Heterogeneity: From Structure Refinement to Protein Folding

A Dissertation Presented

by

Guanglei Cui

to

The Graduate School in Partial Fulfillment of the Requirements for the

Doctor of Philosophy

in

Chemistry

Stony Brook University

December 2003

Stony Brook University

The Graduate School

Guanglei Cui

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

Carlos Simmerling, Ph. D., Advisor

Stony Brook University

Daniel Raleigh, Ph. D., Chairperson

Stony Brook University

Benjamin Chu, Ph. D., Third Member

Stony Brook University

Wendy Cornell, Ph. D., Outside Member

The Novartis Institutes for Biomedical Research, Novartis, Inc.

This dissertation is accepted by the Graduate School

Dean of the Graduate School

November 2003

Abstract of the Dissertation

Heterogeneity: From Structural Refinement to Protein Folding

by

Guanglei Cui

Doctor of Philosophy

in

Chemistry

Stony Brook University

2003

Structure determines everything.

Biomolecules, such as proteins and nucleic acids, are heteropolymers, the repeating units of which are chosen from a finite set of simple organic molecules that share certain physiochemical properties. Biomolecules are unique molecules because their structure is hierarchical and also because low level structure (their sequences) identifies high level structure, their accessible three dimensional conformations. This is often referred to as conformational heterogeneity. However, this is not all that makes biomolecules special. Kinetic processes of a polymeric molecular system can often be heterogeneous as well (kinetic heterogeneity), which again can depend on the sequence. The comprehension of the details of these heterogeneities is critical to understanding functions and mechanisms of biological processes.

In this dissertation, both conformational and kinetic heterogeneities were addressed computationally. First, the local conformational heterogeneity of a modified DNA duplex was investigated with molecular dynamics and the locally enhanced sampling technique. Calculations were compared to those of a regular

DNA duplex with similar sequence. The combined results indicated that the conformational abnormality of an adenine base was introduced by the nearby modified “base pair”. This conformational abnormality has been long recognized as a major drawback to conventional structure determination approaches, such as NMR spectroscopy and X-ray crystallography. Computer modeling was proven to be a valuable tool in this study and provided a logical explanation that was missing in the NMR studies.

Next, the kinetic heterogeneity of folding for a model peptide was thoroughly examined with an ensemble of folding simulations. Three different timescales of folding were found, covering a wide range from tens of picoseconds to tens of nanoseconds. This complicated folding scenario was then justified by subsequent thermodynamic studies with the replica exchange method, from which the free energy landscape of folding was generated. The welding of kinetic and thermodynamic findings solidifies our understanding from simple lattice simulations that protein folding is in general a very heterogeneous process, which may degrade into a simple biphasic behavior under certain circumstances.

At last, the inhibition of the E. Coli enoyl-reductase – FabI by triclosan and its analogs were quantitatively calculated using free energy calculation techniques, as our first step towards new inhibitor design for the M. Tuberculosis enoyl-reductase – InhA. Molecular dynamic studies of the FabI:NAD$^+$:ligand complex revealed that the tetrameric interaction and ligand binding may be closely related. Based on this, a truncated model system was created for the free energy calculations, which reasonably represented the original tetramer system. However, a decent agreement with the experimental relative binding affinities was still difficult to achieve even after taking into account the protonation state of the bound ligand. The uncertainty of the binding needs further investigation.

To Xixi and my loving parents

Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank my advisor, Prof. Carlos Simmerling, for opening the door to a new field for me, for sharing his experience and expertise within and beyond scientific research, for four years of constant support and inspiration, and for always being patient to such an impatient student like me.

I would like to thank the other committee members, Prof. Daniel Raleigh and Prof. Benjamin Chu as well, especially Prof. Daniel Raleigh who was always there when I needed advice. I am also greatly obliged to Dr. Wendy Cornell who not only took on the role of the outside member but also helped me with postdoc position search.

Many thanks to my dear group members, Dr. Viktor Hornak, Bently Strockbine, Asim Okur, Xiaolin Cheng, Raphel Geney, Dan Roe, Salma Rafi, Kun Song, Lauren Wickstrom, Kerri Goldgraben, and Melinda Layten. This is absolutely the best people that one could dream of. I am so fortunate to have spent so much wonderful time with such a group of great colleagues and friends.

It is Viktor that shared his small office with me in the early days, taught me everything about running simulations, introduced me to the great world of Linux, withstood all my questions, and pointed me better solutions. I could not even imagine how I could go through my Ph. D. study without his help.

Thanks to all my Warcraft buddies, Ben, Asim, Raph, Dan, Kun, Melinda, and Kevin who have been not only supporting each other in their research projects but also bringing so much fun to the group. I enjoyed every moment that we were together.

Thanks to all the dear friends from nearby groups, Dr. Karsten Theis, Margaret Luk-Paszyc, Song Xiang, Nils Schrader, and Katrin Fischer. There are so much to remember, afternoon volleyball break, movie nights, green cactus. Without all these, my graduate student life would have been so boring.

Thanks to Xiaolin for many discussions we have had and ideas that were born afterwards. I wish the best for you and your family.

Thanks to Prof. Peter Tonge and Dawn Marin for all the trust in my research even when the calculations did not work as well as planned.

Thanks to my dear Xixi who has been always with me over the past two years, which is the greatest thing that has ever happened to me. It is your smiles and tears that awakened me and stopped me from drifting away and getting lost.

Finally, thanks to my loving parents to whom I owe so much that I could not possibly give back. Your wordless understanding and blessing always stay deep in my heart.

Chapter 1

Introduction

1.1 Structural Biology Overview

Structural biology, which provides structural information of biomolecular systems so as to better understand or explain what is observed in molecular biology, has marched a long journey over the past hundred years, from watching cells in the earliest days of microscopy in the late 17th century to the first atomic-detailed picture of ribosome, consisting of 57 different molecules (3 rRNAs and 54 proteins), and become one of the most important facets of current biological research. A detailed overview of the history of the whole field can be found in Reference [1].

In principle, three-dimensional structures of bio-molecular systems can be studied by theoretical calculations if the underlying physics of how atoms interact is understood adequately. Such an attempt was first made by McCammon and Karplus in the late 1970's, which was based on a classical description (Molecular Mechanics) of atomic interactions [2, 3, 4], and so was opened a door to a brand-new field – computational structural biology. After this groundbreaking work, molecular modeling has been constantly developed over the past 30 years by many (for recent reviews, see [5, 6, 7]) and become one of the essential approaches in structure studies, next to X-ray/neutron crystallography, electron microscopy (EM) and nuclear magnetic resonance (NMR).

One of the early mis-concepts in structure biology is that proteins have relatively rigid structures. But it was quickly discovered that proteins as well as other bio-macromolecules exhibit, more or less, conformational flexibilities that are crucial to their biological functions. The importance of their internal motions or conformation complexity has been illustrated both computationally and experimentally in many studies. One recent example attempted to address the antibody multi-specificity issue by both X-ray crystallography and pre-steady-state kinetics and proposed an preexisting equilibrium between different antibody conformers [8]. Resolving different conformations and evaluating their contributions to the function are of high-priority, and yet the most challenging.

X-ray crystallography is central in our structure determination process and makes the largest contribution to all structures solved so far (85% of the 20,868 structures deposited in Protein Data Bank as of May 2003). However, it is a well-known fact that X-ray crystallography generally only depicts an averaged picture of molecular system that is in ceaseless motion. Investigating the influence of different conformers using X-ray crystallography needs not only passions and patience, but also some blessing. On the other hand, NMR spectroscopy has proved itself in this regard a better choice of tool when there exists mobile regions in the studied system. Although NMR studies are often limited by the system size, its great capability of detecting structural changes makes it a very good complement to X-ray crystallography.

In parallel to the advancements in experimental techniques, significant progress has been achieved in theoretical calculation of biomolecular systems during the past 30 years. Equipped with refined physical models, improved algorithms, and faster computers, "modelers" are now no longer satisfied with being able to reproduce results that are handed over, but eagerly take on to the next stage, making quantitative predictions and building the groundwork for new theories and experiments. Comparing to the above-mentioned traditional tools in structural biology, computer simulations, especially molecular dynamics (MD), have certain advantage in studying conformational complexity because it can, in principle, provide the most direct examination of motional phenomena with ultimate details. Therefore, when experimental approaches are limited as to the information that can be obtained, simulation methods can often become helpful to supply the missing part when the simulation is indicated meaningful.

1.2 Structural Biology *In Silico*

Computational chemistry has achieved a huge success in applying quantum mechanics (QM) to studying small molecule organic reactions and spectroscopy. This work won the Nobel Prize in Chemistry in 1998 for Walter Kohn and John Pople. Full quantum mechanical treatment of biological molecular systems that contains thousands of atoms is possible theoretically, but remains impractical for large systems due to the acute demanding on computer resources. Even though whole protein minimization recently became possible with semiempirical approaches [9], classical molecular mechanics (MM) is by far the only practical way to study conformational changes in a large scale with the premise of an accurate and adequate description of key interactions, or force field.

1.2.1 Force Field

The importance of an accurate description of key interactions is beyond all doubt. However, only recently did its profoundness become realized. Most of modern descriptions have the empirical form of a sum of several terms, characterizing bonded (such as bond stretching, bending and rotation) and non-bonded interactions (Coulombic and van der Waals contributions). Each term is a function of system coordinates with certain adjustable attributes, which are primarily determined by numerical fitting to equilibrium physical properties. For example, van der Waals interaction strength and range are obtained by parameter tuning to reproduce thermodynamic properties of various pure liquids through simulations and atomic partial charges are calibrated to give reasonable electrostatic potential for a few selected conformations calculated using quantum mechanics. Significant efforts have been made in force field development over the past 20 years [10, 11, 12], the criticality of which makes it a high-risk and high-return research direction in the whole field[1]. At present, MD simulations can be routinely carried out for proteins and nucleic acids and tend to maintain a reasonable agreement with experimental structures in the nanosecond range.

As molecular dynamics techniques come of age, applications become more resource-demanding and more focused on conformational flexibilities, such as in low-resolution structure refinement, antibody loop conformation prediction, and protein folding, etc. The hidden side of available force fields and long-used parameterization philosophy begins to expose slowly, in part due to the fact that only the most probable states or conformations are used during parameterization, which does not guarantee a correct description of other states that might be important for conformation morphing. One example is the observed over-stabilization of helical conformations with AMBER and CHARMm force fields. Although its origin and solution are still not clear, several attempts have been made to rectify this by more careful backbone torsion parameter fitting[13, 14] or a numerical lookup in a pre-calculated correction table [15] and how well they perform are being examined. The other important part of current classical MM force field that needs to

[1] Papers describing force field development are among the best cited articles in literature

be improved is possibly the fixed-charge model in electrostatic calculation, which may be a questionable approximation in some cases. Although polarizable force field are in active development, the additional computational cost introduced may mitigate the potential advantage of better calculation, considering that force field is by no means the only improvement that is very much needed. Due to the empirical nature of MM force field, a force field that is highly transferable and highly effective may take a long time to come. At present, it may have to be accepted that problem-fitted force field, which targets a particular problem, is a better trade-off.

1.2.2 Conformational Sampling

The other aspect that cannot be emphasized enough is how to handle the conformational complexity, which is created by the enormous number of degrees of freedom of biomolecular systems. This inherent conformational complexity is directly linked to the multiplicity of maxima (transition states) and minima (well-defined thermodynamic states) on potential energy surface. Monte Carlo (MC) and molecular dynamics (MD) are the most commonly used approaches to produce low-energy conformations that are subject to predefined constraints (temperature, pressure, volume, energy, etc.). In general, Monte Carlo approach directly manipulates internal degrees of freedom and therefore is considered as the most convenient way to stochastically generate conformations under a set of given constraints; on the other hand, molecular dynamics approach calculates the time-dependent chain of conformations that are determined by equations of motion (such as Newtonian). Because of this, molecular dynamics approach is often preferred, though arguably, in studying kinetics-related problems, such as protein folding.

In the process of understanding structure-function relationship, the most frequently encountered question is the relative stabilities of different thermodynamic states, which are separated by energy barriers. Experimentally, this is evaluated by measuring the populations of these states characterized by one or two geometric features (typically referred as “reaction coordinates”) that can be easily detected. Different thermodynamic states can be identified in Monte Carlo or molecular dynamics simulations using the same reaction coordinates, but more directly. However, obtaining a well converged distribution of states can be very difficult, especially when the separating barriers are high comparing to thermal-fluctuations. The typical length of molecular dynamics simulation is usually not adequate to obtain quantitative results when the transition timescale is much longer, e.g. microseconds in the folding of small proteins, even though simulation length are coming closer.

Advanced Sampling Techniques

Various techniques have been invented to enhance the sampling efficiency of poorly populated thermodynamic states. According to Boltzmann law of distribution, the probability of being in a particular state depends on the temperature (T) and the Hamiltonian of the system (H).

$$P(H, T) \sim \exp\left(-\frac{H}{kT}\right)$$

In fact, both H and T can be manipulated to favor the sampling of low probability states. However, Boltzmann distribution may not be preserved and special weighting correction might be needed to obtain relevant thermodynamics. Several sampling techniques are described below.

High Temperature Simulations The simplest way to boost the sampling. MD or MC simulations are performed at elevated temperatures, at which barrier crossover occurs more frequently because of the increased kinetic energy that molecules possess. It is a very common practice in studying conformational transitions that do not take place easily at room temperature. Meaningful understanding can be accessible assuming raising temperature does not alter kinetics in a significant way.

Simulated Annealing A very popular optimization technique that is theoretically guaranteed to locate the global minimum regardless the complexity of the phase space. Simulations are carried out for certain duration at elevated temperature, which is then decreased in a pre-programmed way until the target temperature is reached [16].

Targeted or Steered MD The Hamiltonian of the system is modified with an extra term, which corresponds to certain kinds of driving force. Target MD (TMD, [17]) involves the selection of a reference state, a target state and a reaction coordinate that are used to define the driving force, which often takes the harmonic form. In Steered MD [18], the driving force is defined using the sampling history, e.g., the average experienced forces of atoms.

Locally Enhanced Sampling Sampling is focused on parts of the system that are more interesting, e.g., antibody loops or a particular mobile region. The Hamiltonian of this "hot" area is modified so that the interaction with the rest of the system is down-scaled by a factor of N, but the total Hamiltonian of the system is kept unchanged by duplicating the "hot" area N times. Because of weakened interactions, transition barrier of each copy is only one Nth of the original system. The original thermodynamics is ensured unaltered when all duplicated areas adopt the same conformation [19]. More detailed description is given later.

Umbrella Sampling A Hamiltonian-modifying method used to obtain potential of mean force (PMF). When transitions between two different states can be identified by a characteristic geometric feature (reaction coordinate), sampling along this coordinate can often be enforced by applying a biasing term to the system Hamiltonian, similar to target MD [20, 21, 22]. In fact, TMD becomes umbrella sampling when a series of intermediate targets are inserted along the transition pathway. The samplings collected at each intermediate step are combined with reweighting to give appropriate Boltzmann distribution.

Generalized Ensemble Approaches A large class of methods recently introduced to biological system studies, including multicanonical algorithm (MUCA, also referred as entropic sampling, and adaptive umbrella sampling), simulated tempering (ST), 1/k-sampling, Tsallis sampling, replica-exchange method (REM) and its variant REM/MUCA (review, see [23]). In general, non-Boltzmann sampling is performed so that each state is visited with equal probability regardless its potential energy. Weighted histogram analysis method (WHAM, [24]) is often used to reconstruct a proper canonical ensemble average. The replica exchange method is to be further discussed in Chapter 3.

Improved sampling can also be achieved from using reduced physical models with less detailed potentials. What has become very popular in recent years is to replace explicit solvent model with continuum solvent model, avoiding the calculation of the evolvement of water molecules and approximating electrostatic screening effect of solvent with theoretical treatments, such as distance-dependent dielectric models (DDD), Poisson-Boltzmann model (PB), and Generalized-Born model (GB), etc. These models can often estimate reasonably well the electrostatic contribution to the solvation energy, and non-electrostatic part is usually approximated by a surface-area (SA) term. Continuum solvent model, especially GB, has gained quite some popularity (a recent review, see [25]) over the past few years largely due to its reasonable reproduction of solvation free energy [26] and fast calculation speed with respect to more expensive PB and explicit solvent model. However, GB tends to perform poorly in the cases of large systems with buried charges. Simulations of proteins using GB model appear not as stable as explicit solvent simulations of comparable length, but using explicit solvent simulations as a stability benchmark for GB simulations may not be a rigorous approach since explicit solvent simulations are known to suffer from inadequate sampling. In addition, questions on force field accuracy start to surface when longer simulations can be afforded with GB solvent model. Nevertheless, GB and other continuum solvent models are not bad choice when applied cautiously. In fact, they have been widely used in many conformational energy analysis methods and peptide folding studies, providing a fast and approximate way to estimate solvation free energy.

1.3 Outlines of Research Projects

This dissertation summarizes the results of three independent studies that mainly focused on the heterogeneities of biological molecular systems. The conformational complexities investigated here include local structure refinement of a double-strand DNA molecule, thermodynamics and kinetics of folding of several small peptides, and the calibration of a working model system for an enzyme-inhibitor binding complex. Several computational methods, e.g. molecular dynamics with GB and explicit solvent models, locally enhanced sampling, replica-exchange method, and thermodynamic integration, were prudently exercised to meet the posed challenges during the studies, their effectiveness carefully compared. A short summary of each study is given below.

1.3.1 Local Structure Refinement of a Double-Strand DNA

The conformational ambiguity of one base in a modified double-strand DNA was examined in MD simulations. NMR experiments conducted on this DNA with an incorporated synthetic base (pyrene), bearing a similar volume of a regular base pair, failed to resolve the conformation of a neighboring adenine. Both *anti* and *syn* conformers resulted from NOE-based structure calculations. In MD simulations carried out with both GB and explicit solvent models, several different NMR structures were used to avoid a biased conclusion. Both *anti* and *syn* conformers seemed to be stable in regular molecular dynamics, regardless of the solvent models used. However, direct evaluating the relative thermodynamic stability of the two conformers was impeded because the *syn/anti* interconversion, a seemingly easy local rearrangement that involves one base pair flipping, was barely observed in regular MD simulations. Particularly interesting is the observation of large structural changes that led to different conformers from the same initial structure, depending on the solvent model used. The reversible *syn/anti* interconversion was only available from the locally enhanced sampling simulations, allowing the assessment of thermodynamics and a plausible explanation to be given.

1.3.2 Peptide Folding Studies

The thermodynamics and kinetics of folding were studied for several short peptides with stable and identifiable native states. Peptides are often used as model systems for studying protein folding, which is somewhat controversial among protein folding community. Peptide is certainly much simpler and cannot adopt very sophisticated topologies unlike protein. Therefore, the influence of topology on folding cannot be captured by peptide systems. But peptide and protein follow the same underlying physics, the interplay between thermodynamics and kinetics can still be valuable in general to validate our concepts and understandings. One system studied here, a nonapeptide fragment from influenza hemagglutinin Ha1, is a very good example in this regard. A unified study of thermodynamics and kinetics was achieved with all atomic simulations in GB solvent model. The folding takes place in three very different timescales (100ps, 1ns and 100ns), which indicates that at least three different folding mechanisms may exist. This was further confirmed by the free energy landscape of folding constructed from replica-exchange simulations and principal component analysis (PCA). Additionally, folding process was found different even within the same timescale, depending on the conformation of certain particular residues, which is consistent with kinetic partitioning theory proposed by Thirumalai et. al. and a number of experimental findings. Full-length analysis and discussion on some unsolved issues in studying protein folding can be found in Chapter 2.

1.3.3 Understanding Enzyme-Inhibitor Binding

Quantitative calculation of relative receptor-ligand binding affinity is another big challenge to computational structure biology and attracts huge interests from pharmaceutical industry. In this project, attempts have been made to obtain quantitative agreement with relative experimental binding affinities of FabI (the enoyl-acyl reductase of E. coli) and triclosan analogs through thermodynamic integration (TI), one of the standard free energy calculation methods available, aiming at validating the model system and paving the way for the inhibitor design of InhA (the enoyl-acyl reductase of Mycobacterium tuberculosis, highly homologous to FabI). Preliminary calculation results indicate that the tetrameric form of FabI might be important for the ligand-binding. Free energy calculations carried out also showed that the binding affinity may depend on the protonation state of the ligands, which suggest pH-dependency assay of binding to conducted.

Chapter 2

Conformational Heterogeneity Observed in Simulations of a Pyrene-Substituted DNA

2.1 Introduction

2.1.1 Biological Background

Pyrene DNA is a non-conventional DNA motif, which contains a bulky “base”, deoxypyrene nucleotide, paired with an abasic site (Figure 2.1). The pyrene is roughly equivalent in size to a traditional base pair. Observations that it is selectively assembled into a duplex DNA opposite abasic sites demonstrate that the Watson-Crick hydrogen bond pattern may not be required to entail the fidelity of DNA replication [27] and suggest a strategy for characterization of abasic lesions in damaged DNA. The structure of this unconventional DNA duplex has been studied using NMR methods [28].

This particular system attracted our interest not only due to the importance of the system in contributing to the understanding of nucleic acid structure, but also

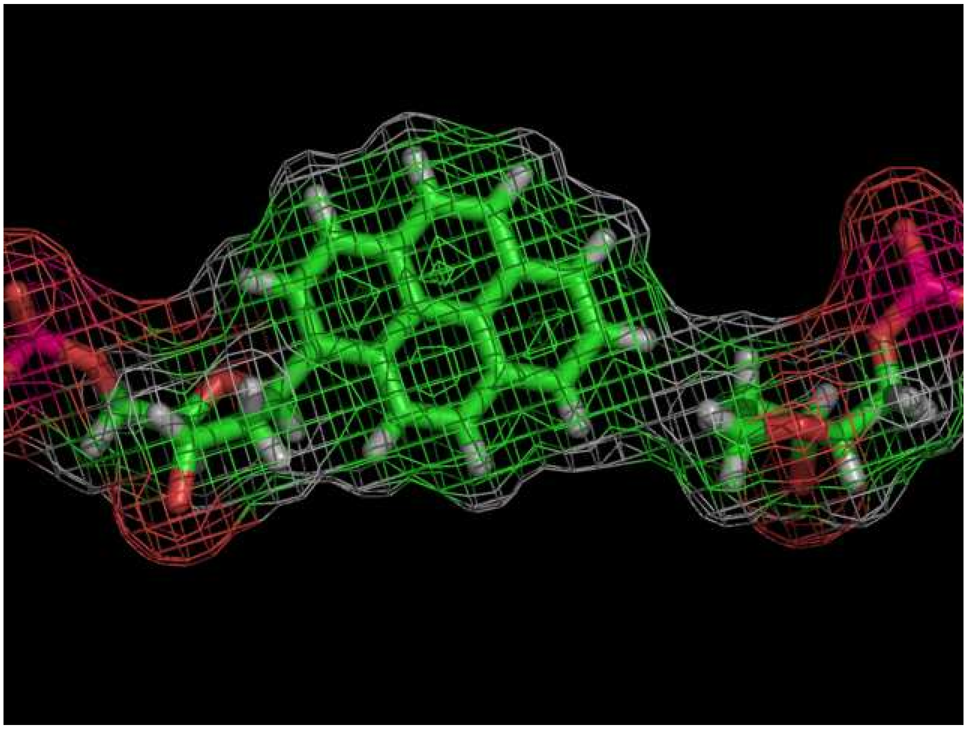

Figure 2.1: A pyrene - tetrahydrofuran “base pair”, shown with a solvent-accessible surface. This pair is nearly as large as a standard base pair.

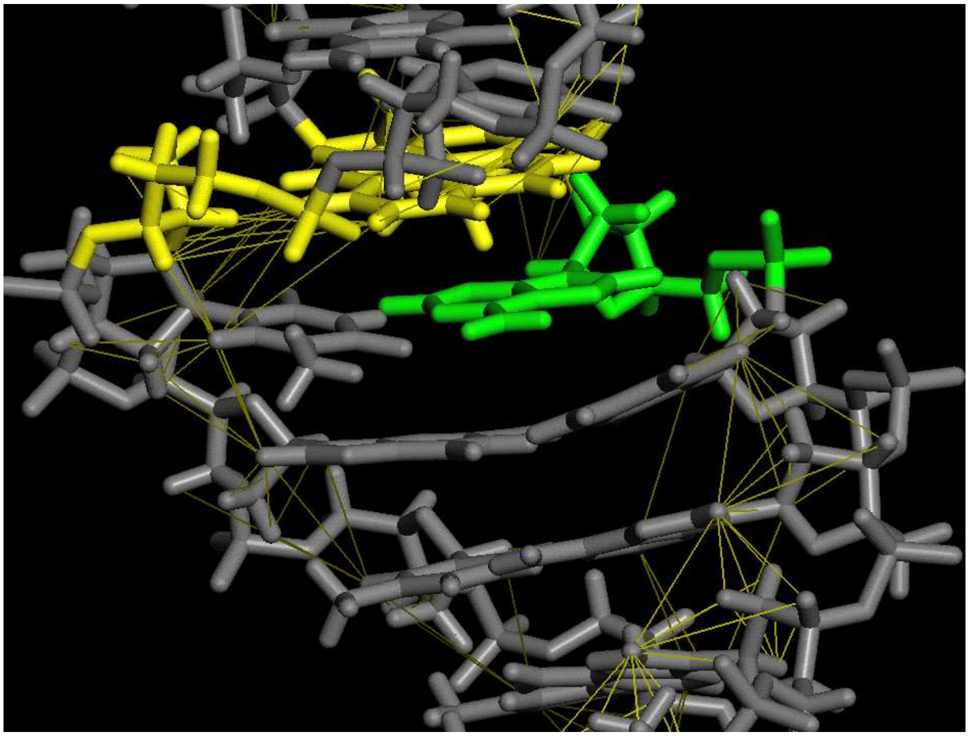

Figure 2.2: A member of the family of NMR-derived structures, with the NOE restraints shown as yellow lines. Adenine 8 (green), 5′ to the pyrene (paired with abasic furan, both in yellow), has no NOEs involving the base.

because of a lack of potentially important structural information. The structural ensemble derived from the NMR data failed to reach agreement on the conformation of adenine 8 (ADE8), located at the 5′-end of the pyrene "base", due to the lack of NOE information involving that base (Figure 2.2). In fact, a wide variety of conformations for ADE8 are present in the family of structures, including both *anti* and *syn* ADE8 conformations, with ADE8 often observed in the major or minor groove without hydrogen bonding to the partner THY19. However, observed chemical shift data suggests that the THY19 imino proton is involved in hydrogen bonding interactions. Thus, while the NMR data provides a detailed view of the conformation of the pyrene, the influence of this motif on the local structure and dynamics of the DNA remains unclear.

To contribute to the understanding of the influence of pyrene on DNA structure, and also to investigate the possibility of using GB solvent model in biomolecular structure refinement, a series of molecular dynamics simulations were carried out for the pyrene-DNA system.

2.1.2 Nucleic Acid Simulations

Molecular dynamics simulations of nucleic acids used to be considered very difficult comparing to those of proteins, largely due to the highly charged backbone phosphate that requires accurate treatment for electrostatic interactions. When not available, the DNA double helical structure can be easily distorted. Charge scaling or base pair restraints used to be typical practice of nucleic acid simulations. The situation did not change [29] until Ewald summation method [30] was introduced, which is a technique for evaluating electrostatic potential of a lattice of point charges, subject to periodic boundary conditions. With fast Ewald summation methods [31] and explicit solvent model, unrestrained simulations have successfully modeled structural changes, such as the conversion between A-DNA and B-DNA [29, 32, 33, 34, 35], the description of sequence-dependent DNA structural properties [36, 37], the influence of abasic sites on the twisting and bending of DNA [38, 39, 40], and individual base-pair "breathing" events [41, 42]. Recently Generalized-Born solvent model has also been applied in nucleic acid simulations and appeared to be a promising alternative approach [43].

2.1.3 Generalized-Born Solvent Model

Generalized-Born solvent model calculates the electrostatic contribution to the free energy of solvation G_{pol}. The model comprises a system of particles with radii a_i and charges q_i. The energy of a point charge in its reaction field is one half of the product of the charge and the reaction potential,

$$G_{ion} = \frac{1}{2}q\phi_{reaction} = -\frac{q^2}{8\pi a}\left(\frac{1}{\varepsilon_0} - \frac{1}{\varepsilon}\right)$$

where $\varepsilon_0$and ε are the permittivity of the ion and the medium around it. This equation is known as the Born equation. For a system with N arbitrary point charges, this free energy of solvation can be approximately expressed as

$$G_{pol} = -\frac{1}{8\pi}\left(\frac{1}{\varepsilon_0} - \frac{1}{\varepsilon}\right)\sum_{i,j=1}^{N}\frac{q_i q_j}{f_{GB}}$$

where

$$f_{GB} = \sqrt{r_{ij}^2 + a_{ij}^2 e^{-D}},\ D = r_{ij}^2/\left(2a_{ij}\right)^2,\ a_{ij} = \sqrt{a_i a_j}$$

proposed by Still and coworkers [44]. This functional form has been used with considerable success to efficiently evaluate solvation free energies of small molecules.

Its application to biological systems only came in recent years because of the computational cost of using explicit solvent condition, especially with large systems. For example, roughly 3000 water molecules were used in the simulation of 13-base pair DNA duplex, which is to be described below. It usually takes a few days to simulate one nanosecond (ns) of molecular motion with multiple fast CPUs, which is rather "useless" considering that most interesting conformational transitions occur in the range of sub-microseconds to milliseconds. Using implicit solvent model like GB reduces simulation cost significantly (3 to 5 times) and increases sampling efficiency in several different ways: it removes the computational expense of calculating the motion of explicit solvent molecules, directly provides a solvation free energy without the need for averaging over solvent configurations, and the lack of solvent friction can accelerate escape from local energy minima during dynamics simulations. This makes it attractive for the study of conformational changes and structural refinement when electrostatic interactions are dominant, such as in the case of nucleic acids. Although limited success with proteins has been reported [45], the utility of GB solvation has been demonstrated on nucleic acid systems [46, 47, 48, 25, 49].

2.1.4 Locally Enhanced Sampling

Locally Enhanced Sampling (LES) is a mean-field approach based on time-dependent Hartree approximation, which allows the regions of interest to be sampled more extensively than the region of less interest [19]. In Hartree approximation, the system is divided into J subsystems. Assuming their motions are independent of each other, the probability density of the original system is replaced by the product of probability densities of subsystems,

$$P(X) = \prod_{j=1}^{J} p_j(X_j)$$

In LES, subsystems of high interests, for example, an antibody loop, can be sampled with multiple copies that do not interact within the copied subsystems. Non-copied subsystems interact with the subsystems in an average way. The effective energy of the system is defined as

$$U = U_{non-LES} + \frac{1}{J}\sum_{j=1}^{J} U_j$$

To keep system energy unchanged, the interaction of copied subsystems with the rest need to be scaled accordingly. As a result of this, transition barriers are lowered for any individual copies and low probability states become more accessible. Moreover, the global energy minimum of the LES system is identical to that of the original system[1]. This enhanced sampling has been demonstrated in a few recent publications [50, 51, 52].

2.2 System and Calculation Setup

The system studied (PDB code 1FZL) consists of 13 base pairs (5′-D(Cp-Ap-Cp-Ap-Ap-Ap-Cp-Ap-(PYP)p-Gp-Cp-Ap-C)-3′ and 5′-D(Gp-Tp-Gp-Cp-(FUR)p-Tp-Gp-Tp-Tp-Tp-Gp-Tp-G)-3′), with 4 and 8 base pairs on each side of the pyrene. In the remainder of this chapter we refer to these as the short and long ends, respectively, and refer to the system as "pyrene DNA". Root mean square deviations (RMSD) were calculated based on all heavy atoms of the 9 central base pairs, since the terminal 2 base pairs on each end showed significant fluctuation among the family of NMR-based structures and during simulations. Groove widths were calculated as the distance between the closest phosphorus pairs across the groove. Another sequence was simulated as a non-pyrene control with 10 base pairs: (5′-D(Cp-Cp-Ap-Ap-Cp-Gp-Tp-Tp-Gp-G)-3′ and 5′-D(Cp-Cp-Ap-Ap-Cp-Gp-Tp-Tp-Gp-G)-3′). This will be referred to as "standard DNA", since it is composed entirely of the 4 standard bases (A, T, G, and C). Simulations for this control sequence have been reported using the same force field as employed in the present study. All molecular dynamics simulations presented in this study were performed using the sander module in AMBER version 6 [53].

2.2.1 Initial Structures

The family of structures resulting from refinement with NMR-derived restraints shows significant diversity in the conformation of ADE8 (Figure 2.3). Four representative structures of the family were selected, corresponding to all four possible

[1] Therefore, every copy of the LES system will have the global minimum conformation when it is located.

conformations and subject to MD calculations described in the following sections:

antiHB ADE8, in *anti* conformation, forms canonical Watson-Crick hydrogen bonds with THY19 with pyrene stacked on the 3′ side.

antiMinor ADE8, in *anti* conformation, resides in the minor groove with no hydrogen bonds formed.

syn1 ADE8, in *syn* conformation, resides in the minor groove with no hydrogen bonds formed.

syn2 ADE, in *syn* conformation, forms Hoogsteen hydrogen bonds with THY19, pyrene stacking from the 3′ side.

2.2.2 Force Field Parameters

Partial charges for the pyrene and tetrahydrofuran analog nucleotides were obtained using the restrained electrostatic potential fitting method (RESP, [54, 55]), in a manner similar to that employed for the standard DNA residues in the ff94 force field [11, 55]. Each nucleotide was excised from the NMR duplex structure and optimized in Gaussian98 [56] (HF/6-31G*). The electrostatic potential was then calculated and used in a two-stage RESP fit. The partial charges of all atoms, except the bases, C1′ and H1′ on the sugar ring, were held fixed at their values in the ff94 force field. Multiple conformations were not used in the charge derivation for both nucleotides, considering the limited flexibility and highly conjugated structure of pyrene, and the simplicity of tetrahydrofuran. Assignment of atom types and missing bond angle and torsion angle parameters were made by analogy to existing atom types in the ff94 parameter set. The resulting partial charges, atom types and additional parameters are listed in Table 2.1.

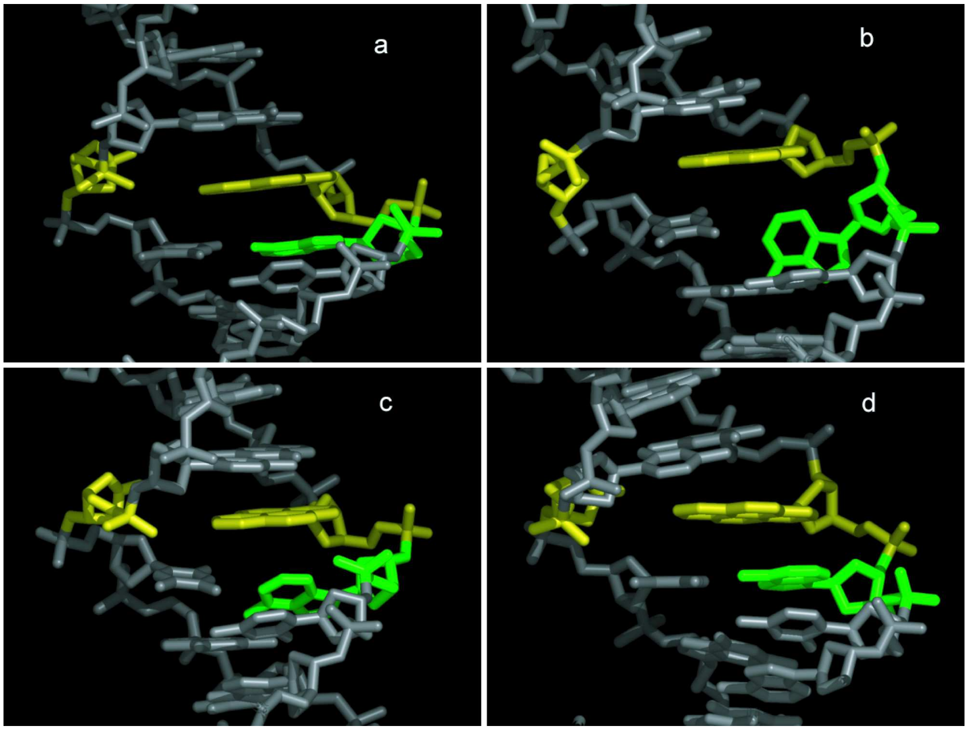

Figure 2.3: Closeup of the region near the pyrene in several structures from the refined NMR family that were used in the simulations: a) antiHB, b) antiMinor, c) syn1, d) syn2.

Atom Name	Atom Type	Partial Charge	Atom Name	Atom Type	Partial Charge
Pyrene			C11	CA	0.0731
P	P	1.1659	C12	CA	-0.0019
O1P	O2	-0.7761	C13	CA	0.0036
O2P	O2	-0.7761	C14	CA	0.0364
O5′	OS	-0.4954	C15	CA	0.0523
C5′	CT	-0.0069	C16	CA	0.1176
H5′1	H1	0.0754	C2′	CT	-0.0854
H5′2	H1	0.0754	H2′1	HC	0.0718
C4′	CT	0.1629	H2′2	HC	0.0718
H4′	H1	0.1176	C3′	CT	0.0713
O4′	OS	-0.3691	H3′	H1	0.0985
C1′	CT	-0.2056	O3′	OS	-0.5232
H1′	H1	0.2043	Tetrahydrofuran		
C1	CA	-0.1420	P	P	1.1659
H1	HA	0.1712	O1P	O2	-0.7761
C2	CA	-0.1694	O2P	O2	-0.7761
H2	HA	0.1367	O5′	OS	-0.4954
C3	CA	-0.1467	C5′	CT	-0.0069
H3	HA	0.1425	H5′1	H1	0.0754
C4	CA	-0.2208	H5′2	H1	0.0754
H4	HA	0.1564	C4′	CT	0.1629
C5	CA	-0.1241	H4′	H1	0.1176
H5	HA	0.1401	O4′	OS	-0.3691
C6	CA	-0.2348	C1′	CT	-0.2619
H6	HA	0.1610	H1′	H1	0.1917
C7	CA	-0.1396	H1″	H1	0.1917
H7	HA	0.1456	C2′	CT	-0.0854
C8	CA	-0.2185	H2′1	HC	0.0718
H8	HA	0.1455	H2′2	HC	0.0718
C9	CA	-0.2239	C3′	CT	0.0713
H9	HA	0.1867	H3′	H1	0.0985
C10	CA	0.0759	O3′	OS	-0.5232

Table 2.1: Force Field parameters for pyrene and tetrahydrofuran moieties

2.2.3 Simulation Protocol

Each MD simulation consisted of two stages, equilibration and production, neither of which used the experimentally determined NOE distance restraints. All MD simulations in the equilibration stage used one femtosecond time step size to ensure better stability, which could be harmed otherwise by steric clashes in poor quality structures. The time step size was doubled in the production stage. In the production stage, structure snapshots were saved every 10 picoseconds for subsequent analysis.

2.2.3.1 Explicit Solvent Simulations

The particle mesh Ewald (PME) method [31] was used in all explicit solvent simulations to evaluate electrostatic interactions. The default parameter values in AMBER were used in PME calculations (8Å cutoff for the real-space nonbonded interactions, and a reciprocal space grid spacing of approximately 1Å). The NMR structures were solvated in a roughly $55\times70\times50\text{Å}^3$ box of TIP3P [57] water molecules with a clearance of at least 9Å between the DNA atoms and each side of the box. The number of water molecules required to solvate each of the systems varied, with total system sizes of 12,000-14,000 atoms. Sodium ions were added to neutralize the system. All bonds with hydrogen atoms involved were constrained with SHAKE. Rigid body motion of the system as a whole (not just the solute) was removed [58]. First, 60 picoseconds of dynamics was carried out at 300K and 1 atm pressure, in which only water molecules and counterions were allowed to move. The whole system was then minimized for 5000 cycles with incrementally reduced positional restraints. Four 10 picosecond restrained MD simulations allowed both solvent and solute to reach local equilibrium by carefully releasing the positional restraints imposed to zero.

2.2.3.2 GB Simulations

An implementation of GB solvent model [59, 60] has been incorporated into AMBER version 6, and testing of this model has been carried out for DNA systems [49]. All of the GB simulations described in this dissertation use a similar approach as that study, employing a modified set of Bondi radii [61] (only hydrogen atoms are modified) for all bases, screening parameters taken from the Tinker program [62] and 0.13Å offset for the effective Born radii. Explicit counterions were not used; an ionic strength of 0.2M was employed in the continuum solvent calculations. SHAKE was used in all dynamics calculations to fix the length of covalent bonds in which hydrogen atoms were involved. Non-bonded interactions were fully evaluated every time step with no cutoff distance used.

