Stony Brook University

Quantitative Study of Protein Folding and Conformational Switch with Molecular Dynamics Simulation and Multi-Dimensional Spectroscopy

A Dissertation Presented

by

Zaizhi Lai

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Chemistry

(Chemical Physics)

Stony Brook University

August 2014

Stony Brook University

The Graduate School

Zaizhi Lai

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Jin Wang–Dissertation Advisor

Associate Professor of Chemistry and Physics

Trevor Sears–Chair Person of Defense

Professor of Chemistry

Joint with Brookhaven National Laboratory

David F. Green–Third Member of Defense

Associate Professor of Applied Mathematics & Statistics

Thomas C. Weinacht–Outside Member of Defense

Professor of Physics and Astronomy

This dissertation is accepted by the Graduate School

Charles Taber

Dean of the Graduate School

Abstract of the Dissertation

Quantitative Study of Protein Folding and Conformational Switch with Molecular Dynamics Simulation and Multi-Dimensional Spectroscopy

by

Zaizhi Lai

Doctor of Philosophy

in

Chemistry

Stony Brook University

2014

This dissertation presents the quantitative approaches towards the study of dynamics and nonlinear spectroscopy of protein folding and conformational switches. Molecular dynamics (MD) simulations and two-dimensional spectroscopy computations were employed in the investigation of two protein systems: Glutamine-binding protein (GlnBP) and the Trp-cage. GlnBP is one of the periplasmic binding proteins that carry small ligands from the periplasmic space into the cytoplasmic space. In the process of the conformational transition, GlnBP exhibits two stable states, that is, the ligand-free open state and the ligand-bound closed state. Traditionally, the potential energy shape in the molecular dynamics simulation is one basin. In this work we applied a structure-based two-well potential energy model to study the properties of the kinetics and statistical distributions for the conformational transition of GlnBP. The analysis shows that below the melting temperature, the open and closed basin of attractions emerge and the kinetic analysis

through the mean and distribution of the first passage time as well as the auto-correlation function implies the complexity and the hierarchical structure of the underlying energy landscape. The multi-dimensional diffusion dynamics of GlnBP conformational change were also investigated in this work. We found that the diffusion is anisotropic and inhomogeneous. The directional and positional dependence of diffusion have significant impacts on the protein conformational kinetics: the dominant kinetic path of conformational change is shifted from the naively expected steepest descent gradient paths. The kinetic transition barrier with considering coordinate-dependent diffusion coefficient is shifted away from the transition barrier without considering the coodinate-dependent effect.

This work also proposes the use two-dimensional infrared spectroscopy(2DIR) and two-dimensional ultraviolet spectroscopy(2DUV) to characterize the folding mechanism of the mini-protein Trp-cage. In this study the Trp-cage was folded by atomistic MD simulation and intermediate conformational ensembles were clustered along the dominant folding pathway of energy landscape. The nonchiral and chiral two-dimensional coherent spectra were calculated for the intermediate and folded states of the mini-protein. A direct structure-spectra relationship was determined by the analysis of conformational properties. The structural origins of diagonal and off-diagonal peaks in the 2DIR spectrum were identified for the folded and intermediate conformational ensembles in the folding mechanism and isotope-labeling was used to reveal residue-specific information. Besides, the complexity of 2DUV signals decreases as the conformational entropy decreases during the folding process, implying that the approximate entropy of the signals provides a quantitative marker of the protein folding status. These works support the implementation of computational techniques in conjunction with experimental two-dimensional spectroscopy

to study the folding mechanism of proteins.

Contents

List of Figures

List of Tables

List of Abbreviations

ApEn approximate entropy

CG Coarse Grained

CI Chirality-Induced

DC Diffusion Coefficient

DFT Density Functional Theory

EHEF excitation Hamiltonian with electrostatic fluctuations

FEL Free Energy Landscape

FPT First Passage Time

FNC Fractions of Native Contact

GlnBP Glutamine-Binding Protein

MD Molecular Dynamic

MFPT Mean First Passage Time

NUV Near-Ultraviolet

FUV Far-Ultraviolet

RMSD Root Mean Square Deviation

2DIR Two-Dimensional Infrared

2DUV Two-Dimensional Ultravoilet

VCD vibrational circular dichroism

Acknowledgments

I would like to acknowledge the following for their helpful discussions and support of my work. Firstly, I would like to thank my advisor Prof. Jin Wang. Without his inspiration, support and continually guidance, this work could not be accomplished. In the last five years, I learned a lot from him and I will always remember that. Professor Shaul Mukamel from University of California, Irvine, who provided helpful thoughts about my work. His kind help and guidance made my research work possible and better. I would like to thank my committee, Prof. Trevor Sears, Prof. David Green and my outside member Prof. Thomas Weinacht. I am grateful for their insights, criticisms and support over the years. I would also like to thank staff from Wang and Mukamel's groups, who have been helpful in discussions of my works and provided valuable cooperation, feedback and support on a number of topics. They include Dr. Qiang Lu, Weixin Xu, Cong Chen and Dr. Yuan Yao in Jin's group and my friend Xiaoyu Jin and Prof. Jun Jiang, Hao Ren and Dr. Nicholas Preketes from Shaul's group. Finally, I heartly thank my family for their great love and support.

Publications

[1]. Jiang, J.; **Lai, Z. Z.**; Wang, J.; Mukamel, S. *J. Phys. Chem. Lett.* **2014**, 5, 1341-1346.

[2]. **Lai, Z. Z.**; Zhang, K.; Wang, J. *Phys. Chem. Chem. Phys.* **2014**, 16, 6486-6495.

[3]. Ren, H.; **Lai, Z. Z.**; Biggs, J. D.; Wang, J.; Mukamel, S. *Phys. Chem. Chem. Phys.* **2013**, 15, 19457

[4]. **Lai, Z. Z.**; Preketes, N. K.; Jiang, J.; Mukamel, S.; Wang, J. *J. Phys. Chem. Lett.* **2013**, 4, 1913-1917.

[5]. **Lai, Z. Z.**; Preketes, N. K.; Mukamel, S.; Wang, J. *J. Phys. Chem. B* **2013**, 117, 4661-4669.

[6]. Xu, W. X.; **Lai, Z. Z.**; Oliveira, R. J.; Leite, V. B. P.; Wang, J. *J. Phys. Chem. B* **2012**, 116, 5152-5159.

[7]. **Lai, Z. Z.**; Lu, Q.; Wang, J. *J. Phys. Chem. B* **2011**, 115, 4147-4159.

[8]. **Lai, Z. Z.**; Su, J. G.; Chen, W. Z.; Wang, C. X. *Int. J. Mol. Sci.* **2009**, 10, 1808-1823.

[9]. **Lai, Z. Z.**; Jiang, J.; Mukamel, S.; Wang, J. Exploring the protein folding dynamics of beta3s with two-dimensional ultraviolet (2DUV) spectroscopy. **2014** (Accepted by *Isr. J. Chem.*).

Chapter 1 Introduction

1.1 Protein and Protein folding

Proteins are heteropolymer chain biomolecules, built by the assembly of the twenty amino acids occurring in nature through the chemically stable peptide bond. Proteins perform and control most functions in almost all living organisms, which include assisting molecules to cross cell membranes, transmitting information between specific cells and organs, catalyzing biochemical reactions, transporting and storing of a variety of nutrition elements, regulating the activity of the immune system etc. In order to perform their biological functions, proteins have to fold their unique and stable three-dimensional structures, also known as the native conformations. The process in which a protein reaches its native conformation starting form the loosen structure or random coil is called protein folding, which is a complicated process that has attracted the attentions of scientists for more than half of a century. Protein folding is one of the important processes in the genetic central dogma. Solving the problem of protein folding has important academic significance. Moreover, protein misfolding is often the root cause of diseases. As such proteins are the primary target of pharmaceuticals developed for the treatment of human disease. Specifically, the activity of misfolded proteins has been implicated in diseases including Alzheimers, Diabetes, Parkinsons[1–3], etc. Although numerous efforts have been made to investigate this problem, the mechanism of this physical chemistry process remains elusive. There are essentially two prospects involving in this incompletely resolved problem:(a) the thermodynamic question of how a native structure results from the inter-

atomic forces acting on an amino acid sequence; (b) the kinetic problem of how a native structure can fold so fast. Let us consider a protein which has only 100 amino acids and assume that there are only three possible orientations per residue, we obtain 3^{100} possible conformational states. If one assumes that an jump from one conformation to the another one requires 100 picoseconds, although we exclude the inaccessible conformation due to steric reason, it still would take around 10^{27} years which is longer than the age of universe to randomly explore all other conformations before acquiring the native state. However, in reality, typical folding times range from microseconds to seconds. This puzzle is known as Levinthals paradox[4].

1.2 Anfinsen Assumption

How can proteins reach their native state among the abundant diversity of conformational space? The first attempt to address this question came from Anfinsen, whose studies on the re-folding of ribonuclease[5] showed that protein sequences under physiological conditions can automatically find their native state by minimizing the free energy. Anfinsen assumption[5] suggests that all the information for protein folding is coded in its amino acid sequence and the native state exists as the state in normal physiological conditions which minimizes its Gibbs free energy globally. This means that the native state is a unique minimum determined by the amino acid sequence as well as the environment it is in. Anfinsen's assumption gives the thermodynamic aspect of protein folding and has an enormous impact on molecular biology. Yet, kinetically, the question remains. The Anfinsen assumption does not tell how a protein folds in a reasonable biological timescale. Thus, the study of protein folding should focus on finding a physically based underlying mechanism responsible for guiding the process thermodynamically and kinetically.

1.3 Energy Landscape Theory

The theory of the energy landscape[6–9] provides not only a simply way of understanding why the Levinthal paradox is not a real problem, but also a conceptual framework for understanding the various scenarios of protein folding and other molecular motions, for instance, protein binding. According to this theory, protein folding proceeds on the a moderately rough energy surface, whose major features are the local minimum and the overall funnel shape sloping toward the folded state. Random hetero-polymer, that is, a polypeptide chain consisting of random amino acid sequences, may have either a very rugged energy landscape with too many local minima or a very flat energy landscape[10]. Such proteins can either easily trap in a local minima and cannot jump out or never find the global minimum when walking randomly on the flat energy surface. Therefore, because of evolution, the real proteins always contain optimized sequences so that they can fold rapidly and efficiently into native conformational states[11]. In order to fold in the reasonably biological time, proteins should have funnel-like energy landscape[11]. Although the energy landscapes are high dimensional, they are usually projected into the two dimensions in which the vertical axis represents the energy and the horizontal axis represents the conformational entropy, as shown in the figure 1.1. The energy landscape model is widely considered as the most realistic model for protein folding. It provides a quantitative description of protein conformational space including the native state, various unfolded or denatured states, and folding intermediate state.

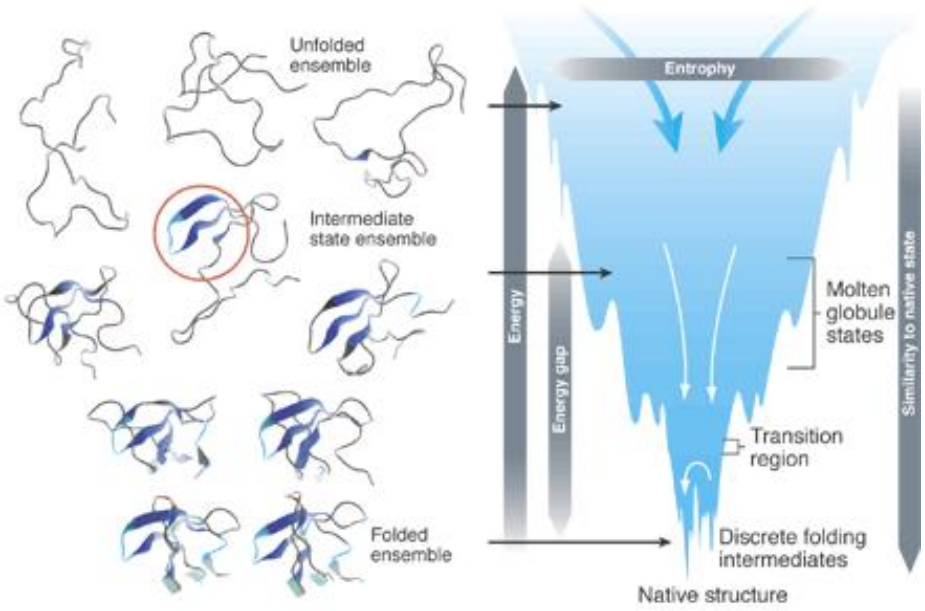


Figure 1.1: A rugged funnel-like free energy landscape that describes the detailed processes and various intermediates in the folding reaction. The width of the funnel represents the conformational entropy; the depth of the funnel represents the change in energy between the denatured and native states. This figure is from reference[12].

1.4 Protein Binding

In addition to the problem of protein folding, another fundamental principle of biological processes is the molecular recognition and binding. Essentially, proteins need to interact with other molecules such as peptides, ligands, and substrates in order to perform their functions. Thus, a better understanding of biological processes requires the study of both the protein folding mechanism and also the mechanism of protein-ligand binding. Protein folding and binding are similar processes because of their common essence: the recognition and organization/reorganization of amino acides. The difference between folding and binding is the presence and absence of the chain connectivity between their

components, leading to two different terms, i.e., the intramolecular and the intermolecular recognition. Therefore, the energy landscape theory may provide a consistent theoretical framework to describe the mechanisms of protein folding and binding. In this thesis, we will use the energy landscape theory to investigate the mechanisms of protein folding and the protein binding process with large conformational switches.

1.5 Molecular Dynamic Simulation

1.5.1 Equations of Motions in Molecular Dynamic

Although the a sophisticated theoretical framework to study protein folding, binding or other molecular motions has been provided, we still need the quantitative studies to demonstrate the theories and connect to experiments. Computer simulations provide a bridge between microscopic length and time scales and the macroscopic world. Basically there are two main families of computer simulation techniques: Molecular Dynamics(MD) and the Monte Carlo(MC). In this thesis we use the MD simulation approach. Based on the empirical potential energy functions, the motion of a system is simulated through step-by-step calculation of particles interactions, coordinates and velocities according to the classical Newtonian equation:

$$m_i \frac{d^2\mathbf{r}_i}{dt^2} = \mathbf{F}_i, \quad i = 1, 2, ..., N. \tag{1.1}$$

$$\mathbf{F}_i = -\frac{\partial U}{\partial \mathbf{r}_i}, \quad i = 1, 2, ..., N. \tag{1.2}$$

where N is the number of atoms, $\mathbf{r}_i$ is the vector of Cartesian coordinates of the i-th atom, $\mathbf{F}_i$ is vector of forces acting on the i-th atom, and U is empirical potential energy and also called force field in the molecular dynamic simulation.

1.5.2 Force Field

The force field plays a key role in a molecular dynamic simulation. In Equation 1.2, the potential forces are calculated by using the empirical potential functions, which can be provided by some mature software such as CHARMM(Chemistry at Harvard Molecular Mechanics)[13], AMBER(Assisted Model Building with Energy Refinement)[14] and so on. In this thesis, the systems were studied by both CHARMM and AMBER software. Force field functions and parameter sets are derived from both experimental work and high-level quantum mechanical calculations. The design and parameterization of force fields for use in protein simulations is a complex task, involving many decisions concerning which data to emphasize in the fits. There are some review papers in this field[15,16]. The commonly used protein force fields incorporate a relatively simple potential energy function:

$$\begin{aligned} U &= \sum_{bonds} K_b(b-b_0)^2 + \sum_{angles} K_\theta(\theta-\theta_0)^2 \\ &+ \sum_{dihedrals} K_\phi[cos(n\phi+\delta)+1]^2 \\ &+ \sum_{nonbondpairs} (\frac{A}{r_{ij}^{12}} - \frac{C}{r_{ij}^{6}} + 4\pi\epsilon_o \frac{q_i q_j}{r_{ij}}) \end{aligned} \tag{1.3}$$

The first three summations are over bonds (1-2 interactions), angles (1-3 interactions), and dihedrals (1-4 interactions). The dihedrals term can also include so-called improper torsions, where the four atoms defining the angle are not all connected by covalent bonds. The last term includes the dispersion and exchange repulsion forces that are represented by a Lennard-Jones 6-12 potential; this is often called the "van der Waals" term. The electrostatic interaction assumes partial charges q_i on each atom that interact via Coulombs law. If all the atoms of a protein are simulated by using the above potential energy function, we call it an all-atom model[17]. Although currently workers have developed powerful computational resources, the simulations of protein molecular dynamic are still very time-

consuming. When running a folding or unfolding trajectory, we usually need to run tens of millions of steps, and for each step, we need to solve the Equations1.1 and 1.2 for thousands of atoms in a general protein, when the all-atom model is considered. If we further explicitly consider the solvent effect, the simulations may become impossible for a large protein system. To overcome this limitation, we can the implicit-solvent model[18], which does not explicitly include the solvent molecules in the simulation but rather implicitly considers the whole solvent as a continuum. This approach can dramatically reduce the computational time. However, the computations are still expensive. Therefore, many coarse-grained or simplified models have been developed to study large and interesting molecular systems.

1.5.3 Reduced Models

Reduced, simplified or course-grained models of proteins in which each amino acid is represented by a few interaction sites, provide an extension of the timescale of simulations compared with that of all-atom model. The general simplified models usually consider a single interaction site per residue, and the potentials between the sites are Lennard-Jones types or other simpler contact potentials. The energy surfaces that are produced by the simplified potential functions may not offer as many details as all-atom models do, nevertheless, the reduced models have made significant contributions to our understanding of molecular structure and function[19–21]. One of the reduced models is the structure-base model[22](also called Go model) in which the potentials for those pairs of residues in contact in the native structure are attractive and those between other pairs repulsive. This model is based on the assumption that the native-state structure largely determines the process of folding or functional motion. In this thesis, part of the work is to apply a modified

structure-base model to study the protein binding process.

1.6 Two-Dimensional Spectroscopy

The MD simulation has provided a very useful tool to study the biological molecular at atomic level, yet to further connect the theories and experiments, development of other tools is necessary. Multidimensional nonlinear spectroscopy[23–26] is a powerful tool for the study of vibrational and electronic excitations of molecular. It can provide detailed information on the dynamics and structure with high temporal and spatial resolution by using femtosecond laser pulse sequences that interact with the molecule and generate coherent nonlinear signals, which include the rich information about the couplings between different parts of the molecule, providing a multidimensional view of molecular structure, interactions, and motion processes. Multidimensional spectroscopy techniques provide a novel tool for studying protein folding. Moreover, studying the dynamical process of the big systems will further inspire development of new methods and techniques of multidimensional nonlinear optical spectroscopy. Here we apply the techniques of multidimensional nonlinear optical spectroscopy, including two-dimensional infrared(2DIR), two-dimensional ultraviolet(2DUV) to investigate the folding process on the energy landscape. Details of muiltidimensional spectroscopy will be described in the fourth and fifth chapters.

1.7 Summary

In Chapter 1 the background of protein folding and some methods and tools which will be used in this thesis was introduced. Levinthal paradox was derived from the problem of protein folding. Anfinsen explained the thermodynamic aspect of this paradox and

proposed the Anfinsen assumption, which suggested that the process of protein folding is controlled by minimizing free energy globally. However, Anfinsen assumption did not explain the kinetical aspect of the paradox. The theory of free energy landscape provides a consistent and sophisticated theoretical framework for explaining the protein folding thermodynamically and kinetically. To quantitatively study the free energy landscape, in the thesis, we used the methods of molecular dynamics simulation and two-dimensional spectroscopy.

Chapter 2 Investigating Protein Conformational Transition with a Double-Well Model

2.1 Abstract

The structure of the glutamine-binding protein (GlnBP) from Escherichia coli is representative of periplasmic binding proteins. A large functional conformational transitions in the Glutamine-binding protein (GlnBP) are very important to bind and transfer the ligand Glutamine. Upon ligand binding, the GlnBP changes conformation such that it is subsequently recognized by a specific inner membrane transporter. A functional GlnBP has two stable states, that is, open state and closed state. Here a structure-based double-well model is applied to investigate the kinetic and dynamic properties of the GlnBP conformational transition. Free energy landscape essentially is a function of temperature. To investigate the various behaviors at different temperatures, the underlying free energy landscape of the conformational transition with different temperatures are constructed and the analysis shows that, below the melting temperature, they have two basins corresponding to the open state and the closed state of the protein, respectively. We also investigated the first passage time distribution and the auto-correlation function to probe the kinetic properties of the conformational switch. The complicated kinetic implies the complexity and the hierarchical structure of the underlying energy landscape. The contact maps of the structures are built up to probe the structural evolution of the conformational transition. Finally, the ϕ values of the residues are calculated to illustrate the important residues of the transition state.

2.2 Introduction

Many proteins need to process large conformational transitions between different stable states to perform their biological functions. The flexibility and plasticity of a protein allows it to bind ligands, form oligomers, and perform mechanical work. Current experiments, including X-ray diffraction and NMR spectroscopy, are able to characterize the structure of a biomolecule. The dynamical properties of a biomolecule can be probed by spectroscopic techniques[27,28]. Although they may provide exact static structures or information on the local environment surrounding the probe, global time-dependent structural information is also difficult to be obtained directly by experiments, providing fewer details on the dynamic and kinetics of conformational transitions. To more fully understand the mechanism of ligand-induced conformational change one would also like to characterize the structural dynamics of the transition. The gap between theory and experiment can be filled by the computational simulations which potentially provide full time-dependent structural information on biomolecules. In the last couples of decades, advances in computational science have been significant development, modeling the whole processes of protein motions, including binding and folding, at all-atom level for the large systems. Most processes of interest occur on time scales (microsecond to second) inaccessible to standard all-atom molecular dynamics simulations. One approach to overcome the problem of long-time scales in simulation is to use a simplified model. We propose to achieve a long-time molecular dynamics simulation by describing the protein interactions in a coarse grained way at the residue level in which the water molecules are not explicitly included and native interactions are preferred.

Functional conformational transitions require a biomolecule to have at least a pair

of conformational states. For the conformational transition, one important question is how to describe numerous sub-conformations fluctuating around the two stable states and the transition between them. The natural and simple way of modeling is so-called energy landscape theory[6–9], which was originally applied to solve the problem of protein folding: the underlying energy landscape is funnel-like which leads the faster kinetics and promises the thermal stability and specificity. In a perfect funneled landscape, the global characteristics of the structural heterogeneity in the transition state ensemble seems mostly determined by the native structure. The structure-base model[22] was proposed to emphasize the importance of the native structure by generalizing structure-base type interactions[22], that is, the attraction interactions are assigned to the pair of the residues that interact in the native structure and repulsive interactions are endowed to the other contacts. The model was expected to provide useful information about the topology of the energy landscape and, from the previous studies[29–31], has been proven to be an effective model and consistent with many experiments. In our work, a new interaction was considered in the standard structure-based model to simulate the large conformational changes and to study the dynamic and kinetic properties controlled by the underlying multiple-basin energy landscape of proteins. The new interaction forms a two-well potential. One corresponds to the open state and the other corresponds to the closed state.

GlnBP was used as the model protein to simulate the conformational transitions. Two reference structures were supplied by X-ray crystallography[32,33], as shown in Fig. 2.1(a) and Fig.2.1(b). GlnBP contains a single polypeptide chain of 226 residues. The tertiary structure of GlnBP consists of 35 percent α-helix and 37 percent β-sheet. The GlnBP is composed of two similar globular domains. The large domain includes two separate peptide segments, residues 1 to 84 and residues 186 to 226, and the small domain

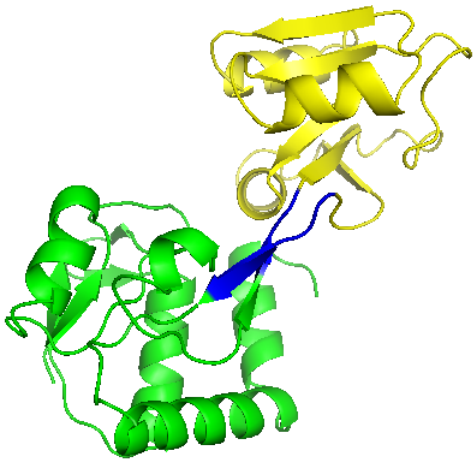
includes residues 90 to 180. These two domains are linked by toe peptide hinges, which are residues 85 to 89 and residues 181 to 185.These two domains look like two arms that can open and close thus binding and releasing the ligand Glutamine. For the ligand-free open structure of the GlnBP, the PDB code is 1GGG; for the ligand-bound closed structure of the GlnBP, the PDB code is 1WDN.

