BOSTON UNIVERSITY
GRADUATE SCHOOL OF ARTS AND SCIENCES

First Draft of Dissertation

PROTEIN FOLDING AND AGGREGATION
– Molecular Dynamics Studies

by

Feng Ding

Molecular Dynamics Studies of Protein Folding and Aggregation

(Order No.)

Feng Ding

Boston University Graduate School of Arts and Sciences, 2004

Major Professor: H. Eugene Stanley, Professor of Physics

ABSTRACT

This thesis applies molecular dynamics simulations and statistical mechanics to study: (i) *protein folding*; and (ii) *protein aggregation*.

Most small proteins fold into their native states via a first-order-like phase transition with a major free energy barrier between the folded and unfolded states. A set of protein conformations corresponding to the free energy barrier, $\Delta G >> k_B T$, are the folding transition state ensemble (TSE). Due to their evasive nature, TSE conformations are hard to capture (probability $\propto \exp(-\Delta G/k_B T)$) and characterize. A coarse-grained discrete molecular dynamics model with realistic steric constraints is constructed to reproduce the experimentally observed two-state folding thermodynamics. A kinetic approach is proposed to identify the folding TSE. A specific set of contacts, common to the TSE conformations, is identified as the folding nuclei which are necessary to be formed in order for the protein to fold. Interestingly, the amino acids at the site of the identified folding nuclei are highly conserved for homologous proteins sharing the same structures. Such conservation suggests that amino acids that are important for folding kinetics are under selective pressure to be preserved during the course of molecular evolution. In addition, studies of the conformations close to the transition states uncover the importance of topology in the construction of order parameter for protein folding transition.

Misfolded proteins often form insoluble aggregates, amyloid fibrils, that deposit in the extracellular space and lead to a type of disease known as amyloidosis. Due to its insoluble and non-crystalline nature, the aggregation structure and, thus the aggregation mechanism, has yet to be uncovered. Discrete molecular dynamics stud-

ies reveal an aggregate structure with the same structural signatures as in experimental observations and show a nucleation aggregation scenario. The simulations also suggest a generic aggregation mechanism that globular proteins under a denaturing environment partially unfold and aggregate by forming stabilizing hydrogen bonds between the backbones of the partial folded substructures. Proteins or peptides rich in α-helices also aggregate into β-rich amyloid fibrils. Upon aggregation, the protein or peptide undergoes a conformational transition from α-helices to β-sheets. The transition of α-helix to β-hairpin (two-stranded β-sheet) is studied in an all-heavy-atom discrete molecular dynamics model of a polyalanine chain. An entropical driving scenario for the α-helix to β-hairpin transition is discovered.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Proteins are among the most important biomolecules. They are the building blocks of life playing remarkable roles: as structural materials and as machines that operate on the molecular level. They are responsible for many functions in cell organization, reproduction, signal transduction, and cell death (apoptosis). Proteins carry out transport and storage in living cells, e.g. Myoglobin & Hemoglobin. Proteins inhibit or catalyze chemical reactions in the form of enzyme. Therefore, proteins are ubiquitous and essential for life! The functional properties of proteins depend upon their three-dimensional structures (see Fig 1.1d,e).

Proteins are linear heteropolymers composed of twenty different amino acids and jointed by peptide bonds (see Fig 1.2). The sequence of amino acids for the proteins is the primary structure of proteins (see Fig 1.1a). Due to the complicated interactions among amino acids and between amino acids and the solvents, the heteropolymer with particular sequence of amino acids folds to generate, from linear chains, compact domains with specific three-dimensional structure (see Fig 1.1b,c). The 3D structures of proteins — tertiary structure — are composed of several segment of regular secondary structures, α-helices and β-sheets (see Fig 1.1d,e), connected by some short unstructured segments—random coil.

To understand the biological function of proteins we would, therefore, like to be able deduce and predict the three-dimensional structure from the amino acids sequence, which is encoded in the gene according to the CENTRAL DOGMA. Given

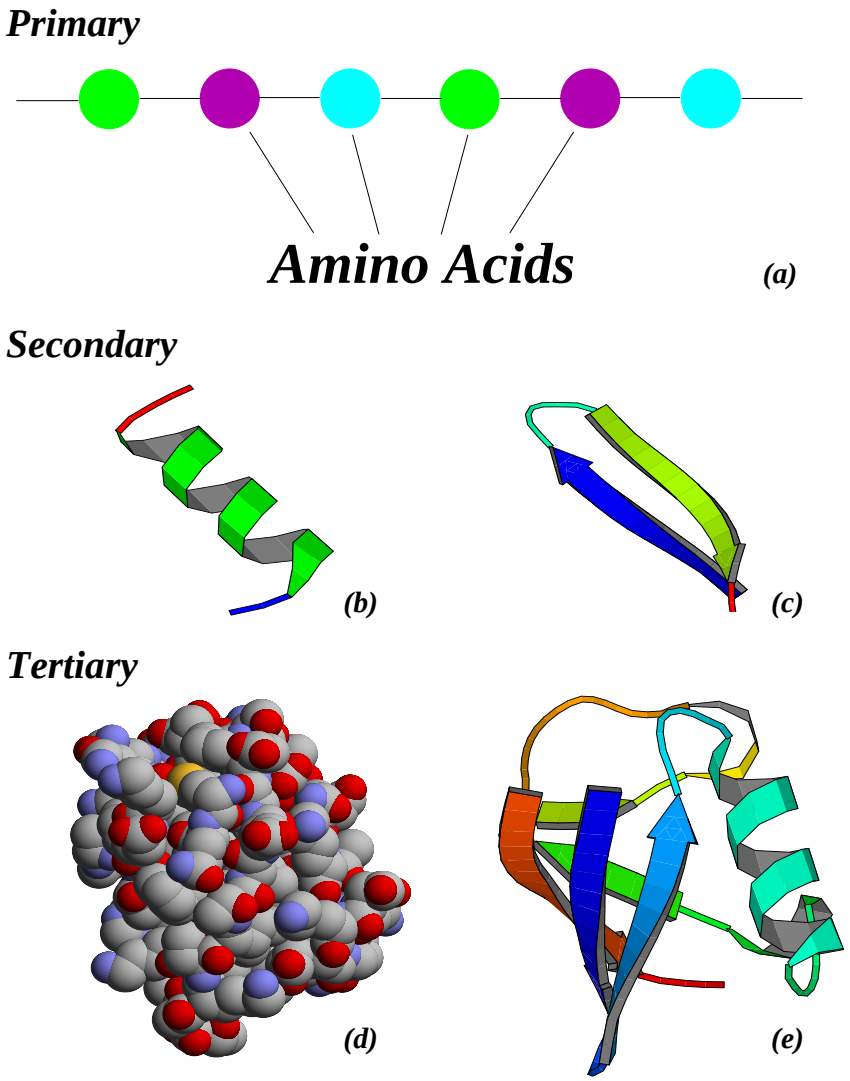


Figure 1.1: The hierarchical composition of proteins.
(a) The amino acid sequence of a protein's polypeptide chain is called its **primary** structure. Different segments of protein form **secondary** structures: (b) α-helix and (c) β-sheet. These secondary structures are stabilized by hydrogen bonds between peptide backbones. The **tertiary** structure is formed by packing several secondary structure elements into the compact globular units. Protein ubiquitin (PDB access code: 1UBQ) are presented in space-filled (d) and cartoon (e) representations for the purpose of illustration.

a sequence of protein with 100 amino acids, and assuming that each residue can adopt two possible conformations, namely α-helix or β-sheet[1], the number of possible three-dimensional conformations of such a protein will be $2^{100} \approx 10^{30}$. The shortest time need for protein to make a conformational change is picoseconds, therefore the folding by random search in the conformation space will take 10^{18} seconds. However, most proteins fold in the order of milliseconds to seconds. This paradox was first described by Levinthal [1]. Therefore, there must be a conformational "information" stored in the primary structure of proteins which drives the protein toward the native state. Ever since Anifsen [2] first shows that protein can fold *in vitro* without any other help such as folders or shapers, the self-assembly property of proteins with a small amount of atoms have been fascinated scientists for almost half an century. The protein folding problem — how does a protein with certain amino acids sequence fold into the specific 3D structure? — have been the subject for extensive theoretical and experimental studies.

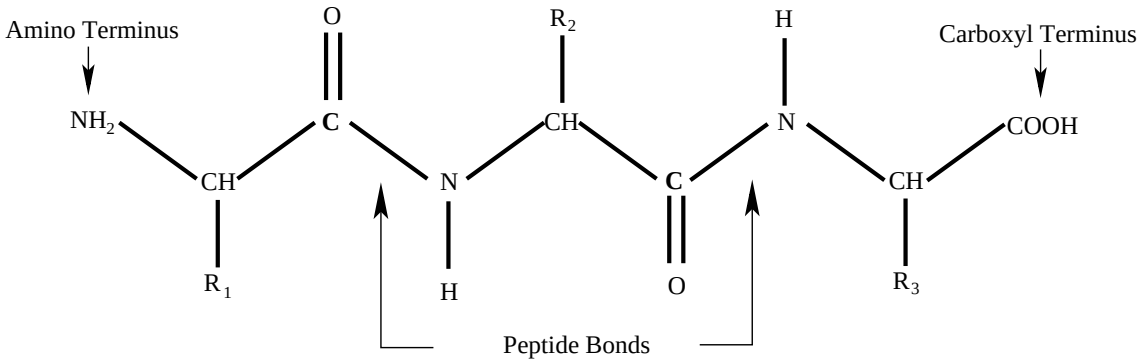


Figure 1.2: Schematic representation of the polypeptide chain.
R_1, R_2, etc. are the side chain groups attached to α-carbons (C_α) of the amino acid.

A natural question upon protein folding problem is "does protein ever fold incorrectly?". The answer is YES. Proteins could fold into some misfolded states usually under conditions other that physiological conditions. When we boil an egg, the proteins in the white unfold. But when the egg cools, the proteins do not return to

[1]This number is obviously a big underestimation

their original shapes. Instead, they form a solid, insoluble (but tasty) mass. This is misfolding. Similarly, biochemists have always complained the tendency of some proteins to form the insoluble deposits in the bottom of their test tubes. We now know that these, too, were proteins folded into the wrong shapes. The misfolded proteins have a strong tendency to form insoluble aggregations. However, it is astonishing how rarely misfolding occurs in the cell. Nature has devised a host of error-correcting mechanisms, most of which are barely understood. However, one type of misfolding and aggregation happens very frequently, causing a process known as amyloidosis forming long-stretched fibrils (see Fig. 1.3), which is observed in a number of diseases, such as Alzheimer's disease, prion disease and Amyotrophic Lateral Sclerosis (Lou Gehrig's disease). Therefore, the studies of protein misfolding and aggregation are of both scientific and medical interests.

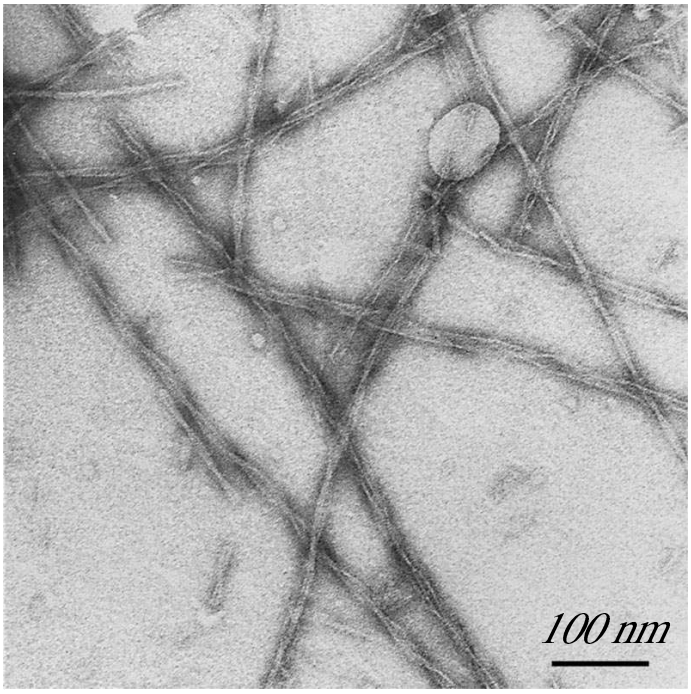


Figure 1.3: The electron micrography of amyloid fibrils formed by SH3 domain.

The problems of protein folding (phase transition) and protein aggregation is of particular interest to physicists. There are a set of theories and models developed in condensed matter physics about phase transition and aggregation in complex

systems, such as the coil-globular transition theory in polymer physics, the random-energy model in describing the disordered systems, the university theory of phase transition near critical point, and various aggregation theories near and/or far way from equilibrium. Many of these phenomena can be understood by focusing on the interplay of only a few important processes, or *driving forces*. The remaining forces only slightly affect the system and may not be critical to understand the phenomena. Therefore, by grasping only the driving force and eliminating the unimportant degree of freedom, the coarse-grained physics models successfully describe the behavior of large scale system. However, while studying the small systems as proteins, many of the assumptions in these physics theories and models do not hold. Lacking the exact knowledge of the *driving force* governing the protein folding and aggregation, it is a great challenge for physicists to develop effective models in describing the behavior of proteins. Therefore, with fast development of computation ability in the past few decades, computer simulations have become increasingly important in studies of proteins. This thesis will focus on the application of discrete molecular dynamics simulations and statistical physics on coarse-grained model proteins to understand the thermodynamics and kinetics of protein folding as well as the mechanism of protein aggregation. The structure of this thesis is as following: the second chapter of this thesis is the discrete molecular dynamics methods. The third chapter presents the studies of the thermodynamics and kinetics of protein folding. In the fourth chapter, the attempt to unveil protein aggregation mechanism is described.

Chapter 2

Discrete molecular dynamics

2.1 Introduction

With dramatic increase of computer power in recent decades, it became possible to study the behavior of relatively large biological molecular systems by computer simulations, such as Monte Carlo (MC) and molecular dynamics (MD) simulations. Monte Carlo simulations on the lattices appear to be very useful to study the theoretical aspects of protein folding [3–5]. The Monte Carlo algorithm is based on a set of rules for the transition from one conformation to another. These transitions are weighted by some transition matrix, which reflects the phenomena under study. The simplicity of the algorithm and a significantly small conformational space of the protein models (due to the lattice constraints) make Monte Carlo on-lattice simulations a powerful tool to study the equilibrium dynamics of the protein models. However, lattice models impose strong constraints on the angles between the covalent bonds, thus, greatly restricting the conformational space of the protein-like model. The additional drawback of this restriction lies in the poor capability of these models to discern the topological properties of the proteins. The time in Monte Carlo algorithms is estimated as the average number of moves (over an ensemble of the folding $\rightleftharpoons$ unfolding transitions) made by a model protein. It was pointed out that Monte Carlo simulations are equivalent to the solution of the master equation for the dynamics and, hence, there is a relation between physical time and computer

time, which is counted as the number of Monte Carlo steps. However, a number of delicate issues, such as the dependence of the dynamics on Monte Carlo move set, remain outstanding and hence an independent test of the dynamics using the MD approach is needed.

The traditional all-atom molecular dynamics with realistic force-field in a physiological solution (which would be ideally used to study protein folding and aggregation) is not computationally accessible with current technology. The complexity and vast dimensionality of the protein conformational space make the folding time too long to be reachable by direct computational approaches. The biological process as allowed by all-atom molecular dynamics, can only be studied on time scales of up to 10^{-7}s using such advanced technologies as world-wide distributed computing [6, 7]. However, the folding and aggregation process happens at least in milliseconds. Therefore, simplified models became popular due to their ability to reach reasonable time scales and to reproduce the basic thermodynamic and kinetic properties of protein folding such as: *(i)* unique native state, i.e. there should exist a single conformation with the lowest potential energy; *(ii)* cooperative folding transition (resembling first order transition); *(iii)* thermodynamic stability of the native state; *(iv)* kinetic accessibility, i.e. the native state should be reachable in a biologically reasonable time.

Recently, a new approach for simulations of model proteins, discrete molecular dynamics [8, 9], has been implemented to study the dynamics of proteins. This approach permits the rapid testing of the folding properties of proteins with reasonable processor time. This MD algorithm has proved to be a powerful tool to study the thermodynamics and kinetics of the folding $\rightleftharpoons$ unfolding transition [8–13] as wall as aggregation of simplified models of proteins [14].

2.2 Discrete Molecular Dynamics

In general, discrete molecular dynamic (DMD) simulations are based on pairwise spherically symmetrical potentials that are discontinuous functions of an interatomic distance r. Each atom has a specific type — A, B, C,... — that determines its inter-

action with other atoms. Each type is characterized by its mass m. The interaction potential between atoms A and B is a step function of their distance r, characterized by distances $0 < r_{min}^{\mathrm{AB}} < r_1^{\mathrm{AB}} \ldots < r_{max}^{\mathrm{AB}}$. If the distance r between two atoms A and B satisfies the inequality $r_i^{\mathrm{AB}} < r < r_{i+1}^{\mathrm{AB}}$, the pair potential has a value of u_i^{AB}. If $r < r_{min}^{\mathrm{AB}}$, $u^{\mathrm{AB}} = \infty$ and r_{min}^{AB} is the hardcore collision distance. If $r > r_{max}^{\mathrm{AB}}$, $u^{\mathrm{AB}} = 0$ and r_{max}^{AB} is the maximal range of interaction. If atoms A and B are linked by a covalent bond, they interact according to a different potential characterized by values $\tilde{r}_i^{\mathrm{AB}}$ and $\tilde{u}_i^{\mathrm{AB}}$. In this case, if $r > \tilde{r}_{max}^{\mathrm{AB}}$, $\tilde{u}^{\mathrm{AB}} = \infty$, which indicates that the bond is permanent and cannot be broken under any conditions.

In DMD all atoms move with a constant velocity unless their distance becomes equal to r_i^{AB}. At this moment of time their velocities change instantaneously. This change satisfies the laws of energy, momentum, and angular momentum conservation. When the kinetic energy of the particles is not sufficient to overcome the potential barrier $\epsilon_i^{\mathrm{AB}} = u_{i-1}^{\mathrm{AB}} - u_i^{\mathrm{AB}}$, the atoms undergo a hardcore reflection with no potential energy change. The main difficulty of this method is the effective sorting and updating of the collision times. However, it is possible to make the speed of the algorithm inversely proportional to $N \ln N$ where N is the total number of atoms [15]. For a sufficiently large number of steps, the method becomes equivalent to a regular MD based on Newtonian dynamics.

2.2.1 Algorithm

In order to effectively simulate the collisions, the system is divided into cells. The dimension of the cell is assigned to be the largest interaction range of all the atom pairs. Thus, all possible interacted atoms of a specific atoms are within the neighbor cells, $n_{cell} = 3^d$ where d is the dimensionality of the simulation system. In addition, traveling of different atom from one cell into another cell has to be included as an additional collision events. In order to determined the soonest collision time t_i for atom i, the calculation only need to between taken over the atoms in the neighbor 3^d cells. Then the smallest t_i will be the soonest collision of the system.

