Theoretical Study of the Folding Kinetics of the Ultra-fast Folding Trp-cage Protein

By

Junmin Liu

A thesis submitted to the faculty of graduate studies
Lakehead University
in partial fulfillment of the requirements for the degree of
Masters of Science in Physics

Department of Physics
Lakehead University
August 2006

Your file Votre référence
ISBN: 978-0-494-21517-3
Our file Notre référence
ISBN: 978-0-494-21517-3

Abstract

Native-centric coarse-grained models, termed Cα Go models, have been widely used for computational simulation of small protein to study their folding kinetics and thermodynamics. The limitations of these models come from the lack of non-native interactions and neglect of the specificity of amino acid residues. On the other hand, the simulations of protein folding in atomistic details using accurate energy force field, termed *ab-initio* models, have been proven to be too computationally expensive, even with the most powerful computers. Therefore, many new models at intermediate level have been developed, such as multi-bead model, Go-like model, and all-atom Go model. These improved Go-like models retain some of the specificity of amino acids and more importantly are still able to fold proteins starting from completely unfolded states to their unique native structures.

Our aim is to develop more realistic all-atom Go-like models from a previous all-atom Go model by incorporating specificity to selected structural elements and to non-native interactions. The template for this development is the ultra-fast folding 20-residues Trp-cage protein. We begin by reanalyzing simulation data of the previous highly optimized Go model of the Trp-cage. This reveals three distinct folding pathways: diffusion collision; hydrophobic collapse; and downhill path. In the latter, proteins fold downhill, at ultra-fast speed, toward the native state without having to surmount an entropic barrier. Second, the interaction energies of key structural elements are tuned to examine the folding pathways in detail. This found that the folding pathways taken by the protein are determined by the balance between the stability of the α-helix and the hydrophobicity of the Trp-cage core, while the folding speed is determined by the salt bridge between residues Asp9 and Arg16. Finally, homogeneous and knowledge-base non-native interactions are incorporated into the Go model. For models with homogeneous non-native interactions the folding pathways is steered toward a pseudo-downhill pathway where proteins are trapped in misfolded conformations. In contrast, models using knowledge-based non-native interactions observe enhanced downhill folding, leading, in some cases, to a modest increase in the folding speed.

Acknowledgement

I am extremely grateful for the successes I have been able to enjoy while receiving my academic training at the Lakehead University. These successes are only possible due to Dr. Linhananta who has mentored and trained me over the last two years. Dr. Linhananta has been not only a great adviser of my thesis but also a good friend, keeping me focused on important scientific questions, finding a Ph. D program and searching a potential job.

The Department of Physics has excellent environment to do science, and the persons in this department are very helpful to graduate student, including Dr. Gallagher, Dr. Hawton, Dr. Keeler, Dr. Das, Dr. deGuise, Mr. Lachaine and Ms Carey.

Of course, I benefited from countless close interactions with students in the Department of Physics, including Ian, Wei, Laura, Ben, Dylan, and Jesse.

Finally, I wish to thank the professor, Dr. Malek, Dr. Leung and Dr. Mallik, in Department of Biology.

Contents

List of Graphs

List of Tables

Chapter 1

Introduction to protein folding problem

Proteins play a key role in almost all biological processes. They take part in, for example, maintaining the structural integrity of the cell, transport and storage of small molecules, catalysis, regulation, signaling and maintenance of the immune system. In biological systems, newly synthesized protein molecules start out in linear random coil structures, but they quickly fold to unique compact native structures. Their abilities to perform their intended functions rest on their folding to these unique native structures. As a result, there have been many efforts, both experimental and theoretical, in determining their native structures, and understanding their folding mechanisms.

1.1 Protein structure and orientation conformations

1.1.1 Primary, secondary, and tertiary structure

Proteins are heteropolymers made up of 20 molecular units, which called amino acids (Branden and Tooze 1999; Petsko and Ringe 2004). All amino acids have the same backbone structure (also called main chain), and are distinguished by the side chain R (see Fig. 1.1). They are usually referred to by acronyms or alphabets (Fig. 1.2). *In vivo* (in nature) and *in vitro* (in the lab) amino acids form peptide bonds to form amino acid sequences called polypeptides. Proteins are amino acid sequences that have been optimized, by natural evolution, to fold to specific structures and to perform specific biological functions. The amino acid sequence (sequence of alphabets, Fig. 1.2) of a protein is referred as the protein's **primary structure** (Fig. 1.3(a)). Proteins readily form ***secondary structure***, (Fig 1.3(b)) stabilized by backbone hydrogen bonds between N–H and C=O groups, which are either alpha helices (Fig. 1.4) or beta sheets (Fig. 1.5). In a folded protein the secondary structure elements are packed to form a specific compact globular structure called the ***tertiary structure*** (Fig. 1.3c). In these form, a protein is said to be folded into its unique **native structure**.

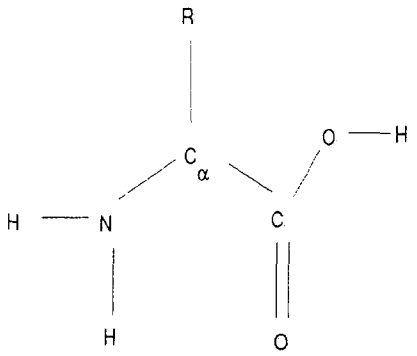


Figure 1.1 The chemical structure of an amino acid. The backbone is the same for all amino acids and consists of the amino group (NH2), the alpha carbon and the carboxylic acid group (COOH). Different amino acids are distinguished by their different side chains, R.

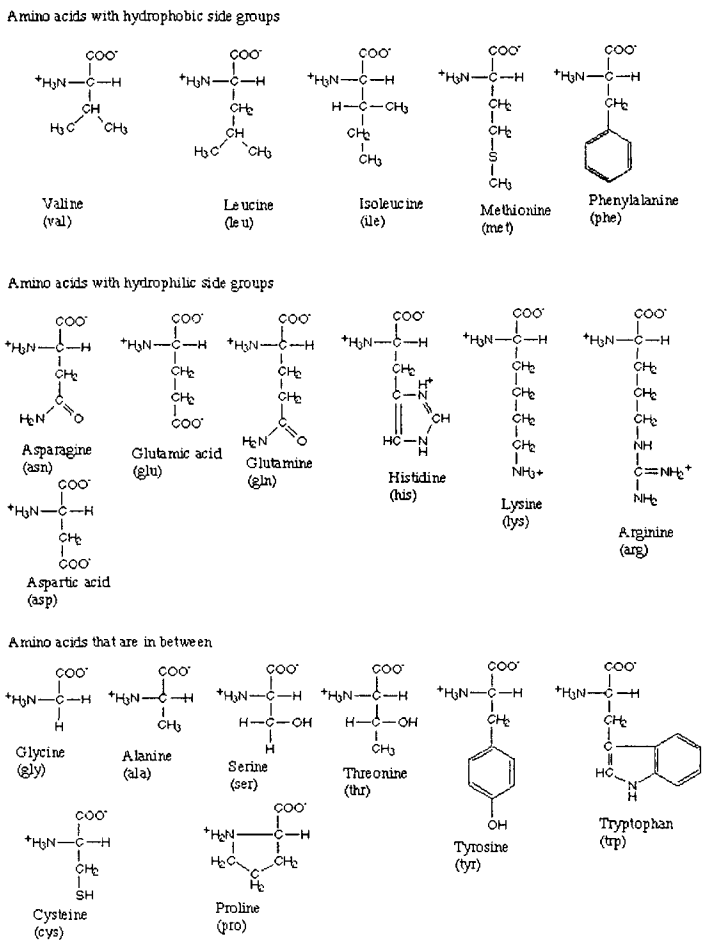


Figure 1.2 **Amino-acid structure and the chemical characters of the amino-acid side chains**

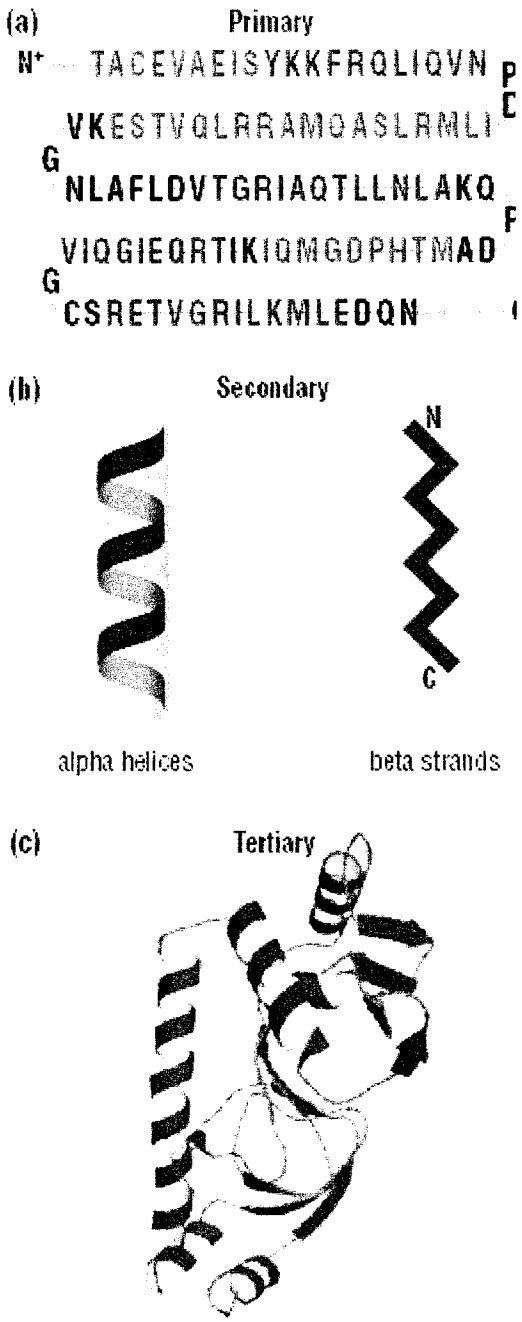


Figure 1.3 **Levels of protein structure illustrated by the catabolite activator protein (a)** The amino-acid sequence of a protein (primary structure) contains all the information needed to specify **(b)** the regular repeating patterns of hydrogen-bonded backbone conformations (secondary structure) such as alpha helices (red) and beta sheets (blue), as well as **(c)** the way these elements pack together to form the overall fold of the protein (tertiary structure) (PDB 2cgp). (adopted from Petsko and Ringe 2004)

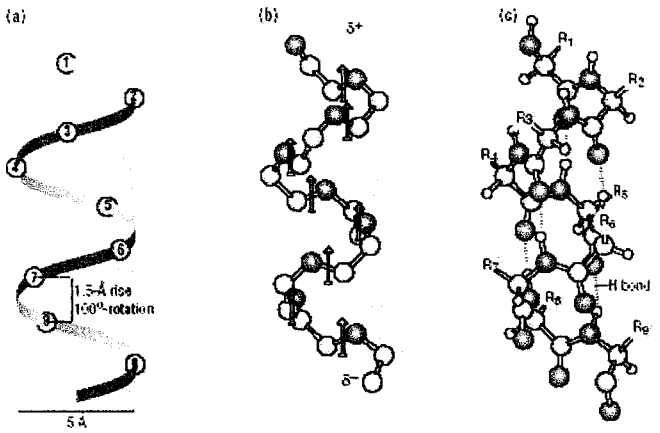


Figure 1.4. The **alpha helix** The chain path with average helical parameters is indicated showing **(a)** the alpha carbons only, **(b)** the backbone fold with peptide dipoles and **(c)** the full structure with backbone hydrogen bonds in red. All three chains run from top to bottom (that is, the amino-terminal end is at the top). Note that the individual peptide dipoles align to produce a macrodipole with its positive end at the amino-terminal end of the helix. Note also that the amino-terminal end has unsatisfied hydrogen-bond donors (N–H groups) whereas the carboxy-terminal end has unsatisfied hydrogen-bond acceptors (C=O groups). Usually a polar side chain is found at the end of the helix, making hydrogen bonds to these donors and acceptors; such a residue is called a helix cap. (adopted from Petsko and Ringe 2004)

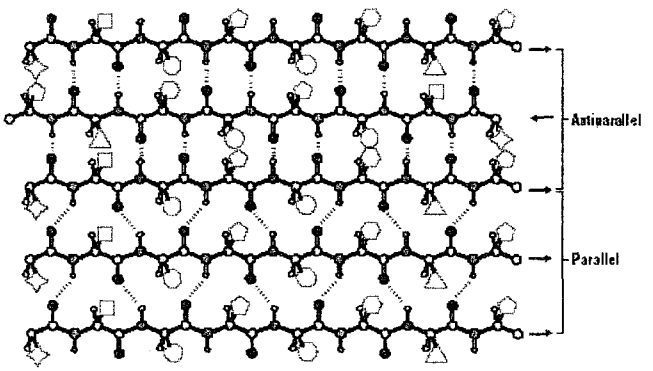


Figure 1.5 The structure of the **beta sheet** The left figure shows a mixed beta sheet, that is one containing both parallel and antiparallel segments. Note that the hydrogen bonds are more linear in the antiparallel sheet. On the right are edge-on views of antiparallel (top) and parallel sheets (bottom). The corrugated appearance gives rise to the name "pleated sheet" for these elements of secondary structure. Consecutive side chains, indicated here as numbered geometric symbols, point from alternate faces of both types of sheet. (adopted from Petsko and Ringe 2004)

1.1.2 Rotational conformations, the Ramachandran plot, and the Levinthal's paradox

In 1968, Ramachandran (Ramachandran et al. 1963) showed that since the bond length between atoms comprising the backbone of a polypeptide is effectively constant (Fig. 1.1), the only conformational freedom available to proteins is the rotation of the side chains about the backbone structure. This is illustrated in Fig. 1.6, where the orientations of a side chain of an amino acid with respect to the backbone are classified by the dihedral angles ψ (psi) and ϕ (phi). Experimental works showed that amino acid tends to have conformational preferences. This is summarized by the Ramachandran plot (Fig. 1.7) which displays the observed ψ and ϕ backbone conformational angles of proteins. Particular noteworthy are the class of amino acids (such as Ala, Glu, Leu and Met) with the range $\psi \sim -60°$ to $-80°$ (or $\phi \sim 60°$ to $80°$) characteristic of α-helix, and those (Val, Ile, Tyr, Cys) with extended range $\psi \sim 100°$ to $180°$ and $\phi \sim -90°$ to $90°$ characteristic of β-sheet (Branden and Tooze 1999; Petsko and Ringe 2004).

This conformational freedom is the basis of **Levinthal's paradox**, named after the French biophysicist C. Levinthal (Levinthal 1968) which is a crude calculation that concluded that a protein in an extended random coil state would take to the order of $\sim 10^{50}$ years to fold to its unique native state (Ploktin and Onuchic 2002a). That calculation made the simplifying, but reasonable, assumptions that an amino acid can assume two conformations and that it takes about one picoseconds to interconvert between conformations. But Levinthal also made the extreme assumption that the search to locate this unique native state is a random search. Later we shall explain that this assumption is erroneous, and describe the resolution of this paradox.

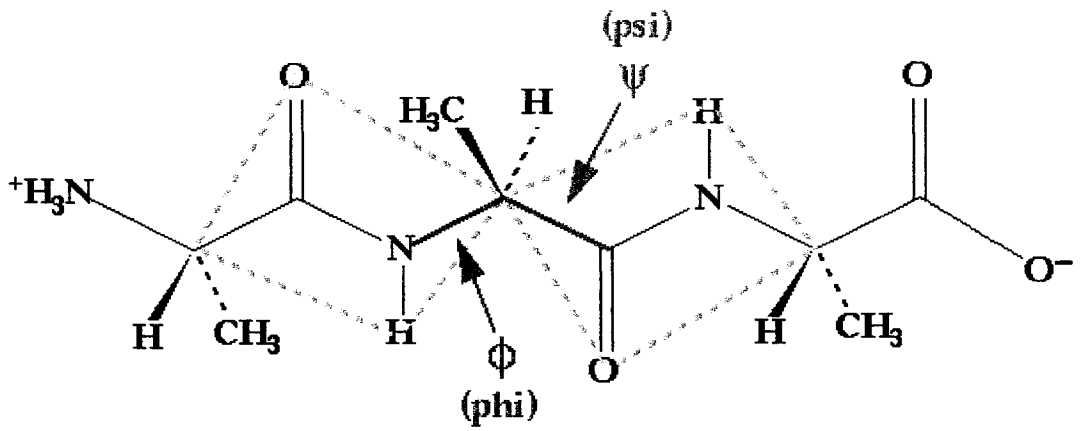


Figure 1.6 A tripeptide of alanine. Note that the peptide bonds on either side of the central alpha carbon act to create rigid plates which rotate about phi and psi. (adopted from Petsko and Ringe 2004)

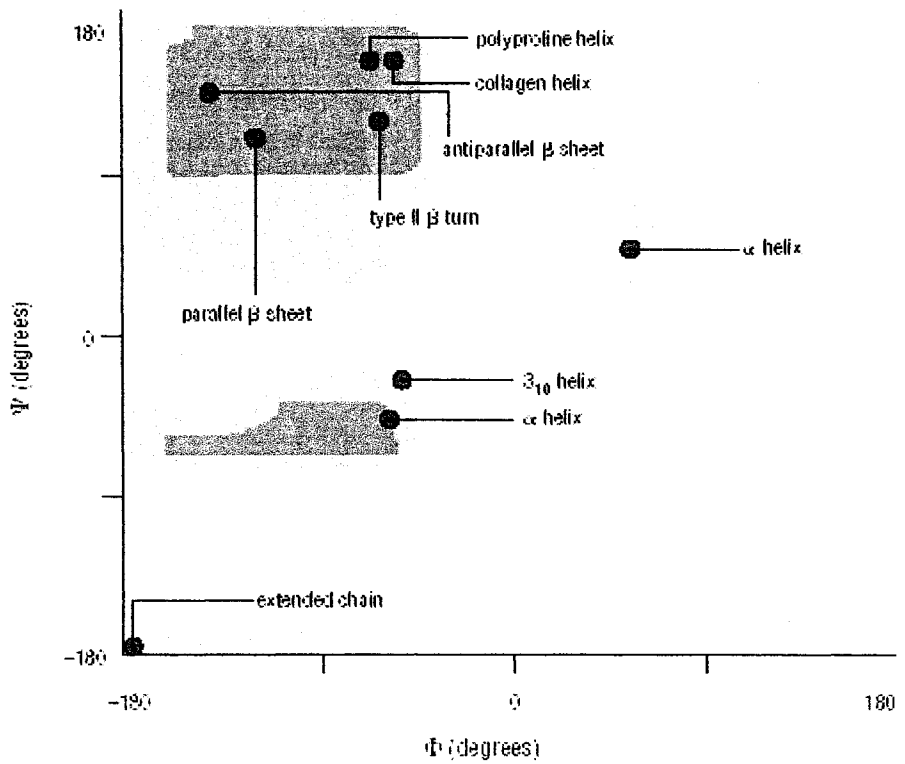


Figure 1.7 A Ramachandran plot for the tripeptide. The plot is similar to a topographical map, where energy, instead of altitude, is shown with the contours. Surrounding phi=0, psi=0 there is a high energy "plateau" which drops into valleys of stability with minima for alpha helices and beta sheet, noted in the figure. In large proteins, the large majority of non-glycine residue possesses phi, psi combinations that reside in these valleys. (adopted from Petsko and Ringe 2004)

1.2 Protein folding kinetics

1.2.1 Main driving forces

In vivo and *in vitro*, proteins exist in aqueous solution of varying pH, often with ions and other macromolecules. It is believed that electrostatic forces arising from charged and polar side chains of proteins as well as from ions are neutralized by water (Dill 1990). Instead, it also is believed that the ***hydrophobic force (effects)*** drives folding. This arises from the tendency of polar water molecules to form strained clathrate networks of hydrogen bonds. As a consequence, in aqueous solutions, nonpolar molecules, tend to aggregate together to avoid interactions with water, minimizing any disruption to the clathrate networks. Hence even though it is often referred to as the hydrophobic force, the term hydrophobic effects is more accurate. As mentioned, an amino acid is classified as hydrophobic if their side chain is nonpolar and as hydrophilic if its side chain is polar or charged. Hydrophilic side chains interact favorably with water molecules, and hence are soluble in water. In the vast majority of proteins whose native structures have been resolved, the side chains of hydrophobic amino acids are packed to form a solidlike hydrophobic core, leaving the side chains of hydrophilic amino acids on the surface of the protein, where they are exposed to water (Fig. 1.3c).

Another important driving force in folding is ***hydrogen bonds*** which, as mentioned, are the main stabilizing factor of secondary structures (Figs.1.4 and 1.5). This usually involves bonding between polar N–H and C=O groups. Two important aspects are that a stable hydrogen bond requires near parallel orientations between the N–H and C=O groups, and that water molecules can also form hydrogen bonds with proteins. The latter process competes with intra-protein backbone hydrogen bonds. For these reasons, secondary structures are not completely stable until the final stages of folding when their stability is increased by tertiary structures. The role of secondary structures in folding is a contentious issue with some researchers believing that they are partially formed at the early stages, while others believing that they are not formed until the final stages (Skolnick 2005). A common technique for probing this issue is to increase

helix stability with the addition of triuorethanol (TFE) and to observe its effect on the overall folding of the protein (Main and Jackson 1999). Later a computational method that mimics these experiments will be described.

1.2.2 Conformation entropy and free energy folding funnel

There are two opposing factors in protein folding: energy consideration favors low-energy native state, while entropy favors the high-energy random coil state. The latter is an unfolded state with high conformational entropy related to the high number of accessible backbone and side-chain conformations (Baker 2000). As mentioned, the Levinthal paradox estimated that it would take the order of ~ 10^{50} years for a protein to fold, even though proteins usually fold within a few seconds. The paradox makes the erroneous assumptions that during folding all conformations, including the native conformations, are equally probable, which, given the large number of available conformations (the order of 10^{60}), resulted in the unphysical folding time. The paradox has been resolved by the folding-funnel picture in which proteins (Onuchic and Wolynes 2004), whose primary structures have been optimized by nature, are guided along funnel like pathways, becoming increasingly native-like, towards the native states. More refined theories include a roughened folding-funnel picture in which the funnel-like free-energy landscapes are decorated with metastable intermediate states and unstable transition states.