2.3 Model

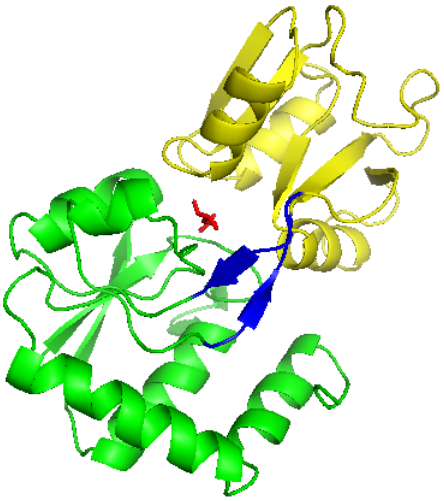
As described above, the GlnBP has two stable states and it is reasonable to use a two-well model to simulate this system. In our simulations, we used a modified potential energy U[34] for a given protein conformation Γ:

$$\begin{aligned} U_{whole}(\Gamma,\Gamma_1,\Gamma_2) &= \sum_{bonds}^{N-1} K_b(b_i-b_{0i})^2 + \sum_{angles}^{N-2} K_\theta(\theta_i-\theta_{1i})^2(\theta_i-\theta_{2i})^2 \\ &+ \sum_{dihedrals}^{N-3} K_\phi[cos(\phi_i-\frac{\phi_{1i}+\phi_{2i}}{2})-cos(\frac{\phi_{1i}-\phi_{2i}}{2})]^2 \\ &+ \sum_{|i-j|>3}^{nonnative} \epsilon\left(\frac{M}{r_{ij}}\right)^{12} + \sum_{|i-j|>3}^{native} V_{nat}(r_{ij}) \end{aligned} \tag{2.1}$$

Where $\Gamma_1(\Gamma_2)$ represents the native open(closed)state. The first three terms describe the bond lengths, the bond angles and the torsion angles, respectively. For the bond term, since the change between the open state and closed state is small, a single harmonic potential was used; for the angle and dihedrals term, the changes between the two states are so significant that the two-well from potential were applied to calculate the interactions. The fourth term and fifth term give the non-bond interaction potential and the native interaction, respectively. M is a constant number. In principle, the parameters of each term can be calculated and fitted by quantum chemical calculations[15], or can be calibrated from the experimental data of NMR[35], X-ray[36], B-factor[37] and so on. For convenient, in our study, we applied the similar values that have been used in the several



(a) Unbound-open structure



(b) Bound-closed structure

Figure 2.1: (a)The unbound-open structure of GlnBP. The region with the green color is the large domain of GlnBP, and the region with the yellow color is the small domain. The hinge region is labeled by blue color. (b) The bound-closed structure of GlnBP. The green region and yellow region represent the large domain and the small domain of GlnBP, respectively. The hinge region is labeled by the blue color, and the ligand glutamine is marked by the red color.

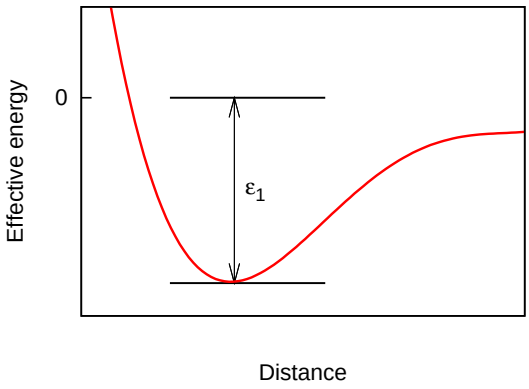
folding studies[6,34,38]. The explicit representation of the native interactions(the fifth term) is following[39].

$$\begin{cases} \epsilon_1 Z(r)(Z(r)-a) \text{ with } Z(r)=\left(\frac{r_1}{r}\right)^k \text{ if } r<r_1 \\ CY(r)^n \frac{\frac{Y(r)^n}{2}-(r_h-r_1)^{2n}}{2n}+\epsilon_2 \text{ with } \begin{cases} Y(r)=(r-r_h)^2 \\ C=\frac{4n(\epsilon_1+\epsilon_2)}{r_2-r_h} \end{cases} \text{ if } r_1 \le r<r_h \\ -B\frac{Y(r)-h_1}{Y(r)^m+h_2} \text{ with } \begin{cases} B=\epsilon_1 m(r_2-r_h)^{2(m-1)} \\ h_1=\frac{\epsilon_h(m-1)(r_2-r_h)^2}{m(\epsilon_h+\epsilon_2)} \\ h_2=\frac{\epsilon_2(m-1)(r_2-r_h)^{2m}}{\epsilon_h+\epsilon_2} \end{cases} \text{ if } r_h \le r<r_2 \\ \epsilon_2\left[5\left(\frac{r_2}{r}\right)^{12}-6\left(\frac{r_2}{r}\right)^{10}\right] \text{ if } r_2 \le r \end{cases} \tag{2.2}$$

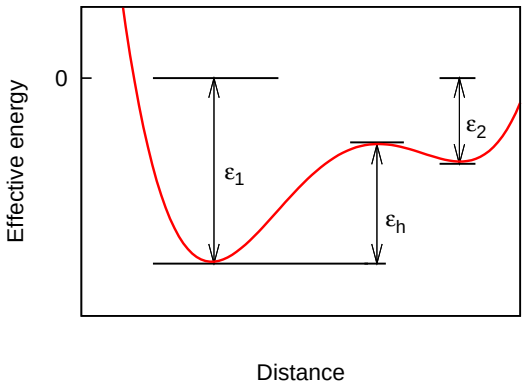
where m=5, k=8, and n=1. The smooth and continuous curve can be obtained by using these parameters, as shown in Fig.3.1. Using this format of potential energy was inspired by the work of Margaret et al[39] in which the authors applied a phenomenological two-well model to the study the water desolvation effects in the protein folding, and such effects exhibit multi-minimum feature in terms of potential energy. Their results are consistent with experiments well, implying the model could be proper for studying the problems of multi-minimum. Also, this two-well model has been applied to study the conformational dynamics of adenylate kinase and obtained well results[34]. The two stable states of GlnBP correspond to the two wells of potential formula. The first well corresponds to the closed structure, and the second well corresponds to the open structure. It is worth to mention that this two-well model is microscopic since it represents the interaction between every two residues. This is different from other types of the two-well models[40–43]. Kei-ichi Okazaki *et al*[41,42] considered the topological characteristic of the energy landscape of proteins and created two independent structure-based potentials and connected them smoothly to make a double-well model. The same spirit can be found in the work of Best *et al*[40]. Comparing to these models, one important feature of our model is to consider the microscopic situation of each contact pair. According to the known structures, we specified the form of interaction for each pair. In this work, we actually used the mixture of single-

well model and two-well model to illustrate the profile of the GlnBP's conformational switch. Three situations were distinguished in our work due to the distance between two residues in the open state r_{ij}^{open} and closed state r_{ij}^{close}. When r_{ij}^{open} equals r_{ij}^{close}, the interactions between the residue i and j have no difference in the open state and closed state. Therefore, a single-well potential is applied to the residues i and j, as shown in Fig.2.2(a); when $0 < |r_{ij}^{open} - r_{ij}^{close}| < 2\mathring{A}$, which means the differences in the open state and closed state are not very significant, a two-well potential with a shallow barrier is applied to the residues i and j, and the form of this kind of two-well potential is very similar to the single well, as shown in Fig.2.2(b); in the case $|r_{ij}^{open} - r_{ij}^{close}| > 2\mathring{A}$, we use the two-well potential with a significant barrier, as shown in Fig.2.2(c). We can use the CSU software[44] to calculate the native contact pairs of the open and the close states.

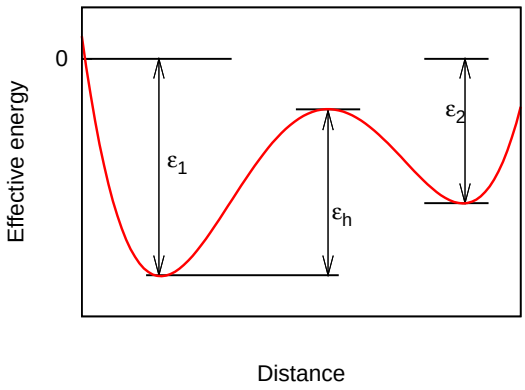
After assigning different groups of residues with different shapes of two-well models, we need to fix the shape of the two wells. There are three parameters ϵ_1, ϵ_2, and ϵ_b that we should choose to decide the form of the two-well potential, as shown in Fig.2.2(c). The ϵ_1 represents the depth of the first well. This depth mainly controls the melting temperature of the protein. The ϵ_2 is the depth of the second well. The difference between ϵ_1 and ϵ_2, that is, $\Delta\epsilon = |\epsilon_1 - \epsilon_2|$, controls relative stability of the two states. The last parameter ϵ_b is the energy barrier between the two states. In our model, the melting temperature is close to 355K. So from the calibration of the simulations, we set $\epsilon_1$to 0.58 kcal/mol. For convenience, we adjusted these two parameters so that the conformational switch happened between the open state and the closed state occurs in a reasonable time. First we kept on changing the value of $\Delta\epsilon$ until the protein spends almost equal time in each state. Then we gradually increased the value of ϵ_b to control the transitions within the reasonable computation time. In this work, $\Delta\epsilon$ was set to 0.12 kcal/mol and ϵ_b was set



(a) One well



(b) Shallow barrier



(c) Two well

Figure 2.2: The Schematic representation of the potentials. (a) Two-well potential with a significant barrier. ϵ_1 is the depth of the first well; ϵ_2 is the depth of the second well; ϵ_b is the height of the barrier. (b) Two-well potential with shallow barrier. The height of barrier is half of that in the plot (a). (c) Single well potential. In this paper, $|\epsilon_1| > |\epsilon_2|$. Those well shapes are just the schematic representation.

.

to 0.35 kcal/mol. After all these models and parameters were set up, we embedded these models into a modified AMBER software and ran the MD simulations.

2.4 Methods and Results

In order to study the different behaviors of the GlnBP at different temperatures, the simulations were performed at four temperatures. In our model, the GlnBP began to melt at 355K. Besides the melting temperature, the other three temperatures chosen were 275K, 310K, and 336K. For each of these four temperatures, fifty AMBER trajectories with 50 million steps for each trajectory were simulated in order to obtain the reliable statistical calculation. Typical simulation trajectories for these four temperatures are shown in Fig.2.3. In Fig.2.3(d), one can observe that the protein begins to melt at 355K. Also, below the melting temperature, we can see the reversible transition between two states. The conformational switch occurs very rapidly, without any intermediate state. In other words, the breaking of the contacts in the initial structure and formation of the contacts in the finial structure occur almost at the same time. This observation is consistent with the previous study[41]. The RMSDs of the two states exhibit that the protein in the open state has greater conformational fluctuation than that in the closed state. This implies that, with the ligand binding, the protein-ligand complex may be more stable than the ligand-unbound structure. Interestingly, from Fig.2.3(a) to Fig.2.3(c), one can observe that the protein prefers to stay in closed state at low temperature and biases to the open state at high temperature. This phenomenon can also be observed clearly in Fig.2.4(a), which describes quantitatively the percentage of staying in the open state or the closed state of the protein at different temperatures. At 275K, the protein has not enough energy so that the frequency of transition is low, and the protein remains in the

closed state for 80% of the simulation time. At 310K, the protein spends almost the same time in the both states. When the temperature reaches 336K, the conformational change occurs frequently and the open state is dominant. The essential properties of the protein determines its thermodynamical behavior. In the two-well potential, the closed state has lower energy than the open state. At high temperature, the thermal energy of the protein can break all of the ligand-induced contacts quickly when they are formed. At low temperature, the protein has not enough energy to break the ligand-induced contacts. In other words, the protein has difficulty surmounting the first barrier ϵ_1. Therefore, most of the time it is trapped into the first well at low temperature. As the temperature increases, the energy of the protein increases, and it has more possibilities to cross the barrier, so the percentage of staying in the open state increases. Additionally, from Fig.2.4(b), one can see that, at high temperature, the average residence time in the closed state is shorter than that of the open state, and the average residence time of both states rises when the temperature decreases.

The free energy was considered as a function of RMSD1 and RMSD2, and obtained by $F = -log(P)$, where P was the statistical population obtained from all fifty trajectories for each temperature. The two-dimensional free energy profiles, as shown in Fig.2.5, were constructed by the trajectories with the temperature 275K, 310K, 336K and 355K. In Fig.2.5(a), Fig.2.5(b), and Fig.2.5(c), there are two local minimum in each profile corresponding to the open state (right basin)and the closed state(left basin) of the protein. In Fig.2.5(a), there are very few conformations connecting the two basins, implying the transition is very rare at this low temperature. And the basin of the closed state is deeper than that of the open state, which suggests the protein prefers dwelling in the closed state at low temperature. In Fig.2.5(b), the depth of the two minima are almost the same and

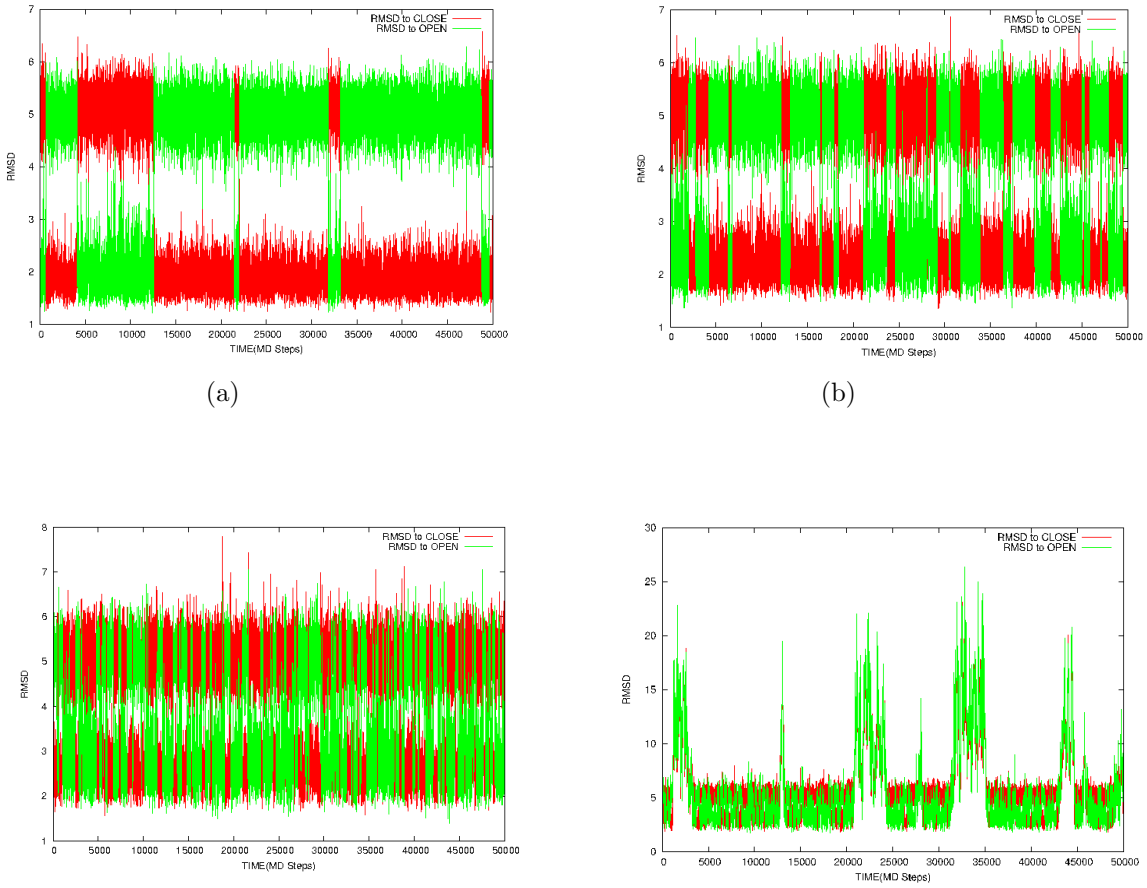


Figure 2.3: Typical trajectories of the simulations at temperature (a) 250K, (b) 310K, (c) 336K and (d) 355K. The red color represents the RMSD to the closed state, and the green color represents the RMSD to the open state. The unit of RMSD is angstrom.

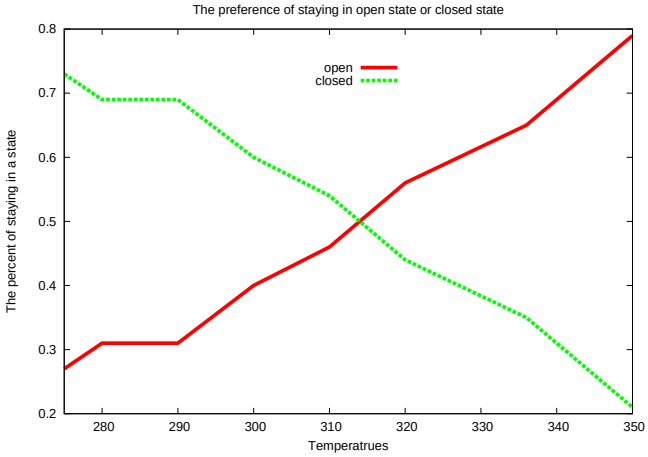
the height of the single free-energy barrier between the two minima equals to 4.4 k_BT. One feature of this free energy landscape is that the right basin corresponding to the open state is broader than the left basin corresponding to the closed state, suggesting the relatively larger conformational fluctuation in the open state. The other distinct feature of the free energy profile in Fig.2.5(d) is that it has a long "tail", which is caused by the large values of root mean square deviation, suggesting the melting state at this temperature. The steepness of the basins is also an interesting topological property on the energy landscape. In Fig.2.5(b), the left basin is steeper than the right basin. This topological characteristic illustrates the following transition dynamic. When the conformational transition commences from the open state, it will undergo relatively more extensive pathway to reach the transition state then downhill to closed state rapidly on the energy surface.

The autocorrelation function $c(k)$, where k is the simulation time step, could be a good property to represent the dynamical motion of the conformation switch. We can either use the closed state or open state to calculate the auto-correlation function. They give basically the same results. Therefore, we just show the results for the closed state. The formula that we used to calculate the $c(k)$ is following:

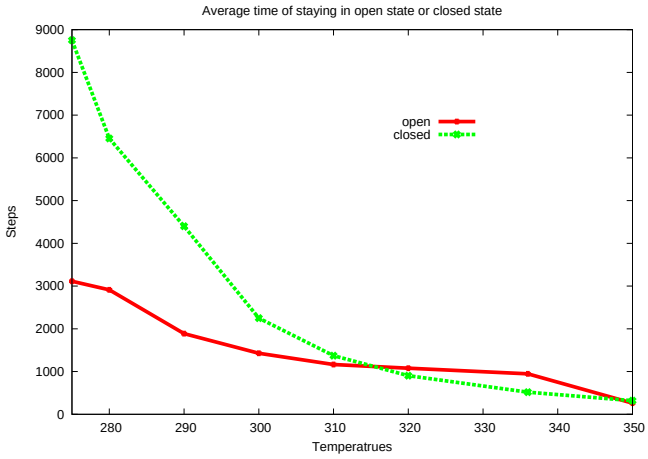
$$c(k) = \frac{\sum_{t=1}^{n-1}(x_t - \langle x \rangle)(x_{t+k} - \langle x \rangle)}{\sum_{t=1}^{n-k}(x_t - \langle x \rangle)^2} \tag{2.3}$$

where x_t is the value of the RMSD from the native state at time t, and $\langle x \rangle$ is the average RMSD value of the whole trajectory. Using the above formula to calculate the $c(k)$ requires large n, the number of the data. Therefore, we simulated long trajectories with 500 million steps at 310K to calculate the auto-correlation function $c(k)$.

Fig.2.7 shows the auto-correlation function in the linear coordinates and the semi-logarithmic coordinates, respectively. In the semi-logarithmic coordinates, y value is logarithmic, and x value remains linear. We see from the Fig.2.7(a) that the auto-correlation

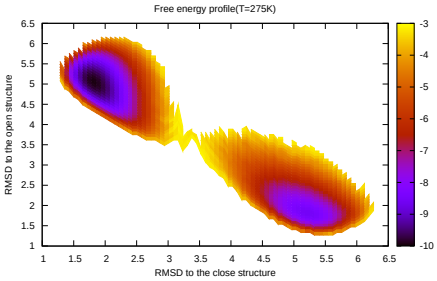


(a)

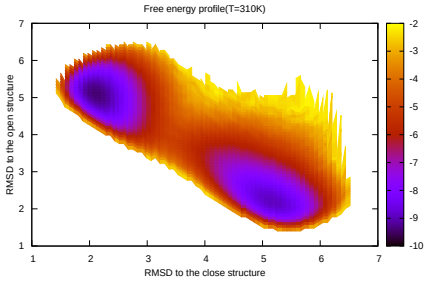


(b)

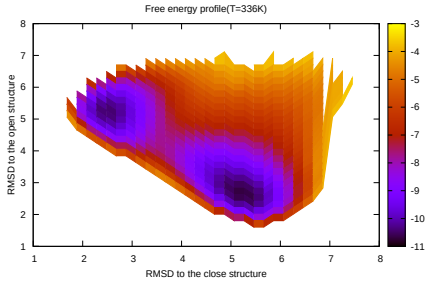
Figure 2.4: (a) Percentages of staying in two states at different temperatures. The red line corresponds to the open state and the green line corresponds to the closed state. (b) Average residence time of two states at different temperatures. The red line corresponds to the open state and the green line corresponds to the closed state. The temperature unit is K.



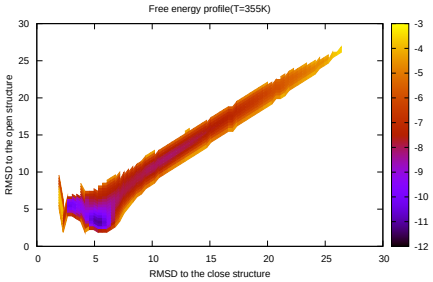
(a)



(b)



(c)

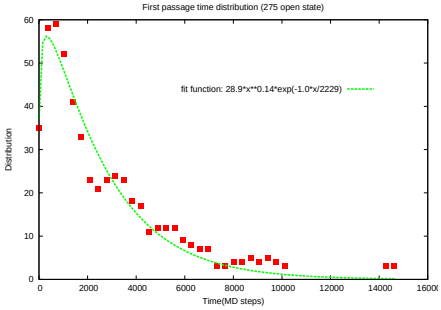


(d)

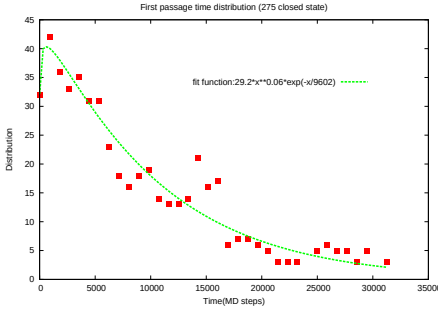
Figure 2.5: The free energy profile of the conformational transition of GlnBP at temperatures (a) 250K, (b) 310K, (c) 336K and (d) 355K. The x and y axis are the RMSDs to the closed structure and open structure, respectively. The unit of RMSD is angstrom.

function decays with time interval k. Also, we can clearly observe from Fig.2.7(b) that the decaying obeys the exponential behavior at the short time limit, $k < 1200$. When $k > 1200$, the decaying deviates the single exponential function. The multiexponential function can fit the auto-correlation function well in the semi-logarithmic coordinates, indicating that the kinetic of the whole conformation-transition process is multiexponential. Since the kinetics can be considered as a good probe to the underlying energy landscape, the multiexponential kinetics implies the complexity of the energy landscape. When a complex system with these characteristics diffuses on the corresponding underlying energy landscape from one state to the other, it will depend on a series of sub-processes. The conformational switch can be recognized as a consequence of these many sub-processes, all of which make the transition possible[45,46].