Since each atom moves with constant velocity in between the collisions evolving

1. Initialize the system, construct the table of all possible collisions

2. Determine the soonest collision, between q and p;

3. Taken away from the collision table the outdated collision related to p and q;

4. Update the state of p and q;

5. Recalculated the new possible collisions of p and q and update them into the collision table;

6. If time is smaller than maximum time, repeat step 2,3,4,5,6. Otherwise, quit the program.

Figure 2.1: The schematic diagram of DMD algorithm.

itself, it is the state — the position of previous collision, the velocity as well as the time of previous collision — that has to be kept track of. The DMD simulation maintains a set of all possible collisions, collision table, and determines the soonest collision. Once the soonest collision is determined between q and p after time δt. The states of atoms p and q will be updated accordingly by satisfying the conservation of energy and momentum. The system time is proceeded by δt. Then all the outdated collision events related to p and q will be taken out from the collision table. The new possible collisions of p and q will calculated by taken their neighboring atoms in account. The new calculated collisions will be inserted into the collision table to find the next soonest collision. Therefore, during each collision event, only the evolved atoms pairs need to be updated to keep track of their new state and the rest of the system is not need to update. The schematic chart diagram of the DMD algorithm is presented in Fig. 2.1. To facilitate the searching of soonest collision, a priority tree date structure can be applied [15].

2.2.2 Temperature Control

One of the important issues in molecular dynamics simulations is to control the temperature. In order to simulate the system in the constant temperature, there must be a thermostat to keep the temperature around the target temperature, T_{target}.

The temperature of a system is defined by the kinetic energy of the system,

$$\frac{3}{2}k_BT \equiv \frac{1}{N}\sum_{i=1}^{N}\frac{mv_i^2}{2}\,, \tag{2.1}$$

where N is number of particles in the system. We used two different ways to control the temperature, "ghost"-particle method and Berendsen thermostat [16].

In the "ghost"-particle method, we introduces a large number of non-interacting "ghost"-particles in addition to the proteins under simulation. These particles only experience hard-core interaction among themselves and also with atoms in the proteins under study. Therefore, the additional particles will not contribute to the total potential energy. The fluctuation of potential energy due to the protein system will be evenly distributed into the whole system including the "ghost"-particles. Upon potential energy change δE, the temperature will also change with the amount of $3\delta E/2N$. Once the number of "ghost"-particles are large enough comparing to the number of atoms in the protein system, the system temperature fluctuation is neglectable. Therefore, by setting the initial temperature to T_{target}, the temperature of the system will keep around that value.

The Berendsen thermostat algorithm is proposed by Berendsen et al. [16] to maintain the system temperature for molecular dynamics simulations. By coupling the system with an external bath with a coupling constant α, the algorithm effectively keeps the temperature constant. The algorithm is as the following: at every time step of δt, the velocities of system with temperature T is rescaled such that the new temperature T' becomes,

$$T' = T + (T_{target} - T)(\alpha\delta t)\,, \tag{2.2}$$

with the scaling coefficient $\chi = \sqrt{T'/T} = \sqrt{1-(1-T_{target}/T)\alpha\delta t}$. Assuming that initially the temperature is T_0 and there is not other input into the kinetic energy, then time dependence of the temperature is

$$T(t) = T_{target} + (T_0 - T_{target})\exp(-\alpha t)\,. \tag{2.3}$$

Therefore, any fluctuation of temperature away from the target temperature will decay exponentially.

However, since the discrete molecular dynamics algorithm is event-driven. Once the velocities of the whole system is rescaled, all the calculations about the pairwise collision time needs to be recalculated, which will affect the efficiency of the algorithm. Therefore, instead of rescaling the velocity by coefficient χ, thus the kinetic energy being rescaled by χ^2, we rescale all the pair potential strengths (depth of all potential wells) by $1/\chi^2$. Therefore, the total energy after rescale becomes

$$E_{rescaled} = \frac{E_{potential}}{\chi^2} + E_{kinetic} = \frac{1}{\chi^2}(E_{potential} + E_{kinetic}\chi^2)\,. \tag{2.4}$$

Also, the Boltzmann factor, $\exp((E_{ij}/\chi^2)/k_BT) = \exp(E_{ij}/k_BT')$, keeps the same as in Berendsen's approach. Therefore, the rescaling of potential energy is equivalent to the rescaling of the kinetic energy (Berendsen thermostat) in the dynamics except that the total energy is scaled, which indicates that the time units is also rescaled. In order to calculate the "real" physics variables such as temperature, time and potential energy which are equivalent to those in Berendsen's approach, we need to take into account of the rescaling factor and accumulates the scaling factors in such a way that everything is equivalent to Berendsen's approach

$$T^{mes} = T^{sim}\prod_i \chi_i^2\,, \tag{2.5}$$

$$t^{mes} = \sum_i \frac{\delta t_i^{sim}}{\chi_i}\,, \tag{2.6}$$

$$E_{potential}^{mes} = E_{potential}^{sim}\prod_i \chi_i^2\,, \tag{2.7}$$

where i denotes ith rescaling. Thus, our approach is fully equivalent to Berendsen's thermostat.

2.2.3 Discussion

The speed of the algorithm also decays linearly with the number of steps in the potential, and strongly decays with the density of the system. This method is very effective in simulating proteins (where the density is small and the majority of the interactions can be modeled using either a hardcore or a simple square well) and allows us to observe protein folding transitions and aggregations [9–11, 14].

Discrete molecular dynamics simulation methodology is a step in simplification of molecular modeling with respect to traditional molecular dynamics simulations. The principal drawback of the discrete molecular dynamics simulations is its difficulty to represent forces. Instead, system's dynamics is realized through ballistic collisions between particles. Interactions between particles are modeled by square-well potentials. Despite its simplicity, discrete molecular dynamics has been proved to be a powerful tool not only to study protein folding thermodynamics [8–11] and kinetics [10–12], but to identify the evasive protein transition state ensembles [10] and to witness aggregation of multiple proteins into amyloid fibrils [14]. The latter two goals have yet to be directly approached with traditional molecular dynamics simulations. In addition, the traditional all-atom molecular dynamics simulations are also a simplification of the quantum mechanics simulations, in which quantum interactions are replaced by approximate Newtonian interactions. The latter, in turn, are approximated by a large number of empirical parameters. The advantage of the discrete molecular dynamics simulations versus traditional molecular dynamics simulations is its ability to resolve larger time scales — 10^6 orders of magnitude. The traditional molecular mechanics simulations have similar advantage over quantum mechanics simulations. The traditional molecular dynamics simulations are based on several decades of improving and testing of model force field, while applications of discrete molecular dynamics simulations have been limited until recently to colloids and hard spheres. Despite of this we believe that modifying and improving parameters of discrete molecular dynamics simulations for proteins by testing them on simple systems such as the polyalanine chain studied here will eventually lead to models with quantitative predictive power.

Chapter 3

Protein Folding Problem

3.1 Introduction

One of the intriguing questions in biophysics is how do protein sequences determine their unique three-dimensional structure. This question, known as protein folding problem [2–5, 17–35], is of great importance because understanding protein folding mechanisms is a key to successful manipulation of a protein structure and, consequently, function. The perspective of manipulation of protein's function is, in turn, crucial for effective drug discovery.

Understanding the mechanisms for protein folding is also crucial for deciphering imprints of the evolution on protein sequence and structural spaces. For example, it has been shown [36] that some positions along the sequence in a set of homologous proteins are more conserved in a course of evolution than others. Such conservation can be attributed to evolutionary pressure to preserve amino acids that play crucial role in: *(i)* protein function, *(ii)* stability, and *(iii)* folding kinetics — the ability of proteins to rapidly reach the native state [37]. Interestingly, function is not conserved among non-homologous proteins that share the same fold, so we can assume that the functional pressure to preserve functionally important amino acids is "weaker" than those that are involved in protein stability and folding kinetics. It has been shown [37] that up to 80% of conservation of amino acids in the course of evolution can be explain by pressure to preserve protein stability. Thus, in order to understand

the role of evolutionary pressure to preserve rapid folding kinetics we need to be able to quantify the importance of amino acids for protein folding kinetics.

3.1.1 Nucleation scenario

Two-state proteins are characterized by fast folding and the absence of stable intermediates at physiological temperatures. If we follow the folding process for an ensemble of initially unfolded proteins, both the average potential energy and the entropy of the ensemble decrease smoothly to their native state values. The absence of energetic and topological frustrations defines a "good folder" [38,39]. Various measures have been proposed to determine if a protein sequence qualifies as a two-state folder, either relying on kinetic [40] or thermodynamic [41] properties.

The free energy landscape of the two-state proteins at physiological temperatures is characterized by two deep minima [3, 24, 42–46]. One minimum corresponds to a unique native state with the lowest potential energy and zero entropy, while the second minimum corresponds to a set of unfolded or misfolded conformations with higher values of the potential energy and non-zero entropy (see Fig. 3.1). At the folding transition temperature T_F these minima have equal depth and a protein coexist in two states with equal probability. The two minima are separated by a free energy barrier. A set of conformations that belong to the top of this barrier, having the maximal values of the free energy, are called the *transition state ensemble*.

At equilibrium, the free energy of a conformation, ΔG, translates to the probability of a given conformation to have a state with a given free energy, $p \sim \exp(-\Delta G/k_B T)$, where k_B is the Boltzmann constant and T is the temperature of the system. Since at T_F free energies of native and unfolded/misfolded ensembles are equal, the probability to exist in each of these states is the same. The probability to find a conformation at the top of the free energy barrier is minimal. Therefore, if we consider any protein conformation at the top of the free energy barrier, such conformation most likely unfold or reach its native state with equal probabilities 1/2. So, the transition state ensemble is characterized by probabilities of the conformations to reach the native state equal to 1/2 [10, 11, 30].

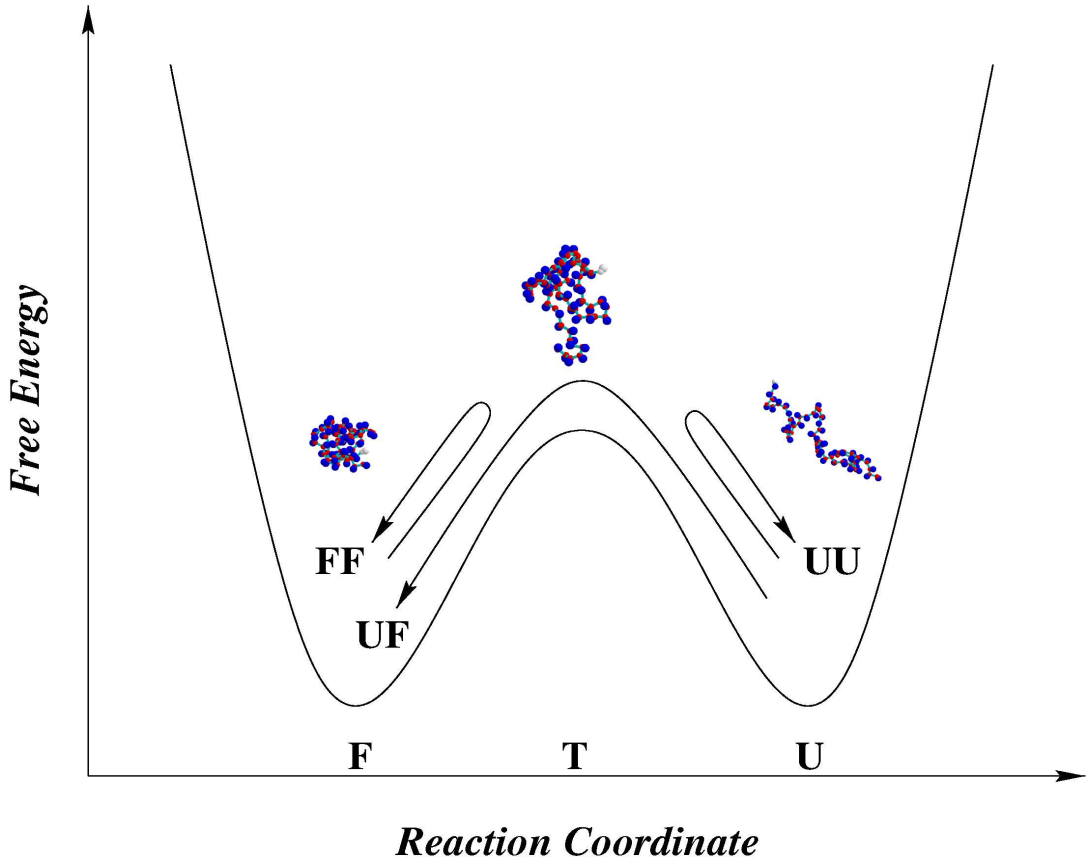


Figure 3.1: Schematic diagram of free energy landscape for the protein at T_f.

The questions then are: "What happens at the top of the free energy barrier?" "Are there any specific mechanisms that are responsible for the rapid folding transition?" There have been proposed numerous folding scenarios to answer these questions [22,23,28,47–52]. The mechanisms that we advocate here is called a *nucleation scenario* [3, 27]. According to nucleation scenario, there is a specific obligatory set of contacts at the transition state ensemble, called a *specific nucleus*, formation of which determines the future of a conformation at the transition state ensemble. If the specific nucleus is formed, a protein rapidly folds to its native conformation. If the specific nucleus is disrupted in the transition state, the protein rapidly unfolds. Thus to verify the nucleation scenario we must determine the nucleus and the transition state ensemble of a protein.

3.1.2 Protein engineering experiments

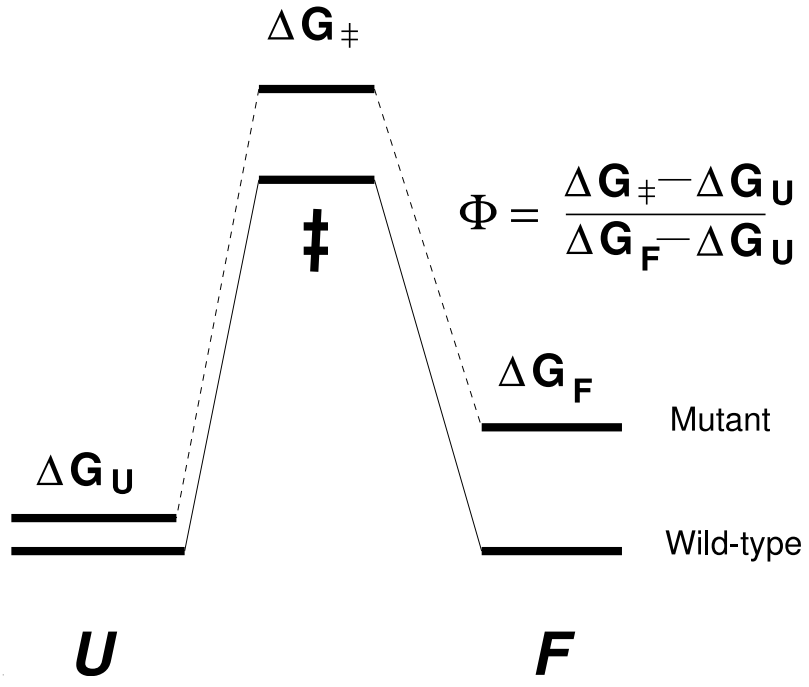


Figure 3.2: Illustration of Φ-value analysis

Schematic diagram about the free energies of different states for both wild type and mutant. U means the unfolded states, ‡ means the transitional states and F means the folded states.

An elegant approach to examine the transition state ensemble in experiments was proposed by Fersht el al. [53, 54]. The method, called *protein engineering* or Φ-value analysis, is based on the engineering a mutant protein with amino acids under consideration replaced by other ones. The site-directed mutations from one amino acid into another one will perturb the contacts formed at this site. The perturbation by the site-directed mutation changes the free energy landscape of wile-type protein (see Fig. 3.2). By measuring the equilibrium rate and the kinetic folding rate for the mutant and for the wild type, the free energy changes in transition state $\Delta G^{\ddagger}$, folded state ΔG^{F} can be determined by assuming that the unfolded state has no stable structure and resembles random coil conformations such that the free energy change for the unfolded states ΔG^{U} (Fig. 3.2) is close to zero [53, 54].

Under the assumption that the mutation does not vary the protein structure as well as the folding pathway, i.e. the change of free energy profile is solely due to the contact changes from mutations in different states, a parameter Φ-values are defined to character the transition state as

$$\Phi = \frac{\Delta G^{\ddagger} - \Delta G^{U}}{\Delta G^{F} - \Delta G^{U}} \,. \tag{3.1}$$

Φ-values are close to 0 if the mutation does not affect the transition state, which indicates that the site does not have substantial structure in the transition state. Φ-values are close to 1 if the substitution affect the transition states to the same extent as the folded states, which indicates that this site have a native-like structure already in the transition state. Thus, these amino acids are most important for the protein folding kinetics.

Due to its simple interpretation, protein engineering has become a popular tools to study protein folding transition state in experiments. However, there are many subtle issues in the interpretation of the experimental data. One example is that the perturbation of single mutation might not be able to disturb the important interaction by the backbone so that the method will not see the importance of these type interactions in the transition state. Another example is that some proteins might form some non-native structures in which case protein engineering does not apply. Moreover, the assumption about the unfolded state is under debate in the field. More and more evidences show that the unfolded state is far from random coil with a certain persistent residual structures. Therefore, precautions need to be taken to interpret the experimental data.

3.1.3 Deriving the folding kinetics from crystal structures

Due to difficulties and cost of actual experimental studies, it is important to develop rapid tools to identify folding kinetics of a given protein from its crystal structure. The ultimate goal is to be able to predict protein folding kinetics of a given protein from its sequence. However, this goal requires the solution of the protein folding problem, i.e. understanding of how a given amino acid sequence folds into native protein structure. Since protein crystal structures provide invaluable information

about amino acid interactions, it is possible to reduce the problem to identifying protein folding kinetics from its structures. Surprisingly, such an approach has already yielded promising and robust results.

Developments in the last decade in protein purification and structure-refining methods [55–58] have led to publication of high resolution proteins' crystal structures. This set of data boosted theoretical studies of protein folding beyond the general heteropolymer models [20, 59–61]. Early studies targeting important amino acids for protein dynamics applied the available crystal structure data in two different approaches: structures were used (i) as reference states (decoys) for theoretical predictions [62–65], and (ii) as a source of dynamical information [66–68]. Studies relied in the developed theoretical framework [69] that explained the folding of relatively small proteins as a chemical reaction between two sets of species – folded and denatured protein states, separated by transition states and by possibly a set of metastable intermediates. Transition states control the rate of the folding reaction, and solving for the portions of the protein that provide structural coherence to these transition states became a major effort in determining the kinetically important amino acids.

Computational power limitation and the inaccuracies in the inter-atomic force-field [70] forced all-atom folding simulations to be performed under extreme conditions favoring denaturation, typically very high temperatures [62–65, 71]. This approach assumes that folding of the protein can be described by running the unfolding simulation backwards in time, and that folding at high temperatures is comparable to folding at room temperatures. These assumptions are questionable, since folding experimental studies are performed under conditions favoring the native state. Furthermore, the low stability of proteins at physiological conditions –only a few kcal/mol [21], indicates that folding of the protein to its native structure is the result of a delicate balance between enthalpic and entropic terms. This balance is distorted at high temperatures, where folding becomes a rare event and the transition state may change drastically [72, 73].

