opaa.dat", ct_6];
Export["corr-5nb-opal.dat", ct_7];
```

APPENDIX 6. CALCULATION OF HOPPING RATES

```
(*SUBROUTINE TO CALCULATE THE NUMBER OF HOPS PER ns*)

(*see the appendices above to calculate the basin occupations;
 s_i=1 for pp-II, s_i=2 for extended, s_i=3 for helix and s_i=4 for other basins;*)

(* d puts the values of s in a table, for residue 2 in a dipeptide;
 the total trajectory is 15 ns therefore the index goes upto 3000 points
  i.e. 1500 points for each residue*)

d = Table[s_j, {j, 2, 3000, 2}];

(*index p represents each ns, while q counts the number of hops in that ns*)

For[p = 1, p < 1501, p = p + 100, {
   For[q = p, q < (p + 100), q++, {
      d[[0]] = 0;
      b_(p-1) = 0;
      (*defines a hop, i.e. if the value of d in the instance before
        is not eqaul to the value of d now, it is a hop, if its the same,
       there is no hop*)

      If[(d[[q]] == d[[(q - 1)]]), b_q = 0, b_q = 1];

      b_q = b_q + b_(q-1);
      hop_p = b_q;
     }

    ];
  }];
```

```
(*x gives the table for all the hop values;
 one can use a barchart to see the hopping per ns*)