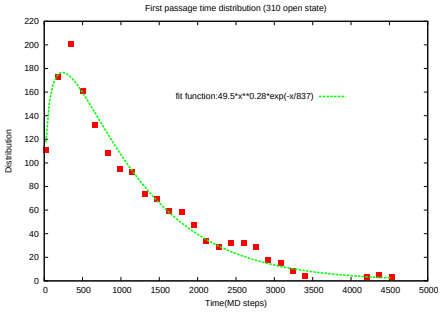
We next investigated the first passage time (FPT) of the conformational transition since the kinetics of the transition between the open state and closed state can be characterized by the first passage time (FPT)distribution of the open state, $P(T_o)$, where T_o is the dwell time of the open state, and closed state, $P(T_c)$, where T_c is the closed residence time, states, respectively. In particularly, the mean first passage times of the open state and closed state, that is, $\langle T_{o(c)} \rangle \propto \int_0^\infty T P(T_{o(c)}) dT$, are significant quantifiers of the transition kinetics. Specially, they determine the mean opening (closing)rates $r_{o(c)} \propto \langle T_{c(o)} \rangle^{-1}$. Furthermore, one may connect the distribution of the FPT to the other kinetic properties, such as the barrier of the transition, controlled by the underlying energy landscape. Fig.3.11 shows the first passage time distributions for the open state and closed state at temperature 275K, 310K, 336K. The first passage time distributions $P(T)$ take the asymptotic form $P(T_{o(c)}) \propto exp(-\lambda_{o(c)} T_{o(c)})$ for large $T_{o(c)}$, where the $\lambda_{o(c)}$ are the exponents corresponding to the open state and closed state, respectively. From the



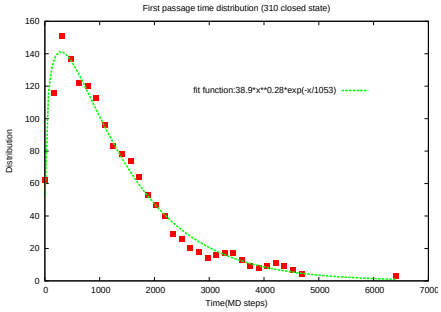
(a) T=275K(open state)



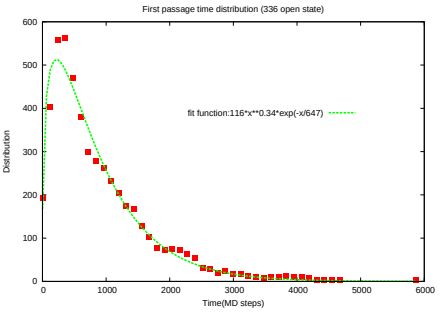
(b) T=275K(closed state)



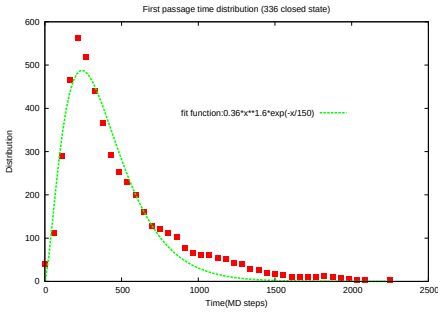
(c) T=310K(open state)



(d) T=310K(closed state)



(e) T=336K(open state)



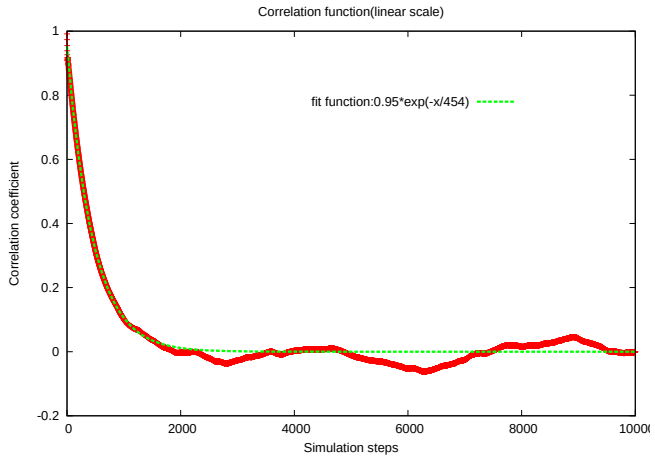
(f) T=336K(closed state)

Figure 2.6: The distributions of the first passage time of the different temperatures. The x axises in plot (a), (b),(c),(d),(e),and (f) represent the number of the molecular simulation steps of staying in the open state and closed state, respectively.

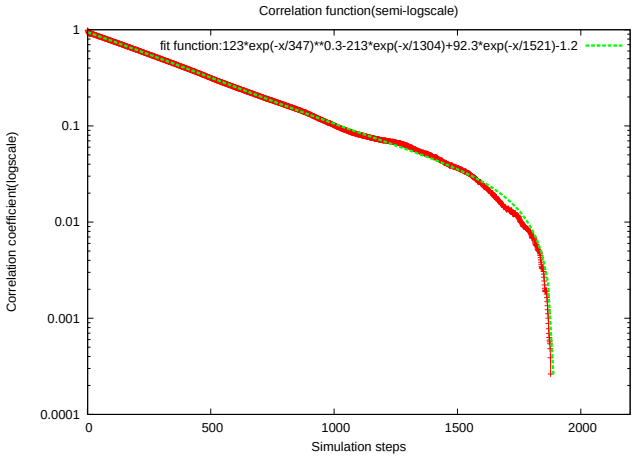
Fig.3.11, one can observe that at high temperature, the difference between the value of mean FPT and the maximal point of the FPT distribution is much smaller than that at low temperature, which means, for the low temperature, the FPT distribution provides more detailed description about the kinetic properties of the conformational transition. As the temperature increases, both λ_o and λ_c increase, suggesting the rate of the conformational switch increases. Yet we should note that those fittings may be influenced by the barrier parameter of the two-well model, and that could be an interesting topic in future.

Our simulations also can catch the structural changes during the process of conformational switch. The residue contact pairs of three different structures on the pathway from the open state to the close state were analyzed by the residue contact map[34]. First, we chose three points (states) on the pathway of the conformational transition on the free energy landscape. These three points are closed state, open state and transition state and the probabilities of the contact pairs are calculated in the different states. The probability means how closely a specific pair is to its native distances (open state or closed state). The formula used is $P_{ij}^k = n_{ij}^k/n^k$, where P_{ij}^k represents the probability of the pair ij in the state k, n^k is the number of the conformations around the state k, and n_{ij}^k is the weight of the pair ij at the state k. The weight was calculated by the distances of the pair ij in the conformations around the state k.

Fig.2.8 shows three contact maps and their corresponding structures. In the contacts maps Fig.2.8(b), Fig.2.8(d), and Fig.2.8(f), each point represents one contact between two residues. A point with red color means the corresponding contact is near to its closed native contact, and a point with blue color implies the corresponding contact is close to its open native contact. The distances of three pairs of residues (10-115, 50-118, 68-157) to



(a)



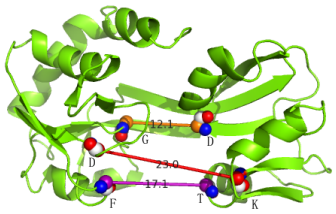
(b)

Figure 2.7: Autocorrelation function distribution of conformational transition. (a) The distribution of autocorrelation function in Cartesian coordinates system. The single-exponential function is applied to fit the data for comparing. (b) The distribution of autocorrelation function in Semi-logarithmic coordinates system in which y value is logarithmic, and x value remains linear. A multi-exponential function is applied to fit the data.

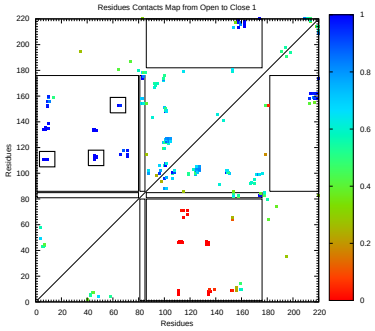
illustrate the meaning of points in the contact maps. The residues Asp10 and Lys115 and the distance between them are marked by the red color; the residues Phe50 and Thr118 and the distance between them are labeled by the purple color; the residues Gly68 and Asp157 and the distance between them are labeled by the orange color. In the contact maps Fig.2.8(b), (d), and (f), each plot has three points enclosed by rectangles. They correspond to the three contacts (10-115, 50-118, 68-157). From the open state to the closed state, the distances of two contacts decrease, and the color of the corresponding points in the contact maps change from blue to red. For comparison, the points under the diagonal in each contact map represent the contacts of the structure of the closed state. From the Fig.2.8(b) to Fig.2.8(f), we can see that the interaction between two groups of residues, which are located in the region 10-75 and region 115-160, changes significantly when conformational transition occurs. These two groups of residues form forceps of the protein to clamp the ligand when it comes in. Specially, there are some interesting residues in these two groups. Previous study[33]has shown that when GlnBP closes, the ligand (Glutamine) is stablized by the residues included in the large domain(Asp10, Ala67, Gly68, Thr70, Arg75) and the small domain(Lys115, Gly119, His156, Asp157). Our simulations show that when conformational transition occurs, Asp157 gets close to Thr70 and Gly68. These three residues stabilize the α-amino group of the ligand[33]. The simulations show that Asp157 moves faster to Thr70 than Gly68. Additionally, the residues Thr70, Arg75 and Gly119 interacting with the α-carboxyl group of the ligands also migrate together. Arg75, Thr70 approach to Gly119 very rapidly. At the transition state, they have formed contacts which are very similar with that of the closed state. The van der Waals interaction plays an important role. The simulations imply that, in the process of the transition, Phe13 comes near to Lys115, Ile139, His156, Asp 157 and Asn160. Simultaneously, Phe50

forms contacts with Ser116, Gly117, Thr118, Pro137 and Asn138. An interesting characteristic of the binding is the "doorkeeper" [33] formed by residues Asp10 and Lys115, which lock the ligand within the binding pocket. The simulations also capture this feature. Additionally, Asp10 may form a contact with Asn138 as well in the conformational transition. Another interesting feature of the contacts formation is that, from open state to closed state, most of the important contacts form slowly before the transition state, and after transition state, the contacts form more quickly. It implies that basin of the closed state on the energy landscape is steeper than that of the open state, as shown in Fig.2.5.

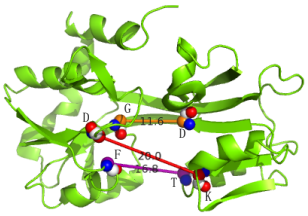
In experiments, ϕ values provide an approach to quantify the strength of native interactions in the transition state[47,48]. Through calculating the ϕ value of a specific residues, one may know its importance in the protein dynamic processes, such as protein folding, conformational transition, and so on.The ϕ values of each residues of the protein in the transition state are shown in the Fig.2.9. Since we are interested in the residues which have significant effect on the transition state, we only exhibited the residues with high ϕ values. It should note that ϕ values obtained from simulations of the simplified model may not correlate well with the experimental values. In our work, the following formula[49] was used to calculate the simulated ϕ values. $\phi_i = \frac{\langle n_i \rangle_{tran} - \langle n_i \rangle_{open}}{\langle n_i \rangle_{closed} - \langle n_i \rangle_{open}}, where \langle n_i \rangle$ is the average value of the number of contacts for residue i, and $tran$, $open$, $closed$ subscripts represent the transition state, open state and closed state, respectively. We can see that some important residues have high ϕ value at the transition state. Thr70, Arg75, and Gly119 play important role in stabilizing the ligand, and their ϕ value are 0.74, 0.77 and 0.71, respectively. Gln183 and Tyr185 from the second hinge may also participate in the conformational transition. Their ϕ values are 0.80 and 0.79, respectively. Other residues with high ϕ values, such as Thr72 (ϕ value 0.88), Tyr86 (ϕ value 0.86), Asp122



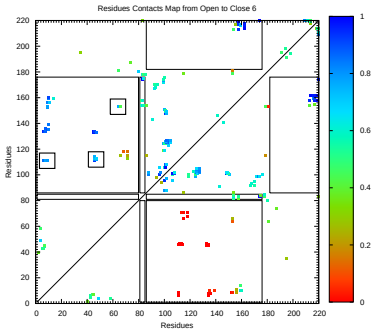
(a) Open state



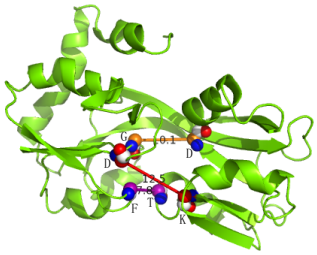
(b) Contact map for open state



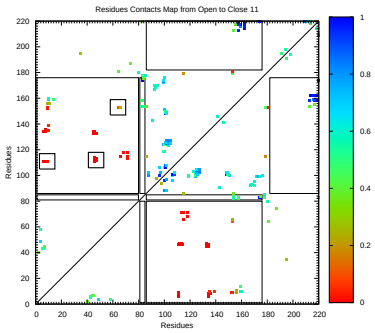
(c) Transition state



(d) Contact map for transition state



(e) Closed state



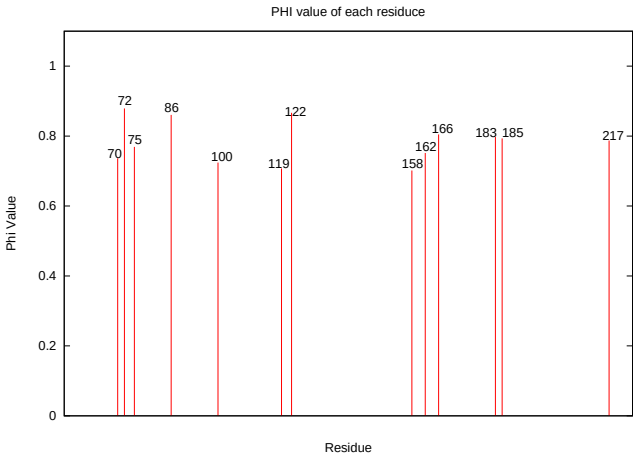
(f) Contact map for closed state

Figure 2.8: Residues contact maps and the corresponding structures on the pathway from the open state to the closed state. The plots (b),(d), and (f) in right column show the contact maps of the open state, transition state and closed state, respectively. The plots (a),(c),and (e) in the left column represent the corresponding structures.

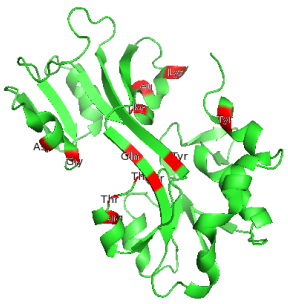
(ϕ value 0.87), Leu162 (ϕ value 0.75), and Lys166 (ϕ value 0.80), can be observed in the simulations, implying that these residues may be involved in the conformational transition as well and play important roles.

2.5 Discussion and Conclusion

In this work, a developed structure-based two-well model was applied to study the properties of the kinetics and statistics distributions for the conformational transition of the GlnBP, which is one of periplasmic binding proteins which carries small ligands form the periplasmic space into the cytoplasmic space. In the process of conformational transition, the GlnBP exhibits two stable states, that is, ligand-free open state and ligand-bound closed state, implying that the protein may go through two distinct local minimum on the potential surface. We constructed the free-energy landscapes of the conformational transition with different temperatures and analyzed their topological characteristics. Two basins were observed on the free-energy landscapes at the temperatures under the melting point. One corresponds to the open state, the other represents the closed state. In our simulation, there was no detectable intermediate state in the process of conformational transition. The topological properties of the free-energy landscape shows that the protein prefers to stay in closed state at low temperature and tends to dwell in open state at high temperature. We also studied the first passage time distribution. Both of the closed and open dwelling times exhibit the Gamma distribution. With the different temperatures, the scale parameters λ_o and λ_c of the Gamma distributions decrease or increase monotonously. The analysis of the autocorrelation coefficient shows that the conformational transition may be the multi-exponent process, which implies the complexity and hierarchical structure of the underlying energy landscape. Finally, the contact maps and



(a) ϕ value



(b) Transition structure

Figure 2.9: (a) ϕ-values of residues with high value. (b) A typical structure in transtion state. The residuces with high ϕ values in the large domain (Thr70, Thr72, Arg75), the small domain (Gly119, Asp122, Lys166) and the hinge region (Gln183, Try185, Tyr86) are marked by red color. These residues are close to the ligand pocket in the complex GlnBP-Gln[33]

ϕ value of each residue were carried out to illustrate the structural evolution and the important residues in the conformational transition. Some residues that are critical for binding and stabilizing the ligand show high ϕ values and significant transient between the open state and the closed state. The model in work of Kei-ichi Okazaki *et al*[41,42] has been further developed to deal with four states, that is, unbound open state, bound open state, unbound closed state, and bound closed state, of the protein. They found that binding ligand could induce in the shape of the energy landscape.

Although there are various benefits by using coarse-grained models, current coarse-grained methodologies still may not be as predictive as all-atom simulations. The validation of associated force-fields is still not very advanced. Intrinsically, the parameterization of coarse-grained force-fields is still difficult related to the fact that complex and diverse interactions must be described by a small number of parameters. Improving the predictivity and accuracy of coarse-grained approaches is still a stimulating challenge that needs more efforts from scientists.

2.6 summary

In chapter 2 a two-well model is employed to investigate the conformational switches of Glutimine-binding protein. Specifically, this work models the GlnBP's transition behaviors at different temperatures and also investigate the kinetical parts, such as first passage time and correlation function, to uncover the complexity and hierarchical structure of GlnBP's underlying free energy landscape. The primary tool facilitating an residue level description used in this study is molecular dynamics simulation the fundamentals of which are discussed in Chapter 1.

Chapter 3 Two-Dimensional Coordinate-Dependent Diffusion of a Conformational Switch

3.1 Abstract

Diffusion on a low-dimensional free-energy surface is a successful model for the motion dynamics of single-domain proteins. Complicating the interpretation of both simulations and experiments is the expectation that the effective diffusion coefficient will in general depend on the position along the reaction coordinate, and this dependence may vary for different coordinates. The multidimensional diffusion dynamics of protein conformational change was explored in this work and we found in general the diffusion is anisotropic and inhomogenous. The directional and positional dependence of diffusion has significant impacts on the protein conformational kinetics: the kinetic transition state with considering the coordinate-dependent diffusion is shifted away from the transition state without coordinate-dependent effect. Also, the dominant kinetic path of conformational change is shifted from the naively expected steepest descent gradient paths. Furthermore, the effective kinetic energy barrier height determining the kinetic rate of the conformational change is shifted away from the one estimated from the thermodynamic free energy barrier. The shift of the transition state in position and value will modify the ϕ value analysis for identification of hot residues and interactions responsible for conformational dynamics.

3.2 Introduction

According to the energy landscape theory, the complex chemical reactions, such as conformational dynamics, can be properly modeled as a diffusive motion of a few collective reaction coordinates that represent the progress from the initial basin to the final basin of states. In such a diffusive model, the diffusion coefficient (DC) is often coordinate-dependent[50–66]. The origin of the coordinate-dependent diffusion is the fact that the underlying energy landscape is multidimensional in nature. Since the diffusion coefficient is a quantitative measure of the ability of local escape, the projection of the high-dimensional landscape into one or a few collective coordinates will in general lead to a dependence of the diffusion coefficient[50]. The local barrier distribution changes along the progression of conformational changes or reaction coordinates. At different coordinates, there will be different local characteristics of the conformational landscape reflected through the coordinate-dependent diffusion. In the previous studies[56–59,61,62], the position-dependent diffusion coefficient along one reaction coordinate has been estimated on the energy landscape of protein folding. However, the energy landscape of the dynamic processes is complex and multidimensional. The energetic character of the local environments (local energy barriers) varies along the different coordinates. Therefore, the use of two or more collective reaction coordinates may provide us with more information when we study the complexity of the diffusive dynamics on the energy landscape. Also, studying two-dimensional diffusion can provide some insights of interaction between the diffusion along the different reaction coordinates. Such interaction may influence the kinetic aspects, such as barrier, transition state, and pathway of the conformational motion, providing much more information than the situation in one dimension.

In those approaches, umbrella samplings with a harmonic bias on one reaction coordinate were often used to collect the simulation sampling locally. However, the energy landscape of the biological dynamic processes is multidimensional and often rough. At each position on the energy landscape which is not necessarily smooth, the distribution of the energy barrier caused by the roughness of the local energy landscape can exhibit inherent inhomogeneous and anisotropic properties along the collective coordinates(or directions), implying the diffusion at this position may show different values along different directions and become inhomogeneous along the coordinates. The anisotropic property of diffusion is not just a theoretical curiosity. The experiments of diffusive dynamics of ligand binding[67], anisotropic ion mobility in ion channel[68], and the diffusion on the cell surface[69] have been performed to understand the characteristics of the diffusive dynamics along the directions of reaction coordinates.

Biological function is often linked with the associated conformational changes, such as ligand binding[32,33]. Transitions between different conformational states are the key for the biological function of many biomolecules. A natural theoretical framework to formalize these kinetic processes is provided by the energy landscape landscape theory[6–9]. According to this theory[8,9], the protein folding is driven by a decrease in free energy, which is dictated by a mechanism of entropy-enthalpy compensation. The requirement for lowering free energy while reducing conformational space determines that the energy landscape of a protein folding should be funnel-like. Analogously, the spontaneous protein-ligand association also lowers the free energy of the system composed of protein, ligand, while reducing the entropy, which is similar to the protein folding. The topology of the energy landscape determines the thermodynamic stability, the dynamic behavior of biochemical processes, and the function. The relationship between stability, dynamics and function

as well as the structure of the protein can be quantitatively understood by the detailed description of the energy landscape. The energy landscape theory has been successful in explaining qualitatively and quantitatively, the process of protein folding in which one important concept emerged as the underlying funneled landscape which controls the kinetic and thermal stability of folding. The conformational transition involves the large-amplitude conformational switches between two or more native structures[70–72], suggesting the underlying energy landscape should accordingly have the multi-basin topography.

In this work, we continue using the structure-based model with the double-well potential[34] to model the conformational dynamics of the GlnBP and investigate its multi-dimensional diffusive dynamics on the underlying energy landscape. Results show that the anisotropic and inhomogeneous diffusion and its impacts on protein conformational kinetic in terms of the dominant kinetic paths away from the steepest descent gradient one, shift of the transition state and the change of the kinetic barrier height.

3.3 Model and Methods

3.3.1 Molecular Modeling of Protein Conformational Change

Similar to the previous work, the potential energy U for a given protein conformation Γ in the simulation was given by the Equation2.1. The schematic representation of the potential is shown in Fig.3.1. We can see that this is a microscopic model which includes the two minima between any two contact residues, one for the open and the other for the closed structure. The first well corresponds to the closed structure, and the second well corresponds to the open structure. The details of how to set the parameters of the double-well potential has been described above. Briefly, in this chapter, the simulation temperature was fixed at 315K and the three parameters ϵ_1, $\Delta\epsilon = |\epsilon_1 - \epsilon_2|$,and ϵ_h were

set to 0.58 kcal/mol, 0.11 kcal/mol and 0.35 kcal/mol, respectively.

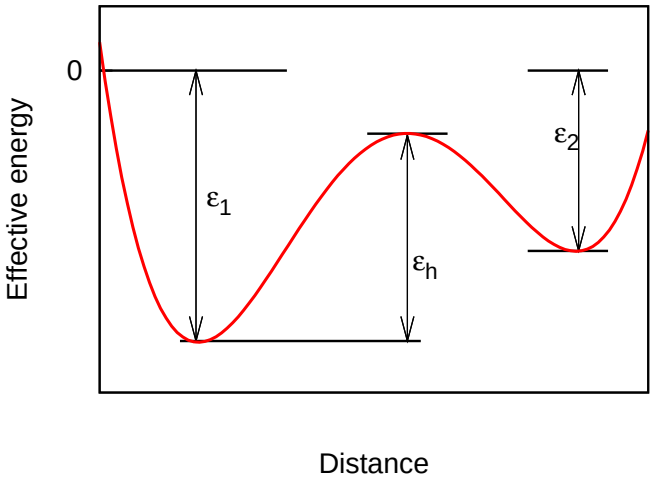


Figure 3.1: Schematic representation of the two-well potential. ϵ_1 is the depth of the first well; ϵ_2 is the depth of the second well; ϵ_h is the height of the barrier. In this paper, $|\epsilon_1| > |\epsilon_2|$.

3.3.2 Diffusion Calculation of Protein Conformational Change

For investigating the multidimensional diffusion behaviors, a single diffusion coefficient parameter is not sufficient and the diffusion should be quantified by diffusion coefficient (DC) tensor given by the formula[73,74],

$$\begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix} = Cov(Q_1, Q_2)/\tau_{corr} \tag{3.1}$$

Q_1 and Q_2 are the fractions of native contact (FNC) to the closed state and open state, respectively. The diagonal elements of DC tensor scale diffusion along Q_1 and Q_2, giving the auto-correlation of the same coordinates; off-diagonal elements quantify diffusion along

Q_1 and Q_2 in a coupled way, giving cross correlation of the different coordinates. The numerator $Cov(Q_1, Q_2)$ is the covariance matrix of the reaction coordinates, and the denominator τ_{corr} is the correlation time of the reaction coordinates Q_1 and Q_2 and can be obtained by the formula defined as,

$$C(Q_1, Q_2, k) = \frac{\sum_t (Q_{1,t} - \langle Q_1 \rangle)(Q_{2,t+k} - \langle Q_2 \rangle)}{[\sum_t (Q_{1,t} - \langle Q_1 \rangle)^2]^{1/2} [\sum_t (Q_{2,t} - \langle Q_2 \rangle)^2]^{1/2}} \tag{3.2}$$

where $Q_{1,t}(Q_{2,t})$ is the FNC relative to the closed(open)state at time t. $\langle Q_1 \rangle (\langle Q_2 \rangle)$ is the average of the fraction of native contact to the closed (open) state for a whole trajectory as the result of the MD simulations with structure-base model[22] When only one variable is considered, the covariance matrix reduces to the variance and the correlation function is substituted by the auto-correlation function.

We employed a harmonic bias potential to restrain the simulation at a specific position on the free energy landscape and calculate the local diffusion coefficients. The biasing potential was defined as,

$$V_{bias}(Q_1, Q_2) = K(Q_1 - Q_1^*)^2 + K(Q_2 - Q_2^*)^2 \tag{3.3}$$

where Q_1, Q_2 are the reaction coordinates that can be calculated by the Eq.3.4, as defined below, Q_1^* and Q_2^* are the specific values on the free energy surface, and K is the strength of the bias.