1.2.3 Folding mechanisms

The ***diffusion-collision model*** or ***framework model*** (Islam et al. 2002) states that local secondary structures (α-helix or β-sheet) form early in the folding pathway and these pre-formed structures diffuse and collide forming late tertiary contacts, or adhesions. Adherents of this model believe that early secondary structure formation is crucial in reducing of the accessible conformations along the folding pathways. The ***hydrophobic collapse model*** (Dill et al. 1993) contends that prior to any secondary structure formation, hydrophobic side chains coalesce forming a nascent core of distant tertiary contacts. Upon this hydrophobic core nucleus, local secondary

structures propagate. These models differ exactly in the temporal order of secondary and tertiary structure formation along the folding pathway. These extremely polarized theories about the nature of the folding pathway are probably overstated, because not all proteins will fold according to the same set of rules, and most likely the actual folding pathways of a protein will include aspects of both theories.

The *Nucleation condensation model* posits that neither pre-formed helix nor excess hydrophobic collapse dominates the early steps of protein folding (Fersht 1995; Itzhaki et al. 1995; Gianni et al. 2003). Instead what limits the process of protein folding is a random search to form the minimal number of specific contacts, which define the transition state ensemble of protein folding. This search is accomplished by a trial-and-error process, wherein partial-folded structures are formed and broken. If, by chance, the partial structure is part of the transition state ensemble, the protein has a 50/50 chance of proceeding downhill towards the folded native state.

1.2.4 Folding rate, folding mechanisms and native topology

A key breakthrough in protein-folding research is the discovery of a mathematical relationship between the topological complexity of a particular protein and its experimentally determined average folding speed (Plaxco *et al.*, 1998). Plaxco *et al.* defined the relative contact order:

$$CO = \frac{1}{L \bullet N} \sum^{N} \Delta S_{i,j} \qquad \text{(Eq. 1.1)}$$

where N is the total number of contracts, $\Delta S_{i,j}$ is the sequence separation, in residues, between contacting residues i and j, and L is the total number of residues in the protein. Since the value of *CO* is large for a protein with many native contacts between far distanced residues ($\Delta S_{i,j}$), it quantifies the topological complexity of a protein structure. It was found that there is a strong correlation between the folding rate of two-state folding proteins and the relative contact order. This led to the proposal that the folding kinetics of a protein is determined by its native structure. This view is consistent with experiments that observe that, with a few exceptions,

different proteins (i.e. proteins with different sequence structures) with the same native structural topology have similar folding mechanisms (Alm and Baker 1999; Baker 2000). This led to the wide use of coarse-grain Cα Go model, where an amino acid residue is represented by a single bead centered at the central backbone carbon (Cα) of the residue, and where the interaction potential is based on the native topology of the protein. A key point is that these models neglect the atomic details of the native structures, which are not believed to play important roles in folding. In a recent landmark study, folding simulations of 18 small proteins using Go-like protein models (Koga and Takada 2001) predict folding kinetics that are similar to those observed in experiments.

The folding-funnel picture provides a partial explanation of why folding mechanisms are dominated by native topology. In funnel-like theory, entropic cost, which is related to the complexity of a protein, is the major contribution to the folding barrier (Baker 2000; Plotkin and Onuchic 2002a). The larger the sequence separation between two residues that are in contact in the native state, the larger the entropic cost for forming that contact. Thus, proteins of simple topologies with mostly local interactions formed more rapidly than those of complex topologies with more non-local interactions (Baker 2000).

1.3 Computational simulation of protein folding by using Go model

Ab-initio computer-simulation models (CHARMM, AMBER, or OPLS) are commonly used to study protein's system (Snow et al. 2005). In these models, all atoms are represented. The "classical" force field of these models accurately represents the atomic bonds, hydrogen bonds, dihedral conformations, electric-static interactions, non-bonded van der Waals interactions as well as the aqueous environments of proteins. However, even with the most powerful parallel computers, it is only feasible to perform *ab-initio* protein simulations to about a few microseconds of protein kinetics. Hence, *ab-initio* models are usually employed to study functional processes that occur on a time-scale of a few nanoseconds, or to the

study of very small fast-folding proteins (Duan and Kollman 1998). For the latter cases, most simulations do not always produce folded proteins, and specialized techniques are usually employed to extract the "theoretical" folding pathways (Snow et al 2002; Pitera and Swope 2003).

Alternatively, we can use simplified model, such as Go model (Go 1983), which assumes that the folding rate and kinetics of a protein are determined by its unique native structure (Takada 1999). This assumption inspired the developments of protein-folding models that employ Go potentials, which bias the conformations of the model proteins to the known native structures of proteins. Though these knowledge-based models are not as accurate as *ab-initio* models, molecular dynamics simulations usually are able to fold a protein a sufficient number of times to be able to extract meaningful statistics on the folding mechanisms of a protein.

The vast majority of Go models are homogeneous Cα models, in which a residue is represented by a single bead and in which the pair-wise interaction energies are the same for all pairs. Staring from native structure, coarse-grained Go-like models have played key roles in advancing the understanding of how proteins fold from extended random coil conformations to their unique native states (Vendruscolo and Paci 2003, Snow et al. 2005). Another important class of Cα models discriminates between hydrophobic **H** and polar **P** residues (Dill et al. 1993). Cα Go models in one, two, and three dimensions are instrumental in the development of key concepts and in the resolution of issues such as the free-energy folding funnels, the two-state cooperativity of small proteins (Kaya and Chan 2003), the interpretation of the Φ-value (Clementi et al. 2000; Ozkan et al. 2001; Koga and Takada 2001), and the controversial diffusion-collision versus condensation-nucleation debate (Dill et al. 1993).

However, Cα Go models are limited by the lack of non-native interactions and the absence of specificity of amino acid residues. To overcome the second limitation,

modified Cα Go models have been constructed in which native contacts interact with residue-dependent potentials (Karanicolas and Brooks 2002; Onuchic and Wolynes 2004; Sutto et al. 2006). An alternative approach is to use an all-atom Go model, where all atoms are represented and the interaction potential energy is based on the detailed all-atom structure of the native state. Clementi et al. constructed an all-atom Go model to study the interplay among tertiary contacts, secondary structure formation and side-chain packing (Clementi et al. 2003). In their model, an attractive Lennard Jones (LJ) potential was assigned to all pairs of atoms with a native-structure state distance less than 4 Å not participating in common bond or bond angle terms. A repulsive LJ interaction was introduced to all atom pairs not participating in native state (non-native atom-pair interactions). Their simulation results reproduced different folding mechanisms of protein G and L, which are proteins with different primary structures but identical native-state topology. Shakhnovich and co-workers used Monte Carlo simulations of their all-heavy-atom models to probe the folding of Crambin and to verify the transition state of chymotryspin inhibitor 2 (Shimada and Shakhnovich 2002).

Another class of all-atom Go models based on discontinuous square-well interactions has also been constructed (Zhou and Linhananta 2002; Linhananta and Zhou 2002; Linhananta et al. 2002, 2005). In contrast to the model of Clementi et al., all native atom-atom interactions are of equal strength, and the atomic parameters are scaled to van der Waals parameters. Zhou and Linhananta used discontinuous molecular dynamics (DMD) to probe the folding of an all-atom model of the second β-hairpin fragment of protein G. This yielded a hydrophobic collapse folding mechanism consistent with other MD simulations based on established energy force field in implicit or explicit solvents. More impressively, the model predicted that the collapse is initiated by two specific hydrophobic and hydrophilic contacts, in agreement with a nuclear magnetic resonance (NMR) experiment (Wolynes 2004).

The positive results of these all-atom Go models highlight the importance of detailed

side-chain packing and atomic-level contacts in folding mechanisms. The importance of explicit side-chain representation was recently assessed by Karanicolas and Brooks, who found that Ising-like models of proteins without explicit side chains can lead to energy landscapes that differ significantly from those obtained from higher resolution models (Karanicolas and Brooks 2003a).

1.4 The aim of this work

As mentioned, the main advantage of Go-like models is their ability to fold model proteins a sufficient number of times to allow meaningful analyses of their kinetic behaviors. In this thesis an all-atom Go-like model of the ultra fast folding mini protein Trp-cage is constructed, and multiple folding simulations will be performed. The major aims include: 1) revealing its multi-folding-pathways; 2) investigating the folding kinetics by tuning the relative stability of different structural elements; 3) and examining the role of non-native interactions in small protein folding process:

1) Revealing the multi-folding-pathways. The structural features of the unfolded states and metastable folding intermediates will be identified using the commonly used cluster analysis method (Shortle et al. 1998, Betancourt et al. 2001, Zhang et al. 2004), as well as by the time-evolution of reaction-coordinate parameters (Zhou and Karplus 1999b). This will allow the classifications of trajectories into different "average" pathways, and to detect commonly occurring early events in folding. It is noteworthy that the structures of proteins during the initial stage of folding are very difficult to detect by experiments, and computer simulations are often the only way to link unfolded states with folded states.

2) Tuning the relative stability of structural elements by constructing a heterogenous all-atom Go model. In previous homogenous all-atom Go models, all atomic-pair native interactions have the same strength (Linhananta et al. 2005). Here, a new variable, R, is introduced to represent the relative

stability of specific structural elements:

$$R = U_{ij} / B_{ij}^{Go} \qquad \text{(Eq. 1.2)}$$

Where $B_{ij}^{Go} = -1$ for all atomic-pair native interactions, $B_{ij}^{Go} = 0$ for non-native contacts. For $R > 1$, the structural elements of interest will be strengthened while for $R < 1$ they will be weakened. For protein Trp-cage, the relative stability of α-helix, hydrophobic-core, salt-bridge and Trp6-Pro12 interactions will be systematically tuned. These simulation results will reveal the effects of the coupling between secondary and tertiary structures on protein folding kinetics. These effects are usually studied by experimentalists using point–mutation and varying the solvent environments (Nerweiler et al. 2005). To our knowledge, this is the first computational study of these effects on the Trp-cage.

3) The role of non-native interactions in protein folding processes. Go models are often criticized because their interaction potentials do not include non-native interactions. Unlike native interactions which bias proteins to their folded structures, non-native interactions can lead to misfolded protein conformations. Different non-native interactions potentials, which include homogenous repulsive, attractive non-native potential and knowledge-based (specific) non-native interactions, are used to detect the effect of these interactions on folding rate and the stability of native states, transition state ensemble and unfolded states.

Chapter 2

All atom Go model of Trp-cage, Discontinuous Molecular Dynamics Algorithms, and Thermodynamics Quantities

2.1 Trp-cage protein

The Trp-cage (Tc5b) protein with the amino acid sequence (i.e. its primary structure) NLYIQWLKDG GPSSGRPPPS was first synthesized by Anderson group (Neidigh et al. 2002). Its primary structure is based on the naturally occurring EX4 protein, whose structure is stabilized mainly by a hydrophobic core. By selective mutations (i.e. amino acid sequence alterations) of EX4, the hydrophobic core was strengthened and the helical tail shortened. The result is the Trp-cage, which at 20 residue long is one of the smallest proteins that can fold to stable unique native structures *in vivo* or *in vitro*. The hallmark of this protein is the hydrophobic core, with the residue Trp6 at the center of the cage formed by 4 proline-residues (Pro12, 17, 18, 19). The secondary structure of Trp-cage protein consists of a short α-helix from residue 2 through 8, a 3_{10}-helix from residue 11 through 14, and a C-terminal polyproline (PPII) helix that packs against the central tryptophan (Trp6). In addition a salt-bridge between Asp9 and Arg16 is an important stabilizing factor of the folded state. A temperature jump (T-jump) experiment by Qiu et al. has determined the Trp-cage's folding time to be 4 μs (Qiu et al. 2002), making it one of the fastest folding proteins. Nuclear magnetic resonance (NMR) and circular dichorism (CD) suggests that it fold by a two-state folding mechanism, where only stable folded or unfolded states are detectable. (Neidigh et al. 2002). However, recently, experimental results of UV-resonance Raman spectroscopy (Ahmed et al. 2005) and fluorescence correlation spectroscopy (FCS) (Neuweiler et al. 2005) observed residual helical structures in the denatured state, as well as metastable intermediate states during the folding process.

The small size and ultra fast folding speed (4μs) of the Trp-cage protein makes it an ideal model for folding simulations, and to date there has been numerous studies using

ab-intio and simplified models (Slimmerling et al. 2002; Snow et al. 2002; Zhou R. 2003; Nikiforovich et al. 2003; Pitera et al. 2003; Chowdhury et al. 2003 & 2004; Linhananta et al. 2005). To achieve the aims (as stated in chapter 1) of this thesis we will construct all-atom discontinuous molecular dynamics (DMD) models of the Trp-cage.

2.2 All atom Go model of Trp-cage

2.2.1 All-atomic representation

The general setup of the all-atom Go model can be found in many papers (Linhananta and Zhou 2002; Zhou and Linhananta 2002; and Linhananta et al. 2002). The initial heavy-atom positions of Trp-cage were obtained from NMR structures (Structure 1 of PDB ID 1L2Y). The NH2 terminus (Asn-1) and N-methyl COOH terminus (Ser-20) were capped with acetyle and amine groups, respectively. The initial positions of polar hydrogen were generated by CHARMM , and the structure was minimized for 100 steps, with fixed heavy-atom positions, by the steepest descend method using polar hydrogen parameter set 19 with distance-dependent dielectric constant. The total number of atom is 189.

2.2.2 Hard sphere potential (discontinuous model)

All heavy atoms and polar hydrogens are represented by hard spheres. Two bonded atoms *i* and *j*, as well as any 1, 3 angle-constrained pair and 1, 4 aromatic carbon pair, are constrained to a center-to-center distance between $0.9\sigma_{ij}$ and $1.1\sigma_{ij}$, where σ_{ij} is the separation of the *i, j* pair in the CHARMM (Brooks et al. 1983) minimized structure. This constraint is accomplished by an infinitely deep square-well potential

$$u_{ij}^{bond} = \begin{cases} \infty, r < 0.9\sigma_{ij} \\ 0, 0.9\sigma_{ij} < r < 1.1\sigma_{ij} \\ \infty, r > 1.1\sigma_{ij} \end{cases} \qquad \text{Eq. (2.1)}$$

As in previous works (Linhananta and Zhou 2002; Zhou and Linhananta 2002;

Linhananta et al. 2002, 2005) a bond flexibility parameter of 0.1 is used. It has been shown in a previous study that the variation of this parameter does not affect the folding mechanism (Zhou and Karplus 1999a). The model also includes a discontinuous improper dihedral potential (Brooks et al. 1983) to maintain chirality about tetrahedral heavy atoms and certain planar atoms. The potential has the form

$$u_{\omega}^{improp} = \begin{cases} \infty, \omega > \omega_0 + 20^\circ \\ 0, \omega_0 - 20^\circ < \omega < \omega_0 + 20^\circ \\ \infty, \omega > \omega_0 - 20^\circ \end{cases} \quad \text{(Eq. 2.2)}$$

A 20° angle flexibility is used to decrease the folding time of the protein. It has been shown previously (Linhananta and Zhou 2002) that a smaller flexibility value increases the folding time, but does not affect the folding mechanism. The improper angles v are obtained from CHARMM potential set 19. In Equation ω_0=35.26439° for chiral-constrained atoms such as a carbon without explicit hydrogen and ω_0=0° for planar-constrained atoms (such as a carbonyl carbon). The improper dihedral potentials u_{ω}^{impro} preserve the *L*-form chirality of amino acids and mimic some of the rigidity of peptide units. It was introduced to eliminate "unphysical" local misfold conformations in our all-atom model of BpA (Linhananta and Zhou 2002), which can prevent folding to the native state. A nonbonded *i*, *j* pair interacts by a hard-core and square-well potential

$$u_{ij}^{Go} = \begin{cases} \infty, r < 0.8\sigma_{ij}^{vdW} \\ B_{ij}\varepsilon, 0.8\sigma_{ij}^{vdW} < r < 1.2\sigma_{ij}^{vdW} \\ 0, r > 1.2\sigma_{ij}^{vdW} \end{cases} \quad \text{(Eq. 2.3)}$$

where σ_{ij}^{vdw} are the van der Waals diameters from the CHARMM polar hydrogen parameter set 19 and *Bij* are the interaction strengths. The factor of 0.8 for hard-core diameters is typical for ratio between the diameter of a hardsphere reference system and the van der Waals diameter found in the Weeks–Chandler–Anderson perturbation

theory (Weeks et al. 1971), while a ratio of 1.5 between the square-well and hard-core diameters is typical for systems of small molecules (Sherwood and Prausnitz 1964).

2.2.3 Homogenous Go potential of protein Trp-cage

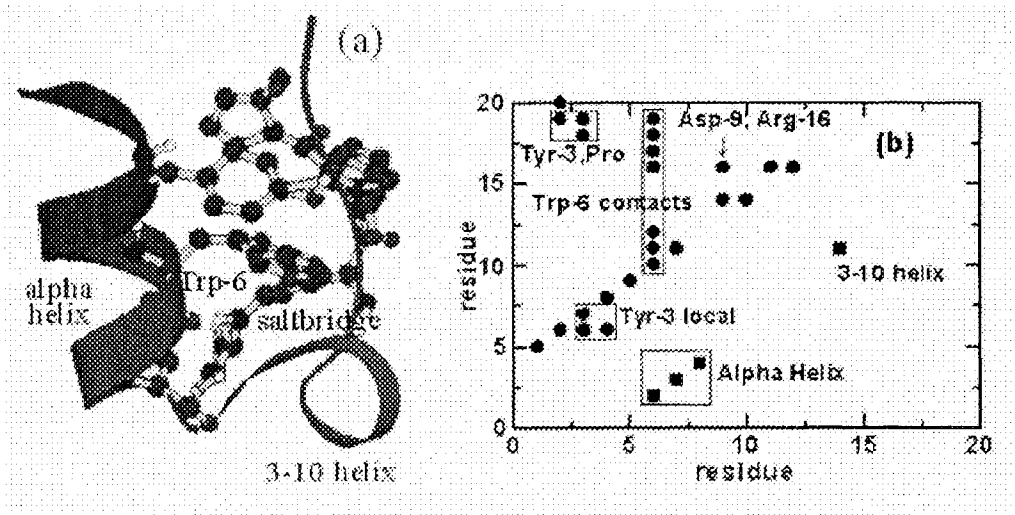


FIG. 2.1. (a) Ribbon representation of the global minimum structure of an all-atom off-lattice model of fragment *B* of the Trp-cage. Residues Tyr-3, Trp-6, Asp-9, Arg-16, Pro-17, Pro-18, and Pro-19 are shown in atomic details. The following features are labeled: ahelix, residue Trp-6, the Asp-9-Arg-l6 salt bridge, and the 310 helix. Drawn using molscript (Kraulis 1991) (b) The native residue-residue contact map of the model Trp-cage. The native tertiary contacts and native secondarystructure hydrogen bonds are above and below the diagonal, respectively. A residue-residue pair is in contact if there is at least one square-well atomic contact between them. (Adopted from Reference Linhananta et al.2005)

Initial hard-core overlaps in the CHARMM minimized structure are removed by a short DMD simulation, where the square-well interaction in Eq. 2.3 is set to large negative values ,-100ε. The deep square wells preserve NMR native contacts as overlapping contacts are removed. The initial hard-core diameters between any two overlapping atoms are set to their separation distances in the energy minimized structure. The hard-core diameters are adjusted at each time step until the true hard-core diameters are reached and the simulation is continued until all overlaps are removed. The resulting structure, shown in Fig. 2.1(a), has a root-mean-square deviation (RMSD) from the NMR structure of 0.65 Å, and is the global minimum structure of the Trp-cage model.

The main features of the Trp-cage are the short a-helix in residues 2–8, the 3_{10} helix in residues 11–14, and the Tyr-3 and three *C*-terminal prolines sresidues (17–19) that pack against the central tryptophan, Trp-6. Also visible in Fig. 2.1(a) is the salt bridge between residues Asp-9 and Arg-16. In all there are 1267 native square-well atom-atom contacts, which include both side-chain and backbone contacts. Figure 1(b) shows the native state contact map that highlights the main structural features. Tertiary contacts between residue pairs (defined for i, j residue pair with $j \geq i+4$) and the Tyr-3-Trp-6 contact are drawn above the diagonal, while hydrogen bond contacts are drawn below the diagonal. The α-helix is stabilized by hydrogen bonds between residue pairs Leu-2 and Tyr-6 sH1d, Tyr-3 and Leu-7 (H2), and Ile-4 and Lys-8 (H3). An α-helix (3_{10}-helix) hydrogen bond is defined, as in previous studies (Linhananta and Zhou 2002; Zhou and Linhananta 2002), by the cutoff distance of 2.88 Å between a main-chain carbonyl oxygen of residue i and anamide hydrogen of residue $i+4$. This definition does not include the orientational constraint used in some all-atom MD studies (Gsponer and Caflish 2001; Tsai et al. 1999). The orientational constraint of hydrogen bonds is difficult to implement in discontinuous models (Smith and Hall 2001), and its use has been shown to significantly compromise the efficiency of DMD simulations (result not shown). This model, just like previous all-atom G*o* models (Linhananta and Zhou 2002; Zhou and Linhananta 2002; Linhananta et al. 2002), does not include hydrogen bond orientations, but instead, is focused on the roles of side-chain packing and detailed atomic contacts. However, overlaps between amide hydrogens and carbonyl oxygens are still valid indicators of secondary structure formation, since such contacts would not be possible without secondary structures. This point will be illustrated explicitly in later sections.