In the equilibration stage, the starting structures were first minimized for 1000 cycles, with atoms restrained to the starting positions with a harmonic force constant 5.0 kcal/(mol·Å^2). Incrementally reduced force constants were used in four subsequent 1000-cycle minimizations, which gradually brought the system to the closest energy minimum. The temperature of the resulting system was raised to 300K over 60 picoseconds while the minimized structure was positionally restrained. The restraints were released incrementally in a subsequent 750-picosecond molecular dynamics (MD) with a 1fs time step. This is longer than the equilibration procedure we typically use with explicit solvent, but simulations with GB appeared to be more sensitive since large oscillating changes in the structure were observed if less careful equilibration was carried out (data not shown). In the production stage, temperature coupling was not employed and the simulations were carried out in the microcanonical ensemble.

2.2.4 Explicit Solvent Simulations with Locally Enhanced Sampling

We replaced the ADE8-THY19 base pair (excluding P, O1P, O2P, O3′, O5′, C5′, H5′1, and H5′2) with 10 copies by using the addles module in AMBER6. The template structure was taken from equilibration dynamics. All copies belonged to the same region and were given the identical initial conformation as the template structure but altered initial velocity information to permit divergence. The AMBER code was modified to permit coupling of LES and non-LES regions to separate thermal baths so that their temperatures could be controlled independently.

2.2.5 MM-PB Calculation

MM-PB/SA (Molecular Mechanics-Poisson-Boltzmann with Surface Area) [63, 64] is a relatively new method to estimate free energy differences by adding the contribution from intra-solute interactions, solvation energy and hydrophobic effect. It was used in this study to compute the free energy difference of two alternate conformations of the system. For each conformation, 100 equally spaced snapshots were collected from a one-nanosecond explicit solvent simulation. The MM energy was calculated as the average of the sum of all bonded and non-bonded interactions using ff94 force field parameters. The electrostatic contribution to the solvation free energy of each conformation was calculated by solving the Poisson-Boltzmann equation using the Delphi program [65]. The cubic grid was constructed so that each side was twice as long as the longest dimension of the structure. A grid spacing of 0.25Å was used. The cavitation, van der Waals and hydrophobic contributions to the solvation free energy can be estimated using solvent accessible surface area, but this contribution is typically small compared to polarization and screening effects for highly charged systems such as nucleic acids. We

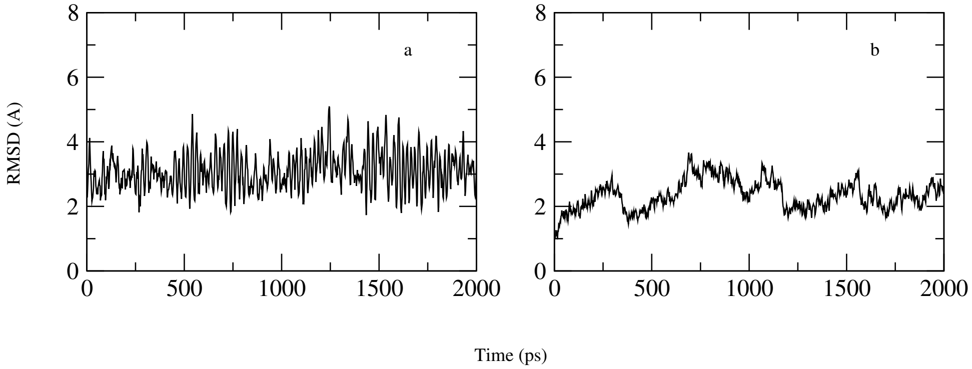

Figure 2.4: The graphs show results of simulations starting from the Watson-Crick structure: a) antiHB with GB, b) antiHB with explicit solvent. Average RMSD values are ~3Å (compared to the initial structure) in both cases, and the average structures from the two solvent models differ by 1.7 Å.

found that the surface areas for average *anti* and *syn* conformations differed by less than 0.1$Å^2$ (out of 5200 $Å^2$); therefore this contribution was not included in subsequent MM-PB/SA calculations.

2.3 Results and Discussion

2.3.1 GB and Explicit Solvent Simulations

The structure with conventional Watson-Crick base pairing for ADE8 (Figure 2.3a, hereafter denoted "antiHB") was selected and examined with unrestrained GB and explicit solvent simulations. Despite the expected slightly larger positional fluctuations and RMSD, the 2-nanosecond GB simulation generally reproduced the results of the explicit solvent simulation of the same length. The RMSD values in each case average $\sim$3 Å from the initial structure (Figure 2.4). Helicoidal parameters in the two solvent models are similar (Figure 2.5), with increased fluctuation in pucker of furan17. The average structures differ by 1.7Å with a slight widening of the major groove in GB (Figure 2.6). As one might expect for a continuum solvent model without frictional terms, the fluctuations in RMSD and helicoidal pa-

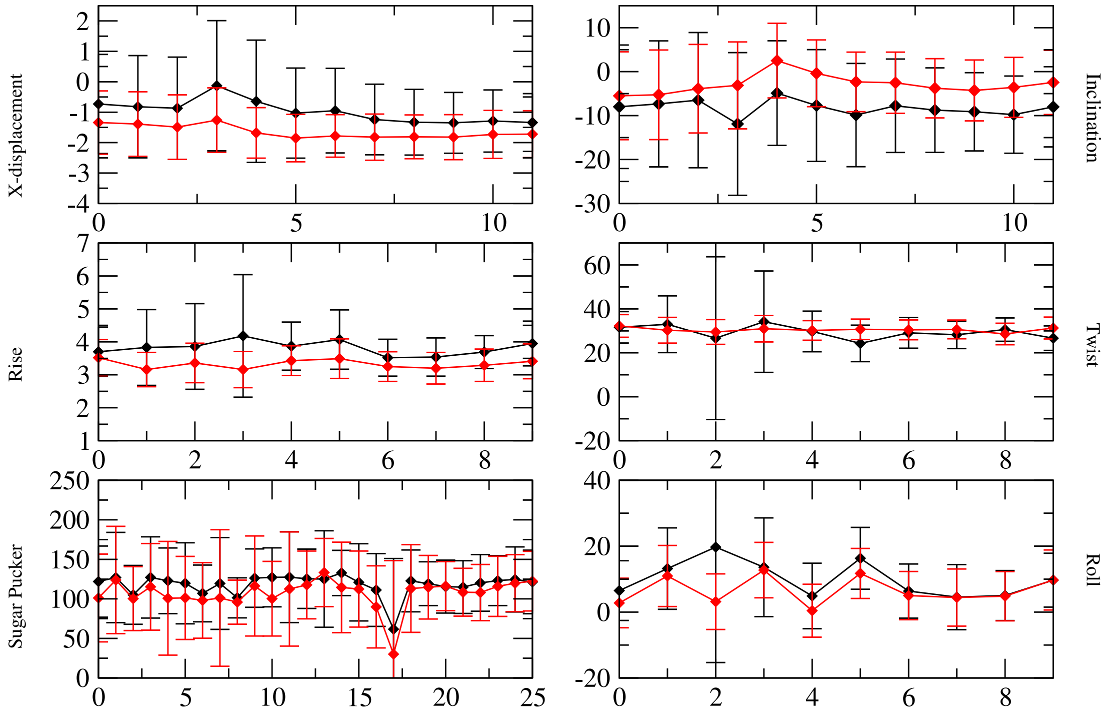

Figure 2.5: Helicoidal parameter comparison of antiHB GB (black) and explicit solvent (red) calculations. Similar results are obtained for the two solvent models.

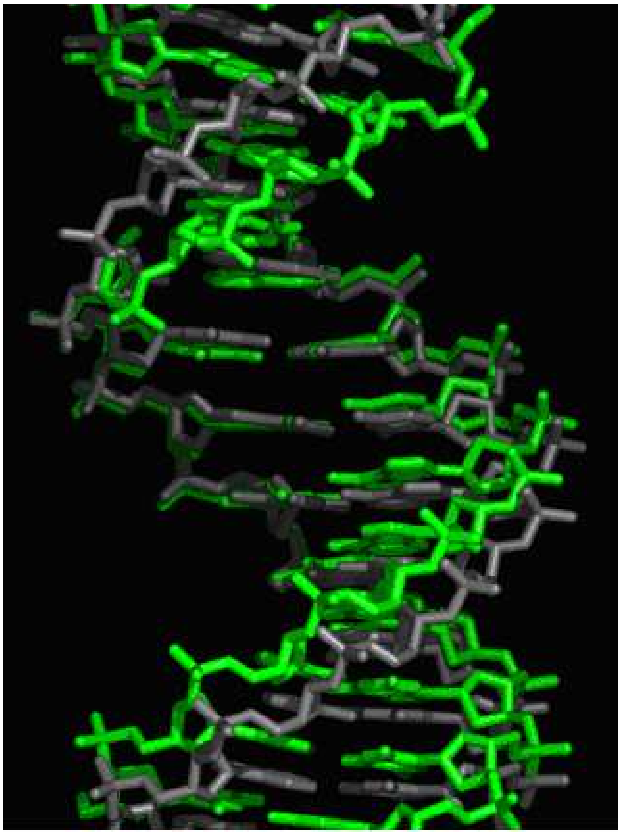

Figure 2.6: Average structure comparison of antiHB GB (grey) and explicit solvent simulation (green), with an RMSD of 1.7 Å. A slight widening of the major groove is observed with GB.

rameters are significantly larger with GB. These observations are consistent with the results reported previously for GB simulation of a standard DNA sequence [49]. The modified intrinsic radii for hydrogen atoms of different types appeared to work well with this modified DNA even though they were previously tested using only standard nucleic acid sequences. With both solvent models, the A:T pair was stable in the initial Watson-Crick conformation during the entire simulation. Even if alternate conformations should be sampled, the timescale for breaking this interaction was inaccessible in these simulations.

This process was repeated for a member of the NMR family in which no A:T pair was present. A structure with ADE8 in the minor groove of the DNA (Figure 2.3b, hereafter denoted "antiMinor") was selected. This structure was chosen because it has the lowest RMSD (1.5Å) among all structures in the family as compared to the Watson-Crick (antiHB) structure described above. The small deviation suggested that the transition to a base-paired structure might be confined to a local region of conformational space and therefore may be accessible during molecular dynamics.

Similar to the antiHB simulations, the general behavior was comparable during 4 nanosecond GB and 16.8 nanosecond explicit solvent simulations. While the antiHB model was stable regardless of solvent model, the antiMinor model is likewise unstable in both solvent models. The deviations from the NMR structure during both simulations (Figure 2.7) clearly show that the unpaired minor-groove ADE8 in the NMR structure is non-optimal. The DNA underwent significant conformational changes in both cases, with a maximum RMSD compared with the initial structure of 7.8Å in GB and 6.4Å in explicit solvent, with both simulations converging to RMSD values near 3Å from each other. This implies that our initial assumption in choosing antiMinor (structures with very localized differences in structure may have a simple transition pathway) was not necessarily valid. The

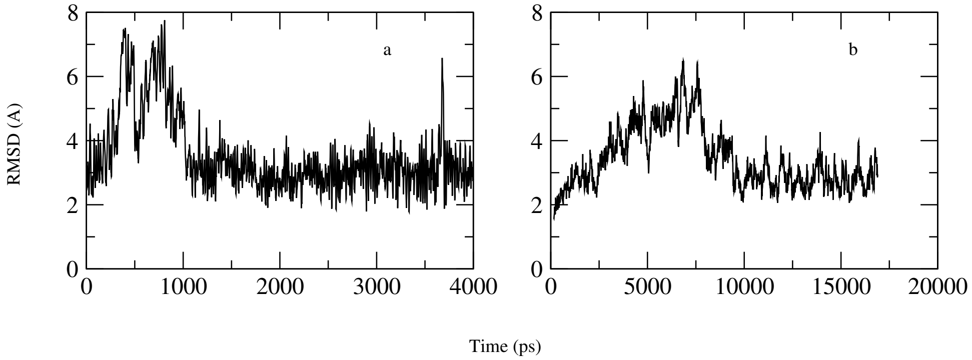

Figure 2.7: The graphs show results of simulations starting with *anti* ADE8 in the minor groove (antiMinor) with: a) GB solvation and b) explicit solvent. Similar results are obtained: a large increase in RMSD values, followed by convergence to lower values. The timescale of the GB transition is much shorter.

RMSD of average structures (after the transition) to average structures starting from antiHB are only 0.4 Å (GB) and 1.4 Å (explicit solvent).

Several important differences were noted when the results from GB and explicit solvent simulation were compared in detail. The most obvious was that when the durations of the transitions were compared, changes with GB occurred nearly an order of magnitude more rapidly than in explicit solvent (1ns vs. 10 ns, Figure 2.7). This is not surprising due to the lack of solvent-based friction in the GB model and has been reported in the past for nucleic acid simulations [66, 49]. In this case such comparison would not be very meaningful if the stable structures after the transitions were dissimilar. In this case, the resulting average structures over the last 1-nanosecond were very similar (Figure 2.8), and differed by only 1.2Å.

The more rapid convergence to a base-paired structure in GB simulations is certainly impressive. However, a closer look at the final structures revealed that ADE8 was in the *anti* conformation forming a Watson-Crick-type base pair with GB, but in explicit solvent the *syn* conformation was located, forming a Hoogsteen-type base pair. This difference is subtle but important, because either hydrogen bond pattern could result in the imino proton chemical shift observed in the NMR

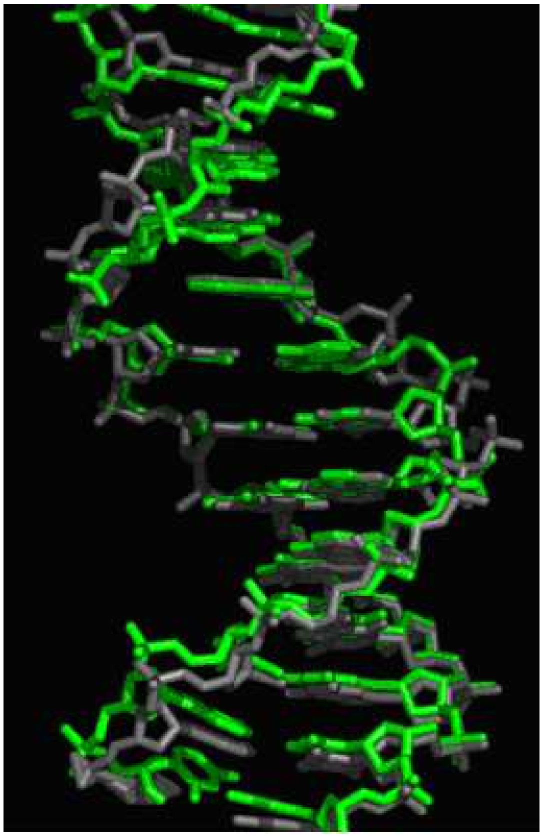

Figure 2.8: Average structure comparison of antiMinor GB (grey) and explicit solvent simulations (green), after the formation of the A:T base pair. The RMSD between the two structures is 1.2 Å.

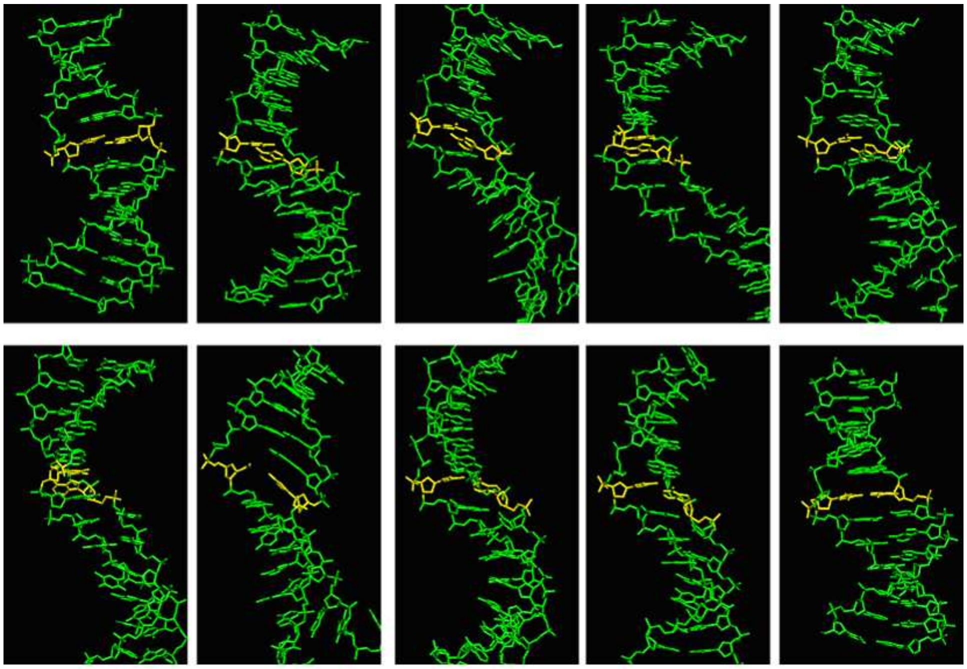

Figure 2.9: Transition snapshots (read from left/top to right/bottom) from antiMinor GB simulations, spaced equally 80ps apart. Large distortions extending several base pairs beyond the ADE8 region are evident during formation of the A:T base pair, but the final structure is similar to the initial structure with the exception of the A:T base pair geometry.

experiment (but neither alone explains the lack of NOE data for ADE8).

Distinctive structural features from the transition process in GB are displayed via 8 snapshots (Figure 2.9) and the time dependence of base pair hydrogen bonds (Figure 2.10). The two initial hydrogen bonds present when ADE8 occupied the minor groove, ADE8:N7-GUA20:N2 and ADE8:N6-THY21:O4′, were broken quickly during equilibration. ADE8 then moved to be coplanar with THY19, and formed reverse-Hoogsteen type hydrogen bonds to THY19 (ADE8:N6-THY19:O2 and ADE8:N7-THY19:N3).

As the simulation continued, the widths of the minor and major grooves changed continuously and simultaneously in multiple places, with the major groove widened by ∼8 Å and the minor groove narrowed by ∼4 Å, except at

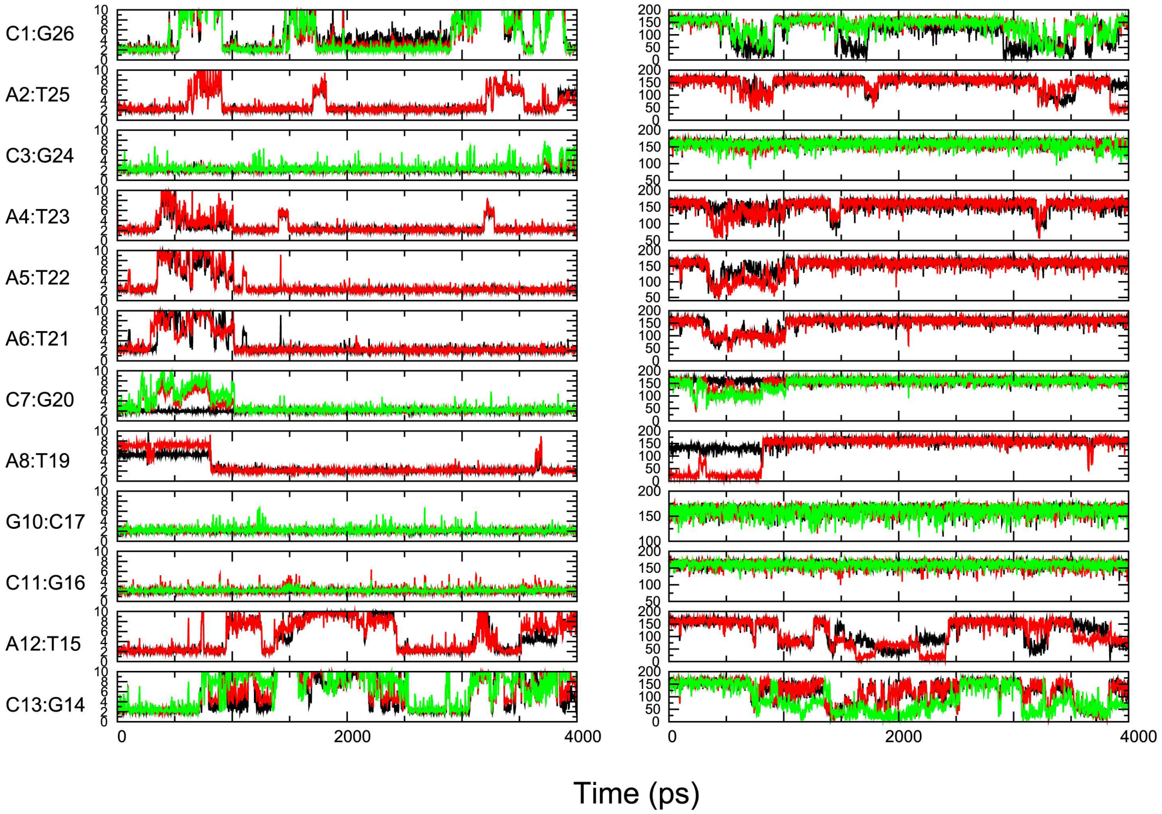

Figure 2.10: Hydrogen bond breaking and reforming events illustrated by the acceptor-hydrogen distances and acceptor-hydrogen-donor angles in the antiMinor GB simulation. Six of 7 base pairs in the long end are lost, and then reestablished after formation of the A:T base pair.

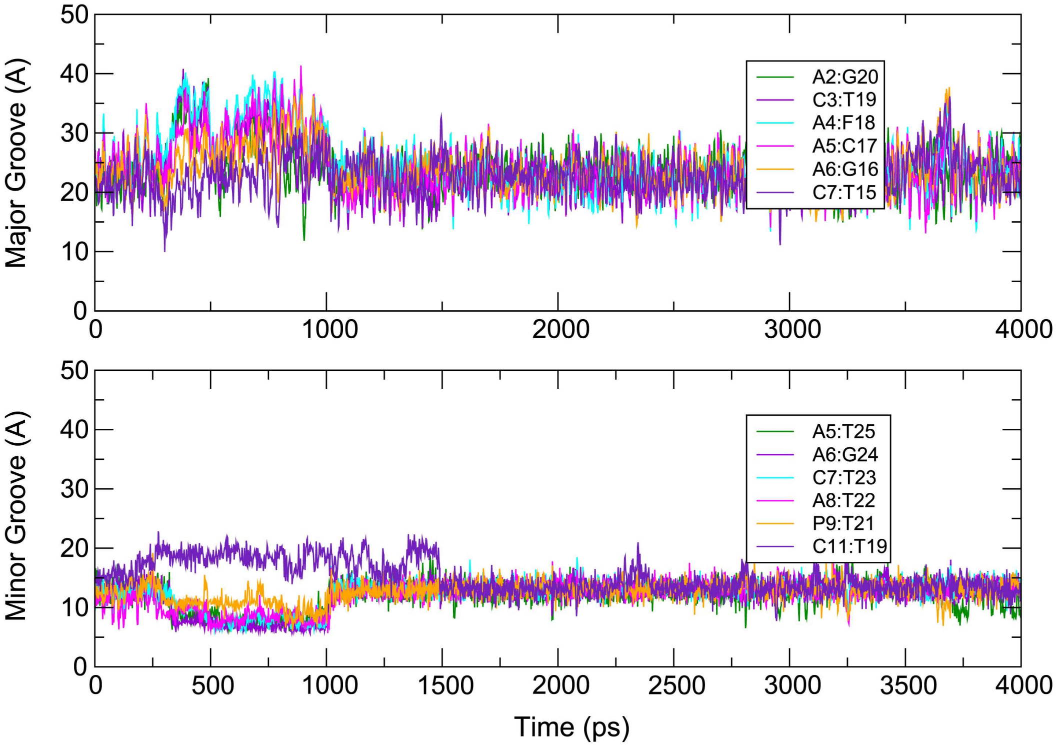

Figure 2.11: Groove width changes in antiMinor GB simulation.

C11-T19 and A8-T22, where ADE8 entered the double helix (Figure 2.11). Severe stretching and unwinding was accompanied by the complete loss of 13 out of 17 Watson-Crick hydrogen bonds, involving 6 of the 7 base pairs in the longer end of the DNA (Figure 2.10). Perhaps surprisingly, the short end remained intact, even though the whole structure deviated from the NMR model by as much as 8Å.

Next, the reverse-Hoogsteen hydrogen bonds between ADE8 and THY19 that had formed earlier were broken, and the bases shifted in-plane to a Watson-Crick configuration. Shortly after this local structure was formed, all of the base pairs previously lost were dramatically re-established, and the RMSD value compared to the average structure obtained from the antiHB control simulation was reduced to only 0.4Å. During the final 3 nanoseconds of this simulation, the overall structure remained stable.

It is remarkable that such a striking series of conformational changes could be observed in unrestrained molecular dynamics simulation of DNA with full atomic detail. As far as we know, it is the first time that such a transition has been reported. What is unusual in this case is not the relatively minor difference between initial and final states, but rather the very large changes seen during the transition pathway. The system moved from the initial NMR model to a region quite distant in phase space, and then returned to the neighborhood of the initial structure with a change in base pair geometry. In this sense the process is more comparable to an unfolding/refolding event than simple base pair formation. It is very exciting to see that simulations are starting to be able to model, in atomic detail, transient but key events that involve significant structural changes.

The transition to base-paired conformation was also observed in the explicit solvent simulation, but in this case it occurred over an almost 10-nanosecond period. Snapshots from the simulated transition are shown in Figures 2.12 and 2.13. The rearrangement started with *syn* ADE8 forming reverse Watson-Crick hydrogen bonds using ADE8:N6-THY19:O2 and ADE8:N1-T19:N3. The major groove was again widened (by $\sim$10 Å, a greater extent than the GB simulations), but in contrast to the GB results the minor groove was also widened (Figure 2.14). At $\sim$3ns, THY19 was flipped out into the solvent in the major groove. This motion opened sufficient space to permit a shift in the position of ADE8, allowing formation of Hoogsteen hydrogen bonds with the returning THY19. In the GB simulations, the space for a similar ADE8 shift had been provided by loss of the proximal C7:G20 base pair rather than motion in THY19. Here the base pair also re-formed at $\sim$10ns and remained stable for the remaining 6.8 nanoseconds of the simulation. The other portions of the DNA did not show the instability observed during the GB transition; this could be a result of the spatial restrictions imposed by the explicit solvent cavity, deficiencies in the GB model or associated parameters, or possibly due to the different final structures. The RMSD of all pairs of average structures

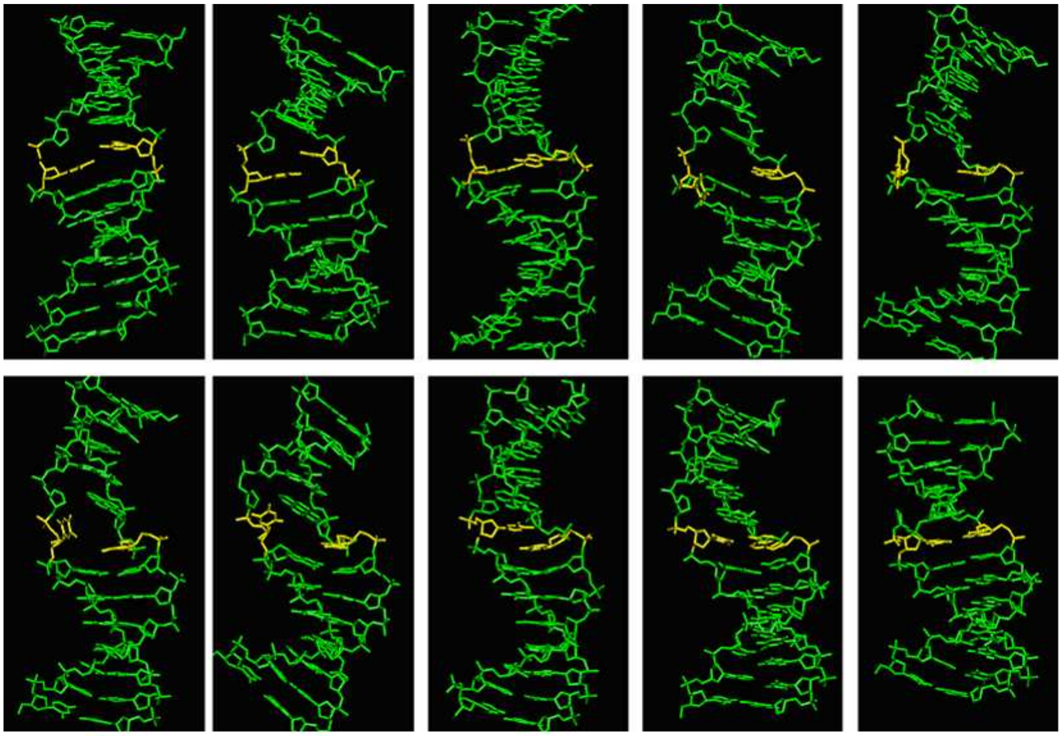

Figure 2.12: Transition snapshots (750ps apart) from antiMinor simulations in explicit solvent. Distortions are evident during the A:T transition, but are less dramatic than observed in the GB simulation.

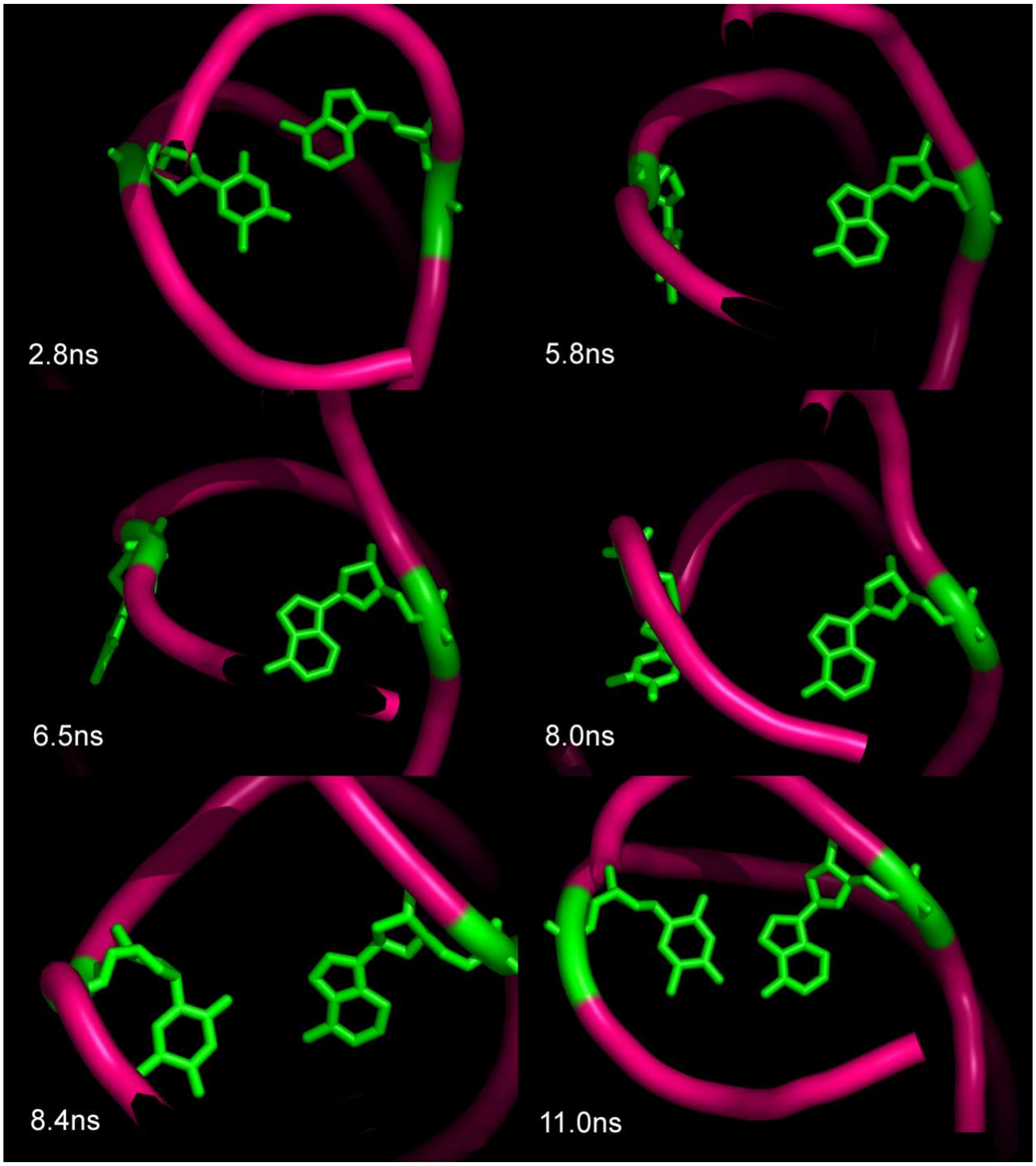

Figure 2.13: Snapshots showing the relative positions of ADE8 and THY19 during the A:T rearrangement in explicit solvent. Only the backbone and bases for this pair are shown.

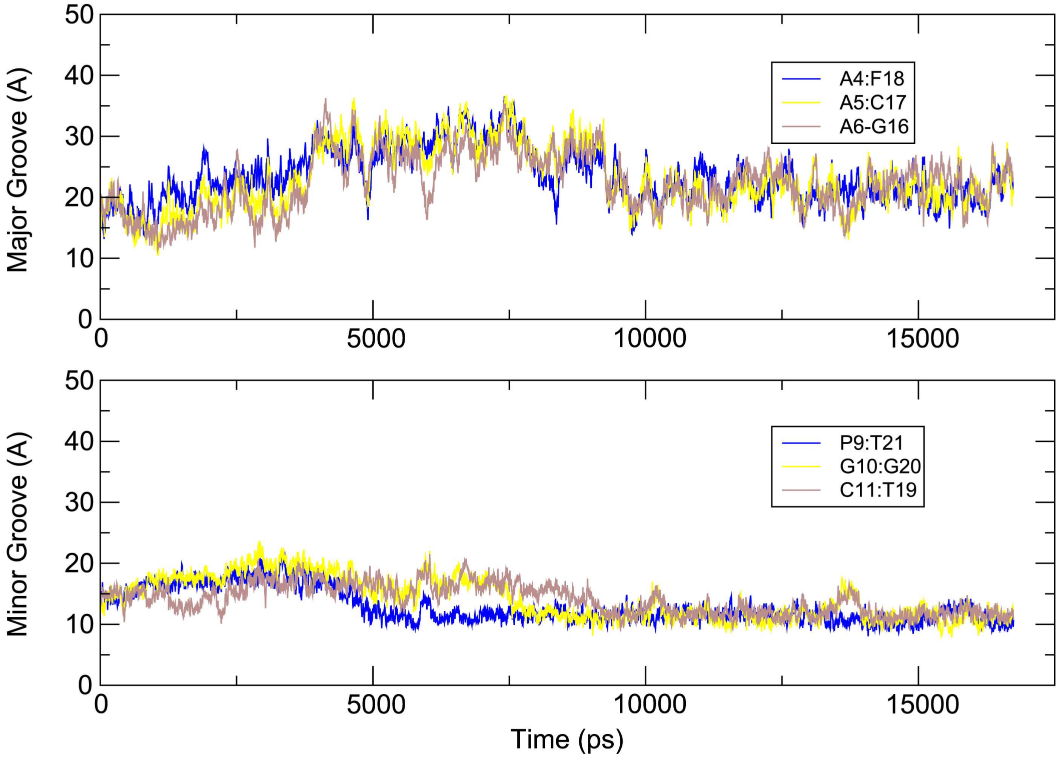

Figure 2.14: Groove width changes in antiMinor simulation in explicit solvent. In contrast to GB simulations, no narrowing of the minor groove is seen during the transition.

	antiHB (GB)	antiHB(explicit)	antiMinor (GB)	antiMinor(explicit)
antiHB (GB)		1.7	0.4	1.3
antiHB (explicit)			1.6	1.4
antiMinor (GB)				1.2
antiMinor (explicit)				

Table 2.2: RMSD (in Å) of all pairs of average structures from the antiHB and antiMinor simulations in GB and explicit solvent. The average structures are similar in all cases.

from the 4 simulations are given in Table 2.2; all are similar, with deviations of only ∼0.5 to 1.5 Å.

Our attention was immediately drawn to the inconsistent occurrence of the *anti* and *syn* ADE8 in the GB and explicit solvent simulations, respectively, because bases in *syn* conformation are not usually observed in standard base pairs and *anti* is generally considered to be thermodynamically more stable than *syn*. In this case it is possible that the observed *syn* conformation was simply kinetically trapped and does not reflect a conformation that is significantly populated in the equilibrium ensemble. To investigate this possibility, 100 structures for each conformation were collected from explicit solvent simulations, stripped of solvent and counterions and used in MM-PB free energy evaluation. MM-GB energies were also calculated for comparison, and even though the absolute energy values differ from those of MM-PB by almost 300 kcal/mol, the correlation is good with a correlation coefficient of 0.956 and slope of 0.980. This provides additional reassurance that the results of the GB model are satisfactory in this case.

The energy distributions, shown in Figure 2.15, indicate no significant preference for either *anti* or *syn*, regardless of whether GB or PB was employed. This suggests that both *anti* and *syn* conformations should be similarly populated at equilibrium in this specific environment. Among the individual components of MM-PB energies for this particular system, electrostatic and solvation terms are

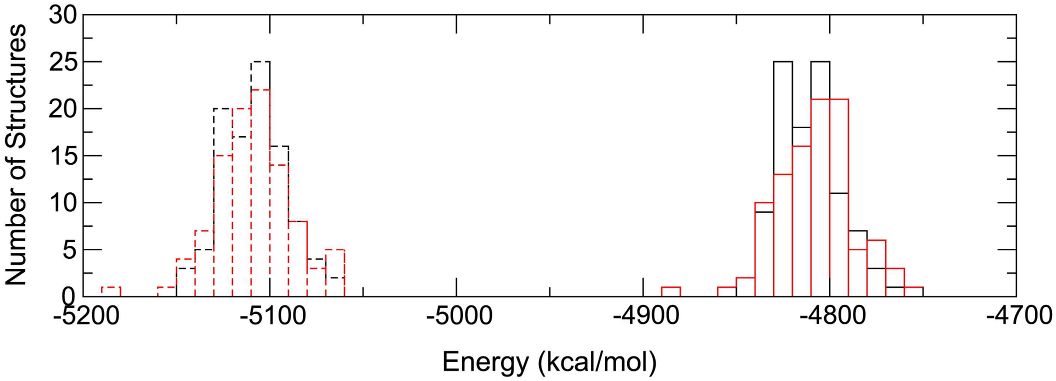

Figure 2.15: MM-PB (dashed lines on left) and MM-GB (solid lines on right) energy calculations for *anti* (black) and *syn* (red) conformations. The PB and GB energies are shifted, but both show no significant difference in energies for the *anti* and *syn* conformations.

the largest in magnitude, but favored different conformations. The effects nearly cancel; this is a common trend in such calculations and likely reflects the enhanced solvation of structures with unfavorable intramolecular electrostatics (such as parallel dipole alignment).

We speculated that the presence of the neighboring pyrene contributes to the unusually high stability of *syn* ADE8, and investigated this interaction in greater detail. The interactions of ADE8 with the rest of the system (Table 2.3) were calculated for 4 average structures: *anti* and *syn* adenine in an A:T base pair, neighboring either the pyrene-abasic pair or surrounded by standard bases. The *anti* conformation is seen to be significantly more stable than *syn* in non-pyrene DNA (2.3 kcal/mol), consistent with typical experimental observations. However, the *anti*/*syn* energy difference is much smaller in the pyrene DNA (0.8 kcal/mol). The *syn* conformation in the pyrene DNA is stabilized by local interactions, predominantly from van der Waals contacts with the bulky pyrene residue.