According the principles of MD simulation, one needs to calculate the first derivative of the potential in the simulations. Therefore, a good designed biasing potential which has a well-defined first derivative should be considered when including this potential as a function of reaction coordinates. To avoid the discontinuity in Q_1 and Q_2, we defined the reaction coordinates to be a continuous exponential function

$$Q_1 = \frac{1}{N_1} \sum_{contact} a e^{-(r - r_1^{nat})^2 / b} \tag{3.4}$$

$$Q_2 = \frac{1}{N_2} \sum_{contact} a e^{-(r-r_2^{nat})^2/b} \tag{3.5}$$

where r is the distance of a contact in the MD simulation, r_1^{nat} is the distance of this contact in the closed state, r_2^{nat} is the distance of this contact in the open state, N_1 is the number of all native contacts in the closed state(open state), and N_2 is the number of all native contacts in the open state. The constants $a = 1$ and $b = 10$ were chosen so that Q_1 and Q_2could decay from one to zero reasonably when the difference between r and r_1^{nat} or r_2^{nat}increases. When all the parameters were set, we can implement the Eq.3.3 and Eq.2.1 into AMBER software and run MD simulations.

3.3.3 Analytical model of stochastic diffusion dynamics of the protein conformational dynamics

A projection of the true high-dimensional dynamics of folding onto a single or more coordinates leads to a dependence of the diffusion coefficient on position along that coordinate. The diffusive dynamics of protein conformational change is intrinsically multi-dimensional involving many residues as illustrated by the molecular dynamics simulation and the analysis of the underlying energy landscape performed in previous studies and in this work[34,75]. We can model the high dimensional diffusive dynamics of the protein conformational dynamics in a coarse-grained way to capture the essence. With a few key reaction coordinates, such as native contact number Q or RMSD for structure displacements, we can study the diffusive dynamics of the protein conformational dynamics in these reduced dimensions.

From the stochastic view of diffusion behavior, we considered the Brownian dynamics with an underlying potential.

$$m\ddot{\boldsymbol{Q}} = -\gamma\dot{\boldsymbol{Q}} - \nabla U(\boldsymbol{Q}) + \xi(t) \tag{3.6}$$

where $\mathbf{Q}$ is a multi-dimensional vector (two dimensional in our study) (Q_1, Q_2), $U(\mathbf{Q})$ is two dimensional free energy, and $\xi(t)$ is a Gaussian white noise with zero average. The term on the left side in the Eq.3.6 is the inertial term which is damped on a time-scale $t \geq m/\gamma = \tau_D$. Therefore, the inertial term can be ignored when $t \gg \tau_D$[76–78]. The protein folding, binding and conformational dynamics are in this over-damped regime. Combined with the Einstein relationship of the fluctuation-dissipation theorem, the Eq.3.6 becomes Langevin equation and can be re-written as

$$\frac{d\boldsymbol{Q}}{dt} = -\frac{D(\boldsymbol{Q})}{k_B T}\nabla U(\boldsymbol{Q}) + \xi(t) \tag{3.7}$$

where $D(\mathbf{Q})$ is the position-dependent diffusion coefficient and $\xi(t)$ is a Gaussian white noise which obeys the fluctuation-dissipation relation, $< \xi(t)\xi(t') >= 2D\delta(t-t')$, where δ is a Dirac delta function. Furthermore, in order to consider the important physical effects of the systems that follow the over-damped Langevin equation, we can use the following effective Langevin equation with the special consider of position dependent diffusion coefficients:

$$\dot{\boldsymbol{Q}} = -\frac{D(\boldsymbol{Q})}{k_B T}\nabla U(\boldsymbol{Q}) + (1-\alpha)\nabla \cdot D(\boldsymbol{Q}) + \xi(t) \tag{3.8}$$

where $\xi(t)$ is a Gaussian noise and $D(\boldsymbol{Q})$ is the diffusion coefficient which depends on the variable $\boldsymbol{Q}$. The different values of α define the different stochastic calculuses. $\alpha = 0$, for instance, represents an Ito Calculus, and $\alpha = 1/2$ leads to a Stratonovich Calculus.

The above effective Langiven equation is intrinsically describe the stochastic dynamics and thus less useful in contrast to the case in the deterministic Newtonian dynamics. Instead, the probability distribution pattern can capture the essence of the stochastic dynamics. Furthermore the time evolution of the probability distribution of the observables is predictable and follows the corresponding linear Fokker-Planck equation. So the more

appropriate and full quantification of stochastic dynamics is realized by the probability distribution. Any choice of α in the addition of the term $(1-\alpha)\nabla \cdot D(\mathbf{Q})$ of the specific effective Langevin equation satisfy the same probability distribution evolution equation namely the Fokker-Planck equation.

$$\frac{\partial}{\partial t}P(\boldsymbol{Q},t) = \nabla \cdot [D(\boldsymbol{Q}) \cdot (\frac{1}{k_B T}\nabla U(\boldsymbol{Q}) + \nabla)P(\boldsymbol{Q},t)] \tag{3.9}$$

A well-known solution of the Fokker-Planck is the Boltzmann's distribution in the long-time limit:

$$P(\boldsymbol{Q},t)\big|_{t\to\infty} = const. \times exp\left(-\frac{U(\boldsymbol{Q})}{k_B T}\right) \tag{3.10}$$

The Ito convention $\alpha = 0$ was chose to discuss the path integral framework in this chapter.

The mean first passage time (MFPT) of the protein conformational transition was also considered in this chapter. The MFPT from any state to a specific final state usually obeys the following adjoint diffusion equation[79]

$$\mathbf{f} \cdot \nabla\tau + \nabla\tau \cdot \mathbf{Q} \cdot \nabla\tau = -\mathbf{1} \tag{3.11}$$

where τ is mean first passage time. The system should satisfy the absorbing boundary condition $\tau = 0$ at the given final space position and the reflecting the boundary conditions $\mathbf{n} \cdot \tau = \mathbf{0}$ on the outer boundary of system. Through solving this partial differential equation, the MFPT from open state to closed state (or from closed state to open state) can be calculated.

The detailed process of the protein conformational dynamics can also be characterized by the analysis of kinetic paths. To quantify the kinetic paths, we will use a path integral method[76,77]. In Ito convention, the path integral representation of the conditional

probability $P(Q_f, t \mid Q_i, 0)$ that can be formulated as the following form[76,77],

$$P(Q_f, t \mid Q_i, 0) = D_0 Exp(-\frac{U(Q_f) - U(Q_i)}{2k_B T}) \int_{Q_i}^{Q_f} \hat{\mathcal{D}}\boldsymbol{Q} Exp\left[-S(\boldsymbol{Q})\right], \tag{3.12}$$

where the action function $S(\boldsymbol{Q})$ is,

$$S(\boldsymbol{Q}) = \int_0^t d\tau \left(\frac{1}{4}\dot{\mathbf{Q}} \cdot D^{-1}(\boldsymbol{Q}) \cdot \dot{\boldsymbol{Q}} + V_{eff}(\boldsymbol{Q})\right) \tag{3.13}$$

In the Eq.3.13, the effective potential $V_{eff}(\boldsymbol{Q})$ can be written as

$$V_{eff}(\mathbf{Q}) = \frac{\mathbf{1}}{\mathbf{4}}\mathbf{f} \cdot \mathbf{D^{-1}}(\mathbf{Q}) \cdot \mathbf{f} + \frac{\mathbf{1}}{\mathbf{2}}\nabla \cdot \mathbf{f} \tag{3.14}$$

where $\boldsymbol{f} = -\frac{\boldsymbol{D(Q)}}{k_B T}\nabla U(\boldsymbol{Q}) + \nabla \cdot D(\boldsymbol{Q})$. The integral over $\hat{\mathcal{D}}\boldsymbol{Q}$ in the Eq.3.12 represents the sum over all possible paths with the boundary conditions Q_i at time $t = 0$ and Q_f at time t. Different pathways have different weights. The most probable pathway comes from the minimize the action $S(\boldsymbol{Q})$ and decides the optimal protein conformational transition pathways. In the Hamilton-Jacobi(HJ) description[76,77,80,81], the dominant pathways with given boundary condition can be obtained by minimizing the following effective action S_{HJ} along a one-dimensional line,

$$S_{HJ} = \int_{Q_i}^{Q_f} \boldsymbol{p} \cdot d\boldsymbol{Q} = \int_{Q_i}^{Q_f} \sqrt{E_{eff} + V_{eff}(\boldsymbol{Q})} dl \tag{3.15}$$

where $\boldsymbol{p}$ is a general momentum , $dl = \sqrt{d\boldsymbol{Q} \cdot D^{-1}(\boldsymbol{Q}) \cdot d\boldsymbol{Q}}$ is an infinitesimal length in Q space, and

$$E_{eff} = \frac{1}{4}\dot{\boldsymbol{Q}} \cdot D^{-1}(\boldsymbol{Q}) \cdot \dot{\boldsymbol{Q}} - V_{eff}(\boldsymbol{Q}) \tag{3.16}$$

The jump time of the protein conformational transition can be determined by the following relationship

$$t = \int_{Q_i}^{Q_f} \frac{dl}{\sqrt{4(E_{eff} + V_{eff}(\boldsymbol{Q}))}} \tag{3.17}$$

The effective energy V_{eff} is defined in the Eq.3.15. The jump time gives a quantitative measure of the prefactor of the kinetic rate[81].

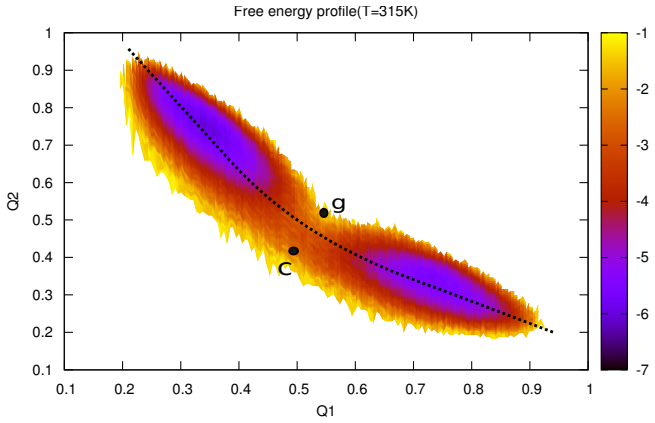
3.4 Results and Discussions

3.4.1 Free Energy Landscape of Conformation Dynamics

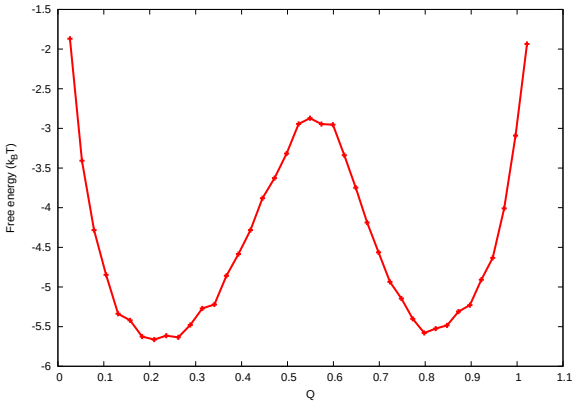
Fig.3.2(a) shows the two-dimensional free energy profiles which is considered as a function of the native contacts relative to the close state, Q_1 and the native contacts relative to the open state, Q_2. It is obtained by $F = -log(P)$, where P is the statistical population obtained from the histograms of all the trajectories in the MD simulations with the confinement parameter $K = 0$. There are two local minima that correspond to the open state (upper left basin)and the closed state(lower right basin) of the protein, respectively. This result is consistent with the second chapter of this thesis. If projecting the 2D free energy surface onto the one-dimensional reaction coordinate along a curve represented by the black dash line in Fig.3.2(a), we can obtain the one-dimensional free energy profile, as shown in Fig.3.2(b), which clearly shows the two minima separated by a solitary free-energy barrier. The minimum with low Q value (< 0.55)corresponds to the open state, and the other minimum with high Q values (> 0.55)represents the closed state. This energy landscape was obtained by $K = 0$. Yet, the value of K should not be equal to zero when studying the diffusive behaviors along the one- and two-dimensional free energy profile.

3.4.2 Diffusion with different values of parameter K

The value of K should be large enough so that the sampling can be constrained on a specific regime on the free energy surface. Meanwhile, the value of K must be in a range in which the quasi-harmonic approximation is valid and diffusion coefficient is independent on K. From the view of the energy landscape, if the biasing strength is small, which means



(a) two dimensional free energy profile



(b) One dimensional free energy profile

Figure 3.2: (a)Two-dimensional free-energy profile of the conformational transition of GlnBP. The two points were marked as c and g, and their diffusion coefficients will be compared in the two-dimensional situation. They locate on the edges of the free energy profile, and their positions are around (Q_1, Q_2)=(0.5,0.4) and (0.55, 0.5), respectively. The black-dash line presents an artificial pathway that crosses the two basins and barrier. (b) One-dimensional free-energy profile along the black-dash line in (a).

the sampling is performed on the large region of the free energy landscape, the diffusion coefficients calculated from such simulations can not represent the diffusive behavior on that constrained point. To determine the value of K, we scanned the range of K from 1 to 120 to test the several points (Q_1, Q_2) which locate in the open state and closed state on the 2D energy surface.

Fig.3.3(a) displays the relationship between the diffusion coefficient and the biasing strength K of the points (0.25, 0.75), (0.30, 0.75), and (0.35, 0.75) which locate in the open state on the free energy surface. A number of features are worth noticing. First, the diffusion coefficient along the Q_2 (top three lines) is larger than the diffusion coefficient along the Q_1 (bottom three lines) for $K > 10$. Second, as Q_1 increases, the values of DC along the Q_1 increase for each K. Interestingly, one can find the opposite behavior of the DC along the Q_2 for $K > 10$. Third, the values of DC are very similar in the region $K > 10$. In other words, the DC is independent of K when $K > 10$. In the region $K < 10$, however, the values of DC increase rapidly as K increases. When K is small, the simulation could not be fixed on the narrow regime on the free energy landscape. The correlation function decays slowly to zero. In other words, the scale time of the correlation function is long when the biasing strength is small. As the biasing strength enhanced, the simulation is constrained to the narrower region, and the scale time of the correlation function decreases and finally gets close to the characteristic relaxation time of that point, which is independent of K. Therefore, the value of K should be larger than 10.

On the other side, overlarge biasing strength K may freeze the system, especially for the closed state which is less flexible and more compact than the open state. The points (0.75, 0.27),(0.80, 0.27), and (0.85, 0.27) are in the closed state, and the relationship between the diffusion coefficients and the K of these points are shown in Fig.3.3(b).

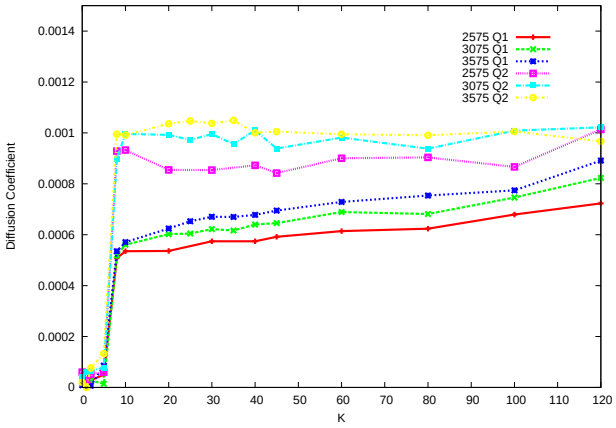
Many characteristics and trends of these data points can be observed. Interestingly, the DC along Q_1 are larger than the DC along Q_2 for each K. These observations are in contrast to the points in the open state. Besides, the values of DC along Q_2 are nearly the same with the different Ks. In other words, the DC along Q_2 is independent of K. In addition, as Q_1 increases, the values of DC along Q_2 increase for each K. In contrast, the values of DC along Q_1 decrease as Q_1 increases for each K. Importantly, the DC along Q_1 is independent of K. In the region of $K < 60$, the values of DC are very similar, In the region of $K > 60$, however, the values of DC start to significantly increase when K increases, suggesting that the value of K being more than 60 is improper. To make sure K is strong enough to constrain the simulations, we chose $K = 60$ in our model. The simulations below on the different points of the energy landscape also justify that $K = 60$ is proper for the simulations. Therefore, the following results in this chapter are all with $K = 60$, unless specified otherwise.

3.4.3 Diffusion Coefficient in Effective One Dimension of Conformational Dynamics

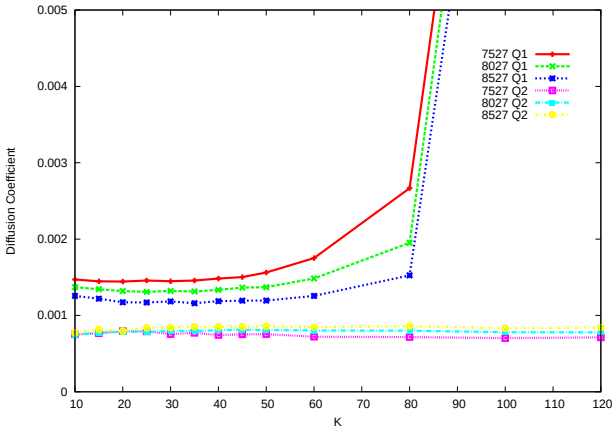
First we investigated the diffusive behavior along the Q_1 and Q_2, respectively, as shown in Fig.3.2(b). To describe the process of dynamical conformational change of proteins along one dimension, the reaction coordinate or order parameter is not only very useful for the experiments[82,83] but also for the theoretical studies[50–53,56,57]. Recently, some studies[82] have shown that the diffusion coefficient in the protein folding process varied strongly along the reaction coordinate when the folding dynamics was projected onto a 1D coordinate and described as a diffusive process. This property was not only in the protein folding process. When describing the diffusive behavior of the conformational

switch along the 1D reaction coordinate, one can find that the diffusion coefficient is significantly position-dependent as well, as shown in the Fig. 3.4.

Furthermore, the diffusion coefficient shows very different behaviors along the different directions (Q_1 or Q_2) on the main conformational change pathway. In Fig.3.4, The red-solid line represents the diffusion coefficient along Q_1, the fraction of native contact (FNC) to the closed state, and the green-dash line exhibits the diffusion coefficient along Q_2, the FNC to the open state. First we discuss the red-solid line. When Q increases from 0 to 0.55, which corresponds to the open basin, the diffusion coefficient along Q_1 first increases, then decreases. One can find a maximum when Q is around 0.3, corresponding to the bottom of the open-state basin on the free energy profile. In the closed-state basin, where Q extends from 0.55 to 1, as shown in Fig.3.2(b), there is another maximum of the diffusion coefficient along Q_1 when Q is around 0.8, corresponding to the bottom of the closed-state basin. The maximum in the closed-state basin is larger than the one in the open-state basin. Interestingly, the values of DC along Q_2 are larger than the ones of DC along Q_1 in the open-state basin, and the opposite phenomenon can be observed in the closed-state basin, as shown in Fig.3.4. However, these two types of DC both have a minimum when Q is around 0.55, which corresponds to the barrier of the free energy landscape. On the main thermodynamic gradient pathway of the conformational change, DC shows the different behaviors in the different directions (Q_1 or Q_2). In the Q_1 direction, there is a faster diffusion from the closed state towards the open state. However, on the Q_2 direction, the diffusion from the closed state to the open state is slower, implying the different local escape time or diffusion. The origin of the different behaviors is because of projecting the multidimensional energy profile into a few dimensions, or coordinates.



(a)



(b)

Figure 3.3: (a) The diffusion coefficient with different K values for the points in the open-state basin. (b) The diffusion coefficient with different K values for the points in the closed-state basin. The unit of the diffusion coefficient is $\mathrm{cm}^2 s^{-1}$.

3.4.4 Diffusion Coefficient in Two Dimensions of Conformational Dynamics

Fig.3.5 shows the 2D position-dependent DCs, which can be calculated by the Eq.3.1. The diagonal elements of diffusive tensor D_{11} and D_{22} are shown in Fig.3.5(a) and Fig.3.5(d), respectively. To describe the properties of the diffusion tensor on the energy landscape, we separate the energy surface into three parts, that is, the open-state region, the barrier region, and the closed-state region. They are marked as "open","barrier", and "close", respectively, in Fig.3.5. Along the $edge1$ around the open-state basin, as shown in Fig.3.5(a), the value of D_{11} increases slightly from the open-state region to the barrier region; after entering the closed-state basin, the value of D_{11} climbs up sharply and then descends, forming a maximum value around the closed region. Along the $edge2$, the values of D_{11} show the similar tendency. Interestingly, although these two edges show the same tendency, the value of D_{11} at the point c ($\sim 4.3\times 10^{-3}$) is larger than that at the point g ($\sim 2.9 \times 10^{-3}$). Given the free energy barrier, the rates of diffusion along the pathways that are close to the point c are larger than the ones that are near to the point g, implying that the dynamic process of the key ligand binding of the GlnBP prefers to go through the pathways near to the point c on the free energy landscape. From the 1D diffusive profile, it is difficult to obtain this property, implying the 2D diffusive profile carries more detailed information regarding the diffusive dynamic of the conformational switches on the energy landscape.

The shape of the D_{22} is showed in Fig.3.5(d). On the open-state basin of the free energy surface, the values of the D_{22} gradually increases and then declines to the barrier domain. The maximal value of the D_{22} locates at the bottom region of the open-state

basin. When turning into the closed-state basin, the values of the D_{22} bounce from the minimum value in the barrier region, and slowly ascends to bottom region of the closed-state basin. Although the overall change of the D_{22} on the free energy landscape follows the above tendency, on the different points on the free energy landscape, the small but important differences can be observed. The value of D_{22} at point c ($\sim 4.6\times 10^{-3}$) is larger than one of the point g($\sim 2.0\times 10^{-3}$). This phenomenon is similar to the D_{11}. The DC along the Q_1 direction, (D_{11}), reveals significant differences compared to the DC along the Q_2 direction, (D_{22}), on the free energy landscape. In the open-state basin, the surface of the D_{11} is flat, as shown in the Fig.3.5(a), which implies the values of the DC along the Q_1 direction are very similar in this basin, yet the surface of the D_{22} in the open-state basin is bending, as shown in the Fig.3.5(d), suggesting the obvious variation of the D_{22} in this area. In contrast, in the closed-state basin,the surface of the D_{22} is flat, while the surface of the D_{11} is bending. Moreover, the highest value of the D_{11} for the overall free energy landscape can be found at the bottom area of the closed-state basin, but for the D_{22}, the maximal value locates in the open-state basin. Although the D_{11} and D_{22} exhibit the opposite behaviors in the two basins, they show the similar properties around the barrier area. Both of them have minimal values locating in the barrier area. Also, for both D_{11} and D_{22}, the value on the point c is larger than the value of the point g on the free energy landscape. The configurational diffusion coefficient describes the ruggedness of the underlying free energy landscape and is also influenced by the shape of the entire free energy landscape. In the realistic dynamics, the underlying free energy landscape is not smooth, and has local distribution of barriers. Therefore, the diffusion tensor is highly anisotropic. In other words, many possible escape-time scales may coexist. For the realistic dynamics the diffusion anisotropic can be particularly pronounced along the

lower-edge direction on the free energy landscape. When projecting the open-state basin into Q_1 and Q_2 coordinates in Fig.3.4, one may observe that the length of the open-state basin in the Q_2 coordinate is longer compared to the Q_1 coordinate, implying the protein has more mobility along the Q_2 direction than that along the Q_1 direction in the open-state basin. The strong anisotropy in this region will dynamically lead the DC along Q_2 to be dominant. Therefore the DC value along Q_2, that is, D_{22}, is larger than the D_{11}, the DC value along the Q_1 in the open-state basin.