In simulations, Daggett et al. [62, 64, 71] unfolded target proteins starting from their crystal structures, and monitored the time evolution of a parameter represent-

ing the structural integrity of proteins during simulations. Abrupt changes in the parameter pinpointed denaturation of these protein, and analysis of the trajectories revealed disrupted native amino acid interactions. The amino acids involved in these key interactions were identified as kinetically important, and the authors found good correlation to experimental folding results.

The issue of the limited statistical significance of the results [62, 64, 71] due to a small number of unfolding simulations was addressed by Lazaridis et al. [63], who performed a larger series of unfolding simulations starting from conformations slightly different from the initial crystal structure. A wealth of simulations allowed authors to extract the common set of key interactions and identify the important amino acids with higher accuracy. Other attempts to circumvent the poor statistics rested on the discretization of a representative unfolding simulation, followed by long equilibrium simulations of the protein around each of the discretized steps [65]. This method assumes that a protein is at equilibrium at every step in the folding process, but given that at high temperatures folding is a rare event, caution must be taken when interpreting the results.

Recent all-atom simulations were also used to increase the efficiency of protein engineering experiments in a self-consistent experimental+computational approach toward determination of the TSE [74]. This method is most useful for proteins for which only a small fraction of the residues play a key role. Such method may also serve as a refining tool of the protein engineering results.

Protein databases [75] of crystal structures have been widely used as a source of dynamical information with application to folding simulations. In their pioneer study, Wilson et al. [76] computed effective pairwise amino acid contact potentials from the frequencies of spatial proximities between pairs of amino acids obtained in the database of structures. Authors used these potentials to reproduce with modest success the folding process of a one-atom crambin [77] model on the square lattice. Skolnick and Kolinski [66, 67] developed a statistical potential using two-atom representation of apoplastocyanin [78]. Folding simulations on a finer lattice than that used in previous studies allowed authors to fold a model protein with a root mean square deviation (rmsd) of 6Å with respect to the crystal structure. However, the

propensity of the amino acids to adopt a specific crystal structure prevented authors from generalizing the applicability of the model to more than one protein at a time.

Since all information necessary to fold a particular protein is precisely encoded in the protein structure, the crystal structure can be used as the sole source of information, with no regard to the protein database. This approach was taken by Dill et al. [68] in their study of the folding mechanisms of crambin and chymotripsin inhibitor. Dill et al. assigned attractive interactions between all pairs of hydrophobic amino acids that were in geometrical proximity from each other in the crystal structure, neglecting other amino acid interactions. The folding dynamics was implemented through a sequence of folding events in Monte Carlo search. Authors found that one every 4000 simulations ended in the crystal structure and proposed a folding pathway for the two proteins. This technique, although able to find a folding event, cannot reproduce a statistically significant ensemble, since the sequence of folding events is forced in the simulation. Thus, only when the proposed sequence of events coincides with the most probable ones, can the results be representative of the folding of the protein.

We combine effective dynamic algorithm as well as coarse-grained protein models with crystal structure based interaction potential to study the protein folding kinetics. Despite the simplicity, the combinatorial approach gives encouraging results [10, 14, 79].

3.1.4 Why is it difficult to determine the folding kinetics?

Some other theoretical approaches [45, 52, 80, 81] have been proposed to predict the transition states in protein folding and obtained significant correlations with experimental ϕ values for several proteins. However, each of these models involves drastic assumptions. For example, each amino acids can only adopt two states—native or denatured, and the ability to be in the native state was considered to be independent of other residues. Such an assumption is normal for one-dimensional systems, but may be inappropriate for three-dimensional proteins, because the native state of a residue depends on its contacts with its neighbors. Moreover, the dynamics is only

derived from thermodynamics in these works.

The principal difficulty to select TSE conformations is the identification of the reaction coordinate for protein folding. The fraction of native contacts Q [82,83] has been proposed as the reaction coordinate to study the TSE. However, the reaction coordinate for folding is not well defined [24, 30, 84], so in principle it is difficult to determine the folding TSE from *equilibrium* sampling. The probability for the protein conformation to fold into the native states p_{fold} [30], is proposed as the robust criteria of TSE. Thus, the TSE can be determined from the *kinetic* simulations as the set of conformations representing the kinetic separatrix between native and unfolded basins of attraction [30, 84].

Here we propose an approach to identify TSE from molecular dynamics simulations. Our approach unifies a number of concepts that has been developed in the protein folding community [5, 85–88]. We test this approach on the folding kinetics of the C-Src SH3 domain (PDB access code: 1NLO), within the Gō model approximation for the amino acid interactions [5, 86]. We introduce a coarse-grained representation of C-Src SH3 domain which includes the C_α and C_β atoms, and a set of additional specific constraints that allow us to mimic protein flexibility. Next, we will describe our model and methods.

3.2 Protein models

3.2.1 Model Geometry

We model the protein by beads representing C_α and C_β (Fig. 3.3). There are four types of bonds: *(i)* covalent bonds between $C_{\alpha i}$ and $C_{\beta i}$, *(ii)* peptide bonds between $C_{\alpha i}$ and $C_{\alpha(i\pm1)}$, *(iii)* effective bonds between $C_{\beta i}$ and $C_{\alpha(i\pm1)}$, *(iv)* effective bonds between $C_{\alpha i}$ and $C_{\alpha(i\pm2)}$. In order to determine the effective bond length, we calculate the average and the standard deviation of distances between carbon pairs of types *(iii)* and *(iv)* for 10^3 representative globular proteins obtained from the PDB. We find that the average distances are 4.7Å and 6.2Å for type *(iii)* and type *(iv)* bonds respectively. The ratio σ of the standard deviation over the average for bond types

(iii) and *(iv)* are, respectively, 0.036 and 0.101. The standard deviation of bond type *(iv)* is larger than that of bond type *(iii)* because it relates to the angle of two consecutive peptide bonds. Thus, the bond lengths of type *(iv)* fluctuate less than the that of type *(iii)*. The effective bonds impose additional constraints on the protein backbone, so our model mimics closely the stiffness of the protein backbone, and can give rise to cooperative folding thermodynamics.

In our simulation, the four types of bonds are realized by assigning an infinitely high potential well barriers [89]:

$$V_{ij}^{\text{bond}} = \begin{cases} 0, & D_{ij}(1-\sigma) < |r_i - r_j| < D_{ij}(1+\sigma) \\ +\infty, & \text{otherwise} \end{cases} , \tag{3.2}$$

where D_{ij} is the distance between atoms i and j in the native state, $\sigma = 0.0075$ for a bond of type *(i)*, $\sigma = 0.02$ for a bond of type *(ii)*, $\sigma = 0.036$ for a bond of type *(iii)* and $\sigma = 0.101$ for a bond of type *(iv)*. The covalent and peptide bonds are given a smaller width and the effective bonds are given a wider width to mimic the protein flexibility.

3.2.2 Non-bonded interaction potential: Gō model

To model interaction between non-bonded atoms, we use a modified Gō model similar to one described in [89], in which interactions are determined by the native structure of proteins. In our model, only C_β atoms that are not next to each other along the chain interact with each other. The cutoff distance between C_β atoms is chosen to be 7.5Å. Thus, the total potential energy

$$E = \frac{1}{2} \sum_{i,j=1}^{N} U_{i,j} \tag{3.3}$$

where i and j denote residue i and j. $U_{i,j}$ is the matrix of pair interactions

$$U_{i,j} = \begin{cases} +\infty, & |r_i - r_j| \le a_0 \\ (1 - 2\Delta_{i,j})\epsilon, & a_0 < |r_i - r_j| \le a_1 \\ 0, & |r_i - r_j| > a_1 \end{cases} , \tag{3.4}$$

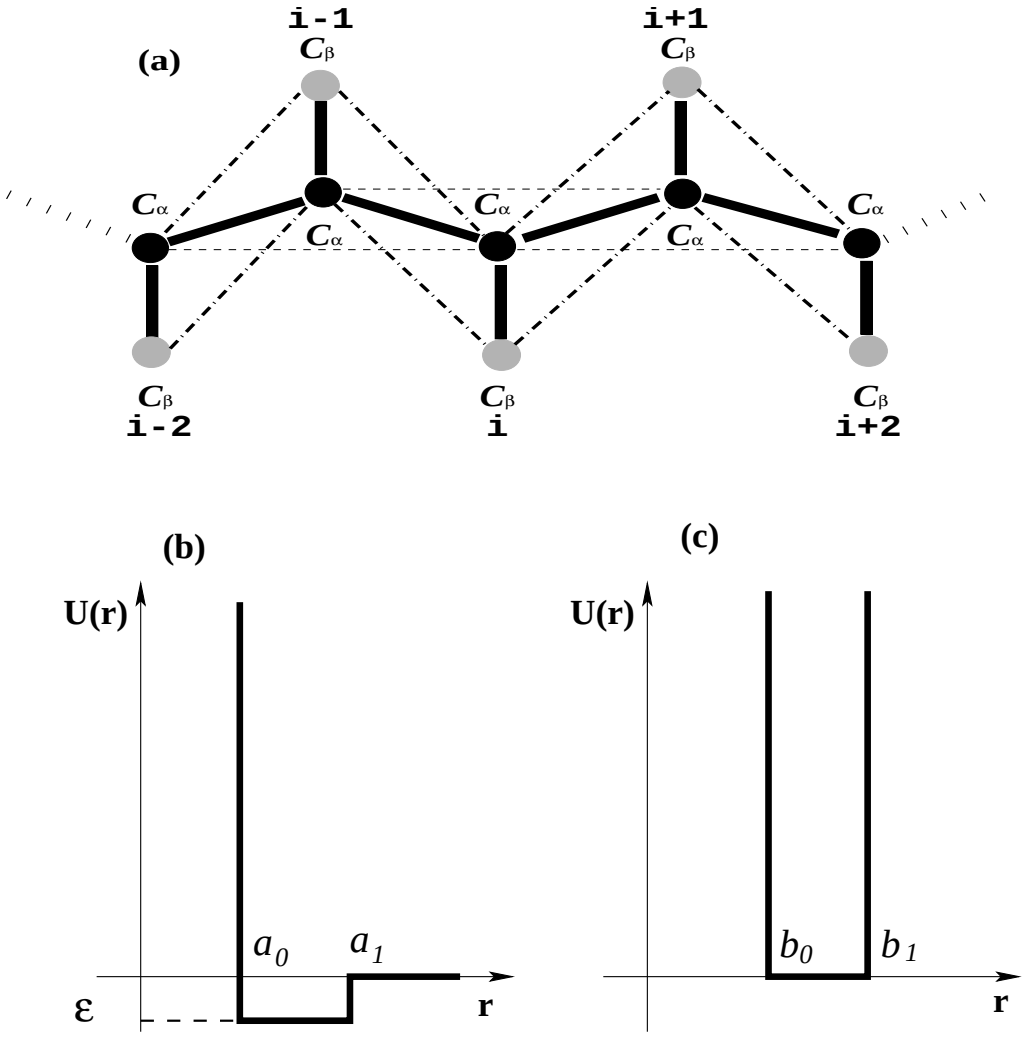


Figure 3.3: The coarse-grained protein model

(a) Schematic diagram of the protein model. Grey spheres represent alpha carbons, black ones represent beta carbons (for Gly alpha and beta carbons are the same). In the present model only the interaction between side chains are counted, so that the interaction only exists between β carbons, and the α carbon only plays the role of the backbone. (b,c) The potential of interaction between (b) specific residues; (c) constrained residues. a_0 is the diameter of the hard sphere and a_1 is the diameter of the attractive sphere. $[b_0, b_1]$ is the interval where residues that are neighbors on the chain can move freely. ϵ is negative for native contacts and positive for non-native ones.

where, a_0 is hard-core distance between the beads and a_1 is the interaction distance used to define a contact. $||\Delta||$ is a matrix of native contacts with values of either 1 forming contact at native state, or 0 without contact at native state. In this model, only C_β atoms that are not next to each other along the chain interact with each other. The matrix $||\Delta||$ is usually termed as contact map. The contact map of C-Src SH3 domain protein is repent in Fig. 3.4.

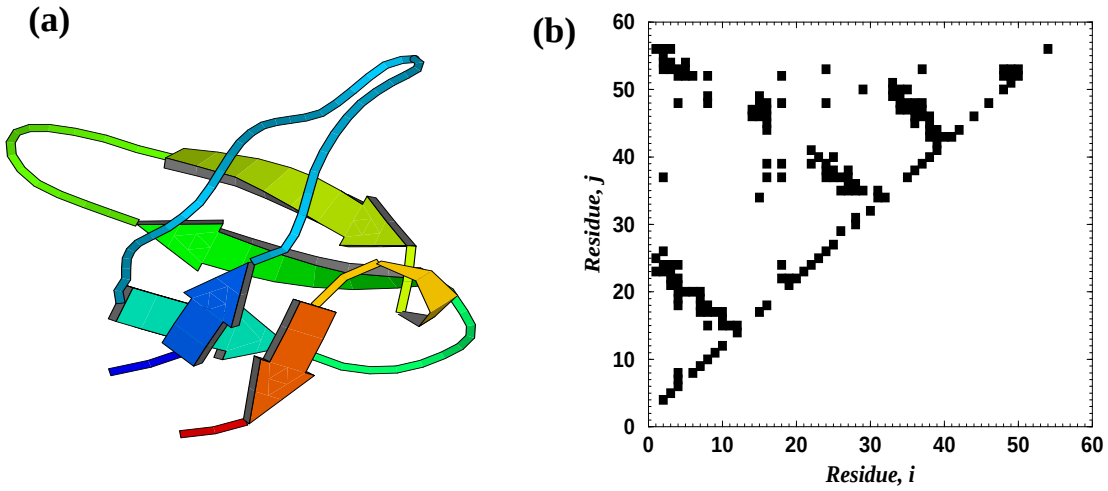


Figure 3.4: The native state of C-Src SH3 domain.
(a) The cartoon representation of C-Src SH3, which is 56 amino acids long protein. It contains mainly β-sheets. (b) The native contact map of SH3 domain.

Despite the drawback of the Gō model, associated with the prerequisite knowledge of the native structure, it has important advantages. It is the simplest model that satisfies the principal thermodynamic requirements for a protein-like model: *(i)* the unique and stable native state, *(ii)* a cooperative folding transition resembling a first-order phase transition. Further, it has been widely applied in the past to study various aspects of protein folding thermodynamics and kinetics [52, 86, 87]. In addition, experimental works [31, 32, 90] show that transition state ensemble of many two-state fast folding proteins is primarily determined by native states topologies.

3.3 Folding Thermodynamics

A successful protein model should reproduce the basic thermodynamic properties [3] such as: *(i)* unique native state, i.e. there should exist a single conformation with the lowest potential energy; *(ii)* cooperative folding transition (resembling first order transition); *(iii)* thermodynamic stability of the native state; *(iv)* kinetic accessibility, i.e. the native state should be reachable in a biologically reasonable time. To test whether the model faithfully reproduces the experimentally observed [31, 91] thermodynamic properties of C-Src SH3 domain, we first perform the discrete molecular dynamics simulations of the model C-Src SH3 domain at various temperatures. At each temperature we calculate the potential energy E, the radius of gyration R_g, the rms deviation from the native state $RMSD$ [92], and the specific heat $C_v(T) \equiv < (\delta E)^2 > /T^2$.

Definition 3.1 *The radius of gyration R_g of a polymer with N monomers is defined as:*

$$R_g^2 = \frac{1}{2N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} (\vec{r}_i - \vec{r}_j)^2 \,, \tag{3.5}$$

where $\vec{r}_i$ and $\vec{r}_j$ are position of the ith and jth monomers. The value R_g measures the overall size of the polymer.

Definition 3.2 *The $RMSD$ of two structures, $\Gamma = \{\vec{r}_i\}$ and $\Gamma' = \{\vec{r}'_i\}$, measures the difference between these two structures. The definition is as following:*

$$RMSD^2 = \min_A \frac{\sum_{i=1}^{N} (\vec{r}_i - A \cdot \vec{r}'_i)^2}{N} \,, \tag{3.6}$$

where A is an arbitrary rotation matrix, and N is total number of atoms in both of the structures.

At low temperatures, the average potential energy $\langle E \rangle$ increases slowly with temperature, and the $RMSD$ remains below 3Å, indicating that the protein is within the basin of native state. Near the transition temperature $T_f = 0.91$, the quantities E, R_g, and $RMSD$ fluctuate between values characterizing two states, folded and unfolded, yielding bimodal distribution of potential energy (Fig. 3.5c). Potential

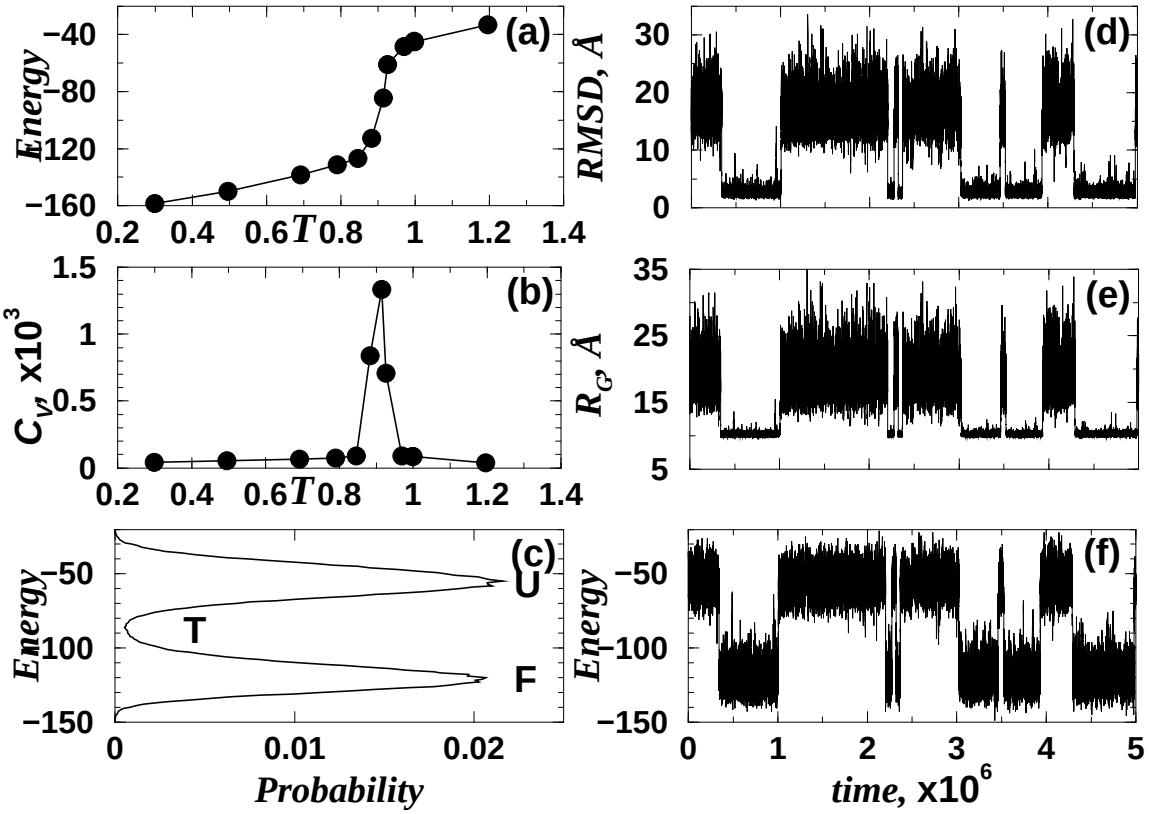


Figure 3.5: Thermodynamics of C-Src SH3 domain.
(a) The average potential energy and (b) the specific heat dependence on temperature. There is a sharp transition at the folding temperature T_f=0.91. (c) The probability distribution of the potential energy at T_f. It is bimodal, with a low probability between the peaks corresponding to folded (F) and unfolded (U) states, which corresponds to the putative TSE (T). (d) The radius of gyration, (e) $RMSD$, and (f) potential energy of the protein at folding temperature T_f respectively. A typical run is shown. In folded states, the $RMSD$ is around 2Å. The energy difference between the folded state and unfolded state is about 70 energy units.

energy fluctuations at T_f give rise to a sharp peak in $C_v(T)$ (see Fig. 3.5b), which is characteristic of a first order phase transition for a finite system. Our findings are consistent with experimental observations for C-Src SH3 domain [31, 91].