To further explore this issue, GB simulations were initiated using 2 NMR structures with *syn* ADE8 (Figures 2.3c and 2.3d), with the expectation that a more favored *anti* conformation in GB would lead to a *syn*/*anti* transition. However, the

	VDW (kcal/mol)	bonding (kcal/mol)	electrostatic (kcal/mol)	tot. (kcal/mol)
pyrene *anti* ADE8	-51.4	43.1	17.5	-25.9
pyrene *syn* ADE8	-52.2	43.1	-15.9	-25.1
std *anti* ADE	-54.3	42.7	-16.6	-28.2
std *syn* ADE	-51.6	41.0	-15.3	-25.9

Table 2.3: Group contributions to the local interaction of AT base pair. Electrostatics favor *anti* in both pyrene and standard DNA, but the van der Waals interactions with pyrene preferentially stabilize the *syn* conformation, nearly canceling the electrostatic effects in the pyrene system.

syn conformation was preserved in both simulations despite the relatively long lengths of 75.8 and 37.4 nanoseconds long respectively. GB simulations at 325K were also performed for both *anti* and *syn* conformations in order to determine if the higher temperature would facilitate *syn*/*anti* interconversion during the timescale accessible by these simulations. However, the process was not observed until after the DNA structure became unstable; the *anti* conformation unfolded after ∼11ns and the *syn* after ∼7ns. This suggests that the *anti*/*syn* transition may have barriers comparable to the stability of the double helix, and thus increased temperature was abandoned as an approach to simulate this interconversion.

2.3.2 LES Simulations

The calculations described above lead us to a possible interpretation that both *anti* and *syn* conformers may exist in the solution and the flexibility in this portion of the structure was responsible for the difficulty in obtaining NMR-based restraints for ADE8. However, this interpretation may be invalid if our statistics are unconverged. Although both structures could be formed in simulations, and energy analysis suggests that they may have similar population, direct simulation of an *anti*/*syn* transition was not observed. The energy barrier to such transitions in a tightly packed duplex is likely to result in a timescale beyond that currently accessible by standard simulation techniques.

LES (Locally Enhanced Sampling, [19]) has been demonstrated in several applications to improve sampling when multiple energy minima are present [67, 68, 52]. Particularly relevant to this study was the observation by Simmerling and Kollman that LES simulations of an RNA tetraloop spontaneously converted from incorrect to correct structure [69] despite the inability of otherwise identical non-LES simulations to do so [70]. LES was employed here not to aid in locating a single optimal structure, but rather to explore the possibility of directly simulating interconversion between *anti* and *syn* ADE8. Ten copies of ADE8 and THY19 were placed in the pyrene DNA duplex in explicit solvent using one snapshot from the equilibration where the base pair had not yet formed.

Differences in the behavior of the LES system could be the result of the pyrene, or an artifact of LES. Two control simulations were therefore used: in the first, an A:T base pair, ADE4:THY17, was duplicated in the non-pyrene DNA 10 base pair system in an identical manner as was done for the pyrene-containing 13 base pair sequence. This provides a test for the use of LES by comparing to our non-LES results for these same sequences. It is also possible that the change in *anti*/*syn* equilibrium is due to sequence-specific effects rather than the pyrene; an additional control was therefore carried out for the 13 base pair system in which ADE8 and THY19 were again replaced with 10 LES copies, but additionally the pyrene:furan pair was replaced with a non-LES A:T pair. Differences in these two simulations should be due to the pyrene substitution.

In the 4.8-nanosecond control for the 10 base pair system, all copies remained in the *anti* conformation during the entire simulation and *syn* was never sampled. Very similar results were obtained during 4.5ns simulation of the 13 base pair system without pyrene: *syn* conformations represented <1% of the total. The glycosidic torsion angle (χ) of each copy sampled during dynamics of the non-pyrene

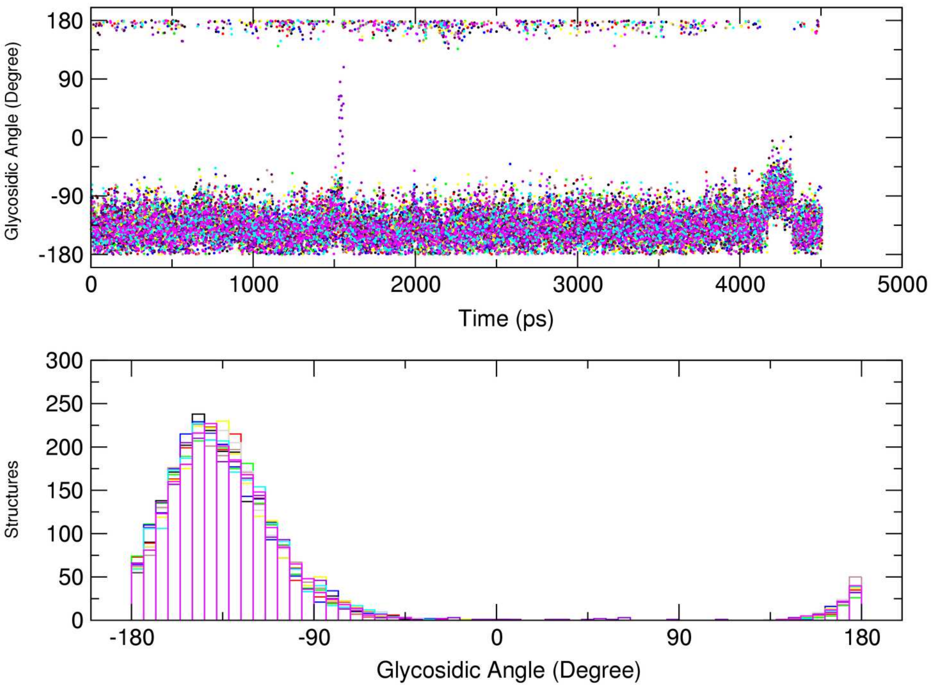

Figure 2.16: Time dependence (upper) and histogram of the adenine glycosidic torsion in LES simulation of non-pyrene DNA. LES copies are colored differently. Only *anti* is significantly sampled.

13 base pair system, as well as distribution histograms, are shown in Figure 2.16.

LES results for the pyrene-containing system differed dramatically from the control simulations. All 10 LES copies in the pyrene DNA spontaneously formed A:T base pairs, consistent with the non-LES simulations described above. However, as we anticipated, more than one LES copy was able to sample the *syn* conformation in the 14-nanosecond pyrene DNA simulation (Figure 2.17). A snapshot showing simultaneous LES population of *anti* and *syn* is shown in Figure 2.18. The base pair hydrogen bond patterns for each χ rotamer are consistent with those located at the end of the transition period in *anti*/*syn* non-LES simulations (in contrast to the less stable patterns observed shortly after base pair formation in the

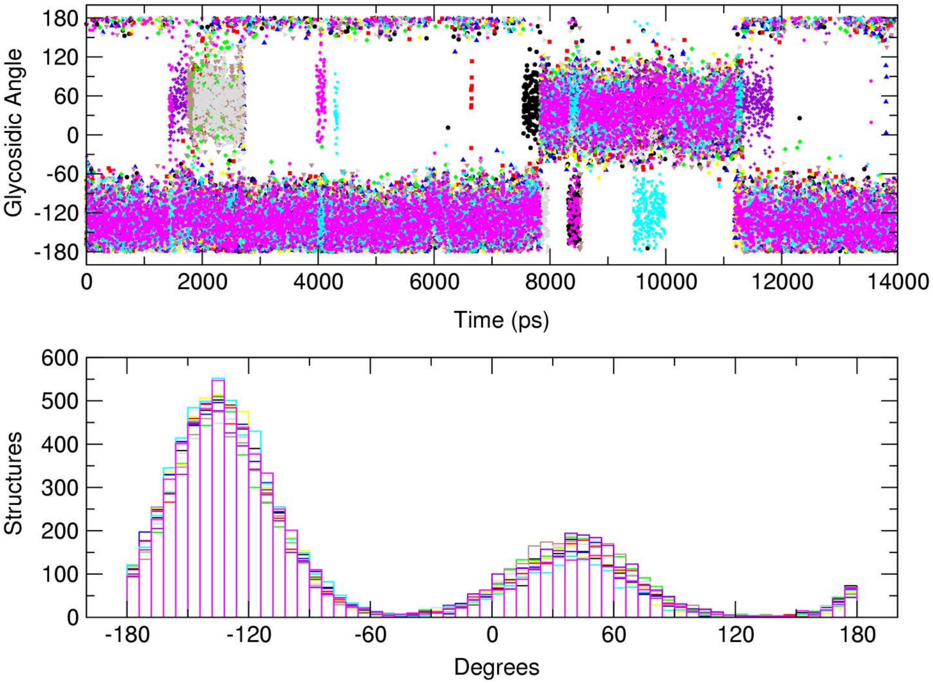

Figure 2.17: Adenine glycosidic torsion distribution, similar to Figure 2.16 but taken from a 14ns LES simulation of pyrene DNA. In contrast to the results of LES with standard DNA, the LES copies (different colors) of ADE8 neighboring pyrene sample significant *syn* population. *Syn* and *anti* are clearly separate minima.

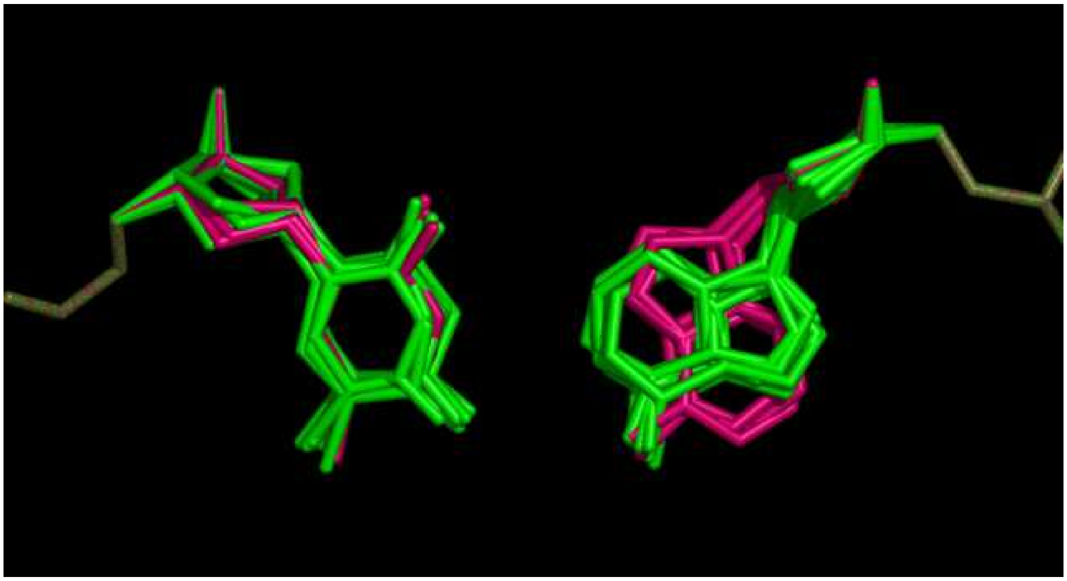

Figure 2.18: MD snapshot showing heavy atoms in the LES A:T pair neighboring pyrene. Copies are colored by conformation: *anti* with Watson-Crick hydrogen bonds (green) and *syn* with Hoogsteen pattern (magenta).

single copy MD). A single THY19 conformation is compatible with both ADE8 possibilities, consistent with the observation of NOE data for THY19. Each conformation was often stable for multiple nanoseconds before the next transition occurred. The total *syn* population was roughly 30% of the total structures sampled by all the copies, providing an estimated free energy difference of 0.5 kcal/mol favoring the *anti* conformation. This is consistent with the 0.8 kcal/mol determined by the energy analysis described above; in combination with the control simulation for the non-pyrene 13 base pair system with LES, this strongly suggests that the shift in *anti*/*syn* equilibrium is due to the presence of the pyrene.

With increased LES temperature (250K), the distribution of angles was broader and transitions became more rapid (data not shown), but the conclusions obtained from the lower temperature remained valid. Exclusively *anti* conformation was still present in all copies for standard DNA at this temperature. Since LES is an approximate method, the transition frequency and *anti*/*syn* ratio are also approximate yet the influence of the pyrene is clearly perceivable. This provides further evidence that a distribution of *anti* and *syn* may indeed be present in this region, and provides a potential explanation for the difficulties encountered in the NMR structure determination.

2.4 Conclusion

Lack of specific detail for portions of the structure is not a unique feature of this system. Structure determination for biomolecular systems is usually accomplished through X-ray diffraction or NMR spectroscopy. Both methods regenerate the structure from the measureables of the experiments, typically through the use of computational approaches that generate structures consistent with this data. However, many factors can influence the accuracy of the final structure. In particular, conformational heterogeneity in the sample studied can dramatically reduce the quality of the final structure. Such disorder typically results in averaged or incomplete data, which complicates the reconstruction process and can result in partially solved or low-resolution structures, which can be of little value.

Structure determination using NMR techniques relies heavily on the accurate determination of relative geometry information, such as interatomic distances and dihedral angles. When such information is abundant, reliable structure models can be built using restrained energy minimization or molecular dynamics techniques, and sophisticated treatment of solvation and electrostatics may not be necessary. However, difficulties in data assignment or interpretation, or motion on the timescale of the data collection can result in a lack of data for these regions and models of reduced quality. In such cases, simulations with more accurate treatment of inter and intra-molecular interactions may provide insight into the missing details.

More advanced biomolecular simulation approaches can be complementary to experimental data, providing an atomic-detail picture of molecular behavior in a manner that need not be time or ensemble averaged. For flexible molecules in the condensed phase, such as biomolecules in aqueous solution, molecular mechanics potential functions have made calculations tractable. One of the goals of computational structural biology is to reduce the amount of experimentally obtained data needed for successful refinement and state-of-the-art biomolecular simulation methods can give reliable insight into the atomic structures of complex systems.

In the present case, the conventional restraint-based model construction procedure failed to define the conformation of the adenine base proximal to the moiety of particular interest (pyrene). All simulations that have been done strengthen the hypothesized presence of the A:T base pair, and additionally provide a possible explanation for missing experimental data due to increased conformational heterogeneity in this region.

The experience gained with this pyrene DNA suggests that investigation of sequence-dependent structure flexibility can be a very challenging task; transitions as simple as base flipping occur on a timescale that is currently inaccessible by standard molecular dynamics simulations. Transient base-pair opening events have been previously observed in simulations; in one case, a standard base pair at the end of the DNA strand opened, and in another an adenine-difluorotoluene interaction in the DNA interior was temporarily lost. The simulations in explicit solvent demonstrate not only loss and reformation of a standard A:T base pair, but transition between hydrogen bond patterns during the process. A continuum treatment of solvation appears to be very promising in this regard, with the potential to dramatically reduce the computational effort required to simulate a given time period, and also provide more rapid equilibration due to the lack of solvent friction. While the structural conclusions drawn from continuum and explicit solvent are quite similar in this case, the agreement may not be expected to generally hold. One must therefore be cautious in relying solely on continuum solvation until more experience are gained with the relative merits of the many variants of these solvent models.

Simulations with the GB model show transient coupled loss of 6 consecutive base pairs, resulting in a dramatic unfolding-refolding event for the majority of the DNA double helix. However, direct simulation of *anti*/*syn* transitions was not possible even with the help of a continuum solvent model, and this process was not observed in any of our standard MD simulations despite evidence of thermodynamic accessibility.

The use of approximations (such as LES) that reduce the energetic barriers to such transitions and permit direct observation may be a critical component of complex structural studies. In the present study, the LES approximation with explicit solvation provided multiple observations of the base pair formation and reopening events, as well as *anti*/*syn* transitions and ensemble populations that are consistent with energy analysis of single copy simulation data. LES also gave results that were independent of starting model, unlike single copy simulations that can be poorly converged.

The effects of incorporation of unusual "bases", such as pyrene, into DNA systems continues to provide new insight into the determinants of nucleic acid structure and flexibility, and warrants further investigation. The introduction of the bulky pyrene residue not only preserves the fidelity of DNA replication, supporting the hypothesis that van der Waals interactions can be as important as specific hydrogen bonding in DNA replication, but as revealed in this study can also influence the behavior of nearby residues. In this case the pyrene appears to stabilize a *syn* conformation in the adenine 5′ to the pyrene, providing a potential explanation for difficulty in determining a single structure for this region with a restraint-based refinement procedure. Similar effects on local structure and flexibility may have other biological implications in the recognition and repair of DNA lesions. Minor conformers have been observed experimentally for polycyclic aromatic hydrocarbon (PAH) substituted nucleic acid systems, such as benzo[a]pyrene diol epoxide (BPDE) DNA adducts, and these may be responsible for tumorigenic activity [71]. Characterization of these ensembles at the atomic level may be critical to understand both the effects of the damage and recognition by cellular repair machinery. One hopes that approaches such as those that are presented here will extend the range of conformational variability that is accessible in computer simulation, and provide valuable models that complement experimental data for these systems.

Chapter 3

Protein Folding Studies with Replica Exchange Method

> "There is a difference between experimentalists and theoreticians: experimentalists observe the minima and maxima in free energy profiles - the experimental entities of intermediates and transition states - whereas theoreticians wish to calculate the entire energy surface of a reaction. Experimentalists talk about pathways, theoreticians about energy landscapes. Experiment and theory touch base around the ground and transition states that provide the milestones in the energy landscapes for the theoreticians to benchmark their calculations. The two views are to be reconciled." [72]

3.1 Protein Folding

3.1.1 Overview

Protein molecules can form, in a matter of microseconds to seconds, unique and well-defined three dimentional structures, which have certain tolerance to external disturbance, such as pH, temperature and denaturants, and conduct their biological functions within their life span. It has been a well-acknowledged fact that the most stable conformation of a protein, known as the "native structure", is largely determined by its amino acid sequence and amino acid sequence alone [73] and this native structure corresponds to the lowest free energy state in thermodynamics[1]. However, the origin of protein stability and the mechanism of protein folding have not been clearly understood, although progress is being made from both experimental and theoretical point of view [74, 75, 76, 77].

The volume of the literature on protein folding is huge and constantly growing. This is not just the reflection of the intense interests of the science community, but also represents the urgent needs of fast protein structure discovery in the era of post-genomics. While hundreds of thousands of protein sequences having been discovered from genome sequencing efforts[2], the traditional structure determination process itself has not changed much to meet the demands. The ultimate goal of protein folding studies is to be able to predict the thermodynamics and kinetics of a protein given its sequence and environment, providing guides for protein engineering and protein design.

However, achieving this is never an easy task.

Protein folding is a complicated process, which typically involves thousands of atoms. It is almost unimaginable for a protein molecule to enumerate all possible configurations to find the native state (Levinthal's paradox [78])[3]. This process needs to be strongly biased, through the acquisition of favorable interactions once the native state is reached, which are mostly dominated by the aggregation of

[1] Proper environments may need to be considered in some cases, but generally this is true.

[2] As of year 2002, 22,318,883 DNA sequences have been deposited at GenBank (http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html)

[3] If only protein backbone dihedral angles (ϕ and ψ) are considered and each dihedral angle can only have three different values, it will take a 100-amino-acid protein molecule 10^{29} years (longer than the present age of the universe) to search through all 10^{49} configurations with an average rotation frequency.

“hydrophobic” residues. Because of this driving force, the entropy cost in losing structural diversity can be compensated and the overall process is spontaneous.

To address protein folding, ideas had been drawn from our comprehension of the kinetics of chemical reactions, in which covalent bonds are made or broken during the transition from reactant to product. Protein folding was then described as another kind of reaction, folded state being the “product” and unfolded states being the “reactants”. Certain routes connecting the product and reactants exist and define how the folding occurs. Many experimental techniques were devised to seek this existence as well as transition states and possible intermediate states. Three mechanisms emerged to explain what these routes should be like [72].

Famework Model proposed that native local secondary structures could be formed, independent of the tertiary structure. These local elements would diffuse until they collided, associated, and coalesced to give the tertiary structure.

Nucleation Model hypothesized that folding process was initiated by the formation of a nucleus composed of certain native secondary structure from some neighboring residues of sequence. Once the nucleus was formed, folding would be continued in a stepwise manner.

Hydrophobic Collapse Model suggested that a protein would collapse rapidly due to the aggregation of its hydrophobic side chains and then rearrange the rest of the molecule to form the native state.

However, protein folding differs greatly from chemical reactions of small molecules [79, 80], which only involves a few strong bonds. During protein folding, hundreds of “weak” bonds might be formed or broken simultaneously. Therefore, only a few reaction progress indicators (“reaction coordinates”) are needed to describe the advance of chemical reactions. But this might not be a valid approach when studying a progress that proceeds in a high dimension space and can be potentially very heterogeneous. Unfortunately, high-dimensional data (>3) are difficult to visualize, rather folding process is commonly projected to one or two dimensions, investigated in a reduced or integrated phase space. In fact, looking for the best progress indicators is a significant part of folding studies. Any proposed folding mechanisms need to be cross-confirmed by different experiments.

Experimentally, conducting protein folding studies usually involves protein preparation, structure determination if necessary, thermodynamic measurements by thermal unfolding (measured by calorimetry and spectroscopy) or solvent denaturation using urea or guanidinium chloride (GdmCl), and kinetics model creation by relaxation experiments. The formation of secondary and tertiary structures can be monitored by various spectroscopic methods, such as circular dichroism (CD), NMR, fluorimetry, etc. To initiate folding events, the native state of protein needs to be perturbed either thermally or by denaturants and then the folding condition is quickly restored by quenching or diluting. Various techniques have been applied to follow the dynamic process of folding, covering different timescales [81]. Stopped-flow methods are ideal for conventional rapid mixing experiments, but are limited to millisecond timescale or greater. Specialized continuous-flow apparatus sometimes can reach for tens of microseconds, but usually relaxation methods or flash photolysis are more suitable for studying fast folding events. Especially, laser-induced temperature-jump experiments can be used to observe early folding events that take place in the range of nanoseconds to microseconds. NMR line-width broadening analysis is also a promising method in the range of 10 to 100 microseconds [72].

With all the technical advancements, experimental folding studies are able to provide us with a macroscopic picture of protein folding, however, the microscopic rationalization can be very different from case to case and becomes rather elusive. For the general principles of folding to be found, experiments and the following analysis should be conducted at microscopic level. The recent advent of single molecule spectroscopy [82], in which only one protein molecule is observed at a time, is one step toward this direction. Unlike experimental studies, computer simulation of protein folding is not limited by the issues of time or ensemble averaging. Perturbation of the native state is far more easier to do in computers than in beakers. Computer-simulated folding can even start with any desired conformation. The folding trajectory can be saved and replayed repeatedly for different analysis or "probing" with the ultimate structural details. With increased computer speed and better algorithms, there is much to expect as folding experiments and computer simulations are combined.

3.1.2 Computational Studies of Protein Folding

Computer simulations of protein folding mostly fall into two categories. The first category utilizes statistical tools developed in the studies of disordered systems, polymers, and phase transitions of finite systems, proposes general folding mechanism, and performs computer simulations using minimalist models, e.g. lattice models [83]. In the simplest lattice model, amino acid residues are represented by a single atom or "bead". The resulting chain of beads adopts a self-avoiding walk on a cubic lattice. Only very few kinds of basic interactions are defined. The major outcome of this category is that protein folding can be understood using statistical energy landscape theory [84, 85]. In the energy landscape theory, it emphasizes that folding is a very heterogeneous process and well-defined discrete pathways are not likely to occur early in the folding process. The folding landscape has a funnel shape with some roughness, which guides protein molecules through many different conformations toward the native state. The advantage of using minimalist models is that all conformations of a given sequence can be enumerated exhaustively and similar structure motifs (α-helix and β-sheet) found in real proteins exist as well. Many ideas have been obtained from these simulations, such as minimally frustrated protein model and kinetic partitioning of folding.

The second category is between the minimalist models and the real world, simulating protein folding with full atomic descriptions and sophisticated force field [86]. Although this appears to be the most realistic approach, the computational costs are increased dramatically and the accuracy of force fields are subject to more strict examination. Because of this, direct folding of large proteins is rarely performed. On the other hand, the reverse process, unfolding or denaturation, has been systematically studied through force-induced unfolding, high temperature and denaturant-added molecular dynamics simulations [74, 87]. Assuming detailed balance principle applicable, folding process may be understood as well from unfolding studies. Twenty-four unfolding simulations of chymotrypsin inhibitor 2 (CI2, 64-residue) allowed the identification of transition state, which was in "excellent" agreement with experiments [88]. However, the validity of applied the detailed balance principle under "extreme" non-equilibrium conditions is often questioned and conclusions drawn from this approach might be biased because of the funnel-shaped landscape in the vicinity of the native state where discrete pathways are likely to exist. To avoid drawing an incomplete picture of folding, multiple folding simulations should be conducted from random conformations. This is obviously very difficult to do at full-atomic level for large proteins with more than 50 residues.

Many short peptides (natural or engineered, less than 40 residues) with stable tertiary conformations appear especially appealing to both experimental and theoretical studies. They are often used as model system for the folding of large proteins. A recent example is "tryptophan cage" [89], a 20-residue engineered peptide from exendin-4 (EX4) of Gila monster saliva, which is not only thermodynamically stable but also kinetically a fast folder ($4\mu s$, [90]). An *ab initio* structure prediction from its sequence, using continuum solvent molecular dynamics, successfully rendered the correct native conformation [13]. Computational folding studies of other model peptides like "tryptophan cage" can also help to bridge the gap between experiments and theoretical work.

Thermodynamics of folding is equally important and is required to explain the folding kinetics. Obtaining reliable thermodynamics data from simulations of all-atom model is often impeded by the transition timescales and the sampling efficiency of regular MD simulations. This has been demonstrated in the previously discussed pyrene DNA studies, in which regular MD simulations were unable to give estimates of the relative stability of two distinct conformations [91]. The sampling difficulty in protein folding simulations is still greater than that can be comfortably handled even if only the degrees of freedom of solute are considered explicitly. One common practice is to use elevated temperature to boost the transition probability, which would be rather low otherwise. Recently, computational folding studies using multiple loosely coupled MD simulations is becoming increasingly popular. Umbrella or biased sampling technique, in which intermediate steps are minimally coupled, has been applied by Brooks and colleagues to generate the effective energy landscape of folding along two predefined progress indicators, the radius of gyration and the percent of native contacts [86]. Both folding and non-equilibrium unfolding approaches were studied for fragment B of Staphylococcal protein A, segment B1 of Streptococcal protein G, cold shock protein A, and src-SH3 domain. The results were benchmarked against the folding energy surface generated from umbrella sampling techniques. It was clearly demonstrated in their calculations, from theoretic point of view, that the non-equilibrium unfolding approach appear to be quite unsuitable for the study of protein folding, especially when protein cannot be simply described by a two-state model.

Another approach that is attracting more attention recently is replica exchange method (REM, [92]), which was developed as an extension of simulated tempering or parallel tempering. Although in principle analogous to the biased-sampling methods, practically it is much easier to conduct. Several peptide folding studies with REM have been reported [93, 94, 95, 96, 97, 98, 99, 100]. This chapter describes in detail the first implementation of REM in AMBER molecular modeling package at the level of message-passing interface (MPI), and the thermodynamic investigations of several small peptides.

3.2 Replica Exchange Method

3.2.1 Theory

In REM [92], a generalized ensemble is constructed, consisting of N non-interacting replicas of the studied system $x_i, i = 1, 2, \cdots, N$. Each system i is studied at a different temperature T_i. The state of this generalized ensemble is denoted as $X = \{x_1, x_2, \cdots, x_N\}, x_i = \{p_i, q_i\}$. Because replicas are independent of each other, the weight factor for the state X is then simply the product of Boltzmann factor for each replica:

$$W_{REM}(X) = exp\left\{-\sum_{i=1}^{N} \beta_i H\left(p_i, q_i\right)\right\}$$

where β_i is the inverse temperature, p_i is the momentum vector, and q_i is the coordinate vector. A random process takes place within the generalized ensemble, in which two replicas exchange their temperatures. Hence, the state of the generalized ensemble changes from X to X':

$$X = \{\cdots, x_{i,T_i}, \cdots, x_{j,T_j}, \cdots\} \rightarrow X' = \{\cdots, x_{i,T_j}, \cdots, x_{j,T_i}, \cdots\}$$

where T is the temperature. In molecular dynamics, this can be done naturally by scaling the momentum vector p:

$$\begin{cases} p_{i,T_j} = \sqrt{\frac{T_j}{T_i}} p_{i,T_i} \\ p_{j,T_i} = \sqrt{\frac{T_i}{T_j}} p_{j,T_j} \end{cases}$$

The detailed balance condition needs to be fulfilled to achieve an equilibrium condition:

$$W_{REM}(X)\, w(X \rightarrow X') = W_{REM}(X')\, w(X' \rightarrow X)$$

where $w(X \rightarrow X')$ is the transition probability from X to X'. After substituting W_{REM} and Hamiltonian H, the exchange probability are obtained:

$$\frac{w(X \rightarrow X')}{w(X' \rightarrow X)} = exp(-\Delta)$$

where

$$\Delta = (\beta_i - \beta_j)(E_{j,T_j} - E_{i,T_i})$$

and E is the potential energy of a replica. This can be satisfied by applying the usual Metropolis criterion:

$$w(X \rightleftharpoons X') = \begin{cases} 1, & for\, \Delta \leq 0, \\ exp(-\Delta), & for\, \Delta > 0, \end{cases}$$

In general, an REM simulation with N replicas repeats the following two steps:

1. Each replica is simulated under canonical ensemble condition at fixed tem-

perature simultaneously and independently for certain MD steps.

2. Periodically a pair of replicas with neighboring temperatures are chosen for exchange. The exchange is accepted or rejected based on the probability calculated from the equation above.

Any thermodynamic quantities at any replica temperatures can then be calculated by direct arithmetic averaging over generated replica configurations. For intermediate temperatures that are not simulated, multiple-histogram reweighting techniques is needed instead.

3.2.2 Implementation in AMBER

REM can be implemented in different ways. The simplest solution is to combine scripting and molecular dynamics modules, which requires the least modification to available MD code. When exchange attempts are due, exchange probabilities are calculated by scripts usually written in interpreting languages (e.g. Perl) and MD modules are restarted with reassigned velocities according to the exchanging temperatures. Another solution is to take advantage of inherent multiple communication domain mechanism in the message-passing interface (MPI), the parallel programming interface standards. Although modest modification to the original MD code is necessary, the resulted code is completely binary, which may outperform the script solution on certain computer architectures, permits new types of communication, and appears more appealing, both aesthetically and practically[4]. In this implementation (AMBER 6), the second approach was adopted and will be ported to the coming new release – AMBER 8.

[4] This is true, especially when the code is run at national supercomputing centers where job priority is mostly decided by the number of requested CPUs.

Multi-Loader Code

The key to the implementation is the capability of running multiple MD simulations simultaneously within a single process. This was done by utilizing the multi-communication mechanism in MPI, which allows the creation of multiple independent communication domains. Each domain runs a separate MD simulation with its own input and output file namespaces. Domain creation was done by calling the following standard MPI subroutines:

```
    MPI_COMM_SPLIT(COMM, COLOR, KEY, NEWCOMM, IERROR)
```

Sander, the MD module of AMBER, was properly modified and transformed into a subroutine that is called in each communication domain. The master process of each domain is responsible for file I/O. The interprocess communication in the original sander module is hence contained within each sander domain. This code, named `multiloader`, forms the framework for the later REM addition. Although `multiloader` is a only a simple large-scale parallel "device", it proves itself a very useful tool to set up hundreds of similar simulations.

REM Extension

Built on top of `multiloader`, REM came as a natural extension. An additional communication domain (REM domain) is created, in which exchange probability is calculated, and includes the master processes from all sander domains. The organization of communication domains is demonstrated in the scheme below (Figure 3.1). When exchange is attempted, replicas are paired by their temperatures and exchange probabilities as well as velocity scaling factors are calculated by one replica from collected system energy and temperatures, which is then sent to the other. Therefore, it was decided that replicas with odd indices initiate the exchange process and form replica pairs alternatively with two neighbors.

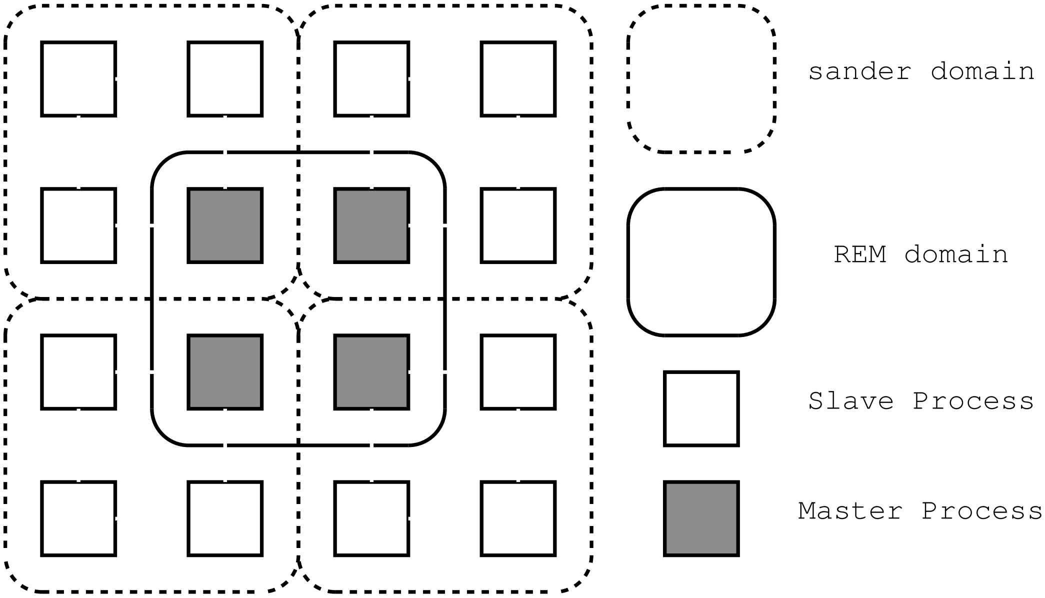

Figure 3.1: Domains and processes in the REM implementation

3.2.3 Usage

This REM implementation does not require extra types of input files other than those of standard sander module. However, a separate set of input files does need to be created for each replica except *prmtop*. All input files of the same type (e.g., *mdin* or *inpcrd*) have a common root name, which is used in sander command line. After the command line is processed, a suffix of a 3-digit number will be automatically added to the root name of each input file based on the replica number starting with 000. For example, sander command line argument *-i rem.in* for an 8-replica REM job will require eight *mdin* files with the names of *rem.in.000*, *rem.in.001*, · · ·, *rem.in.007*. The only difference between these *mdin* files is typically the target temperature (**TEMP0**) of replicas and nothing else. All output files (*mdout*, *restrt*, *mdinfo*, and *mdcrd*) are generated for each replica following the same naming rule and ascendingly sorted by their temperatures[5]. Modifications to command line argument list and each type of input and output files are summarized hereafter.

[5] So *mdcrd.000* will have all the snapshots of the lowest replica temperature.

command line Two new flags, *-numexchg <integer>* and *-saveexchg <integer>*, are added, specifying the number of exchanges and how often output files are updated in terms of number of exchanges respectively. One new option for output file appending other than overwriting (*-O*), *-A*, is available, affecting *mdcrd* and *mdinfo*.

mdin is not changed in any noticeable way[6]. However, small differences do exist. *ntave* can be used to control if exchange probability is calculated from averaged or instantaneous energy and temperature. A new namelist variable *freqrem* is defined for the ease of coding and it is not supposed to be set by users.

inpcrd/restrt The target temperature of each replica is added to *inpcrd/restrt* files on the second line after **NATOM** and **TIME** with format E15.7. This will overwrite the value read from *mdin*. The reason is to facilitate resubmitting jobs automatically at certain supercomputing centers where a wallclock time policy is imposed and job restarting is frequent.

mdout/mdinfo *mdout* will not contain system energetics for a whole REM simulation, which instead can be found in *mdinfo*.

mdcrd The only change to *mdcrd* is that the title line is not written if the file is appended.

rem.log is new and has the information about job setup and every exchange attempt. It is created or appended if it already exists. The filename is hardwired in the code. An entry is placed for each replica and records current scaling factor (negative if attempt is rejected), current temperature, current energy, current target temperatures of this and next MD runs. Exchange success ratio can be calculated from these entries.

[6] It means that regular sander *mdin* needs no modification before used in an REM simulation.

3.3 Thermodynamics and Kinetics Studies of a Nonapeptide

3.3.1 Background

Short peptides and protein segments are generally considered unstructured in aqueous solution [101], exhibiting large conformation entropy by sampling the allowed region of the Ramachandran plot. In opposite to this understanding, early UV-CD studies on poly-L-Lysine by Tiffany and Krimm [102] suggested that the "random-coil" state of peptides and proteins might prefer certain local order, which is described as an extended left-handed 3_1 helix. Attention was brought back to this disagreement by a series of more recent studies on tripeptides by Schweitzer-Stenner [103, 104], which supports the view of Tiffany and Krimm that the "coiled" states of peptides and proteins may be more structured than what people thought. If this is true, the structure preference of protein segment would certainly have more or less impact on the process of folding.

The conformational preference in aqueous solution of a 9-residue peptide (YDVPDYASL) from influenza hemagglutinin Ha1 (residue 100-108) [105] was studied using GB solvent model. This peptide was originally used as a model antigen for the study of antibody-antigen recognition mechanism [106]. A preliminary CD spectroscopic study did not show any measurable secondary structures at room temperature[7]. A particular conformation was visited frequently in several MD simulations, which was mostly helical at the C-terminus with a turn near the N-terminal region that facilitates the stacking of aromatic rings from two tyrosine side chains (Figure 3.2). In this study, the thermodynamic stability of this conformation was estimated to be -0.6 kcal/mol at room temperature, suggesting a very weak conformational preference. This peptide was used as a model system

[7] T. Canseco, D. Raleigh, C. Simmerling, unpublished data.

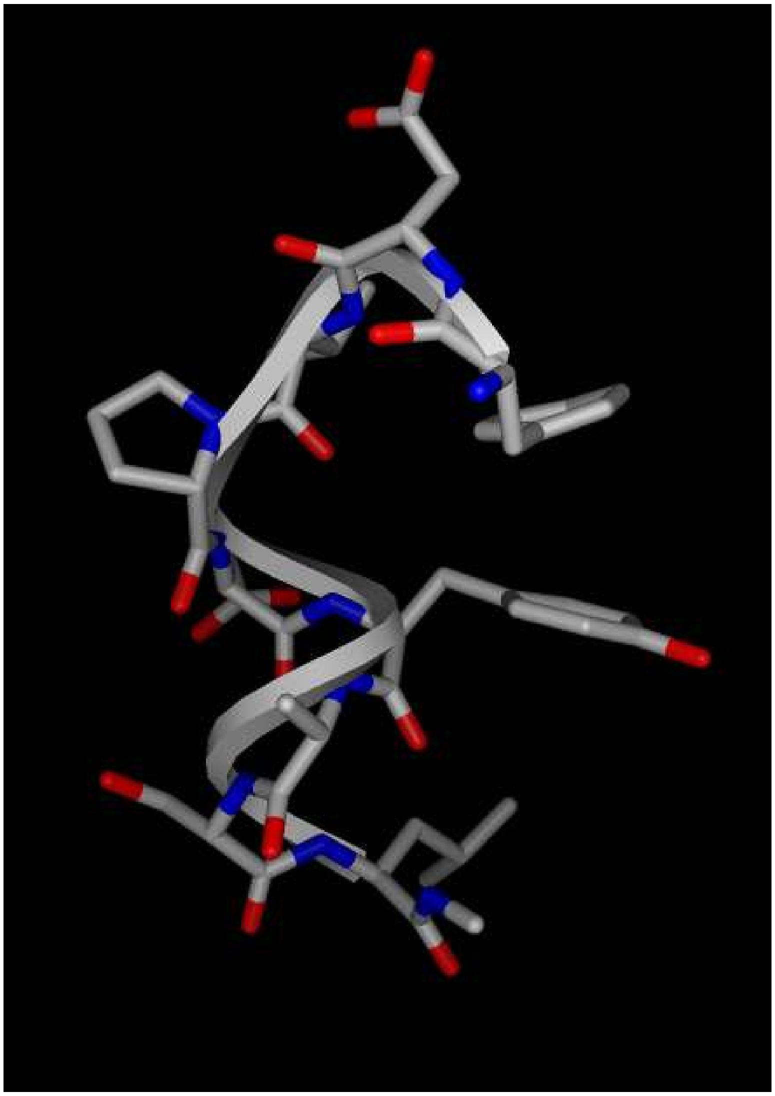

Figure 3.2: The preferred structure of the 9-residue influenza hemagglutinin fragment

to study protein folding.