Fig.3.5(b)and (c) show the off-diagonal element of the DC tensor D_{12} and D_{21}, respectively. From these two plots, one can find that the values of the D_{12} and D_{21} are almost the same. Physically, the off-diagonal elements of the diffusion tensor arise from the interaction of the gradient-induced fluxes. Take a orthogonal field configuration as an example. A density flux in the Q_2 direction will cause the flux in the Q_2 direction. This flux will interact with field and induce the flux in the other orthogonal direction, for instance, Q_1 direction. Therefore, a gradient in the Q_2 direction can cause a flux in the Q_2 direction, described by D_{22}, and also induce a flux in the Q_1 direction, described by D_{12}, whose negative values imply that the flux caused by the gradient in the Q_2 direction may have opposite direction of the Q_1 coordinate. In the barrier region in the Fig.3.5(b), the maximal value of the D_{12} can be found, implying the interaction between the fluxes induced by the gradients in the Q_1 and Q_2 directions respectively are strong. This strong interaction may have effective influence on the thermodynamic free energy and shift the position and height of the barrier so that the actual kinetic paths of the conformational switch may not go through the thermodynamic transition state but pass through the effective transition state determined by both kinetic diffusion and thermodynamics. This may be verified by the properties of the D_{11} and D_{22} around the barrier region. In the

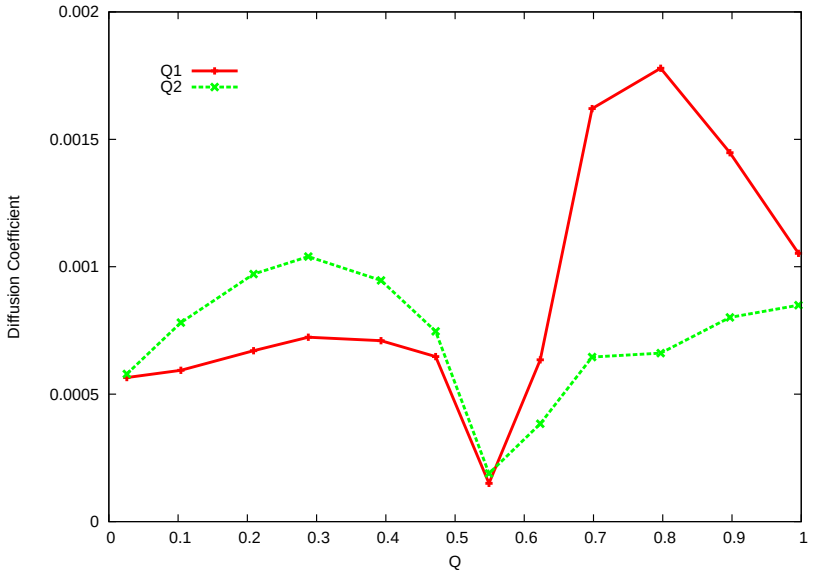


Figure 3.4: Diffusion coefficent along the one-dimensional free energy profile in Fig.3.2(b). Red line represents the diffusion coefficient along the Q_1 direction, and green line represents the diffusion coefficient along the Q_2 direction. The unit of diffusion coefficient is $\mathrm{cm}^2 s^{-1}$.

Fig.3.5(a), the value of the D_{11} at point c is larger than the one at point g and the same behavior can be found for the D_{22} in the Fig.3.5(d), implying the conformation transition may not pass the thermodynamic transition state on the energy landscape, but pass through the pathway which is close to the point c around the barrier region.

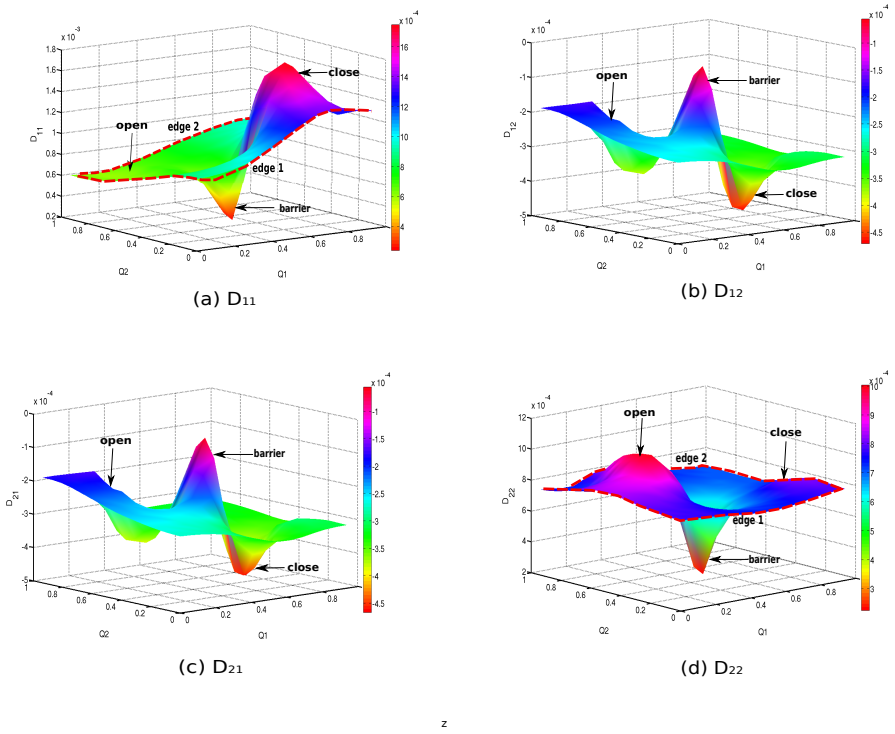


Figure 3.5: Two-dimensional diffusion coefficient (a) D_{11}, (b)D_{12}, (c)D_{21}, and (d)D_{22} on the free energy landscape. Three regions were labeled as "open", "close", and "barrier", respectively. In the subfigures (a) and (d), two edges were represented by the red dash lines. The edge on which the point c of the Fig.3.2(a) locates was marked as "edge 1", and the edge on which the point g of the Fig.3.2(a) locates was labeled as "edge 2". The unit of the diffusion coefficient is $\mathrm{cm}^2 s^{-1}$.

3.4.5 The Influence of Inhomogeneous and Anisotropic Diffusion on the Kinetic Paths, Rates, and Barrier

Fig.3.6 shows the dominant conformational transition pathways on the free energy landscape. It is well known that the most probable pathway should follow the steepest descent gradient of the underlying free energy landscape going through the saddle point in a constant diffusion coefficient, which is illustrated by the red line in the Fig.3.6. In the calculation of most probable pathway, the second-order error decays rapidly with exponential scale, so in terms of first-order approximation, the pathway obtained here is

accurate. After taking into the account of the effects of spatial-dependent diffusion, the dominant kinetic pathway, represented by the black line in the Fig.3.6, shifts from the red line, implying the spatial diffusion shifts the position of the effective kinetic free energy barrier away from the thermodynamic saddle point or transition state. In other words, the dominant conformational transition pathways with the coordinate-dependent diffusion coefficient deviating from the ones of the naively expected steepest descent gradient paths and not going through the saddle point.

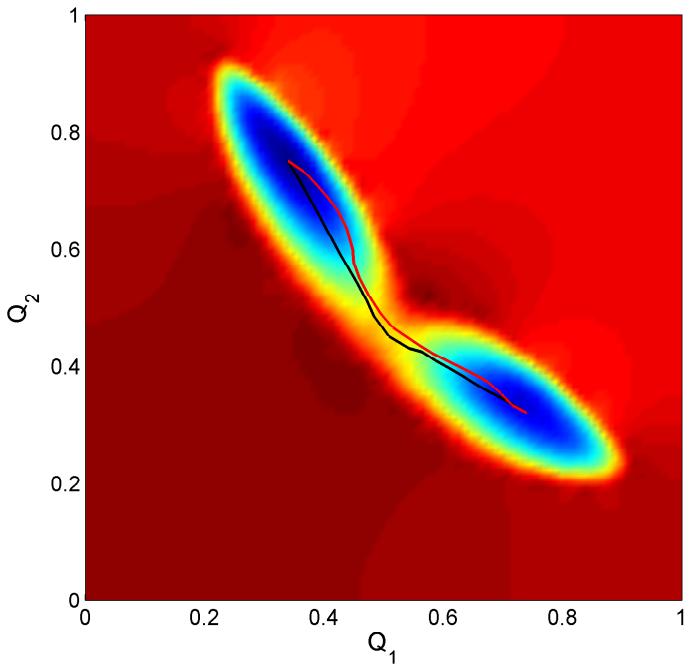


Figure 3.6: Conformational transition pathways on the free energy profile. Red line is the transition pathway without spatial-dependent diffusion effect; The pathway with black line considers the spatial-dependent diffusion effect.

In order to figure out how the spatial diffusion influences the effective free energy barrier height, we should work out the conformational transition time τ by solving the

adjoint Fokker-Planck diffusion Eq.3.14. When the diffusion coefficient is constant, the rate is proportional to the exponential of the free energy barrier height $\tau_c = \tau_0 exp[\frac{\delta U}{k_B T}]$; when the effect of the spatial diffusion is considered, the rate should be calculated by the effective barrier height $\tau_{eff} = \tau_0 exp[\frac{\delta U_{eff}}{k_B T}]$, where τ_0 is a constant which is not sensitive to the coordinate. We assume the similar value of τ_0 for τ_c and τ_{eff} for simplicity here. Combining these two rate equations and eliminating the τ_0, one can obtain $\delta U_{eff} = k_B T ln(\tau_{eff}/\tau_c) + \delta U$. According to MD simulations, the thermodynamic free energy barrier from closed state to open state is U $\sim$ 2.6kT, and from open state to closed state is U $\sim$ 3.0. The transition time τ_c of the conformation change from close state to open state is 756ps (from open to close is 624ps) with the constant diffusion coefficient around 1.8$\times$ 10^3 ps which was calculated at the bottom of the thermodynamic close state basin. When the spatial and anisotropic diffusion is considered, the transition time τ_{eff} from close state to open is 1562ps(from open to close is 1136ps). Therefore, the effective kinetic barrier δU_{eff} from close to open conformation transition due the presence of spatial and anisotropic diffusion is around 3.33kT (from open to close is 3.6 kT), which is larger than the thermodynamics barrier height around 2.6(from open to close is 3.0)kT . The kinetic barrier from close conformation to open conformation transition is shifted from the original thermodynamic barrier by by 0.73kT, and the barrier shift from the open state to close state is 0.6 kT. Therefore, the transition time is slower when the spatial and anisotropy in diffusion are taken into consideration. As we can see both the barrier position and height are shifted due to the spatial and anisotropic diffusion. The effective barrier caused by the spatial and directional diffusion is significant.

The ϕ value of a residue is defined as the ratio of the change in kinetic speed of conformational change and thermodynamic stability upon mutations and the ϕ value

analysis has been widely used in the protein folding and conformational dynamics to identify the important residues and interactions responsible for the underlying mechanisms and function[34,47,72,75,84]. Quantitatively, the ϕ value of a residue is defined as the ratio of the change in difference between the free energy of transition state and initial reactant state versus the change in difference between the free energy of final product state and initial reactant state upon the mutation of this particular residue. The kinetic barrier under constant diffusion is the same as the thermodynamic barrier. However, when the diffusion becomes spatial dependent and anisotropic, the kinetic barrier starts to deviate from the thermodynamic barrier both in position and in value as shown above. Therefore the corresponding ϕ value analysis should be modified and carried out using the kinetic barrier rather than the thermodynamic barrier.

3.5 Conclusion

We developed a two-dimensional diffusion model to estimate the diffusion coefficients of the GlnBP on the energy landscape. In the previous studies[50–53,56,57,60,61], the position-dependent diffusion folding dynamics along one reaction coordinate has been studied. However, the energy landscape of the dynamic processes is complex and multidimensional. The energetic character of the local environments (local energy barriers) varies along the different coordinates. Therefore, two or more collective reaction coordinates are often needed to study the complexity of the diffusive dynamics on the energy landscape. In this work, two-dimensional position-dependent diffusion coefficient on the free energy landscape of the conformational switch of the GlnBP was investigated by a structure-based two-well model.

After calculating the 2D diffusion coefficient tensor, we found that the local DC on

the energy landscape is anisotropic, which is very hard to be found in a one dimensional diffusion study.The diagonal elements D_{11} and D_{22} of the DC tensor exhibited almost opposite behaviors. In the open-state basin, the surface of D_{11} was flat, but the surface of D_{22} was nonflat. One can find the maximal value of the D_{22} in the open-state basin. In comparison, in the closed-state basin, the surface of D_{11} was non-flat, but the surface of D_{11} was flat, and the highest value of the D_{11} is located in the closed-state basin. Although with opposite trends of the DC in two different basins, similar behaviors at the barrier region of the energy landscape were observed. Around the barrier region, both the D_{11} and D_{22} have the lowest values. Also, interestingly, the value of the D_{11} at point c on the free energy landscape is higher than that at point g. One also can observe the similar shape of the D_{22}. For the off-diagonal elements D_{12} and D_{21} of the DC tensor, their maximal values are located in the barrier region, suggesting the correlation between the two direction-dependent diffusion is strong. The anisotropy and inhomogeneous coordinate dependent diffusion can shift the thermodynamic pathway of the conformational transition of the protein away from the naively expected steepest descent gradient path on the free energy landscape. Furthermore, the dominant kinetic paths do not necessarily go through the transition state. Both the position and the value of the barrier height are shifted by the inhomogeneous and anisotropic diffusion. The inhomogeneous and anisotropic diffusion will therefore modify the phi value analysis for identification of hot residues for conformational change dynamics[34,47,72,84].

The diffusion analysis has been carried out for folding with only one stable state (funnel) and with one reaction coordinate. Qualitative and semi-quantitative discussions on the effects of spatial diffusion on kinetics have been carried out[57,60,61]. Our multi-dimensional diffusion studies here enrich the qualitative and quantitative pictures

obtained in previous studies in several ways. First, we now can consider diffusions not only in one dimension but also in higher dimensions. Second, we can now explore the effects of anisotropy not only from the differences of the diagonal elements of the diffusion but also from the couplings among diagonal elements and off-diagonal elements of the diffusion. Third, we can quantify the effects of the multidimensional anisotropic spatial inhomogeneous diffusion on the dominant kinetic paths explicitly. This can not be realized in one dimension. Fourth, we can also quantify the effects of the multi-dimensional anisotropic spatial inhomogeneous diffusion on the kinetics: the position as well as the height shift of the associated kinetic barrier. Fifth, even for one stable state as in the case of the protein folding funnel, we can now begin to explore the effects of anisotropy and inhomogeneity of multidimensional diffusion on the dynamics.

Reduced representation models, may neglect some components that influences the accuracy of simulation of the proteins. Nevertheless, previous studies[51–54] have shown that coarse-grained model have significant impact on the protein model study. Compare to the limitation of the computational resource, the error of using the simplified model is sustainable. The effects of one dimensional spatial dependent diffusion on the conformational dynamics of protein folding has been investigated experimentally[82]. Our current study provides a basis for further experimental studies on the two dimensional diffusive dynamics on the molecular conformational changes. In the future work, we will consider some other systems to study the multi-dimensional diffusive dynamics on the biomolecular energy landscape.

3.6 Summary

In this chapter, we further use the two-well model that was applied in the Chapter 2 to investigate the two-dimensional diffusive characteristics of GlnBP on its free energy landscape. The behaviors of anisotropy and inhomogeneous coordinate-dependent diffusion are found and such diffusive properties can shift the thermodynamic pathway of the conformational transition of the protein away from the naively expected steepest descent gradient path on the free energy landscape. Furthermore, the dominant kinetic paths do not necessarily go through the transition state. Both the position and the value of the barrier height are shifted by the inhomogeneous and anisotropic diffusion.

Chapter 4 Exploring Trp-cage Folding with Two-dimensional Infrared Spectroscopy

4.1 Abstract

Probing the underlying free energy landscape, pathways, and mechanism is the key for understanding protein folding in theory and experiment. Recently time-resolved two-dimensional infrared (2DIR) with femto-second laser pulses, has emerged as a promising tool to investigate the dynamical process of protein folding on fast timescales. In this work, we calculated 2DIR spectroscopy of Trpcage structures along the free energy profile. Non-chiral and chiral 2DIR signals illustrate the variation of the spectrographic patterns when the protein evolves on the underlying free energy landscape. Isotope-labeling is used to reveal the residue-specific information. We showed that the high resolution structural sensitivity of 2DIR can differentiate the ensemble evolution of the protein, and thus provides a microscopic picture of the folding process. This work provides a protocol for applying multidimensional IR spectroscopy to study protein folding

4.2 Introduction

Protein folding is one of the most fundamental problems in modern molecular biology. According to the current view, protein folding is envisioned to proceed along a moderately rough funnel-like energy landscape[6,38]. Many local minima on the energy landscape form due to the competition between the "downhill" pathway towards the native state and the accumulation of misfolded and/or partially folded states. The important issues of folding

pathways have been explored from both theoretical and experimental perspectives[85–87]. Uncovering the detailed folding mechanism requires methods that can monitor the structures at high temporal and spatial resolution. Many conventional spectroscopic methods can only provide averaged information due to the lack of high temporal resolution. For example, atomic resolution structures can be directly determined by NMR spectroscopy, but only on around microsecond timescales. Nanosecond measurements in NMR are based on the frequency dependence of relaxation rates and are therefore indirect.

Two-dimensional infrared (2DIR) spectroscopy[23–26,88–99] is a novel approach we can apply to study transient molecular structure and dynamics. As a vibrational spectroscopy, it can investigate the vibrations of chemical bonds and how the vibrations interact with one another. 2DIR spectroscopy spreads a vibrational spectrum over two frequency axes, allowing to reveal structural and kinetic correlations[91,92]. Compare to linear, absorption spectroscopy, the obvious advantages of 2DIR is the feature of correlating excitation and emission frequencies to allow for a separation of homogenous and inhomogeneous line shape components, and to give rise to structurally sensitive cross-peaks. Cross peaks in the spectrum encode the couplings and orientation between vibrations. Modeling this spectrum reveals a structure in terms of connectivity, distance or orientation between molecules. Meanwhile, since the measurement is made with a picosecond or faster laser, which captures information on molecular structure in solution on a time scale fast compared to most dynamics, it offers an effective avenue to directly reveal protein folding dynamics which accompany the conformational changes in the pico- to nanosecond time scale. 2DIR spectroscopy achieves its time resolution through the use of femtosecond pulse sequences that interact with the protein and generate coherent nonlinear signals, which are generated by three laser pulses with wavevectors $\mathbf{k}_1$, $\mathbf{k}_2$, $\mathbf{k}_3$. The coherent

signal field is emitted along the phase-matching directions $\mathbf{k}_4 = \pm\mathbf{k}_1 \pm \mathbf{k}_2 \pm \mathbf{k}_3$ and is detected by interference with a fourth "local-oscillator" pulse with wavevector $\mathbf{k}_4$. The pulses interact with the protein and produce a coherent nonlinear signal which depends on three time delays $S(t_3, t_2, t_1)$. A two-dimensional Fourier transform generates a 2DIR spectrum $S(\Omega_3, t_2, \Omega_1)$, where Ω_3 and Ω_1 are the frequency conjugates to t_3 and t_1, respectively. By choosing different polarization configurations, one can obtain non-chiral (i.e. *xxxx*) and chirality-induced (CI) (i.e. *xxxy*) signals[26], where *ijkl* represents the polarization configuration of the four pulses in chronological order. The corresponding one-dimensional analogue of two-dimensional chirality-induced signals is circular dichroism spectra[100]. Compare to the CD spectra, the CI 2D signals can exhibit the correlations between different parts of a protein through enhancing cross-peak contributions and delegate them to structural characters. The cross-peaks are very sensitive to the secondary structure variation, and the chiral configuration between different chromophores can be determined from the signs of the corresponding cross-peaks. Although CI 2DIR spectroscopy has not yet been implemented experimentally as the signal fields are much weaker than nonchiral 2DIR signals, the cross peaks in the CI 2D signals are explicitly coordinate-dependent and are therefore particularly sensitive to structural changes.

The amide I band, primarily associated with the peptide bond carbonyl stretch, is the most widely studied by the 2DIR technique because it is sensitive to the hydrogen bonding, dipole-dipole interactions, and geometry of the peptide backbone, thus providing a good indicator of secondary structure and dynamics. The cross-peaks (off-diagonal features) of the amide bands carry signatures of intra- and intermolecular couplings. Site-specific isotope-labeling, where the frequency of the amide I transition is modified by substituting $^{12}C{=}^{16}O$ by $^{13}C{=}^{16}O$ or $^{13}C{=}^{18}O$ can be used to isolate structurally important

residues, providing site specific information on peptide folding[98]. 2DIR has been successfully applied to study many chemical and biological processes such as hydrogen bonding dynamics[101], fast chemical exchange in molecular complexes[102], and protein folding[90].

In this work, we calculated the of 2DIR spectroscopy signatures of the ultrafast folding process of the 20-residue Trp-cage peptide (Asn1-Leu2-Tyr3-Ile4-Gln5-Trp6- Leu7-Lys8-Asp9-Gly10-Gly11 -Pro12-Ser13-Ser14-Gly15-Arg16-Pro17-Pro18-Pro18-Ser20), which is one of the fastest folding mini-proteins. Although the Trp-cage is small and relatively simple, the mechanism of its folding remains elusive. Recent UV-resonance raman experiments[103] show that the Trp-cage is not a simple two-state miniprotein. Additionally, the folding time determined by tryptophan fluorescence and recent 2D ^{1}H NMR spectra experiment suggests downhill folding mechanism[104]. On the other hand, some studies[105,106] have suggested that it follows a simple two-state folding mechanism. It is very interesting that even for such a small system we still have conflicting views of its folding mechanism. we generated its folding free energy landscape by the molecular dynamics (MD) simulations. The observations of the conformational evolution on the folding pathway through 2DIR spectroscopy provide a detailed picture of the structure and dynamics of the peptide along the pathway and the folding mechanism.

4.3 Methods

4.3.1 Molecular dynamics (MD) simulations

We carried out all the MD simulations and part of the analysis using AMBER 10 software package[14] with the AMBER ff99SB protein force field[107]. A constant temperature of 315 K was maintained in the MD simulations. An Generalized Born implicit solvation model[108] with a collision frequency of 1 ps^{-1} was used to simulate the solvent environment.

The SHAKE algorithm[109] was used to constrain covalent bonds involving hydrogen atoms. The time step was set to 2 fs. 50 trajectories were simulated for 200 ns each. The initial structure of each trajectory was given by an extended conformation and the different atom velocities from a Gaussian distribution were assigned to the different trajectories to start the simulations. The total 10 μs simulations provided enough data to construct the free energy landscape (FEL). Five folding states were chosen along the dominant folding pathway from the unfolded state to the folded state on the FEL. For each folding state, 200 snapshots around that location were harvested to calculate the 2DIR signals.

4.3.2 Calculation of 2DIR spectra

The effective vibrational Hamiltonian of the system was needed to calculated first, and the details of the calculations can be found somewhere else[25]. After constructing the vibrational Hamiltonian, absorptive 2DIR spectra were simulated for non-chiral (*xxxx*) and CI (*xxxy*) polarization configurations. The 2DIR spectra were computed using the quasiparticle approach based on the nonlinear exciton equations[110–113], as implemented in SPECTRON[114]. Absorptive signals were defined as the addition of the rephasing ($\mathbf{k}_I = -\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3$) and non-rephasing ($\mathbf{k}_{II} = \mathbf{k}_1 - \mathbf{k}_2 + \mathbf{k}_3$) spectra. Homogeneous broadening was set to 5.5 cm^{-1} for all transitions. All signals were calculated in the inhomogeneous limit by averaging over 200 configurations extracted from each location on the FEL. For each snapshot, the simulated peptide was explicitly resolvated with TIP3P water model[115] and equilibrated for 10 ps. The explicitly solvated and equilibrated structure was used to compute all the signals. We assumed short impulsive pulses and t_2=0.

4.4 Results and Discussion

4.4.1 Folding Mechanism

The free energy landscape of the Trp-cage folding is shown in the Fig.4.1. One axis is root mean square deviation (RMSD) and the other is radius of gyration (Rg). According to the statistic physics, the free energy is determined by calculating F =-log(P), where P is the population obtained from the 10 μs MD simulated data, including 50 MD trajectories of 200 ns each. The FEL reveals several interesting features of the folding mechanism. First, it is smooth and there are no apparent thermodynamic barriers or intermediate states, implying that it may follow a downhill folding mechanism. In addition, there is an obvious dominant pathway connecting the unfolded to the folded state (black curve in Fig. 4.1). We have calculated the 2DIR spectra at five locations along the folding pathway, L1, L25, L50, L75, and L100. Before L50, the peptide structure does not change significantly and retains the extended linear or coil structure. After the peptide passes L50, the folding process seems to accelerate and the peptide rapidly reaches the folded state.

We next calculated the average number of hydrogen bonds at each location to illustrate the folding process (Table 4.1). The number of hydrogen bonds in the entire peptide and particularly the α-helix region (residues 1-9) abruptly increases between L25 and L50, The α-helix hydrogen bonds increment is largely caused by the increasing number of inter-α-helix rather than intra-α-helix hydrogen bonds. At the same time, the number of hydrogen bonds in the coil region shows a large increase while the number of hydrogen bonds in the "other" region only increases slightly. These observations suggest that the α-helix tends to form hydrogen bonds with the coil region to form a "two-strand"

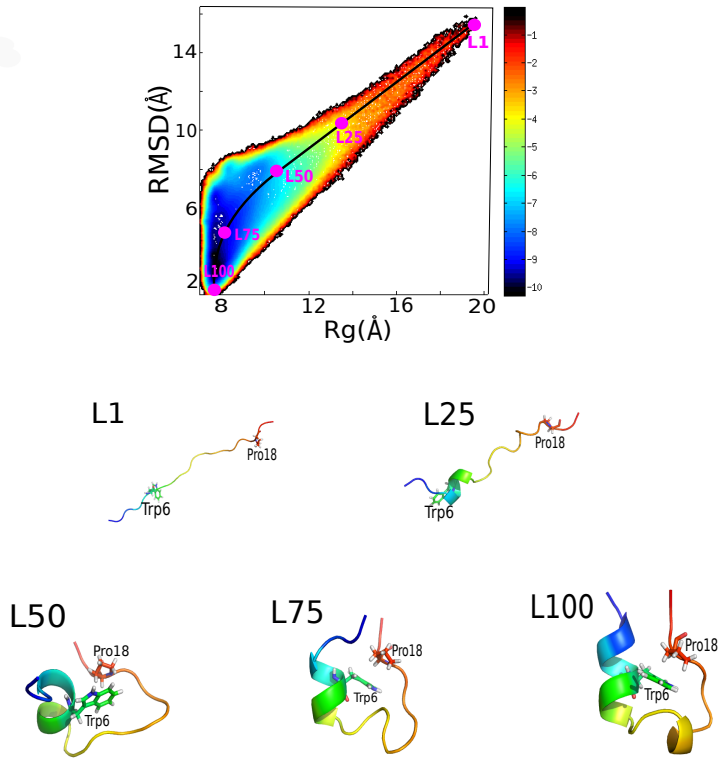


Figure 4.1: Free energy profile of Trp-cage folding vs the RMSD and the radius of gyration (Rg). (top). Five structures along the folding pathway are labeled L1, L25, L50, L75, and L100. The corresponding structures are shown (bottom). Trp6 and Pro18 are shown by stick representation.