x = Table[hop_p, {p, 1, 1500, 100}];
BarChart[x, Ticks → {{1, 5, 10, 15, 20}, {0, 10, 20, 30, 40, 50, 60}}];
```

REFERENCES

Allen, M. P. T., D. J. (1987). *Computer Simulation of Liquids*, Oxford University Press, Oxford.

Alonso, D. O. V. & Daggett, V. (1998). Molecular dynamics simulations of hydrophobic collapse of ubiquitin. *Protein Science* **7,** 860-874.

Amzel, L. M. (1997). Loss of translational entropy in binding, folding, and catalysis. *Proteins* **28,** 144-9.

Anderson, H. C. (1983). Rattle: A velocity version of the shake algorithm for molecular dynamics calculations. *J. Comp. Phys.* **52,** 24-34.

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science* **181,** 223-30.

Barker, J. A. (1963). *Lattice Theories of the Liquid State*, MacMillan Press, New York, NY.

Bolhuis, P. G., Dellago, C. & Chandler, D. (2000). Reaction coordinates of biomolecular isomerization. *Proc. Natl. Acad. Sci. USA* **97,** 5877-82.

Brady, G. P. & Sharp, K. A. (1997). Energetics of cyclic dipeptide crystal packing and solvation. *Biophys. J.* **72,** 913-27.

Bryant, Z., Pande, V. S. & Rokhsar, D. S. (2000). Mechanical unfolding of a beta-hairpin using molecular dynamics. *Biophys J* **78,** 584-9.

Bryngelson, J. D., Onuchic, J. N., Socci, N. D. & Wolynes, P. G. (1995). Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins* **21,** 167-195.

Cantor, C. R. & Schimmel, P. R. (1980). *Biophysical Chemistry: Part III*, W. H. Freeman and Co., New York, NY.

Chan, H. S. & Dill, K. A. (1998). Protein folding in the landscape perspective: chevron plots and non-Arrhenius kinetics. *Proteins* **30,** 2-33.

Chen, B. L. & Schellman, J. A. (1989). Low-temperature unfolding of a mutant of phage T4 lysozyme. 1. Equilibrium studies. *Biochemistry* **28,** 685-91.

Colubri, A. & Fernandez, A. (2002). Pathway Diversity and Concertedness in Protein Folding: An ab-initio Approach. *J. Biomol. Struct. Dyn.* **19,** 739-64.

Creamer, T. P. & Rose, G. D. (1994). Alpha-helix-forming propensities in peptides and proteins. *Proteins* **19,** 85-97.

D'Aquino, J. A., Gomez, J., Hilser, V. J., Lee, K. H., Amzel, L. M. & Freire, E. (1996). The magnitude of the backbone conformational entropy change in protein folding. *Proteins* **25,** 143-56.

Deber, C. M. & Behnam, B. A. (1984). Role of membrane lipids in peptide hormone function: binding of enkephalins to micelles. *Proc Natl Acad Sci U S A* **81,** 61-5.

Despa, F. & Berry, R. S. (2001a). Inter-basin dynamics on multidimensional potential surfaces. I. Escape rates on complex basin surfaces. *Journal of Chemical Physics* **115,** 8274-8278.

Despa, F. & Berry, R. S. (2001b). Relaxation dynamics in the presence of unequally spaced attractors along the reaction coordinate. *European Physical Journal D* **16,** 55-58.

Despa, F., Fernandez, A., Berry, R. S., Levy, Y. & Jortner, J. (2003). Interbasin motion approach to dynamics of conformationally constrained peptides. *Journal of Chemical Physics* **118,** 5673-5682.

Dobson, C. M. (2001). Protein folding and its links with human disease. *Biochem Soc Symp*, 1-26.

Duan, Y. & Kollman, P. A. (1998). Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* **282,** 740-4.

Dukor, R. K. & Keiderling, T. A. (1991). Reassessment of the random coil conformation: vibrational CD study of proline oligopeptides and related polypeptides. *Biopolymers* **31,** 1747-61.

Englander, S. W. & Mayne, L. (1992). Protein folding studied using hydrogen-exchange labeling and two-dimensional NMR. *Annual Review of Biophysics & Biomolecular Structure* **21,** 243-65.

Englander, S. W., Mayne, L. C., Bai, Y. & Sosnick, T. R. (1997). Hydrogen exchange: the modern legacy of Linderstrom-Lang. *Protein Sci.* **6,** 1101-1109.

Fernandez, A., Colubri, A. & Appignanesi, G. (2001). Semiempirical prediction of protein folds. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* **64,** 021901.

Flory, P. J. (1953). *Statistical Mechanics of Chain Molecules*, Cornell University Press, Ithaca, NY.

Flory, P. J. (1969). *Statistical Mechanics of Chain Molecules*, Wiley, New York.

Garcia, A. E. & Sanbonmatsu, K. Y. (2002). Alpha-helical stabilization by side chain shielding of backbone hydrogen bonds. *Proc. Natl. Acad. Sci. USA* **99,** 2782-7.