A Go potential is employed to bias the energy of the model protein to the global minimum structure in Figure 2.1. Atomic i, j pairs with square-well overlaps in the gloal minimum structure are designated as native contacts. For these native pairs, a square-well overlap between two atoms results in an interaction energy $Bij=-1$. All other pairs are designated as non-native, and the overlap energies are set to zero $Bij=0$

in Chapter 3 and Chapter 4. This differs from many Cα and all-atom Go models (Shimada et al. 2001; Karanicolas and Brooks III 2002 and 2003; Clementi et al. 2000, 2001, and 2003), in which non-native interactions are repulsive. The choice not to include repulsive non-native interactions is consistent with recent studies that concluded that non-native interactions can be attractive and may assist the folding process (Plotkin 2001; Paci et al. 2002; Clementi 2004). But, in Chapter 5, different kinds of non-native potential will be introduced the all-atom Go model of Trp-cage.

2.3 Discontinuous Molecular Dynamics Algorithms

Molecular dynamics simulation algorithms for chains interacting with discontinuous potentials such as hard-sphere and square-well potentials are different from those for chains interacting with soft potentials such as LJ interactions. Unlike soft potentials, discontinuous potentials only exert forces when particles collide. The binary collision dynamics for discontinuous potentials can be solved exactly. Thus, the DMD algorithm (Alder and Wainwright 1959; Rapaport 1980) involves searching for the next collision time and collision pair, moving all beads for the duration of the collision time, and then calculating the velocity changes of the colliding pair.

Molecular dynamics simulations are often performed in a micro-canonical ensemble, that is, an ensemble containing a constant number of particles, constant total energy, and constant volume. The details of the use of the DMD algorithm for the case of a square-well chain in a micro-canonical ensemble can be found in a book written about "Computer Simulation of Liquids" (Allen and Tildesley 1987). In a constant energy ensemble, however, a short isolated chain is not an argotic system at low energies because the chain can be trapped permanently in a low-energy configuration if its kinetic energy is not high enough to overcome the energy barrier. This problem may be remedied by placing the isolated chain in a constant-temperature bath. The collision between the bath particles and the chain can help the chain to get out of low energy traps. For this reason, we simulate the isolated square-well chain in a

''canonical'' ensemble, that is, an ensemble with constant number of particles, N, and constant temperature, T; for an isolated protein molecule, there is no volume constraint.

Current constant-temperature MD techniques include velocity-scaling, stochastic collision methods, constraint methods, extended system methods. The Anderson stochastic collision method (Anderson 1980) is best suited for the present case. In the Anderson method, an isolated chain is immersed in a constant temperature bath of imaginary ghost particles, a system that can be easily handled by DMD techniques (Zhou et al. 1997; Zhou and karplus 1999).

The basic idea of the Anderson method is that particles experience random collisions with imaginary heat-bath particles ghost particles, which do not appear explicitly in simulations. The Anderson method can be incorporated into DMD techniques by introducing a new type of collision—the ''bead-ghost'' collision—in addition to core, bond, and square-well collisions. The ''collision free'' time for a particle satisfies an exponential distribution. Thus, the time at which a ''bead-ghost'' collision occurs is calculated from an exponential-distribution-random-number generator, is the time since the previous ghost or real collision and n is the mean bead-ghost collision rate. Most bead-ghost collisions occur at t, 1/ n. Assuming that the ghost heat-bath particles are hard-spheres which have the same size and mass as the polymer bead, we can calculate the mean collision rate n with a given bead from the kinetic theory of gases.

2.4 Thermodynamics Quantities

All quantities are reported in terms of reduced units unless specified otherwise. The equations for reduced energy, temperature, and time, are $E^* = E/\varepsilon$, $T^* = k_B T/\varepsilon$, and $t^* = t * ((\varepsilon/M\sigma_L^2)^{0.5})$, respectively (Linhananta et al. 2005). These reduced formulas are the same as those used for Lennard-Jones systems (Allen and Tildesley 1987), where all units can be determined in terms of basic units of mass, energy and length. Given

these values for the parameters, each reduced time unit t* corresponds approximately to 1 ps, so that a folding simulation that lasts $t^* \approx 10^6$ is formally equivalent to a simulation of 1 ms in ``physical'' time. However, since the collapse process is significantly faster (by a factor of 10^2 to 10^3), than that observed experimentally (Hagen et al., 1996). Due to the simplicity of the model, a more meaningful conversion factor is t* close to 1 ns; in this case, $t^* \approx 10^6$, would correspond to 1 ms, a very reasonable scale for the folding time.

The heat capacity, C, is determined by the standard statistical mechanics formula

$$C = \frac{< E^2 > - < E >^2}{k_B T^2} \qquad \text{(Eq. 2.4)}$$

and the scaled heat capacity C*, is defined by the formula

$$C^* = \frac{C}{k_B} = \frac{< E^{*2} > - < E^* >^2}{T^{*2}} \qquad \text{(Eq. 2.5)}$$

where < > denotes the conformational average. E and E* denotes the internal energy and reduced internal energy, respectively. k_B is the Boltzmann constant. T and T* are temperature and reduced temperature, respectively. In this work, equilibrium simulations are performed from T*=1.0 to T*=6.0 in steps of $\Delta T^*=0.2$. Each of these simulations is started from the global minimum structure (Fig. 2.1), equilibrated for a reduced time of 20 000 and continued for 100 000 reduced time steps, during which data are recorded every 100 steps. At each temperature, five independent runs with different random initial velocities are performed to estimate errors.

The partition function Z can also be obtained via DMD simulations. Since square-well chain models have discrete energy levels, the partition function Z can be expressed as sum over all energy levels as follows (Pathria 1996, p55)

$$Z = \sum_{k} g_k e^{-\beta E_k} \qquad \text{(Eq. 2.6)}$$

where E_k and g_k are the energy and degeneracy factor for the energy level k, respectively. The degeneracy factor g_k is the sum of contributions from the various configurations that have the same energy. The contribution of each configuration depends on the volume of configurational space in which the chain is free to move without changing its energy. The partition function Z can be calculated for any temperature if the temperature-independent degeneracy factors g_k are known. If all energy levels have statistically significant populations at one temperature, it is possible to extract g_k from the probability distribution in simulation runs at that temperature (Ferrenberg and Swendsen 1989). The details can be found in the Appendix of Zhou et al. paper (Zhou et al. 1997). Once the partition function is known, thermodynamic quantities can be calculated from the relations (Pathria 1996, p53-54):

$$A^* = -T^* \ln Z \qquad \text{(Eq. 2.7)}$$

$$E^* = (T^*)^2 \, \partial \ln Z / \partial T^* \qquad \text{(Eq. 2.8)}$$

$$C^*_v = \partial E^* / \partial T^* \qquad \text{(Eq. 2.9)}$$

Where A* is the reduced Helmholtz free energy.

Chapter 3

Exploring Multiple Folding Pathways of the Trp-cage with a Homogenous All-Atom Go Model Using the Cluster Analysis Method

Abstract:

Recently, experimentalists have focused on engineering ultra-fast folding proteins whose folding speeds approach the "speed limit", where folding from the random coil to the native state proceeds without having to overcome an "entropic barrier". The downhill folding pathway is distinguished by the absence of a free-energy folding barrier, as well as a very high folding rate compare to the other pathways. In this work, we studied folding process of the ultra-fast folding Trp-Cage mini-protein by using homogenous all-atom Go potentials. By classifying 100 long-time molecular simulations with the cluster analysis method, we observed three different folding mechanisms: downhill (12%), diffusion-collision (80%) and hydrophobic collapse (8%). The observation of a downhill folding pathway partly explains the ultra-fast folding speed of the Trp-cage.

3.1 Introduction

It is well known that *in vitro* and *in vivo* proteins readily fold from an unfolded (random coil) state to their unique native states. In the random coil state, a protein assumes an extended state where it strongly fluctuates among the large number of available conformations. In the native state, a protein assumes a compact structure that fluctuates weakly. The elucidation of how proteins fold has been the topic of intense research for the past three decades. Most of these works focused on small proteins, which are believe to fold by a two-state mechanism, in which the two stable states are the random coil and native states. To fold to its native state a protein must surmount an entropic barrier known as the transition state. This is a bottleneck ensemble of conformations that all folding pathways must pass through to reach the native state ensemble.

The assumption of two-state mechanism is often used in experimental studies of protein folding. In such experiments, single-site mutations in which a single amino acid residue of the original primary structure of a protein (wild type) is substituted by a different amino acid to create a mutant amino acid sequence. For example, consider a hypothetical wild type protein sequence RKDE, where the four amino acid residues are Arginine (R), Lysine (K), Aspartic Acid (D), and Glutamic acid (G). The mutant sequence RKWE can be created by changing the third residue from D to W (Tryptophan). By observing how single-site mutations affect the stability, and the folding and unfolding rate of the protein, it is possible to determine the so-called Φ-values of each amino acid residue of the protein. The Φ-values quantify the amount of native structures formed in the transition state (Fersht 1999). The determination of Φ-values, from such protein mutations experiments, assumes that proteins fold by a two-state mechanism. This is a contentious issue since it is usually not possible to experimentally observe the actual folding pathways. In addition, computer simulation studies often predict "uncooperative" behaviour where metastable intermediate states are observed in addition to the random coil and native states. Hence computer simulations is a valuable tool in linking the pathways proteins take from the random coil to the transition state, and in assessing the validity of the method of Φ-value analysis (Snow et al. 2005).

One of the main challenges in computer simulation research on protein folding is the extraction of useful information from the large amount of datum. In a typical study hundreds of simulations, under different thermodynamic conditions, are performed. Various methods have been used to determine the average dominant folding pathways of the proteins. In the reaction-coordinate method the time variation of reaction coordinates are used to determine the folding pathways (Caflisch 2005). Usually, these reaction coordinates quantify the amount of native structure. Examples are the radius of gyration and the root-mean-squares deviation (RMSD). In the cluster analysis method protein structural data recorded during simulations are grouped into clusters based on structural similarity, which is determined by comparing the energy and/or RMSD of the structures. This enables the detection of folding intermediates, which are metastable structures whose native contents are intermediate between the unfolded random coil structures and the folded native structures. Clustering techniques have been previously used in the analysis of conformational data, particularly in identifying recurring conformations in folded protein structures (native states), obtained from experimental and computational results. However, it also has been used to identify folding intermediates in the folding processes (Duan and Kollman 1998; Chowdhury et al. 2003; Chowdhury et al. 2004).

This chapter describes computer simulations results of the homogeneous all-atom Go model of the Trp-cage. Frequently occurring unfolded states and metastable intermediate states are identified by cluster analysis based on pairwise root-mean-squared distance (pRMSD) (Shortle et al. 1998; Betancourt et al. 2001; Zhang et al. 2004). Three "average" folding pathways are observed, including the elusive downhill folding pathway in which a protein folds "downhill" from its unfolded random coil state to its native state, without having to cross an entropic barrier.

3.2 Method

3.2.1 Simulation Detail

One hundred 120000-reduced-time-step folding discontinuous molecular dynamics (DMD) simulations were performed at $T^* = 3.0$ for Trp-cage starting from 100 different initial random-coil conformations, which were produced by a short DMD simulation at

high temperature T*=5.0, where the protein is unfolded. In a previous study it was shown that the transition temperature, at which the random coil and native state are equally stable, is T*=4.0 for homogenous all-atom Go model of Trp-cage (Linhananta et al. 2005). In that work, it was estimated that 100 reduced time unit scales to ~ 20 ns to 70 ns.

3.2.2 Root Mean Square Deviation (RMSD)

The quantity *Root Mean Square Deviation* (RMSD) is often used to compare the suctural similarity between two conformations of molecule (protein). Consider two sets of molecular conformations **V** and **W**. Let v_i and w_i be the coordinates of the i^{th} atom of the molecule in the **V** and **W** conformations, respectively. Here i = 1, 2, ..., N, where N is the number of atoms in the molecule. The RMSD is defined as follows:

$$\mathrm{RMSD}(\mathbf{v},\mathbf{w}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|v_i - w_i\|^2}$$

$$= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(v_{ix} - w_{ix})^2 + (v_{iy} - w_{iy})^2 + (v_{iz} - w_{iz})^2}$$

where in this work the RMSD value is in Ångström (Å), 1 Å is equal to 10^{-10}m. The standard method is to superimpose the two conformations being compared so that the value of RMSD is minimized. In protein folding analysis, unfolded protein conformations are usually compared to the known ground state structure, which in this work corresponds to the lowest energy structure of the all-atom Go model. In this thesis, an RMSD value between a conformation and ground state structure of Trp-cage less than 1.6 Å indicates a folded protein in the native state; RMSD value between 3.5Å and 8.0Å indicates an intermediate structure with a moderate amount of native content; while RMSD larger than 8 Å indicates an unfolded random-coil structure with little native content.

In this thesis, RMSD denotes the main-chain RMSD (side chain atoms are not included in the calculation) of a Trp-cage conformation from simulation data compare to the Trp-cage ground state (lowest energy NMR structure) structure. pRMSD denotes the main-chain RMSD between any two conformations obtain from simulation data.

3.2.3 Structural Cluster Analysis

All folding simulation data are analyzed by the following method:

a) Begin by calculating the main-chain RMSD (with respect to the ground state NMR structure) of all recorded conformations. Select the region of interest by setting the lower limit RMDS = R_{min} and upper limit RMSD = R_{max}. The conformations of interest include all Trp-cage conformations within the range $R_{min} < RMSD < R_{max}$. Only these conformations, referred to as the ensemble of conformations, are analyzed by cluster analysis.

b) Set up the threshold value R_{cut} to classify structural similarity. Two Trp-cage conformations with $pRMSD < R_{cut}$ are considered to be similar in structure.

c) The main-chain pRMSD is calculated for each pair of structures in the ensemble of conformations. Then gives a matrix pRMSD(i,j) ($i \neq j$) (i and j varying from 1 to the total number of conformations in the ensemble of conformations).

d) For a given conformation, say the i^{th}, count the total number of j conformations with $pRMSD(i^{th},j) < R_{cut}$. The conformation with the greatest number of pRMSD $< R_{cut}$ pairs, say i_1, is the center of the first cluster. The first cluster includes i_1, and all conformation j for which $pRMSD(i_1,j) < R_{cut}$.

e) Remove the conformations belonging to the first cluster from the ensemble of conformations. Repeat step d) to calculate the second cluster.

f) Repeat to determine higher order clusters (third, fourth,....).

The first cluster includes the most frequently occurring similar conformations and, using the basic principles of statistical mechanics, is the most stable intermediate state. It follows that the second cluster is the second most stable intermediate state and so on. In this work, we set the parameters for structural cluster analysis as: R_{min}=3.5Å and R_{max}=8.0Å. The reason is shown in Fig. 3.1, which plots the free-energy functional vs. RMSD of the 100 folding simulations at T* = 3.0, where the transition state (TS) is located at RMSD ~ 3.2Å, and where conformations with RMSD > 8.0Å have high free energy and are unstable. Hence stable intermediate states are most likely to include

conformations within the range 3.5Å < RMSD < 8.0Å. This work uses R_{cut}=2.5Å, which is a similar values as those used by other researchers (Chowdhury et al 2004).

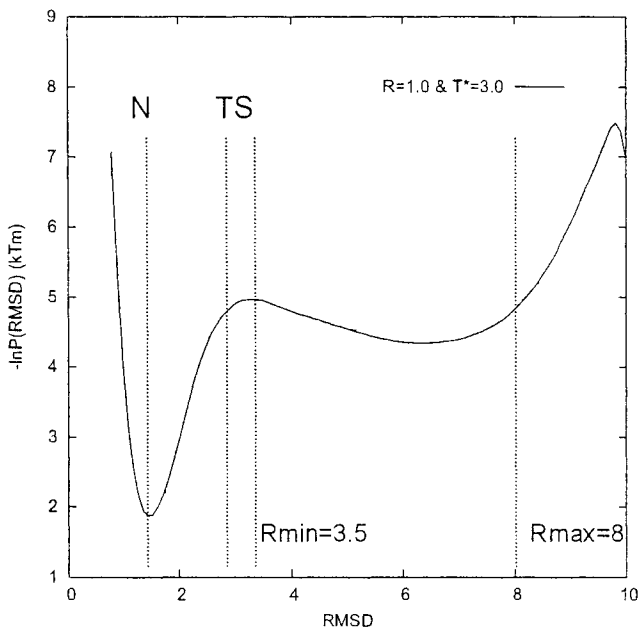


Figure 3.1. Free energy profile of homogenous Go model of Trp-cage 100 folding simulations at T*=3.0. In the figure, N denotes the native state and TS the transition state.

3.2.4 The method of reaction coordinates

In computer-simulation studies of protein folding reaction coordinates are often used to monitor the progress of folding. The most commonly used coordinates are the main-chain RMSD (see 3.2.2) and the fraction of native contact (Q) (Celmenti et al. 2000, 2002, and 2003; Zhou 2003). The RMSD monitors the overall folding while Q monitors the formation of tertiary contacts. This work employs RMSD and Q. The secondary structure formations are monitored by calculating the hydrogen bond probabilities of the α-helix and 3_{10} helix as folding progresses along the reaction coordinates.

The quantities are defined as follow. The total nonlocal native contact, M_{nat} is defined as

$$M_{nat}=(1/2)\sum M_{i.nat} \qquad \text{(Eq. 3.1)}$$

where $M_{i.nat}$ is the number of nonlocal atomic contacts of the i^{th} residue with all other residues when the Trp-cage is in the lowest-energy NMR state. A nonlocal contact is

defined as a square-well overlap between two atoms, belonging to residue i and j, respectively, such that $|i-j| \geq 4$. We define the fraction of native contact of residue i as

$Q_i = m_i / M_{i.nat}$, (Eq.3.2)

where m_i is the nonlocal native contact number for the i^{th} residue in the conformation of interest. For a given conformation, Q_i quantifies the native content of the i^{th} residue with $Q_i = 0$ ($Q_i = 1$) denoting that the i^{th} residue is completely unfolded (folded). Finally the total fraction of native contact is defined as

$Q = \sum m_i / Mnat$. (Eq.3.3)

It is clear that Q quantifies the native content of the whole protein with Q=0 (Q=1) denoting that the Trp-cage is completely unfolded (folded).

As mentioned in chapter 2 (see Fig. 2.1) the key structural elements are the α-helix from residue 2 through 8, the hydrophobic core centered about the Trp6 residue, the salt-bridge between residues Asp9 and Arg16, and the 3_{10}-helix from residue 11 through 14. The following variables are used to monitor the native content of these key structural elements. Q_{helix} is the fraction of three hydrogen bonds (Leu2-Trp6, Tyr3-Leu7 and Ile4-Lys8) that stabilized the α-helix. Q_{hyd} is the fraction of nonlocal native contacts in the hydrophobic-core (Tyr3, Trp6, Leu7, Pro12, 17, 18 and 19). Q_{sb} is the fraction of native contacts between the salt-bridge stabilizing residues Asp9 and Arg 16. Q_{310} is the fraction of the 3_{10}-helix hydrogen bond formation.

3.2.5 Transition states ensemble (TSE)

As mentioned earlier, the transition state ensemble (TSE) is a bottleneck ensemble that all folding pathways must pass through to reach the native state. It is associated with the main entropic barrier that divides the random coil and native states. For an appropriately chosen reaction coordinate a free energy functional profile along the coordinate would show a prominent maximum corresponding to the TSE. This is illustrated in Fig. 3.1, where the TSE maximum is located at RMSD ~ 3.2Å. Many researchers employed Q as the reaction coordinate to locate the TSE maximum from equilibrium simulation data (Clementi et al. 2000). However, it has been shown that the location of TSE determined

in this way depends on the choice of reaction coordinates (Ozkan et al. 2001). A more unambiguous method for determining the TSE is to define the transition state ensemble as the conformations from where a protein has equal chances of folding to the native state or unfolding to the random coil state (Du et al. 1998). In this work, to locate the TSE, 200 conformations with the reaction coordinates (Q~0.4-0.55, RMSD~2.8-3.5) were selected from the folding simulation data. These conformations are deemed to be good candidates for transition state ensemble. To identify the TSE conformations among these candidates, 40 short simulations, of duration t*=2000, staring from each of the candidate structures. Conformations that fold to the native state 16-24 times (probability = 0.4-0.6) are identified as transition-state conformations. 31 transition states are identified. Simulations of some of the candidate structures for t*=4000 give similar results, illustrating that the results are not sensitive to the duration of the simulations. Similar methods have been previously employed by the other researchers (Du et al. 1998; Ding et al. 2002). Typical transition state structures of the Trp-cage at T*=3.0 are shown in Figure 3.9A and 3.9B.

3.2.6 Mean First Passage Time τ and Folding rate, k

In this work a protein is considered to have folded to its native state if its main-chain RMSD is less than 1.6 Å. The first passage time (FTP) is defined as the time it takes for a protein to fold starting from an initial unfolded conformation. The Mean First Passage Time (MFPT), denoted by τ, is simply the average of the FTP, which in this chapter is simply the average FTP of the 100 Trp-cage simulations at T*=3.0. In this work we shall use the term average folding time and MFPT interchangeably. Finally, the folding rate is defined as $k = 1/\tau$.

3.3 Results

3.3.1 Classifying the trajectories

Previous computational works on Trp-cage reported the observation of multiple folding pathways:

a) Diffusion collision, partial α-helical structure formed first (Chowdhury et al. 2004);

b) Hydrophobic collapse, hydrophobic-core formed first (Zhou 2003);

c) Ultra-fast folding, α-helix and hydrophobic-core formed cooperatively. (Chowdhury et al. 2004).