Location	Whole peptide	Whole α-Helix	Intra α-Helix	Inter α-Helix	Coil
1	0.89	0.62	0.58	0.04	0.23
25	2.27	1.65	1.06	0.58	0.57
50	6.12	4.84	1.95	2.88	3.49
75	7.76	5.22	2.56	2.66	3.70
100	8.33	5.48	3.03	2.45	2.82

Table 4.1: The average number of hydrogen bonds of the whole peptide and its parts for L1, L25, L50, L75, and L100. The hydrogen bonds were calculated using VMD[116] with an acceptor-donor distance cutoff of 3.5 $\AA$ and an acceptor-donor-hydrogen angle cutoff of 30 degrees. Residues 1-9 and 15-19 were defined as the α-helix and coil region, respectively, to calculate the corresponding hydrogen bonds. Intra α-helix includes hydrogen bonds among the α-helix region and inter α-helix includes hydrogen bonds between the α-helix region and all other residues.

structure between L25 and L50. The formation of this "two-strand" structure can dramatically reduce the conformational searching in the huge configuration space and help the peptide fold into its correct native structure. After L50, the number of intra-α-helix hydrogen bonds continues to increase smoothly, implying the growth of the α-helix. Interestingly, the number of inter-α-helix hydrogen bonds decreases when moving from L50 to L100, indicating that residues 1-9 tend to form hydrogen bonds among themselves, out-competing the inter-α-helix hydrogen bonds as the α-helix gradually grows.

We also present the couplings between the ranges of α-helix region and the coil region (residues 15-19) in Fig.4.2 to demonstrate the structural changes that occur during folding. There is almost no coupling between these two groups at L1 and L25, yet the two groups become strongly coupled at L50. This is consistent with the analysis of hydrogen bond above which suggests there are few hydrogen bonds between the "two strands" (N-terminal α-helix region and C-terminal coil region) at L1 and L25 but the hydrogen bonds

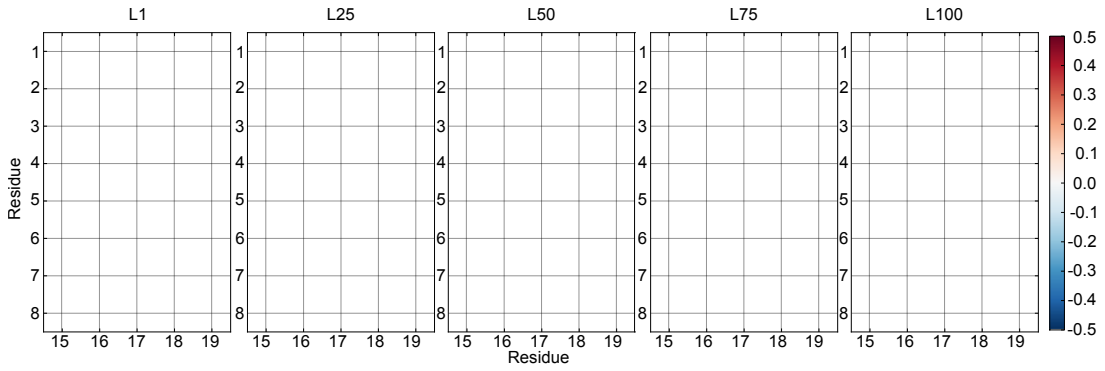


Figure 4.2: Average coupling between residues 1-9 and residues 15-19 in cm^{-1}.

increase after L50. Furthermore, at L75, the coupling between these two groups changes, indicating an orientational rearrangement. Comparing L75 and L100, the coupling pattern remains largely the same, however, with the larger magnitudes. This indicates that the structural changes between L75 and L100 involves only minor rearrangements. Also, the transition dipole couplings between residues are shown in Fig.4.3. At both L1 and L25, the couplings are weak and primarily nearest-neighbor. This is indicative of the random coil structure at these locations. At L50, the α-helix extending from residues 2 to 9 has formed as seen by the strong positive nearest neighbor coupling and the strong negative 1-3 coupling in this region[94]. The same coupling pattern was observed in simulations of the Villin headpiece, which contains three α-helices[94]. At L75, we observe the formation of the short 3_{10}-helix-like structure from residues 11 to 14. This can be seen by the strong positive nearest-neighbor coupling in this region. At L100, the system has reached the native state and we see the N-terminal α-helix and the 3_{10}-helix-like structure from residues 11 to 14.

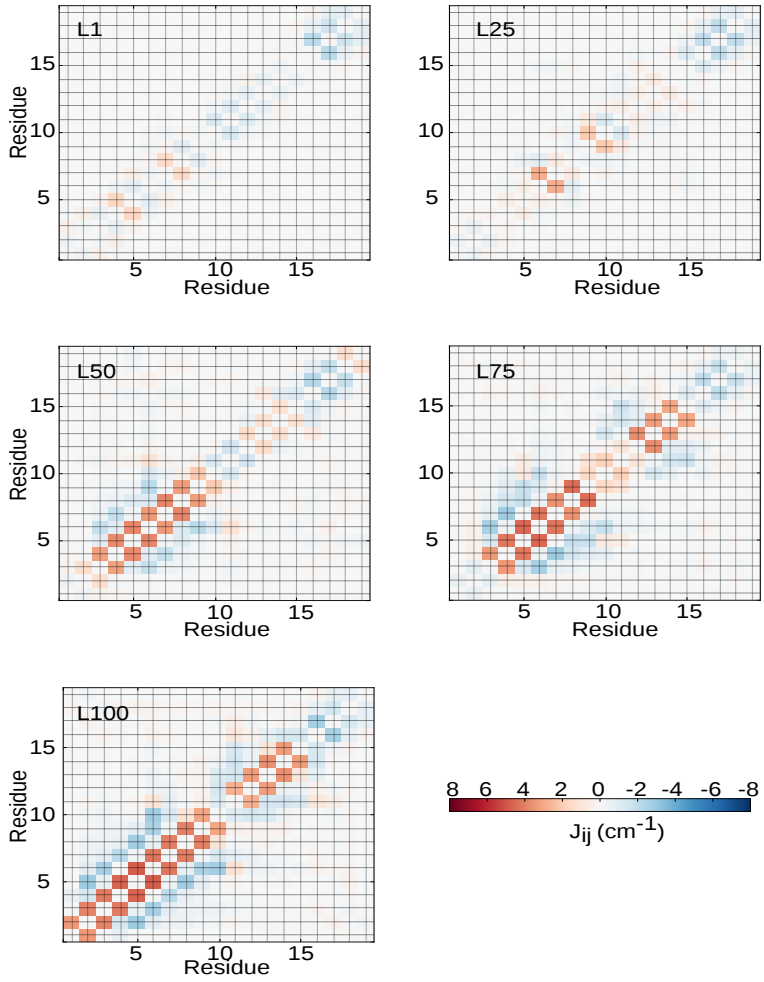


Figure 4.3: Average transition dipole couplings among the different amide I vibrational modes for L1, L25, L50, L75, and L100.

4.4.2 2DIR Spectra of Peptide Folding

Unless specified otherwise, in the following simulations, we considered an isotopomer of the Trp-cage where the Trp6 and the Pro18, which is an important link that stabilizes the native state[117], were $^{13}C{=}^{18}O$ isotope labeled. To account for the isotope-labeling, the field free frequency of the isotope-labeled amide I modes is red-shifted by 65 cm^{-1}compared to the unlabeled modes[25]. The double isotope-labeling scheme is used to obtain information on the local dynamics of Trp6 and Pro18, which are on opposite "strands" of the peptide. Therefore, their coupling may provide information on the formation of the tertiary structure of the peptide.

Linear Absorption and Nonchiral 2DIR spectra

Fig.4.4(a) shows the amide I absorption spectra of the five FEL locations. The amide I absorption of the unlabeled group results in a single peak, The peak intensity decreases from L1 to L50 and then increases from L50 to L100, which indicates that the ordered structures, including the linear extended structure and folded structures, tend to enhance the intensity. The maximum red-shifts from 1650 cm^{-1}at L1 and L25 to 1640 cm^{-1}at L100. This is due to the formation of the secondary structure and hydrogen bonds, as shown in Table 4.1, which weakens the C=O bond and reduces its vibrational frequency[88]. The 10 cm^{-1}redshift is consistent with recent one-dimensional time-resolved IR experiments on the Trp-cage[118]. Since the contributions from the isotope-labeled residues are much weaker than those from the unlabeled residues, we enlarge the spectral region corresponding to isotope-labeled amide I modes in Fig.4.4(b)-(f). There are two bands in the isotope-labeled region of the linear absorption spectrum. One is near 1560-1570 cm^{-1}, and the other is around 1580-1590 cm^{-1}. To determine the origins of these two bands, we have calculated the projected density of states[94] which shows that the higher frequency band in the isotope-labeled region originates from Pro18 while the lower frequency band originates from Trp6, as shown in Fig.4.4(b)-(f).

The absorptive 2DIR nonchiral spectra are displayed in Fig.4.6. All spectra are dominated by an inhomogeneously (diagonally) broadened peak centered near (-1640, 1640) cm^{-1}. The diagonal L100 peak is red-shifted by $\approx$ 10 cm^{-1}compared to L1, consistent with the linear absorption spectrum and the previous study[97]. The similarity of the 2DIR nonchiral spectra of the unlabeled amide groups indicates that the nonchiral signals are not very sensitive to protein secondary structure motifs without the use of

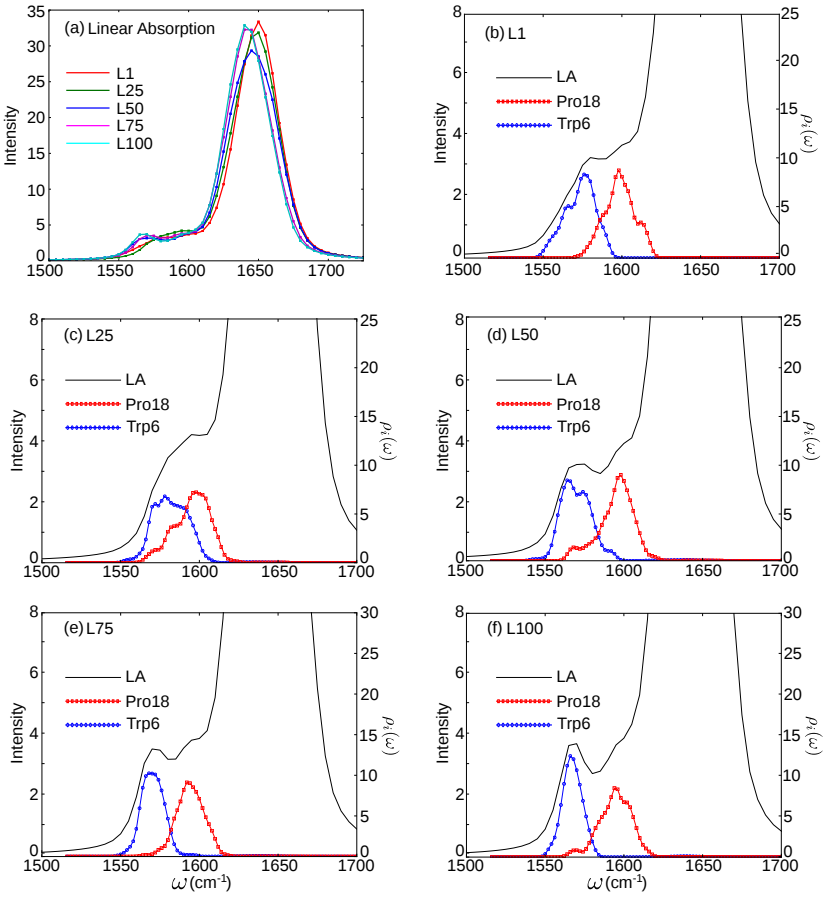


Figure 4.4: (a) Amide I absorption spectra for L1, L25, L50, L75, and L100 where Trp6 and Pro18 are isotopically labeled. (b)-(f) Isotope-labeled region of the linear absorption spectra and projected density of states for L1, L25, L50, L75, and L100.

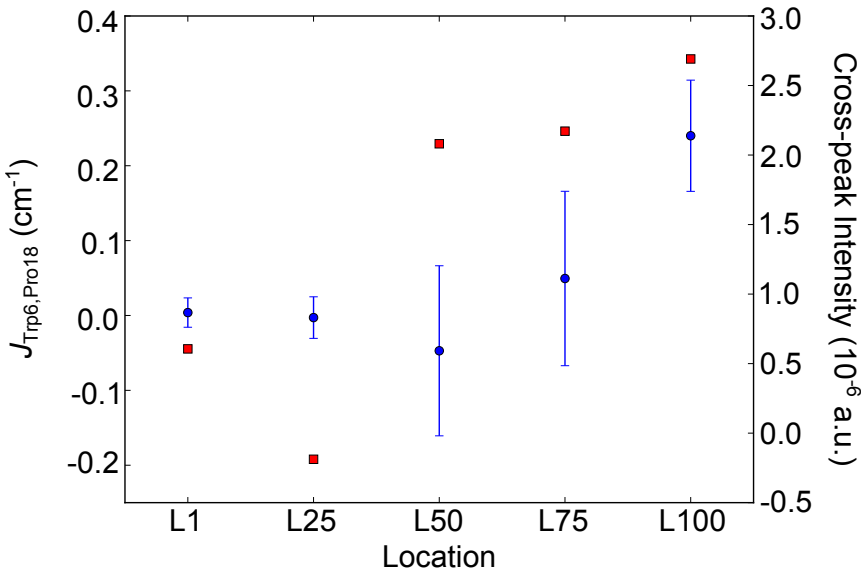


Figure 4.5: Correlation plot of the coupling of Trp6 and Pro18 and the cross peak intensity. The average coupling is displayed along with the standard deviation in blue. The cross-peak intensities are displayed in red.

site-specific isotope-labeling. The 2DIR nonchiral spectra in the region of the isotope-labeled residues shows some interesting features during folding (Fig.4.6). Starting at L50, two isotope-labeled bands clearly begin to emerge at approximately (-1570,1570) $cm^{-1}$and (-1590,1590) cm^{-1}. The band around (-1570,1570) cm^{-1}gradually increases from L50 to L100 and the intensities of the band around 1590 $cm^{-1}$are almost unchanged from L50 to L100. After the two bands appear at L50, the cross peak at (-1570 cm^{-1}, 1590cm^{-1}) emerges. At L1 and L25, this cross peak is extremely weak and the coupling between the two isotope-labeling residues is nearly zero, as shown in Fig.4.5. At L50, the magnitude of the coupling increases by nearly an order of magnitude while the cross peak intensity also increases. Between L50 to L100, both the coupling and the cross peak intensity continue to increase. It should be noted that the couplings at L50 and L75 have both positive and negative values due to the varying relative orientation between Trp6 and Pro18, while at

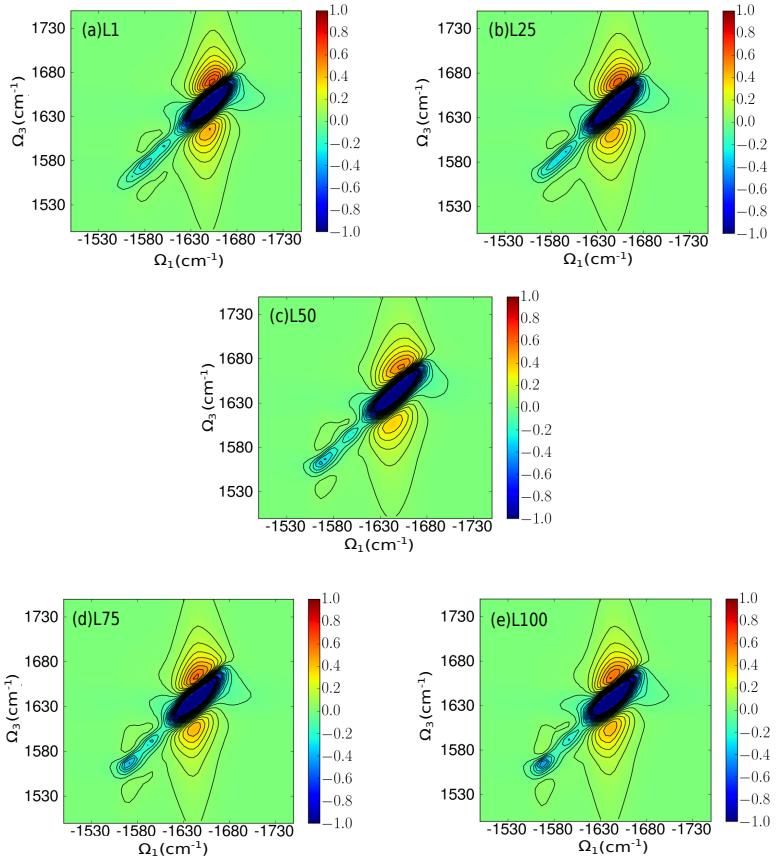


Figure 4.6: Isotope-labeled nonchiral (xxxx) 2DIR $\mathbf{k}_I + \mathbf{k}_{II}$ amide I spectra for L1, L25, L50, L75, and L100. Trp6 and Pro18 are isotopically labeled.

L100, the coupling is always positive as the relative orientation of the two strands has been stabilized.

VCD and CI 2DIR Spectra

The vibrational circular dichroism (VCD) and CI 2DIR spectra are shown in Figs. 4.7 and 4.8, respectively. The unlabeled amide I band in the VCD spectra has one negative and one positive peak (Fig.4.7(a)). However, these peaks are red-shifted as the peptide moves from L1 to L100. At L25, the VCD signal nearly vanishes due to the cancellation of various random coil configurations upon ensemble averaging. After the formation of the compact form at L50, the VCD intensity ascends from L50 to L100. The isotope-labeled

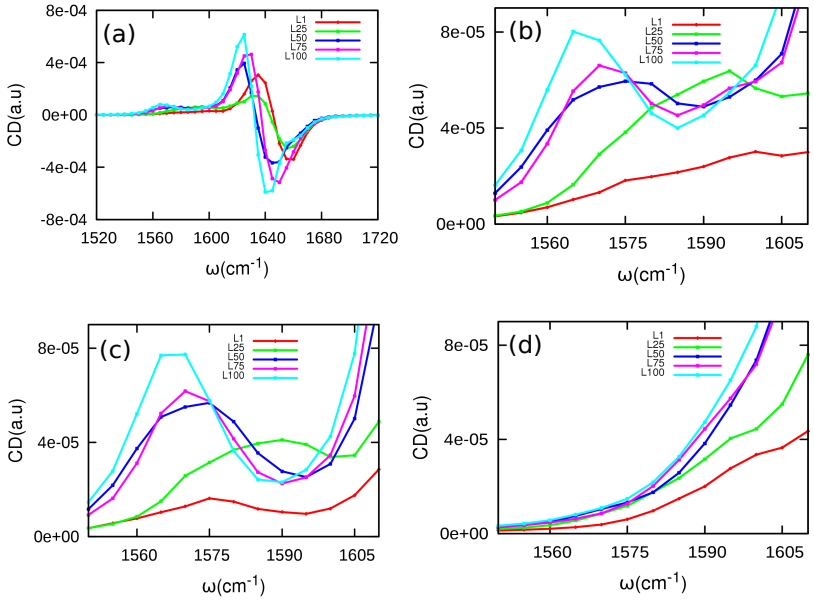


Figure 4.7: (a) VCD spectra for L1, L25, L50, L75, and L100 with Trp6 and Pro18 both isotopically labeled. (b) Isotope-labeled region of the vibrational circular dichroism spectra with Trp6 and Pro18 both labeled. (c) Isotope-labeled region of the VCD spectra with only Trp6 labeled. (d) Isotope-labeled region of the VCD spectra with only Pro18 labeled

band of the VCD spectra is shown for three different isotopomers in Fig.4.7(b)-(d). To determine the contributions of Trp6 and Pro18 to the VCD spectrum, we calculated the VCD spectra of the isotopomers where only Trp6 (Fig. 4.7c) or Pro18 (Fig.4.7d) are labeled. In the double-labeled spectrum (Fig.4.7(b)), the peak is redshifted from 1600 cm^{-1}at L1 to 1565 cm^{-1}at L100. At L50, L75, and L100, the single-labeled spectra of Trp6 closely resemble the double-labeled spectra, demonstrating that Trp6 dominates the double-labeled spectrum after L50. However, at L1 and L25, the double-labeled spectra is dominated by the contribution of Pro18, resulting in a peak at $\approx$ 1590-1600 cm^{-1}.

For the CI 2DIR spectra, one should consider two types of chirality in proteins. One is related to global structure and the other is associated with the local chirality originating from the individual vibrational modes. In this work we consider the former because it dominates the response in extended systems[26]. The Fig.4.8 displays the absorptive chiral signals of five states. The chiral spectra (Fig.4.8) show two inhomogeneously broadened diagonal peaks that are surrounded by four symmetrically distributed cross peaks. The stronger diagonal peak is initially located at (-1645,1645) $cm^{-1}$and redshifts to (-1630,1630) cm^{-1}during folding, while the weaker peak is initially located at (-1625,1625) $cm^{-1}$and redshifts to (-1610,1610) cm^{-1}. Both peaks contain contributions from several highly delocalized transitions which cannot be assigned to specific sites. The cross peak of these two transitions is seen at approximately (-1620,1645) $cm^{-1}$and is related to the coupling between these two classes of delocalized transitions. We note that in the non-chiral spectra, there was only a single diagonal peak for the unlabeled amide I band, presenting an obvious advantage for CI 2DIR.

During folding, the cross peaks are also red-shifted along the diagonal, similar to the diagonal peaks. At L1 and L25, the four cross peaks are weak. When the peptide evolves to L50, the intensities of the four cross peaks increase due to the increased coupling caused by the relatively compact conformation. After L50, the peak locations remain similar while the intensities of the four cross peaks increase(Fig.4.8(d) and (e)). The increase in the intensities of the cross peaks is caused by the increased coupling between residues (Fig.4.3). Another interesting feature is the isotope-labeled band around 1560-1590 cm^{-1}. At L1 and L25, this band is extremely weak due to the cancellation of contributions from various random coil configurations, as in the VCD spectra. After L50, a weak band appears that gradually enhances and red-shifts as the peptide folds, indicating that the

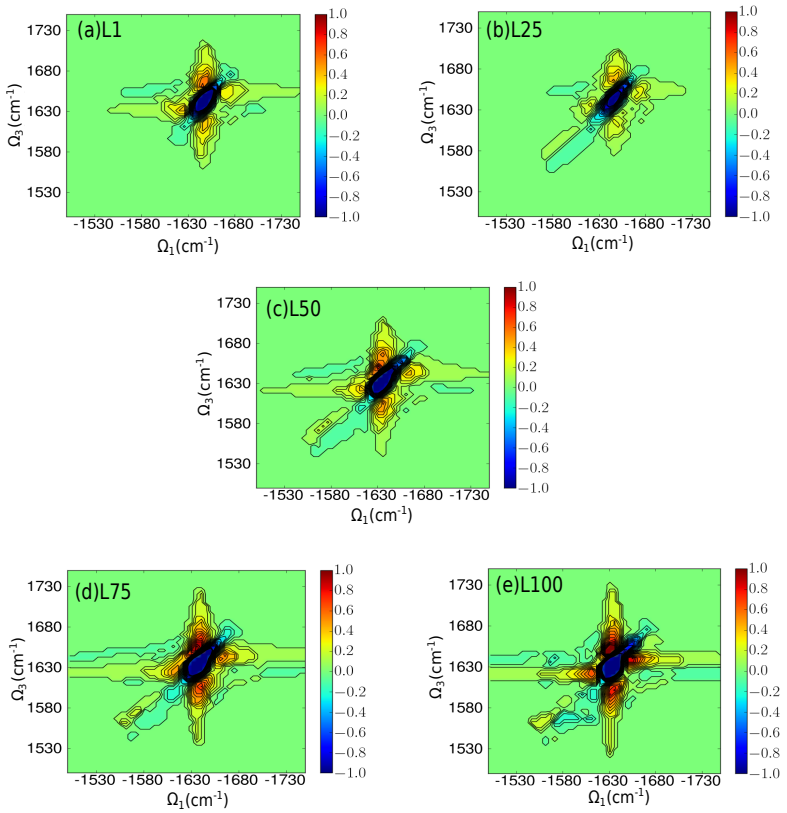


Figure 4.8: Isotope-labeled chirality-induced (xxxy) 2DIR spectra for L1, L25, L50, L75, and L100.

two isotope-labeled residues, Trp6 and Pro18, get close to each other and maintain a particular relative orientation from L75 to L100.