The trajectory at T_f (see Fig. 3.5c,d,e,f) clearly behaves a two-state folding ther-

modynamics. At this temperature, the protein has equal probability to be seen in the folded and unfolded states (Fig. 3.5c). In the histogram of potential energy at T_f, there is a region with low probability between the folded and unfolded state. The protein conformations in this region are unstable. Then, the question is "whether this region are the true transition state"? Next, we examine this question.

3.4 Folding Kinetics

3.4.1 Construction of putative TSE

Next we determine for the C-Src SH3 domain the folding TSE, a set of conformations with p_{fold} equal to 1/2. It is computationally impossible to find p_{fold} for every single conformation of a protein. Thus, following Ref. [87], we limit the search for TSE conformations to the energy range $\{E_{TS}\}$, defined to be $-91 < E < -80$, corresponding to the unstable region with the lowest probability in the potential energy histogram at T_f (Fig. 3.5c). Not all conformations from $\{E_{TS}\}$ belong to the TSE, so we partition these conformations into four kinds of fluctuations that bring the protein to the unstable state within the range of $\{E_{TS}\}$ (see Fig. 3.6): (i) FF, when the folded protein unfolds to $\{E_{TS}\}$ and then rapidly refolds to its native state, (ii) UU, when the unfolded protein partly folds into $\{E_{TS}\}$ and then rapidly unfolds, (iii) FU, when the folded protein unfolds to $\{E_{TS}\}$, and then proceeds unfolding further, and (iv) UF, when the unfolded protein traverses the energy range $\{E_{TS}\}$ on its way to folded conformations.

Remark 3.1 *As definition, the value p_{fold} measures the probability of a certain conformation to fold into the native state. The TSE conformations corresponds to the free energy barrier separating the folded and unfolded states so that has %50 probability to convert into native state. However, (i) the free energy landscape are very sensitive to temperature, for example slightly deviate T_f the free energy landscape will have only one basin either folded or unfolded. Therefore, the p_{fold} analysis should be performed at temperature T_f. (ii) At T_f, each unfolded conformation could have high probability fold into folded basin by thermo-fluctuations if the observation time is*

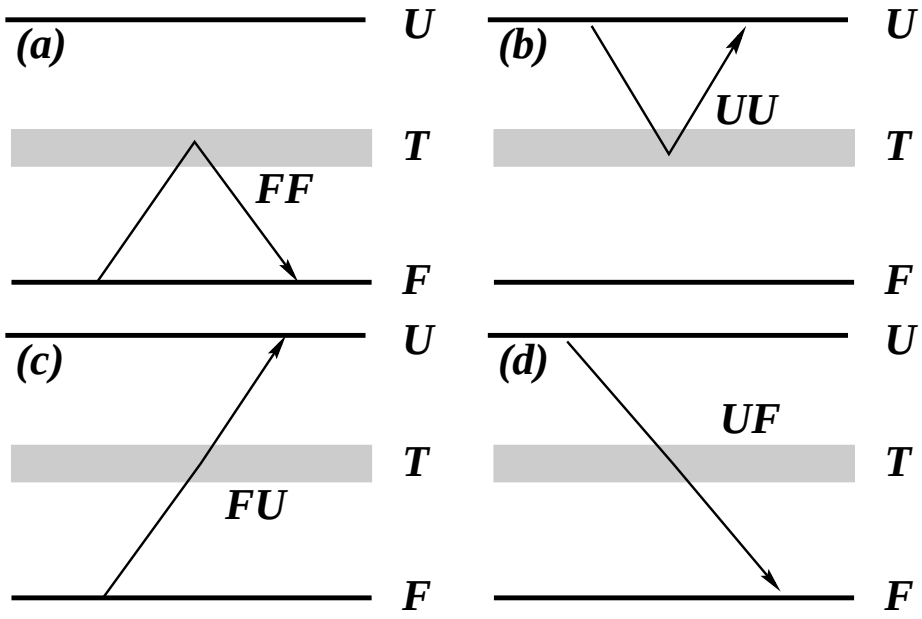


Figure 3.6: Schematic diagram of four types of fluctuations to the putative transition state region
FF (a), UU (b), FU(c) and UF(d). The upper line corresponds to the average energy of unfold states (U), and the lower line corresponds to the average energy of folded states (F). The cyan shaded region indicates the putative transition state (T) energy range $\{E_{TS}\}$, $-91 < E < -80$. All the fluctuations are selected along the trajectory and are partitioned according their history and their future.

comparable of typical folding time. Therefore, the observation time should be much less than the typical folding time.

We determine p_{fold} from 100 simulation runs for conformations out of these four UU, FF, FU and UF ensembles. For each ensemble, we randomly select 10 conformations to calculate the corresponding p_{fold} values. In each run, we reassign the initial velocities of each residue keeping the temperature unchanged at T_f. Because the initial state is unstable, it rapidly evolves to a stable folded or unfolded state. Indeed, p_{fold} varies greatly between starting conformations, despite the fact that their energies are similar: FF (Fig. 3.7b) conformations have $p_{\text{fold}} \approx 1$, while UU (Fig. 3.7c) conformations have low p_{fold}, UF (Fig. 3.7d) and FU (data not shown) conformations exhibit $p_{\text{fold}} \approx 1/2$, and thus belong to the TSE. UU and FF conformations represent

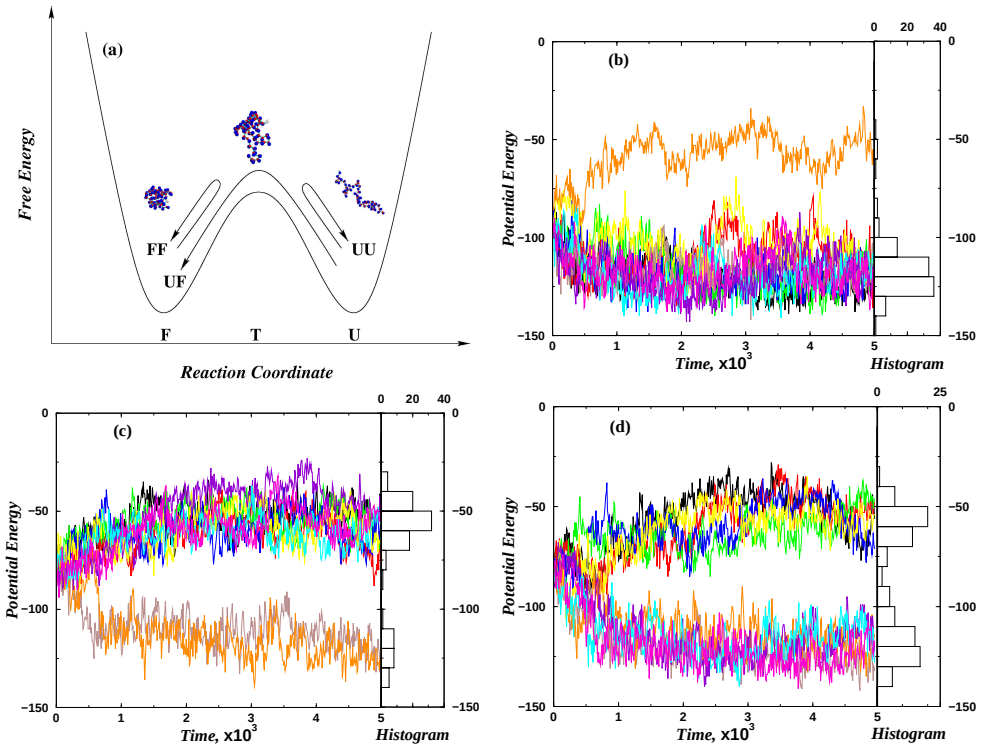


Figure 3.7: Relaxation trajectory for different fluctuations extended to putative transition region.

(a) A schematic representation of TSE conformations. TSE conformations belong to the top of the free energy barrier between folded and unfolded states, and have 50% probability to descend to the folded state and 50% probability to descend to unfolded states. (b) The evolution of potential energy for simulations starting from a conformation from the native state basin of attraction (FF conformations). Most simulations fold (see histogram). (c) The evolution of potential energy of the protein for simulations starting from a conformation that belongs to an unfolded basin of attraction (UU conformations). Most simulations unfold (see histogram). (d) The fluctuations of potential energy starting at time zero from a conformation belonging to the TSE: there is $\approx 50\%$ probability to fold, and $\approx 50\%$ to unfold. All three classes of fluctuations shown in (b)–(d) start from conformations of the same potential energy, and only 10 out of the 100 energy trajectories are shown.

basins of attraction of unfolded and native states respectively, so the energy and also the fraction of native contacts Q, which is related to energy in the Gō models, are *not* appropriate reaction coordinates for folding. For simplicity, we construct our TSE only of UF conformations from the energy window $\{E_{TS}\}$, i.e., conformations that are collected only along trajectories that traverse this energy range on the way from the unfolded state to the folded state. We analyze 200 independent folding transitions to create the TSE. For a control, we randomly select 10 of them, and for each we determine p_{fold}. For all 10 conformations, p_{fold} is approximately 1/2, verifying our selection of conformations representing the TSE.

Remark 3.2 *It is important, that even though we perform thermodynamic simulations, we study the protein folding kinetics because we select UU, FF and UF conformations based on their past and future states. It is due to kinetic selection of the UU, FF, and UF conformations we observe difference in p_{fold} values, even though their energetic (potential energy) and structural (RMSD, R_g) characteristics are close to each other.*

3.4.2 Characterization of TSE

"Virtual screening" method

We use a technique similar to experimental ϕ-value analysis to predict the TSE via computer simulations. We assume that the mutation does not give rise to significant variation of the three-dimensional structures of folded state and transition state ensembles, the same assumption that is made in protein engineering experiments. In our simulations, the free energy shifts due to mutation can be computed separately in the unfolded, transition, and folded state ensembles:

$$\Delta G_x = -kT \ln\langle \exp(-\Delta E/kT)\rangle_x \,. \tag{3.7}$$

Here x denotes a state ensemble: folded, F, unfolded, U, and transition, T, ΔE is the change of potential energy due to the mutation, and the average $\langle \ldots \rangle_x$ is taken over

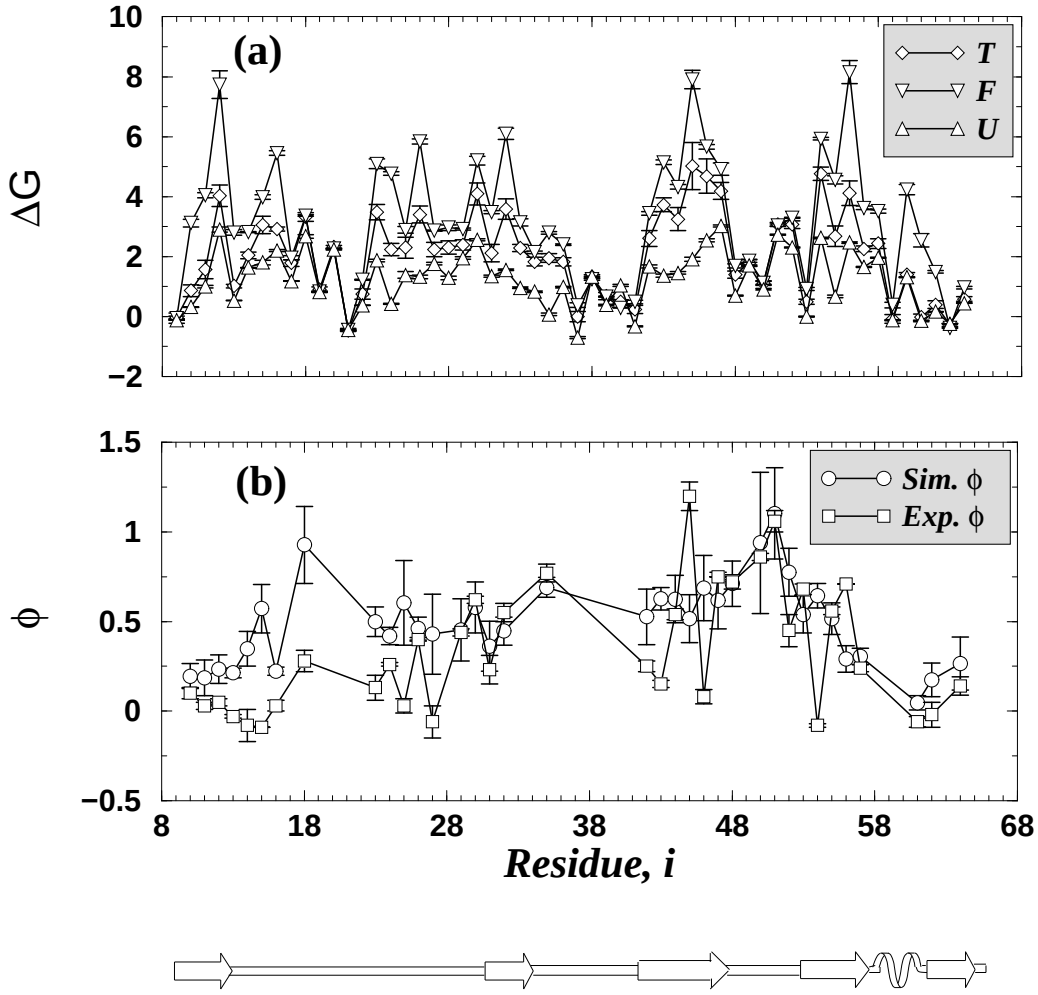


Figure 3.8: Φ-values for C-Src SH3 domain.

(a) The values of ΔG for folded (F), transitional (T), and unfolded (U) conformations determined in simulations at T_c. (b) ϕ-values determined from simulations by the virtual screening method (•) and by experiment (□) [31,90]. Only residues for which experimental ϕ-values are known are shown. The statistical errors of ΔG values are estimated as the standard deviations. The errors of ϕ-values are derived from that of ΔG by the error propagation. Below the x axis, the linear structure of C-Src SH3 domain is shown. The arrows denote the β strands, the spaces between the arrows are RT loop, n-src loop and distal β hairpin, respectively, and the short spiral denotes the 3_{10} helix.

all conformations of unfolded, transition, and folded state ensembles. We compute

$$\begin{aligned}\phi &\equiv \frac{\Delta G_T - \Delta G_U}{\Delta G_F - \Delta G_U} \\ &= \frac{\ln\langle\exp(-\Delta E/kT)\rangle_T - \ln\langle\exp(-\Delta E/kT)\rangle_U}{\ln\langle\exp(-\Delta E/kT)\rangle_F - \ln\langle\exp(-\Delta E/kT)\rangle_U}\,. \end{aligned} \tag{3.8}$$

The same equation has been applied to calculate ϕ values in Ref. [83]. Interestingly, if one adopts a simplified definition of ϕ-value used in recent work [93] as proportional to the number of contacts a residue makes in the TSE, the correlation coefficient between theoretical and experimental ϕ-values is reduced to 0.27. An approximation to the ϕ-value, the *difference* between the average number of contacts residues form in the TSE and in unfolded states, $\phi \approx (\langle N_i\rangle_T - \langle N_i\rangle_U)/(\langle N_i\rangle_F - \langle N_i\rangle_U)$, provides a better correlation coefficient between predicted and experimentally observed ϕ-values (0.48) than does the approximation of Ref. [93]. The reason why a thermodynamic definition of the ϕ-value yields better agreement with experiments can be inferred from the ΔG plot (Fig. 3.8a), which shows that $\Delta G_F - \Delta G_U$ for most of the amino acids is *not* negligible. Indeed, there are several amino acids that make persistent short-range contacts in the unfolded states.

Next, we determine the ϕ-values for each residue using the "virtual screening" method. The correlation coefficient between experimental [31, 90] and simulated ϕ-values is 0.58 (Fig. 3.8b). In Fig. 3.8b, there are regions that our determined ϕ-values apparently mismatch the experimental ones, such as residues 10-20, residues around 24, residues 43-46 and residue 54. One of the main reasons is that the mutations can not probe all the surrounding interactions, especially the backbone interactions, as what we do by "virtual screening" method. For example, residues 10-20 belong to the N-terminal strand of RT-loop, which is mostly stabilized by backbone interactions and persistent [94] in the partially unfolded states, and thus the mutations in this region produce low ϕ values. For residues 43-46, we predict all intermediated ϕ-values around 0.6 so that the corresponding β strand adopts the native-like structure in TSE, which is consistent with that fact that residue A45 having the highest experimental ϕ-value. The reason why we can not capture the fluctuations of experimental values is because our simplified Gō model does not consider the specific nature of different amino acids. In addition, the mutation on the same site to dif-

ferent amino acids may yield different ϕ-values while the "virtual screening" method does not consider the specificity of mutations. For residues L24 and G54, which we found to be crucial for the folding kinetics, will be discussed later in this paper. For comparison, we have also calculated the ϕ-values by using potential energy as the reaction coordinate, i.e., selecting all the UU, FF, FU and UF conformations as transition states. The correlation coefficient between thus calculated and experimental ϕ-values reduces to 0.49 (data not shown). Our results allow us to directly evaluate the relative importance of various interactions in the TSE — an insight difficult to obtain solely from experiments, which report on the structure of TSE only implicitly, via ϕ-values, the interpretation of which is too complex in some cases [95, 96].

By comparing the number of contacts N_C that an amino acid makes in the TSE with that number in the unfolded state (Fig. 3.9a), we select amino acids that are most important for the formation of the TSE by setting the cutoff as 2 (Fig. 3.9b): A12, N23, L24, F26, L32, V35, W43, A45, H47, G54, Y55 and I56. These amino acids have high calculated ϕ-values except A12. The low calculated as well as experimentally derived ϕ-value for A12 indicates that it still need to form many more contacts to be native like by noticing that it forms 13 contacts in the native states (Fig. 3.10a). In general, the majority of the residues from that list have also high experimental ϕ-values; remarkably, residue A45 which has the highest number of contacts in the TSE with respect to the unfolded states has the highest experimental ϕ-value — 1.2. Notable exceptions are N23, L24, W43 and G54, which have ϕ-values that are either small or negative as in the case of G54. For residue G54, mutation destabilizes the protein while accelerating folding, strongly suggesting that it indeed participates in the TSE [95, 97].