For the diffusion-collision pathway, it is expected the α-helix forms first. Hence a cluster analysis of the data should reveal that the most populated cluster (first cluster) composed of structures with well formed α-helix. In contrast, for the hydrophobic collapse pathways, the hydrophobic-core forms before the α-helix, and the first cluster is expected to have little α-helical content. In the diffusion collision and hydrophobic collapse mechanisms, there is an entropic barrier that must be crossed before folding to the native state. To cross the barrier the protein must locate the transition state. This is often referred to as the rate-limiting step. In contrast, for ultra-fast downhill pathway, the entropic barrier does not have to be surmounted, and the protein proceeds "downhill" toward the native state. Here no intermediate states exist and cluster analysis should not detect any cluster.

From a total of 100 runs for homogenous Trp-cage model, all 100 trajectories reach the native state (RMSD < 1.6Å). Cluster analysis of all 100 runs shows that simulations can be classified into three mechanisms: (I) diffusion collision (79 runs), MFPT≈ 16000steps; (II) hydrophobic collapse (9 runs), MFPT≈ 27800steps; (III) downhill (12) MFPT≈ 5900steps. The average folding time of proteins that fold by the diffusion collision mechanism (τ=16000) is significantly lower than the hydrophobic collapse pathways (τ=27800), which agree with a previous study (Linhananta et al. 2005). The ultra-fast folding pathway has very a low average folding time (τ=5900) compare to the other pathways, reflecting the "downhill" nature of this pathway. Similar results were reported by Duan group (Chowdhury et al. 2004), which found 3-15% of all trajectories followed the fast folding pathway (25 to 60 ns).

To verify the above results, we also classify the pathways by checking the time-evolution of α-helix and hydrophobic-core formation for individual trajectory. This method has been used by other groups (Zhou and Karplus 1999a and 1999b). Figure 3.2 is a typical trajectory in which the protein folds by the diffusion collision folding pathways. The

average fraction of native contacts in α-helix rises quickly to 80% within the first 1000 time steps. This is followed by a long random search to form the hydrophobic-core contacts necessary to reach the transition state. Noteworthy is the decrease in α-helix content that occurs prior to significant hydrophobic core formation. The typical trajectory sorted into the hydrophobic collapse folding pathway is shown in Figure 3.3. Here, the hydrophobic-core forms before the α-helix. An interesting feature is the large decrease in the hydrophobic core content before the appearance of significant α-helical structures.

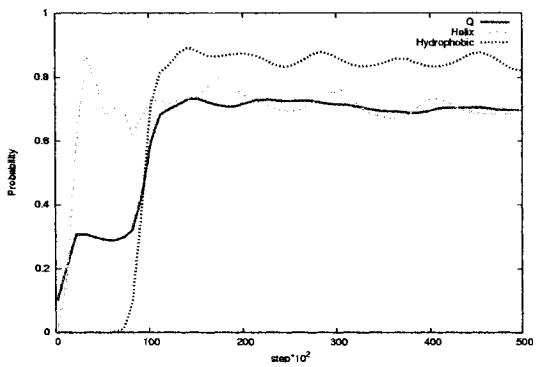


Figure 3.2. A folding trajectory of Trp-cage at T*=3.0 in diffusion collision mechanism. Fraction of native content vs. time.

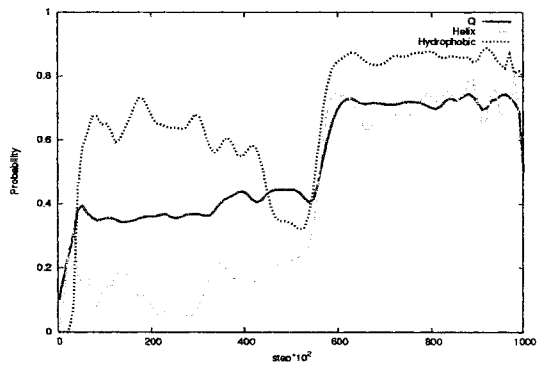


Figure 3.3. A folding trajectory of Trp-cage at T*=3.0 in hydrophobic collapse mechanism. Fraction of native content vs. time.

3.3.2 Statistical analysis of the folding trajectories

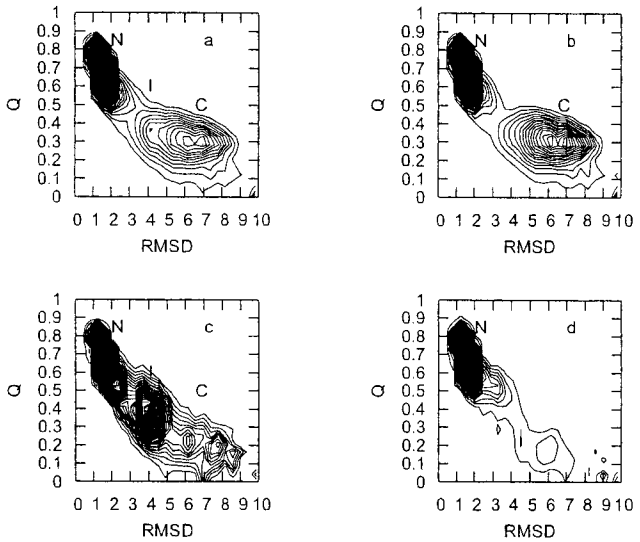


Figure 3.4. Probability distribution as a function of RMSD and the fraction of nonlocal contacts Q. (a) 100 simulations at t*=3.0 of homogenous Go model of Trp-cage. The vicinities of the native N state, intermediate I state, and the random coil C state are indicated. Regions where contour lines are closely spaced indicate steep changes in the probability distribution. The very dark region corresponds to sharp peak about the native state; (b) 79 out of 100 simulations by diffusion-collision pathways; (c) 9 out of 100 runs by hydrophobic-collapse pathways; (d) 12 out of 100 trajectories by downhill folding pathways.

Figure 3.4 shows contour maps for Trp-cage folding simulation at T*=3.0. Regions where the lines are closely spaced indicate high occupation probability. Figure 3.4(a) is the contour map for all trajectories and shows two main regions of high occupation probability: one associated with the unfolded random coil state (C), the other with the folded native state (N). Also evident is a metastable intermediate state region (I), which is similar to the intermediate state observed by Zhou using an ab-initio model (Zhou 2003). In that work, the intermediate state is hypothesized as a key component in the folding mechanism. Figure 3.4(b) shows the contour map of all trajectories that fold by the dominant diffusion collision pathway, which is the pathway followed by 79% of all trajectories. The distribution shows a purely two-state landscape with stable native and random coil phases, with no metastable intermediate state. Figure 3.4(c) shows the

contour map of trajectories that follow the hydrophobic collapse pathway. It shows three main stable states N, I, and C. Structural analysis shows that the intermediate phases in Figure 3.4(a) and (c) are identical. Hence, the intermediate state I does not play a role in the dominant folding pathway, which contradicts the conclusion of Zhou. This illustrates that the analysis of protein folding simulation data is complicated when there are multiple folding pathways. For trajectories folding by the downhill mechanism, Figure 3.4(d) shows that only the native state region is well populated. Outside this region, unfolded states are unstable and, consequently, are sparsely populated.

Figs. 3.5a and 3.6a show the free energy profile for the three folding pathways along the reaction coordinates Q and RMSD respectively. The free energy are calculated by the formula $f(RMSD,Q) = \Sigma\ e^{-\beta Ei}$, where the summation i is over all conformations with values RMSD = 1 Å to RMSD = 10 Å, and Q=0.1 to Q= 0.9, and E_i is the energy of the i^{th} conformation. It is clear that all three pathways end up near the same native minimum at Q ~ 0.73 and RMSD ~ 1.5 Å. However, they all start from different partially folded states. For the diffusion-collision mechanism, there is a very broad RMSD minimum at RMSD = 5-8 Å, and a comparatively sharp Q minimum at Q ~ 0.3. This is consistent with an unfolded state with substantial α-helical secondary structure, and partial and fluctuating core tertiary contacts. In contrast, the hydrophobic collapse mechanism show a sharp RMSD minimum (Figure 3.5a) at RMSD ~ 4 Å and a broad Q minimum at Q=0.3-0.55. This is consistent with a compact structure with varying amount of secondary and tertiary structure. For both the diffusion-collision and hydrophobic-collapse pathways, the free energy minimum is separated from the native state minimum by a transition state maximum. The behavior is very different for the downhill-folding path, where the energy profiles (pink lines in Figs. 3.5a and 3.6a) shows only a native state minimum (Q≈ 0.73 for Fig. 3.5a and RMSD≈ 1.5Å for Fig. 3.6a), with all unfolded conformations having about equal values of free energy. A remarkable feature is the absence of a entropy-driven maximum separating the native states from the unfolded states. This is why trajectories following this pathway can fold unhindered at ultra-high speed.

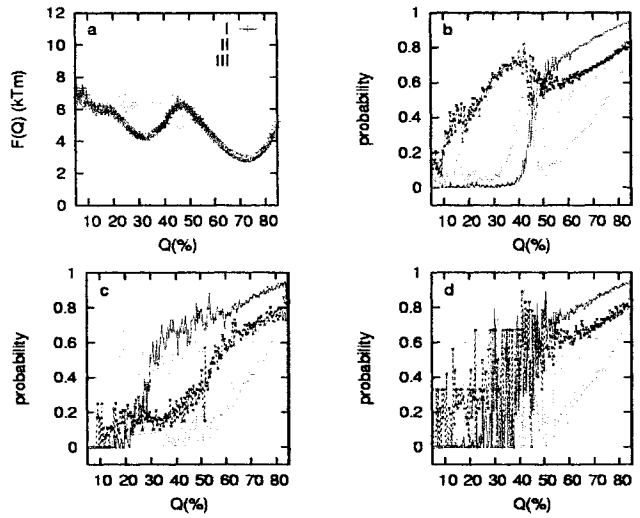


Figure 3.5 Free energy profiles (a) and probability of key pair interactions formed as a function of Q (b, c and d)at T*=3.0 for different folding mechanism.(a), I for diffusion-collision, II for hydrophobic collapse, and III for downhill; (b), diffusion-collision pathways;(c) hydrophobic-collapse pathways; (d) downhill folding pathways (in (b),(c) and (d): blue for helix; green forTrp6-Pro12; red for Trp6-Pro18;pink for Asp9-Arg16, salt-bridge).

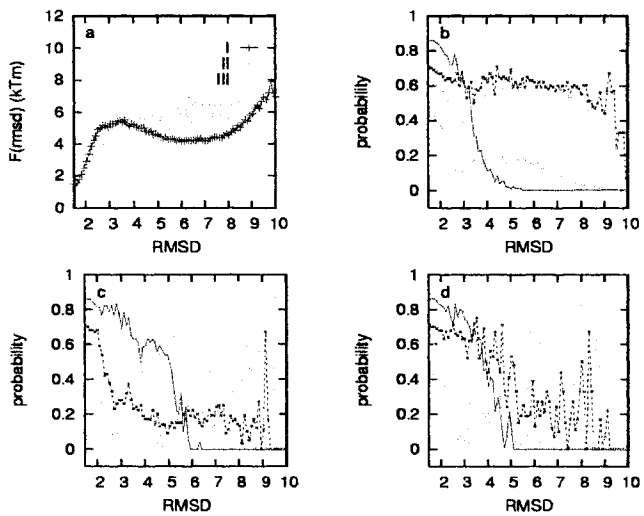


Figure 3.6 Free energy profiles (a) and probability of key pair interactions formed (b, c, and d) as a function of RMSD at T*=3.0 for different folding mechanism. Symbols and color codes are the same as for Fig. 3.5.

3.3.3 Diffusion collision mechanism

The probability of the key non-local native contacts between two residues formed, Q_{ij}, as a function of Q and RMSD is shown in Figure 3.5b-d and 3.6b-d respectively. Similar methods have been used by other group (Clementi et al. 2003) to extract specific details from simulation data. These results clearly reveal the "average" time evolution of specific structural elements of the Trp-cage model folding along the three main pathways.

A detail analysis of diffusion collision mechanism, seen in 79 trajectories is given in Figure 3.5b and 3.6b. Starting from a completely unfolded state, the early folding events ($Q<0.3$) are the formation of α-helix and Trp6-Pro12 residue pairs. Then, from Q • 0.35 to 0.4, the probability of salt-bridge formation rapidly increases to the peak-point (0.5). This value is higher than that of the native structure (0.42). In the region of Q increasing from 0.45 to 0.5, the tertiary contacts between Trp6, Pro 17,18 and 19 increase. Meanwhile, the probability of salt-bridge formation and of Trp6-Pro12 decreases and a drop in the helical content is observed. Similar drop of helical content as folding proceed was reported in a computational study of the villin headpiece domain (Duan Y. and Kollman P.A. 1998). These results demonstrate the crucial role of the coupling between the secondary and tertiary structure in protein folding processes. When Q reaches 0.5, both hydrophobic-core and α-helix are well formed, but the probability of salt-bridge formation is very low. At this Q value, the structure of the Trp-cage model is characterized by the formation of native contacts of the pairs Tyr3-Pro18, Tyr3-Pro19, Trp6-Pro17, Trp6-Pro18 and Trp6-pro19. It is noteworthy, that these Trp-cage native contacts have been designated as key contacts by an NMR NOE experiment (Neidigh 2002). After the formation of these key contacts, all native contacts show a steady rise toward the native state. These results suggest that conformations in the transition state ensemble are in the reaction coordinate range $Q\approx0.45$ to 0.5 and RMSD≈3Å (Linhananta 2005).

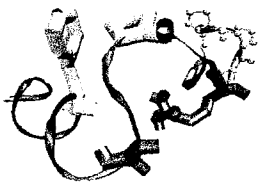
Table 3-1: Summary of the most populated clusters for metastable intermediate states in ***Diffusion Collision folding path way (6638 conformations)***

	size	RMSD	Q	Qhelix	Qsb	Q_{310}
1	50	5.35	0.36	0.80	0.40	0.08
2	36	5.46	0.35	0.81	0.29	0.19
3	36	4.95	0.36	0.80	0.30	0.11
4	35	5.25	0.36	0.79	0.35	0.03
5	33	4.94	0.37	0.81	0.47	0.09
6	29	4.68	0.37	0.83	0.45	0.10
7	28	6.17	0.35	0.78	0.36	0.07
8	24	5.44	0.35	0.79	0.44	0.08

Cluster analysis was used to determine the major metastable intermediate state of Trp-cage protein folding by the diffusion-collision mechanism. The eight largest clusters are summarized in Table 3-1. The average RMSD is from 5 to 5.5, and Q = 0.36. But, Q_{helix} = 0.8, which actually above the native value of Q. The representative simulated structures of the first four clusters are shown in Figure 3.7. These structures demonstrate the following structural fingerprint: (1) well formed α-helix; (2) no hydrophobic-core; (3) Trp6-Pro12 contacts but Trp6 is flipped away from the center of the cage; (4) Asp9-Arg16 salt-bridge occupies in the centre of the cage. The results are consistent with the contact order theory, which states that for a given topology, local interactions are more likely to form early in folding than non-local interactions (Baker D 2000). The presence of the Asp9-Arg16 salt-bridge in the intermediate state and its absence in the latter stage of folding (discussed earlier) suggests that the early formed salt bridge must be broken in order for folding to proceed. A similar conclusion was reached by Zhou (Zhou 2003), where it was suggested that the early formed salt bridge impedes the ability of the Trp6 residue from forming the hydrophobic core. This is consistent with the conclusion of Duan et al. (Chowdhury et al. 2004) that states that the incorrect burial of Trp6 impairs folding. Taken together, our results indicate the orientation of the side chain of Trp6 is involved in the rate-limiting steps.

(A)

(B)



(C)

Figure 3.7. Represent structures of the metastable intermediate states in diffusion-collision mechanism. (Blue for α-helix; Red for Salt-bridge; Golden for Trp6 and Pro12; CPK for Pro17 to 19)

3.3.4 Hydrophobic-collapse folding mechanism

The hydrophobic-collapse mechanism is followed by 9 out 100 folding trajectories. The time-variation plots are shown in Figure 3.5c and 3.6c. They show that a partial hydrophobic core involving residues Trp6, Pro17, 18, 19 are formed at Q~0.3, at the early stage of folding. For comparison sake, in the diffusion-collision mechanism, Trp-cage core contacts do not appear until Q>0.4. The plots also show low probability of α-helical structure (Qhelix <0.2) and salt-bridge (Qsb < 0.2). At a later stage, Q=0.4-0.6, the helical content increases significantly. At the final stage of folding (Q>0.6) there is a steady increase in the native content of all key structural elements.

Table 3-2: Summary of the most populated clusters for metastable intermediate states in ***Hydrophobic Collapse-folding pathway (1486 conformations)***

	size	RMSD	Q	Qhelix	Qsb	Q_{310}
1	29	3.90	0.42	0.16	0.05	0.28
2	21	3.94	0.45	0.18	0.01	0.33
3	19	3.93	0.48	0.64	0.00	0.42
4	16	3.96	0.44	0.18	0.06	0.25
5	16	3.78	0.41	0.17	0.00	0.25
6	12	3.78	0.35	0.54	0.00	0.25
7	11	4.05	0.37	0.22	0.00	0.45
8	11	4.56	0.40	0.16	0.00	0.64

Cluster analysis is used to determine the intermediate state of Trp-cage protein that folded by the hydrophobic-collapse mechanism. The properties of the eight largest clusters are summarized in Table 3-2. The average of the reaction coordinates are RMSD = 4A, Q = 0.42. But, note that $Q_{helix} < 0.25$. The representative simulated structures of the first four clusters are shown in Figure 3.8. These structures demonstrate the following structural fingerprint: (1) only the first turn of α-helix are formed; (2) hydrophobic-core are formed except for the Tyr-PPII interactions; (3) Trp6-Pro12 are formed and Trp6 is at the cage-centre; (4) the possibility of salt-bridge formation is very low. Figure 3.9B shows the structure of transition states in hydrophobic-collapse folding mechanism. By comparing these structures, we find that the significant difference between the transition state and intermediate states is the absence in the latter of significant salt-bridge and 3_{10}-helix. Therefore, the formation of Asp9-Arg16 (salt-bridge) pair contacts and 3_{10}helix should be the rate-limiting steps in the hydrophobic collapse pathway.

(A)

(B)

(C)

Figure 3.8. Represent structures of the metastable intermediate states in hydrophobic collapse mechanism. (Blue for α-helix; Red for Salt-bridge; Golden for Trp6 and Pro12; CPK for Pro17 to 19)

(A) Transition state in diffusion collision mechanism.

(B) Transition state in hydrophobic collapse mechanism.

(C) Ground state (lowest energy) of Trp-cage.

Figure 3.9. Represent structures of the transition states and ground state. (Blue for α-helix; Red for Salt-bridge; Golden for Trp6 and Pro12; CPK for Pro17 to 19)

3.3.5 Downhill folding mechanism

In downhill folding mechanisms, proteins in the random coil states rapidly fold to the native states without the hindrance of entropic barriers (Eaton 1999; Kubellka et al. 2004). Due to its small size and ultra-fast folding speed, several researchers (Zuo et al. 2006; Bunagan et al. 2006) have proposed that the Trp-cage is a good candidate for downhill folding. As discussed, 12 out of 100 trajectories satisfied the "downhill" criteria of encountering no entropy barrier in the free energy profile (see Figs 3.5a and 3.6a). They also fold very fast with average folding time of τ=5900 compared to the trajectories that fold by diffusion-collision and hydrophobic-collapse mechanisms which have τ=16000 and 27800, respectively. In the diffusion and hydrophobic-collapse mechanisms, folding is hierarchical in that certain native contacts tend to form before others. This means that certain contacts (unfolded conformations) are more likely than others. In the downhill folding pathway, there is no hierarchy. This is supported by the free energy profiles (Figs. 3.5a and 3.6a) of downhill folding which show uniform probability away from the native state, and by the contour map of Fig. 3.4d which shows uniformly low probability for the unfolded conformations. This is also supported by cluster analysis, which failed to detect any metastable intermediate state.

Table 3-3: transition state and folded state

	size	RMSD	Q	Qhelix	Qsb	Q_{310}
1	36	3.15	0.49	0.66	0.07	0.48
2	10	3.00	0.51	0.30	0.41	0.30
3	1000	1.51	0.71	0.82	0.42	0.28

Note:
1:TSE on diffusion collision mechanism
2:TSE on hydrophobic collapse mechanism
3:Folded states for wild type at T*=3.0.

3.4 Discussion and conclusion

The Trp-cage is a recently synthesized protein (Neideigh et al. 2002) with one of the fastest known folding time of ~4μs (Qiu et al. 2004). There have been numerous computer simulation studies (Slimmerling et al. 2002; Snow et al. 2002; Zhou 2003; Nikiforovich et al. 2003; Pitera et al. 2003; Chowdhury et al. 2003 & 2004). Even before its structure was resolved by a NMR experiment, *ab-initio* simulations were performed

(Slimmerling et al. 2002). In that work, two 100ns simulations succeeded in folding the Trp-cage correctly to its native structure within 5ns and 20ns. The Trp-cage was also studied by the **Folding@home** group using distributed computing technique, in which thousands of simulations were performed on "home computers" (Snow et al. 2002). However, the longest simulation was only 80ns, and not all simulations successfully folded the protein. In another study, the Amber ab-initio software was used to perform 77 100ns simulations, only 10 of which resulted in folded Trp-cage. The key point is that most computer simulation studies on the Trp-cage have employed ab-initio codes in which proteins are represented in atomic detail using accurate interaction force field. Due to the complexity of these models, it is only possible to perform simulations for time durations that are equivalent to only about 5% of the average Trp-cage folding time of 4μs. Hence, in these works, the folding mechanisms of the Trp-cage are extrapolated by approximate methods.

Our all-atom Go model found three distinct folding pathways: diffusion-collision; hydrophobic collapse; downhill. In the dominant diffusion-collision pathway, the α-helix forms early follows by the formation of the hydrophobic core anchored by the Trp6, Pro12 residues. This is the same conclusion reached by the ab-initio simulations of Duan group (Chowdhury et al. 2004). The main difference is that in our work all simulation trajectories successfully fold the Trp-cage to its native structure. Hence our conclusion on folding mechanisms was reached by direct observations of the data.