4.5 Conclusions

The folding of a 20-residue peptide Trp-cage was simulated to build up the FEL which suggests one dominant pathway that connects the unfolded and native state. We demonstrated that valuable conformational information about the structural evolution on the folding pathway can be revealed by multidimensional IR spectroscopy. The amide I absorption, VCD, non-chiral 2DIR, and CI 2DIR spectra were simulated to structurally illustrate the characters of the folding process. The spectra were calculated for an isotopomer where Trp6 and Pro18 were $^{13}C{=}^{18}O$ labeled. The linear absorption spectra shows a 10 cm^{-1}redshift of the unlabeled amide I band, consistent with experimental results. In the isotope-labeled region of the linear absorption, there are two peaks around 1560 $cm^{-1}$and 1590 cm^{-1}, which are caused by Trp6 and Pro18, respectively. The diagonal peak of the unlabeled amide I band in the 2DIR xxxx spectrum redshifts 10cm^{-1}during folding. The cross peak intensity between the $^{13}C{=}^{18}O$ labeled Trp6 and Pro18 amide I transitions increases during folding. The intensity of this cross-peak is correlated with the coupling between these two groups and is indicative of the formation of the peptide's tertiary structure, which is consistent with the analysis of transition dipole coupling and hydrogen bonds. The VCD spectra reveal the two peaks for the unlabeled amide I band which is redshifted during the folding. The overall intensity of the VCD spectra increases during the folding because the cancellation of various random coil conformations upon ensemble averaging of the signals. The isotope-labeled band reveals two peaks around 1560 $cm^{-1}$and 1600 cm^{-1}which are red-shifted, similar to the linear absorption. The CI

2DIR spectra show two distinct diagonal peaks for the unlabeled amide I band, whereas the non-chiral spectra only show a single diagonal peak in this region. Strengthening of the cross peaks of the unlabeled amide I band can be observed during folding which is directly correlated with the increase in coupling.

One should consider that the accuracy of the parameters in MD simulation software(in our case is AMBER) could influence the 2D spectra. The study of another peptide named Beta3s[119] has shown that the under-stabilization of PPII conformatons[120] in MD simulations of unfolded peptides may cause random coil structure to display a positive couplet. The simulations indicate that only 2 percents of conformations exist in the PPII state at random-coil structure. Newer force fields which have been optimized to accurately calculate properties for unfolded peptides.

Hamm *et al.* had measured 2DIR signals of a photoswitchable isotope-labeled α-helix[92] whose structure is very similar to the Trp-cage in the folded state. Our simulation results are consistent with their findings for the changes between the folded and unfolded conformations. They also found the two bands for some residues in the isotope-labeled region. Note that the folding time scale in our simulations is not necessary the real experimental folding time scale since the Generalized Born solvent model was applied to run the simulations and this implicit solvent model reduced the folding time dramatically. Nevertheless, the conformational transition of the folding process and corresponding spectrum changes are our main concern, and the implicit solvent model and spectrum calculations have proven to give reasonable results in previous studies[26,94,108]. While CI 2DIR experiments have yet to be performed, our simulations show that CI 2DIR measurements may reveal changes in cross peaks which may be difficult to see in non-chiral 2DIR measurements.

4.6 Summary

In Chapter 4 the free energy landscape and two-dimensional spectroscopy are employed in the study of the folding mechanism of Trp-cage. Here we models the free energy landscape generated by the molecular dynamics simulation and calculate the signatures of 2DIR spectra towards the structural identification and characterization of the intermediate state ensemble a critical component of the folding pathway. The tools we use here are molecular dynamics simulation and two-dimensional infrared spectroscopy.

Chapter 5 Probing the Peptide Trp-cage Folding with Two Dimensional Ultraviolet Spectroscopy

5.1 Abstract

Ultraviolet (UV) spectra of proteins originate from the electronic excitations of their backbone chromophore and aromatic side chains and thus provide a sensitive probe of the secondary structures. Recently developed femtosecond lasers allow the multidimensional spectroscopy to be extended into the UV regime. Two-dimensional UV (2DUV) techniques, with short pulses, provide a promising tool to study the structures and dynamics of proteins. We combined 2DUV spectroscopy and molecular dynamics generated free energy profiles to simulate the protein electronic transitions and UV photon echo signals to monitor the peptide folding process of a mini-protein Trp-cage. The ultraviolet signals illustrate the variation of the 2D correlation plots when the protein evolves along the underlying free energy landscape. The complexity of signals decreases as the conformational entropy decreases during the folding process. We show that the approximate entropy (ApEn) of the signals, which is accessible by both theoretical calculations and experiments, provides a quantitative indicator of the protein folding status.

5.2 Introduction

Theoretical and experimental studies have provided considerable insights into the protein folding process. However, monitoring protein folding dynamics is still challenging. Experiments on the kinetics and thermodynamics of folding do not usually provide

atomic-level structural changes, and thus, the microscopic dynamics characterizing the folding process. Computer simulations performed at various levels of complexity, from simple lattice models to all-atom models, fill in some of the gaps in our knowledge of protein folding. Unfortunately, because of the current computational capacity, most folding processes of interest occur on timescales (microsecond to second) that are inaccessible to standard all-atom molecular dynamics (MD) simulations[121,122]. Therefore, faster folders may be used to overcome this limitation. In recent years, many fast-folding proteins have been characterized to fill this need. With modern supercomputers, long simulation times with fast-folding protein may provide the first direct insight into the mechanism of protein folding. According to the current view, protein folding is envisioned to proceed along a moderately rough energy landscape, the major features of which are the local minima and overall funnel-like downhill slope toward the native state[123].

Optical spectroscopy is a powerful tool to probe the structural details and transformation of biological molecules[124–127]. Experimental techniques, including linear optical, Raman, fluorescence spectroscopy X-ray scattering and Laue diffraction, have been widely used to detect proteins structures[128–130]. However, experimental methods often lack the time resolution for monitoring the whole folding dynamics and structural information at high resolution to observe the ultrafast protein processes, and they result in indirect information about the structures along the folding pathway on the free energy surface of proteins. Recent simulation advances in multidimensional correlation spectroscopy[127,129,130], which provides much higher time resolution, has emerged as a new probe to investigate the mechanism of protein folding. The technique employs sequences of laser pulses to probe the electronic or vibrational degrees of freedom and detects correlated events during controlled time intervals. The resulting multi-point correlation functions contain

detailed structural and dynamics information. Studies so far have shown some success in the IR, near-IR, and visible spectral regions, however, extension to the UV domain is in its infancy. New, intense, and stable femtosecond lasers with high repetition rates allow multidimensional spectroscopy to be extended into the UV range, whose advantages over 2DIR come through shorter pulse durations and higher quality polarization control[131]. Therefore, the study of 2DUV on protein dynamic could give us a new window to investigate the mechanism of folding. The 2DUV signals require the computation of the system Hamiltonian, including the environment at the quantum mechanics (QM) level, which is extremely expensive. Traditionally, some empirical methods, such as the dipole approximation and the map method[132] have been applied to reduce computational cost, yet these methods require empirically fitted parameters. The EHEF (excitation Hamiltonian with electrostatic fluctuations) algorithm[133], free from empirical parameters, provides an efficient approach to calculate the accurate system Hamiltonian involving the environment at the QM level. Herein, we report first principles simulation of 2DUV spectra of the protein folding based on EHEF algorithms. Additionally, we have applied the chirality-induced pulses technique to probe the structural changes of a protein. Signals with carefully designed chirality- induced polarization configurations have high sensitivity to protein structural evolution. Although 2DUV techniques have been used to study the protein molecules[131], the studies have only focused on the near-native states or simply single trajectory of folding, and applying 2DUV spectra to monitor the whole process based on the ensemble pathway is still rare.

5.3 Model and Methods

In this work, we continued to use the peptide Trp-cage as the model protein to illustrate the simulations of 2DUV spectroscopy of the folding process. The Trp-cage is a fast folder and thus is one of the good modeling systems used to study the folding mechanism[134]. It contains 20 residues with the sequence "NLYIQWLKDGG PSSGRPPPS". The settings of molecular simulations were similar to the Chapter 4. Briefly, the initial structures were the extended amino-acid chains. 50 200ns trajectories were ran with different initial conditions at 315 K. The 10-μs simulations were enough for building the free energy landscape, which was calculated as F =-log(P), where P was the population obtained from all the 10-μs MD simulated data, as shown in the inset graph of Fig. 5.1 (A). 100 state points along the folding pathway on the energy landscape denoted L1, L2, ..., L100, were selected on the FEL. We chose 200 sub-conformations (MD snapshots) around each state point. The variation of RMSD (root mean square deviation) and Rg (Radius of gyration) along the folding path are shown in Fig. 5.1 (B) and (C), respectively.

The Hamiltonian calculation of the system includes the electronic transition of chromophores including peptide unit, benzene, phenol, and indole can be modeled by the Frenkel exciton Hamiltonian with the Heitler-London approximation, and details of the calculations can be found somewhere else[135,136].

After computing the Hamiltonian of the system, 2DUV calculations were performed for the non-chiral (xxxx, xyyx, and xyxy) and the chirality-induced (xxxy) pulse polarization configurations after the Hamiltonian was calculated. Chiral 2D signals record interferences among transitions at different parts of the whole protein, and thus provide richer spectral features compared to their non-chiral counterparts. The signals are dis-

played on a non-linear scale that interpolates between logarithmic for small values and linear for large values thus revealing both the strong and weak features.

5.4 Results and Discussion

The Trp-cage folding is a packing process from a strand to a compact cage, and the evolution of tertiary structure can be characterized by the protein packing density[137]. We computed the packing density as the average of the number of residue's Cα atoms within a 9 Angstrom radius of the Cα atom of a given residue. The evolution of the inverse of packing density displayed in Fig. 5.1 (C) is consistent with the Rg. The packing density is closely related to the protein conformation entropy, and its evolution suggests that the Trp-cage conformation entropy decreases along the packing (folding) process. This observation is consistent with the concept of energy landscape.

The circular dichroism (CD) spectrum is the standard one dimensional spectroscopic technique widely used for identifying protein secondary structures. In Fig.5.2, we depicted the structures of 5 states: L1, L25, L50, L75, and L100 and the corresponding computed CD signals. The CD spectrum of L100 with the final folded structure agrees well with the experiment[138] of the folded Trp-cage peptide. From L1 to L100, the CD signals reflect the variation of secondary structural elements. The negative feature at $\sim$56000 cm^{-1}($\sim$180 nm) and positive signals at $\sim$43000 cm^{-1}($\sim$230 nm) marked 'RC' are typical of a random coil. They are in L1 yet reverse in L75 and L100, since the decrease of random coil, as shown in Fig. 5.1 (D). Besides, the helix structures increase from L1 to L100, so the CD from 53000 to 58000 cm^{-1}($\sim$190 to 170 nm) marked 'H' changes from negative to positive.

The couplings between electronic transitions and structural variations significantly affect the 2D photo echo signal. 2DUV xxxx (non-chiral) and xxxy (chiral) spectra of the

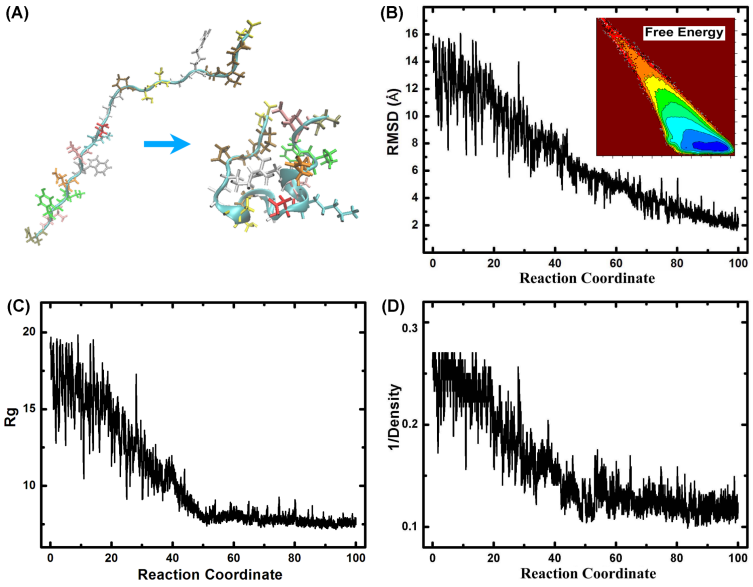


Figure 5.1: (A) From the unfolded strand to folded cage structure of a Trp-cage protein (PDB code: 1L2Y). The backbone trace is shown as a ribbon, the side chains are depicted with wires. The RMSD (B), Rg (C), and inverse of packing density (D) along the free energy landscape of Trp-cage folding process (from L1 to L100). The inset in (B) shows the free energy landscape, which is the same as the one in Chapter 4.

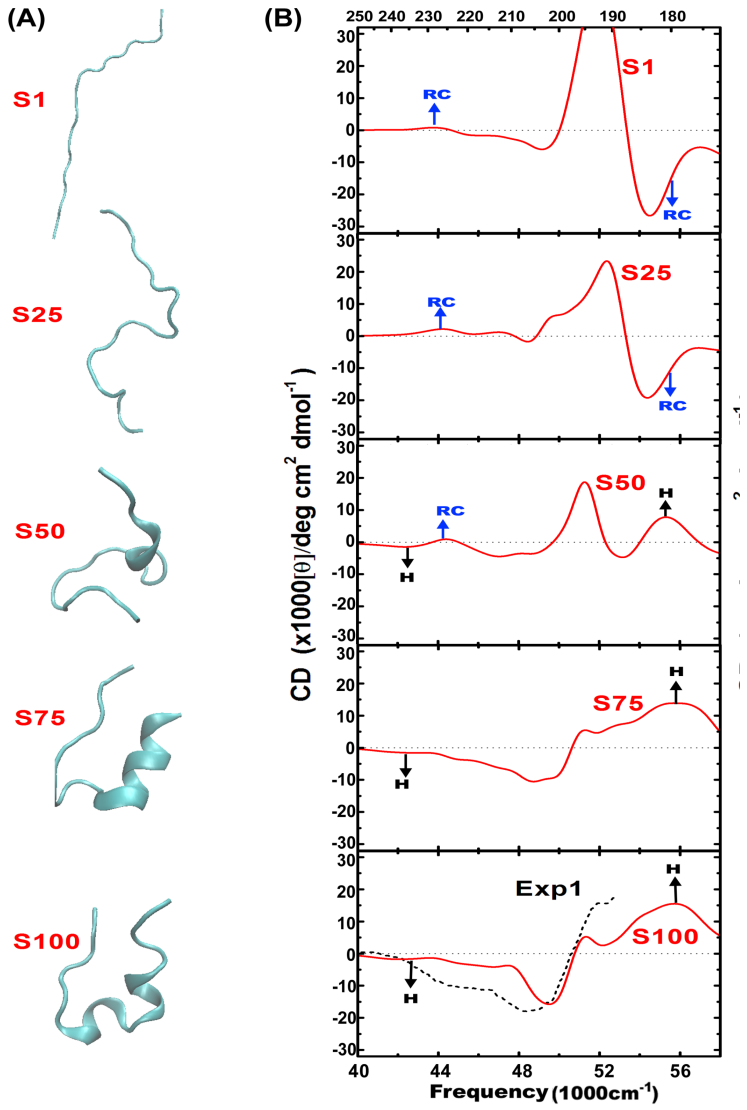


Figure 5.2: Structure (A) and CD spectra (B) of 5 states (from top to bottom:L1, L25, L50, L75, L100) along the trp-cage folding process. Spectra are averaged over 100 MD snapshots for each state. We labeled CD signals of the random coil and Helix as 'RC' and 'H', respectively. Black dotted CD curve is from experiment Exp1[138] for the folded trp-cage protein.

chosen 5 states are displayed in the right and middle column of Fig. 5.3, respectively. From L1 to L100, all non-chiral 2DUV spectrum have strong diagonal peaks ∼52000 cm^{-1} and relatively weak cross-peaks distributed symmetrically around the diagonal peaks. The diagonal peaks are stronger than the cross-peaks indicating that the electrostatic interaction between the different amides is weaker than the amide transition energy. The similarity of the 2D nonchiral spectra indicates the signals are not sensitive to protein secondary structure motifs In contrast, the xxxy signals vary significantly as the peptide moves from L1 to L100. The unfolded states L1 and L25 have negative diagonal peaks at from 48000 to 56000 cm^{-1}, which are typical region for random coil and strand structural motifs[131]. A helical structure normally produces positive diagonal signals in that region[131]. These observations imply that from L50 to L100, the increased helical structure reduces the negative signals and induces additional positive peaks at the diagonal part. The xxxy chiral signals also reflect the tertiary structure. As expected from the decrease of conformational entropy in the folded structure, the xxxy spectral pattern becomes more compact and simple, decreasing the signal complexity. The number of peaks (marked with white dots in Fig. 5.2) in the xxxx and xxxy spectra of the five states are plotted in Fig. 5.4 (A). The number of xxxx peaks remains similar during the folding. The xxxy peak number decreases from L1 to L50, and remains flat from L50 to L100, implying a trend similar to the variation of Rg and the inverse of packing density. The previous study[139] showed that the approximate entropy (ApEn)[140] provides a good measure for the complexity of 2D signals. In the subfigure of Fig. 5.4 (B) we display a scanning line perpendicular to the diagonal of the 2D contour map, starting from the bottom left (lower energy) to the upper right (higher energy) corner. The projections of the xxxy signals of state L1 and L100 along this line are depicted in Fig. 5.4 (B), showing that the L1 signals have

richer structures (such as more peaks) than the ones of L100. Fig. 5.4 (C) depicts the variation of ApEn values of xxxy spectra during the folding process. The xxxy ApEn decreases considerably as the peptide moves from L1 to L50, and keeps almost the same from L50 to L100, consistent with the evolution of Rg and the inverse of packing density (conformational entropy) shown in Fig. 5.1 (C) and (D).

Chiral signals are harder to measure due to their very weak intensities. The technique of difference spectroscopy between two non-chiral spectra with different polarizations can cancel the single exciton contributions such that the correlations of transitions are retained and better resolved. The computed 2DUV xyyx-xyxy difference spectra of our five states are displayed in the right column of Fig. 5.3. The signal complexity is reduced as we move from L1 to L50. This may also be seen from the number of 2DUV xyyx-xyxy spectral peaks (marked with white dots in Fig. 5.3), and the ApEn values shown in Fig. 5.4 (A) and (C). The change of the complexities of difference spectra thus also provide a quantitative indicator of the decrease of protein conformational entropy during the folding process.

5.5 Conclusion

Lately, the multidimensional time-resolved ultrafast spectroscopy has provided a powerful tool in investigating protein folding. Herein we presented a study of the peptide folding of Trp-cage by combining the multidimensional UV spectroscopy. We applied a new technique of computing the system Hamiltonian at the QM level and CI pulses to calculate the 2DUV spectra of the structures, which were chose from the pathway of the underlying free energy landscape. Atomistic MD simulations using the AMBER force field were applied to sample the conformations and build up the underlying free energy surface.

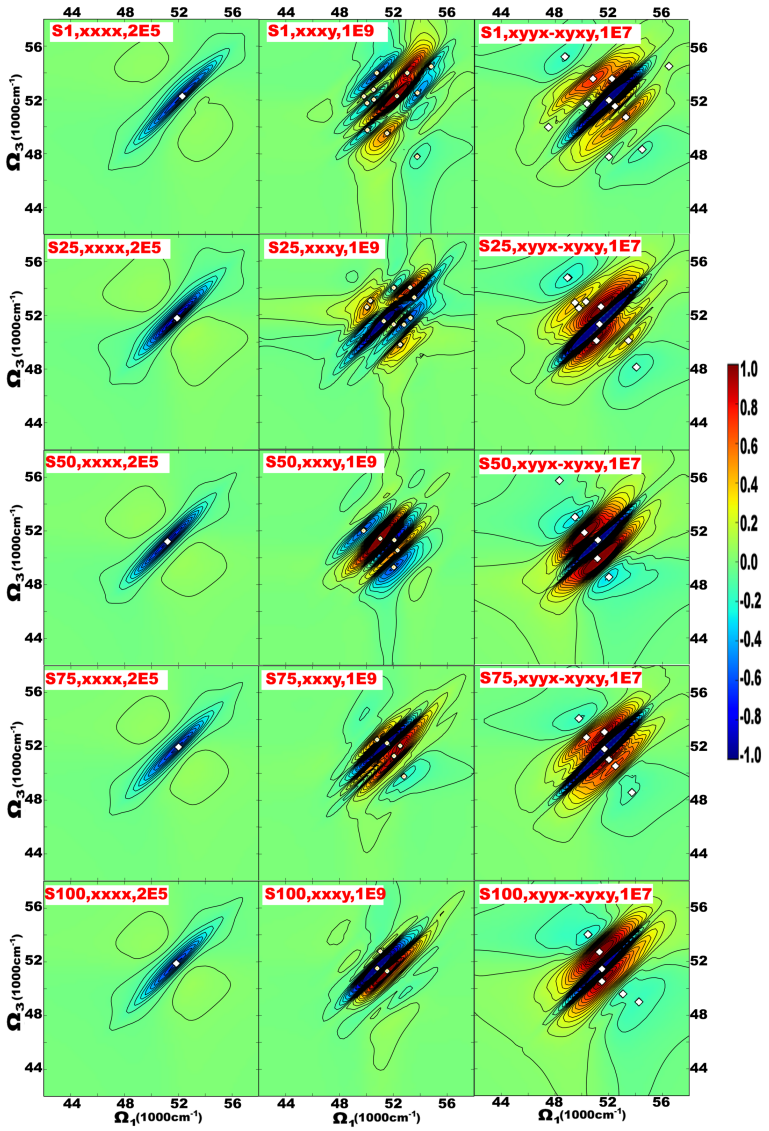


Figure 5.3: From left to right: 2DUV xxxx, xxxy, and xyyx-xyxy spectra of 5 states (from top to bottom:L1, L25, L50, L75, L100) along the trp-cage folding process. Spectra are averaged over 100 MD snapshots for each state. The scale bar is plotted at the right top edge, and signal peaks are marked by white square dots.

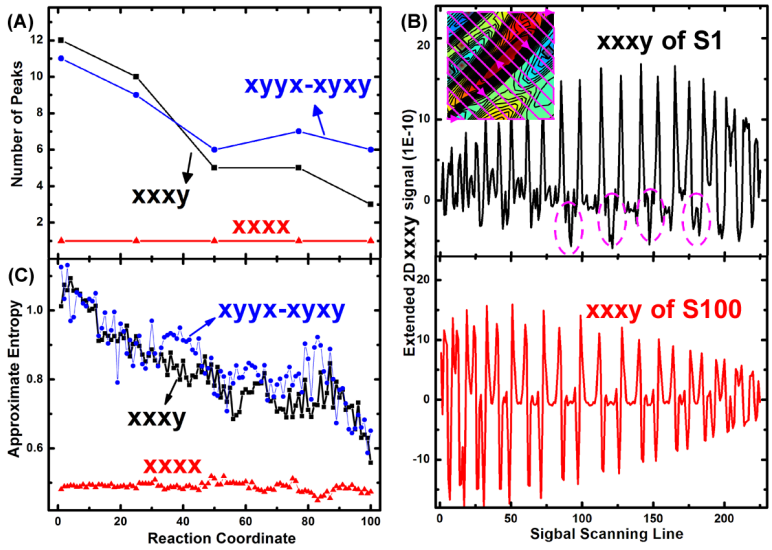


Figure 5.4: The evolution of the number of 2DUV peaks (A) and ApEn (B) during the Trp-cage folding process. Spectra are averaged over 200 MD snapshots for each state. (C): 2DUV xxxy signal evolution curves of states L1 and L100 along the scanning line given in the inset of (C). Purple circles highlight the multiple-peak patterns.

We demonstrated that 2DUV signals are sensitive to the change of peptide secondary and tertiary structure, and especially useful in probing the global structural changes. The complexity of 2DUV spectra of peptide backbone as measured by the ApEn is a potential good indicator for the conformational entropy, and provides a quantitative index of folding status with the same accuracy as the calculated RMSD and Rg values. The RMSD and Rg can only be extracted from the known atomic structures at every time point, which require huge computation resources for most proteins and are not accessible to any existing experimental techniques. As 2DUV signals are becoming feasible[127,129,130,141], protein folding can be measured by the ApEn value of 2DUV chiral signals. Our results provide a direct connection between the experiment and theoretical calculations. Ongoing and future experiments can test the predictions of the calculations. By directly connecting 2D spectroscopy experiments with molecular simulations, we can uncover the underlying mechanisms of folding and see how the folding process actually occurs in time by

monitoring the folding pathways.