3.4.3 Folding Nuclei

A method to identify protein folding nucleus from equilibrium trajectories was proposed in Ref. [89]. The idea is to study ensembles of conformations that have specific history and the future. For example, conformations that originate in the unfolded state, reach a putative transition state, and later unfold, must differ from the confor-

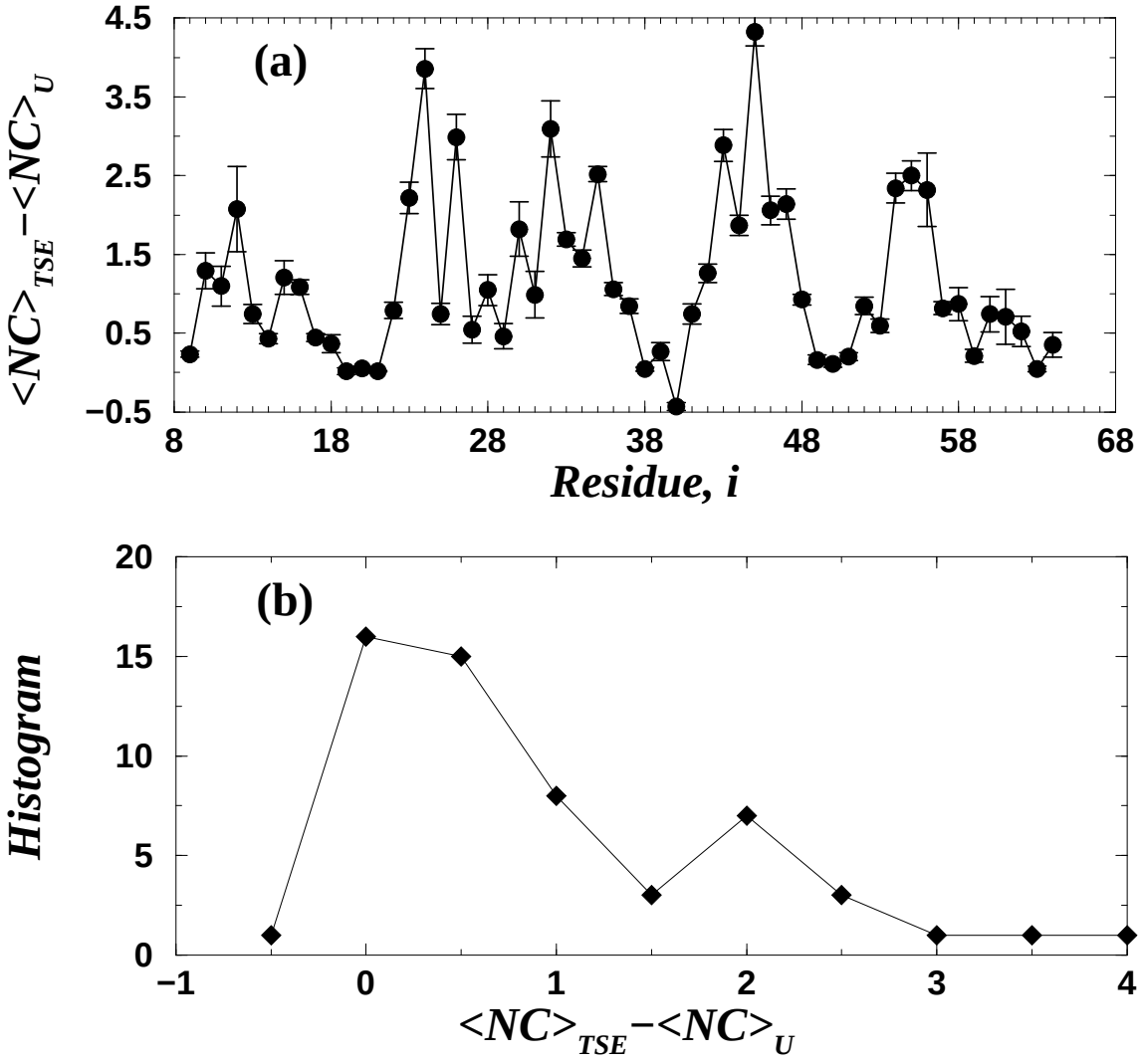


Figure 3.9: Formation of contacts in TSE.

(a) The number of extra contacts that each residue forms in the TSE, compared to the unfolded ensemble. (b) The histogram of the extra contact numbers for each amino acids. There is a second peak at contact number 2. We set the cutoff as 2 for the selection of amino acids that contribute most to the folding TSE. (c) Structure of the native state of C-Src SH3 domain. The color code (from red to white) represents the relative contribution of individual amino acids to the TSE. The brighter colors of red represent the kinetically most important structures.

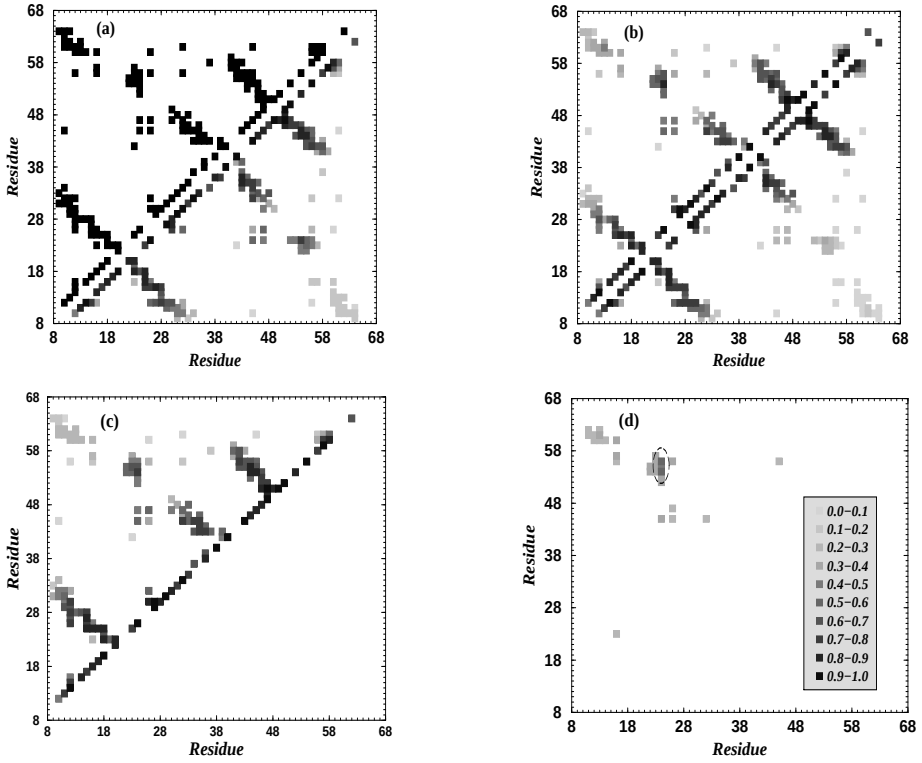


Figure 3.10: Frequency maps of different protein ensemble

(a) Above the diagonal is the contact map of the native C-Src SH3 conformation, while below the diagonal is the map of frequencies of contacts between residues obtained from the averaging over 200 conformations of TSE. (b) Above the diagonal is the map of frequencies of contacts between residues obtained from the averaging over 200 conformations of FF conformations, while below the diagonal is the map of frequencies of contacts between residues obtained from the averaging over 200 conformations of UU conformations. (c) The contact map of putative TSE is calculated by using fraction of native contacts Q to select the TSE conformations [82, 83]. (d) The difference of the frequency maps for FF and UU conformations shows that the key contacts distinguishing FF and UU basins are between L24 and G54-I56 (dashed ellipse). A long range contact L24-G54 occurs with high probability in all conformations that belong to the basin of attraction of the native state. The gray scale represents the frequency scale.

mations that originate in the folded state, reach putative transition region and later fold. Both sets of conformations, which we denote UU and FF correspondingly, are characterized by the same potential energy and similar overall structural characteristics. Nevertheless, there is a crucial kinetic difference between them. According to nucleation scenario, UU conformations lack the folding nucleus. It is not created at the transition state, which leads to the protein unfolding. FF conformation have the nucleus intact at the transition state, so that the protein does not unfold. Thus, in order to determine the nucleus, we propose to compare the average frequencies of contacts between amino acids in UU and FF ensembles of conformations. Amino acid contacts that have the largest frequency difference form the folding nucleus.

In Fig. 3.10 we present the contact map of the native state and maps of frequencies of the relative participation of contacts in TSE, FF and UU conformations and the difference of the frequency maps for FF and UU conformations. For comparison we also show the contact map of the putative TSE (Fig. 3.10c) derived from using the fraction of native contacts Q as the method to select TSE conformations [82, 83]. We find that the three-strand beta-sheet (residues 28–56) forms first — it is already present in the majority of UU conformations (Fig. 3.10b). This is not surprising since the three-strand beta-sheet is just a combination of distal hairpin (residues 44–56) and the n-src loop [90] (Fig. 3.9c). It is a substructure of relatively short range contacts which form fast in accord with general observations [98] and experimental data on the rate of beta-hairpin formation [99]. However, formation of the three-strand beta-sheet is necessary but not sufficient for a conformation to enter the basin of attraction of the native state. Comparison of FF and TSE contact maps with UU contact map in Fig.3.10d reveals a crucial structural element that needs to be formed in order to rapidly fold into the native conformation: specific long-range contact between L24 from RT loop and/or G54 and/or I56 from the distal hairpin (dashed oval in Fig. 3.10d). Interestingly, we do not find any specific role the contact L24 and G54 (Fig. 3.10c) by using the equilibrium sampling method to select TSE by Q [82, 83].

3.4.4 Discussion

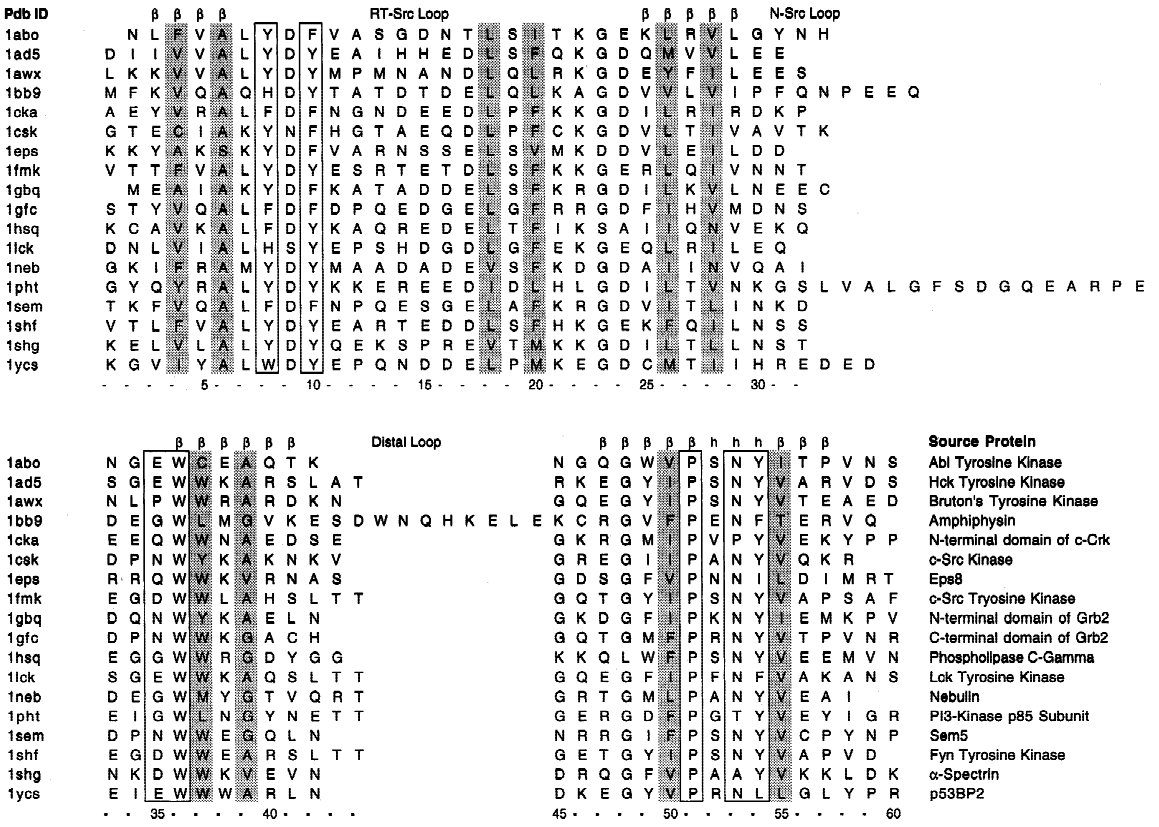


Figure 3.11: Fig. 2. SH3 domain alignment.
An alignment of the sequences of the 18 SH3 domains with solved structures. The PDB accession and parent protein are shown for each sequence. Conserved hydrophobic core positions are shaded, and the most conserved positions involved in peptide binding are boxed. Courtesy of Larson [100].

Remarkably, L24 and especially G54 are two of the most conserved structural residues in SH3 fold family [100] (see Fig. 3.11 in which case, they are residue 18 and residue 48 respectively). Furthermore, Baker and coworkers [31] showed that L24 cannot be diversified in phage-selection experiments, along with other kinetically important residues. The fact that ϕ-values of these two particular residues are not close to unity — despite strong evidence of their participation in the folding nucleus — is similar to the case of I76 in chymotrypsin inhibitor 2 (CI2) protein which

also has a low ϕ-value but appears to participate in the folding nucleus [95]. Such apparent contradiction was explained for CI2 by Fersht and coworkers who showed that the strain in the native structure may account for this anomalous behavior of a residue. This explanation is likely to be also valid for C-Src SH3 domain given the extremely tight packing of C_α of G54 against C_β of L24 in the native structure of C-Src SH3. The site mutation of G54 destabilizes the native state, but may not destroy the backbone interaction and thus can not probe the transition states properly. Importantly, residues that are sequence neighbors of G54 have all large ϕ-values while sequence neighbors of L24 have low ϕ-values, fully consistent with our findings (Figs. 3.8b and 3.10).

We further verify the crucial role of contact between L24 from the RT loop and the C-terminal strand of the distal hairpin by "cross-linking" L24 and G54. As shown in Fig. 3.12, the cross-linking dramatically changes the cooperativity of the folding transition by essentially eliminating the free energy barrier between folded and unfolded states, and shifting equilibrium toward the manifold of folded states. To see if this change can be attributed to non-specific stabilization due to the entropy reduction of the unfolded state caused by cross-link [87, 88], we perform a control simulation with N- and C-termini cross-linked [101] and rule out this possibility (Fig. 3.12). We find that the NC cross-linked protein is indeed more stable (T_f increases) than the wild type, but the barrier between the native and unfolded states remains intact, in sharp contrast to the L24–G54 cross-linked protein.

Thus we reconstruct a comprehensive picture of the C-Src SH3 folding mechanism derived directly from folding kinetics simulations. The three-stranded beta-sheet and diverging turn is present in the TSE, in accord with previous analysis. However, while this structural feature is present in the TSE, it is not sufficient for folding. A key long-range contact between L24 and the distal hairpin (residues 54-56) must be formed in order to enter the basin of attraction of the native state, causing direct and fast descent to the native state. These kinetically relevant amino acid interactions can not be obtained from the thermodynamic approach [82,83] to study the TSE by using some global reaction coordinate (such as Q). We predict that cross-linking these residues (by mutating them to cysteines) [102] would dramatically change the free

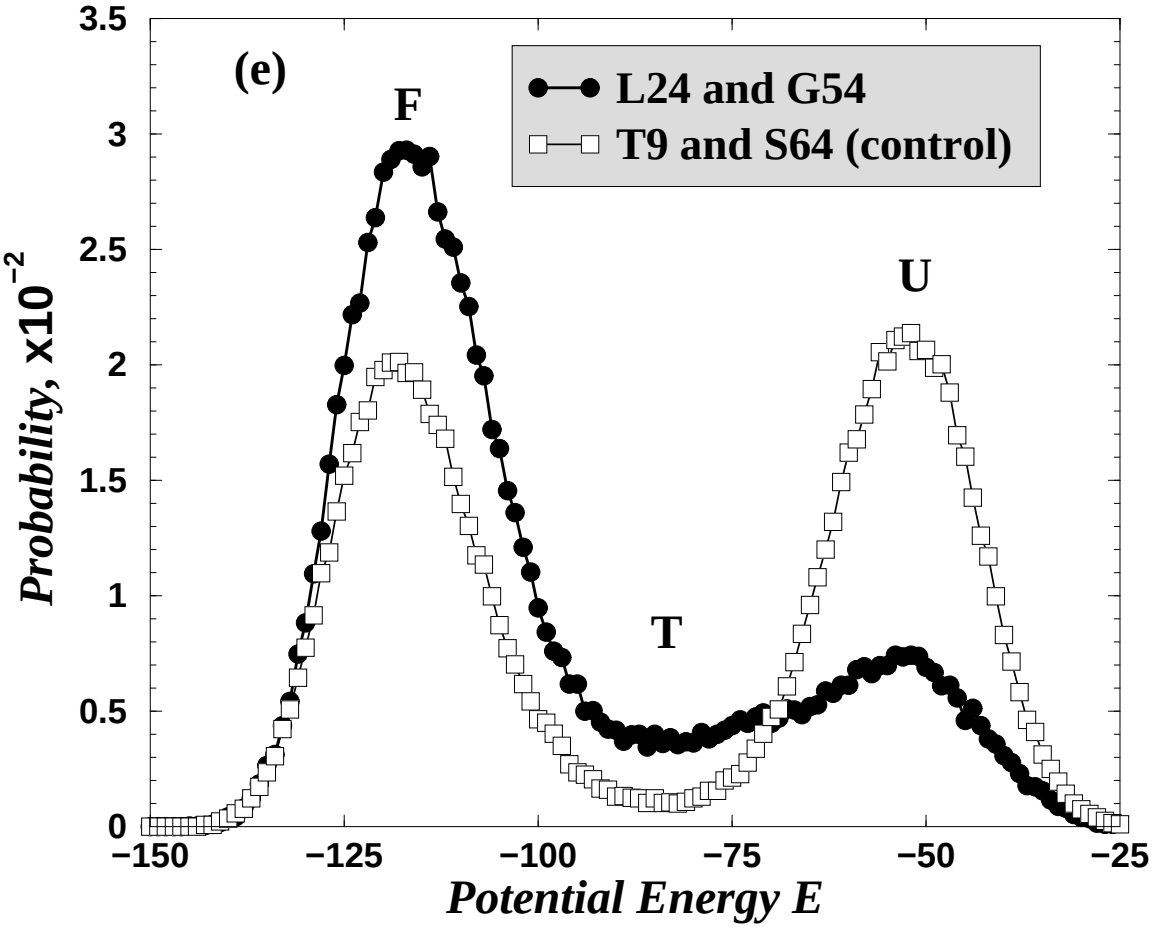


Figure 3.12:

(e) The probability distribution of potential energy E for cross-linked L24 and G54 shows suppressed bimodality. The distribution for NC cross-linked protein (T9 and S64) is as bimodal as for the wild type of Fig. 3.3d.

energy landscape, and it would be interesting to test this prediction experimentally. The crucial kinetic roles of these amino acids, especially G54 may contribute to the high conservatism in SH3 fold family [100]. This model and discrete molecular dynamics simulations used to analyze it represent a combination of structural and dynamic realism with computational efficiency needed to gain statistically significant insights into structural features of the main milestones along the protein folding pathway.