A key feature of the Trp-cage (TC5b) is that it was synthesized from the naturally occuring peptide sequence TC3b (Neidigh et al. 2002). The main difference between TC3b and TC5b, is the presence, in the latter, of an Asp9-Arg16 salt bridge, which is believed to be one of the reasons for the stability of Trp-cage despite its small size. In a landmark study Zhou (Zhou 2003), using the Amber OPLSAA force field in explicit water, performed detailed simulations on the Trp-cage. The simulations observed an intermediate state stabilized by the Asp9-Arg16 salt bridge. It was hypothesized that the salt-bridge stabilized intermediate plays a crucial role in the ultra-fast folding speed of the Trp-cage. But that the salt bridge must be broken at a later stage in order for the Trp-

cage to be able to fold to the native states. These features have been verified by other workers (Ding et al. 2005). The fact that our all-atom Go model is able to produce these specific behaviors suggest that Go models can accurately produce generic and specific detail of protein folding.

The computer simulation results presented in this chapter is the first to observe downhill folding mechanism in the Trp-cage. Downhill folding has been observed in a simulation study on the src SH3 domain (Shea et al. 2002). However, that work employed the importance equilibrium sampling method to extrapolate non-equilibrium folding processes. As far as we know, this is the first work to directly observe downhill folding starting from the random coil state follows by a monotonic increase toward the native state without the hindrance of an entropic barrier.

Chapter 4

Investigating relative roles of α-helix, hydrophobic core and specific pair interactions in the folding mechanism of the ultra-fast folding protein Trp-cage by varying the interaction potential of an all-atom Go model

Abstract

Recent works on proteins suggest that the mechanism of folding is determined by the balance between the stability of secondary structural elements and the amount of hydrophobic core in the native structure. In this chapter, these factors are investigated by altering the strength of the interaction energy of key structural elements of the Trp-cage, such as the α-helix, hydrophobic-core, and salt-bridge. It is found that varying the stability of α-helix and hydrophobic-core strongly alters the folding kinetics and significantly changes the folding rate. In contrast, increasing the stability of the salt-bridge increases the folding rate but does not change the folding mechanisms. Increasing the stability of Trp6-pro12 interactions decreases the folding rate 50% and switches the two-state mechanisms into a more complex folding mechanism. This highlights the importance of energetic balance in the Trp-cage and shows that the folding pathways are controlled mainly by the stability of the α-helix and hydrophobic-core.

4.1 **Introduction**

The role of α-helical propensity and hydrophobicity in determining the rate and mechanism of protein folding has been explored by numerous theoretical and experimental studies (Munson et al. 1996; Bieri et al. 1999; Chiti et al. 1999; Kuhlman et al. 2004; Roder 2004; Meisner et al. 2004; Susanne et al. 2005; Liwo et al. 2005). These studies have shown that the folding rate of proteins that fold by two-state-kinetics is dependent on the content and stability of their secondary structure. For example, α-helical proteins formed from sequences with high local helical propensity have been found to fold by a diffusion-collision mechanism (Dearco et al. 2004), which describes a hierarchical process in which marginally stable elements of secondary structure collide and dock, forming intermediates with increasing stability, ultimately leading to the native state. In other models, hydrophobic collapse dominates the early stages of folding, causing compaction that decreases the accessible conformational space and, consequently reducing the search time required by proteins to reach their native states. A third model, the nucleation-condensation model (Fersht A.R., 1995), proposes that secondary and tertiary structure are stabilized concomitantly with most, if not all, residues contributing towards the stability of the folding nucleus, whose formation characterizes the rate-limiting transition state (Fersht et al. 2004). The folding mechanism that dominates folding for a particular amino-acid sequence is determined by the balance between the intrinsic stability of secondary structural elements and the propensity of the polypeptide chain to undergo hydrophobic collapse. The folding process can thus be described by a variable model in which the balance between the stability of secondary structure and the hydrophobicity of proteins can be tuned to determine the folding pathways (Fersht et al. 2002; Giani et al. 2003).

A number of studies on the role of helix stability in protein folding have shown that increasing the intrinsic stability of helices by amino-acid substitutions of solvent-exposed residues stabilizes the native structure (Munoz et al 1996; Taddei et al. 2000). In such cases, stabilization of helices increases the rate of folding if the helix is formed in the rate-limiting step, but not if it is formed in the unfolded state. Alternatively, if the helices

are formed only after the rate-liming transition state has been crossed, increasing the helical propensity has no effect on the folding rate constant, but results in a reduction in the rate of unfolding. Altering the stability of helix, therefore, is a powerful method of elucidating the role of individual secondary structural elements in the mechanism of folding.

Although the importance of hydrophobic collapse in folding has been known for decades (Dill K.A. 1990), predicting the effects of altering the hydrophobicity on the mechanism of protein folding is difficult. For example, decreasing the size and/or hydrophobicity of amino-acid residues involved in the folding nucleus and in the core of the native state slows folding and decreases protein stability. In contrast, increasing the size of hydrophobic side chain (by amino-acid substitution) within the core can selectively alter the stability of the native state relative to the transition state. Increasing the hydrophobicity of solvent-exposed residues in the native state can also destabilize the native state by the so-called reverse hydrophobic effect, which induces non-native hydrophobic clusters in the unfolded states. The role of hydrophobicity in the mechanism of folding, therefore, depends on the role and environment of individual residues at each stage of folding.

Computational simulations hold great promise as a tool for investigating the effect of the stability of secondary structure and hydrophobic-core on folding rates and mechanisms. This can be implemented by simply varying interaction energy of key structural elements, such as the α-helix and hydrophobic-core. These *in silico* methods have the advantage of being able to vary the secondary-structure stability and hydrophobicity continuously, without affecting the geometry of the original protein referred to as the **wild type**. In contrast, in protein mutation experiments it is only possible to vary these effects discretely, and the accompanying amino-acid substitutions will change the local geometry of the protein – the protein is now a **mutant** with different primary structure from the **wild type**. Various groups have implemented this *in silico* mutation method with C^{α} Go model by removing specific native interactions to destabilize certain structural elements (Li et al. 1994). This work uses an all-atomic model to perform

theoretical mutations at the atomic level (Chowdhury et al. 2004). The interaction energy of key structural elements are varied, instead of removed as in other studies.

Based on a coarse-grained model, Ding and co-works proposed that the folding dynamics of the Trp-cage is governed by a few key structural factors (Ding et al. 2005). These are the short α-helix from residue 2 through 8, and a hydrophobic-core including Tyr3, Trp6, Leu7, Pro12, 17, 18 and 19. Several experimental works proposed that the Pro-Trp interactions may be the key stabilizing factor (Neidigh et al. 2002, Gellman et al. 2002). These works reported that of the four Pro-Trp6 contact pairs, the Pro12-Trp6 is the most important, since it is formed even in the denatured (unfolded) states (Neidigh et al. 2002). The work of Zhou (Zhou R. 2003) concluded that an additional important stabilizing factor of the folded (native) state arises from the salt-bridge between residues Asp9 and Arg16. The salt-bridge was purposed to be the cause of the ultra-fast folding speed.

Here the roles of these key structural elements are studied, by tuning the interaction energy of atomic pairs with native contacts in these elements. The four structural elements to be tuned are the α-helix (residues from 2 to 8), hydrophobic-core (residues Tyr3, Trp6, Leu7, Pro12, 17, 18 and 19), Salt-bridge (Asp9 and Arg16) and Trp6-pro12. This leads to the determination of the structural elements that control the Trp-cage folding pathway, the calculation of the folding-rate, and an improved understanding of the coupling between secondary and tertiary structure.

4.2 **Method**

4.2.1 *Simulating trajectories*

One hundred 120 000 reduced time-step folding discontinuous molecular dynamics (DMD) simulations were performed at T* = 3.0 for every Trp-cage mutants starting from 100 initial random-coil conformations, produced by short simulations at a high temperature, T*=5.0, where the protein is unfolded. In a previous study it was shown that the transition temperature, at which the random coil and native state are equally stable, is T*=4.0 for a homogenous all-atom Go model of Trp-cage (Linhananta et al. 2005). In that work, it was estimated that 100 reduced time unit scales to ~ 20 ns to 70 ns.

4.2.2 *Wild-type, mutants, and the relative stability*

The wild type Trp-cage is defined by the homogeneous all-atom Go model (chapter 3), where all native contacts have the same square-well depth, $B^{w}_{i,j} = -1.0$ (see equation 2.3) or R=1.

The difference between a mutant model and the wild type model is the square-well depth of specific atom-atom native interactions in the structural element of interest. Denoting the atomic pair square-well depth of mutant contacts as $B^{m}_{i,j}$, the relative stability of the mutant structural elements, R, is defined as

$$R = B^{m}_{i,j} / B^{w}_{i,j} \qquad \text{(Eq. 4.1)}$$

Note that only native contacts within the structural elements take on this value, other native contacts of the mutant remains $B^{m}_{i,j} = B^{w}_{i,j} = -1$. For the former if $R > 1$, the specific interactions of the mutant are strengthened (stabilized), while for $R < 1$, it is weakened (destabilized). Setting R=0 mimics the mutation methods of other computational researchers (Li et al. 1994).

4.2.3 *Helicity defined by backbone dihedral angle*

A common method of determine whether a segment of a protein is helical is to calculate the dihedral angles φ and ψ in that segment (see Figs 1.4 and 1.6 and 1.7 and section 1.1.2). Following the prescription of Duan et al. (Duan Y. et al., 1998), define the probability of finding α-helical structures as Q_{dih}, which is a function of the Q,

$$Q_{dih}(Q) = M_{(\varphi\psi)}(Q) / M_{(\varphi\psi),nat} \qquad \text{(Eq. 4.2)}$$

where $M_{(\varphi\psi)}(Q) = \sum M_{(\varphi\psi),i}(Q)$;

$$M_{(\varphi\psi),i}(Q) = 1, \text{ if } |\varphi_i - \varphi_{i,nat}| < \delta \text{ and } |\psi_i - \psi_{i,nat}| < \delta$$

$$\text{or if } |\psi_i - \psi_{i,nat}| < \delta \text{ and } |\varphi_{i+1} - \varphi_{i+1,nat}| < \delta$$

$$= 0, \text{ in the other case.}$$

As was done by previously (Duan Y. et al., 1998) we set $\delta = 30^{\circ}$, for $2 \leq i \leq 8$ (residues in α-helix).

4.2.4 Mean First Passage Time τ and Folding rate, k

In this work a protein is considered to have folded to its native state if its main-chain RMSD is less than 1.6 Å. The first passage time (FTP) is defined as the time it takes for a protein to fold starting from an initial unfolded conformation. The Mean First Passage Time (MFPT), denoted by τ, is simply the average of the FTP, which in this chapter is simply the average FTP of the 100 Trp-cage simulations at T*=3.0. In this work we shall use the term average folding time and MFPT interchangeably. Finally, the folding rate is defined as $k = 1/\tau$.

4.3 Results

4.3.1 *Effect of altering α-helix stability on the folding kinetics and folding rates*

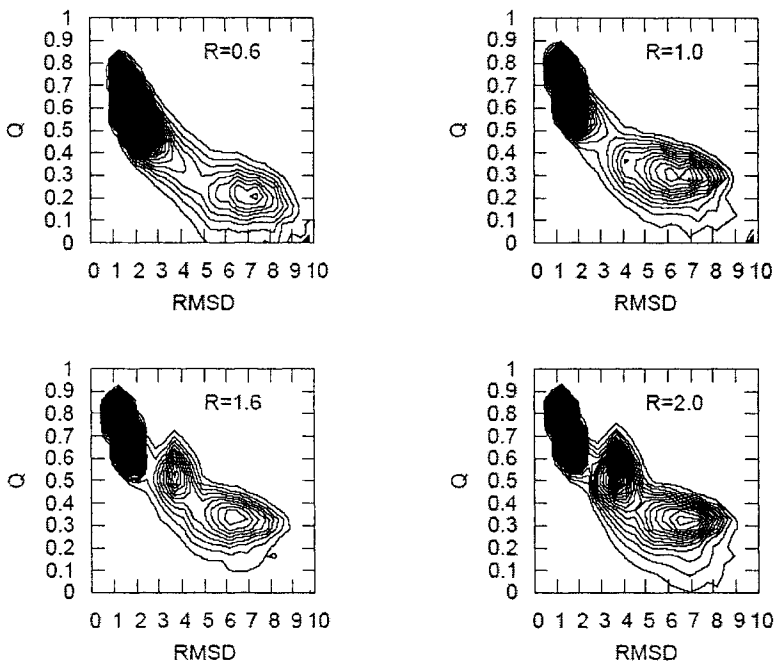


Figure 4.1. Probability distribution contour plots as a function of RMSD and the fraction of the total number of nonlocal native contacts formed, Q, for α-helix mutants. Regions of closely spaced lines indicate high occupation probability.

Fig. 4.1 is a set of contour plots of the probability distribution versus the reaction coordinates, in Q and RMSD space, with different stability of α-helix at the same temperature T*=3.0. The free energy landscape is smooth for low α-helix stability

(Figure 4.1a) indicating a two-state mechanism with the random coil and native states being the only two stable states. In contrast, increasing the stability of α-helix makes the landscape rougher (note the three regions with closely spaced lines in Figure 4.1c,d), inducing the appearance of metastable intermediate states.

The relationship between the folding rate and relative stability R of α-helix is shown in Figure 4.2. The data show that the wild-type Trp-cage (R=1.6) has the maximum value of folding rate. Decreasing or increasing the stability of α-helix will decrease the folding rate. This result is consistent with the experiment of Main et al. which strengthened the helix stability of a protein by adding the compound TFE (Main et al. 1999). Taken together the data suggests that two-state folding mechanisms do not necessary lead to the fastest folding rate, and that in certain cases the presence of metastable intermediate states actually aid folding (Zhou 2003; Linhananta et al. 2006)

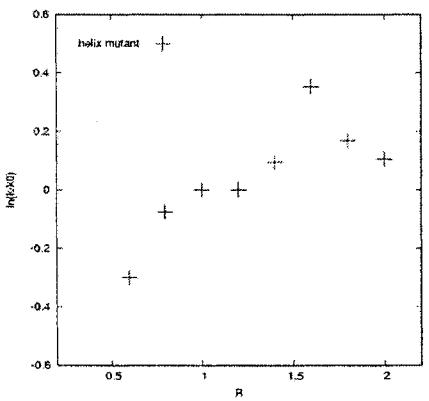


Figure 4.2. Mean folding-rate versus relative stability R of α-helix.

Figs. 4.3a and 4.4a show the probability of finding the α-helical structure (Q_{dih} (Q) of section 4.2.3) for a set of α-helix mutants. It is clear from the Figure 4.3A that α-helix (residue 2-8) reached the maximum as $Q \approx 0.3$-0.4. It is interesting that, in all cases, there is a valley ($Q \approx 0.45$) corresponding to a loss of helical structure prior to the protein folding to the native state ($Q > 0.65$). These results indicate that the early formed α-helix need to be softened in order to allow the formation of tertiary structures necessary to

reach the native state. It also explains why increasing the stability of α-helix too much decreases the folding rate.

Trp-cage was designed by decreasing the stability of α-helix through decreasing the length of helices of its mother protein EX4, which is often observed to be misfolded in water (Neidigh et al. 2002). We did 100 runs for different α-helix mutants at T*=3.0. The simulations results of the R=2.0 mutant demonstrate that only 90% trajectories folded to the native state. This is consistent with the experimental results (Neidigh et al. 2002) that found both EX4 (percentage helix 52.27) and TC3b (percentage helix 11.08) have higher stability of α-helix than that of Trp-cage (TC5b, percentage helix 1.29) but often do not fold to their native structure. (Percentage helix was obtained by using AGADIR, www.Gateway.com) On the other hand, the simulation results of R=0.6 α-helix mutant show that only 60% is folded (Table 4-1). This suggests that for optimum folding efficiency, the stability of α-helix must be finely tuned – it cannot be too high or too low.

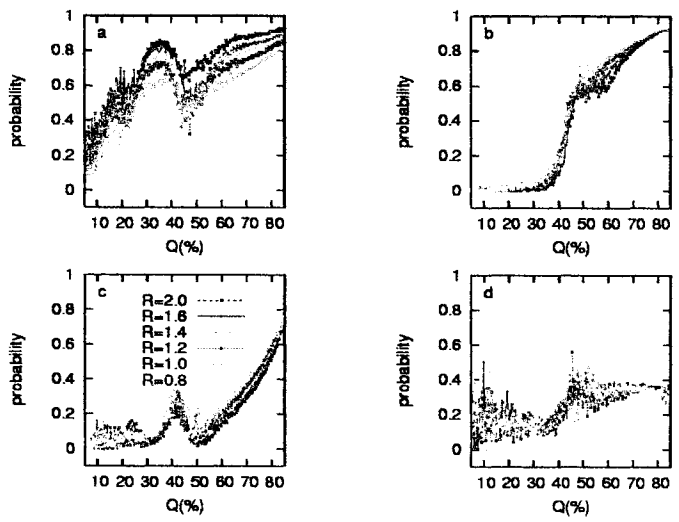


Figure 4.3. α-Helix Mutants affect the probability of key elements as a function of Q: α-Helix (a), Trp6-Pro18 (b), salt-bridge (Asp9-Arg16) (c) and 310-helix (d).

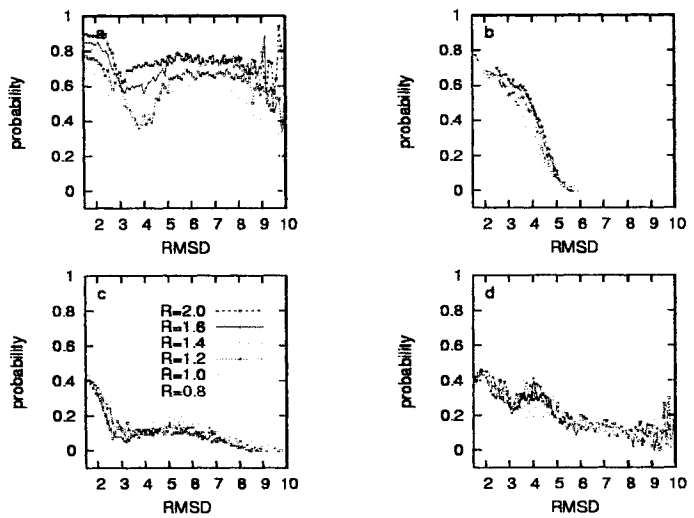


Figure 4.4. α-Helix Mutants affect the probability of key elements as a function of RMSD: α-Helix (a), Trp6-Pro18 (b), salt-bridge (Asp9-Arg16) (c) and 310-helix (d).

A cluster analysis is performed on the 10 trajectories that did not fold to the native states of the R=2.0 α-helix mutant. The representative conformations of the most populated cluster for the 10 unfolded trajectories are shown in Figure 4.5 and Table 4-1A. These conformations are compact with average RMSD ≈ 3.5 and Q ≈ 0.55. Native like α-helix and hydrophobic-core are well formed but the probability of salt-bridge formation is very low (0.1). Close inspection of these intermediate structures show that nonnative Trp6-Pro12 contacts impede further formation of the hydrophobic core, as well as preventing the formation of the salt-bridge between Asp9 and Arg16.

A cluster analysis on the 90 trajectories (of the R=2 α-helix mutant) that did fold found that 80 runs fold in diffusion collision folding mechanism, 9 fold by the downhill folding mechanism, and only 1 folds by the hydrophobic collapse mechanism (Table 4-3). It appears that very stable α-helical structure steers the Trp-cage model to the diffusion collision pathway. The representative conformations of the three most populated clusters in R=2.0 α-helix mutant diffusion collision folding mechanism (80 runs) are shown in Figure 4.6 and Table 4-1B.

Table 4-1A: Summary of the most populated clusters for intermediate states in failed-folding trajectories --cluster analysis (3991 conformations) for $R_{\alpha\text{-helix}}$ = 2.0.

	size	RMSD	Q	Qhelix	Qsb	Q_{310}
1	350	3.63	0.57	0.88	0.09	0.33
2	237	3.92	0.57	0.88	0.13	0.27
3	179	2.97	0.51	0.89	0.00	0.39
4	179	3.81	0.59	0.88	0.16	0.16
5	131	2.89	0.51	0.88	0.00	0.30
6	104	3.67	0.60	0.88	0.19	0.13
7	88	3.48	0.54	0.88	0.03	0.38
8	82	3.62	0.58	0.87	0.21	0.23

Table 4-1B: Summary of the most populated clusters for intermediate states in Diffusion-collision (6557 conformations), $R_{\alpha\text{-helix}}$ = 2.0

	size	RMSD	Q	Qhelix	Qsb	Q_{310}
1	101	3.76	0.43	0.23	0.00	0.39
2	44	5.50	0.38	0.93	0.23	0.05
3	28	5.84	0.40	0.99	0.33	0.07
4	28	5.53	0.38	0.97	0.20	0.04
5	27	5.61	0.39	0.96	0.26	0.07
6	26	5.87	0.36	0.96	0.08	0.04
7	25	4.99	0.41	0.99	0.50	0.08
8	25	5.41	0.38	0.98	0.26	0.04

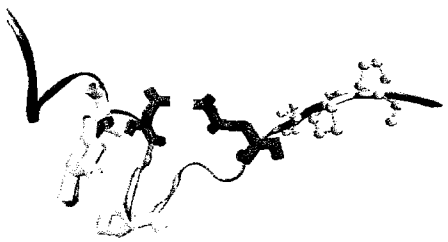
(A)

(B)

(C)

Figure 4.5. Representative intermediate structures of the 10 trajectories that did not fold for R=2.0 α-helix mutant. (Blue for α-helix; Red for Salt-bridge; Golden for Trp6 and Pro12; CPK for Pro17 to 19)

(A)



(B)

(C)

Figure 4.6. Representative intermediate structures of the 10 trajectories that did not fold for R=2.0 α-helix mutants. (Blue for α-helix; Red for Salt-bridge; Golden for Trp6 and Pro12; CPK for Pro17 to 19)

4.3.2 *Effect of altering hydrophobic-core stability on the folding kinetics and folding rates*

Figure 4.7 is a set of contour plots of the probability distribution versus the reaction coordinates, Q and RMSD with different stability of hydrophobic-core at the same temperature T*=3.0. Increasing the stability of hydrophobic-core makes the landscape rougher (Figure 4.7c, d), inducing a metastable intermediate state. An interesting feature is that the intermediate state is very close (as indicated by their RMSD and Q values) to the native state.