To study the formation of this preferred structure, which is then denoted as the "native" or "folded" state of this nonapeptide, 188 folding simulations at room temperature (298K) were conducted from arbitrarily selected structures sampled at 800K. With more than 96.2% of them reached the folded state, the distribution of first passage time of folding was calculated and the kinetics of folding was plotted as the fraction of unfolded simulations against folding time. Surprisingly, it was found that at least three exponentials were required to fit the data. Their timescales vary by three orders of magnitude (80ps, 1.1ns and 55ns). To be able to explain the complex folding process of this nonapeptide, a thermodynamic analysis of the folding landscape through REM was performed. A total of eight replicas were used to obtain equilibrium distributions at temperatures over a range of 200K to 400K. The folding free energy landscape was constructed using the two largest principal components as the reaction coordinates[8]. Four main basins of attraction on the landscape were found, a perfect demonstration of the "kinetic partitioning" mechanism (KPM) proposed by Thirumalai [107], which provides a unified depiction of the folding of heteropolypeptide that is topologically frustrated because of the uniform distribution of hydrophobic residues. The topological frustration creates a rugged free energy landscape consisting the native basin of attraction and many other minima. The denatured ensemble then partitions into fast folders and slow folders that are stuck in the local minima before reaching the native state.

In order to reconcile the thermodynamic and kinetic views of folding for this sequence, structures sampled during the folding kinetics simulations were projected onto the replica-exchange landscape. From this analysis the native state assumed

[8] Principal components are calculated from principal component analysis (PCA), which is a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables (*principal components*). Principal components are usually sorted by their magnitude, which reflects the amount of variation in the data. The first component accounts for as much of the variation as possible, and each succeeding component accounts for as much of the remaining variation as possible.

in previous MD simulations was located and is the global minimum of the free energy landscape. Moreover, each timescale of folding was excellently explained by the landscape and their characteristics were found. One of the most interesting findings is that the slowest kinetic timescale is in fact composed of two different processes of a similar folding rate, corresponding to two distinct non-native basins. More importantly, the intermediate timescale, the majority of all, does not correspond to any single non-native basin and overall there is no single transition state of folding. Rather, a funnel-shaped landscape view of folding is more appropriate and many possible ways of folding are sampled, which is consistent with a wealth of experimental data for protein folding. The full-length analysis and discussion are given after the method section.

3.3.2 Importance of This Work

Why would one care for the *simulated* folding of a nonapeptide that does not even *possess* an experimentally detectable structure?

Protein folding is a multi-dimensional transition of a system dominated by non-specific interactions. Protein sequences are well-optimized to adopt a general funnel-shaped energy landscape that speeds up the process of searching the global energy minimum. However, numerous local energy minimum may exist, created by the incorrect order of forming native contacts or the possibility of forming non-native contacts. Therefore, the folding routes of individual protein molecules may differ, exhibiting the inherent microscopic heterogeneity. Studying this heterogeneity and its sequence dependence is the key to understanding experimental observations and designing new protein sequences. Unfortunately, tracking the folding process of individual protein molecules cannot be achieved easily with available experimental techniques.

Simulated or *in silico* folding is a more convenient approach in this regard. By combining well-studied kinetic results and reliable thermodynamics, this work rendered a complete picture of the kinetic process. A sound justification for this process was derived from the free energy landscape of folding. Thereby, this work convincingly demonstrated the underlying heterogeneity of protein folding. Even though little experimental work has been done on this particular system, similar folding scenarios have been found elsewhere [108, 109, 110]. The importance of this work is not obtaining the direct correlation with any experimental observations, but serving as a scaffold, just like lattice model, upon which deeper understanding may be built. Using the state-of-the-art simulation techniques (ensemble folding simulations and the replica exchange method), this work also demonstrated how crucial good statistics are for computational folding studies. The ultimate goal of computational protein folding is to simulate the folding of larger proteins to similar satisfaction. However, it is far more difficult and challenging.

3.3.3 Methods

All simulations were carried out using the parm94 force field [11] with GB continuum solvent model as implemented in AMBER 6 [53]. Translation and rotation of the center of mass were removed periodically during the simulations. A time step of 2fs was used in all folding simulations, but 1fs was found more appropriate for REM simulations due to the higher temperatures. Unless otherwise noted, all reported root mean square deviations (RMSDs) were measured using backbone atoms of all residues with the "native" conformation as reference.

3.3.3.1 Folding Simulations

The initial structures of 188 folding simulations were chosen randomly from an MD simulation at 800K. The RMSDs of all 188 initial conformations range from 2.3Å to 6.7Å, which covers most of the denatured states as shown in the discussion section. All folding simulations were carried out at 298K and were stopped after it folded to the native conformation, indicated by RMSD less than 0.8Å[9]. Structures were saved every 10ps. Of 188 folding simulations, 96.2% were folded successfully within 100ns. The 188 folding simulations were treated as an ensemble of individual folding events. The unfolded faction of this ensemble was then calculated and used to describe the folding kinetics, which corresponds to the forward folding rate. As would be carried out for experimental data, the time evolution of this fraction of unfolded was fitted to single or multiple exponential equations to determine the folding scenario of this nonapeptide. No constraint on the sum of weights of the exponentials was used.

3.3.3.2 REM Simulations

Eight trial MD simulations were first carried out for 750ps to determine the temperature dependency of average potential energy, which is shown is Figure3.3. To achieve a one-dimensional random walk in temperature space, i.e. a uniform exchange ratio (0.15 used here), replica temperatures need to have exponential distribution (Figure 3.4), which can be calculated from the temperature dependency of average potential energy. The replica temperatures used in this study is 200K, 220K, 243K, 268K, 295K, 325K, 358K, and 394K. Exchange were attempted every 500 MD steps, only the last snapshot saved for later analysis. Each replica was coupled to its temperature bath with a constant of 0.1ps.

A total of 100,000 structures were collected from each replica, which were split into two sets of 50,000 each. Principal component analysis (PCA) was performed on the covariance matrix of backbone coordinates obtained from the second set of structures. All structures were best-fit to the native structure using the peptide backbone (C, CA, N, and O). The two largest principal components (magnitude-

[9] This cutoff value was chosed based on the RMSD histogram analysis of early MD simulations.

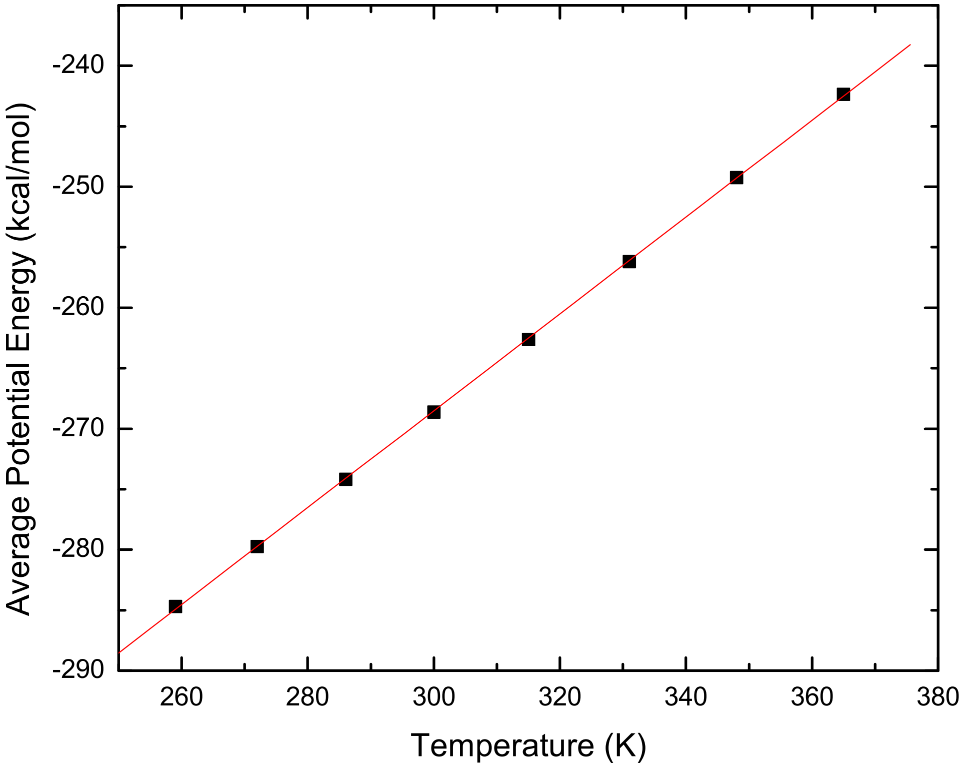

Figure 3.3: Temperature-dependency of average potential energy

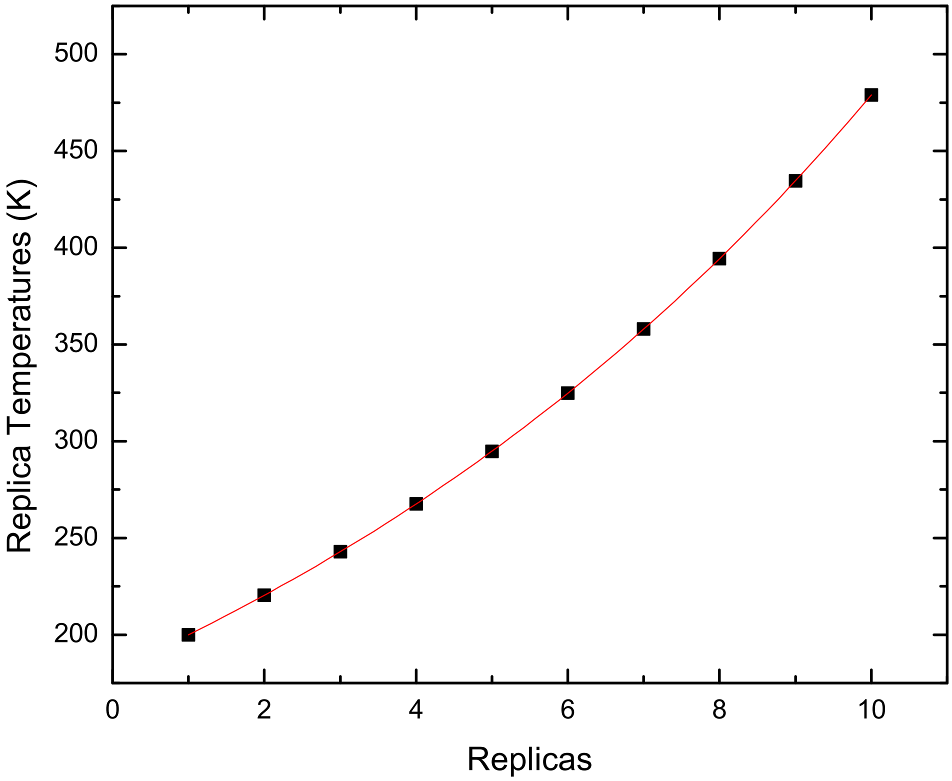

Figure 3.4: Replica temperature distribution

wise) together contribute 61% of the total fluctuation. The sampling along these two components was calculated as the dot product of Cartesian coordinate vector and principle component vector and histogramed using 100 bins in each dimension. Free energy of each bin was calculated relative to the most populated bin. The free energy landscape calculated from the first set of 50,000 structures showed no obvious difference, suggesting sufficient sampling. Landscapes at higher temperatures were constructed similarly. In general, they bear close resemblance to that of 295K. The landscape was colored by the relative free energy values, ranging from blue to red. The color mapping was indicated by a color bar on the right. System properties, such as root mean square deviation, torsion angles, were calculated and then mapped onto the landscape.

3.3.4 Results and Discussion

Folding Kinetics

The sequence of this nonapeptide is neither evolutionarily optimized, nor engineered to stabilize a particular conformation. In spite of the fact that there was no detectable secondary structure in a preliminary CD study, a marginally stable conformation was observed from MD simulations in continuum solvent, which was assumed as the "native" state as mentioned earlier[10]. To one's surprise, the "folding" process of this nonapeptide is unexpectedly complicated. The fraction of unfolded simulations out of 188 folding simulations is plotted as a function of time (Figure 3.5). At least three exponentials ($fraction\,of\,unfolded\% = \sum_{i=1}^{3} w_i \exp(-t/\tau_i)$) were required to accurately fit the curve, with time constants ranging over nearly three orders of magnitude (88ps, 1.1ns and 55ns). The fitting is summarized in Table 3.1.

Interpretation of experimental observation of non-single exponential folding

[10] This was later explained by the thermodynamic studies.

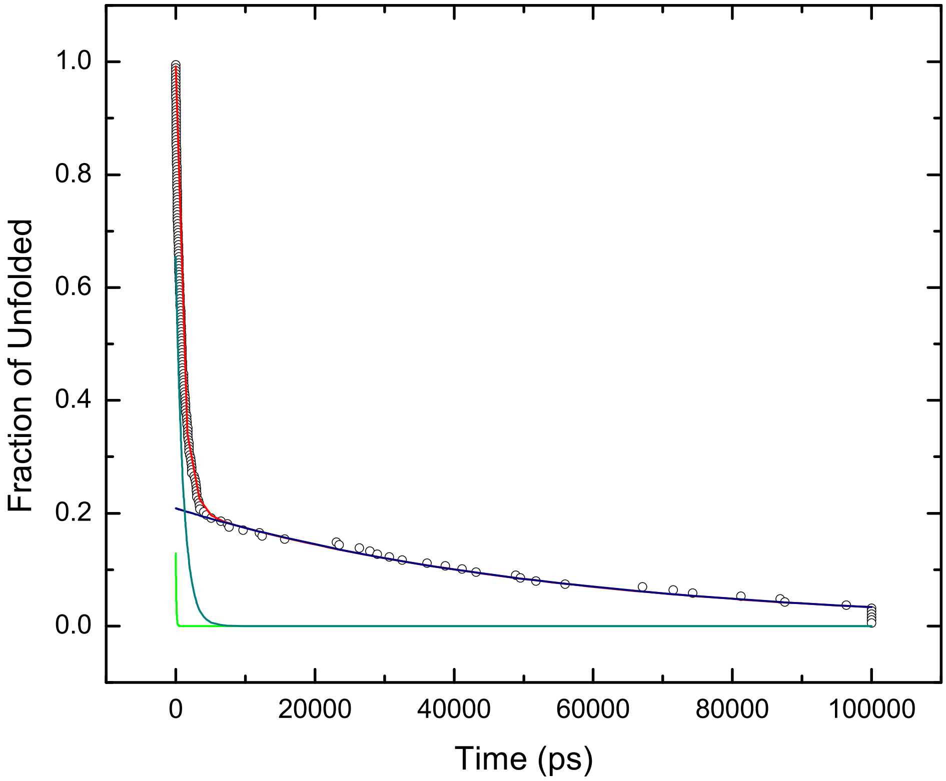

Figure 3.5: Exponential fitting of the fraction of unfolded: simulation data (open circle), fitted curve (red) and three individual components: fast timescale (green), intermediate timescale (dark cyan), and slow timescale (navy).

	Weight (%)	Time Constant τ (ps)
Fast Timescale	13	79±7
Intermediate Timescale	66	1100±17
Slow Timescale	21	55000±1700

Table 3.1: Weights and time constants of the exponential fitting (the correlation coefficient is 0.9991)

kinetics can be challenging due to the ensemble nature of folding. In the present case, however, the simulations provides time-resolved data with single molecule resolution. In principle, one should be able to use this data to validate the use of three exponentials, and ideally to determine what types of transitions or pathways give rise to each of the characteristic times of the folding process. The main challenge in carrying out such analysis lies in the difficulty of assigning a particular folding simulation to one of the three timescales. This is particularly true of the intermediate and slow timescales, in which significant overlap exists between the transitions of each population represented by these curves. The overlap with the slow process becomes less significant beyond 10ns, with fast and intermediate timescales almost completely fading out. Therefore, details of the slow timescale could be possibly picked out first.

The proximity of each residue to its native conformation was measured using residue-wise RMSDs. With the exception of Val3 and Ser8, most residues are able to adopt a native-like local conformation on a short timescale. It is therefore possible that the non-native conformations of Val3 and Ser8 can be used to characterize the slow folding process. Further analysis of hydrogen bonds sampled during these folding simulations showed that the slow folding, non-native local backbone conformations (negative ψ_3 and positive ϕ_8) are correlated with formation of non-native backbone hydrogen bonds, which must be broken for folding to occur.

To elucidate if ψ_3 and ϕ_8 are sufficient to identify the slow timescale, folding simulations with non-native ψ_3 and ϕ_8 were removed. The kinetics of the remaining folding simulations was then reanalyzed and fit to two exponentials with $\tau_1 = 96 \pm 6$ps and $\tau_2 = 1200 \pm 11$ps (Figure 3.6). These relaxation times are quite similar to those obtained for the short and medium timescales from the full ensemble fitting, and the lack of a slow phase in this fit suggests that the entire 55ns timescale was the result of the non-native ψ_3 and ϕ_8 families. Folding simulations

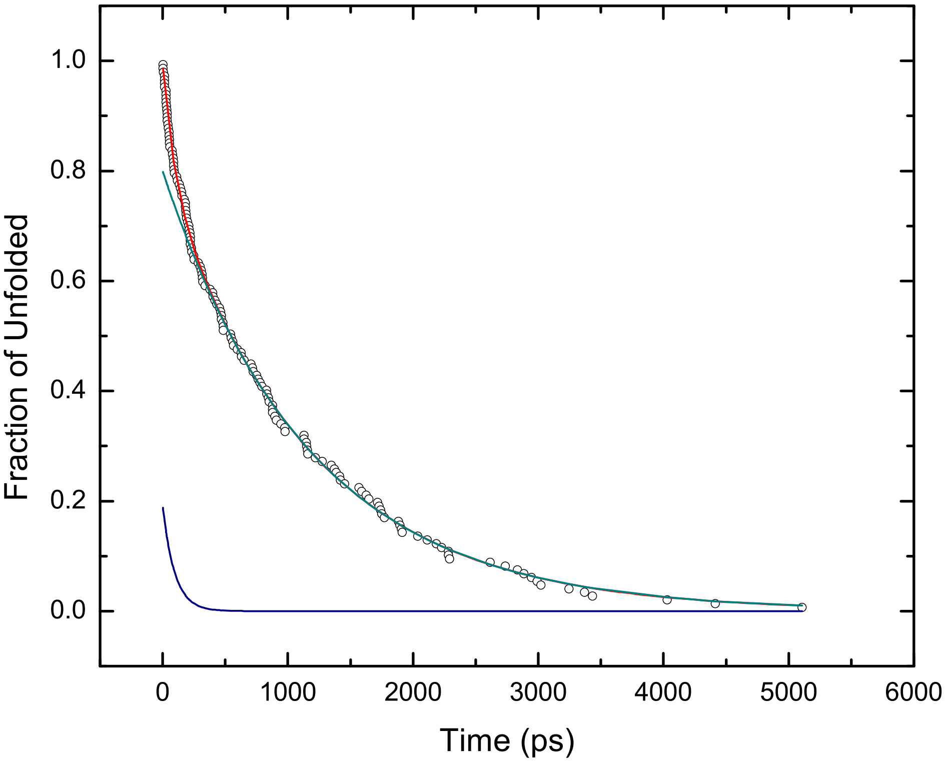

Figure 3.6: Two exponential fitting of the fast and intermediate timescales (the correlation coefficient is 0.9987).

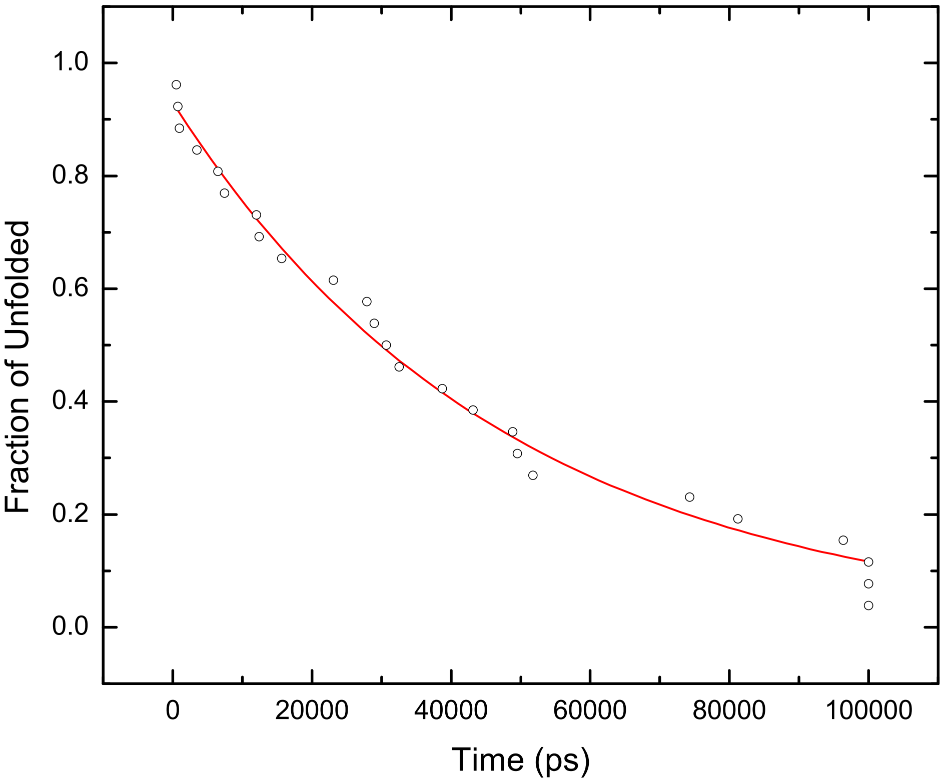

Figure 3.7: The single exponential fitting of non-native ψ_3 (the correlation coefficient is 0.9878).

with non-native ψ_3 and ϕ_8 can be fit individually to a single-exponential with a time constant of 48000 ± 1600ps (Figure 3.7) and 52000 ± 13000ps (Figure 3.8).

Thermodynamics

The free energy landscapes of folding at 295K, 325K, and 394K are shown in Figure 3.9, 3.10, and 3.11. At room temperature (295K), there are four distinct basins of attraction, the centers of which are largely located at $m_1 = (-1.4, -0.4)$, $m_2 = (4, -0.4)$, $m_3 = (0.5, 3)$, and $m_4 = (1, 2.5)$. In the center of the landscape is a large poorly populated area. As the temperature is increased, this area and the barriers that are between minima become smaller and eventually disappear. The global minimum m_1 at the lower-left of the landscape is stable by 2.5 kcal/mol comparing

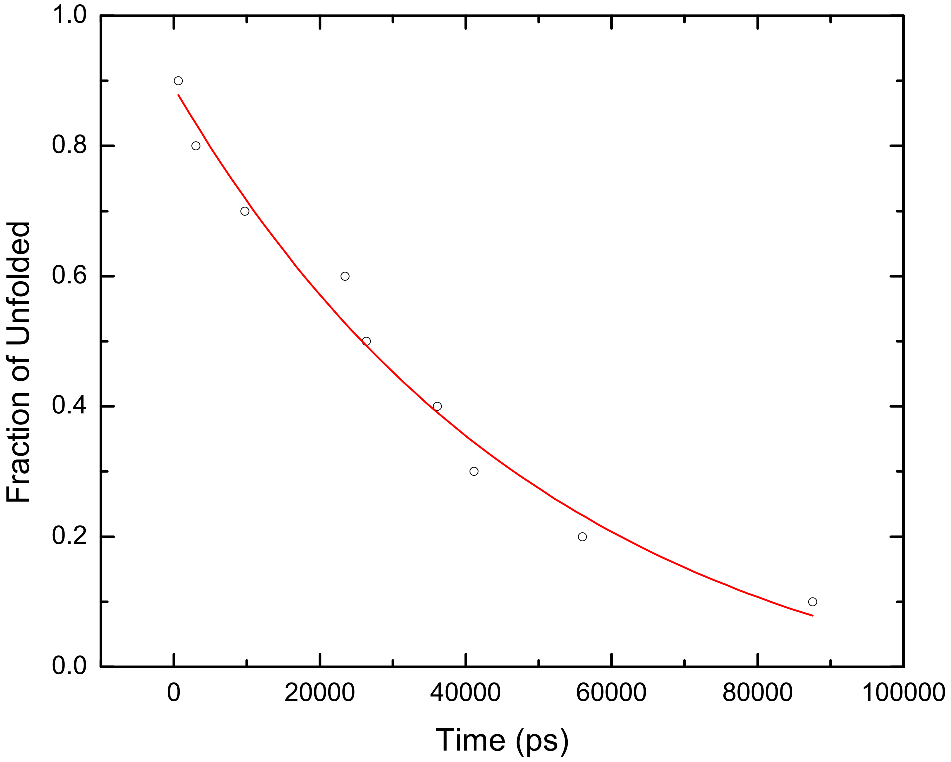

Figure 3.8: The single exponential fitting of non-native ϕ_8 (the correlation coefficient is 0.9818)

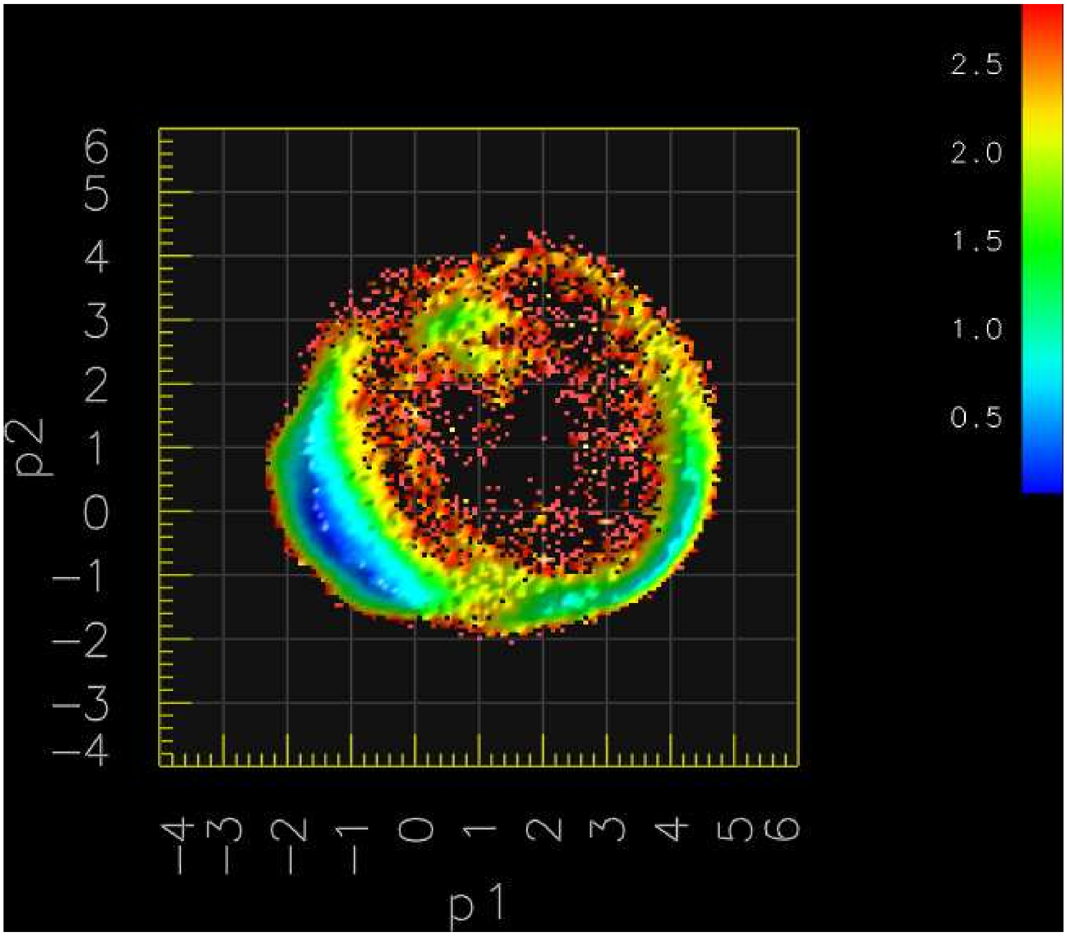

Figure 3.9: The free energy landscape of folding of the 9-residue peptide at 295K.

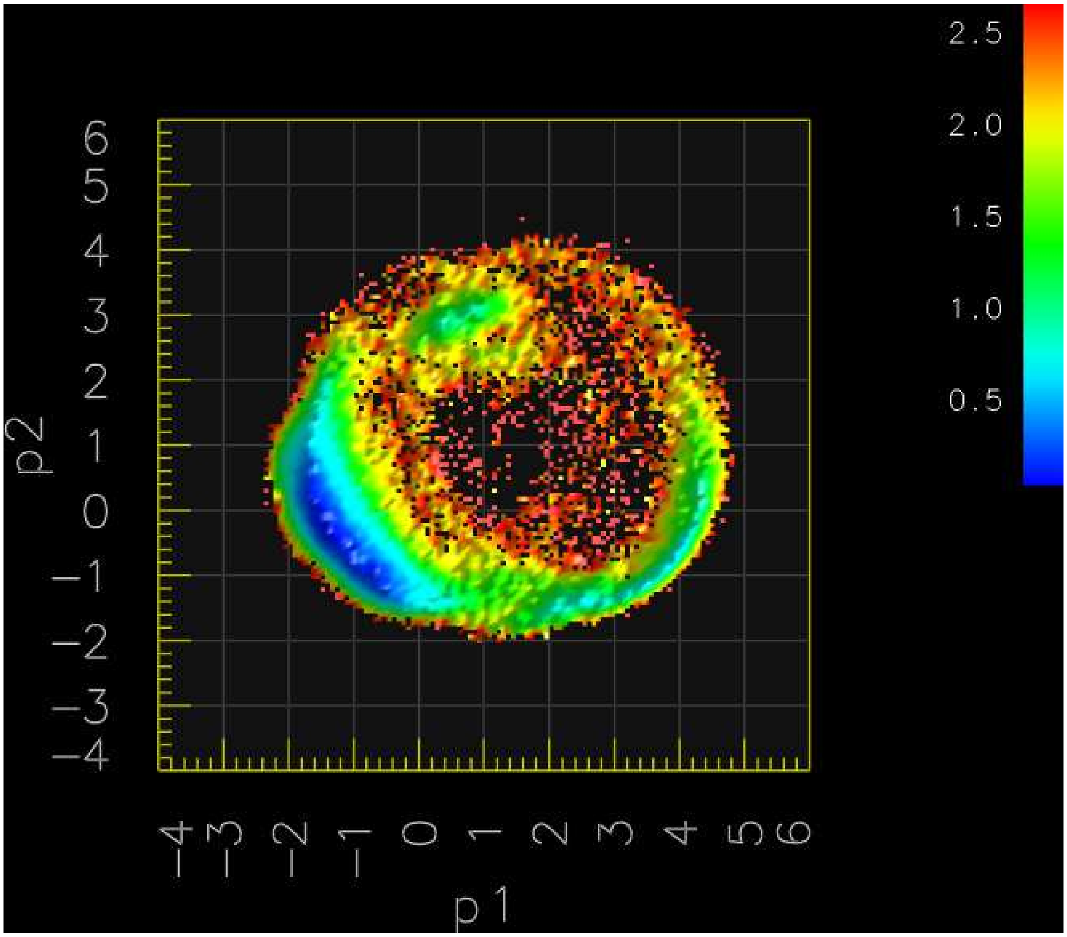

Figure 3.10: The free energy landscape of folding of the 9-residue peptide at 325K.

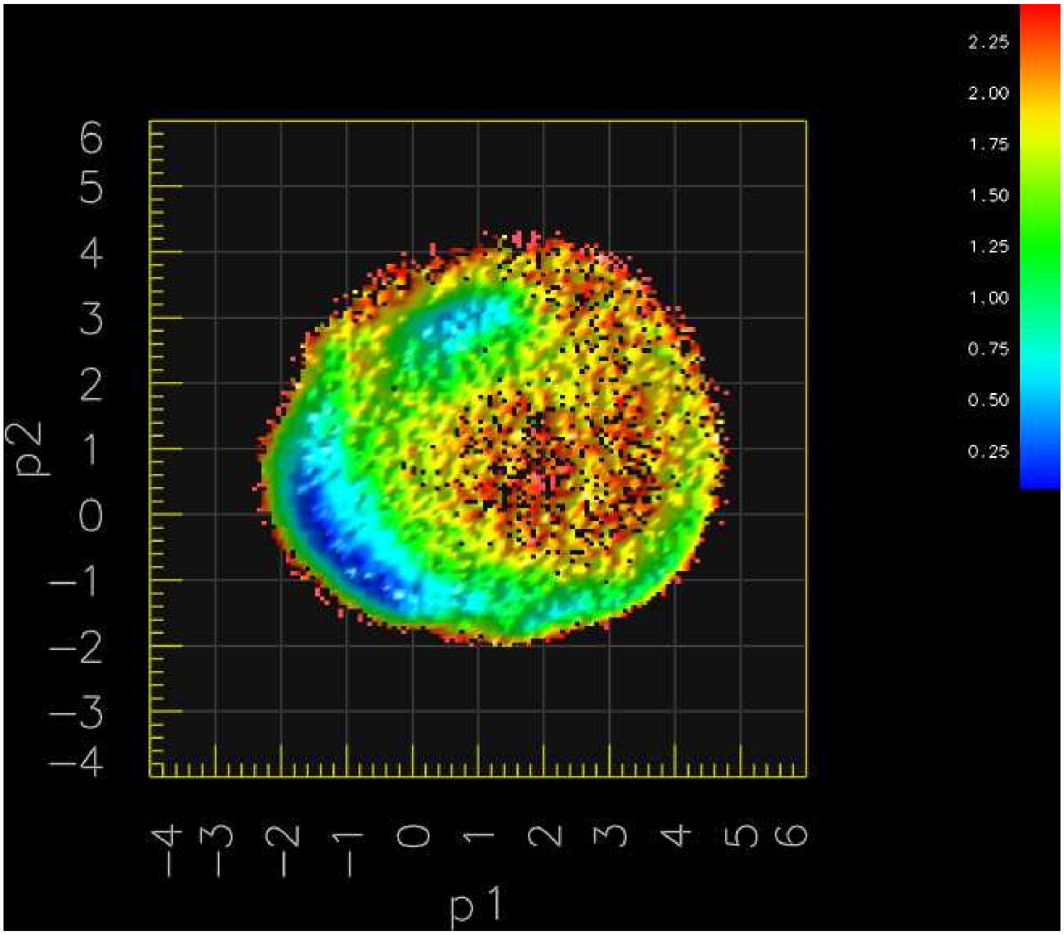

Figure 3.11: The free energy landscape of folding of the 9-residue peptide at 394K.

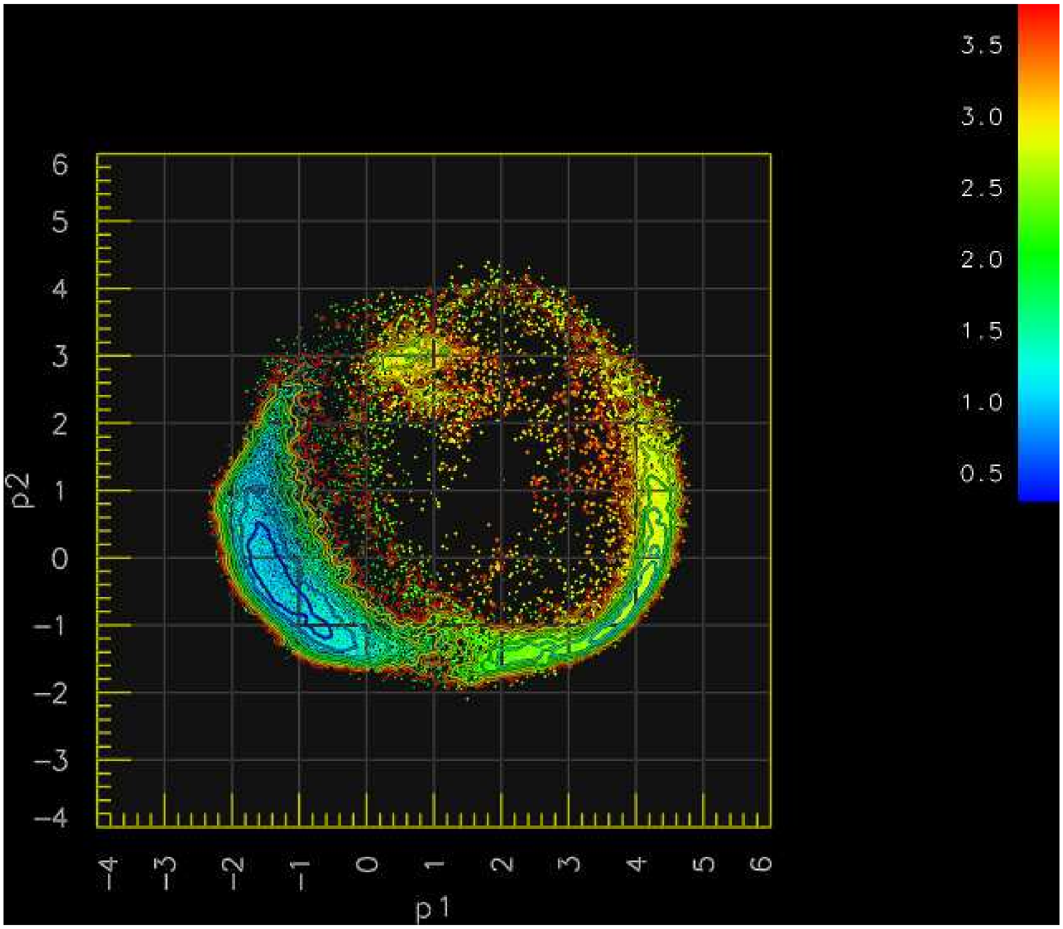

Figure 3.12: Ensemble RMSD map on the free energy landscape, which is shown as contour lines.

to the least populated area, but slightly destabilized at 394K as indicated by the range of the colorbar. The other three minima m_2, m_3 and m_4 are arranged in a circle next to the native basin and about 1.0 to 1.5 kcal/mol higher in free energy. All free energy barriers are no higher than 2.5 kcal/mol.

The ensemble RMSD values can be color-mapped to the landscape, which is shown as contour lines in Figure 3.12. Indeed, the assumed "native" conformation does occupy the global free energy minimum; RMSD values of structures in m_1 are typically below 1.5Å. If this was used to differentiate between "folded" and "unfolded", folded structures approximately accounts for 61% of the total sampling,

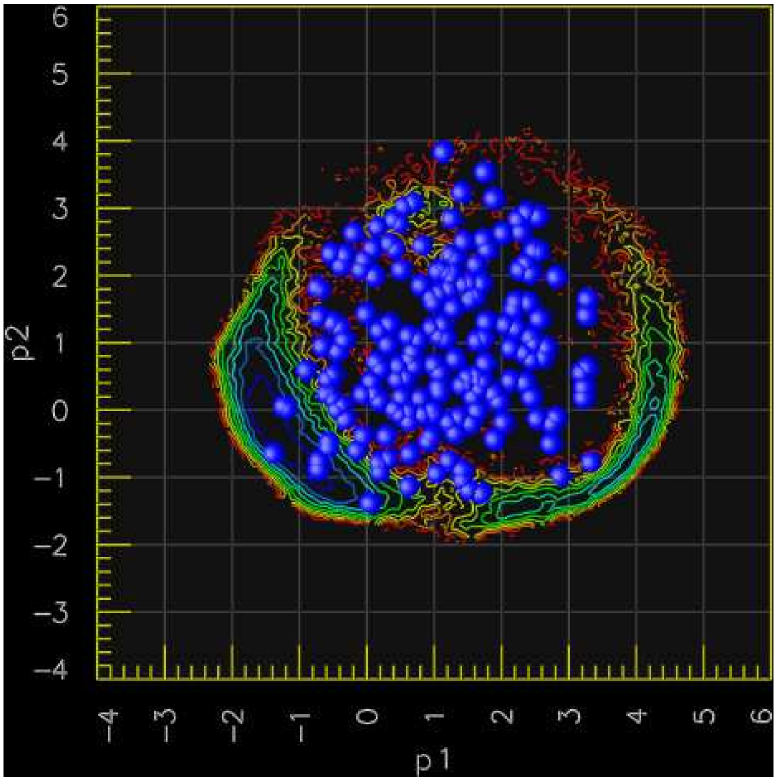

Figure 3.13: Initial folding ensemble coverage on the free energy landscape

which suggests a 0.26 kcal/mol stability of the folded state. This may explain why this peptide was previously considered unstructured from CD experiments.