Gilson, M. K., Given, J. A., Bush, B. L. & McCammon, J. A. (1997). The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophys. J.* **72,** 1047-69.

Go, N. (1983). *J. Stat. Phys.*, 413.

Graham, W. H., Carter, E. S., 2nd & Hicks, R. P. (1992). Conformational analysis of Met-enkephalin in both aqueous solution and in the presence of sodium dodecyl sulfate micelles using multidimensional NMR and molecular modeling. *Biopolymers* **32,** 1755-64.

Gurney, R. W. (1953). *Ionic Processes in Solution*, McGraw-Hill, New York, NY.

Holtzer, A. (1995). The "cratic correction" and related fallacies. *Biopolymers* **35,** 595-602.

Hu, H., Elstner, M. & Hermans, J. (2003). Comparison of a QM/MM force field and molecular mechanics force fields in simulations of alanine and glycine "dipeptides" (Ace-Ala-Nme and Ace-Gly-Nme) in water in relation to the problem of modeling the unfolded peptide backbone in solution. *Proteins* **50,** 451-63.

Hughes, J., Smith, T. W., Kosterlitz, H. W., Fothergill, L. A., Morgan, B. A. & Morris, H. R. (1975). Identification of two related pentapeptides from the brain with potent opiate agonist activity. *Nature* **258,** 577-80.

Jackson, S. E. & Fersht, A. R. (1991). Folding of chymotrypsin inhibitor 2. 1. Evidence for a two-state transition. *Biochemistry* **30,** 10428-10435.

Jacobson, H., Beckmann, C. O. & Stockmayer, W. H. (1950). Intramolecular reaction in polycondensations. II. Ring-chain equilibrium in polydecamethylene adipate. *J. Chem. Phys.* **18,** 1607-1612.

Jacobson, H. & Stockmayer, W. H. (1950). Intramolecular reactions and polycondensation: I. the theory of linear systems. *J. Chem. Phys.* **18,** 1600-1606.

Janin, J. (1996). For Guldberg and Waage, with love and cratic entropy. *Proteins* **24,** i-ii.

Jencks, W. P. (1975). Binding energy, specificity, and enzymic catalysis: the circe effect. *Adv. Enzymol.* **43,** 219-410.

Jencks, W. P. (1981). On the attribution and additivity of binding energies. *Proc. Natl. Acad. Sci. U S A* **78,** 4046-4050.

Jorgensen, W. L. T.-R., J. (1988). The OPLS Potential Functions for Proteins. Energy Minimizations for Crystals of Cyclic Peptides and Crambin. *J. Am. Chem. Soc.* **110,** 1657-1666.

Kaminski, G. A., Friesner, R. A., Tirado-Rives, J. & Jorgensen, W. L. (2001). Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *Journal of Physical Chemistry B* **105,** 6474-6487.

Karayiannis, N., Mavrantzas, V. G. & Theodorou, D. N. (2002). A novel Monte Carlo scheme for the rapid equilibration of atomistic model polymer systems of precisely defined molecular architecture. *Phys Rev Lett* **88,** 105503.

Kauzmann, W. (1959). Some factors in the interpretaion of protein denaturation. *Adv. Protein Chem.* **14,** 1-63.

Kollman, P. D., R.; Cornell, W.; Fox, T; Chipot, C.; Pohorille, A. (1997). *Computer Simulations of Biomolecular System 3* (van Gunstern, W. F., Wiener, P. K., Wilson, A. J., Ed.), Dordrecht.

Krantz, B. A., Mayne, L., Rumbley, J., Englander, S. W. & Sosnick, T. R. (2002). Fast and slow intermediate accumulation and the initial barrier mechanism in protein folding. *J. Mol. Biol.* **324,** 359-71.

Krantz, B. A., Moran, L. B., Kentsis, A. & Sosnick, T. R. (2000). D/H amide kinetic isotope effects reveal when hydrogen bonds form during protein folding. *Nature Struct. Biol.* **7,** 62-71.

Krantz, B. A. & Sosnick, T. R. (2001). Engineered metal binding sites map the heterogeneous folding landscape of a coiled coil. *Nature Struct. Biol.* **8,** 1042-1047.

Levinthal, C. (1968). Are there pathways for protein folding? *Extrait du Journal de Chimie Physique* **65,** 44-45.

Lifson, S. & Roig, A. (1961). On the theory of the helix-coil transion in polypeptides. *J. Chem. Phys.* **34,** 1963-1974.

Linhananta, A., Zhou, H. Y. & Zhou, Y. Q. (2002). The dual role of a loop with low loop contact distance in folding and domain swapping. *Protein Science* **11,** 1695-1701.

MacKerell, A. D., Jr., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T.,

Prodhom, B., Reiher, W. E., III, Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D. & Karplus, M. (1998). All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. In *J. Phys. Chem. B*, **Vol. 102**, pp. 3586-3616.