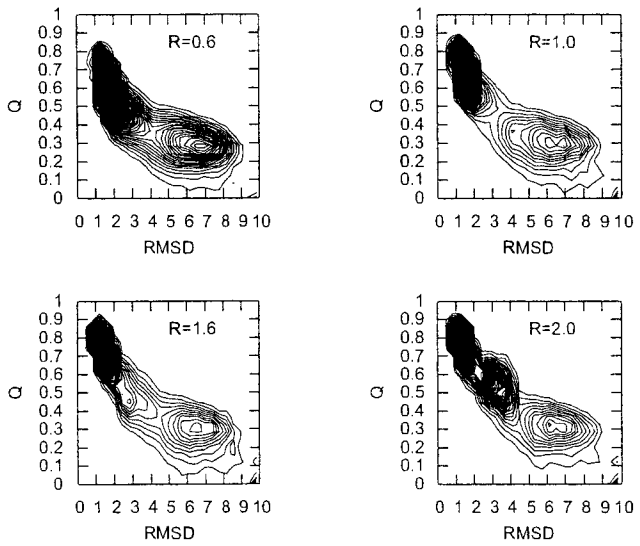


Figure 4.7. Probability distribution as a function of RMSD and the fraction of the total number of nonlocal native contacts formed, Q, for hydrophobic-core mutants.

The relationship between the folding rate and relative stability R of hydrophobic-core is shown in Figure 4.8. The data show that slightly strengthening the stability of hydrophobic-core (R=1.2~1.4) maximizes the folding rate. But, if the hydrophobic-core is too stable (large R), the folding rate will decrease. It is difficult to analyze the effects of stability of hydrophobic-core on folding mechanisms because the core consists of many native contact pairs: Trp6-Pro12, Trp6-Pro17-19. In the wild type Trp-cage model (chapter 3), these pair are formed at different stages of folding. We shall discuss this complex issue later.

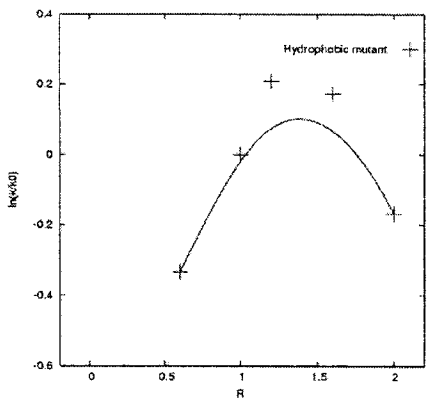


Figure 4.8. Mean folding-rate versus relative stability R of hydrophobic-core.

4.3.3 *Effect of altering salt-bridge on the folding kinetics and rates*

The contour maps of a set salt-bridge mutants are shown in Figure 4.9. Just as for the wildtype contour plot (Fig.3.4a) when the salt bridge is strengthened (R=1.2, 1.6, 2.0) a kinetic intermediate state (Q=0.3-0.4, RMSD=4-4.8) is observed, while when it is weakened (R = 0) the intermediate state is absent (Fig. 4.11a). As R increases the structures of the intermediate state becomes more nativelike. However, there is a significant drop in the probability distribution of the intermediate state as R is increased to 2 (Fig. 4.11d). It is noteworthy that the intermediate state induced by strengthening the salt bridge is relatively unstable when compared to the intermediate states induced by strengthening the α-helix (Fig. 4.1) and hydrophobic core (Fig. 4.7).

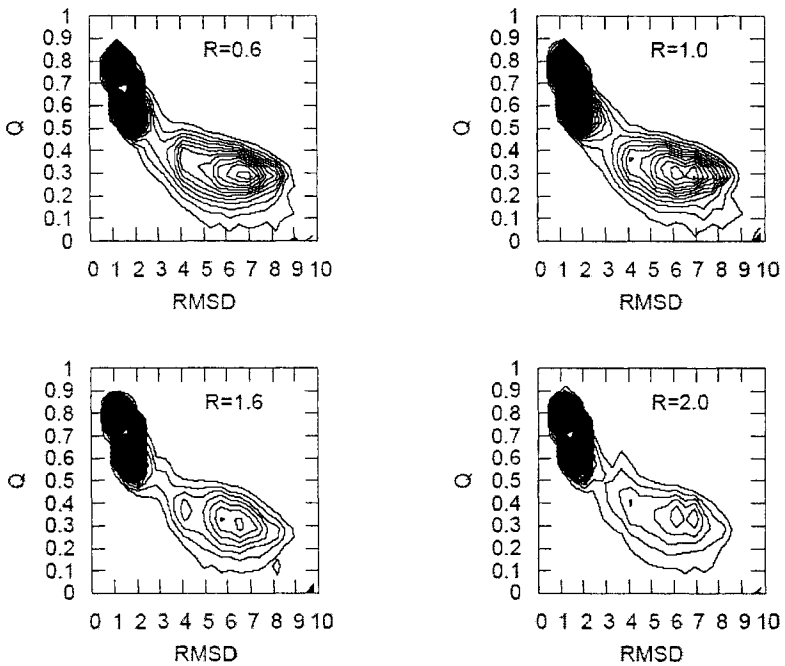


Figure 4.9. Probability distribution as a function of RMSD and the fraction of the total number of nonlocal native contacts formed, Q, for salt-bridge mutants.

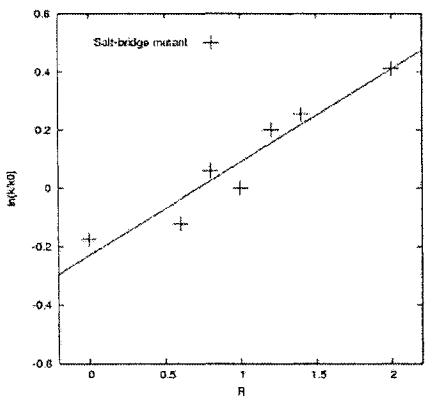


Figure 4.10. Relative mean folding-rate ln (k_m/k_w) vs. relative stability R of salt-bridge.

Fig. 4.10 plots ln (k_m/k_w) vs. R., where k_w is the folding rate of the wildtype Trp-cage and k_m is the folding rate of the mutant at a given R. This shows that the folding rate increases with R. Taken together the data presented in Figs. 4.9 and 4.10 suggests that the

Asp9-Arg16 saltbridge of the Trp-cage alters the free-energy landscape minimally by inducing a weakly stable intermediate state that is very difficult to detect on contour plots. However, these intermediate states have the effect of significantly increasing the folding rate. This is a remarkable agreement with the hypothesis of Zhou (Zhou 2003) that attributed the ultra-fast folding speed of the Trp-cage to the formation of a salt-bridge stabilized intermediate state.

Figs. 4.11 and 4.12 show the effects of changing the strength of Asp9-Arg16 interaction on the folding kinetics of key structural elements. Strengthening the Asp9-Arg16 by increasing R does not appear to have significant effects on the α-helix and the Trp6-Pro18 contacts (Figs. 4.11a, 4.11b, 4.12a and 4.12b). It appears that with increasing R slightly decrease the stability of the 3_{10}-helix (Figs. 4.11c and 4.12c). The main effect of strengthening the Asp9-Arg16 is the large increase in the stability (Figs. 4.11c and 4.12c) of salt bridge. These results is in agreement with the earlier conclusion that the salt bridge does not play an important role in determining the folding pathway of the Trp-cage.

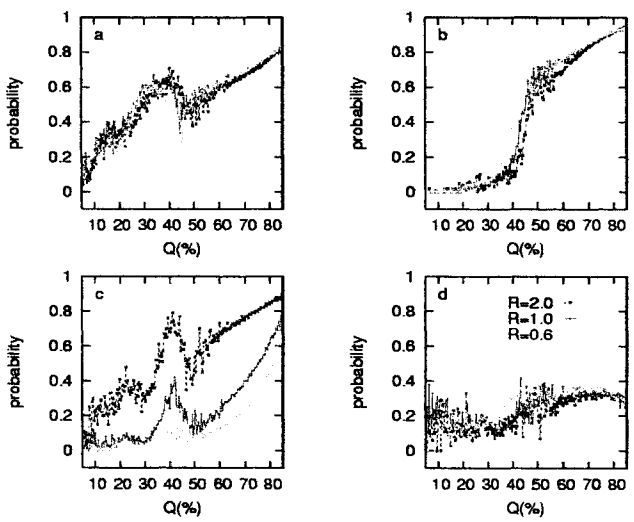


Figure 4.11. Salt-bridge Mutants effect on the probability of key elements as a function of Q: α-Helix (a), Trp6-Pro18 (b), salt-bridge (Asp9-Arg16) (c) and 310-helix (d).

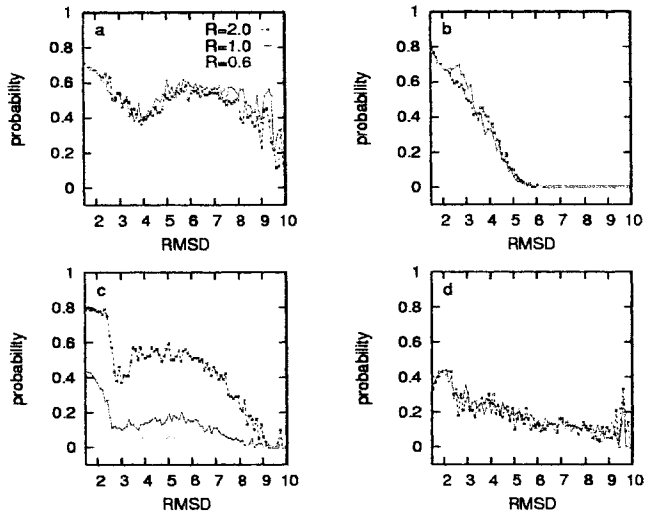


Figure 4.12. Salt-bridge Mutants effects on the probability of key elements as a function of RMSD: α-Helix (a), Trp6-Pro18 (b), salt-bridge (Asp9-Arg16) (c) and 310-helix (d).

4.3.4 *Effect of altering Trp6-Pro12 on the folding kinetics and rates*

As mention the Trp-cage hydrophobic core is stabilized by the native contacts pairs Trp6-Pro12, Trp6-Pro17, Trp6-Pro18, and Trp6-Pro19. Four Pro residues play a critical role in Trp-cage folding mechanism. Among the four Trp6-Pro pairs, Trp6-Pro12 is the only one detected in the denatured state by an NMR study (Neidigh et al. 2002). This has led to the suggestions that the Trp6-Pro12 native contact pair is a key structural element in the folding of the Trp-cage. This is investigated by tuning the strength of Trp6-Pro12 interactions.

The contour maps of a set of Trp6-Pro12 mutants are shown in Figure 4.13. In the case where the Trp6-Pro12 mutant is strengthened to R=2 there is a dramatic change in the free energy landscape with the appearance of two metastable intermediate state. This roughening of the free energy landscape is accompanied by a 50% decrease in the folding rate was observed (Fig. 4.14). Our result suggests that the existence of Trp6-Pro12 in the unfolded state may impede the folding process. Recently, by replacing Pro12 residue with Trp12 to produce the Trp^2-cage mutant, Gai's group claims to almost reach the "folding speed limit" (Bunagan et al. 2006). Our computational result is in general agreement with

a previous computational research that concluded that a slightly roughened free-energy landscape aid folding, but that a very rough landscape contains traps that impede folding.

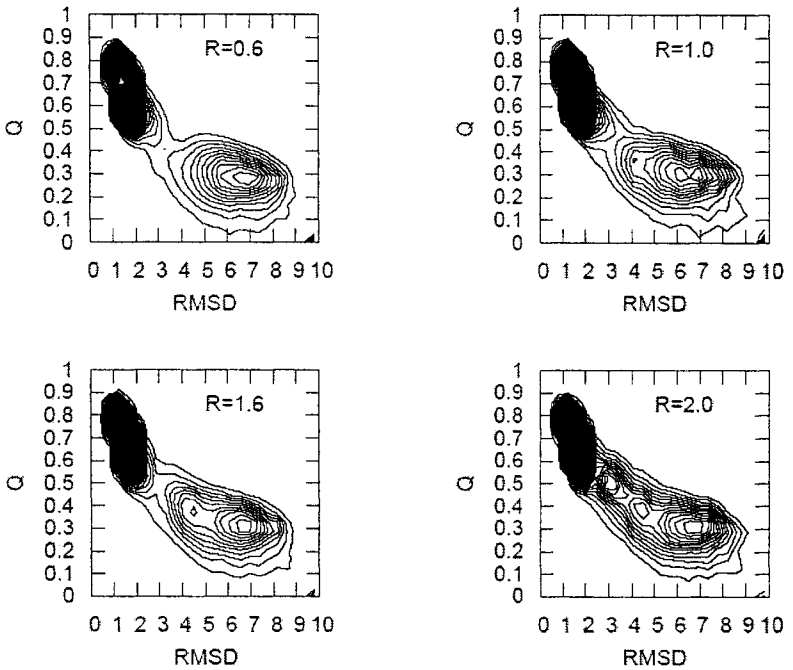


Figure 4.13. Probability distribution as a function of RMSD and the fraction of the total number of nonlocal native contacts formed, Q, for Trp6-Pro12 mutants.

For the R=2 Trp6-Pro12 mutant, 98 out of 100 folding simulations fold to the native state. Cluster analysis of the 98 folded trajectories shows that simulations can be classified into three mechanisms (Table 4-3): (I) diffusion collision (66 runs); (II) hydrophobic collapse (9 runs); (III) downhill (23). Two points are evident: the increase in number of trajectories following downhill folding pathway by more than a factor of two compare to the wildtype; and the significant decrease in the number of trajectories folding by the collision-diffusion mechanism. By comparing cluster analysis result for the wildtype (Tables 3-1 and 3-2) with cluster analysis results for the R=2 Trp6-Pro12 mutant (Table 4-2A and Table 4-2B), the α-helical content of the intermediate states of the diffusion collision and hydrophobic collapse pathways are more similar in the mutant. Compare to the wild type intermediates, the α-helical content of the mutant intermediates is lower in the diffusion collision pathway but higher in the hydrophobic collapse pathway. This shows that strengthening the Trp6-Pro12 interactions unified the folding

pathways such that the diffusion collision and hydrophobic collapse become more similar. It appears that, of all the structural elements tuned in this work, the folding pathways are altered most dramatically by varying the Trp6-Pro12 contacts.

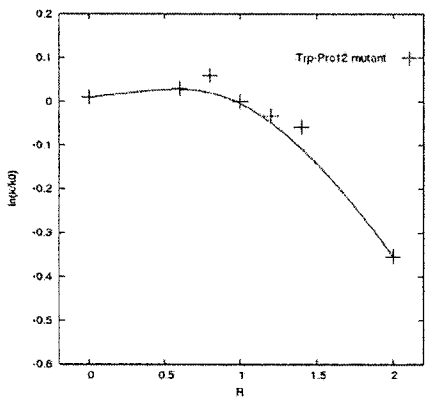


Figure 4.14. Relative mean folding-rate ln (k_m/k_w) vs. relative stability R of Trp6-Pro12 contacts.

Table 4-2A: Summary of the most populated clusters for metastable intermediate states in Diffusion collision mechanism --cluster analysis (6690 conformations) for $R_{Trp6-Pro12}$ = 2.0.

	Size	RMSD	Q	Qhelix	Qsb	Q_{310}
1	73	5.84	0.37	0.79	0.33	0.07
2	43	5.65	0.37	0.75	0.46	0.02
3	38	5.76	0.36	0.77	0.25	0.13
4	30	5.91	0.38	0.82	0.44	0.03
5	29	5.32	0.37	0.78	0.51	0.03
6	29	5.69	0.37	0.81	0.32	0.10
7	28	5.58	0.37	0.78	0.34	0.00
8	28	5.99	0.37	0.80	0.30	0.00

Table 4-2B: Summary of the most populated clusters for metastable intermediate states in hydrophobic-collapse mechanism --cluster analysis (2500 conformations) for $R_{Trp6-Pro12}$ = 2.0.

	Size	RMSD	Q	Qhelix	Qsb	Q_{310}
1	62	3.90	0.47	0.70	0.00	0.48
2	37	4.38	0.46	0.76	0.00	0.46
3	25	3.83	0.49	0.67	0.00	0.12
4	24	3.96	0.49	0.70	0.00	0.25

4.4 **Discussion and Conclusion**

To conclude this chapter we consider again the controversial debate on whether proteins fold by the framework mechanism, which includes both the diffusion-collision and hydrophobic-collapse pathways, or by the nucleation-condensation pathway. To reconcile

this debate, Fersht et al. proposed that decreasing the stability of the secondary structure will steer the folding mechanisms from the framework to the nucleation-condensation mechanisms (Gianni et al. 2003). To use our models to test this hypothesis, consider again Q_{helix} and Q_{hydro}, defined in section 3.2.4. Q_{helix} is the average fraction of the three α-helix hydrogen bonds formation. Q_{hydro} is the average fraction of tertiary hydrophobic contacts in the Trp-cage core. A plot of Q_{helix} vs. Q_{hydro} would reveal the folding mechanisms followed by the protein. In the diffusion-collision mechanism, Q_{helix} would increase more rapidly than Q_{hydro}. In the hydrophobic collapse mechanism, Q_{hydro} would increase more rapidly than Q_{helix}. In the nucleation-condensation mechanism, the Q_{helix} vs. Q_{hydro} curve would be a diagonal line. For a multiple folding process the Q_{helix} vs. Q_{hydro} plot would be complex.

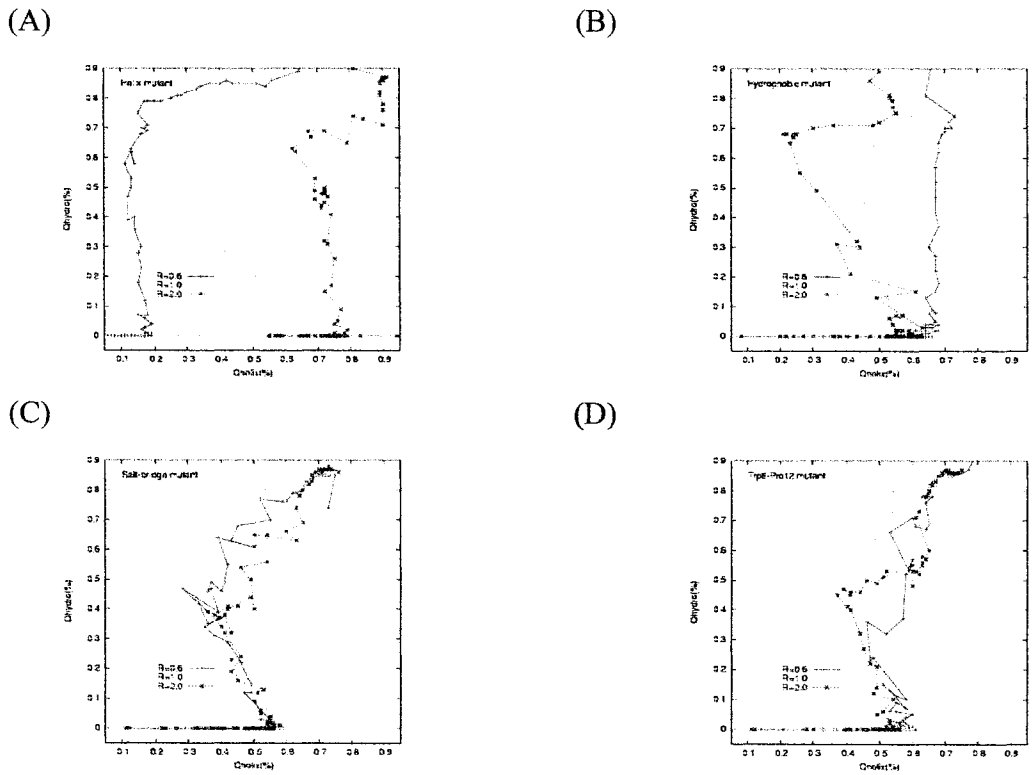


Figure 4.15. The trajectory of Q_{hydro} vs. Q_{helix} for different relative stability of α-helix and hydrophobic-core. (A) for α-helix mutants; (B) for hydrophobic-core mutants ; (C) for salt-bridge mutants; (D) for Trp6-Pro12 mutants. The Qhydro and Qhelix are the average values calculated from 100runs, corresponding certain RMSD. The line is plotted with points connected sequentially in RMSD decreasing.

To test the hypothesis of Fersht, the stabilities of the structural elements are systematically varied as described in section 4.2.2. The results of the Q_{helix} vs. Q_{hydro} plots are shown in Fig. 4.15. Fig. 4.15(A) shows the shift of folding mechanisms by varying the relative stability of α-helix. When the relative stability of α-helix is decreased, it is clear that the folding mechanism is a hydrophobic collapse. When it is increased the folding mechanism is a diffusion collision. However, the nucleation-condensation pathway is not observed for any value of the relative stability (R) of α-helix. Fig. 4.15(B) shows the shift of the folding mechanism by varying the relative stability of hydrophobic-core. When the relative stability of hydrophobic core is decreased, the folding mechanism is a diffusion collision. When it is increased, the Q_{helix} vs. Q_{hydro} plot is complex, indicating that the protein folds by multiple mechanisms. Again the nucleation-condensation pathway is not observed. Figs. 4.15(C) and (D) show that there is no significant shift of folding mechanisms by varying the relative stability of the salt-bridge and the Trp6-Pro12 contacts.