188 initial structures of folding simulations and structures sampled during several folding simulations were projected onto the landscape. As expected, the initial folding ensemble quenched from the 800K MD simulation do cover most of the landscape in a uniform manner (Figure 3.13). Four folding simulations of 26ns (Figure 3.14), 49ns (Figure 3.15), 87ns (Figure 3.16), and 100ns (Figure 3.17) were similarly visualized on the landscape as well as a native MD simulation of 29ns (Figure 3.18), colored by time (blue when $t = 0$ and red in the end). The projections and the landscape agreed almost perfectly with each other, the most frequently sampled areas conforming with free energy minima identified by REM simulations. What appears more interesting is that both folding simulation No. 37

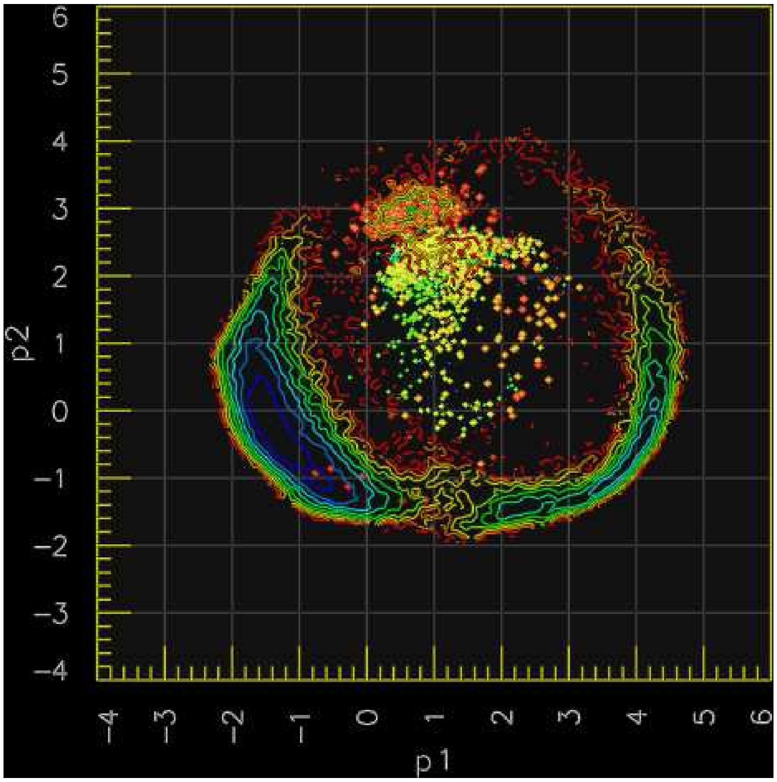

Figure 3.14: The landscape projection of folding simulations No. 35

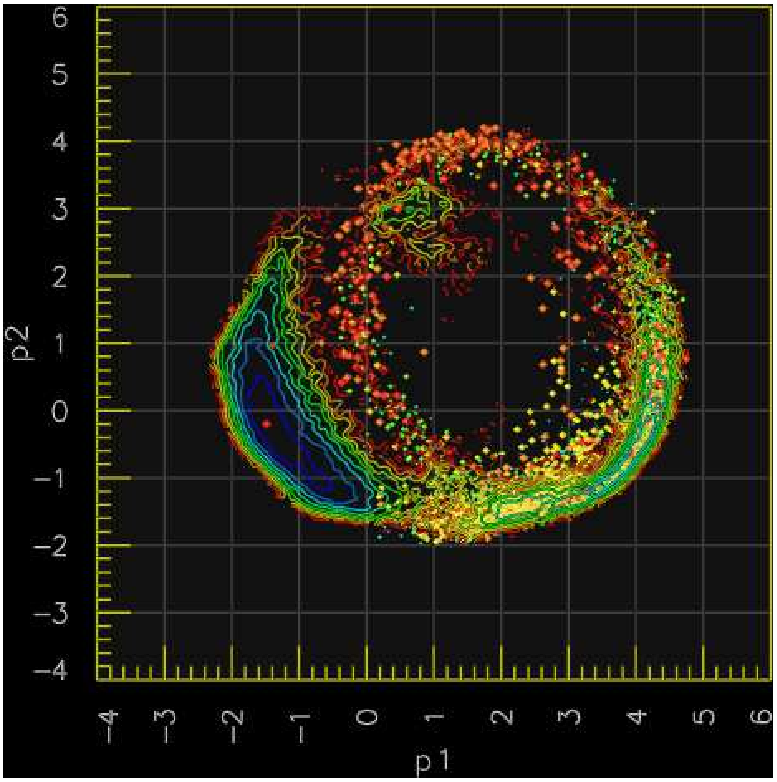

Figure 3.15: The landscape projection of folding simulation No. 37

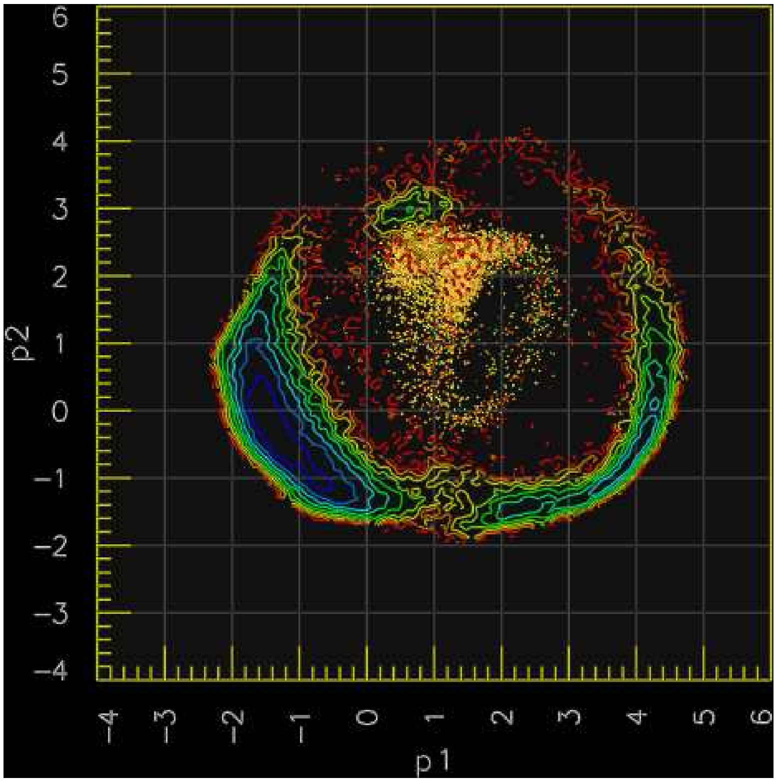

Figure 3.16: The landscape projection of folding simulation No. 60

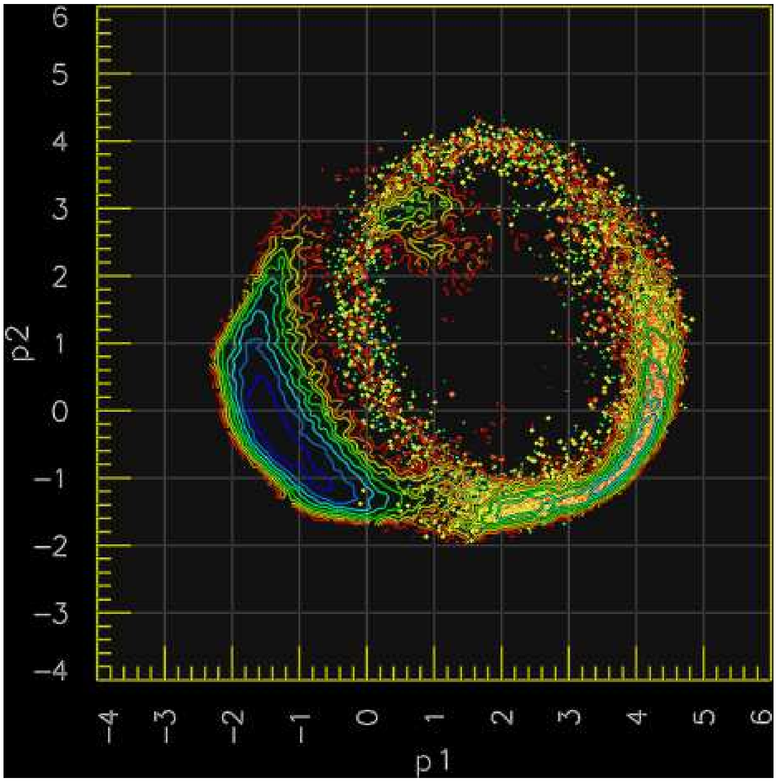

Figure 3.17: The landscape projection of folding simulation No. 80

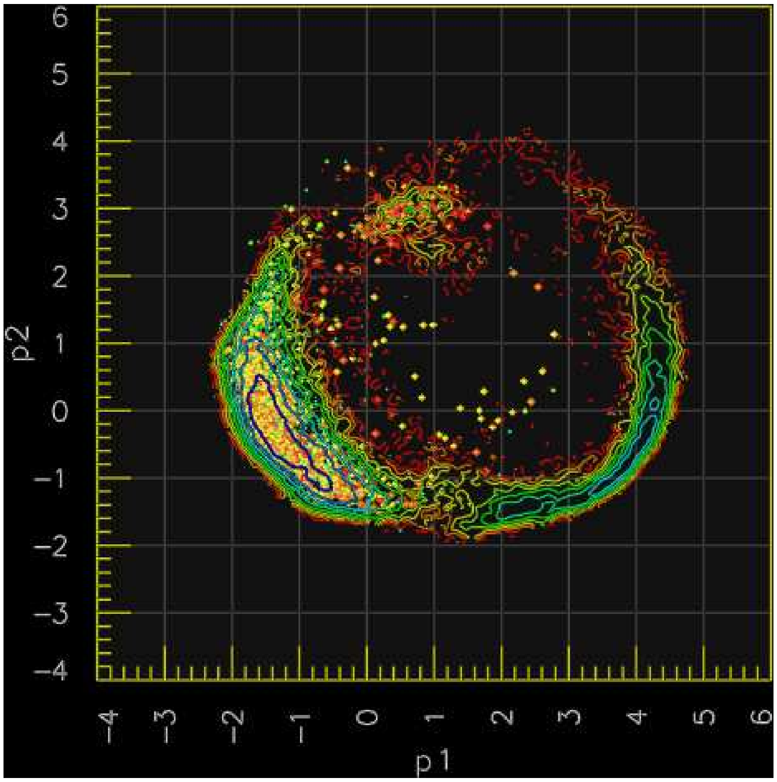

Figure 3.18: The landscape projection of a native MD simulation

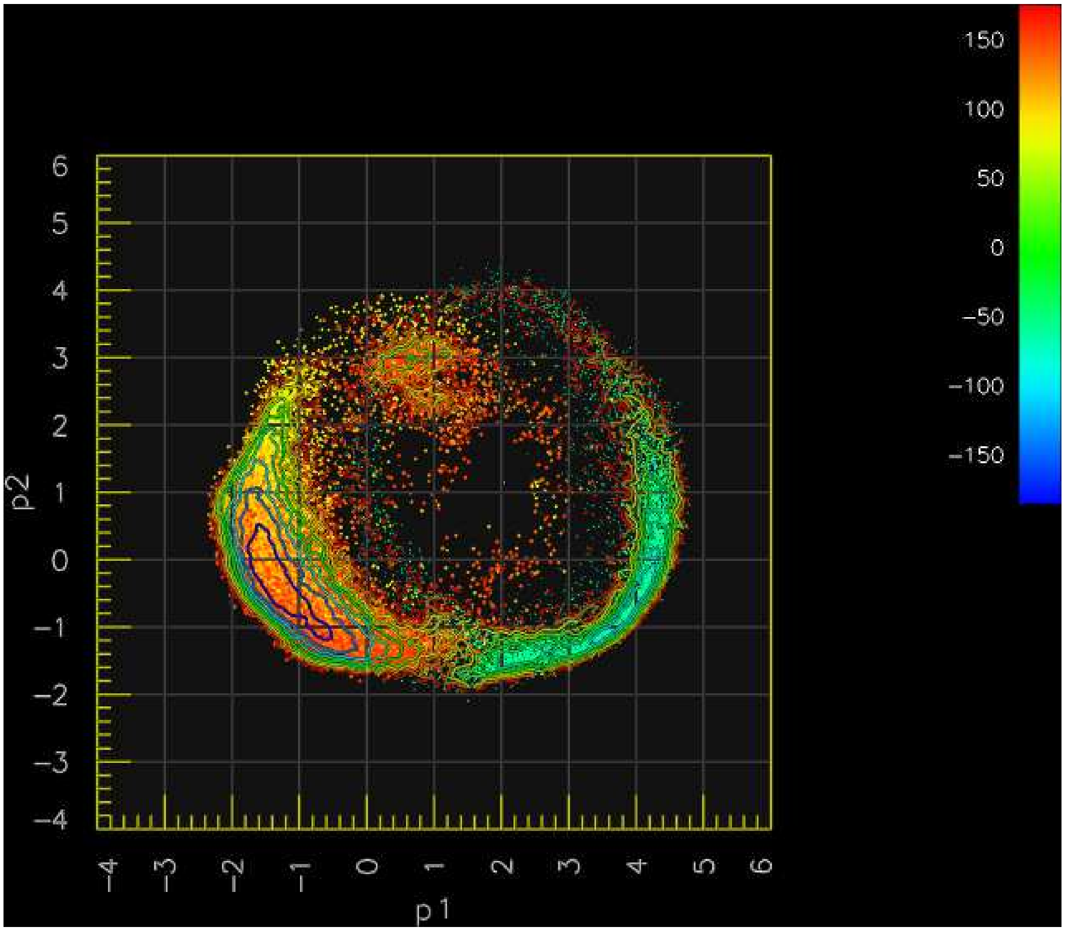

Figure 3.19: ψ_3 map colored by the torsion values indicated by the color bar. Non-native ψ_3 corresponds to m_2.

and No. 80 were similar in their kinetic path, trapped by m_2, but the folding of No. 80 was not achieved within the limit of these folding simulations – 100ns. The projection of all folding simulations suggests that most of the folding events occur by crossing the undersampled long narrow "ridge" next to m_1 (e.g. Figure 3.15) instead of the more feasible barrier between m_1 and m_2. The reason of this is not clear from these landscapes generated from PCA components.

Can one use these landscapes to explain why there are three apparent folding processes? First, ψ_3 and ϕ_8 of REM-sampled structures were mapped to the landscape (Figure 3.19 and 3.20). It clearly indicates that non-native ψ_3 and ϕ_8

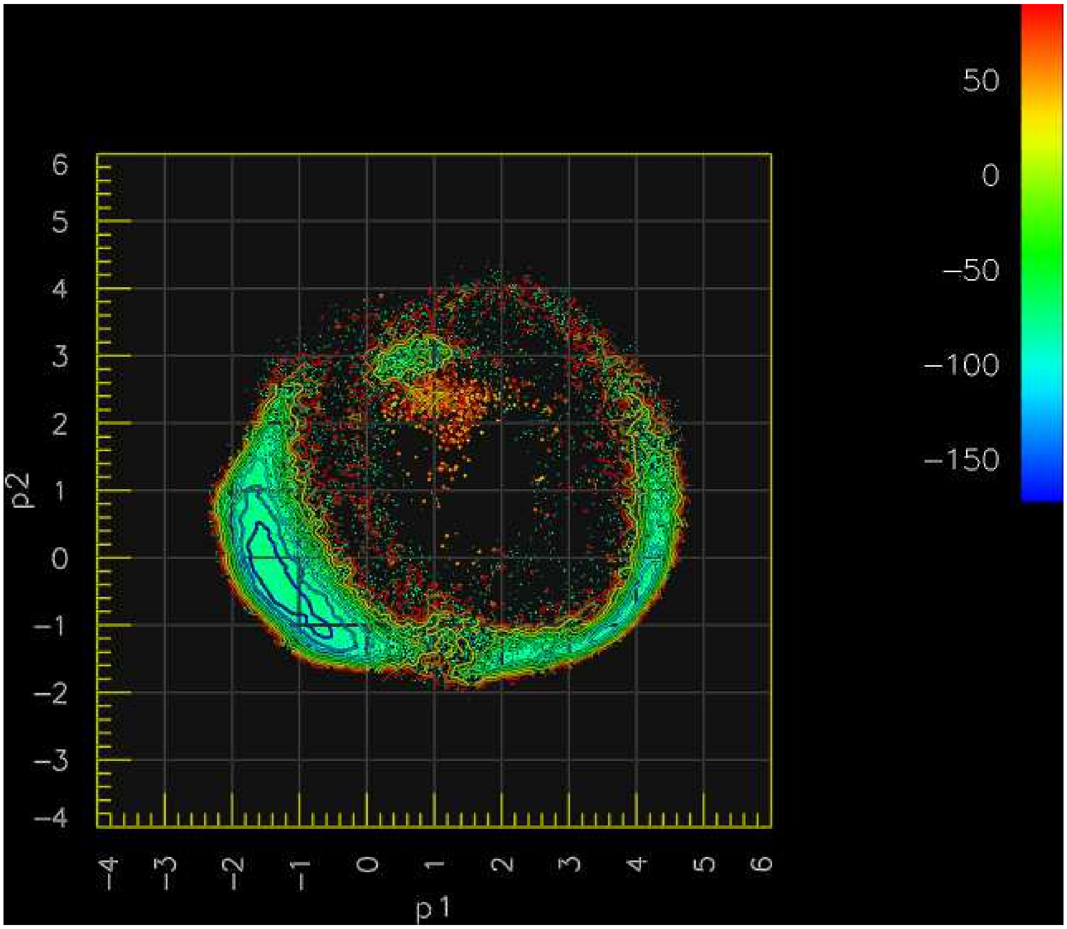

Figure 3.20: ϕ_8 map colored by the torsion values indicated by the color bar. Non-native ϕ_8 corresponds to m_4.

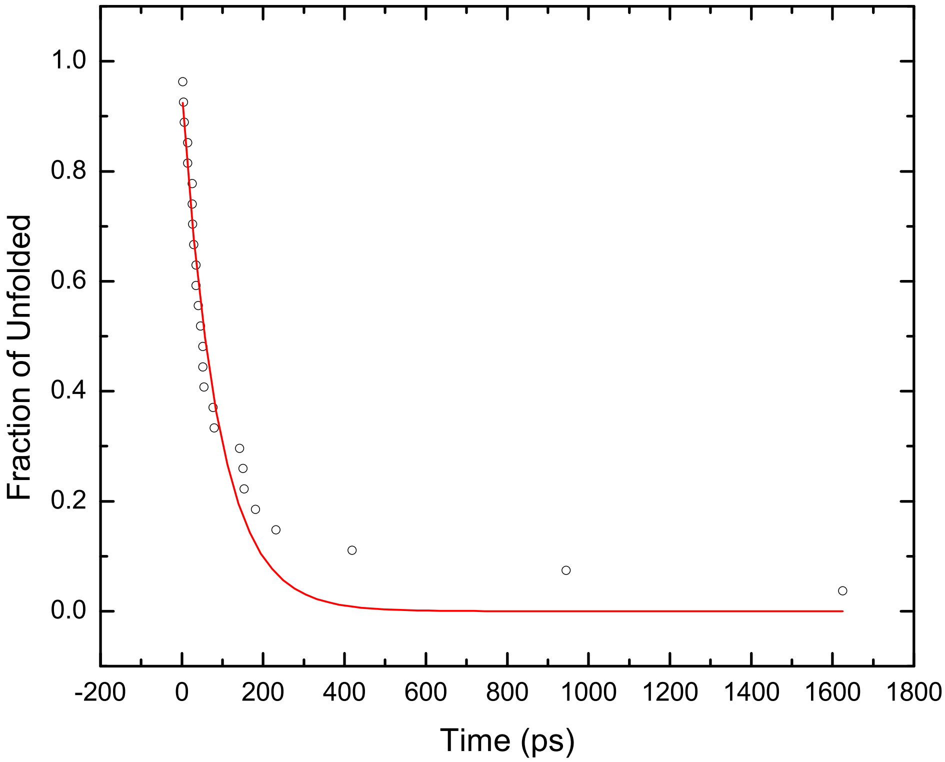

Figure 3.21: The single exponential fitting of near-native folding kinetics (the correlation coefficient is 0.9558)

distribution correspond to alternate local minima m_2 and m_4. The slow timescale reflects their transitions to the global minimum m_1, which have been shown in Figure 3.15 and 3.16.

At this point, two timescales (fast and intermediate) and one obvious non-native minimum m_3 remained unassigned. It is speculated that the fast timescale arose from the folding of near-native initial conformations (Figure 3.13), which were already inside of the native basin. This was subsequently inspected by dividing all folding simulations according to their initial RMSDs (the cutoff is 1.5Å) from the sub-ensemble with no non-native ψ_3 and ϕ_8. The two resulting sets were then fit to a single exponential with time constants of 88 ± 7ps and 1200 ± 7ps (Figure 3.21 and 3.22), which are again matching with those obtained from full-ensemble

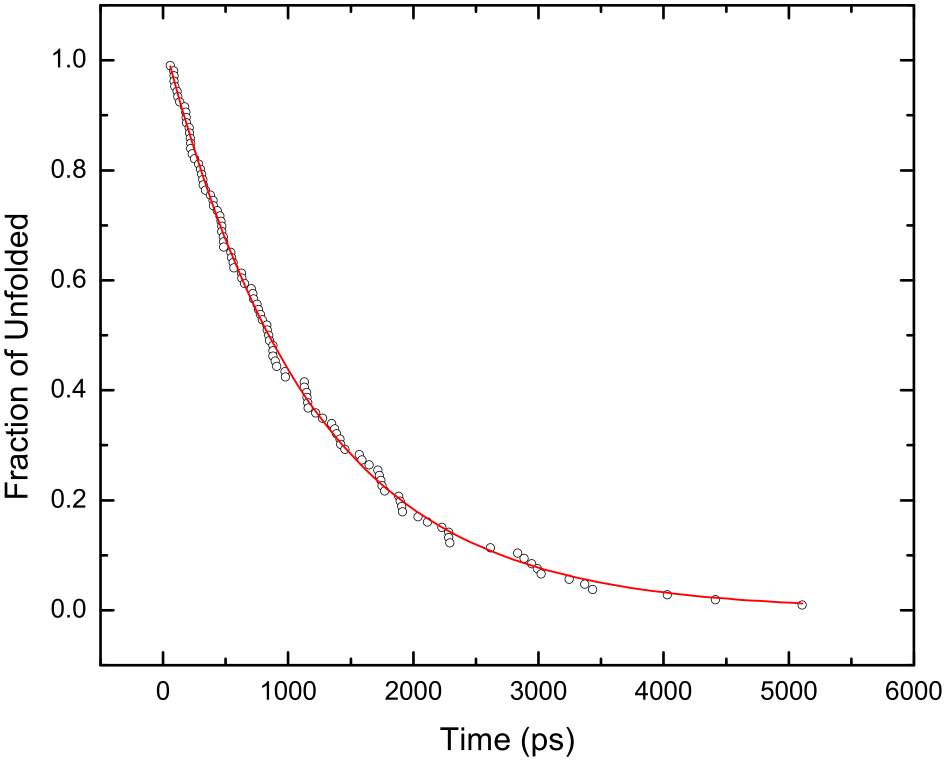

Figure 3.22: The single exponential fitting of intermediate folding kinetics (the correlation coefficient is 0.9984)

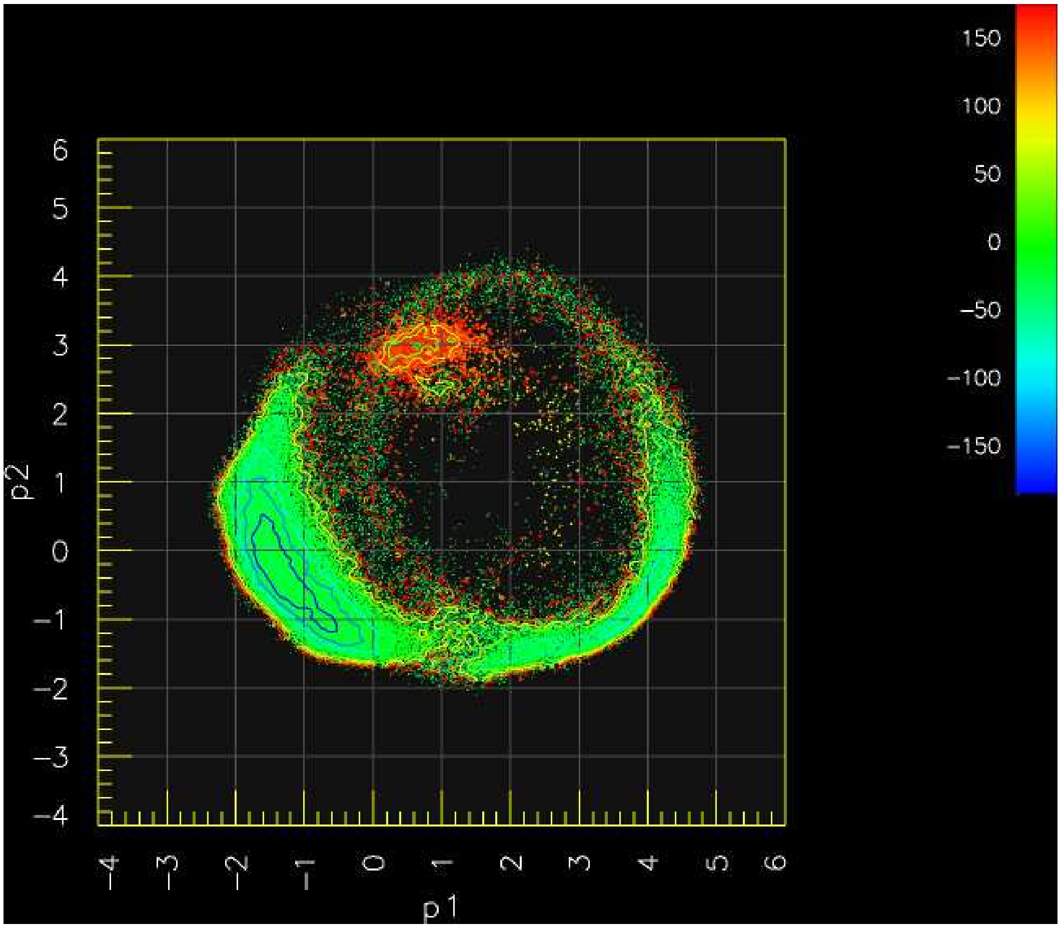

Figure 3.23: ψ_4 map colored by the torsion values indicated by the color bar. Non-native ψ_4 corresponds to m_3.

analysis. The speculation that near-native folds with the fast timescale is largely credible, but certainly not perfect. One can see from the discrepancy beyond 200ps in Figure 3.21 that some intermediate timescale folding can still happen to near-native foldings.

Examinations of ϕ and ψ of other residues allow the identification of the last minimum m_3 (Figure 3.23), which is associated with non-native ψ_4 (Pro4). Folding simulations that sampled this basin can be fit with a single exponential ($\tau = 1400 \pm 33$ps), similar to the intermediate timescale found for the full-ensemble analysis (Figure 3.24). In this case, however, only 37% of the intermediate timescale

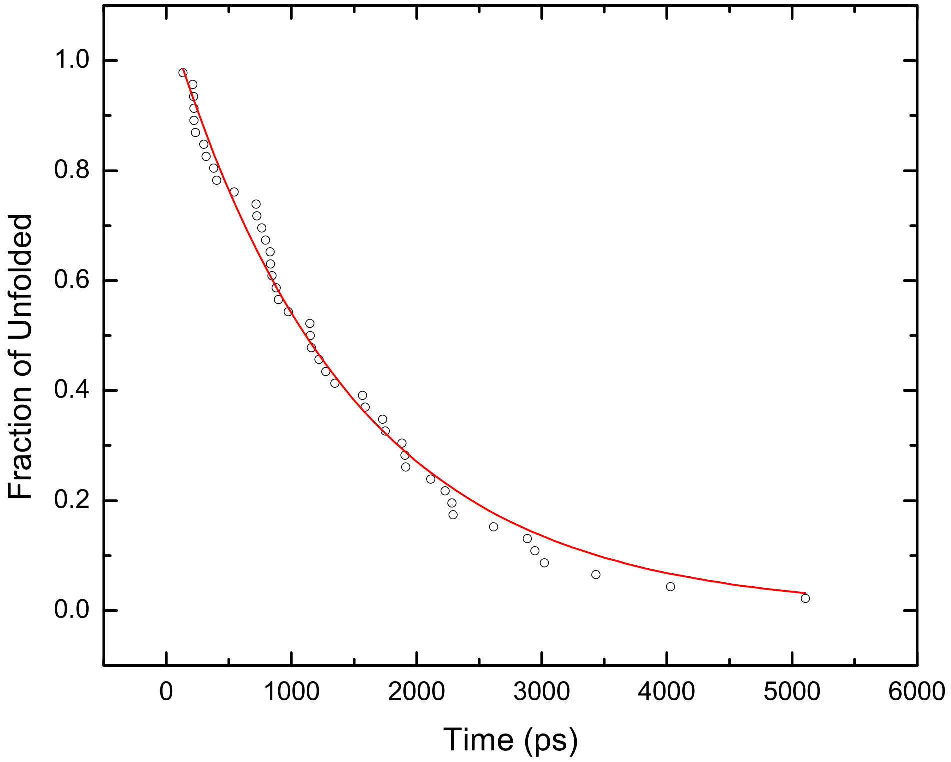

Figure 3.24: The single exponential fitting of non-native ψ_4 folding kinetics (the correlation coefficient is 0.9883)

can be explained by this minimum. The reminder of the intermediate timescale foldings do not appear to be associated with any apparent minima on the free energy landscape. Their projections have little in common, and their folding transitions mostly occur at a variety of locations on the separating ridge right to the global minimum m_1, suggesting the ensemble-nature of the transition states for this particular peptide.

To summarize the above analysis, 188 folding simulations were classified into three categories: fast (near-native), intermediate (including non-native ψ_4, and slow (non-native ψ_3 and ϕ_8). The number of folding simulations in each category is 27 (14.4%), 125 (66.5%), and 36 (19.1%) respectively, which agree well with their kinetic weights from triple exponential fitting (13.2%, 65.5% and 20.9%).

3.4 Thermodynamics Studies of Other Peptides

The thermodynamics of a few other peptides being studied in the lab were investigated as well with replica exchange method to complement their folding studies. What is different from the previous nonapeptide is that these peptides all have rather stable structural motifs (often referred as "mini-proteins") and their solution structures have been determined by NMR techniques. Extensive thermodynamic studies on these peptides can not only help to understand folding kinetics but also serve as "reality" benchmark of solvation models and force fields.

3.4.1 Tryptophan Zipper

3.4.1.1 Background

"Tryptophan zippers" (trpzip) are a series of designed 12 to 16-residue peptides, which adopt stable, monomeric β-hairpin conformations, stabilized by cross-strand tryptophan-tryptophan packing [111]. High-resolution NMR structures

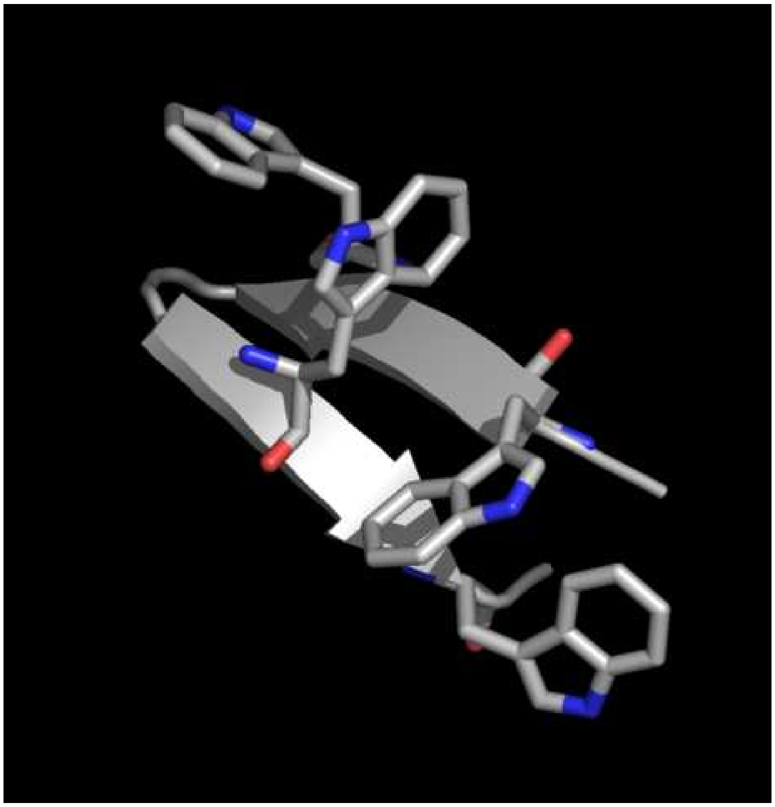

Figure 3.25: Average trpzip2 (1LE1) NMR structure

of three of members of the family (PDB codes: 1LE0, 1LE1, and 1LE3) showed that two cross-strand Trp pairs stack in a zipper-like motif on the surface of the folded peptides (Figure 3.25, 1LE1 shown here). It was claimed that the trpzip peptides are equally stable as much larger proteins on a residue-based comparison ($\Delta G_{unf,max} = 60 - 120\,\mathrm{cal} \cdot \mathrm{mol}^{-1} \cdot \mathrm{residue}^{-1}$). For example, trpzip2 (sequence: SWTWENGKWTWK), the most stable among other 12-residue trpzips, has a melting temperature $T_m = 345.0 \pm 0.1\,\mathrm{K}$ and $\Delta C_p = 281 \pm 2\,\mathrm{kcal} \cdot \mathrm{mol}^{-1} \cdot \mathrm{K}^{-1}$.

The folding and unfolding of trpzip2 have been extensively studied by Okur and Roe using all atomic MD simulations with GB continuum solvent model and a new set of backbone force field parameters [14]. One of the major findings in kinetic studies is that unfolding appears to be a single-exponential process while the

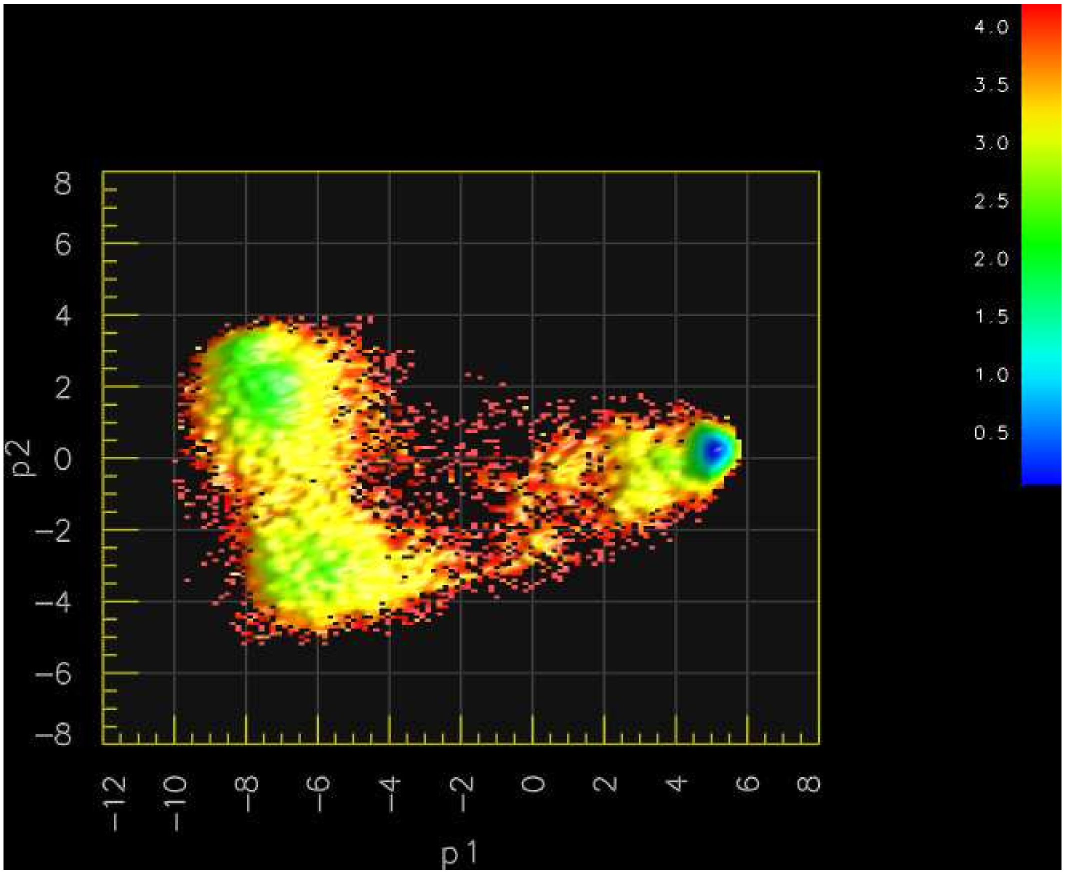

Figure 3.26: Free energy landscape of trpzip2 at 308K

reverse process started with high-temperature ensemble is double-exponential[11] [112]. Similar to the nonapeptide folding studies, REM simulations were carried out for trpzip2 to provide thermodynamic pictures of folding. Replica temperatures are 272, 290, 308, 329, 350, 373, 397, and 423K.

3.4.1.2 Thermodynamics of Folding

The surface representations of the free energy landscapes at 308, 329 and 350K are shown in Figure 3.26, 3.27, and 3.28. The free energy landscape is composed of two large most sampled regions at 308K, separated by a wide and barely sampled

[11] Both folding and unfolding simulations were conducted at 350K, slightly higher than the melting temperature.