Mammen, M., Shakhnovich, E. I., Deutch, J. M. & Whitesides, G. M. (1998). Estimating the Entropic Cost of Self-Assembly of Multiparticle Hydrogen-Bonded Aggregates Based on the Cyanuric Acid.Melamine Lattice. *J. Org. Chem.* **63,** 3821-3830.

Martin, M. G. & Siepmann, J. I. (1998). *J. Phys. Chem. B.* **102,** 2569.

Martin, M. G. & Siepmann, J. I. (1999). *J. Phys. Chem. B.* **103,** 4508.

Mitsutake, A., Kinoshita, M., Okamoto, Y. & Hirata, F. (2000). Multicanonical algorithm combined with the RISM theory for simulating peptides in aqueous solution. *Chemical Physics Letters* **329,** 295-303.

Moran, L. B., Schneider, J. P., Kentsis, A., Reddy, G. A. & Sosnick, T. R. (1999). Transition state heterogeneity in GCN4 coiled coil folding studied by using multisite mutations and crosslinking. *Proc. Natl. Acad. Sci. USA* **96,** 10699-10704.

Mu, Y., Kosov, D. S. & Stock, G. ((in press)). Conformational Dynamics of Trialanine in Water II: Comparison of AMBER, CHARMM, GROMOS, and OPLS Force Fileds to NMR and Infrared Experiments. *J. Phys. Chem. B.*

Neria, E., Fischer, S. & Karplus, M. (1996). Simulation of activation free energies in molecular systems. *Journal of Chemical Physics* **105,** 1902-1921.

Ooi, T., Oobatake, M., Nemethy, G. & Scheraga, H. A. (1987). Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. *Proc Natl Acad Sci U S A* **84,** 3086-90.

Pande, V. S., Grosberg, A. Y. & Tanaka, T. (1997). Statistical mechanics of simple models of protein folding and design. *Biophys J* **73,** 3192-210.

Pappu, R. V. & Rose, G. D. (2002). A simple model for polyproline II structure in unfolded states of alanine-based peptides. *Protein Sci.* **10,** 2437-55.

Pappu, R. V., Srinivasan, R. & Rose, G. D. (2000). The Flory isolated-pair hypothesis is not valid for polypeptide chains: implications for protein folding. *Proc. Natl. Acad. Sci. U S A* **97,** 12565-70.

Pastor, R. W. & Karplus, M. (1988). Parametrization of the Friction Constant for Stochastic Simulations of Polymers. *Journal of Physical Chemistry* **92,** 2636-2641.

Pearlman, D., Case, D. A. , Caldwell, J. W. , Ross, W. S. , Cheatam, I. T. E. , Ferguson, D. M. , Singh, U. C. , Weiner, P. & Kollman, P. (1995). Amber 4.1.

Ponder, J. W. R., S.; Kundrot, C.; Huston, S.; Dudek, M.; Kong, Y.;Hart, R.; Hodson, M.; Pappu, R.; Mooiji, W.; Loeffler, G. (1999). TINKER: Software Tools for Molecular Design 3.7 edit. Washington University, St. Louis, MO.

Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. (1963). Stereochemistry of Polypeptide Chain Configurations. *Journal of Molecular Biology* **7,** 95-&.

Ramachandran, G. N. & Sasisekharan, V. (1968). Conformation of polypeptides and proteins. *Adv Protein Chem* **23,** 283-438.

Ramstein, J. & Lavery, R. (1988). Energetic Coupling between DNA Bending and Base Pair Opening. *Proceedings of the National Academy of Sciences of the United States of America* **85,** 7231-7235.

Raschke, T. M. & Marqusee, S. (1997). The kinetic folding intermediate of ribonuclease H resembles the acid molten globule and partially unfolded molecules detected under native conditions. *Nat. Struct. Biol.* **4,** 298-304.

Robinson, C. R. & Sauer, R. T. (1996). Equilibrium stability and sub-millisecond refolding of a designed single-chain Arc repressor. *Biochemistry* **35,** 13878-84.

Ryckaert, J. P. & Bellemans, A. (1978). *Faraday Discuss. Chem. Soc.* **66,** 95.

Scalley, M. L. & Baker, D. (1997). Protein folding kinetics exhibit an Arrhenius temperature dependence when corrected for the temperature dependence of protein stability. *Proc Natl Acad Sci U S A* **94,** 10636-40.

Schweitzer-Stenner, R. (2002). Dihedral angles of tripeptides in solution directly determined by polarized Raman and FTIR spectroscopy. *Biophys J* **83,** 523-32.

Shen, M. Y. (2002). PhD Thesis, University of Chicago.

Shen, M. Y. & Freed, K. F. (2001). Long time dynamics of met-Enkephalin: Explicit and model solvent simulations and mode-coupling theory studies. *Abstracts of Papers of the American Chemical Society* **222,** U361-U362.

Shen, M. Y. & Freed, K. F. (2002a). All-atom fast protein folding simulations: the villin headpiece. *Proteins* **49,** 439-45.