3.5 Dissect the transition state ensemble

The concept of the protein transition state ensemble (TSE) [103, 104] is the foundation of modern views on protein folding. Conformations of proteins belonging to the TSE are unstable and by definition have a 50% probability to fold to the protein native state, and a 50% probability to unfold or misfold. The TSE conformations belong to the free energy barrier separating native and unfolded or misfolded domains for two-state proteins. The principle difficulty to find the evasive TSE conformation is that the reaction coordinate of folding transition is not well defined. For example, it has been shown from our kinetics studies of C-Src SH3 that potential energy as well as fraction of native contact Q is not a good candidate. The conformations of events extended into the putative transition region defined by potential energy — UU, FF, UF, and FU — are not all true TSE conformations. The UU or "pre-transition" conformations are *en route* to the native domain from the unfolded state but the transition barrier has not been crossed. While the FF or "post-transition" conformations are *en route* to the unfolded domain from the native state but the transition barrier has not been crossed. To understand the structure of the TSE conformations we must determine the difference between "pre-transition" states and "post-transition" states.

The distinguishing kinetic feature between pre- and post-transition conformations is their probability to reach the native state domain, p_{FOLD} [30]. Since in the post-transition conformations the nucleus [87] is not disrupted, these conformations are more probable to fold than pre-transition conformations in which the nucleus is not formed. If both pre- and post-transition states are structurally and energetically close to the TSE, the question is then what global properties distinguish these states from each other?

To answer this question, we systematically the pre- and post-transition states. For C-Src SH3 domain we use the UU and FF conformations. In order to generalize our study, the same calculations are done in parallel for another protein chymotrypsin inhibitor 2 (CI2) by our collaborators in Harvard University. The simulation on CI2 is done by an all-atom Monte Carlo simulations [105] and the pre- and post-

transition states are sample using the method of Vendruscolo et al. [93]. We verify that the p_{FOLD} of the selected pre- and post-transition conformations is ≈ 0 and 1 correspondingly for both proteins (Table 3.1).

We find that such structural properties of protein conformations as radius of gyration (R_G), rms displacement (RMSD) from the native state, solvent accessible surface area, and contact order [98] can not distinguish the pre- and post-transition conformations (Table 3.1). Correspondingly, the entropy of the pre- and post-transition conformations cannot account for the difference between these conformations. We also find that the potential energies (E) and the total number of contacts between amino acids are within error bars from each other in the pre- and post-transition conformations. If the pre- and post-transition conformations are similar to each other structurally, we hypothesize that there may be a difference in the topology of the network of amino acid interactions in these conformations.

Protein	Relation to TSE	Number of conf.	p_{FOLD}	R_G, Å	RMSD, Å	SASA, $\times 10^3$ Å^2	Contact order, %	Number of contacts	E	L
CI2	post-	20	0.89 ± 0.07	13.0 ± 0.5	5.2 ± 0.9	6.5 ± 0.2	19 ± 1	183 ± 5	-102 ± 13	3.5 ± 0.1
	pre-	6	0.02 ± 0.04	13.0 ± 0.2	5.9 ± 0.3	7.1 ± 0.3	19 ± 2	171 ± 4	-130 ± 19	4.4 ± 0.4
C-Src SH3	post-	10	0.96 ± 0.01	11.2 ± 0.3	4.9 ± 0.3	4.5 ± 0.1	22 ± 3	110 ± 4	-85 ± 2	2.73 ± 0.03
	pre-	10	0.26 ± 0.08	11.8 ± 0.3	4.7 ± 0.1	4.4 ± 0.1	16 ± 2	102 ± 7	-84 ± 3	3.31 ± 0.06

Table 3.1: List of different parameters of pre- and post-transition state ensemble.

The structural (R_G, $RMSD$, solvent accessible surface area (SASA), contact order, number of contacts), energetic (E), and topological properties (L) of pre- and post-transition states of CI2 and C-Src SH3 domain proteins. The values of p_{FOLD} correlate only with L-values: the post-transition states are characterized by $p_{\mathrm{FOLD}} \approx 1$ and their L-values are smaller than for the pre-transition states, that are characterized by $p_{\mathrm{FOLD}} \approx 0$.

To study the topology of pre- and post-transition conformations, we construct graphs corresponding to these conformations in which nodes represent amino acids and edges represent those pairs of amino acids that are geometrically located within interaction distance from each other. Vendruscolo et al. [106] have shown recently that the "small-world" feature [107–109] of proteins can be used to identify the key residues that stabilize the structure of the transition state. Our hypothesis is that the network of amino acid interactions in post-transition conformations is more "small-world" like [107–109] than that in pre-transition conformations. The small-world graphs are a special class of random graphs that are as strongly connected as regular graphs (the clusters have a similar structure to regular graphs), but the average path that spans two nodes via a minimal set of graph edges is as low as for random graphs [106, 110] (Fig. 2 of Ref. [107]). The difference between regular, small-world, and random graphs is the "wiring" of these graphs: regular graphs are strongly locally connected with no long-range edges, random graphs are locally disconnected but have many long-range edges, while small-world graphs are the blend of the high local connectivity with a number of the long-range contacts. Small-world graphs are characterized by small separation of nodes from each other, which for proteins means a higher degree of *interaction cooperativity*. Thus, we hypothesize that the wiring of the post-transitional conformation graphs is "tighter" than that of the pre-transition conformation graphs, resulting in a cooperative folding to the native state domain.

Definition 3.3 *The protein graphs are constructed based on the C_α representation of proteins. Each graph node represents an amino acid. Each graph edge connects pairs of nodes that correspond to pairs of amino acids that are geometrically located within an interaction threshold radius, which we set to $R_c = 8.5$ Å. We test graph connectivity properties for various definitions of contacts and find that these properties are qualitatively invariant under contact definitions.*

A simple measure of topological properties of the graph is an average minimal path along the edges between any two nodes of the graph, L, proposed recently by Dokholyan et. al. [106, 111]:

$$L = \frac{1}{N(N-1)} \sum_{i>j}^{N} \ell_{ij} \,, \tag{3.9}$$

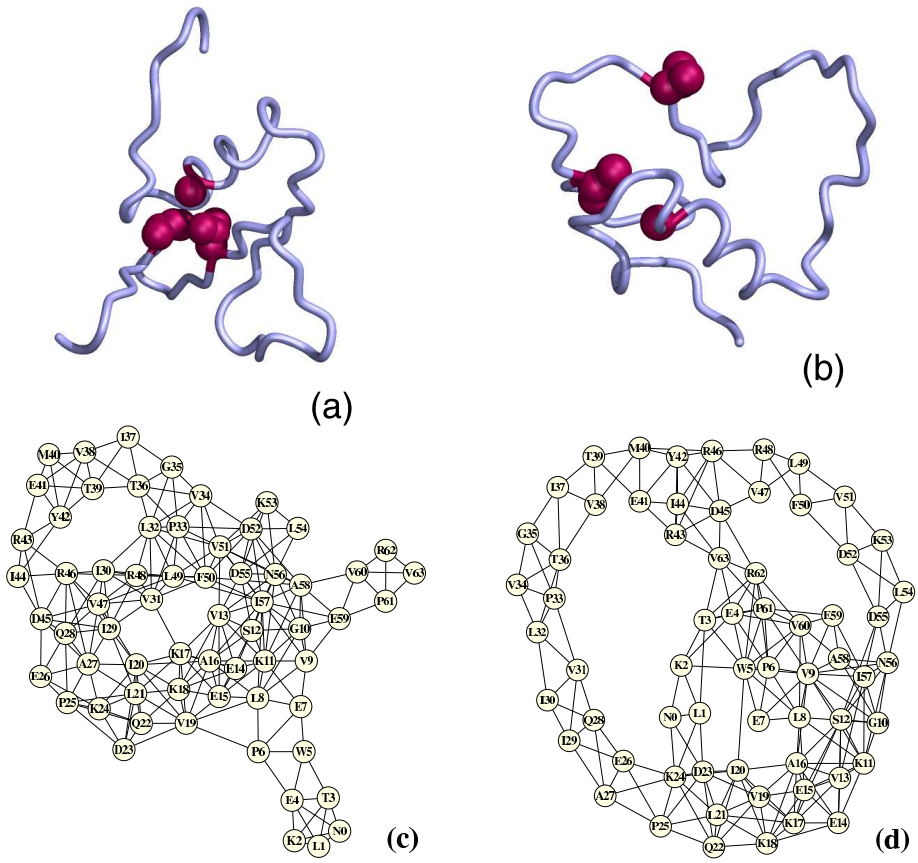


Figure 3.13: The protein graph of the typical pre- and post-transition state. The three-dimensional structure of the CI2 protein in (a) post- and (b) pre-transition states. The protein graphs are constructed based on the structure of (c) post- and (d) pre-transition states. Each node of protein graphs corresponds to an amino acid, while each edge between a pair of nodes corresponds to that pair of amino acids that are geometrically in contact with each other. For both CI2 and C-Src SH3 domain proteins' graph constructions, the contact between two amino acids is considered to be present if the distance between corresponding C_α atoms is less than 8.5 Å. In (a) and (b), residues A16, L49, and I57 belonging to the specific nucleus of CI2 [95] are denoted by red spheres. A16, L49, and I57 form a triad of contacts in post-transition conformations (a), while such contacts are missing in the pre-transition conformations. In both pre- and post-transition states the number of edges (contacts) are approximately the same.

where N is the number of amino acids, ℓ_{ij} is the minimal path between nodes i and j. L-values characterize the "tightness" of the network by computing the average separation of elements from each other.

We compute the average minimal distance L between any pair of nodes of a graph by counting the minimal set of edges that connect these nodes [107]. We find that the L-values for post-transition conformation graphs are distinctly smaller than those for the pre-transition conformation graphs, thus fully supporting our hypothesis (Table 3.1). We also observe that the post-transition conformation graphs have more edges that are of intermediate- and long-range than pre-transition ones (Fig. 3.13), which shortens the minimal path for each node k, $L(k)$, (Fig. 3.14), thus creating a more cooperative network for the former graphs. A similar mechanism was observed by Watts and Strogatz [107], who, by re-wiring circular graphs by removing local edges and creating a few long-range edges, were changing the graph properties from the regular to the small-world. Interestingly, some CI2 pre-transition conformations have N- and C-termini in contact, in contrast to post-transition conformations (Figs. 3.13 and 3.14). Although the contact between the N- and C-termini is of longest-range, the lack of intermediate-range contacts nevertheless makes pre-transition conformation networks less "cooperative" than post-transition ones. The difference between the numbers of long-range contacts in pre- and post-transition conformations is not statistically significant, so that the average contact orders for both conformation ensembles cannot discriminate between pre- and post-transition ensembles.

An important property of the L-values of protein conformations is that they can serve as a structurally reliable determinant of the pre- ($p_{\mathrm{FOLD}} \approx 0$) and post-transition ($p_{\mathrm{FOLD}} \approx 1$) states. The principal difficulty to select TSE conformations — the basis of the protein engineering experiments — is the identification of the reaction coordinate for protein folding. The reaction coordinate for folding is not well defined [24, 30, 84], and has yet to be identified. The fact that average graph connectivity distinguishes the protein pre- and post-transition states, which can be close along the reaction coordinate to the TSE, tells us that any future constructions of the reaction coordinate should strongly depend on the structure of protein interaction networks.

Interestingly, in experimental studies of CI2, the cleavage between amino acids

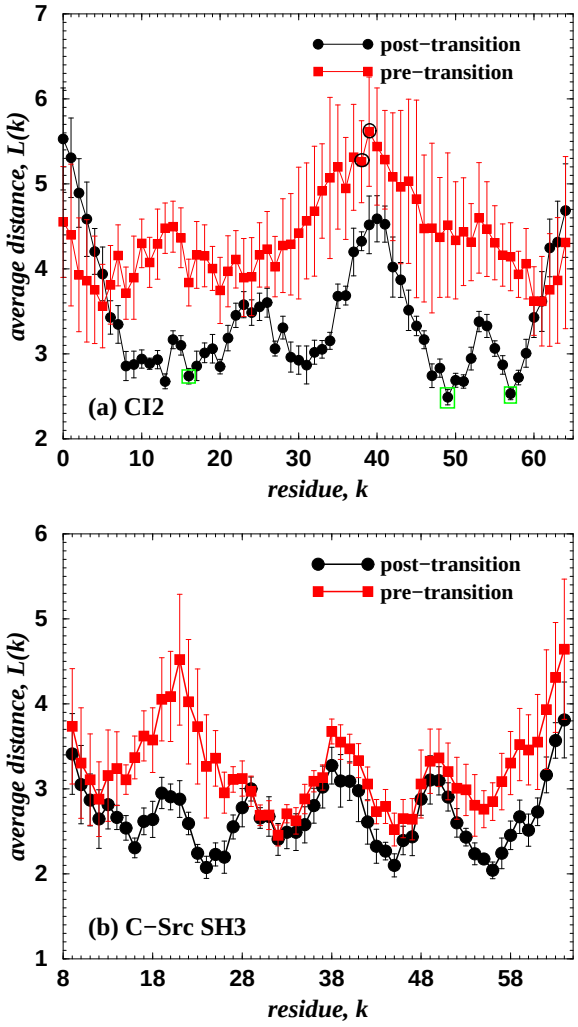


Figure 3.14: The $L(k)$ values of pre- and post-transition states for CI2 and C-Src SH3 domain.

The dependence of the average minimal distance $L(k)$ between a node k and the rest of the nodes on (a) CI2 and (b) C-Src SH3 domain proteins' graphs for post- (•) and pre-transition (▪) states. The error bars represent the standard deviation from the average values of $L(k)$ over all post- and pre-transition states. In (a) by the open circles (◯) we denote amino acids M40 and E41 that do not affect the protein three-dimensional structure after cleavage of 40-41 bond [112]. In (a) by the open boxes (□) we denote the folding nucleus of CI2 [95], A16, L49, and I57.

M40 and E41 is the only one that does not destroy the protein's three-dimensional structure [112]. Neira et al. cut CI2 at M40–E41 (without circular permutation) to separate fragments 1–40 and 41–64 and found that these fragments re-associate into CI2 [112]. We find that amino acids M40 and E41 have the largest values of $L(k)$ in the pre-transition states and among the largest values in the post-transition states (Fig. 3.14a), indicating that these amino acids are the most separated on interaction network from the rest of the amino acids. Weak participation of amino acids in the protein interaction network in pre- and post-transition states means that these amino acids have weak impact on protein folding kinetics and on the final native state of the protein (since the folding pathway is not altered). Thus, our findings are in agreement with [112].

A crucial factor that distinguishes pre- and post-transition states is the protein folding nucleus, the formation of which in the TSE results in the rapid folding transition to the native state, and the disruption of which results in the global unfolding [87]. Pre-transition states lack the folding nucleus, while post-transition states have it intact (Fig. 3.13a,b). Thus, the difference of $L(k)$ between the pre- and post-transition states, $\Delta L(k)$, is most pronounced for those amino acids that are part of the protein folding nucleus. We find that for C-Src SH3 domain (Fig. 3.14b) $\Delta L(k)$ is most pronounced for two fragments, RT-loop (16–26) and β4 (54–61), suggesting a crucial role of the connectivity between these fragments in the TSE. This observation is in agreement with our finding of nucleus, where nucleus of C-Src SH3 domain is identified on the RT-loop and β4. We also find that for CI2 (Fig. 3.14a), the experimentally identified folding nucleus [95] — A16, L49, and I57 — has one of the largest $\Delta L(k)$ values.

We presented a new structure-based topological criterion that appears to be a good predictor of kinetic ability to fold for a given conformation. The fact that this criterion performed equally well for two different proteins, simulated within different models using different techniques suggests its generality. Moreover, the recent all-atom Monte Carlo analysis of TSE of protein G done by our collaborators (J. Shimada and E.I. Shakhnovich, unpublished) also shows consistency with the proposed criterion. Further theoretical understanding of deep connection between

topological properties of protein conformations and their kinetic ability to fold is a challenging task for future studies.

Chapter 4

Protein Aggregation Problem

4.1 Introduction

Proteins carry out various functions in the body because of their specific three-dimensional structures. Nascent proteins can assembly into the their unique native states themselves (sometimes in the help of chaperones). What will happen if this process goes wrong? Of course, the protein can never perform their usual functions. However, the worst thing is that the misfolded proteins will aggregate and condensate into insolvable fibrils or plaques, which will result into the malfunctioning of the cellular machinery [113]. Some of these aggregates are extremely toxic, for example the aggregates in the case of Alzheimer's disease will lead to the neuron cell death. The final forms of these aggregates often have a well-defined fibrillar nature (see Fig. 4.1), and are known as amyloid, hence the term amyloidosis is used to describe many of the clinical conditions with which they are associated.

This group of diseases, of which nearly 20 have been described, includes Alzheimer's and Parkinson's diseases, the spongiform encephalopathies such as Creutzfeldt-Jakob disease, type II diabetes and a range of less well-known but often equally serious conditions such as fatal familial insomnia [115, 116]. These diseases can be sporadic, inherited or even infectious, and are often manifest only late in life. Each disease is associated with a particular protein and aggregates of these proteins are thought to be the direct or indirect origin of the pathological conditions associated with the dis-

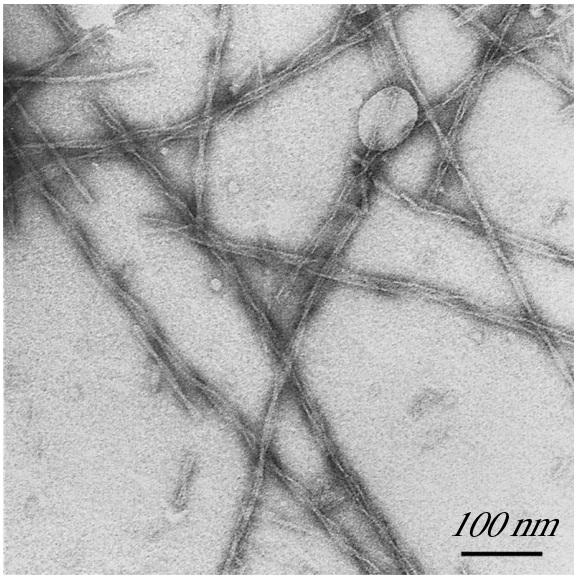


Figure 4.1: The electron micrograph of amyloid fibrils formed by SH3 domain. The electron micrograph of amyloid fibrils formed by SH3 domain, a protein known to have no relationship to any aggregation diseases. Courtesy of Guijarro [114]

ease in question. In some cases, the quantity of material involved is enormous, with several kilograms of protein being deposited in certain manifestations of systemic amyloidosis. Remarkably, despite the range of proteins involved in these diseases, including several well-known proteins such as lysozyme, transthyretin and the prions, all of which have unique and characteristic the fibrils in which they are found in the disease states are extremely similar in their overall appearance [117].