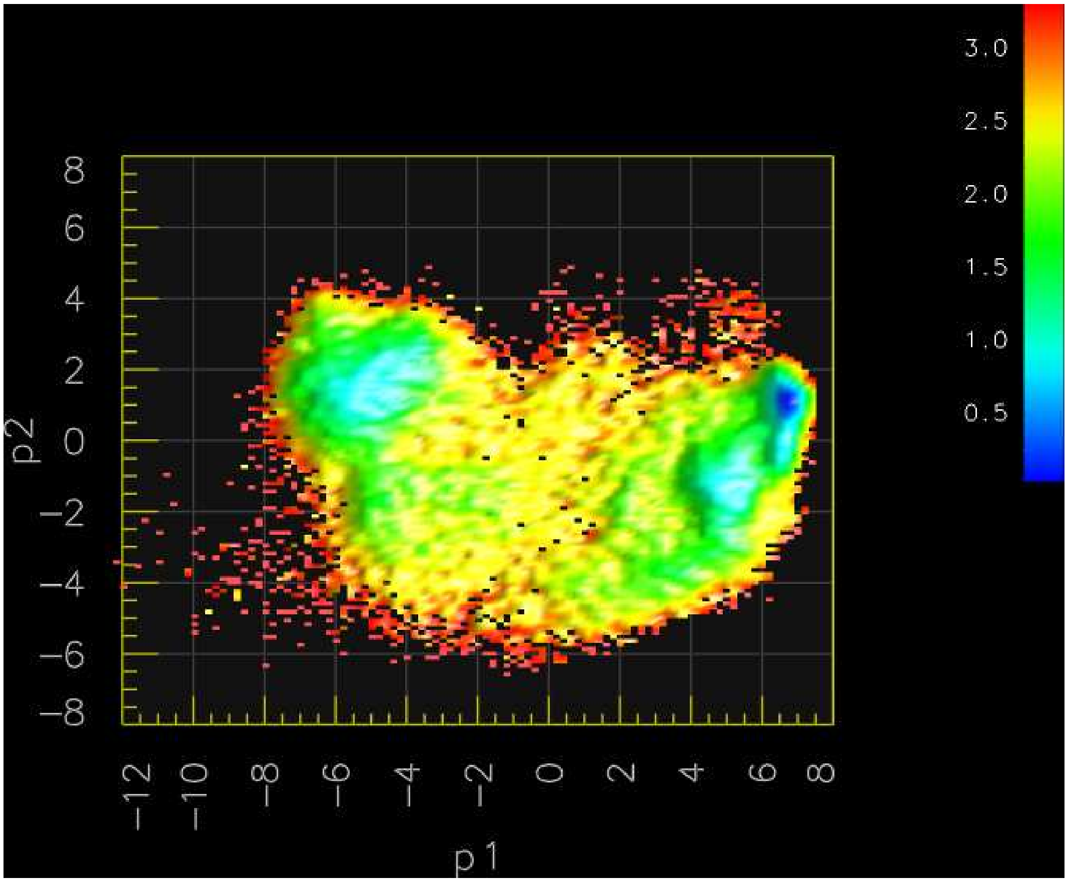

Figure 3.27: Free energy landscape of trpzip2 at 329K

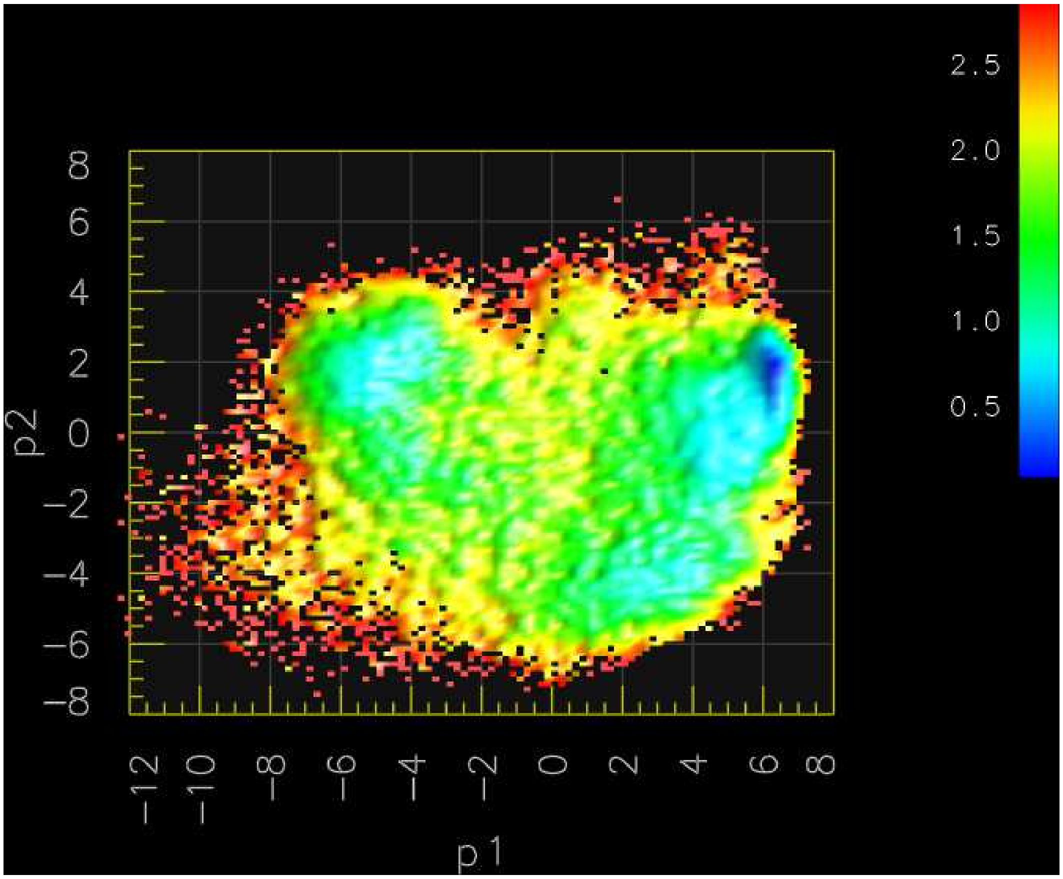

Figure 3.28: Free energy landscape of trpzip2 at 350K

“ridge” region that becomes flat and populated gradually as the temperature is increased, similar to what has been observed in the nonapeptide case. However, the difference between the designed peptide and the nonapeptide is very significant. First, the landscape is more funnel-shaped around room temperature and the global minimum is much deeper (about 4 kcal/mol) and narrower, on contrary to the shallow and wide native state of the 9-residue peptide. Due to this strong bias, trpzip2 folds in a highly cooperative manner. The first principle component (PC) has an magnitude of 75.7% while the second PC only has an magnitude of 6.0%. In the case of the 9-residue peptide, the magnitudes of the first two PCs are 45.6% and 15.6%. This indicates that the folding of trpzip is mainly dominated by one kind of motion, unlike the 9-residue peptide where a few exist[12]. It has been speculated that millions of years of evolution selected a very small portion of protein sequence-space, which can form quickly their desired structures required by their functions without being delayed by local traps. This has been at least partly demonstrated in the computational folding studies of these two peptides of similar length, although quantitative agreements of thermodynamic properties, such as melting temperature, was not reached. The computed melting curve of trpzip2 is shown in Figure 3.29, with a melting temperature about 40K lower than the experimental one. This might be explained by the large local minimum on the left of the “ridge”, which corresponds to a partially left-handed α-helix (Figure 3.30). Oddly enough, none of the folding and unfolding simulations (over 100) sampled this conformation at 350K. But an MD simulation started with this left-handed α-helix stayed and did not fold to the native conformation even after 50ns.

[12] In fact, the third and fourth PCs of the 9-residue peptide shave magnitudes of 11.2%, 7.5% while the second and third PCs of trpzip2 are only 4.2% and 3.9% respectively.

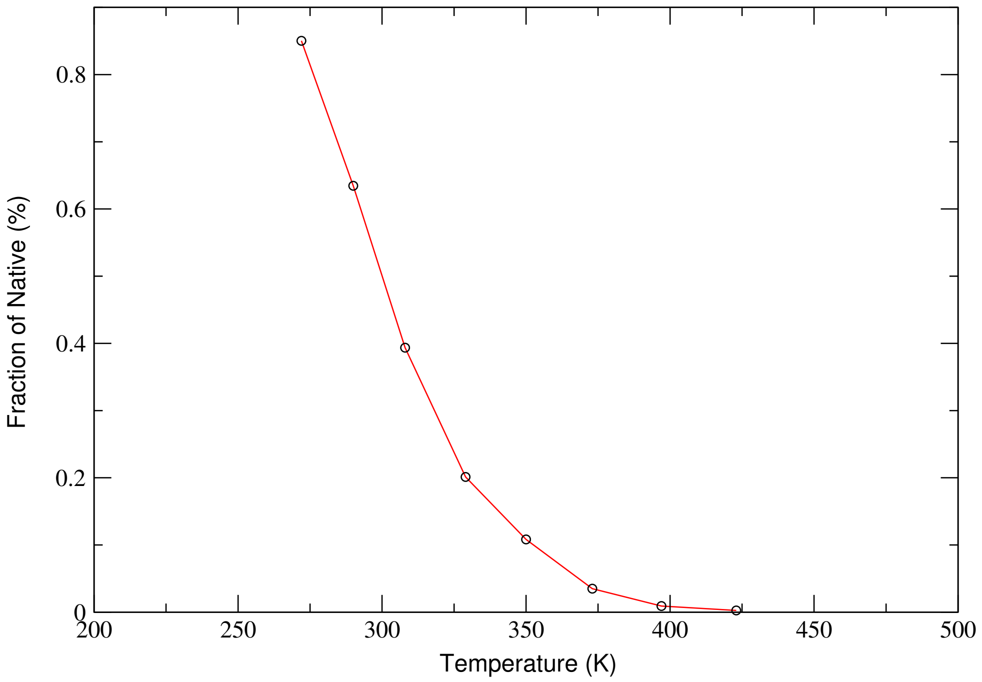

Figure 3.29: The computed melting curve for trpzip2 from REM simulations

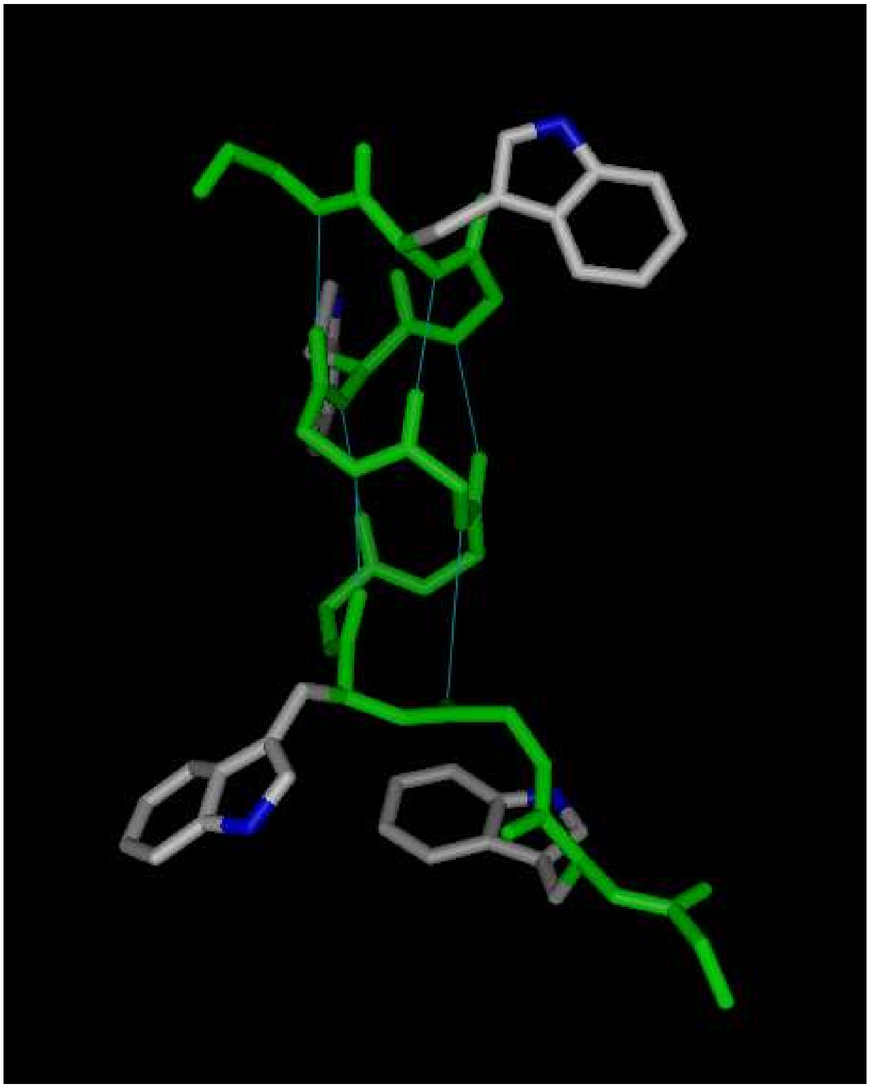

Figure 3.30: A snapshot of "mis-folded" trpzip2 in left-handed α-helix conformation. The backbone atoms are shown in green sticks and (i, i+4) hydrogen bonds are shown in cyan lines. Only four Trp side chains are shown.

3.4.2 Tryptophan Cage and Mutants

3.4.2.1 Background

Recently Neidigh and Andersen [89] reported the successful design of a 20-residue mini-protein – “Trp-cage”, from a poorly folded 39-residue peptide exendin-4 (EX4) of Gila monster saliva. According to their understanding of “self-folded domain”[13], Trp-cage is well qualified and “is significantly more stable than any other miniprotein reported to date”. Trp-cage (sequence: N^1LYIQWLKDGGPSSGRPPPS^{20}) is >95% folded in water at physiological conditions, featured with a short α-helix (residue 2 to 8) followed by a 3_{10}-helix (residue 11 to 14), and a polyproline II helix at C-terminus. The overall fold is stabled by a compact hydrophobic core formed by three proline residues and a glycine packed against the aromatic side chains of Tyr3 and Trp6 (Figure 3.31). Trp-cage folds spontaneously and cooperatively with a folding rate of $4\mu s$ measured by laser temperature-jump spectroscopy [90], being the fastest observed among other complete proteins.

Simmerling and Roitberg [13] performed an *ab initio* structure prediction using molecular dynamics and successfully achieved the NMR structure from only the sequence information (all heavy-atom RMSD of 1.1Å) before it was deposited in PDB (code: 1L2Y). As the very first accomplished full-atom structure prediction, it greatly shortened the gap between experimental and modeling studies. Because of the protein-like nature of Trp-cage and its appealing size, several independent computational folding studies have been carried out to understand the underlying folding mechanism. Alongside of these, Andersen et al. are continuing to push the envelope, further optimizing the sequence to improve the stability of the fold. In this section, the thermodynamic studies of Trp-cage and a few of its variants were

[13] Multiple secondary structure elements; tertiary interactions; well-defined χ_1 and χ_2 values; better protected backbone amide proton exchange.

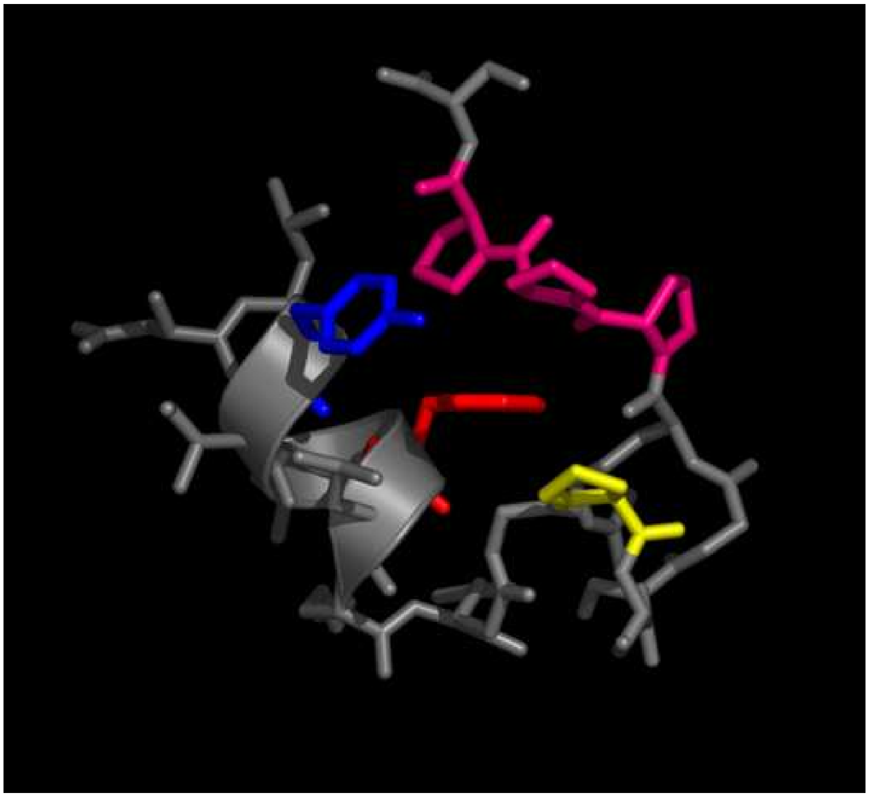

Figure 3.31: Average Trp-cage NMR structure. The hydrophobic cluster are highlighted: Tyr3 in blue, Trp6 in red, Pro12 in yellow, and Pro17-Pro19 in magenta. The α-helix (residue 2 to 8) is shown as cartoon.

reported, including:

tc5b(Trp-cage) N^{1}LYIQWLKDGGPSSGRPPPS^{20}

tc4a D^{1}LFIEWLKNGGPSSGRPPPS^{20}

tc10b D^{1}AYAQWLKDGGPSSGRPPPS^{20}

tc10f D^{1}AYAQWLKDGAPSSGRPPPS^{20}

tc10e D^{1}AYAQWLKDG^DAPSSGRPPPS^{20}

3.4.2.2 Thermodynamics of Folding

REM simulations were performed for each of the five sequences with the same setup: 8 replicas at 267, 283, 300, 318, 338, 358, 380 and 403K, starting with extended conformation built from sequence using GB continuum solvation. Collected structures (over 70,000 per replica for each sequence) were analyzed as in the nonapeptide studies, in which only the second half were used. RMSD calculations were made to the backbone of residue 3 to 18 fitting to the NMR structure of tc5b.

Experimental thermodynamic studies indicated that the relative stability measured by tryptophan chemical shift upon tertiary fold formation (%-folded at pH 7 and 280K, T_m) of the five sequences studied is tc10b (99.7%, 61C) > tc5b (98.5%, 43C) > tc10e (96%, 38C) > tc4a (46%) > tc10f (< 10%). tc10f is suspected to have a different fold, indicated by the low percentage of fold. Here, the primary interest of the following thermodynamic studies is not to reproduce the absolute experimental stabilities, rather, more pertinently, the relative tendency.

To calculate the native population from REM data, RMSD was used as a simple indicator of the nativeness. To choose a proper cut-off value, both RMSD distribution at 300K and the free energy landscape colored by RMSD were consulted. As

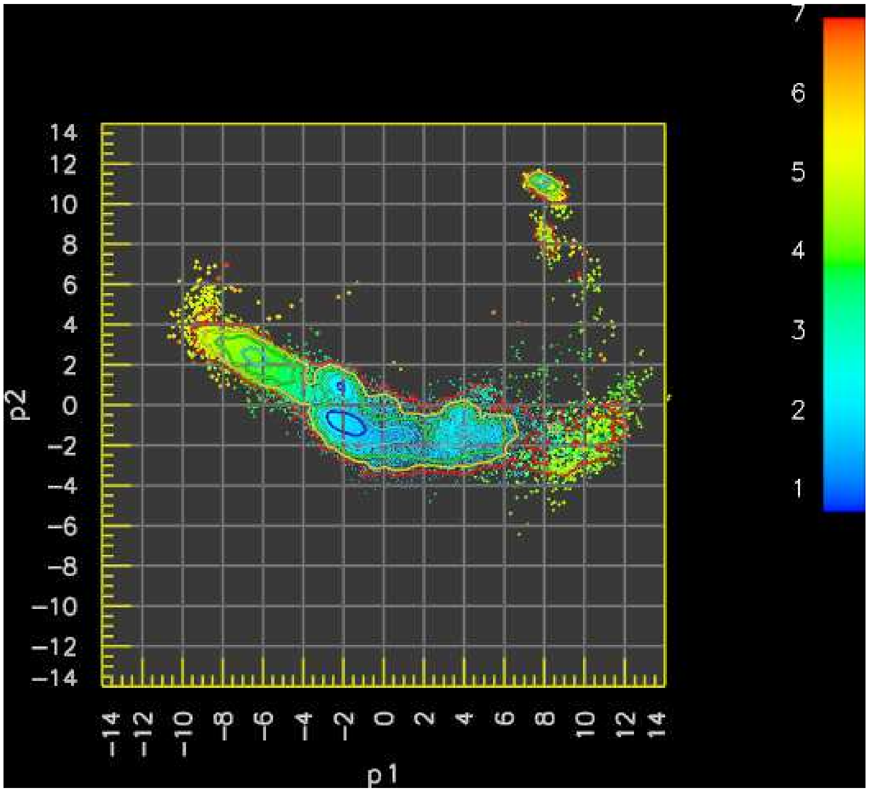

Figure 3.32: RMSD map of tc5b REM samples at 300K. The landscape is shown in contour line. The color bar for RMSDs is shown on the right.

one can see from Figure 3.32and 3.33, the RMSD histogram and its landscape map are rather consistent, with the low RMSD structures close to the global free energy minimum, and 2.5Å seems to be a good cut-off value. The calculated melting curves of the five sequences are shown in Figure 3.34. It is not hard to tell that the overall agreement with %-folded at pH 7 and 280K does not appear attractive. However, the simulation does indicate that tc5b and tc10b are among the most stable sequences. The melting curves of tc5b and tc10b almost coincide with each other and the calculated melting temperature of tc5b is about 30 degrees higher comparing to the experiment, but still much lower than a similar study recently reported by Pitera and Swope [100], in which only 4,000 conformations were collected for each replica. At low temperatures, the calculated fraction of folded was

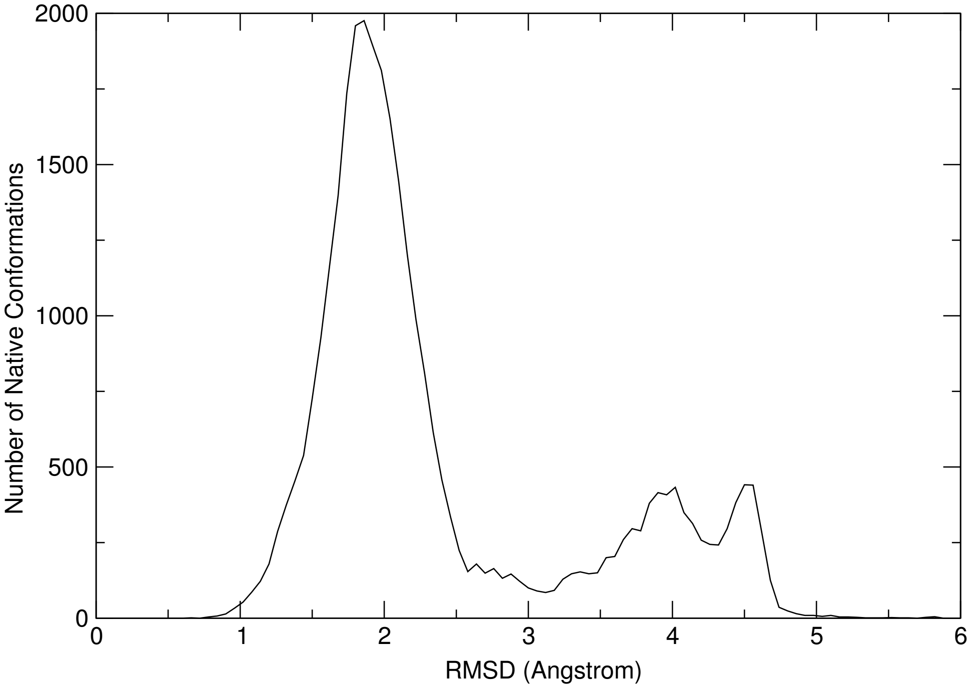

Figure 3.33: RMSD histogram (100 bins) of tc5b 300K REM data

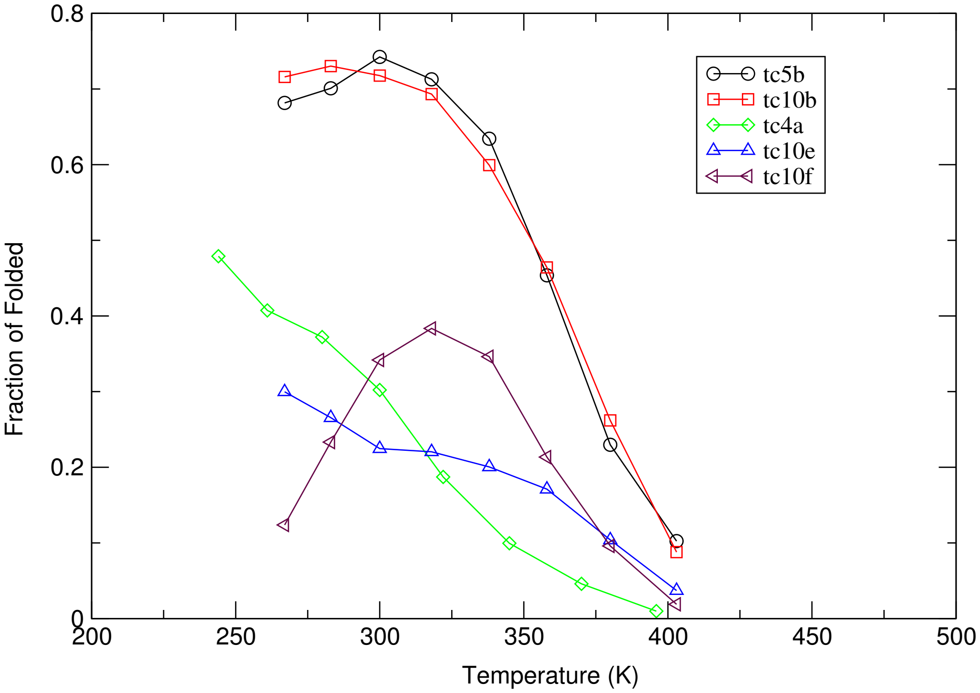

Figure 3.34: Melting curves of five Trp-cage peptides, assuming that all have the tc5b fold.

lower than that of the experiment by nearly 20%, which suggests that the simulated transition is less biased than it should. This might be explained by the lack of entropy contribution from solvent molecules, which is missing in continuum solvation. tc10b was designed to be more stable than tc5b by increasing the stability of the N-terminal α-helix, but this difference was not captured by the force field used in this study. In fact, the N-terminal α-helix population of tc10b and tc5b is 80% and 88% respectively, calculated from backbone RMSD of residue 3 to 8.

All other sequences (tc4a, tc10e and tc10f) have significantly lower melting temperatures, qualitatively in agreement with design. Their melting curves do not possess a sharp drop of fraction of folded. Interestingly, the melting temperature of tc10f exhibits a peak around 320K and a sudden decline of fraction of folded at low temperature range. This would not be surprising if the tc5b native conformation is not dominant on the landscape. Indeed, four other minima coexist with the global minimum (Figure 3.35), which is not much favored energetically and the RMSD landscape mapping (Figure 3.36) clearly reveals that low RMSD structures (relative to tc5b native conformation) do not correspond to the global free energy minimum. Instead, one structure that belongs to the global minimum was filtered off (Figure 3.37), from which the RMSDs were recalculated, and low RMSD structures nicely filled the global minimum (Figure 3.38). With most of the tc5b secondary structure elements present, this tc10f "native" conformation features a strong salt-bridge formed between N-terminal Asp and Arg16, which is largely exposed to the solvent. This questionable salt-bridge is very likely an artifact of the continuum solvent model used, which may underestimate the screening effect between charged residues in certain cases. This phenomenon has been repeatedly observed with the GB implementation in AMBER 6. In fact, it exists in every Trp-cage sequence except tc5b, which does not have a N-terminal Asp. This overstabilized salt-bridge directly affects the tertiary conformation of the peptide, i.e. secondary

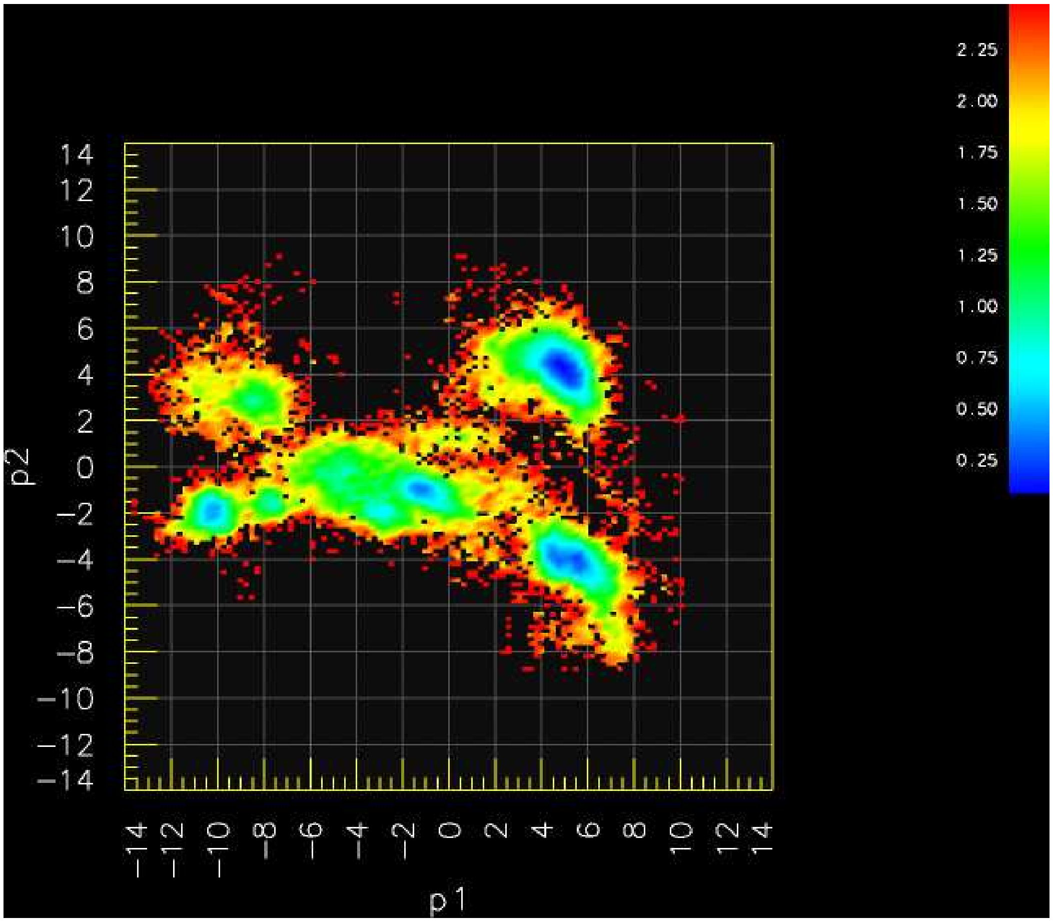

Figure 3.35: The free energy landscape of tc10f at 300K colored by relative free energies.

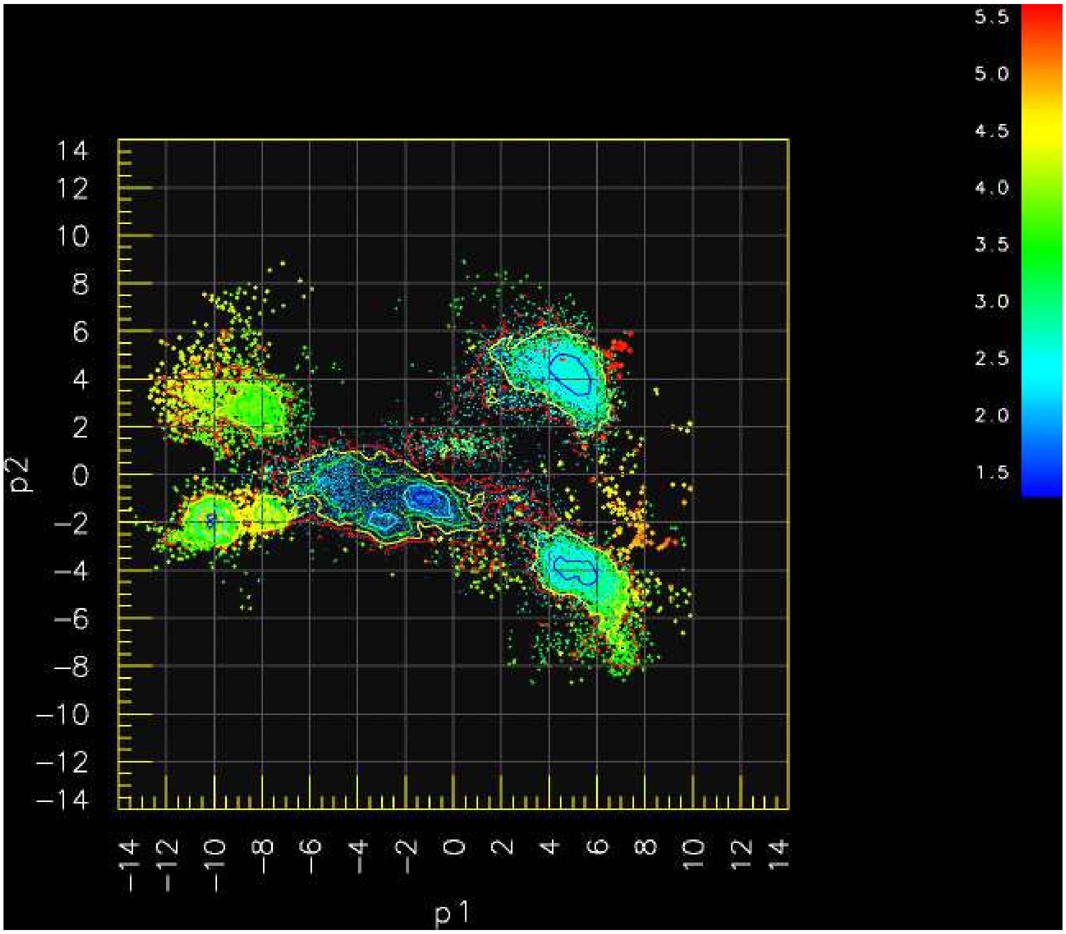

Figure 3.36: The RMSD mapping for tc10f, relative to tc5b native conformation

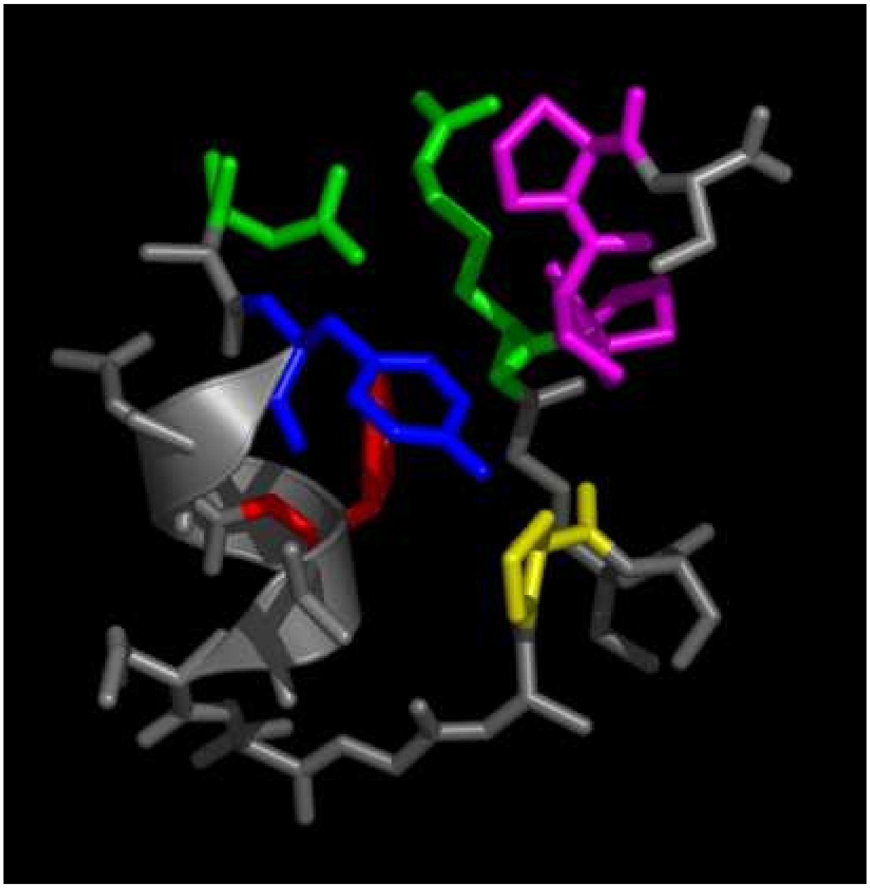

Figure 3.37: The global minimum conformation for tc10f, similarly color-coded as Figure 3.31. The salt-bridge formed between Asp1 and Arg16 are highlighted in green.

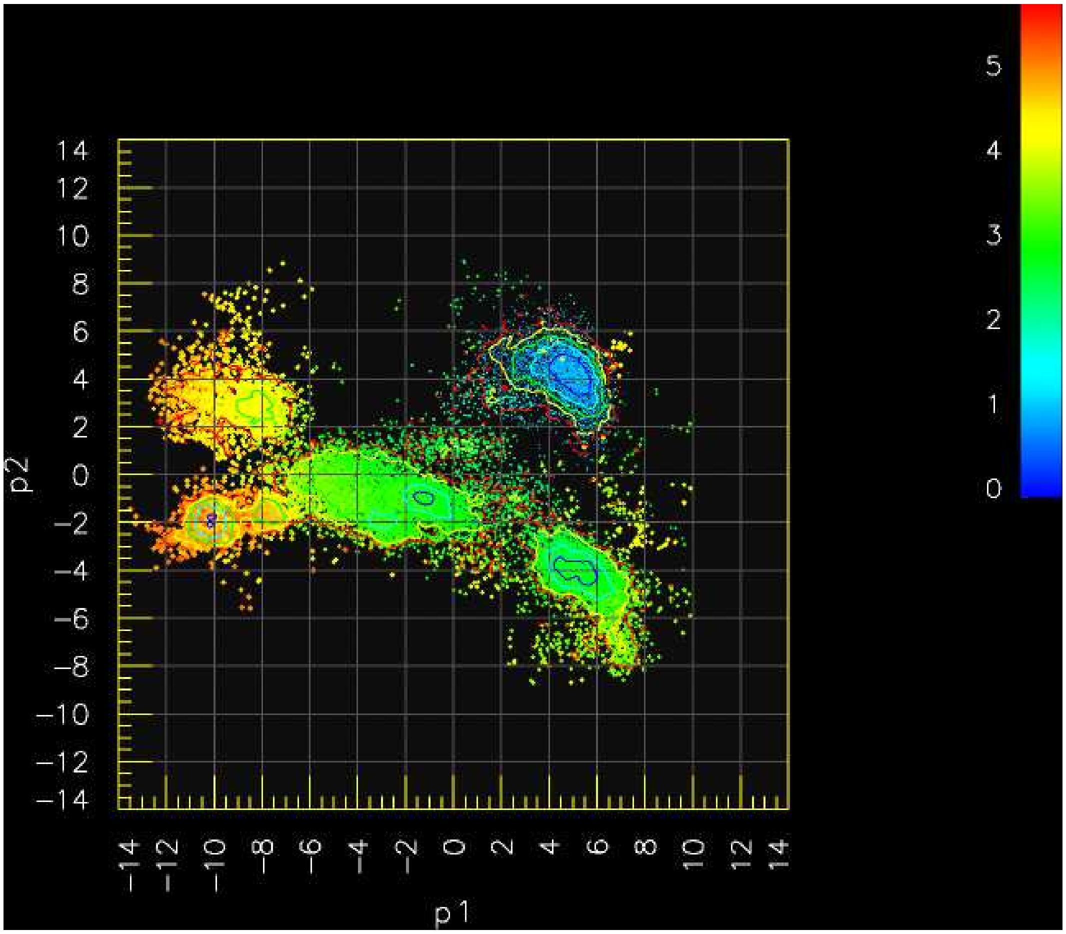

Figure 3.38: The RMSD mapping of tc10f, relative to its global minimum conformation.

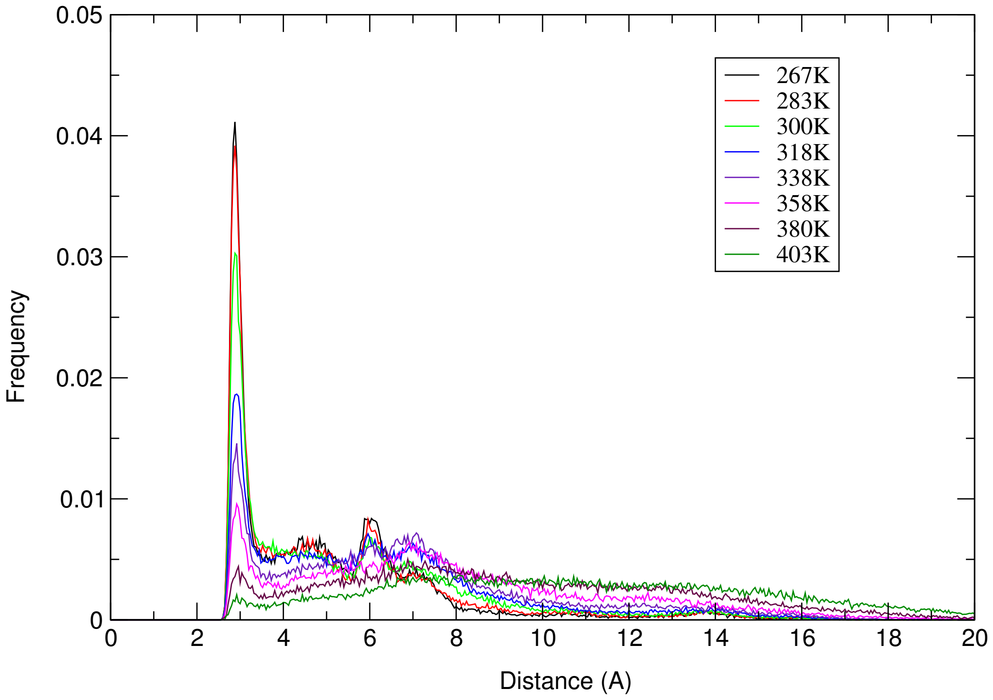

Figure 3.39: The distance (Trp6@NE1–Asp16@O) distribution of tc5b.

element packing. Tertiary structural information was experimentally measured by Trp6 chemical shift upon hydrogen-bonding with carbonyl oxygen of Arg16, and can be equivalently monitored using the distance between the hydrogen bond donor (NE1) and the acceptor (O). A nearly single distribution of this distance was obtained for tc5b over a range of 40Å, centered at about 3Å (Figure 3.39). However, a second peak was found present at about 5Å for all other sequences (see Figure 3.40 for an example)[14].

[14] This distance of the representative structure for tc10f native state mentioned above was measured as 4.7Å.