Shen, M. Y. & Freed, K. F. (2002b). Long time dynamics of met-enkephalin: Comparison of explicit and implicit solvent models. *Biophysical Journal* **82,** 1791-1808.

Shi, Z., Olson, C. A., Rose, G. D., Baldwin, R. L. & Kallenbach, N. R. (2002a). Polyproline II structure in a sequence of seven alanine residues. *Proc Natl Acad Sci U S A* **99,** 9190-5.

Shi, Z., Woody, R. W. & Kallenbach, N. R. (2002b). Is polyproline II a major backbone conformation in unfolded proteins? *Adv Protein Chem* **62,** 163-240.

Shimada, J. & Shakhnovich, E. I. (2002). The ensemble folding kinetics of protein G from an all-atom Monte Carlo simulation. *Proc Natl Acad Sci U S A* **99,** 11175-80.

Shirts, M. R. & Pande, V. S. (2001). Mathematical analysis of coupled parallel simulations. *Phys Rev Lett* **86,** 4983-7.

Sides, S. W., Curro, J., Grest, G. S., Stevens, M. J., Soddenmann, T., Habenschuss, A. & Londono, J. D. (2002). *Macromolecules* **35,** 6455.

Smith, D. & Griffin, J. F. (1978). Conformation of [Leu5]enkephalin from X-ray diffraction: features important for recognition at opiate receptor. *Science* **199,** 1214-6.

Smith, G. D. & Yoon, D. Y. (1994). *J. Chem. Phys.* **100,** 649.

Smith, G. D., Yoon, D. Y. & Matsuda, T. (1993). *J. Chem. Phys.* **98,** 10037.

Snow, C. D., Nguyen, H., Pande, V. S. & Gruebele, M. (2002). Absolute comparison of simulated and experimental protein-folding dynamics. *Nature* **420,** 102-6.

Stock, G. & Mu, Y. (2002). *J. Phys. Chem. B* **106.**

Sugita, Y., Kitao, A. & Okamoto, Y. (2000). Multidimensional replica-exchange method for free-energy calculations. *Journal of Chemical Physics* **113,** 6042-6051.

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

Tamura, A. & Privalov, P. L. (1997). The entropy cost of protein association. *J. Mol. Biol.* **273,** 1048-60.

Thompson, J. B., Hansma, H. G., Hansma, P. K. & Plaxco, K. W. (2002). The backbone conformational entropy of protein folding: experimental measures from atomic force microscopy. *J Mol Biol* **322,** 645-52.

Wang, J., Onuchic, J. & Wolynes, P. (1996). Statistics of kinetic pathways on biased rough energy landscapes with applications to protein folding. *Physical Review Letters* **76,** 4861-4864.

Wang, Y. & Kuczera, K. (1996). *J. Phys. Chem. B* **100,** 2555.

Williams, D. E. (1967). *J. Chem. Phys.* **47,** 4680.

Wolynes, P., Luthey-Schulten, Z. & Onuchic, J. (1996). Fast-folding experiments and the topography of protein folding energy landscapes. *Chem Biol* **3,** 425-32.

Woutersen, S. & Hamm, P. (2001a). Isotope-edited two-dimensional vibrational spectroscopy of trialanine in aqueous solution. *Journal of Chemical Physics* **114,** 2727-2737.

Woutersen, S. & Hamm, P. (2001b). Time-resolved two-dimensional vibrational spectroscopy of a short alpha-helix in water. *Journal of Chemical Physics* **115,** 7737-7743.

Woutersen, S., Mu, Y., Stock, G. & Hamm, P. (2001). Subpicosecond conformational dynamics of small peptides probed by two-dimensional vibrational spectroscopy. *Proc Natl Acad Sci U S A* **98,** 11254-8.

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

Yu, Y. B., Lavigne, P., Kay, C. M., Hodges, R. S. & Privalov, P. L. (1999). Contribution of Translational and Rotational Entropy to the Unfolding of a Dimeric Coiled-Coil. *J. Phys. Chem. B* **103,** 2270-2278.

Zagrovic, B., Snow, C. D., Khaliq, S., Shirts, M. R. & Pande, V. S. (2002). Native-like mean structure in the unfolded ensemble of small proteins. *J. Mol. Biol.* **323,** 153-64.

Zaman, M. H., Berry, R. S. & Sosnick, T. R. (2002). Entropic benefit of a cross-link in protein association. *Proteins* **48,** 341-51.

Zaman, M. H., Shen, M. Y., Berry, R. S. & Freed, K. F. (2003a). Computer Simulations of Met-Enkephalin using axplicit atom and united atom force-fields: Similarities, Differences and Suggestions for Improvement. *J. Phy. Chem. B* **107,** 1685-1691.

Zaman, M. H., Sosnick, T. R. & Berry, R. S. (2003b). Temperature dependence of reactions with multiple pathways. *submitted.*

Zimm, G. H. & Bragg, J. K. (1959). Theory of the phase transition between helix and random coil in polypeptide chains. *J. Chem. Phys.* **31,** 526-535.

Zwanzig, R. & Aliwadi, N. (1969). *Phys Rev* **182,** 280.