Amyloid fibrils are straight, unbranched, usually 70-120Å in diameter, and several thousand Å in length. Observed types of amyloid fibrils consist of different precursor proteins which share no sequence or structure similarity. However, different types of amyloid fibrils explored by x-ray diffraction [118–120] show some common core structural features (see Fig. 4.2a): the presence of a 4.7Å inter-strand spacing along the fibril axis and a 9-10Å inter-sheet spacing perpendicular to the axis. The combination of Cryo-electron microscopy technique and x-ray study [121] has applied to reconstruct the three dimensional structure of amyloid fibrils formed by SH3 domain (Fig. 4.2b).

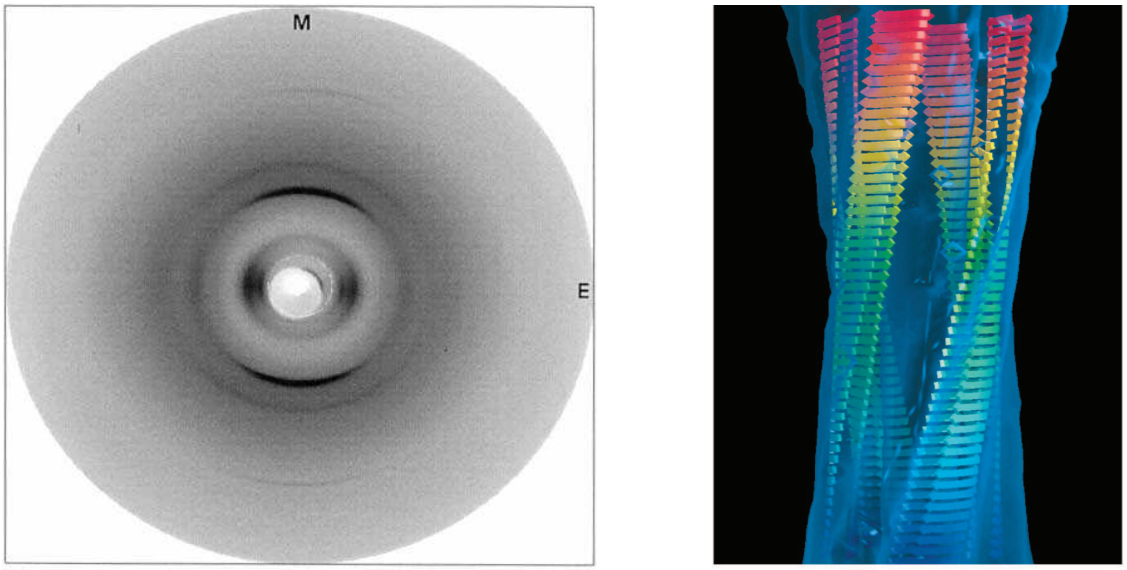


Figure 4.2: X-ray pattern and cartoon of amyloid fibril.
(a) The typical fibril x-ray diffraction pattern. The axis of fibrils is along the meridional (M) direction. There are two peaks observed: the first one is along the meridian corresponding to 4.7Å, the second one is along the equatorial direction corresponding to 7-10Å. (b) The 3D reconstruction of amyloid core structure from Cryo-electron microscopy. Courtesy of Dobson [121].

Studies of the mechanism of the conversion of the normally soluble proteins into amyloid fibrils have benefited from the fact that, in many cases, the structural transitions of the disease-associated molecules can be reproduced under laboratory conditions [116]. In order to achieve this, a common procedure has been to expose the folded proteins to mildly denaturing conditions, such as low pH or elevated temperatures. There are accumulating evidences that the formation of amyloid accounts for the partially unfolded states of globular proteins, some soluble intermediates with a predominantly β structure, while the totally unfolded proteins lead to amorphous aggregation [122]. Interestingly, it has been reported [117] that some proteins with α rich native states can have this amyloid conversion, giving rise to the well-known α-helix-to-β-sheet transition in protein folding.

Recently, a group of proteins unrelated to any human disease were found to be

able to form amyloid fibril structures in *vitro* under denaturing conditions [114, 123]. The ability of proteins with different sequences and native structures to form similar amyloid fibrils suggests that amyloidogenesis is a common feature of proteins in denaturing conditions [124].

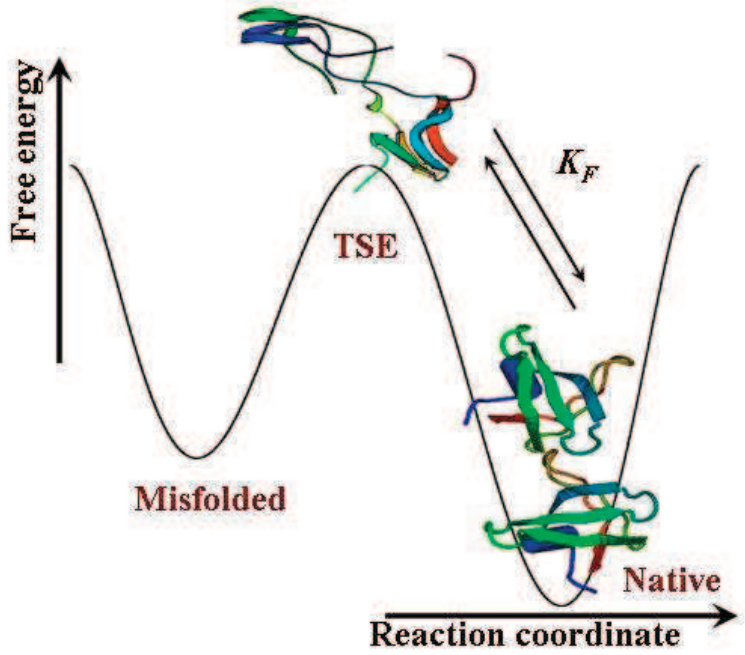


Figure 4.3: The schematic diagram of protein aggregation hypothesis. The deterioration of condition lead to the destabilization of native states and the protein has high probability to stay in the unfolded states.

The *in vitro* study of protein aggregations on many systems under conditions such as low pH value, high temperature and mutants are consistent with an aggregation hypothesis that the destabilization of native states will lead to the aggregation of proteins. The schematics free energy landscape diagram in Fig. 4.3 demonstrate this aggregation scenario. The protein under normal condition has two minima in the landscape (1D projection is shown) corresponding to folded and unfolded states respectively. The barrier in between keeps the protein stable in the folded

states. With deterioration of conditions, the native state becomes unstable such that the protein has high probability to stay in the unfolded states. With exposure of hydrophobic core and the opening of backbone (the hydrogen bond donor and receptor), the aggregation becomes favorable.

The aggregation of proteins with no sequential and structural similarity into the overall similar β-rich amyloid structure suggests that the non-specific backbone hydrogen interaction is important in this process. The possible scenario is that at first stage of aggregation partially folded (still unfolded) meet each other by hydrophobic interactions and the exposure of backbone from different protein in the unfolded states will pack together by forming the hydrogen bonds networks. However, which part of the protein lead to the core structure of amyloid structure remain unknown. The main problem comes from the difficulty to crystallize the amyloid fibril and thus the fine x-ray diffraction has yet to produce the intrinsic structure of amyloid fibrils. Lack of knowledge of the detailed structure of amyloid fibril makes it difficult to understand aggregation mechanisms.

Although many advances have been made in structural characterization of amyloid fibrils and the mechanism of their formation, many aspects of this process remain unclear. Due to difficulties in crystallizing amyloid fibrils, the detailed intrinsic structure has yet to be determined from x-ray diffraction. Lack of knowledge of the detailed structure of amyloid fibril makes it difficult to understand aggregation mechanisms. Some alternative experimental techniques have been applied [125, 126] to understand the structure of amyloid fibril core. H/D exchange of amide protons combined with NMR analysis [125] shows that the β_2-microglobulin amyloid fibril β-sheet core is composed the middle region of the protein, including the loop regions in the native structure, while the N- and C-termini are excluded. Designing different fragments of amyloid β-peptide (Aβ) [126] that can aggregate into fibrils similar to those formed by wild type peptides shows that residues 14-23 are the basic "bricks" composing the Aβ amyloid fibrils. However, no direct observations of the amyloid fibril core structures have been reported.

Recent studies of some protein dimer structure propose that "Domain swapping" [127], can explain the prolongation of amyloid fibrils. The domain swapping

mechanism [127] posits that two or more protein chains exchange identical domains including a helix, a loop, a single β-strand or an entire domain to form a strongly bound oligomer. In a propagational instead of reciprocal manner, the domain swapping mechanism explains the elongation of amyloid fibrils [128–131]. However, the domain swapping hypothesis is based on the aggregation of only two proteins into dimers, so the mechanism for fibril formation from more than two identical proteins is still unclear.

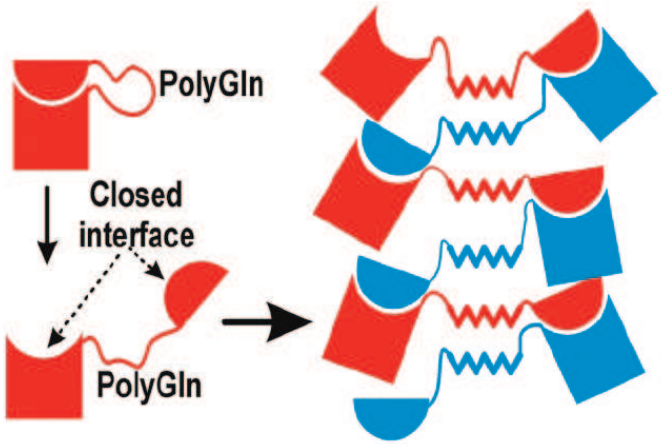


Figure 4.4: Domain swap scenario to form amyloid fibril
The schematic diagram of domain swapping mechanism to form the elongated amyloid fibrils.

4.2 Aggregation of SH3

Due to limitations in computation power to study large protein systems in molecular dynamics simulations, we employ the discrete molecular dynamics algorithm [8, 9] — a computationally fast and dynamically realistic simulation technique for investigations of the protein folding thermodynamics [9] and kinetics [12, 132]. We study the aggregation of Src SH3 domain (Protein Databank entry 1NLO), a globular protein, consisting of 56 amino acids, extensively explored in experiments [31, 32] and computer simulations [10]. The longer SH3 fold family homologue PI3-SH3 has been experimentally shown to aggregate into amyloid fibrils under acidic condi-

tions [114, 122]. PI3-SH3, a 84 residue protein with the insertion of a long helical loop between $\beta 2$ and $\beta 3$, shares the same fold as Src SH3 domain. The rmsd between the structure of Src SH3 and PI3 SH3 is 1.04Å with 48 amino acids used for alignment [133]. Experimental study of PI3-SH3 [134] shows a slow refolding with the time constant 2.8 seconds in water, while the folding kinetics still follows a two-state folding scenario. It is a challenge to simulation such a slow folding protein. With a modified Gō model, we find the 84 residue SH3 domain follows a two-state folding transition and our simulation of two PI3-SH3 proteins shows the same aggregation scenario as Src SH3 domain (data not shown). However, the requirement to simulate more than two proteins in the aggregation study is extremely time-consuming for PI3-SH3. Thus, we use Src SH3 domain as the model system to study the amyloidogenesis process.

The ability of proteins with different sequences to aggregate into common amyloid fibrils suggests that non-specific hydrogen bonding between the main chain carbonyl oxygen and the amide nitrogen may play an important role in amyloid formation. In the present study we use the the coarse-grained protein model as in Ref. [10], with a native state specific Gō interaction potential between C_β atoms representing the side chains, and non-specific interactions between C_α atoms representing the hydrogen bonding interactions between the backbones of proteins.

The folding kinetics of our Src SH3 domain model, a C_α – C_β model with a Gō interaction between side chains, has been studied by discrete molecular dynamics and has shown agreement with experimental observations [10]. We simulate the model proteins near the folding transition temperature, T_f, where the protein has a high probability to be found in the partially folded states. It is most likely to observe amyloid fibril formation under these conditions because complete unfolding was shown to lead mostly to amorphous aggregation [122] while highly stable proteins do not aggregate at all.

The Src SH3 domain under study consists of five β-strands and a long RT-loop (see Fig. 4.6a). It was observed that the RT-loop plays a critical role in the folding kinetics of Src SH3 domain despite the low experimental ϕ values [10]. Experiments and simulations reveal that the RT-loop is flexible. However, the RT-loop itself

is stable and persists in the partially folded states. The unfolding/folding events correspond to the opening/closing of the RT-loop with respect to the rest of the protein [10, 94]. Accordingly, we expect that the RT-loop may play an important role in the amyloid formation of Src SH3 domain.

4.2.1 Two-bead model with hydrogen bond interaction

The principal difficulty to study the protein folding *ab initio* is the lack of knowledge about the energetics between amino acids. The native state specific Gō [8, 17, 83] potential has been successfully used to model amino acid interactions. It has been shown [10] that our coarse-grained model with a Gō interaction potential for the Src SH3 domain can faithfully reproduce the thermodynamic and kinetic properties observed in experiments. Thus, we use the Gō potential to model interactions between C_β atoms for a single protein. In order to reproduce the process of aggregation, we need to simulate more than one protein and to model the interaction between different proteins. For simplicity, we apply the Gō potential for C_β atoms between different proteins by assuming that two amino acids that attract to each other in a single protein will also have attraction in different proteins. The cutoff distance to define a contact is set as 7.5Å.

The ability of proteins with no sequence similarity to aggregate into the same amyloid structure indicates that the non-specific backbone hydrogen bonding interaction may play an important role in the process of amyloidogenesis process. It has been shown [135] that only backbone hydrogen bonds can lead to a cooperative formation of two-dimensional β-sheet. We add to our model a non-specific interaction between any two C_α atoms to model the hydrogen bonding interaction between protein backbones. We add to two-beads model a type of non-specific interaction between any two C_α atoms to model the hydrogen bonding interaction between protein backbones. It has been observed in many globular proteins that the number of backbone hydrogen bonds for the each residue does not exceed two (bifurcated hydrogen bonding is very rare and is not considered here). Another important property of the two hydrogen bonds formed by one peptide block is that they are approximately

parallel to each other. This is a reason for the formation of two-dimensional planar β-sheet. In the present study, *(i)* one C_α atom can not make more than two effective hydrogen bonds, and *(ii)* the two hydrogen bonds must be aligned linearly.

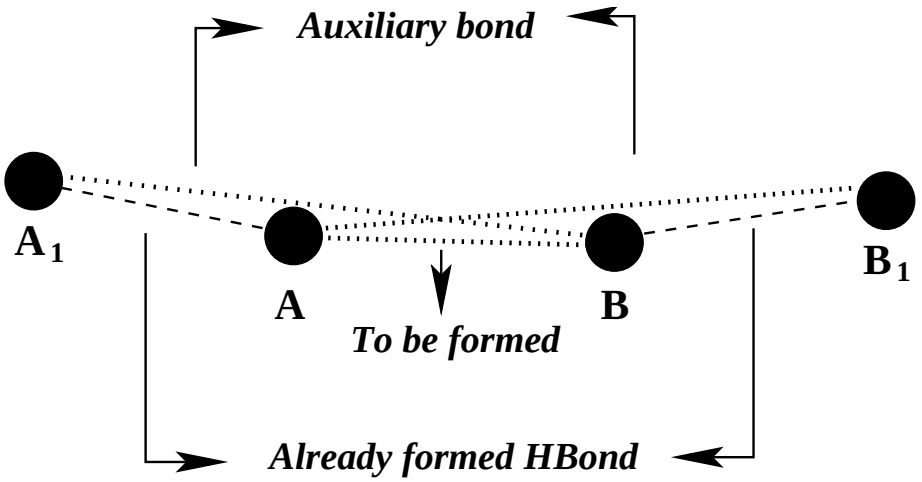


Figure 4.5: Model of a hydrogen bond.

Existing hydrogen bonds AA1 and BB1 are shown in dashed lines. When the beads A and B come to a distance 5Å, a new hydrogen bond (dotted line) may form if the distances A1B and B1A satisfy inequalities 8.7Å$\leq$ A1B$\leq$ 10.0Å and 8.7Å$\leq$ B1A $\leq$10.0Å. If the bond AB is formed, the auxiliary bonds A1B and B1A (dashed lines) are formed simultaneously. These bonds can fluctuate within the interval 8.7–10Å and cannot be broken unless beads A and B move away from each other to a distance 5Å. If the beads A and B have enough kinetic energy to leave the hydrogen bond attraction well, their velocities are changed in order to conserve energy and momentum, and the hydrogen bond AB is destroyed simultaneously with the auxiliary bonds A1B and B1A. The velocities of A1 and B1 do not change at the moment of forming or destroying of hydrogen bond AB. Analogously, if one of the hydrogen bonds, A1A or B1B, breaks before hydrogen bond AB, the corresponding auxiliary bonds A1B or B1A also breaks.

We set the hydrogen bond interaction range between two C beads to $D^{HB} =$ 5.0Å, and their hard-core distance to $D^{HB}_{HC} = 4.0$Å. We use the following procedure in order to satisfy the criteria for the hydrogen bond formation: when two C beads,

A and B, come to a distance D^{HB}, we check for any existing hydrogen-bond partners of A and B. If both beads A and B have no existing hydrogen partners they can form a hydrogen bond automatically. If one of the beads, for example A, already has one partner, A1, and the distance between the bead A1 and the bead B is within the range of 8.7–10Å (i.e. the angle between vectors $\vec{AA}1$ and $\vec{AB}$ is within the range of $120^o - 180^o$), the bead A can form another hydrogen bond with bead B provided that either the bead B has no existing hydrogen bonds or its single hydrogen bond partner, B1, has a distance with bead A in the range of 8.7–10Å (see Fig. 4.5). If one of beads A and B or both already have two hydrogen bond partners, the pair will proceed with a hard-core collision without forming a new hydrogen bond. When a new hydrogen bond is formed between beads A and B, new hydrogen bond partners are recorded for these two beads, and whenever a bead gets two hydrogen bond partners an auxiliary bond is formed between these two partners. Every auxiliary bond can fluctuate within the range of 8.7–10Å to keep two hydrogen bonds within the angle $120^o - 180^o$ and it cannot be broken unless one of the two hydrogen bonds is broken. A hydrogen bond between beads A and B can be broken when these two beads move away from each other to a distance of D^{HB} and their kinetic energies are higher than ϵ_{HB} . When a hydrogen bond is formed or broken, the velocities of the beads A and B change in order to conserve energy and momentum, such that their kinetic energy increases or decreases by the value ϵ_{HB} .

We perform discrete molecular dynamics simulation to model the protein system. We set $\epsilon_{HB} = 3$ and $\epsilon_{G\bar{o}}$=1 so that we favor the formation of non-specific backbone-backbone hydrogen bonding formation. Since the formation of backbone hydrogen bonds comes with a large reduce of entropy comparing to the formation of a Gō contact, therefore we need to assign a larger potential energy gain. We also study the thermodynamics for a monomer SH3 domain and find that the additional hydrogen bonding interaction does not affect the two-state folding transition and the folding transition temperature T_f is slightly increased to $T_f = 0.95$ comparing the model without hydrogen bonds (T_f=0.91).