To conclude, the *in silico* method employed here shows that the Trp-cage can fold by three pathways: diffusion-collision, hydrophobic collapse and downhill. The nucleation condensation mechanism is not observed. Table 4.3 shows that the folding mechanism of the Trp-cage is very sensitive to the stability of specific structural elements. By tuning the relative stability, it is found that the folding pathways of the Trp-cage are controlled mainly by the balance between the stability of the α-helix and of the hydrophobic core. It is found that the high Trp-cage folding rate is maintained by the Asp9-Arg16 salt-bridge. It is also found that the key to reaching the "folding speed limit" may be found in the Trp6-Pro12 native contacts. Finally, the results presented here may be used as a guide by experimentalists to engineer Trp-cage mutants that fold faster or ones that fold by specific mechanisms.

Table 4-3. The population and average FPT for different folding mechanisms with different relative stability of key structural elements. Mechanism I, II, and III denote diffusion collision, hydrophobic collapse, and downhill mechanism, respectively.

α-Helix mutants:.

	Mechanism I	Mechanism II	Mechanism III	Unfolded	FPT
R=2.0	80% (150)	1% (180)	9% (30)	10%	139
R=1.6	86% (112)	1% (177)	9% (60)	4%	108
R=1.2	79% (149)	8% (352)	13% (67)	0%	154
R=1.0	79% (160)	9% (278)	12% (59)	0%	154
R=0.8	61% (175)	17% (515)	22% (117)	0%	166

Salt-bridge mutants:.

	Mechanism I	Mechanism II	Mechanism III	Unfolded	FPT
R=2.0	77% (102)	9% (214)	14% (44)	0%	101
R=1.6	80% (123)	6% (202)	14% (65)	0%	119
R=1.2	76% (124)	9% (258)	15% (72)	0%	126
R=1.0	79% (160)	9% (278)	12% (59)	0%	154
R=0.8	85% (162)	2% (515)	13% (64)	0%	145
R=0.6	76% (176)	9% (374)	15% (49)	0%	174

Trp6-Pro12 mutants:.

	Mechanism I	Mechanism II	Mechanism III	Unfolded	FPT
R=2.0	72% (178)	3% (661)	23% (89)	2%	199
R=1.0	79% (160)	9% (278)	12% (59)	0%	154
R=0.0	77% (169)	8% (167)	13% (50)	0%	145

Chapter 5

The effects of nonnative interactions on the Trp-cage folding kinetics

Abstract

The kinetics of an all-atom model for protein Trp-cage is investigated by varying the relative strength of homogenous and knowledge-based non-native interactions. The cluster analysis method is used to determine the effects of non-native interactions on the folding kinetics of the Trp-cage. It is found that non-native interactions play only a minor role in folding, which is a testament to the high optimization of the Trp-cage.

5.1 Introduction

It is now the consensus that the folding kinetics and thermodynamics of proteins are dominated by the native structure of proteins (Paxco et al. 1999, Baker 2000, Kogo and Takada 2001). This has inspired numerous computer-simulation studies based on Go models, the interaction potentials are biased to the known native structures of the proteins. Most Go models are homogeneous Cα Go models in which residue is presented by a single bead and in which the pairwise interaction energies are the same for all pairs. The successes of these coarse grain Cα Go models in reproducing the folding properties of many proteins illustrate that, for most proteins, it is the native topology that dominates folding mechanisms (Baker 2000, Koga and Takada 2001).

Despite the success of coarse-grained Cα Go models, recent works have shown that certain specific folding details can only be obtained with the use of ab-initio or all-atom Go models (Shimada et al. 2001, Linhananta 2005). In addition, recent experimental results indicate that for certain proteins non-native interactions often lead to the formation of non-native intermediate states, which are globular conformations that contain structural elements not present in the native state. For example, in a folding experiment on β-lactoglobulin, a metastable non-native α-helix, not observed in its native state, is detected (Hamada and Goto 1997). Non-native structural elements in the denatured states of spectrin SH3 domain have also been reported (Blamo et al. 1998; Viguera et al. 2002).

There have been several theoretical works on the effect of non-native interactions on the thermodynamics and folding kinetics of proteins. Karplus et al. using the CHARMM energy force field to assess models of proteins concluded that the non-native contribution to the energy of the TSE is between 12 and 20% (Paci et al. 2002). Zhou et al. use a Cα model with non-native interactions to probe the folding of a three-helix-bundle protein (Zhou and Karplus 1999a). With a lattice model Go-like

model, Shakhnovich and coworkers proposed that the negative or larger than unity Φ-values indicates non-native structures in the TSE (Li et al. 2000). The theoretical and computational results from Plotkin and coworkers demonstrated that weak non-native interactions can speed folding rate up with simulations of an off-lattice coarse-grained protein model (Plotkin 2001, Clementi and Plotkin 2004).

In Go models, a residue pair (or atomic pair) is native if the pair is in contact in the native state. Other pairs are designated as non-native. In the Go interaction potential native pairs are attractive, while non-native pairs are repulsive or do not interact with each other. Here it is clear in such models the protein structures are highly optimized toward the native states. Usually, Go-like models that include non-native interactions reduce the optimization by introducing attractive interactions between non-native pairs. In many cases, the non-native interactions are homogeneous, so that all non-native pairs have the same attractive energy (Zhou and Karplus 1999a; Plotkin 2001; Clementi and Plotkin 2004). This is an extreme assumption, since the specificity of the 21 amino acids suggests that the strength of the pairwise interactions depend on the identity of the interacting pairs. In this work we introduce an all-atom Go-like model with knowledge-based non-native interactions. The strength of the non-native atomic interactions is based on extensive experimental data on amino acids (Zhang et al. 1997). This will be used to determine the roles of non-native interactions in the folding pathways of Trp-cage.

5.2 Method

5.2.1 *Homogenous non-native potential*

Begin by considering the square-well depth of Eq. 2.3 of non-bonded atomic pairs interacting by van der Waals type potentials. Non-bonded atomic pairs with square-well overlaps in the ground state structure are designated native contact pairs. These native pairs have a square-well depth of Bn. All other pairs are designated as

non native a square-well depth of Bo. In the homogeneous all-atom Go model described in chapter 3, Bn = -1 and Bo = 0. Consider now the "bias gap" parameter

$$g = 1 - Bo/Bn \qquad \text{(Eq.5.1)}$$

used by Zhou et al. to studied non-native effects (Zhou and Karplus 1999a). For the highly optimized homogeneous Go model g = 1. Setting g = 0 is equivalent to Bn = Bo = -1, which is appropriate for a model of homopolymers. For $0<g<1$ and with Bn = -1 is equivalent to $-1 < Bo < 0$ for all non-native pairs. The latter parameters are appropriate for a model in which non-native interactions are attractive, but not as attractive as native interactions. In this Go-like model protein structures are still biased towards the ground state structures, but, depending on the strength of the non-native interactions, metastable non-native structural elements may appear along the folding pathways. In this model the strength of non-native interactions increase as g approaches zero.

5.2.2 *Knowledge-based non-native potential*

The main shortcoming of homogeneous non-native potentials is that they treat atomic pairs the same regardless of which amino acids they belong to. However, amino acids are distinct molecular units with heterogeneous interactions. To take this into account we use the "phenomenological" contact energy of atomic pairs of amino acids compiled by Delisi and coworkers paper (Zhang et al. 1997). The contact energies were determined by extensive analysis of experimental protein structures (from the protein data bank). These determined contact energies can be attractive or repulsive and are highly heterogeneous, with the interaction between two atoms of different amino acid residues depending on the identity of the residues. For example, the contact energy between a Cα atom of Leucine and a Cβ atom of Alanine can be different than that between a Cα atom of Serine and a Cβ atom of Isoleucine. To construct our knowledge-base non-native potential we set the non-native square-well depths of a non-native atomic ij pair, B_{ij}^{KB}, to the value in the normalized the matrix

in the reference (Table 3, Zhang et al. 1997). Now define the knowledge-based gap parameter g*

$$1\text{-}g^* = B^{nn\text{-}KB} / Bn \qquad \text{(Eq.5.2)}$$

with

$$B^{nn\text{-}KB} = (\Sigma\,(B_{ij}{}^{KB})^2)^{0.5}/N_{nn},$$

where N_{nn} is the total number of non-native atom-pair contacts which are not appear in the native states. Just like the bias gap model, the knowledge-base non-native interaction strength increases as g* decreases.

5.3 Results

5.3.1 *Homogenous Non-native Interactions*

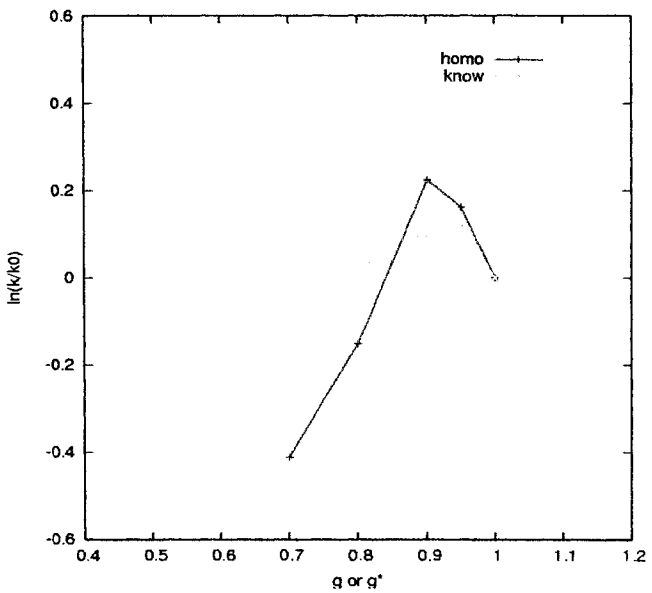


Figure 5.1. Mean folding rates vs. gap parameter (g or g*). Homo stands for homogenous attractive non-native interactions (g); Know stands for knowledge-based non-native interactions (g*).

For the homogeneous non-native interactions, Figure 5.1 shows that at first the folding rate increases, until it reaches its maximum rate at g = 0.9, after which the folding rate decreases linearly with g. we can see that the folding rate increases first with the increase of strength of homogenous attractive non-native interactions, and reaches a maximum at g=0.9. This behaviour can be understood with the contour plots

of Fig. 5.2. For weak non-native interaction g > 0.94 Figs. 5.2a,b show a smooth two-state free energy landscape, while at maximum folding rate parameter of g = 0.9 the landscape becomes rougher with the appearance of metastable intermediate states. For stronger non-native interactions, such as for g = 0.8, Fig. 5.2d shows a very rough landscape with highly stable (long-lived) intermediate state. Similar result has been found with off-lattice simulations of a Cα model of the src-SH3 domain (Clementi and Plotkin 2004) and lattice-simulations of a Go-like model with non-native interactions (Li et al. 2000, Fan et al. 2002). In those studies, it was concluded that weak non-native interactions speed up folding by inducing weakly stable non-native intermediates, but that if the non-native interactions are too strong the intermediate states become "folding traps" that slow folding.

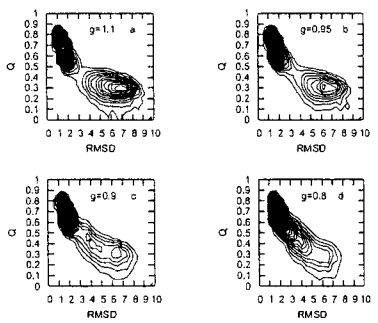


Figure 5.2 Contour maps of simulations at T*=3.0 for homogenous attractive non-native interactions with a) g =1.1, b) g = 0.95, c) g = 0.9, and d) g = 0.8.

Figure 5.3 shows contour maps of different folding pathways for the model with g = 0.8. The decomposition of trajectories into different pathways is determined by the cluster analysis method. The diffusion-collision (Fig. 5.3b) and hydrophobic collapse (Fig. 5.3 c) are similar to those of the wild type pathways discussed in Chapter 3. The appearance of Fig. 5.3d suggests trajectories of pathway III following downhill folding pathways (see Fig. 3.4d). However, cluster analysis reveals the presence of a long-lived intermediate state with partial helix and hydrophobic core at RMSD and Q values very close to that of the native state. The latter feature gives the appearance in Fig. 5.3d that the native state is broadly distributed in the Q and RMSD space. The average folding time of pathway III is 13400, substantially greater than the folding

time of about 5000 for trajectories folding by downhill pathway in the wild type model (g=1), which further indicates that this is not a downhill pathway.

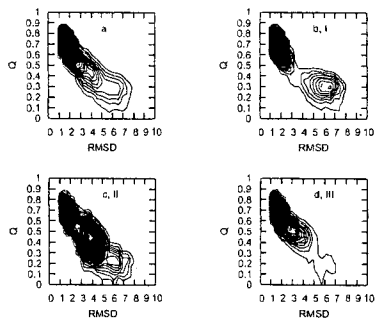


Figure 5.3. Decomposing the contour map of g=0.8 homogenous attractive non-native interactions in three different folding mechanisms determined by clustering analysis. (a) all trajectories (b) I, diffusion collision; (b) II, hydrophobic collapse; (c) III, partial helix and partial hydrophobic core

Table 5.1 reveals that the major effect of the homogeneous interactions is to increase the number of trajectories folding by the pseudo-downhill pathway III. For g < 9.4, the pathway is downhill-like with low MFPT (< 7000). Similar effects of homogeneous non-native interactions on folding mechanisms also have been reported, previously (Zhou and Karplus 1999a and 1999b). These results suggest that homogenous attractive non-native interactions help hydrophobic collapse but impede the formation of α-helix.

Table 5-1. The population and average FPT (MFPT in brackets) for different folding mechanisms with different relative stability of ***homogenous*** non-native interactions. Trajectories are classified by the cluster analysis method.

	path I	path II	path III	Unfolded	MFPT
g=1.1	88% (168)	3% (269)	9% (52)	0	161
g=1.0	79% (160)	9% (278)	12% (59)	0	154
g=0.95	72% (148)	3% (220)	25% (67)	0	131
g=0.9	54% (117)	12% (293)	34% (76)	0	123
g=0.8	49% (137)	12% (515)	38% (134)	1	179
g=0.7	29% (237)	18% (402)	50% (173)	3	232

5.3.2 *Knowledge-based non-native interactions*

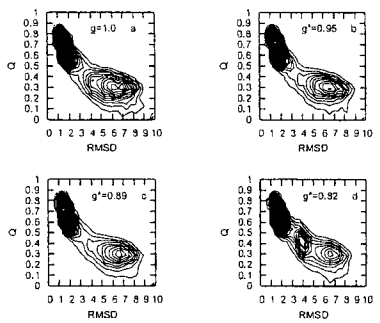


Figure 5.4. Contour maps of simulations at T*=3.0 for knowledge-based non-native interactions with a) g* =1, b) g* = 0.95, c) g* = 0.89, and d) g* = 0.82

Figure 5.4 shows a set of contour maps with different strength of knowledge-based non-native interactions. No significant effects are observed for g* <=0.89 (Figure 5.4(b) and (c)), where the free energy landscapes remain two state. For g*=0.82 the landscape becomes rougher with the appearance of an intermediate state (see Figure 5.4(d)). Table 5.2 summarizes the results of cluster analysis on folding simulations of Go-like models with knowledge-based non-native interactions. Just as for models with homogeneous non-native interactions, the main effect of the knowledge-based interactions is to steer folding trajectories toward the downhill folding pathway. The main difference is that the pathways remain "true" downhill pathways, with cluster analysis detection no intermediate state, prior to the native states. This is corroborated by the fact that MFPT < 7000 along the downhill pathway for all values of g*. Another feature is that the hydrophobic collapse pathway is very slowly folding with MFPT = 41800, which contributes to the slight increase in the overall average folding time of the Trp-cage model.

Table 5-2. The population and average FPT (MFPT in brackets) for different folding mechanisms with different relative stability of ***knowledge-based*** non-native interactions.

	path I	path II	path III	Unfolded	MFPT
g*=1.0	79% (160)	9% (278)	12% (59)	0	154
g*=0.95	72% (149)	6% (201)	22% (71)	0	136
g*=0.89	71% (157)	6% (202)	23% (56)	0	141
g*=0.82	64% (132)	9% (418)	26% (69)	1	149

5.4 Discussion and Conclusion

In conclusion, for weak non-native interactions, the main effects are an increase in the average folding speed, which is due mainly by the fact that folding trajectories are steered into pseudo-downhill pathways. For low g and g*, the pathway is similar to the wild type pathway with not detectable intermediate state prior to the native state, and a low MFPT < 7000. For strong non-native strength, the behaviours of the models diverge. In the homogeneous model with g = 0.8, the pseudo-downhill pathway possesses (III) an intermediate state with values of Q and RMSD similar to the native state, giving the appearance of a broadly distributed native state as seen in Fig. 5.3d. Here the behaviour is reminiscent of glassy behaviour in systems of homopolymers, where trajectories fold to unspecific compact states along unspecific pathways on roughened free-energy landscapes. In contrast, cluster analysis on trajectories following downhill-folding pathway for g*=0.82 (comparable in magnitude to g = 0.8 of the homogeneous model) detects no intermediate, indicating that the pathway remains downhill. Also there appear on the total contour map (Fig. 5.4d) partially folded intermediate states, quite distinct from the native state, which indicate that the protein model exhibits behaviours commonly attributed to real proteins. Taken together, the results of this chapter suggest that knowledge-based non-native interactions be used to improve the accuracy of Go-like models of proteins.

References

Ahmed Z., Beta I.A., Mikonin A.V., and Asher S.A. **2005**. UV-Resonance Raman thermal unfolding study of Trp-cage shows that it is not a simple two-state miniprotein. *JACS* 127: 10943-10950.

Alder B.J., and Wainwright T.E., **1959**. Studies in molecular dynamics I. General method. *J. Chem. Phys.* 31: 459-466.

Allen M.P., and Tildesley D.J., **1987**. Computer simulation of liquid. Oxford Science Publication.

Alm E., and Baker D. **1999**. Matching theory and experiment in protein folding. Current Opinion in Structural Biology 9:189-196.

Anderson H.C., **1980**. Molecular dynamics simulations at constant pressure and/or constant temperature. *J. Chem. Phys.* 72: 2384-2393.

Bai Y., **2003**. Hidden intermediates and levinthal paradox in the folding of small proteins. *Biochemical and Biophysical Research Communications.* 305:785-788.

Baker D., **2002**. A surprising simplicity to protein folding. *Nature* 405:39-42.

Bieri O., and Kiefhaber T., **1999**. Elementary steps in protein folding. *Biol. Chem.* 380: 923-929.

Borreguero J.M., Ding F., Buldyrev S.V., Stanley H.E., and Dokholyan N.V., **2004**. Multiple foldingpathways of the SH3 domain. *Biophysical J.* 87: 521-533.

Branden C., and Tooze J., **1999**. Introduction to Protein Structure (2ed.) *Garland Publishing, Inc.*

Brooks B.R., Bruccoleri R.E., Olafson B.D., StatesD.J., Swaminathan S., and Karplus M., **1983**. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem* 4:187.

Bunagan M., Yang X., Saven J.G., and Gai F., **2006**. Ultrafast folding of a computationally designed Trp-cage mutant: Trp^2-cage. *J. Phys. Chem. B.*

Caflisch A., **2005**. Network and graph analyses of folding free energy surfaces. *Current Opinion in Structural Biology* 16:1-8.

Cavalli A., Haberthur U., Paci E., and Caflisch A., **2003**. Fast protein folding on downhill energy landscape. *Protein Science* 12: 1801-1803.

Chang I., Cieplak M., Banavar J.R., and Maritan A., **2004**. What can one learn from experiments about the elusive transition states? *Protein Science* 13: 2446-2457.

Chavez L.L., Onuchic J.N., and Clementi C., **2004**. Quantifying the roughness on the free energy landscape: Entropic bottlenecks and protein folding rates. *J. Am. Chem. Soc.* 126: 8426-8432.

Chikenji G., Fujitsuka Y., and Takada S., **2004**. Protein folding mechanisms and energy landscape of src SH3 domain studied by a structure prediction tool box. *Chem. Phys.* 307:157-162.

Chiti F., Taddei N., Webster P., Hamada D., Fiashi T., Ramponi G., and Dobson C.M., **1999**. Acceleration of the folding of acylphosphatase by stabilization of local secondary structure. *Nature Struc. Biol.* 6(4):380-387.

Chowdhury S., Lee M.C., Xiong G., and Duan Y., **2003**. An initio folding simulation of the Trp-cage mini-protein approaches NMR resolution. *J. Mol. Biol.*, 327:711-717.

Chowdhury S., Lee M.C., Duan Y., **2004**. Characterizing the Rate-limiting Step of Trp-Cage Folding by All- atom Molecular Dynamics Simulations, *J. Phys. Chem. B* 108:13855-13865.

Clementi C., Nymeyer H., and Onuchic J.N., **2000**. Topological and energetic factors: What determines the structural details of the transition state ensemble and "En-route" intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.* 298: 937-953.

Clementi C., Garcia A.E., and Onuchic J.N., **2003**. Interplay among tertiary contacts, secondary structure formation and side-chain packing in protein folding mechanism: All-atom representation study of protein L. *J. Mol. Biol.* 326: 933-954.

Clementi C. and Plotkin S., **2004**. The effects of nonnative interactions on protein folding rates: Theory and simulation. *Protein Science* 13:1750-1766.

Dagget V., and Fersht A., **2003**. The present view of the mechanism of protein folding. *Nature Review: Molecular Biology* 4:497-502.

DeMarco M.L., Alonso D.O.V., and Daggett V., **2004**. Diffusing and colliding: the atomic level folding/unfolding pathway of a small helical protein. *J. Mol. Biol.* 341: 1109-1124.