5.6 Summary

A QM/MM protocol is applied to simulate the 2DUV spectra of the folding of Trp-cage peptide. 2DUV signals are demonstrated that they are sensitive to the change of peptide secondary and tertiary structure and especially useful in probing the global structural changes. The complexity of 2DUV spectra of peptide backbone as measured by their number of the peaks is a good indicator for the conformational entropy and provides a quantitative index of folding status with the same accuracy as the calculated RMSD and Rg values. 2DUV can offer a fast experimental measurement and theoretical verification of the protein folding state.

Chapter 6 Summary

The energy landscape theory provides an essential framework for understanding protein folding and binding and has been widely used to interpret the folding and binding processes. The theory assumes a funnel-like shape of the surface, which is sufficiently biased to direct the folding or binding so that they can happen on the experimental time scale. The work presented in this dissertation provides insights into the protein folding and binding through the Glutermine-binding protein and Trp-cage model systems by investigating their underlying free energy landscape. The whole thesis has two major parts. In the first part, we applied a course-grained model (residue level) and a developed microscopic two-well potential energy model to explore the thermodynamic landscape, kinetics, and structural evolution of the conformational transition of Glutermine-binding protein. Furthermore, the multi-dimensional coordinate-dependent diffusion dynamics on the energy landscape of GlnBP was investigated as well. In the second part, the all-atom molecular dynamics simulations and the effective and novel two-dimensional spectroscopies, including two-dimensional infrared spectroscopy and two-dimensional ultraviolet spectroscopy, were combined to study the folding mechanism of a peptide, Trp-cage. We will summarize the details of those work below.

6.1 Applying a two-well model to study the conformational switches of GlnBP

GlnBP exhibits the ligand-free open state and ligand-bound closed state during the process of ligand binding. Therefore, a microscopy potential energy model with two wells

was used to describe the conformational transition of GlnBP. The free energy landscapes of the conformational transition at different temperatures were built and analyzed. Two basins were observed on the free energy landscapes and the topological properties of the free energy landscape show that at low temperature most of the time the protein prefers to stay in the closed state. As temperature increases, the protein tends to cross the energy barrier and dwell in the open state. The kinetic aspects of the transition were also investigated by calculating the mean and distribution of the first passage time and also the correlation function. Both the closed- and open staying times exhibit the Γ distribution. The complexity and hierarchical structure of the underlying energy landscape was illustrated by the analysis of the autocorrelation function.

We further developed the umbrella sampling with two harmonic biasing to perform constrained simulations for estimating the diffusion coefficients. After calculating the 2D diffusion coefficient tensor, we found that the local DC on the energy landscape is anisotropic. The anisotropy and inhomogeneous coordinate dependent diffusion can shift the thermodynamic pathway of the conformational transition of the protein away from the naively expected steepest descent gradient path on the free energy landscape. Furthermore, the dominant kinetic paths do not necessarily go through the transition state that does not consider the effect of coodinate-dependent diffusion. Both the position and value of the barrier height are shifted by the inhomogeneous and anisotropic diffusion. The inhomogeneous and anisotropic diffusion will therefore modify the ϕ value analysis for identification of hot residues for conformational change dynamics. Glutamine binding protein is a relatively simple system which only includes the open and closed states, and the intermediate state is rarely observed. Our results suggest that for more complex systems that have intermediate states, the anisotropic and inhomogeneous diffusion may

also have significant influences on the kinetic and dynamic properties. This could be an interesting topic in future.

6.2 Two-dimensional spectroscopy of Trp-cage folding

In this part of work, we illustrated that two-dimensional infrared spectroscopy and two-dimensional ultraviolet spectroscopy can be used to monitor the conformational evolution on the folding free-energy landscape of Trp-cage. First, the folding of Trp-cage was simulated to build up the free energy landscape, which suggests one dominant pathway that connects the unfolded and native state. For the 2DIR spectroscopy, the simulated linear absorption spectra, nonchiral spectra, and chirality-induced 2DIR spectra illustrate the conformational changes during folding. In our simulations, the chiral 2DIR spectra displayed a very strong feature, indicating the formation of the folded state. In the nonchiral signals the formation of the folded state is seen as a similar pattern in the spectra. Comparing the chiral and nonchiral 2DIR spectra, we have found that the chiral signals are more sensitive when the structural changes. However, these signals are much weaker than their nonchiral counterparts and have not yet been observed in experiment. In the chiral 2DIR experiments, a typical box-car or pumpprobe geometry may be used. A highly sensitive detector could be achieved by upconverting the signal to the visible and detecting using a charge-coupled device[142]. For the 2DUV spectroscopy, its signals are sensitive to the change of peptide secondary and tertiary structure and especially useful in probing the global structural changes. We found that the complexity of 2DUV spectra of peptide backbone could be a promising indicator folding status. Coherent 2D experiments are now possible, but require state of the art technology. In future, the 2D spectroscopy experiments can test the models and the models also can provide the guidance for the

experiments. In our future work, we may study the systems which have multi-pathways. It would be interesting if we compare the spectra on the different pathways.

References

[1] Selkoe, D. J. *Neuron* **1991**, *6*, 487–498.

[2] Lundmark, K.; Westermark, G. T.; Olsn, A.; Westermark, P. *Proc. Natl. Acad. Sci. USA.* **2005**, *102*, 6098–6102.

[3] Pepys, M. B. *Annu. Rev. Med.* **2006**, *57*, 223–241.

[4] Levinthal, C. *Proceedings in Mossbauer Spectroscopy in Biological Systems, edited by Debrunner, P. and Tsibris, J. and Munck, E.* **1969**, 22.

[5] Anfinsen, C. B. *Biochem. J.* **1972**, *128*, 737–749.

[6] Levy, Y.; Wolynes, P. G.; Onuchic, J. N. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 511–516.

[7] Shoemaker, B. A.; Portman, J. J.; Wolynes, P. G. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 8868–8873.

[8] Papoian, G.; Wolynes, P. G. *Biopolymers* **2003**, *68*, 333–349.

[9] Wang, J.; Verkhivker, G. M. *Phys. Rev. Lett.* **2003**, *90*, 188101.

[10] Dill, K.; Chan, H. *Nat. Struct. Biol.* **1997**, *4*, 10–19.

[11] Onuchic, J.; Wolynes, P. *Curr. Opin. Struct. Biol.* **2004**, *14*, 70–75.

[12] Charles L. Brooks, C. L. I.; Onuchic, J. N.; Wales, D. J. *Science* **2001**, *293*, 612–613.

[13] Brooks, B. R.; Brooks, I., C. L.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A. e. a. *J. Comp. Chem.* **2009**, *30*, 1545–1615.

[14] Case, D. A.; Darden, T.; Cheatham, T. E. I.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker, R. C.; Zhang, W.; et al, *AMBER 10, University of California, San Francisco.* **2008**,

[15] Ponder, J. M.; Case, D. A. *Adv. Prot. Chem.* **2003**, *66*, 27–85.

[16] Gonzalez, M. A. *Collection SFN* **2011**, *12*, 169–200.

[17] Mackerell, A. D. J. *J. Comput. Chem.* **2004**, *25*, 1584–604.

[18] Dominy, B. N.; Brooks, C. L. *J. Phys. Chem.* **1999**, *103*, 3765–3773.

[19] Shakhnovich, E. *Chem. Rev.* **2006**, *106*, 155988.

[20] Klimov, D. K.; Thirumalai, D. . *Phys. Rev. Lett.* **1997**, *79*, 31720.

[21] Veitshans T, T. D., Klimov D *Fold. Des.* **1996**, *2*, 122.

[22] Go, N. *Annu. Rev. Biophys. Bioeng.* **1983**, *12*, 183–210.

[23] Mukamel, S.; Abramavicius, D. *Chem. Rev.* **2004**, *104*, 2073–2098.

[24] Zhuang, W.; Sgourakis, N. G.; Li, Z.; Garcia, A. E.; Mukamel, S. *Proc. Natl. Acad. Sci.* **2010**, *107*, 15687–15692.

[25] Zhuang, W.; Hayashi, T.; Mukamel, S. *Angew. Chem. Int. Ed.* **2009**, *48*, 3750–3781.

[26] Zhuang, W.; Abramavicius, D.; Mukamel, S. *Proc. Natl. Acad. Sci.* **2006**, *103*, 18934–18938.

[27] Royer, C. A. *Chem. Rev.* **2006**, *106*, 1769–1784.

[28] Lillo, M. P.; Beechem, J. M.; Szpikowska, B. K.; Sherman, M. A.; Mas, M. T. *Biochemistry* **1997**, *36*, 11261–11272.

[29] Clementi, C.; Jennings, P. A.; Onuchic, J. N. *J. Mol. Bio* **2001**, *311(4)*, 879–890.

[30] Shea, J. E.; Onuchic, J. N.; Brooks, C. L. *Proc. Natl. Acad. Sci.* **2002**, *99(25)*, 16064–68.

[31] Cheung, M. S.; Garcia, A. E.; Onuchic, J. N. *Proc. Natl. Acad. Sci.* **2002**, *99(2)*, 685–690.

[32] Hsiao, C. D.; Sun, Y. J.; Wang, B. C. *J. Mo. Biol* **1996**, *262*, 225–242.

[33] Sun, Y. J.; Rose, J.; Wang, B. C.; Hsiao, C. D. *J. Mol. Biol.* **1998**, *278*, 219–229.

[34] Lu, Q.; Wang, J. *J. Am. Chem. Soc.* **2008**, *130*, 4772–4783.

[35] Mehnert, T.; Jacob, K.; Beyer, K. *Biophys. J.* **2006**, *90*.

[36] Maulik, P. R.; Shipley, G. G. *Biochemistry* **1996**, *35*.

[37] Korkuta, A.; Hendricksona, W. A. *Proc. Nati. Acad. Sci. USA* **2009**, *106*.

[38] Shoemaker, B. A.; Portman, J. J.; Wolynes, P. G. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 8868–8873.

[39] Cheung, M. S.; Garcia, A. E.; Onuchic, J. N. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 685–690.

[40] Best, R. B.; Chen, Y. G.; Hummer, G. *Structure* **2005**, *13*, 1755–173.

[41] Okazaki, K. I.; Koga, N.; Takada, S.; Onuchic, J. N.; Wolynes, P. G. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 11844–11849.

[42] Okazaki, K. I.; Takada, S. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 11182–11187.

[43] Chu, J. W.; Voth, G. A. *Biophys. J.* **2007**, *93*, 3860–3871.

[44] Sobolev, V.; Wade, R. C.; Vnend, G.; Edelman, M. *Proteins: Struct. Funct. Genet.* **1996**, *25*, 120–129.

[45] Edman, L.; Mets, U.; Rigler, R. *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 6710–6715.

[46] Palmer, R. G.; Stein, D. L.; Abrahams, E.; Anderson, P. W. *Phys. Rev. Lett.* **1984**, *54*, 958–961.

[47] Fersht, A. R.; Leatherbarrow, R. J.; Wells, T. N. C. *Biochemistry* **1987**, *26*, 6030–6038.

[48] Matouschek, A.; Fersht, A. R. *Methods Enzymol* **1991**, *202*, 81–112.

[49] Ejtehada, M. R.; Avall, S. P.; Plotkin, S. S. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 15088–15093.

[50] Bryngelson, J. D.; Wolynes, P. G. a. *J. Phys. Chem.* **1989**, *93*, 6902–6915.

[51] Lee, C. L.; Lin, C. L.; Stell, G.; Wang, J. *Phys. Rev. E* **2003**, *67*, 041905–041910.

[52] Lee, C. L.; Stell, G.; Wang, J. *J. Chem. Phys.* **2003**, *118*, 959–968.

[53] Wang, J. *Biophys. J.* **2004**, *87*, 2164–2171.

[54] Socci, N. D.; Onuchic, J. N.; Wolynes, P. G. *J. Chem. Phys.* **1996**, *104*, 5860–5868.

[55] Pogorelov, T. V.; Luthey-Schulten, Z. *Biophys. J.* **2004**, *87*, 207–214.

[56] Hummer, G. a. *New J. Phys.* **2005**, *7*, 34–48.

[57] Best, R. B.; Hummer, G. *Phys. Rev. Lett.* **2006**, *96*, 228104–228108.

[58] B., B. R.; Hummer, G. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 1088–1093.

[59] Best, R. B.; Hummer, G. *Phys. Chem. Chem. Phys.* **2011**, *13*, 10692–16911.

[60] Yang, S.; Onuchic, J. N.; Levine, H. *J. Chem. Phys.* **2006**, *125*, 054910–054918.

[61] Chahine, J.; Oliveira, R. J.; Leite, V. B. P.; Wang, J. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 14646–14651.

[62] Oliveira, R. J.; Whitford, P. C.; Chahine, J.; Leite, V. B. P.; Wang, J. *Methods* **2010**, *52*, 91–98.

[63] Yang, S.; Onuchic, J. N.; Garcia, A. E.; Levine, H. *J. Mol. Biol.* **2007**, *372*, 756–763.

[64] Sangha, A. K.; Keyes, T. *J. Phys. Chem. B.* **2009**, *113*, 15886–15894.

[65] Oliveira, R. J.; Whitford, P. C.; Chahine, J.; Wang, J.; Onuchic, J. N.; Leite, V. B. P. *Biophys. J.* **2010**, *99*, 600–608.

[66] Xu, W. X.; Lai, Z. Z.; Oliveira, R. J.; Leite, V. B. P.; Wang, J. *J. Phys. Chem. B* **2012**, *116*, 5152–5159.

[67] Ferreon, J. C.; Hilser, V. J. *Protein Science* **2003**, *12*, 982–996.

[68] Frise, A. F.; Dvinskikh, S. V.; Ohno, H.; Kato, T.; Furo, I. *J. Phys. Chem. B* **2010**, *114*, 15477–15482.

[69] Smithi, B. A.; Clark, W. R.; McConnell, H. M. *Proc. Natl. Acad. Sci. USA* **1979**, *76*, 5641–5644.

[70] Wang, J.; Chu, X. K.; Wang, Y.; Hagen, S.; Han, W.; Wang, E. K. a. *PLOS Comp. Biol.* **2011**, *7*, e1001118.

[71] Wang, Y.; Gan, L.; Wang, E. K.; Wang, J. *J. Chem. Theory Comput.* **2013**, *9*, 84–95.

[72] Wang, Y.; Chu, X.; Longhi, S.; Roche, P.; Han, W.; Wang, E.; Wang, J. *Proc. Natl. Acad. Sci. USA* **2013**, *DOI:10.1073/pnas.1308381110.*

[73] Dettmann, C. P.; Cohen, E. G. D. *J. Stat. Phys.* **2000**, *101*, 775.

[74] Liu, P.; Harder, E.; Berne, B. J. *J. Phys. Chem. B* **2004**, *108*, 6595–6602.

[75] Lai, Z. Z.; Lu, Q.; Wang, J. *J. Phys. Chem. B* **2011**, *115 (14)*, 4147–4159.

[76] Wang, J.; Zhang, K.; Lu, H. Y.; Wang, E. K. *Biophys. J.* **2005**, *89*, 1612–1620.

[77] Wang, J.; Zhang, K.; Lu, H. Y.; Wang, E. K. *Phys. Rev. Lett.* **2006**, *96*, 168101.

[78] Faccioli, P. *J. Chem. phys.* **2010**, *133*, 164106.

[79] Van Kampen, N. G. *Stochastic Processes in Physics and Chemistry.*; Elsevier, 2007; pp 292–322.

[80] Faccioli, P.; Sega, M.; Pederiva, F.; Orland, H. *Phys. Rev. Lett.* **2006**, *97*, 108101.

[81] Wang, J.; Zhang, K.; Wang, E. K. *J. Chem. phys.* **2010**, *133*, 125103.

[82] Schuler, B.; Eaton, W. a. *Curr Opin Struct Biol.* **2008**, *18*, 16–26.

[83] Chung, H.; McHale, K.; Louis, J. M.; Eaton, W. A. *Science* **2012**, *335*, 981–984.

[84] Wang, Y.; Tang, C.; Wang, E.; Wang, J. *PLos Comp.. Biol.* **2012**, *8(4)*, e1002471.

[85] Wang, J.; Onuchic, J. N.; Wolynes, P. G. *Phys. Rev. Lett.* **1996**, *76*, 4861.

[86] Kiefhaber, T. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 9029.

[87] Dyson, H. J.; Wright, P. E. *Chem. Rev.* **2004**, *104*, 36073622.

[88] Xiang, Y.; Duan, L. L.; Zhang, J. Z. H. *Phys. Chem. Chem. Phys.* **2010**, *12*, 15681–15688.

[89] Moran, A. M.; Park, S. M.; Dreyer, J.; Mukamel, S. *J. Chem. Phys.* **2003**, *118*, 3651–3659.

[90] Chung, H. S.; Ganim, Z.; Jones, K. C.; Tokmakoff, A. *Proc. Natl. Acad. Sci.* **2007**, *104*, 14237–14242.

[91] Wang, L.; Middleton, T.; Zanni, M. T.; Skinner, J. L. *J. Phys. Chem. B* **2011**, *115*, 3713–3724.

[92] Backus, E. H. G.; Bloem, R.; Donaldson, P. M.; Ihalainen, J. A.; Pfister, R.; Paoli, B.; Caflisch, A.; Hamm, P. *J. Phys. Chem. B* **2010**, *114*, 3735–3740.

[93] Smith, A. W.; Lessing, J.; Ganim, Z.; Chunte, S. P.; Tokmakoff, A. *J. Phys. Chem. B*

[94] Bagchi, S.; Falvo, C.; Mukamel, S.; Hochstrasser, R. M. *J. Phys. Chem. B* **2009**, *113*, 11260–11273.

[95] Marai, C. N. J.; Mukamel, S.; J., W. *PMC Biophysics* **2010**, *3:8.*

[96] Demirdoven, N.; Cheatum, C. M.; Chung, H. S.; Khail, M.; Knoester, J.; Tokmakoff, A. *J. Am. Chem. Soc.* **2004**, *126*, 7981–7990.

[97] Ganim, Z.; Chung, H.; Smith, A.; DeFlores, L.; Jones, K.; Tokmakoff, A. *Acc. Chem. Res.* **2008**, *41*, 432–441.

[98] Roy, S.; Jansen, T. L. C.; Knoester, J. *Phys. Chem. Chem. Phys.* **2010**, *12*, 9347–9357.

[99] Zanni, M. T.; Gnanakaran, S.; Stenger, J.; Hochstrasser, R. *J. Phys. Chem. B* **2001**, *105*, 6520–6535.

[100] Zhuang, W.; Abramavicius, D.; Mukamel, S. *Proc. Natl. Acad. Sci.* **2006**, *103*, 18934.

[101] Jansen, C.; Dijkstra, A. G.; Watson, T. M.; Hirst, J. D.; Knoester, J. *J. Chem. Phys.* **2006**, *125*, 044312.

[102] Finkelstein, I. J.; Zheng, J. R.; Ishikawa, H.; Seongheun Kim, S.; Kwak, K.; Fayer, M. D. *Phys. Chem. Chem. Phys.* **2007**, *9*, 1533–1549.

[103] Ahmed, Z.; Beta, I. A.; Mikhonin, A. V.; A., A. S. *J. Am. Chem. Soc.* **2005**, *127*, 1094310950.

[104] Halabis, A.; Zmudzinska, W.; Liwo, A.; Oldziej, S. *J. Phys. Chem. B* **2012**, *116*, 6898–6907.

[105] Neidigh, J. W.; Fesinmeyer, R. M.; Andersen, N. H. *Nat. Struct. Biol.* **2002**, *9*, 425–430.

[106] Gellman, S. H.; Woolfson, D. N. *Nat. Struct. Biol.* **2002**, *9*, 408–410.

[107] Hornak, V.; Abel, R.; Okur, A.; Strockbine, A., B. andRoitberg; Simmerling, C. *Proteins* **2006**, *65*, 712725.

[108] Onufriev, A.; Bashford, D.; Case, D. *Proteins* **2004**, *55*, 383394.

[109] Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. *J. Comput. Phys.* **1977**, *23*, 327341.

[110] Abramavicius, D.; Palmieri, B.; Voronine, D.; Sanda, F.; Mukamel, S. *Chem. Rev.* **2009**, *109*, 2350–2408.

[111] Chernyak, V.; Zhang, W.; Mukamel, S. *J. Chem. Phys.* **1998**, *109*, 9587.

[112] Zhang, W.; Chernyak, V.; Mukamel, S. *J. Chem. Phys.* **1999**, *110*, 5011.

[113] Mukamel, S. *Annu. Rev. Phys. Chem.* **2000**, *51*, 691–729.

[114] Zhuang, W.; Abramavicius, D.; Hayashi, T.; Mukamel, S. *J. Phys. Chem. B* **2006**, *110*, 3362–3374.

[115] Neria, E.; Fischer, S.; Karplus, M. *J. Chem. Phys.* **1996**, *105*, 1902.

[116] Humphrey, W.; Dalke, A.; Schulten, K. *J. Molec. Graphics* **1996**, *14*, 33–38.

[117] Bendova-Biedermannova, L.; Hobza, P.; Vondrasek, J. *Proteins* **2008**, *72*, 402–413.

[118] Culik, R. M.; Serrano, A. L.; Bunagan, M. R.; Gai, F. *Angewandte Chemie-International Edition* **2011**, *50*, 10884–10887.

[119] L., Z.; P., N. K.; J., J.; M., S.; W., J. *J. Phys. Chem. Lett.* **2013**, *4*.

[120] Verbaro, D.; Ghosh, I.; Nau, R., W. M.and Schweitzer-Stenner *J. Phys. Chem. B* **2010**, *114*.

[121] Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. *Science* **2010**, *330*, 341–346.

[122] Bowman, G.; Volez, V.; Pande, V. S. *Curr. Opin. Struct. Biol.* **2011**, *21*, 4–11.

[123] Wolynes, P. G.; Onuchic, J. N.; Thirumalai, D. *Science* **1995**, *267*, 1619–1620.

[124] Ostroumov, E. E.; Mulvaney, R. M.; Cogdell, R. J.; Scholes, G. D. *Science* **2013**, *340*, 52.

[125] Mukamel, S.; Abramavicius, D.; Yang, L.; Zhuang, W.; Schweigert, I. V.; Voronine, D. *Acc. Chem. Res.* **2009**, *42*, 553–562.

[126] Chung, H. S.; Ganim, Z.; Jones, K. C.; Tokmakoff, A. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 14237–14242.

[127] Nuernberger, P.; Selle, R.; Langhojer, F.; Dimler, F.; Fechner, S.; Gerber, G.; Brixner, T. *J. Opt. A: Pure Appl. Opt.* **2009**, *11*, 085202.

[128] Wasmer, C.; Lange, A.; van Melckebeke, H.; Siemer, A. B.; Riek, R.; Meier, B. H. *Science* **2008**, *319*, 1523–1526.

[129] West, B. A.; Womick, J. M.; Moran, A. M. *J. Phys. Chem. A* **2011**, *115*, 8630–8637.

[130] Tseng, C.; Matsika, S.; Weinacht, T. *Opt. Express* **2009**, *17*, 18788–18793.

[131] Abramavicius, D.; Jiang, J.; Bulheller, B. M.; Hirst, J. D.; Mukamel, S. *J. Am. Chem. Soc.* **2010**, *132*, 7769–7775.

[132] Hayashi, T.; Zhuang, W.; Mukamel, S. *J. Phys. Chem. A* **2005**, *109*, 9747–9759.

[133] Jiang, J.; Mukamel, S. *Phys. Chem. Chem. Phys.* **2011**, *13*, 2394–2400.

[134] Neidigh, J. W.; Fesinmeyer, R. M.; Andersen, N. H. *nature structural biology* **2002**, *9*, 425–430.

[135] A., D.; M., S. *J. Phys. Chem. B* **2009**, *113*.

[136] Schenkl, S.; van Mourik, F.; van der Zwan, G.; Haacke, S.; Chergui, M. *Science* **2005**, *309*, 917–920.

[137] Liao, H.; Yeh, W.; Chiang, D.; Jernigan, R.; Lustig, B. *Protein Eng. Des. Sel.* **2005**, *18*, 59–64.

[138] Adams, C. M.; Kjeldsen, F.; Patriksson, A.; van der Spoel, D.; Graslund, A.; Papadopoulos, E.; Zubarev, R. A. *J. Mass Spectrom* **2006**, *253*, 263–273.

[139] Jiang, J.; Golchert, K. J.; Kingsley, C. N.; Brubaker, W. D.; Martin, R. W.; Mukamel, S. *J. Phys. Chem. B.* **2013**, *117*, 14294.

[140] Pincus, S. M. *Proc. Nati. Acad. Sci. USA* **1991**, *88*, 2297–2301.

[141] Consani, C.; Aubock, G.; van Mourik, F.; Chergui, M. *Science* **2013**, *339*, 1586.

[142] Nee, M. J.; McCanne, R.; Kubarych, M., K. J.and Joffre *Opt. Lett.* **2007**, *32*, 713–715.