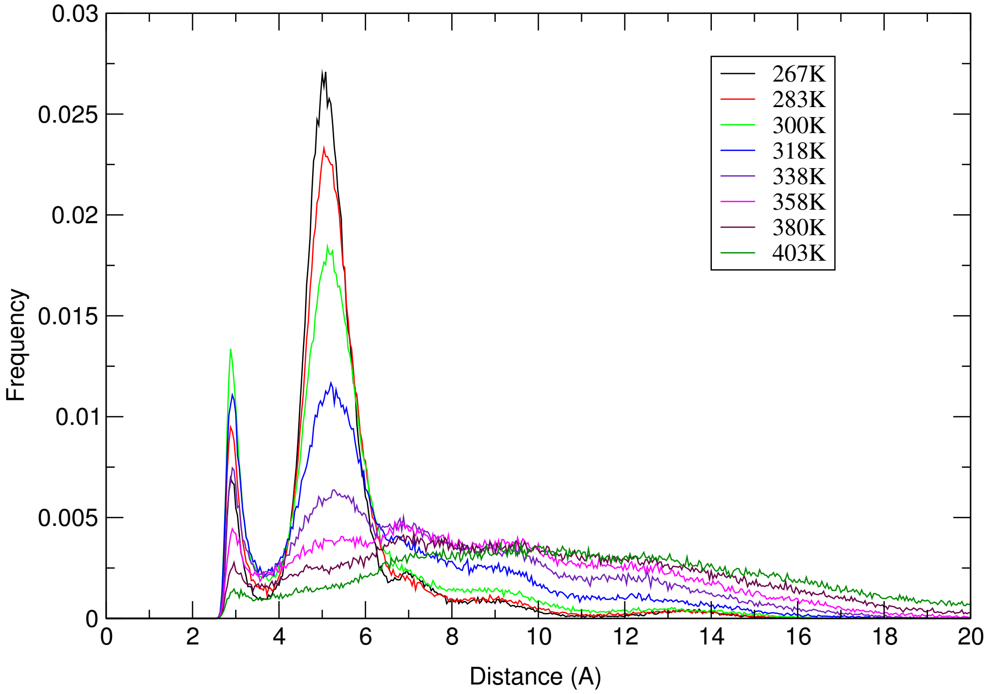

Figure 3.40: An example (tc10f) of the distance (Trp6@NE1–Asp16@O) distribution shows a much larger peak centered at about 5Å.

3.5 Conclusions

The folding of a few peptides that are different in length, fold and thermodynamic stabilities have been studied with replica exchange approach. Much has been learned not only in the conceptual understanding of protein folding but also in the strength and weakness of methodologies that are available for studying protein folding.

3.5.1 Conceptual Understanding

Protein folding is certainly a highly complicated process that involves coordinated movements of thousands of atoms, including solvent. If one universal folding model exists to explain the folding of all protein sequences, it undoubtedly has to be built on the basis of statistic models, which has established that the folded state of protein molecules corresponds to the global free energy minimum of a high dimensional space. This space, when "compressed" or reduced down to only one or two dimensions, bears the shape of a funnel, the bottom of which represents the folded state. Folding process is greatly accelerated due to the shape of the funnel and becomes thermodynamically favorable. Unfolded state is not a single state but an ensemble of microscopic states with similar energies. Because of this, it may not be appropriate to interpret folding process as static pathways and more importantly, the apparent kinetics of folding relies on not only the thermodynamics of folding but also how folding is initiated. The nonapeptide in the first study poses an almost ideal model system that could be examined thoroughly with all atomic details. Owing to its unoptimized sequence, it exhibits at least four different folding scenarios that could be characterized (the fast, two different slow timescales and a part of the intermediate timescale). The weight of each timescale is decided by the 800K MD simulation that generated the initial folding ensemble. The contribution from non-native ϕ_8 might completely disappear or become undetectable if the initial ensemble had been created at a much lower temperature, at which non-native ϕ_8 is only a negligible portion of the system[15]. This has been recently supported experimentally that cold shock protein A exhibits different folding kinetics depending on how the laser induced temperature jump experiments were done [109]. A single exponential kinetics became double exponential when the temperature was raised 5 degrees higher.

Then why can so many natural proteins be described with a simple two-state model? The apparent two-state folding model can be simply a result of how the folding process is observed, in other words, the choice of progress variables. In the nonapeptide case, if only ψ_3 or ϕ_8 had been used, the folding kinetics would have been an apparent two-state. Computer modeling of protein folding is obviously more advantageous comparing to experimental studies in this regard.

To have a correct picture of the folding process of a particular protein, it is important to obtain both unbiased kinetics and thermodynamics with sufficient statistics. Unlike small molecule reactions, knowing only the thermodynamics does not imply knowledge of the actual kinetics of protein folding. Instead, many folding simulations that start from representative locations in the phase space are needed and required to proceed to complete. The amount of folding simulations can be systematically identified following the strategy proposed by Brooks [86]. A significant portion of the folding simulations must reach the folded state. Otherwise, slow timescales are likely to be overlooked, therefore resulting in a premature conclusion.

[15] 500 out of 100,000 structures collected at 300K using REM have the non-native ϕ_8 conformation.

3.5.2 Methodology

Replica exchange method is a much more effective sampling approach that regular molecular dynamics cannot match. This has been adequately demonstrated in this chapter. In the nonapeptide study, twice data collected in regular MD simulations did not cover as much as the REM simulations. Some of the folding simulations were trapped by local minima for 100ns (equivalent to the length of the REM simulations) and failed to find the folded state (see Figure 3.17 for an example). In the trpzip2 study, the left-handed α-helix conformation identified on the REM landscape was never found in almost 100 folding and unfolding MD simulations. However, the improved sampling efficiency comes with a cost. Running REM simulations poses huge requirement on computational resources[16], which increases along with the problem size[17], and yet usually only a few replicas out of N has a relevant temperature that are most interesting. Therefore it is not very practical to study the folding of large proteins unless much simpler models are used. Many early REM applications are typically very short (only a few nanoseconds per replica), which might not be very careful practice. System properties, such as fraction of folded (Figure 3.41), calculated from short samplings may not represent that of an equilibrated system.

The increased sampling efficiency not only advances the understanding of conformational heterogeneity, but also starts to reveal flaws in protein force field and solvent models. Accurate description of conformations other than the native is crucial to study processes that involve conformational transitions. However, due to the limitation of sampling efficiency of regular MD simulations, the traditional force field benchmark approach, the description of non-optimal or transient con-

[16] Needless to say, it also needs significant amount of graduate student's time to analyze the data.

[17] As larger systems are studied, more replicas will be needed to cover a useful temperature range so that there is enough energy overlap between adjacent pairs that ensures a reasonable exchange ratio.

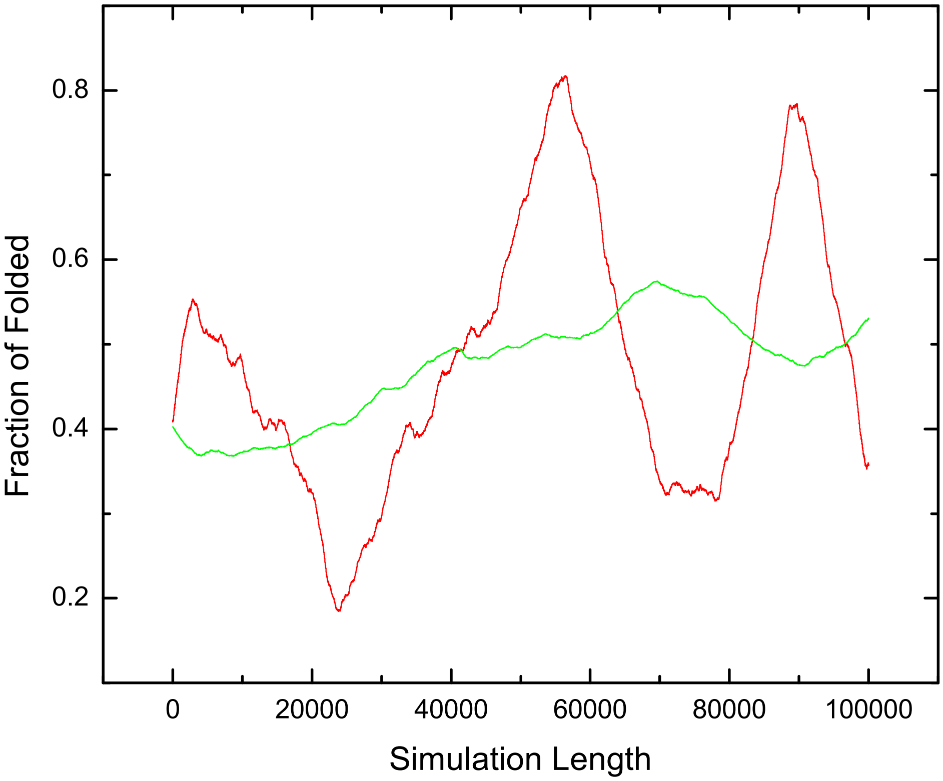

Figure 3.41: The fraction of folded of trpzip2 300K replica, calculated using data sets of different length (red: 10,000, green: 50,000)

formations might be inadequate. REM approach can be helpful in evaluating force field modifications since the convergence of thermodynamics is not a concern and energy profile can be obtained rather quickly.

The use of principal components as the coordinates for constructing folding landscape did a great job in resolving different minima. Characterization of them is not straightforward, but can be easily done by property mapping. Although principal components themselves are not progress variables and do not correspond to physical observables, they have the advantage of being truly independent and no pre-assumptions of the native state, therefore suitable for structure prediction. The convergence of major components and the implications of their magnitudes are worth further investigation.

Chapter 4

FabI Inhibition Studies with Free Energy Calculation

4.1 Introduction

NADH-dependent enoyl reductase enzyme catalyzes the reduction of long chain trans-2-enoyl-ACPs in the type II dissociated fatty acid biosynthesis pathway (FASII) in bacteria, which produces long chain fatty acid, a key component of bacteria cell wall. Interruption of this process can be deadly because the integrity of cell wall is crucial to the survival of invading bacteria against macrophages inside the target organism [113, 114, 115]. The FASII pathway has been heavily studied in several bacteria and substantial evidence now supports the notion that FASII is an attractive target for antibacterial drug development [116, 117, 118]. NADH-dependent enoyl reductase (FabI/InhA[1]) is one of the most common targets.

Tuberculosis, a disease that has been fought against worldwide[2] [119, 120, 121], is caused by Mycobacterium tuberculosis (TB) infection. Attempts to control the

[1] InhA is the enoyl reductase of Mycobacterium tuberculosis. The homologue of InhA in other bacteria is often referred to as FabI

[2] One third of the world population are infected and over two million people die every year. Tuberculosis is also a major opportunistic pathogen in patients with HIV/AIDS.

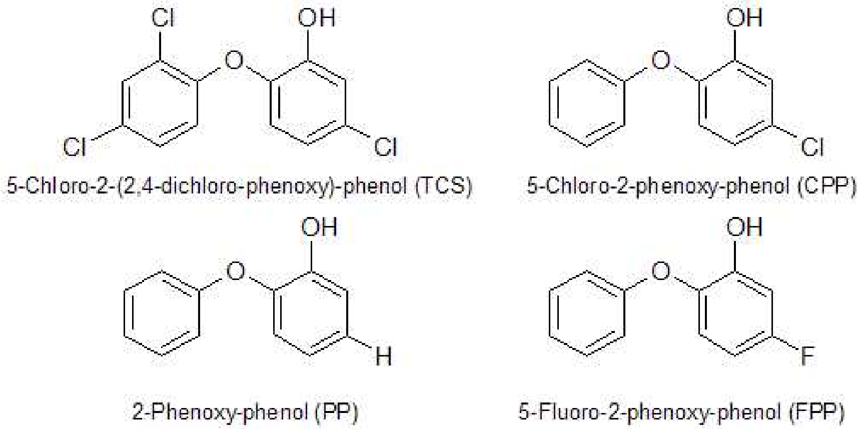
Cl
OH
O
Cl
Cl
5-Chloro-2-(2,4-dichloro-phenoxy)-phenol (TCS)
OH
O
Cl
5-Chloro-2-phenoxy-phenol (CPP)
OH
O
H
2-Phenoxy-phenol (PP)
OH
O
F
5-Fluoro-2-phenoxy-phenol (FPP)

Figure 4.1: Four triclosan analogs that were studied

spread of this disease are severely hampered by the emergence of multi-drug resistant strains of TB [122, 123]. The need of developing high affinity inhibitors is urgent. Recent studies have revealed that the "nonspecific" biocide triclosan (TCS) is an high affinity (picomolar) FabI inhibitor and can be used as a potential lead for InhA inhibitor design, which has been proposed by Tonge etc. Preliminary structural and inhibition studies showed that triclosan binds weakly to the complex of InhA and NAD$^+$ (micromolar, [124]) in a very similar manner as observed in the case of FabI. However, a loop (residue 196–205) near the binding site of FabI becomes ordered upon triclosan binding while this is not observed in the InhA:triclosan structure [124].

The long term goal of the present modeling study is to predict changes to the triclosan structure that will improve binding to InhA. In order to validate/refine the approach the initial effort has been focused on using molecular dynamics and free energy calculation methods to reproduce the changes in FabI-binding affinity observed for four of the triclosan analogs (Figure 4.1) [125, 126]. Among them, a small change in structure results in a large alteration in binding affinity (Table 4.1). Reproducing this sensitivity not only calibrates force field parameters and

Compound	pK$_a$	FabI Inhibition (K$_1$)
TCS	7.8±0.1	7±1pM
PP	9.12±0.03	0.50±0.02μM
CPP	8.13±0.02	1.1±0.1pM
FPP	8.12±0.06	3.2±0.4nM

Table 4.1: The pK$_a$ and binding affinity of four triclosan analogs

computational techniques that are to be applied, but also serves to provide insight on the origin of this sensitivity, which is more important to future inhibitor design work.

Comparing the binding affinities of different ligands computationally involves the calculation of binding free energy change ($\Delta G_{binding}$). The direct calculation of absolute binding free energy change is possible albeit extremely difficult due to the cost of sampling both the bound and unbound states, therefore is not generally practical. On the other hand, binding free energy change ($\Delta\Delta G_{binding,A\rightarrow B}$) between different ligands can be readily calculated provided that structural and free energy changes between ligands are relatively small. Free energy calculation techniques (free energy perturbation or thermodynamic integration) have become mature [127, 128] and can provide accurate estimation of $\Delta\Delta G_{binding,A\rightarrow B}$ that is comparable to experimental values, but the computational cost is still rather substantial and rarely used routinely in rational ligand design. In the early stage of this long term investigation, the good agreement between modeling and experiments is of first order and free energy calculation is the only reliable approach available to achieve it. Fast and approximate ligand scoring approaches will only become suitable after an appropriate model is established. The main goal of this study is to acquire this model and pave the way for the coming ligand design process.

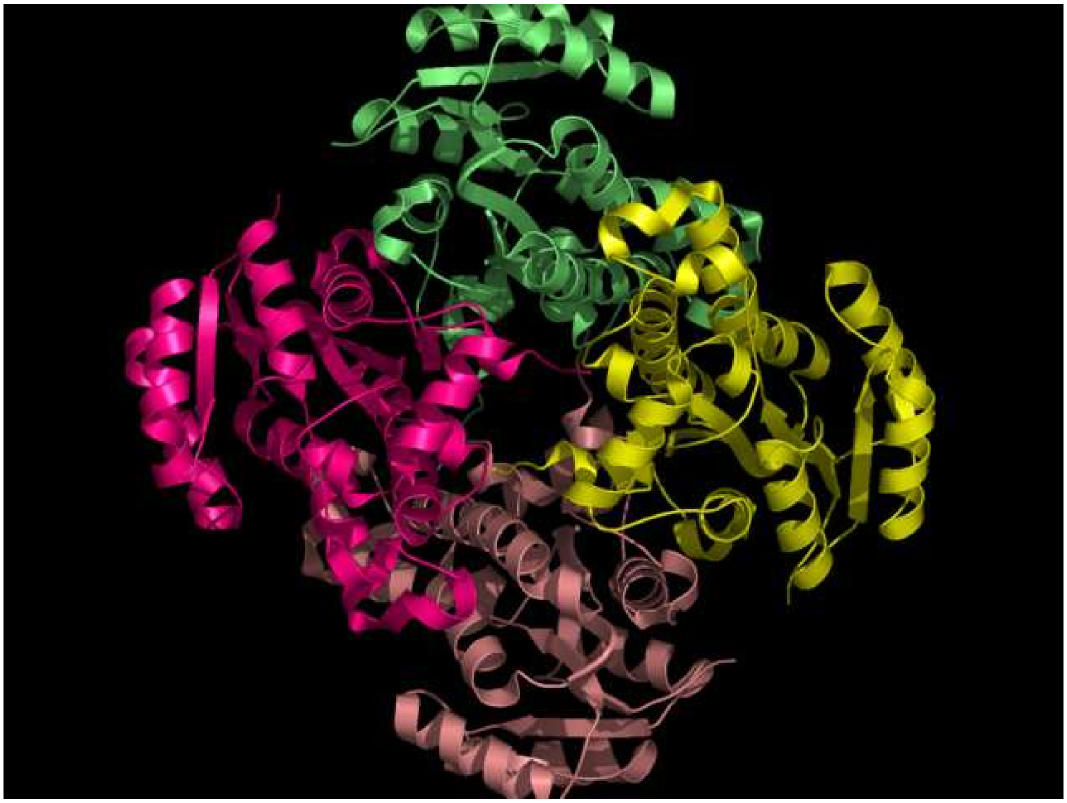

Figure 4.2: FabI:NAD$^+$:triclosan complex crystal structure colored by monomer shown in ribbon diagram

4.2 System and Setup

The crystal structure of FabI:NAD$^+$:triclosan complex has been solved by Stewart and Kisker with 1.75Å resolution [129]. The complex packs together with three identical copies and forms a tetramer (Figure 4.2) with large interface shared between monomers. Molecular dynamics simulations were first performed as the preparation step for the following free energy calculations. This step is necessary to adjust the crystal conformation to the simulated aqueous environment and the changes introduced when triclosan is replaced with other three analogs. All MD simulations and free energy calculations took into account the solvation effect explicitly by adding water molecules around the FabI:NAD$^+$:ligand complex. Force field parameters for all four triclosan ligands, specifically the atomic charges and

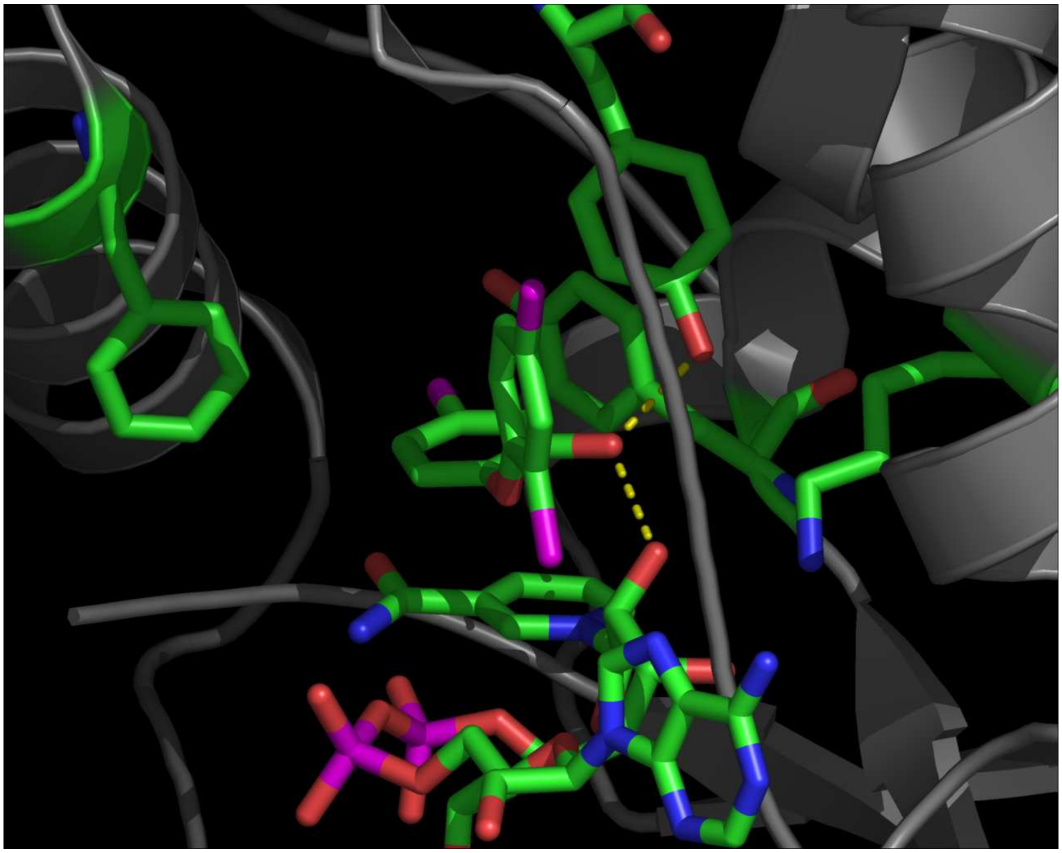

Figure 4.3: FabI:NAD$^+$:Triclosan binding site

two torsion angles involved in the ether linkage between two rings, were carefully derived following the procedures described below.

4.2.1 Triclosan Analogs

Triclosan and three analogs in Figure 4.1 all share the same skeleton of two benzene rings connected by ether bond. For later convenience, the phenol ring is denoted as "**A** ring" while the phenoxy ring is denoted as "**B** ring". In the crystal structure of the ternary complex (Figure 4.3), the hydroxy group of A ring forms hydrogen bonds with Y156 and the NAD$^+$ ribose. **A** ring, deep inside of the binding pocket, stacks on top of the nicotinamide fragment of NAD$^+$. The plane of **B** ring is roughly perpendicular to that of **A** ring and is less buried. Substitution of the two chlorine groups on **B** ring affects very little the binding affinity (comparing the binding constants of TCS and CPP in Table 4.1), however, **A** ring is extremely sensitive to even very slight changes, binding affinity losing almost one million fold when the chlorine is replaced with a hydrogen.

4.2.2 Force Field Parameters

Atomic charges for all four triclosan compounds as well as their deprotonated forms were calculated following the same procedure as in the pyrene-DNA case described in Chapter 2. The electrostatic potentials were first calculated with quantum mechanics and then reproduced by two-stage charge fitting. The needed torsion terms for the ether linkage were calculated with the same philosophy. First, the potential energy surface with respect to the two torsion angles (a 10 by 10 grid) was scanned with quantum mechanics calculations at HF/6-31G* level (15 days 11 hours). The torsions terms were then fitted to reproduce the potential energy surface relative to a reference conformation chosen arbitrarily. Standard *ff94* force field [11] was used for FabI.

4.3 Molecular Dynamics Simulations of FabI and Triclosan Complex

4.3.1 Monomer Calculations

A 6.4ns explicit solvent molecular dynamics was first calculated for the monomer of the binding ternary complex at room temperature (300K). Following the protocol described in Section 2.2.3.1, the positional restraints applied on the ternary complex were gradually released altogether in five 10ps MD simulations to give the relaxed conformation for the next 6.4ns production run. The backbone root

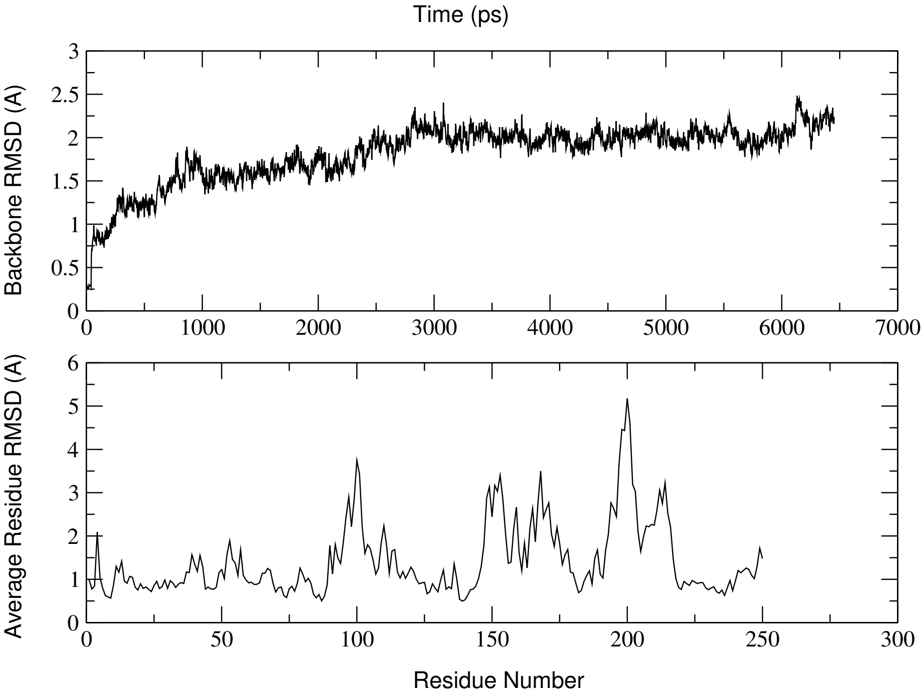

Figure 4.4: The backbone and averaged residue RMSD for the FabI:NAD$^+$:Triclosan monomer complex.

mean square deviation (RMSD) with respect to the crystal conformation and average residue RMSD were shown in Figure 4.4. The structure started to adopt a stable conformation at 2Å after roughly 3ns. Significant deviation from the crystal conformation was observed in three regions that can be identified from the average residue RMSD. Although most part of the protein backbone prefer the crystal conformation, residue 94–110, residue 146–178, and residue 197–217 exhibits rather large difference (generally over 2Å). All three regions are located on the interface between monomers and possibly contribute to the ligand binding. (Figure 4.5). The final snapshot of the simulation was compared to the crystal structure, showing an expanded binding site and an outbreaking triclosan (Figure 4.6). This large conformational deviation of the binding site is particularly worrisome be-

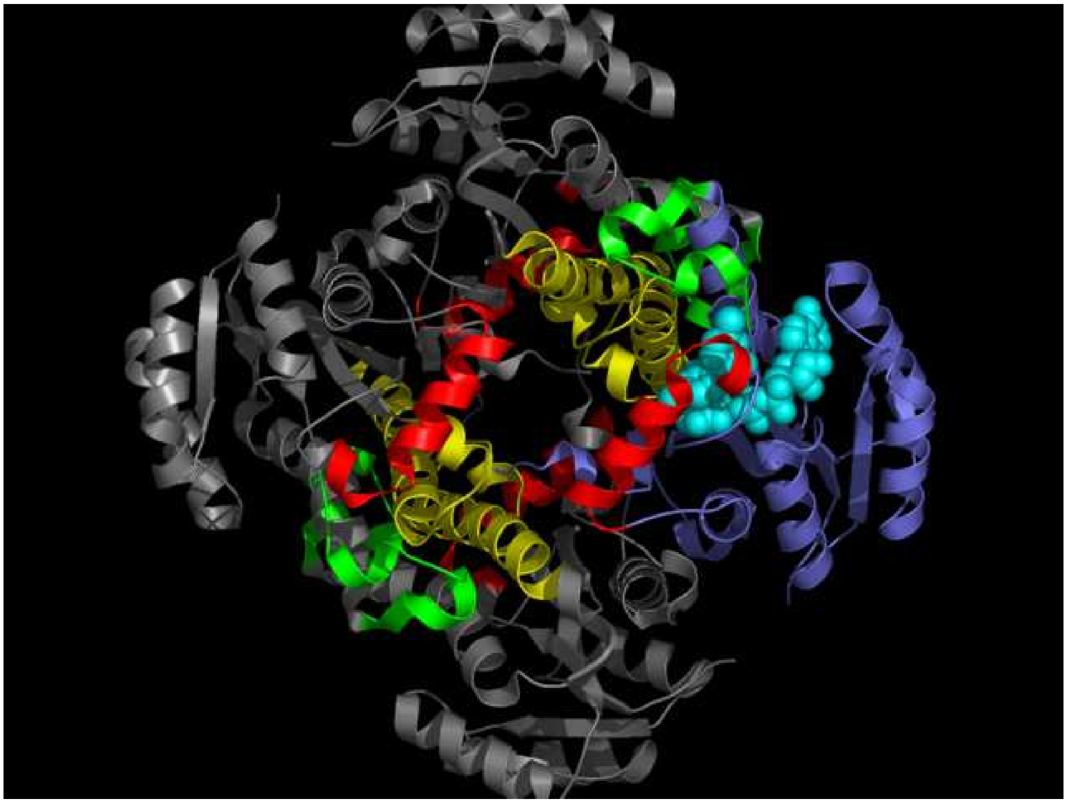

Figure 4.5: Three regions (residue 94–110, green; residue 146–178, yellow; residue 197–217, red) on the monomer interface show large deviation from the crystal conformation.

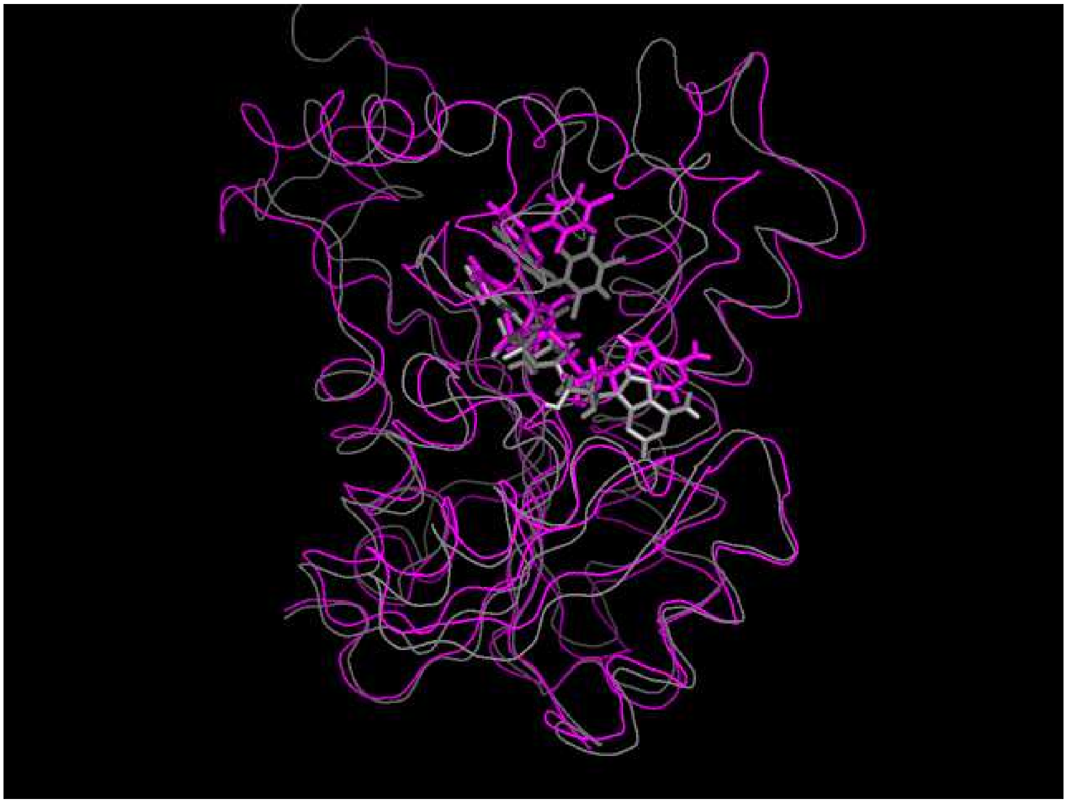

Figure 4.6: The final snapshot of the monomer (purple) simulation is compared to the crystal structure (gray). The FabI backbone is shown in backbone trace and the ligand and NAD$^+$ are shown with sticks.

cause of possible involvement of other monomers in the ligand binding through binding interface stabilization, which may complicate the free energy calculations. It is not clear from available experimental data that FabI:NAD$^+$:triclosan exists monomerically in the solution. More cautiously, in the second MD simulation of the monomer ternary complex, the positional restraints on triclosan, NAD$^+$, and FabI were released separately. However, this did not help to remove the binding site conformational deviation.

4.3.2 Tetramer Calculations

The explicit solvent simulation of the tetrameric ternary complex was only carried out for 100ps due to the limitation of computational resources. The whole system consists of almost 68,000 atoms when solvated and required 76 hours to calculate the 100ps MD trajectory using 10 250 MHz SGI-2000 CPUs. Although the simulation does show lowered RMSD (below 1Å) and residue fluctuations (mostly below 1.5Å) for backbone atoms, the length of the simulation is certainly too short to conclude that monomer interactions are necessitated for the binding. Nonetheless, the MD simulations clearly indicate that the monomer is not an appropriate system in order to best address the binding free energy difference between triclosan analogs.

4.4 Relative Binding Free Energy Calculations of Triclosan analogs

4.4.1 Free Energy Calculation Setup

Performing free energy calculation on the tetramer, the closest representation to reality, is prohibitively expensive if not impossible. Additionally, the computing time for the same amount of sampling in the current free energy calculation routine is almost twice as much as regular MD simulations. To reduce the cost and make calculations tractable for a series of compounds that are under study, an intermediate model was used that was similar in size to the monomer but included an additional 115 amino acid residues from neighboring monomers that were within 20Å from triclosan. In this downsized model system, all atoms beyond 15Å from triclosan were weakly restrained with a force constant, 0.2 or 0.5 kcal/(mol·$\mathring{A}^2$)[3] to their MD-equilibrated coordinates. This model system was then solvated in a truncated-octahedral box with an 8Å padding on each side, resulting in a system of ∼27,000 atoms. All ligand perturbations in solvent were done in a cubic box with 20Å padding on each side. Counterions were not used in any calculations.

The relative binding free energy calculation of two ligands is typically done in two perturbations using thermodynamic integration (TI) approach [130], one in protein ($\Delta G_{A \rightarrow B,protein}$) and the other in solvent ($\Delta G_{A \rightarrow B,solvent}$). The difference between the two corresponds to the binding free energy difference of the two ligands. In the current study, each perturbation was done with 12 consecutive windows, the λ and weight of each window can be found in AMBER7 manual. Each perturbation started with $\lambda = 0$ toward $\lambda = 1$. For each window, a short initial integration (10ps) generated the starting conformation under NTP condition. The long production integration was then continued for 100ps under NVT condition, the average of the free energy derivative being taken every picosecond. Therefore, 100ps production integration is equivalent to 100 measurements and the statistical error was then calculated accordingly.

[3] 0.5 kcal/(mol·$\mathring{A}^2$) was used in cases that free energy calculations failed due to the outbreak of unrestrained solvent molecules.

4.4.2 Systematic Errors

The systematic error introduced by using the truncated model system was first estimated by examining $\Delta G_{TCS \rightarrow CPP,solvent}$ and $(dG/d\lambda)_{TCS \rightarrow CPP,solvent}$, the results of which was compared against those from the full system calculated with periodic boundary condition (PBC). For this test, solvent molecules beyond 15Å of TCS were removed and the outer 5Å layer of this truncated solvated system was gently restrained. The free energy derivatives with respect to each λ were then plotted in Figure 4.7. It is clear that the approximate setup increased the uncertainty of the measurement with a standard deviation of 11 to 12 kcal/mol comparing to 1 to 2 kcal/mol in the more rigorous setup. The weighted sum of free energy derivatives of all windows, corresponding to the free energy cost of mutating TCS to CPP in solvent, only differed by 0.4kcal/mol. The standard deviation of this sum for the truncated setup is about 1.1 to 1.2 kcal/mol, relatively small.

The same test was conducted for $\Delta G_{protein}$ and $(dG/d\lambda)_{protein}$ as well. In the reference calculation, all four ligands in the tetramer were perturbed at the same time and the average of the overall perturbation free energy was compared with that of the truncated system. The integration was performed for 50ps at each λ for the tetramer calculation. Similarly, the uncertainty of the 15Å-truncated setup was substantially larger than the reference. To reduce the uncertainty, a larger truncation (20Å) with periodic boundary condition was employed in all subsequent free energy calculations. Although longer production integration (>100ps) is certainly preferable for better confidence of the results, compromise had to be made between reliability and computational cost. In fact, this compromise is perhaps not "terrible" as can been seen from the next section.

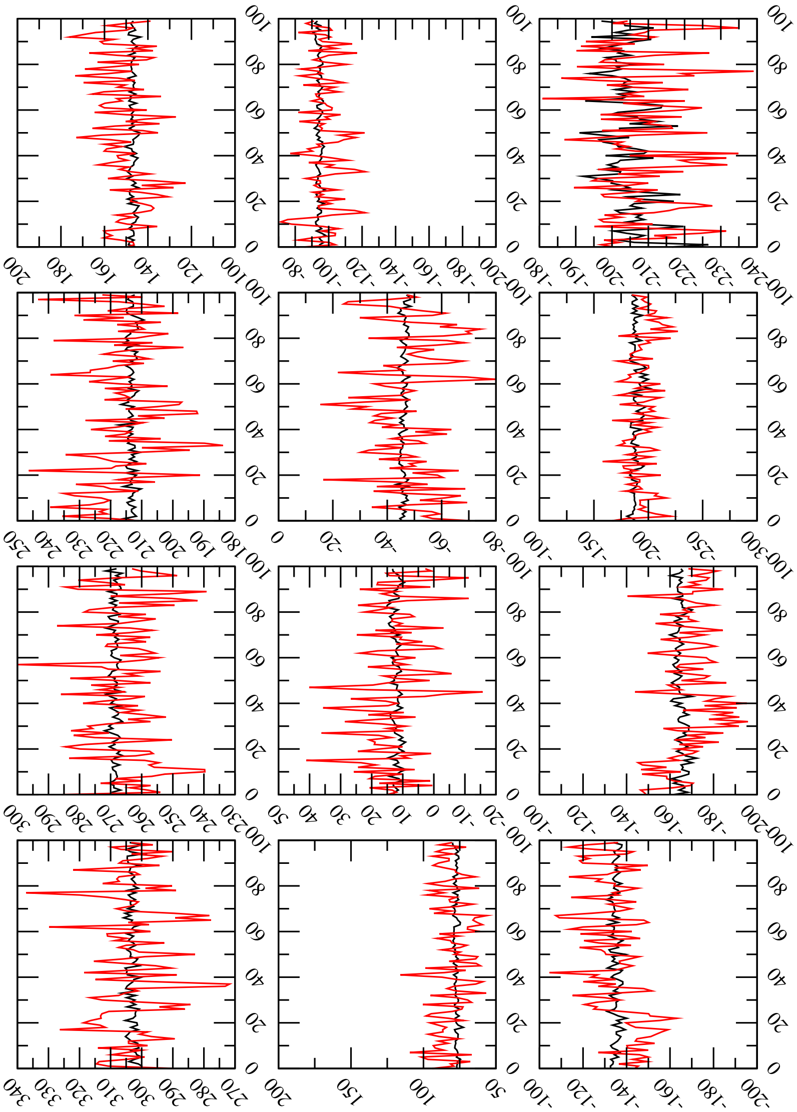

Figure 4.7: Free energy derivatives of each λ $(dG/d\lambda)_{TCS\rightarrow CPP,solvent}$. Periodic boundary condition is colored black, the truncated system is colored red.

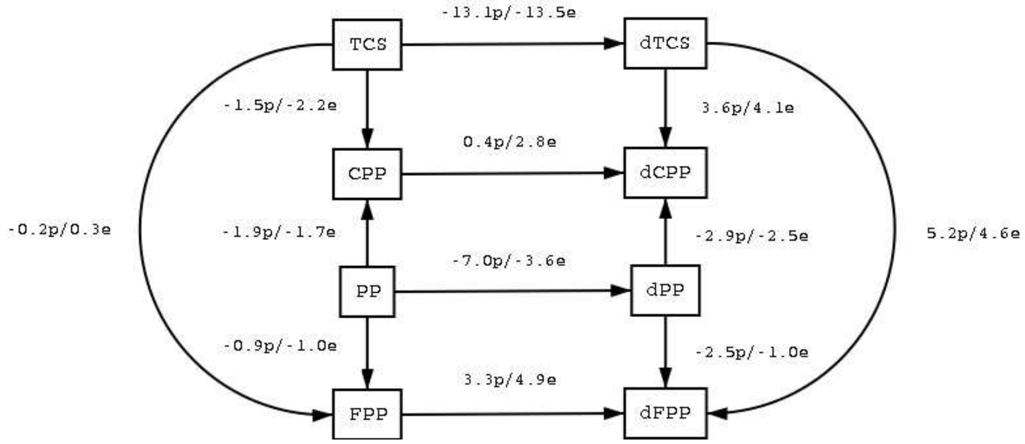

Figure 4.8: Conducted relative binding free energy calculations: the direction of the arrow indicates the target of each perturbation. The values accompanying each arrow correspond to the associated binding free energy change[5] (*p*: production; *e*: equilibration).