Dill K.A., **1990**. Dominat forces in protein folding. *Biochemistry* 29(31):7133-7155.

Dill K.A., Fiebig K.M., and Chan H.S., **1993**. Cooperativity in Protein-Folding Kinetics. PNAS 90: 1942-1946.

Dill K.A., Bromberg S., Yue K., Fiebig K.M., Yee D.P., Thomas P.D., and Chan H.S., **1995**. Principles of protein folding – A perspective from simple exact models. *Protein Science* 4:561-602.

Dill, K. A., and Chan, H. S. **1997**. From Levinthal to pathways to funnels. *Nature Struct. Biol.* 4: 19.

Dill K.A., **1999**. Polymer principles and protein folding. *Protein Science* 8: 1166-1180.

Ding F., Dokholyan N.V., Buldyrev S.V., Stanley H.E., and Shakhnovich E.I., **2002**. Direct molecular dynamics observation of protein folding transition state ensemble. *Biophysical J.*, 83:3525-3532.

Ding F., Buldyrev S.V., and Dokholyan N.V., **2005**. Folding Trp-cage to NMR resolution native structure using a coarse-grained protein model. *Biophysical J.* 88:147-155.

Dobson C.M., **2003**. Protein folding and misfolding. *Nature* 426:884-890.

Dokholyan N.V., **2005**. Studies of folding and misfolding using simplified models. *Current Opinion in Structural Biology* 16: 1-7.

Du R., PandeV.S., Grosberg A.Y., Tanaka T., and Shakhnovich E.I., **1998**. On the transition coordinate for protein folding. *J. Chem. Phys.* 108: 334-350.

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

Duan Y., and Kollman P.A., **2001**. Computational protein folding: from lattice to all-atom. *IBM Systems J.* 40(2): 297-309.

Eaton W.A., **1999**. Searching for "downhill scenarios" in protein folding. *PNAS* 96: 5897-5899.

Ejtehadi M.R., Avall S.p., and Plotkin S.S., **2004**. Three-body interactions improve the prediction of rate and mechanism in protein folding models. *PNAS* 101(42): 15088-15093.

Fan K., J. Wang, and W. Wang, **2002**. Folding of lattice protein chains with modified Go potential. *Eur. Phys. J. B.* 30: 381-191.**Feng** H., Vu N.D., Zhou Z., and Bai Y., **2004**. Structural examination of value analysis in protein folding. *Biochemistry* 43: 14325-14331.

Ferara P., and Caflisch **2001**. Native topology or specific interactions: What is more important for protein folding? *J. Mol. Biol.* 306:837-850.

Fersht, A. R., Itzhaki, L. S., elMasry, N. F., Matthews, J. M. & Otzen, D. E. **1994**. Single versus parallel pathways of protein folding and fractional formation of structure in the transition state. *PNAS* 91: 10426-9.

Fersht A.R., **1995**. Optimization of rates of protein folding : the nucleation-condensation mechanism and its implications. *PNAS* 92: 10869-10873.

Fersht A.R. **1999**. Structure and mechanism in protein science. (W.H. Freeman, New York).

Fersht A.R., and Daggett V., **2002**. Protein folding and unfolding at atomic resolution. *Cell* 108:1-20.

Fersht A.R., and Sato S., **2004**. Φ-values analysis and the nature of protein-folding transition states. *PNAS* 101(21): 7976-7981.

Gellman S.H., and Woolfson D.N., **2002**. Mini-proteins Trp the light fantastic. *Nature Struct. Biol.* 9(6):408-410.

Gianni S., Guydosh N.R., Khan F., Caldas T.D., Mayor U., White G.W.N., DeMarco M.L., Daggett V., and Fersht A.R., **2003**. Unifying features in protein-folding mechanism. *PNAS* 100(23):13286-13291.

Go N, **1983**. Theoretical studies of protein folding. *Annu. Rev. Biophys. Bioeng.* 12: 183-210.

Gruebele M., **2005**. Downhill folding: evolution meets physics. C.R. Biologies 328: 701-712.

Gsponer J., and Caflisch A., **2002**. Molecular dynamics simulations of protein folding from the transition state. *PNAS* 99(10): 6719-6724.

Hagen S. J., Hofrichter J., Szabo A., and Eaton, W. A., **1996**. Diffusion-limited contact formation in unfolded cytochrome c: estimating the maximum rate of protein folding. *PNAS*, 93:11615-11617.

Hagen S.J., Qiu L., Pabit S.A., **2005**. Diffusion limits to the speed of protein folding: fact or friction? *J. Phys.: Condens. Matte*r 17 S1503-S1514.

Hamada D., Chiti F., Guijarro J.I., Kataoka M., Taddel N., and Dobson C.M., **2000**. Evidence conserning rate-limiting steps in protein folding from the effects of trifluoroethal., *Nature Struc. Biol.* 7(1):58-61.

Honig B., **1999**. Protein folding: from the Levinthal paradox to structure prediction. *J. Mol. Biol.* 293:283-293.

Islam S.A., Karplus M., and Weaver D.L., **2002**. Application of the Diffusion–Collision Model to the Folding of Three-helix Bundle Proteins. *JMB* 318: 199-215.

Itzhaki L.S., Otzen D.E., and Fersht A.R., **1995**. The Structure of the Transition State for Folding of Chymotrypsin Inhibitor 2 Analysed by Protein Engineering Methods: Evidence for a Nucleation-condensation Mechanism for Protein Folding. *JMB* 254 :260-288.

Jemth P., Gianni S., day R., Lin B., Johnson C.M., Dagget V., and Fersht A.R., **2004**. Demonstration of a low-energy on-pathway intermediate in a fast-folding protein by kinetics, protein engineering, and simulation. *PNAS* 101(17):6450-6455.

Jewett A.I., Pande V.S., and plaxco K.W., **2003**. Cooperativity, smooth energy landscapes and origins of topology-dependent protein folding rates. *J. Mol. Boil.* 326: 247-253.

Karanicolas J., and Brooks III C.L., **2003a**. Improved Go-like mod4ls demonstrate the robustness of protein folding mechanisms towards non-native interactions. *J. Mol. Biol.* 334: 309-325.

Karanicolas J., and Brooks III C.L., **2003b**. The structural basis for biphasic kinetics in the folding of WW domain from a forming-binding protein: lessons for protein design? *PNAS* 100(7): 3954-3959.

Karpen M.E., Tobias D.J., and Brooks III C.L., **1993**. Statistical clustering techniques for the analysis of long molecular dynamics trajectories: analysis of 2.2-ns trajectories of YPGDV. *Biochemistry* 32: 412-420.

Kaya H., and H.S. Chan, **2003**. Solvation effects and driving forces for protein thermodynamic and kinetic cooperativity: how adequate is native-centric topological modeling? *J. Mol. Boil.* 326: 911-931.

Knott M., Kaya H., and Chan H.S., **2004**. Energetics of protein thermodynamic cooperativity: Contribution of local and nonlocal interactions. *Polymer* 45: 623-632.

Koga N., and Takada S., **2001**. Roles of native topology and chain-length scaling in protein folding: a simulation study with a Go-like model. *J. Mol. Biol.* 313:171-180.

Kouza M., Li. M.S., O'Brien E.P., Hu C, and D. Thirumalai, **2006**. Effect of finite size on cooperativity and rates of protein folding. *J. Phys. Chem.*. A 110: 671-676.

Kraulis P. **1991**. *MOLSCRIPT*: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* 24:946-950

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

Kuhlman B., and baker D., **2004**. Exploring folding free energy landscapes using computational protein design. *Current Opinion in Structural Biology* 14:89-95.

Laughlin R.B., Pines D., Schmalian J., Stojkvic B.P., and Wolynes P., **2000**. The middle way. *PNAS* 97(1):32-37.

Lazaridis T., and Karplus M., **1997**. "New view" of protein folding reconciled with the old through multiple unfolding simulations. *Science* 278:1928-1931.

Levinthal C., **1968**. Are there pathways for protein folding? *J. Chim. Phys.* 65:45-45.

Li A. and Daggett V., **1994**. Characterization of the transition state of protein unfolding by use of molecular dynamics: Chymotrypsin inhibitor 2. *PNAS* 91: 10430-10434.

Li H. and Zhou Y., **2005**. SCUD: fast structure clustering of decoys usinf reference state to remove overall rotation. *J. Comput. Chem* 26:1189-1192.

Li L., Mirny L.A., and Shakhnovich E.I. **2000**. Kinetics, thermodynamics and evolution of non-native interactions in a protein folding nucleus. *Nature Struct. Biol.* 7(4): 336-342.

Linhananta A., and Zhou Y., **2002**. The role of sidechain packing and native contact interaction in folding: discontinuous molecular dynamics folding simulations of an all-atom Go model of fragment B of staphylococcal protein A. *J. Chem. Phys.* 117(19): 8983-8995.

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

Linhananta A., Boer J., and MacKay I., **2005**. The Equilibrium Properties and Folding kinetics of An All-atom Go Model of the Trp-cage, *J. Chem. Phys.* 122, 114901-114914.

Liwo A., Khalili M., and Scheraga H.A., **2005**. Ab initio simulations of protein-folding pathways by molecular dynamics with united-residue model of polypeptide chains. *PNAS* 102(7): 2362-2367.

Main R.G. E. and Jackson E., **1999**. Does trifluoroethanol affect folding pathways and can it be used as a probe of structure in transition states? *Nature Struc. Biol.* 6(9):831-835.

Meisner W.K., and Sosnick T.R., **2004**. Fast folding of a helical protein initiated by the collision of unstructural chains. *PNAS* 101(37): 13478-13482.

Merlo C., Dill K.A., and Weikl T., **2005**. Φ-values in protein-folding kinetics have energetic and structural components. *PNAS* 102(29):10171-10175.

Neidigh J.W., Fesinmeyer R.M., and Anderson N.H., **2002**. Designing a 20-residue protein. *Nature Struc. Boil.* 9: 425-430.

Munoz V., **2002**. Thermodynamics and kinetics of downhill protein folding investigated with a simple statistical mechanical model. *Int. J. Quantum Chemistry*, 90: 1522-1528.

Munson M., Anderson K.S., and Regan L., **1996**. Speeding up protein folding: mutations that increase the rate at which rop folds and unfolds by over four orders of magnitude. *Folding & Design* 2191):77-87.

Neuweiler H., Doose S., and Sauer M. **2005**. A microscopic view of miniprotein folding: enhanced folding efficiency through formation of an intermediate. *PNAS* 102(46): 16650-16655.

Nguyen H., Jager M., Moretto A., Gruebele M., and Kelly J.W., **2003**. Tuning the free-energy landscape of a WW domain by temperature, mutation, and truncation. *PNAS* 100(7): 3948-3953.

Nikiforovich G.V., Anderson N.H., Fesinmeyer R.M., and Frieden C., **2003**. Possible locally driven folding pathways of TC5b, a 20-resdue protein. *Proteins: Structure, Function, and Genetics* 52:292-302.

Onuchic J.N., and Wolynes P.G., **2004**. Theory of protein folding. *Current Opinion in Structural Biology* 14: 70-75.

Oliveberg M., Tan Y., and Fersht A.R., **1995**. Negative activation enthalpies in the kinetics of protein folding. *PNAS* 92: 8926-8929.

Oliveberg M., **2001**. Characterisation of the transition states for protein folding: towards a new level of mechanistic detail in protein engineering analysis. *Current Opinion in Structural Biology* 11:94-100.

Ota M., Ikeguchi M., and kidera A., **2004**. Phylogeny of protein-folding trajectories reveals a unique pathway to native structure. *PNAS* 101:17658-17663.

Ozkan S.B., Bahar L. and Dill K.A., **2001**. Transition States and the Meaning of Φ–values in Protein folding Kinetics, *Nature Structural Biology* Vol. 8(9):765-769.

Paci E., Vendruscolo M., and Karplus M., **2002**. Native and Non-Native Interactions Along Protein Folding and Unfolding Pathway, *PROTEINS: Structure, Function, and Genetics* 47: 379-392.

Pande V.S., Grosberg A.Y., Tanaka T., and Rokhsar D.S., **1998**. pathways for protein folding: is a new view needed? *Current Opinion in Structural Biology* 8: 68-79.

Pande V.S., **2003**. Meeting halfway on the bridge between protein folding theory and experiment. *PNAS* 100(7): 3555-3556.

Park S.H., O'Neil K. T., and Roder H., **1997**. An early intermediate in the folding reaction of B1 domain of protein G contains a native-like core. *Biochemistry* 36: 14277-14283.

Pathria R.K., **1996**. Statistical Mechanics. *Butterworth-Heinemann.*

Petsko G.A., and Ringe D., **2004**. Protein Structure and Function. *New Science Press Ltd.*

Pitera J.W., and Swope W., **2003**. Understanding folding and design: Replica-exchange simulations of "Trp-cage" miniproteins. *PNAS* 100(13): 7587-7592.

Plaxco K.W., and Baker D., **1998a**. Limited internal friction in the rate-limiting step of a two-state protein folding reaction. *PNAS* 95: 13591-13596.

Plaxco K.W., Simons K.T., and Baker D., **1998b**. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Boil.* 277: 985-994.

Plotkin S.S., and Onuchic J.N., **2000**. Investigation of routes and funnels in protein folding by free energy fundtional methods. *PNAS* 97(12): 6509-6514.

Plotkin S.S., **2001**. Speeding protein folding beyond the Go model: How a little frustration something helps. *Proteins: Structure, Function, and genetics* 43:337-345.

Plotkin S.S., and Onuchic J.N., **2002a**. Understanding protei folding with energy landscape theory, part I: basic concepts. *Quarterly Review of Biophysics* 35(2): 111-167.

Plotkin S.S., and Onuchic J.N., **2002b**. Understanding protei folding with energy landscape theory, part II: Quantitative aspects. *Quarterly Review of Biophysics* 35(3): 205-286.

Qiu L., Pabit S.A., Roitberg A.E., and Hagen S.J. **2002**. Smaller and faster: The 20-residue Trp-cage protein folds in 4μs. *J. Am. Chem. Soc.* 124: 12952-12953.

Qiu L., and Hagen S.J., **2005**. Internal friction in the ultrafast folding of the trptophan cage. *Chemical Physics* 312:327-333.

Raleigh D.P. and Plaxco K.W., **2005**. The protein folding transition state: What are Φ–values really telling us? *Protein and Peptide letters* 12:117-122.

Ramachandran G.N., Ramakrishnan C., and Sasisekharan V., **1963**. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* 7:95-99.

Rao F., Settanni G., Guarnera E., and Caflisch A., **2005**. Estimation of protein folding probability from equilibrium simulations. *J. Chem Phys.* 122: 184901-5.

Rapaport D.C., **1980**. The event scheduling problem in molecular dynamics simulation. *J. Comput. Phys.* 34: 184-201.

Roccatano D., Colombo G., Fioroni M., and Mark A.E., **2002**. Mechanism by which 2,2,2-trifluoroethanol/water mixtures stabilize secondary-structure formation in peptides: A molecular dynamics study. *PNAS* 99(19):12179-12184.

Roder H., **2004**. Stepwise helix formation and chain compaction during protein folding. *PNAS* 101(7): 1793-1794.

Sadqi M., Lapidus L.J., and Munoz V., **2003**. How fast is protein hydrophobic collapse? *PNAS* 100(21): 12117-12122.

Sanchez I.E. and Kiefhaber T., **2003a**. Hammond Behavior versus ground state effects in protein folding: Evidence for narrow freeenergy barriers and residual structure in unfolded states. *J. Mol. Biol.* 327:867-884.

Sanchez I.E. and Kiefhaber T., **2003b**. Origin of unusual Φ-values in protein folding: Evidence against specific nucleation sites. *J. Mol. Biol.* 334: 1077-1085.

Schonbrun J., and Dill K.A., **2003**. Fast protein folding kinetics. *PNAS* 100(22): 12678-12682.

Schymkowita J., Rousseau F., and Serrano L. **2002**. Surfing on protwin folding energy landscapes. *PNAS* 99(25): 15846-15848.

Settanni G., Rao F., and Caflisch A., 2005. Φ-values analysis by molecular dynamics simulations of reversible folding. *PNAS* 102(3): 628-633.

Shakhnovich, E. I. **1998**. Folding nucleus: specific or multiple? Insights from lattice models and experiments. *Fold. Des.* 3: R108-11.

Shea J.E., Onuchic J.N., and Brooks C.L., **1999**. Exporing the origins of topological frustration: Design of a minimally frustrated model of fragment B of protein A. *PNAS*, 96(22):12512-12517.

Shea J.E., Onuchic J.N., and Brooks C.L., **2000**. Energetic frustration and the nature of the transition state in protein folding. *J. Chem. Phys.* 113: 7663-7671.

Shea J.E., Onuchic J.N., and Brooks III C.L. **2002**. Probing the folding free energy landscape of src-SH3 protein domain. *PNAS* 99(25): 16064-16068.

Sherwood A.E., and Prausnitz, **1964**. Intermolecular Potential Functions and the Second and Third Virial Coefficients. *J. Chem. Phys.* 41:429

Shimada J., and Shakhnovich E.I., **2002**. The ensemble folding kinetics of protein G from an all-atom Monte Carlo simulation. *PNAS* 99: 11175-11180.

Shortle D., Simons K.T., and Baker D., **1998**. Clustering of low-energy conformations near the native structures of small proteins. *PNAS* 95: 11158-11162.

Skonick J., **2005**. Putting the pathway back into protein folding. *PNAS* 102(7): 2265-2266.

Slimmerling C., Strockbine B., and Roitberg A.E., **2002**. All-atom structure prediction and folding simulations of a stable protein. *J. AM. Chem. Soc.* 124:11258-11259.

Smith A.V., and Hall C.K., **2001**. Protein refolding versus aggregation: computer simulations on an intermediate-resolution protein model. *J. Mol. Biol.* 312:187-202.

Snow S.D., Zagrovic B., and Pande V.S., **2002**. The trp Cage: Folding kinetics and Unfolded State Topology via Molecular Dynamics Simulations, *J. Am. Chem. Soc.* 124:14548-14549.

Snow S.D., Sorin E.J., Rhee Y.M., and Pande V.S., **2005**. How well can simulation predict protein folding kinetics and thermodynamics? *Annu. Rev. Biophys. Biomol. Struct.* 34:43-60.

Southhall N.T., Dill K.A, **2002a**. Potential of mean force between two hydrophobic solutes in water. *Biophysical Chemistry* 101-102: 295-307.

Southhall N.T., Dill K.A, and Haymet A.D.J., **2002b**. A view of the hydrophobic effect. *J. Phys. Chem. B* 106: 521-533.

Srinivasan R., and Rose G.D., **1999**. A physical basis for protein secondary structure. *PNAS* 96(25): 14258-14263.

Susanne C.M., Clarie T.F. and Sheena E.R., **2005**. Helix stability and hydrophobicity in the folding mechanism of the bacterial immunity protein Im9, *Protein Engineering, Design & Selection*, 18(1):41-50.

Sutto L., Tiana G., and Broglia R.A., **2006**. Sequence of events in folding mechanism: Beyond the Go model. *Protein Science*. 15:1638-1652.

Takada S., **1999**. Go-ing for the prediction of protein folding mechanism. *PNAS* 96(21): 11698-11700.

Tozzini V., **2005**. Coarse-grained models for proteins. *Current Opinion in Structural Biology* 15:144-150.

Tsai J., Levitt M., and Baker D., **1999**. Hierarchy of structure loss in MD. simulation of src SH3 domain unfolding. *J. Mol. Biol.* 291: 215-225.

Vendruscolo M., Paci E., Dobson C.M., and Karplus M., **2001**. Three key residues form a critical contact network in a protein folding transition state. *Nature* 409: 641-645.

Vendruscolo M., Paci E., **2003**. Protein folding: bring theory and experiment closer together. *Current Opinion in Structural Biology* 13:82-87.

Wagner C., and Kiefhaber T., **1999**. Intermediates can accelerate protein folding. *PNAS* 96: 6716-6721.

Weeks J.D., Chandler D., and Andersen H.C., **1971**. Role of repulsive forces in fluid argon. *J. Chem. Phys.* 56: 5237-5247.

Wolynes P.G., **2004**. Latest folding game results: protein a barely frustrates computationalists. *PANS* 101(18):6837-6838.

Zagrovic B., and Pande V.S., **2004**. How does averaging affect protein structure comparison on ensemble level? *Biophysicsl J.* 87: 2240-2246.

Zhang C., Vasmatzis G., Cornette J.L., and Delisi C., **1997**. Determination of atomic desolvation energies from the structures of crystallized proteins. *JMB* 267: 707-726.

Zarrine-Afsar A., and Davidson A.R., **2004**. The analysis of protein folding kinetic data produced in protein engineering experiments. *Methods* 34:41-50.

Zhou R. **2003**. Trp-cage: Folding free energy landscape in explicit water. *PNAS* 100(23): 13280-13285

Zhou Y., Karplus M., Wichert J.M., and Hall C.K., **1997**. Equilibrium thermodynamics of homopolymers and clusters: Molecular dynamics and Monte Carlo simulations of systems with square-well interactions. J. Chem. Phys. 107(24): 10691-10708.

Zhou Y., and Karplus M., **1999a**. Folding of a model three-helix bundle protein: a thermodynamic and kinetic analysis. JMB 293: 917-951.

Zhou Y., and Karplus M., **1999b**. Interpreting the folding kinetics of helical proteins. *Nature* 401: 400-403.

Zhou Y., and Linhananta A., **2002**. Thermodynamics of an all-atom off-lattice model of the fragment B of staphylococcal protein A: Implication for the origin of the cooperativity of protein folding. *J. Phys. Chem B.* 106: 1481-1485.

Zuo G., Wang J. and Wang W., **2006**. Folding with downhill behavior and low cooperativity of proteins. *Proteins: Structure, Function, and Bioinformatics.*