4.4.3 Results of TI Calculations

All TI calculations that were performed are shown in the diagram as well as the binding free energy change for each perturbation (Figure 4.8). Overall 12 perturbations were calculated. The left part of the diagram are the perturbations between neutral ligands with the **A** ring phenol group protonated. Although each perturbation was only calculated from one direction, the sum over the closed circle (TCS → CPP → PP → FPP → TCS) is -0.3 kcal/mol, within the calculation uncertainty to the true value, which is zero. This speaks well for the validity of the calculation and technical treatments. However, the calculated relative binding free energies are rather far away from what experiments had measured (Table 4.2). The calculation suggests that the binding affinity is largely insensitive to the introduced changes *if* ligands remain protonated upon binding.

A closer examination of the crystal structure disclosed a buried lysine side chain (K163) close to the bound ligand (Figure 4.3), which prompts a possibility of deprotonated bound state for the ligand phenol group. The electrostatic potential

$\Delta\Delta G_{binding}$ (kcal/mol)	TCS	PP	CPP	FPP
TCS	—	6.6	-1.1	3.6
PP		—	-7.7	-3.0
CPP			—	4.7
FPP				—

Table 4.2: Binding free energy differences measured in experiments, calculated by subtracting row from column.

of the ternary complex while K163 is charged was calculated with *Delphi* and visualized on the molecular surface (Figure 4.9), which reveals a large electro-positive surface next to the bound triclosan created by K163. Several attempts were made to calculate the effective pK$_a$ of the ligand, K163, and Y156 within the context of the protein by solving Poisson-Boltzmann equation. However, they were not very successful due to the extreme sensitivity of the calculated pK$_a$ to the local hydrogen bond network. pK$_a$ calculation was then forfeited and perturbations that involved the deprotonation state of the ligands were pursued, as shown in the right part of Figure 4.8.

Before the discussion of the perturbation results, it should be noted that the MD simulation of the monomer complex with the deprotonated triclosan (dTCS) was again unstable after a few nanoseconds.

Once the deprotonated state were included in the perturbation, large free energy difference started to emerge. The perturbations represented by the horizontal arrows correspond to the difference between deprotonating ligand in protein and solvent, which is equivalent to relative pK$_a$ calculation (ΔpK$_{a,protein-solvent}$). The pure electrostatic nature of this change makes the calculations difficult to converge over the period of 100ps, therefore not very trustworthy. The perturbations between deprotonated ligands are somewhat promising, although the error in the sum of the closed circle appeared to be bigger (-1.2 kcal/mol). The calculated binding affinity difference agreed rather well to the experimental ones except three

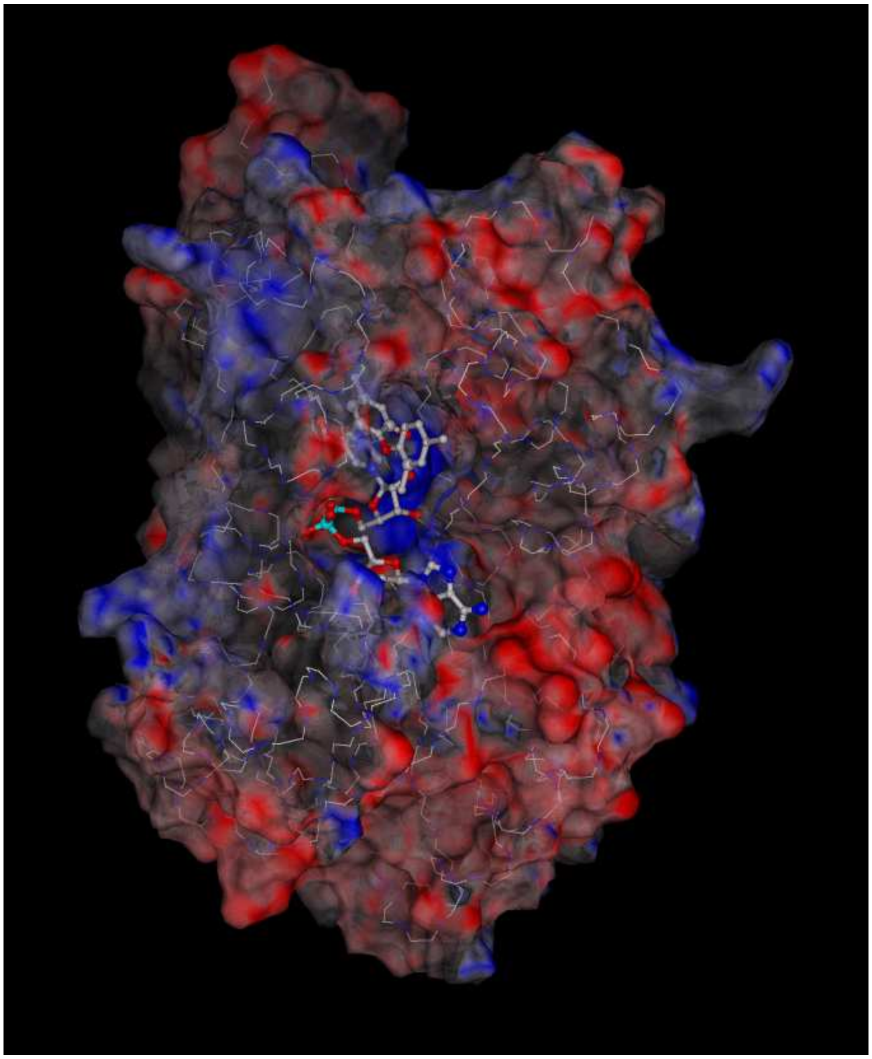

Figure 4.9: The molecular surface colored by the calculated electrostatic potential (negative, red; positive, blue)

kcal/mol	BOND	ANGLE	1-4 NB	1-4 EEL	VDW	EELEC	total
TCS→CPP	-2.0	0	-0.2	0.0	0.5	0.1	-1.5
PP→CPP	0.3	0	0.0	-0.4	-1.9	0.0	-1.9
PP→FPP	0.0	0.0	0.0	0.1	-0.7	-0.3	-0.9
dTCS→dCPP	-0.4	0.0	-0.1	0.1	2.6	1.4	3.6
dPP→dCPP	-0.2	0.0	0.0	0.0	-2.3	-0.3	-2.9
dPP→dFPP	-1.4	0.0	0.0	0.0	-0.8	-0.3	-2.5

Table 4.3: Individual contributions from bond energy (BOND), bond angle energy (ANGLE), 1-4 non-bond interactions (1-4 NB), 1-4 electrostatic interactions (1-4 EEL), van der Waals (VDW), electrostatic interactions (EELEC).

perturbations that dCPP were involved.

This partial discrepancy or agreements raised the question on the reliability of both theoretical calculations, which were based on the assumption that all ligands bind similarly as triclosan, and the experiments, particularly the binding affinity of CPP. The binding affinities in Table 4.1 suggest that **B** ring is of little significance to the binding, removing two chlorine groups (TCS vs. CPP) slightly enhances the binding, but **A** ring is extremely sensitive to modifications. However, this was opposite to the calculation results, two **B** ring chlorine groups contributing 3.6 kcal/mol to the binding, 75% of which is from van der Waals contact lost (Table 4.3). Modifying **A** ring definitely affects the binding as shown in the perturbations from dPP to dFPP or dCPP, but not as much as what experiments had measured in PP→CPP case. The solution to this discrepancy and the final converging explanation of FabI inhibition will be found, but perhaps require longer free energy calculation and re-examination of experiment measurements.

4.5 Conclusions

As a preliminary investigation of InhA inhibition studies and inhibitor design, the relative binding affinities of four ligands against FabI were studied by molecular dynamics and free energy calculation approaches using a simplified model system, in which important insights of the binding nature were obtained. MD simulations of FabI monomer and tetramer provide indirect evidence that the tetramer interactions and ligand binding may be related to the binding. In FabI:NAD$^+$:TCS complex, residue 197–217 becomes ordered upon binding, however, the corresponding region in InhA tertiary complex is disordered[6], giving poor electron density. If this is proved, it may be the key to the success of inhibitor design, nevertheless, it does greatly increases the difficulty of the design.

The simplified model system appears to be a reasonable approximation and can be used in the binding study without too much compromise of the reliability. Although the free energy calculations conducted using this model system did not completely reproduce the relative binding affinities of all four ligands, they did provide strong indication that the protonation state of the ligand play a very important role in the ligand/protein interactions. Deprotonated ligands can be formed with little cost under the condition (pH = 8) that binding assays are conducted. Surprisingly, van der Waals makes the largest contribution to the binding free energy difference in most of the cases, not electrostatics. Instead, electrostatic interactions might provide important steering guidance for the access of ligands to the binding site.

[6] TCS is only a micromolar inhibitor for InhA.

Final Remarks

Heterogeneity in biomolecular systems remains a real challenge to both computational and experimental biophysicists.To address this heterogeneity needs the joint efforts from both sides. Computational approaches based on physical models potentially should be able to predict quantitatively both thermodynamics and kinetics, however, this power is often weakened by the quality of available models and statistics, which is limited by computational resources. Therefore, a significant amount of time in the current computational biophysics research is being spent on sophisticating physical models and improving sampling efficiency.

In the first study, the conformational heterogeneity of the pyrene-modified DNA duplex was resolved because of the better statistics obtained from the locally enhanced sampling technique. However, the kinetic feature of the transition between *anti* and *syn* conformers is still inaccessible to direct simulations with atomic detail.

In the second study, the kinetic heterogeneity of a model peptide folding process was addressed to the best that can be done with the current state-of-the-art simulation techniques, i.e., ensemble folding simulation, continuum solvent model, and the replica exchange method. As a fairly realistic example for protein folding, it described a generic scenario that could occur to any polyaminoacid of arbitrary sequence.

The successful explanation of the folding mechanism above can be attributed to the small size of the peptide and the accessible timescales of the transition. When the timescale of a process far exceeds the limit of computer simulations, kinetics modeling becomes infeasible, which is the case of the first study. Models with low resolution, such as bead model, may prove to be more suitable and successful than the all-atomic models.

Finally, the relative FabI binding affinities of triclosan analogs were calculated by the rigorous free energy technique, but with approximations. Once again, quantitative match with experimental results is still the ultimate challenge. Not only good statistics but also accurate physical and chemical models are critical, which will continue to trouble computational biophysicists for some time.

Bibliography

[1] I. D. Campbell, “The march of structural biology,” *Nature Reviews Molecular Cell Biology*, vol. 3, no. 5, pp. 377–381, 2002.

[2] J. A. McCammon, B. R. Gelin, M. Karplus, and P. G. Wolynes, “Hinge-bending mode in lysozyme,” *Nature*, vol. 262, no. 5566, pp. 325–326, 1976.

[3] J. A. McCammon and M. Karplus, “Internal motions of antibody molecules,” *Nature*, vol. 268, no. 5622, pp. 765–766, 1977.

[4] J. A. McCammon, B. R. Gelin, and M. Karplus, “Dynamics of folded proteins,” *Nature*, vol. 267, no. 5612, pp. 585–590, 1977.

[5] M. Karplus and J. A. McCammon, “Molecular dynamics simulations of biomolecules,” *Nature Structural Biology*, vol. 9, no. 9, pp. 646–652, 2002.

[6] M. Karplus, “Molecular dynamics simulations of biomolecules,” *Accounts of Chemical Research*, vol. 35, no. 6, pp. 321–323, 2002.

[7] M. Karplus, “Molecular dynamics of biological macromolecules: A brief history and perspective,” *Biopolymers*, vol. 68, no. 3, pp. 350–358, 2003.

[8] L. C. James, P. Roversi, and D. S. Tawfik, “Antibody multispecificity mediated by conformational diversity,” *Science*, vol. 299, no. 5611, pp. 1362–1367, 2003.

[9] V. Gogonea, L. M. Westerhoff, and K. M. Merz, “Quantum mechanical/quantum mechanical methods. i. a divide and conquer strategy for solving the schrodinger equation for large molecular systems using a composite density functional- semiempirical hamiltonian,” *Journal of Chemical Physics*, vol. 113, no. 14, pp. 5604–5613, 2000.

[10] A. D. Mackerell, J. Wiorkiewiczkuczera, and M. Karplus, “An all-atom empirical energy function for the simulation of nucleic-acids,” *Journal of the American Chemical Society*, vol. 117, no. 48, pp. 11946–11975, 1995.

[11] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, “A 2nd generation force-field for the simulation of proteins, nucleic-acids, and organic-molecules,” *Journal of the American Chemical Society*, vol. 117, no. 19, pp. 5179–5197, 1995.

[12] D. R. Langley, “Molecular dynamic simulations of environment and sequence dependent dna conformations: The development of the bms nucleic acid force field and comparison with experimental results,” *Journal of Biomolecular Structure and Dynamics*, vol. 16, no. 3, pp. 487–509, 1998.

[13] C. Simmerling, B. Strockbine, and A. E. Roitberg, “All-atom structure prediction and folding simulations of a stable protein,” *Journal of the American Chemical Society*, vol. 124, no. 38, pp. 11258–11259, 2002.

[14] A. Okur, B. Strockbine, V. Hornak, and C. Simmerling, “Using pc clusters to evaluate the transferability of molecular mechanics force fields for proteins,” *Journal of Computational Chemistry*, vol. 24, no. 1, pp. 21–31, 2003.

[15] M. Feig, A. D. MacKerell, and C. L. Brooks, “Force field influence on the observation of pi-helical protein structures in molecular dynamics simulations,” *Journal of Physical Chemistry B*, vol. 107, no. 12, pp. 2831–2836, 2003.

[16] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, “Equation of state calculations by fast computing machines,” *Journal of Chemical Physics*, vol. 21, no. 6, pp. 1087–1092, 1953.

[17] J. Schlitter, M. Engels, P. Kruger, E. Jacoby, and A. Wollmer, “Targeted molecular-dynamics simulation of conformational change - application to the t[–]r transition in insulin,” *Molecular Simulation*, vol. 10, no. 2-6, pp. 291–, 1993.

[18] S. Izrailev, S. Stepaniants, M. Balsera, Y. Oono, and K. Schulten, “Molecular dynamics study of unbinding of the avidin-biotin complex,” *Biophysical Journal*, vol. 72, no. 4, pp. 1568–1581, 1997.

[19] R. Elber and M. Karplus, “Enhanced sampling in molecular-dynamics - use of the time- dependent hartree approximation for a simulation of carbon-monoxide diffusion through myoglobin,” *Journal of the American Chemical Society*, vol. 112, no. 25, pp. 9161–9175, 1990.

[20] G. N. Patey and J. P. Valleau, “Free-energy of spheres with dipoles - monte-carlo with multistage sampling,” *Chemical Physics Letters*, vol. 21, no. 2, pp. 297–300, 1973.

[21] G. N. Patey and J. P. Valleau, “Dipolar hard spheres - monte-carlo study,” *Journal of Chemical Physics*, vol. 61, no. 2, pp. 534–540, 1974.

[22] G. N. Patey and J. P. Valleau, “Monte-carlo method for obtaining interionic potential of mean force in ionic solution,” *Journal of Chemical Physics*, vol. 63, no. 6, pp. 2334–2339, 1975.

[23] A. Mitsutake, Y. Sugita, and Y. Okamoto, "Generalized-ensemble algorithms for molecular simulations of biopolymers," *Biopolymers*, vol. 60, no. 2, pp. 96–123, 2001.

[24] S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg, "The weighted histogram analysis method for free-energy calculations on biomolecules .1. the method," *Journal of Computational Chemistry*, vol. 13, no. 8, pp. 1011–1021, 1992.

[25] V. Tsui and D. A. Case, "Theory and applications of the generalized born solvation model in macromolecular simulations," *Biopolymers*, vol. 56, no. 4, pp. 275–291, 2000.

[26] A. Jeancharles, A. Nicholls, K. Sharp, B. Honig, A. Tempczyk, T. F. Hendrickson, and W. C. Still, "Electrostatic contributions to solvation energies - comparison of free-energy perturbation and continuum calculations," *Journal of the American Chemical Society*, vol. 113, no. 4, pp. 1454–1455, 1991.

[27] T. J. Matray and E. T. Kool, "A specific partner for abasic damage in dna," *Nature*, vol. 399, no. 6737, pp. 704–708, 1999.

[28] S. Smirnov, T. J. Matray, E. T. Kool, and C. D. L. Santos *To be published*.

[29] T. E. Cheatham and P. A. Kollman, "Observation of the a-dna to b-dna transition during unrestrained molecular dynamics in aqueous solution," *Journal of Molecular Biology*, vol. 259, no. 3, pp. 434–444, 1996.

[30] P. Ewald *Annual Physics*, vol. 64, p. 253, 1921.

[31] T. Darden, D. York, and L. Pedersen, "Particle mesh ewald - an n.log(n) method for ewald sums in large systems," *Journal of Chemical Physics*, vol. 98, no. 12, pp. 10089–10092, 1993.

[32] T. E. Cheatham and P. A. Kollman, "Insight into the stabilization of a-dna by specific ion association: spontaneous b-dna to a-dna transitions observed in molecular dynamics simulations of d[acccgcgggt](2) in the presence of hexaamminecobalt(iii)," *Structure*, vol. 5, no. 10, pp. 1297–1311, 1997.

[33] T. E. Cheatham, M. F. Crowley, T. Fox, and P. A. Kollman, "A molecular level picture of the stabilization of a-dna in mixed ethanol-water solutions," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, no. 18, pp. 9626–9630, 1997.

[34] B. Jayaram, D. Sprous, M. A. Young, and D. L. Beveridge, "Free energy analysis of the conformational preferences of a and b forms of dna in solution," *Journal of the American Chemical Society*, vol. 120, no. 41, pp. 10629–10633, 1998.

[35] D. Sprous, M. A. Young, and D. L. Beveridge, "Molecular dynamics studies of the conformational preferences of a dna double helix in water and an ethanol/water mixture: Theoretical considerations of the a double left right arrow b transition," *Journal of Physical Chemistry B*, vol. 102, no. 23, pp. 4658–4667, 1998.

[36] M. A. Young and D. L. Beveridge, "Molecular dynamics simulations of an oligonucleotide duplex with adenine tracts phased by a full helix turn," *Journal of Molecular Biology*, vol. 281, no. 4, pp. 675–687, 1998.

[37] D. Sprous, M. A. Young, and D. L. Beveridge, "Molecular dynamics studies of axis bending in d(g(5)- (ga(4)t(4)c)(2)-c-5) and d(g(5)-(gt(4)a(4)c)(2)-c-5): Effects of sequence polarity on dna curvature," *Journal of Molecular Biology*, vol. 285, no. 4, pp. 1623–1632, 1999.

[38] L. Ayadi, M. Jourdan, C. Coulombeau, J. Garcia, and R. Lavery, "Experimental and theoretical studies of the conformational perturbations induced by an abasic site," *Journal of Biomolecular Structure and Dynamics*, vol. 17, no. 2, pp. 245–257, 1999.

[39] L. Ayadi, C. Coulombeau, and R. Lavery, "The impact of abasic sites on dna flexibility," *Journal of Biomolecular Structure and Dynamics*, vol. 17, no. 4, pp. 645–653, 2000.

[40] D. Barsky, N. Foloppe, S. Ahmadia, D. M. Wilson, and A. D. MacKerell, "New insights into the structure of abasic dna from molecular dynamics simulations," *Nucleic Acids Research*, vol. 28, no. 13, pp. 2613–2626, 2000.

[41] P. Cieplak, T. E. Cheatham, and P. A. Kollman, "Molecular dynamics simulations find that 3′ phosphoramidate modified dna duplexes undergo a b to a transition and normal dna duplexes an a to b transition," *Journal of the American Chemical Society*, vol. 119, no. 29, pp. 6722–6730, 1997.

[42] E. Cubero, E. C. Sherer, F. J. Luque, M. Orozco, and C. A. Laughton, "Observation of spontaneous base pair breathing events in the molecular dynamics simulation of a difluorotoluene-containing dna oligonucleotide," *Journal of the American Chemical Society*, vol. 121, no. 37, pp. 8653–8654, 1999.

[43] V. Tsui and D. A. Case, "Molecular dynamics simulations of nucleic acids with a generalized born solvation model," *Journal of the American Chemical Society*, vol. 122, no. 11, pp. 2489–2498, 2000.

[44] W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson, "Semianalytical treatment of solvation for molecular mechanics and dynamics," *Journal of the American Chemical Society*, vol. 112, no. 16, pp. 6127–6129, 1990.

[45] T. Simonson, “Macromolecular electrostatics: continuum models and their growing pains,” *Current Opinion in Structural Biology*, vol. 11, no. 2, pp. 243–252, 2001.

[46] B. Jayaram, D. Sprous, and D. L. Beveridge, “Solvation free energy of biomacromolecules: Parameters for a modified generalized born model consistent with the amber force field,” *Journal of Physical Chemistry B*, vol. 102, no. 47, pp. 9571–9576, 1998.

[47] D. J. Williams and K. B. Hall, “Experimental and theoretical studies of the effects of deoxyribose substitutions on the stability of the uucg tetraloop,” *Journal of Molecular Biology*, vol. 297, no. 1, pp. 251–265, 2000.

[48] J. Srinivasan, M. W. Trevathan, P. Beroza, and D. A. Case, “Application of a pairwise generalized born model to proteins and nucleic acids: inclusion of salt effects,” *Theoretical Chemistry Accounts*, vol. 101, no. 6, pp. 426–434, 1999.

[49] V. Tsui and D. A. Case, “Molecular dynamics simulations of nucleic acids with a generalized born solvation model,” *Journal of the American Chemical Society*, vol. 122, no. 11, pp. 2489–2498, 2000.

[50] C. Simmerling, J. L. Miller, and P. A. Kollman, “Combined locally enhanced sampling and particle mesh ewald as a strategy to locate the experimental structure of a nonhelical nucleic acid,” *Journal of the American Chemical Society*, vol. 120, no. 29, pp. 7149–7155, 1998.

[51] C. Simmerling, T. Fox, and P. A. Kollman, “Use of locally enhanced sampling in free energy calculations: Testing and application to the alpha ->beta anomerization of glucose,” *Journal of the American Chemical Society*, vol. 120, no. 23, pp. 5771–5782, 1998.

[52] C. Simmerling, M. R. Lee, A. R. Ortiz, A. Kolinski, J. Skolnick, and P. A. Kollman, “Combining monsster and les/pme to predict protein structure from amino acid sequence: Application to the small protein cmti-1,” *Journal of the American Chemical Society*, vol. 122, no. 35, pp. 8392–8402, 2000.

[53] D. Case, D. Pearlman, J. Caldwell, T. C. III, W. Ross, C. Simmerling, T. Darden, K. Merz, R. Stanton, A. Cheng, J. Vincent, M. Crowley, V. Tsui, R. Radmer, Y. Duan, J. Pitera, I. Massova, G. Seibel, U. Singh, P. Weiner, and P. Kollman, “Amber 6,” 1999.

[54] C. I. Bayly, P. Cieplak, W. D. Cornell, and P. A. Kollman, “A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges - the resp model,” *Journal of Physical Chemistry*, vol. 97, no. 40, pp. 10269–10280, 1993.

[55] P. Cieplak, W. D. Cornell, C. Bayly, and P. A. Kollman, “Application of the multimolecule and multiconformational resp methodology to biopolymers - charge derivation for dna, rna, and proteins,” *Journal of Computational Chemistry*, vol. 16, no. 11, pp. 1357–1377, 1995.

[56] M.J.Frisch, G. Trucks, H. Schlegel, G. Scuseria, M. Robb, J. Cheeseman, V. Zakrzewski, J. M. Jr., R. Stratmann, J. C. Burant, S. Dapprich, J. M. Millam, A. Daniels, K. Kudin, M. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G. Petersson, P. Ayala, Q. Cui, K. M. D. K. Malick, A. Rabuck, K. Raghavachari, J. Foresman, J. Cioslowski, J. Ortiz, A. Baboul, B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R. Martin, D. Fox, T. Keith, M. Al-Laham, C. Peng, A. Nanayakkara, C. Gonzalez, M. Challacombe, P. Gill, B. Johnson, W. Chen, M. Wong, J. Andres, C. Gonza-

lez, M. Head-Gordon, E. Replogle, and J. Pople, “Gaussian 98, revision a.7,” 1998.

[57] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, “Comparison of simple potential functions for simulating liquid water,” *Journal of Chemical Physics*, vol. 79, no. 2, pp. 926–935, 1983.

[58] S. C. Harvey, R. K. Z. Tan, and T. E. Cheatham, “The flying ice cube: Velocity rescaling in molecular dynamics leads to violation of energy equipartition,” *Journal of Computational Chemistry*, vol. 19, no. 7, pp. 726–740, 1998.

[59] R. Constanciel and R. Contreras, “Self-consistent field-theory of solvent effects representation by continuum models - introduction of desolvation contribution,” *Theoretica Chimica Acta*, vol. 65, no. 1, pp. 1–11, 1984.

[60] M. Schaefer and M. Karplus, “A comprehensive analytical treatment of continuum electrostatics,” *Journal of Physical Chemistry*, vol. 100, no. 5, pp. 1578–1599, 1996.

[61] A. Bondi, “Van der waals volumes + radii,” *Journal of Physical Chemistry*, vol. 68, no. 3, pp. 441–, 1964.

[62] J. W. Ponder and F. M. Richards, “An efficient newton-like method for molecular mechanics energy minimization of large molecules,” *Journal of Computational Chemistry*, vol. 8, no. 7, pp. 1016–1024, 1987.

[63] J. Srinivasan, T. E. Cheatham, P. Cieplak, P. A. Kollman, and D. A. Case, “Continuum solvent studies of the stability of dna, rna, and phosphoramidate - dna helices,” *Journal of the American Chemical Society*, vol. 120, no. 37, pp. 9401–9409, 1998.

[64] P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. H. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D. A. Case, and T. E. Cheatham, "Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models," *Accounts of Chemical Research*, vol. 33, no. 12, pp. 889–897, 2000.

[65] A. Nicholls and B. Honig, "A rapid finite-difference algorithm, utilizing successive over- relaxation to solve the poisson-boltzmann equation," *Journal of Computational Chemistry*, vol. 12, no. 4, pp. 435–445, 1991.

[66] D. J. Williams and K. B. Hall, "Unrestrained stochastic dynamics simulations of the uucg tetraloop using an implicit solvation model," *Biophysical Journal*, vol. 76, no. 6, pp. 3192–3205, 1999.

[67] A. Roitberg and R. Elber, "Modeling side-chains in peptides and proteins - application of the locally enhanced sampling and the simulated annealing methods to find minimum energy conformations," *Journal of Chemical Physics*, vol. 95, no. 12, pp. 9277–9287, 1991.

[68] C. Simmerling and R. Elber, "Hydrophobic collapse in a cyclic hexapeptide - computer- simulations of chdlfc and caaaac in water," *Journal of the American Chemical Society*, vol. 116, no. 6, pp. 2534–2547, 1994.

[69] C. Simmerling, J. L. Miller, and P. A. Kollman, "Combined locally enhanced sampling and particle mesh ewald as a strategy to locate the experimental structure of a nonhelical nucleic acid," *Journal of the American Chemical Society*, vol. 120, no. 29, pp. 7149–7155, 1998.

[70] J. L. Miller and P. A. Kollman, "Theoretical studies of an exceptionally stable rna tetraloop: Observation of convergence from an incorrect nmr structure to

the correct one using unrestrained molecular dynamics," *Journal of Molecular Biology*, vol. 270, no. 3, pp. 436–450, 1997.

[71] J. L. Schwartz, J. S. Rice, B. A. Luxon, J. M. Sayer, G. Xie, H. J. C. Yeh, X. Liu, D. M. Jerina, and D. G. Gorenstein, "Solution structure of the minor conformer of a dna duplex containing a dg mismatch opposite a benzo[a]pyrene diol epoxide/da adduct: Glycosidic rotation from syn to anti at the modified deoxyadenosine," *Biochemistry*, vol. 36, no. 37, pp. 11069–11076, 1997.

[72] A. R. Fersht, *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*. W H Freeman Co., 1999.

[73] C. B. Anfinsen, R. R. Redfield, W. L. Choate, J. Page, and W. R. Carroll, "Studies on the gross structure, cross-linkages, and terminal sequences in ribonuclease," *Journal of Biological Chemistry*, vol. 207, no. 1, pp. 201–210, 1954.

[74] A. G. Ladurner, L. S. Itzhaki, V. Daggett, and A. R. Fersht, "Synergy between simulation and experiment in describing the energy landscape of protein folding," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 15, pp. 8473–8478, 1998.

[75] U. Mayor, C. M. Johnson, V. Daggett, and A. R. Fersht, "Protein folding and unfolding in microseconds to nanoseconds by experiment and simulation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 25, pp. 13518–13522, 2000.

[76] V. Daggett and A. Fersht, "The present view of the mechanism of protein folding," *Nature Reviews Molecular Cell Biology*, vol. 4, no. 6, pp. 497–502, 2003.

[77] T. Lazaridis and M. Karplus, "Thermodynamics of protein folding: a microscopic view," *Biophysical Chemistry*, vol. 100, no. 1-3, pp. 367–395, 2003.

[78] C. Levinthal, “Are there pathways for protein folding,” *Journal De Chimie Physique Et De Physico-Chimie Biologique*, vol. 65, no. 1, p. 44, 1968.

[79] C. M. Dobson, A. Sali, and M. Karplus, “Protein folding: A perspective from theory and experiment,” *Angewandte Chemie-International Edition*, vol. 37, no. 7, pp. 868–893, 1998.

[80] M. Karplus, “Aspects of protein reaction dynamics: Deviations from simple behavior,” *Journal of Physical Chemistry B*, vol. 104, no. 1, pp. 11–27, 2000.

[81] N. Ferguson and A. R. Fersht, “Early events in protein folding,” *Current Opinion in Structural Biology*, vol. 13, no. 1, pp. 75–81, 2003.

[82] S. Xie, “Single-molecule approach to enzymology,” *Single Molecule*, vol. 2, no. 4, pp. 229–236, 2001.

[83] E. I. Shakhnovich, “Modeling protein folding: The beauty and power of simplicity,” *Folding and Design*, vol. 1, no. 3, pp. R50–R54, 1996.

[84] J. N. Onuchic, Z. LutheySchulten, and P. G. Wolynes, “Theory of protein folding: The energy landscape perspective,” *Annual Review of Physical Chemistry*, vol. 48, pp. 545–600, 1997.

[85] K. A. Dill and H. S. Chan, “From levinthal to pathways to funnels,” *Nature Structural Biology*, vol. 4, no. 1, pp. 10–19, 1997.

[86] J. E. Shea and C. L. Brooks, “From folding theories to folding proteins: A review and assessment of simulation studies of protein folding and unfolding,” *Annual Review of Physical Chemistry*, vol. 52, pp. 499–535, 2001.

[87] A. R. Fersht and V. Daggett, “Protein folding and unfolding at atomic resolution,” *Cell*, vol. 108, no. 4, pp. 573–582, 2002.

[88] S. L. Kazmirski, K. B. Wong, S. M. V. Freund, Y. J. Tan, A. R. Fersht, and V. Daggett, "Protein folding from a highly disordered denatured state: The folding pathway of chymotrypsin inhibitor 2 at atomic resolution," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 8, pp. 4349–4354, 2001.

[89] J. W. Neidigh, R. M. Fesinmeyer, and N. H. Andersen, "Designing a 20-residue protein," *Nature Structural Biology*, vol. 9, no. 6, pp. 425–430, 2002.

[90] L. L. Qiu, S. A. Pabit, A. E. Roitberg, and S. J. Hagen, "Smaller and faster: The 20-residue trp-cage protein folds in 4 mu s," *Journal of the American Chemical Society*, vol. 124, no. 44, pp. 12952–12953, 2002.

[91] G. L. Cui and C. Simmerling, "Conformational heterogeneity observed in simulations of a pyrene-substituted dna," *Journal of the American Chemical Society*, vol. 124, no. 41, pp. 12154–12164, 2002.

[92] Y. Sugita and Y. Okamoto, "Replica-exchange molecular dynamics method for protein folding," *Chemical Physics Letters*, vol. 314, no. 1-2, pp. 141–151, 1999.

[93] A. E. Garcia and K. Y. Sanbonmatsu, "Exploring the energy landscape of a beta hairpin in explicit solvent," *Proteins-Structure Function and Genetics*, vol. 42, no. 3, pp. 345–354, 2001.

[94] R. H. Zhou, B. J. Berne, and R. Germain, "The free energy landscape for beta hairpin folding in explicit water," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 26, pp. 14931–14936, 2001.

[95] R. H. Zhou and B. J. Berne, "Can a continuum solvent model reproduce the free energy landscape of a beta-hairpin folding in water?," *Proceedings of the*

National Academy of Sciences of the United States of America, vol. 99, no. 20, pp. 12777–12782, 2002.

[96] K. Y. Sanbonmatsu and A. E. Garcia, “Structure of met-enkephalin in explicit aqueous solution using replica exchange molecular dynamics,” *Proteins-Structure Function and Genetics*, vol. 46, no. 2, pp. 225–234, 2002.

[97] S. Gnanakaran and A. E. Garcia, “Folding of a highly conserved diverging turn motif from the sh3 domain,” *Biophysical Journal*, vol. 84, no. 3, pp. 1548–1562, 2003.

[98] M. K. Fenwick and F. A. Escobedo, “Hybrid monte carlo with multidimensional replica exchanges: Conformational equilibria of the hypervariable reigons of a llamma v-hh antibody domain,” *Biopolymers*, vol. 68, no. 2, pp. 160–177, 2003.

[99] D. Bratko and H. W. Blanch, “Effect of secondary structure on protein aggregation: A replica exchange simulation study,” *Journal of Chemical Physics*, vol. 118, no. 11, pp. 5185–5194, 2003.

[100] J. W. Pitera and W. Swope, “Understanding folding and design: Replica-exchange simulations of "trp-cage" fly miniproteins,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 13, pp. 7587–7592, 2003.

[101] C. Tanford, “Protein denaturation,” *Advanced Protein Chemistry*, vol. 23, pp. 121–282, 1968.

[102] M. L. Tiffany and S. Krimm, “New chain conformations of poly(glutamic acid) and polylysine,” *Biopolymers*, vol. 6, no. 9, pp. 1379–, 1968.

[103] F. Eker, X. L. Cao, L. Nafie, and R. Schweitzer-Stenner, "Tripeptides adopt stable structures in water. a combined polarized visible raman, ftir, and vcd spectroscopy study," *Journal of the American Chemical Society*, vol. 124, no. 48, pp. 14330–14341, 2002.

[104] F. Eker, K. Griebenow, and R. Schweitzer-Stenner, "Stable conformations of tripeptides in aqueous solution studied by uv circular dichroism spectroscopy," *Journal of the American Chemical Society*, vol. 125, no. 27, pp. 8178–8185, 2003.

[105] D. Fleury, S. A. Wharton, J. J. Skehel, M. Knossow, and T. Bizebard, "Antigen distortion allows influenza virus to escape neutralization," *Nature Structural Biology*, vol. 5, no. 2, pp. 119–123, 1998.

[106] J. M. Rini, U. Schulzegahmen, and I. A. Wilson, "Structural evidence for induced fit as a mechanism for antibody-antigen recognition," *Science*, vol. 255, no. 5047, pp. 959–965, 1992.

[107] D. Thirumalai, D. K. Klimov, and S. A. Woodson, "Kinetic partitioning mechanism as a unifying theme in the folding of biomolecules," *Theoretical Chemistry Accounts*, vol. 96, no. 1, pp. 14–22, 1997.

[108] R. A. Goldbeck, Y. G. Thomas, E. F. Chen, R. M. Esquerra, and D. S. Kliger, "Multiple pathways on a protein-folding energy landscape: Kinetic evidence," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 6, pp. 2782–2787, 1999.

[109] D. T. Leeson, F. Gai, H. M. Rodriguez, L. M. Gregoret, and R. B. Dyer, "Protein folding and unfolding on a complex energy landscape," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 6, pp. 2527–2532, 2000.

[110] R. A. Goldbeck, E. Chen, R. M. Esquerra, Y. G. Thomas, and D. S. Kliger, "Multiple pathways on the protein folding energy landscape: Cd and mcd studies of ultra-fast folding reactions in cytochrome c," *Biophysical Journal*, vol. 80, no. 1, pp. 187A–187A, 2001.

[111] A. G. Cochran, N. J. Skelton, and M. A. Starovasnik, "Tryptophan zippers: Stable, monomeric beta-hairpins," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 10, pp. 5578–5583, 2001.

[112] A. Okur, D. Roe, and et. al. *in preparation*.

[113] P. J. Brennan and H. Nikaido, "The envelope of mycobacteria," *Annual Review of Biochemistry*, vol. 64, pp. 29–63, 1995.

[114] R. E. Lee, P. J. Brennan, and G. S. Besra, "Mycobacterium tuberculosis cell envelope," *Tuberculosis*, vol. 215, pp. 1–27, 1996.

[115] D. G. Russell, H. C. Mwandumba, and E. E. Rhoades, "Mycobacterium and the coat of many lipids," *Journal of Cell Biology*, vol. 158, no. 3, pp. 421–426, 2002.

[116] R. J. Heath, S. W. White, and C. O. Rock, "Lipid biosynthesis as a target for antibacterial agents," *Progress in Lipid Research*, vol. 40, no. 6, pp. 467–497, 2001.

[117] J. W. Campbell and J. E. Cronan, "Bacterial fatty acid biosynthesis: Targets for antibacterial drug discovery," *Annual Review of Microbiology*, vol. 55, pp. 305–332, 2001.

[118] R. J. Heath, S. W. White, and C. O. Rock, "Inhibitors of fatty acid synthesis as antimicrobial chemotherapeutics," *Applied Microbiology and Biotechnology*, vol. 58, no. 6, pp. 695–703, 2002.

[119] A. Kochi, “The global tuberculosis situation and the new control strategy of the world-health-organization,” *Tubercle*, vol. 72, no. 1, pp. 1–6, 1991.

[120] A. Kochi, “The global tuberculosis situation and the new control strategy of the world health organization,” *Bulletin of the World Health Organization*, vol. 79, no. 1, pp. 71–75, 2001.

[121] B. R. Bloom and C. J. L. Murray, “Tuberculosis - commentary on a reemergent killer,” *Science*, vol. 257, no. 5073, pp. 1055–1064, 1992.

[122] D. C. Perlman, W. M. ElSadr, L. B. Heifets, E. T. Nelson, J. P. Matts, K. Chirgwin, N. Salomon, E. E. Telzak, O. Klein, B. N. Kreiswirth, J. M. Musser, and R. Hafner, “Susceptibility to levofloxacin of mycobacterium tuberculosis isolates from patients with hiv-related tuberculosis and characterization of a strain with levofloxacin monoresistance,” *AIDS*, vol. 11, no. 12, pp. 1473–1478, 1997.

[123] A. Rattan, A. Kalia, and N. Ahmad, “Multidrug-resistant mycobacterium tuberculosis: Molecular perspectives,” *Emerging Infectious Diseases*, vol. 4, no. 2, pp. 195–209, 1998.

[124] C. Kisker and P. J. Tonge *Unpublished data*.

[125] S. Sivaraman and P. J. Tonge, “Inhibition of the enoyl-acyl carrier protein reductase from e-coli by triclosan.,” *Biochemistry*, vol. 40, no. 29, pp. 8659–8659, 2001.

[126] S. Sivaraman, J. Zwahlen, A. F. Bell, L. Hedstrom, and P. J. Tonge, “Structure-activity studies of the inhibition of fabi, the enoyl reductase from escherichia coli, by triclosan: Kinetic analysis of mutant,” *Biochemistry*, vol. 42, no. 15, pp. 4406–4413, 2003.

[127] P. Kollman, “Free-energy calculations - applications to chemical and biochemical phenomena,” *Chemical Reviews*, vol. 93, no. 7, pp. 2395–2417, 1993.

[128] T. Simonson, G. Archontis, and M. Karplus, “Free energy simulations come of age: Protein-ligand recognition,” *Accounts of Chemical Research*, vol. 35, no. 6, pp. 430–437, 2002.

[129] M. J. Stewart, S. Parikh, G. P. Xiao, P. J. Tonge, and C. Kisker, “Structural basis and mechanism of enoyl reductase inhibition by triclosan,” *Journal of Molecular Biology*, vol. 290, no. 4, pp. 859–865, 1999.

[130] J. G. Kirkwood, *Theory of Liquid*. 1968.