The number of proteins we studied varies from 2 to 8. The concentration of proteins in our simulation system is usually higher than *in vivo* and *in vitro* condi-

tions, so that the condensation process is much faster and enables us to access the amyloidogenesis process by discrete molecular dynamics. First, we heat the system to high temperatures so that all the proteins are fully unfolded and moving freely inside the system box. Then, we quench the system to the temperature around T_f and wait for the system to equilibrate. The final equilibrium states result in possible amyloid fibril structures of Src SH3 domain.

4.2.2 Dimerization

First, we study the dimerization of two identical Src SH3 domains. The two proteins are confined to a cubic cell with length of 150Å. Starting from fully unfolded states, we quench the system to different temperatures. We observe an aggregation temperature threshold, T_a, below which we find aggregation, and $T_a \approx 1.03$ is only slightly higher than folding transition temperature $T_f \approx 0.95$. Ordered aggregations only occur near T_f. At T_f, the time needed for aggregation, τ_a, is of the order of 10^4 time units. This time is significantly smaller that the time needed for a single protein to fold into native state, τ_f, which is in the order of 10^5 time units, indicating that the kinetic barrier for the two-protein system to aggregate is much smaller than the folding barrier of each individual protein. As temperature decreases, τ_a increases and τ_f decreases. Below a certain temperature threshold $T_c = 0.85$ we observe the separate folding of the two proteins without dimerization.

For ordered aggregation, we observe a *closed* form dimer structure by domain swapping (see Fig. 4.6b), where the two proteins exchange their RT-loops. We also observe an *open* aggregation state (Fig. 4.6c), which relates to the packing of RT-loops. Unlike the usual domain swapping where the swapped part interacts with the *complementary* domain from a different protein, in our *open* aggregation state the first/second strand of the RT-loop from one protein forms contacts with the second/first strand of the RT-loop from another protein. During this process, amino acids in the RT-loops reorient their side chains. The original interactions that stabilize the RT loop are replaced by similar contacts between the complementary strands of RT-loops from different proteins. Stabilized by hydrogen bonds along the back-

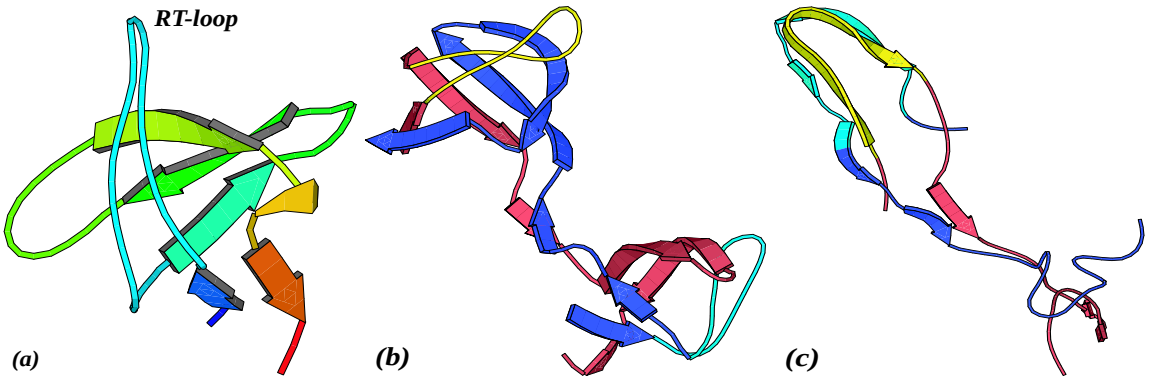


Figure 4.6: Dimerization of Src SH3 domain.
(a) The native state of Src SH3 domain. Molecular dynamics simulations yield two types of aggregates: (b) the *closed* dimer formed by exchanging RT-loops and (c) an *open* aggregation state formed by swapping two parts of RT-loop from different proteins. In (b) and (c) the first protein is red and its RT-loop is yellow, while the second protein is blue and its RT-loop is cyan. The pictures are produced by molscript [136].

bone, RT-loops form a β-sheet structure (Fig. 4.6c).

The *closed* dimer has a stable structure with lower potential energy than the more flexible *open* structure. However, the probability to observe the *closed* form is lower than the *open* aggregation state because the entropy of the *open* structure is higher than that of the *closed* form. As temperature decreases, the probability to form *closed* dimers increases. We observe that when the quenching temperature drops below $T_c = 0.85$, the two proteins fold separately, avoiding aggregation. This process depends on the diffusion coefficient D, and the density of the proteins ρ: as D and/or ρ increase, T_c decreases, and, therefore the aggregation becomes more likely.

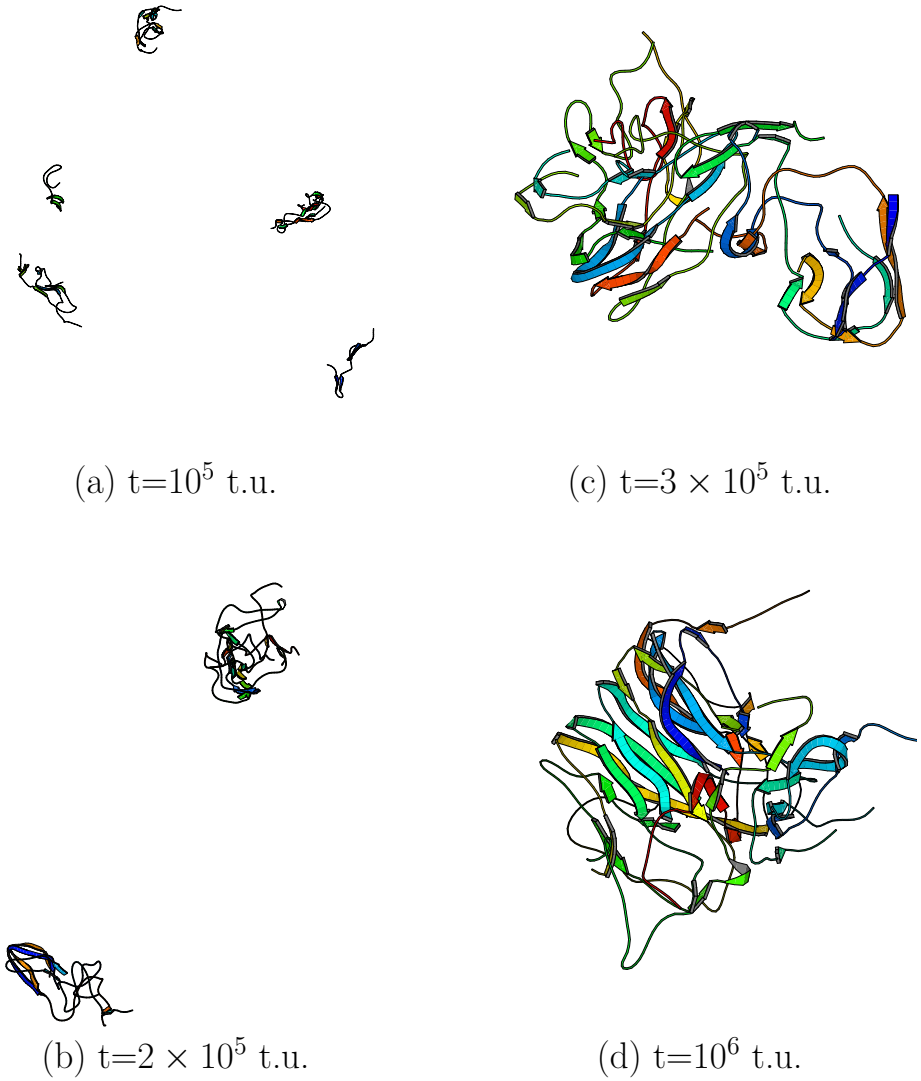
The *closed* dimer has a well-defined 3D structure with the hydrophobic core buried inside, so the *closed* dimer can not further aggregate into elongated amyloid fibril. However, according to amyloidogenesis hypothesis proposed in Refs. [128, 129], domain swapping may lead to elongated amyloid fibrils if the swapping is not reciprocal but propagational. The *open* aggregation state is more flexible and has the hydropho-

bic core exposed. The closely packed RT-loops form a β-sheet structure stabilized by hydrogen bonding interactions and the two exposed ends can accept further condensation to form the fibril structure. In order to test which process leads to the amyloid fibril formation for Src SH3, we study aggregation of more than two proteins in molecular dynamics simulations.

4.2.3 Amyloidogenesis

From the simulation of two proteins, we find that the optimal temperature to observe aggregation is T_f. Next, we perform the molecular dynamic simulations of eight Src SH3 domains in a cubic cell with the length of 300Å at T_f. All simulations from different initial configurations show similar equilibrium structures of SH3 aggregates (Fig. 4.8). Snapshots during the aggregation (Fig. 4.7) demonstrate that the initial step of aggregation is the dimerization and the formation of the *open* states which are the nuclei of aggregation process (Fig. 4.7a). Other partially unfolded proteins grow on these open states. Usually there are more than one nucleus; they merge with each other forming fibrils (Fig. 4.7b,c). Finally all eight proteins form one aggregate. However, the initial aggregate has only short range order (Fig. 4.7c). As system equilibrates, proteins rearrange themselves so that the equilibrium structure shows distinct long range order (Fig. 4.7d and Fig. 4.8).

At equilibrium, all eight proteins exhibit a tendency to form aggregates with a preferred direction of condensation (Fig. 4.8a), which can be identified as the amyloid fibril axis. In the aggregated state, proteins pack their RT-loops on the top of each other by swapping the two parts of the RT-loops. We do not observe the aggregation states formed by propagational domain swapping as proposed in Refs. [128, 129]. Packed RT-loops form a double β-sheet structure (Fig. 4.8b). As discussed above, the aggregation process involves the reorientation of amino acids along the RT-loop. Thus, the side chains (usually the hydrophobic residues) from the two parts of one RT-loop directed to each other and the separation of the two β-sheets is around 10Å(Fig. 4.8b). Due to the saturation and angular dependence property of the backbone hydrogen bonds, only the exposed proteins on the two



(a) t=10^5 t.u. (c) t=3×10^5 t.u.

(b) t=2×10^5 t.u. (d) t=10^6 t.u.

Figure 4.7: Snapshots during the aggregation of eight SH3 domains at T_f. The simulations start from fully unfolded conformations. The snapshots are taken at different times (a) 10^5, (b) 2×10^5, (c) 3×10^5, and (d) 10^6 time units (t.u.).

ends can allow further aggregation by making backbone hydrogen bonds to form an extended β-sheet structure. Thus, by adding more proteins to the two ends of the aggregate, it continues growth to form an elongated fibril structure — amyloid fibril.

Remark 4.1 *The domain swap scenario of amyloid fibril formation implicates that (1) the β-rich core is composed of the hinge region of the monomeric protein; (2) the two complimentary "domains" keep their native structures such that the final fibril*

contain a large number of native-like structures. However, in most of the cases the hinge region of a monomer is usually short which can not explain the participation of a large portion of the sequence into the fibrillar core. Analysis of the mature fibril indicates the lost of their function and structure of original protein. Therefore, the domain swapping, which might be the possible mechanism of protein oligomerization, could not lead to amyloid fibrils.

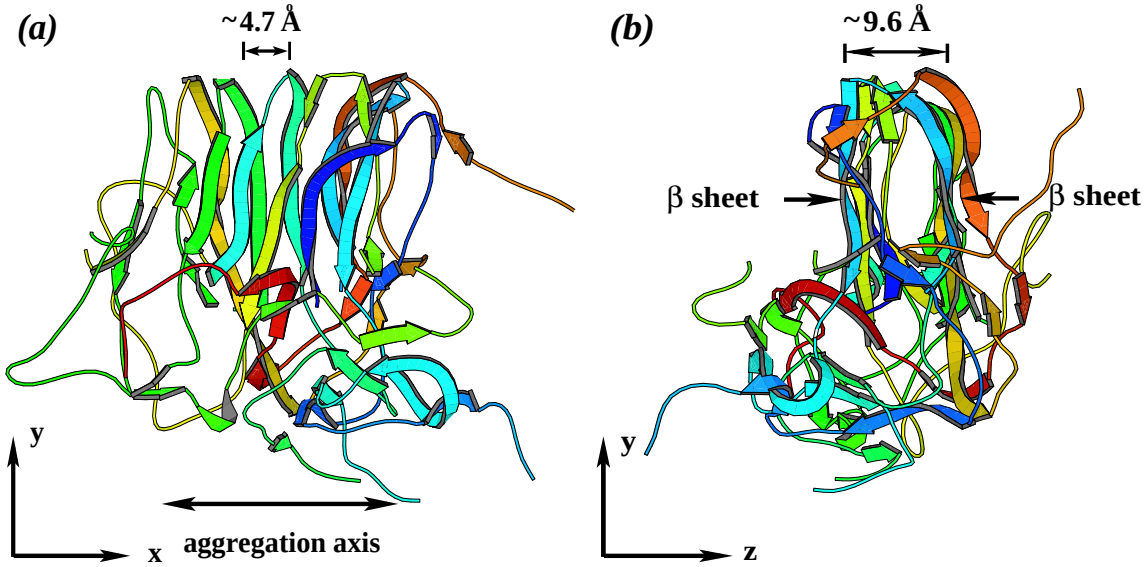


Figure 4.8: The typical equilibrium aggregation state for eight proteins (a) XY and (b) ZY projections of the aggregate. The preferred aggregation direction is along the X-axis.

4.2.4 Characterization of the aggregates

Definition 4.1 X-ray Calculation:

In order to compare with experimental x-ray diffraction patterns, we calculate the intensity of diffraction using the elastic diffraction formula

$$I(\vec{k_f}) = |\sum_j \exp(i(\vec{k_f} - \vec{k_i}) \cdot \vec{r_j})|^2 \,, \tag{4.1}$$

where $\vec{k_i}$ is the wave vector of the incoming x-ray, $\vec{k_f}$ is the wave vector of the diffracted x-ray, $\vec{r_j}$ is the position vector of jth atom, and the summation is over

all the atoms in the structure. We align the aggregation structure along the x axis as in Fig. 4.8 and choose the incoming x-ray with wavelength of 1Å along the y axis. The diffraction intensity is collected in projection of the $y-z$ plane varying the deflecting angle, $\theta = cos^{-1}(\vec{k}_f \cdot \vec{k}_i/k^2)$, from 0.05 to 0.25 in radian. As in the x-ray diffraction experiments, the amyloid fibril has no preferred orientation in the $y-z$ plane. We rotate the aggregation structure around the x axis n times by angle $2\pi/n$ and add all the diffraction intensities. In the current study, we use n=20.

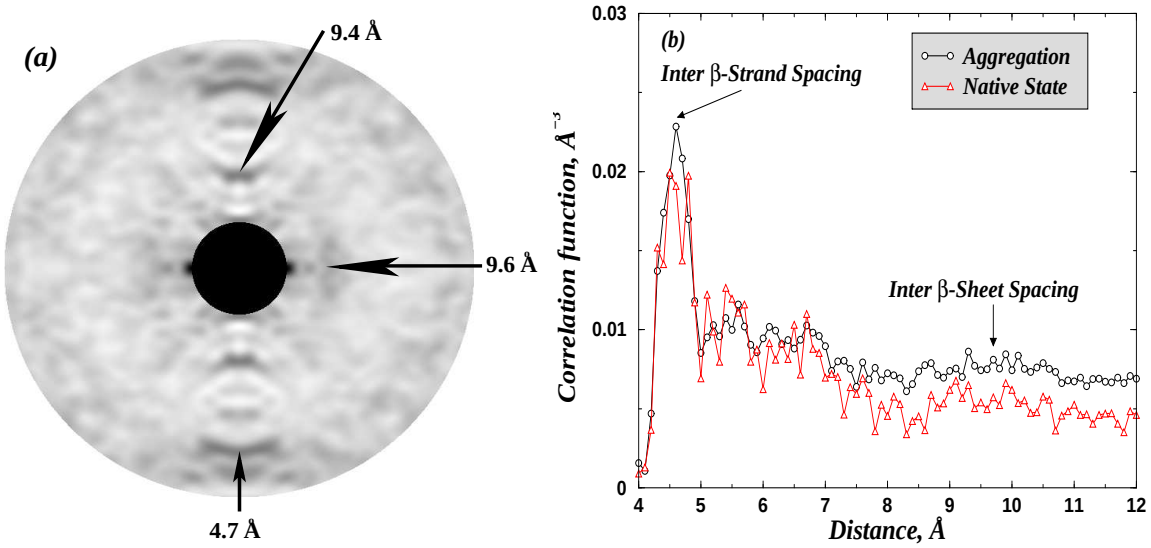


Figure 4.9: Characterization of Aggregates of Src SH3.
(a) Computed x-ray diffraction pattern of the aggregation structure formed by eight Src SH3 domains from the molecular dynamics simulations. (b) The correlation function of the native state of the Src SH3 domain (△) and the aggregate of eight Src SH3 domains (◯) from Fig. 4.8.

In order to characterize the structure of *in vivo* or *in vitro* amyloid fibrils, experimental x-ray scattering analysis has been widely applied [118–120]. The x-ray scattering patterns of different amyloid fibrils share the same features: (i) a relatively sharp and intense 4.7Å meridional reflection, and (ii) a weaker and more diffuse 7-10Å equatorial reflection. The first peak corresponds to the β-strands spacing along the direction of the fibril axis, and the second, much weaker peak, is understood as the spacing between β-sheets. In order to compare our aggregates to experiments,

we compute the x-ray diffraction pattern (Methods) for the aggregation structure obtained from molecular dynamics simulations (Fig. 4.9a). We observe a distinguished peak corresponding to 4.7Å along the meridional direction, which is related to β-strand spacing (Fig. 4.8b). We also find a peak along the same direction with the spacing, 9.4Å, which is due to the doubling of the β-strand spacing. In the equatorial direction, we observe a weak peak of 9.6Å which is related to the separation of the two β-sheets (Fig. 4.8b). Therefore, the structure of the aggregates derived from the molecular dynamics simulations fully agrees with the experimental observations.

Definition 4.2 *The pair correlation function g(r) function measures the average density of atoms in the shell of radius r surrounding an atom,*

$$g(r) = \frac{1}{N} \sum_i \sum_j < \delta(r - (R_i - R_j)) > \ . \tag{4.2}$$

The pair correlation function $g(r)$ is important for many reasons. It tells us about the structure of complex, isotropic systems, and it can be measured in neutron and X-ray diffraction experiments.

We calculate the correlation function of the structure of our aggregate (Fig. 4.9b) and compare it with that of the native structure. Because the native structure of Src SH3 contains five β-strands already, the correlation functions for the aggregation and native state has similar, although weaker, peak at inter β-strand spacing distance, 4.7Å. However, the aggregation state has a weak long range peak corresponding to the inter β-sheet spacing, $\sim$10Å, which is not present in the native state.

4.2.5 Discussion

We perform molecular dynamics simulations to study the aggregation of Src SH3 domain. We find that the proteins start to aggregate at the threshold temperature T_a. Simulations at high temperature usually produce amorphous aggregates. Near the folding transition temperature $T_f = 0.95$, protein conformations in the unfolded states are partially folded and there are two competing processes: folding and aggregation. When the kinetic barrier for aggregation is smaller than that for folding, we observe ordered aggregation. As temperature decreases, the folding kinetic barrier

decreases, so that τ_f decreases. Below some temperature threshold T_c, the partially folded protein states are near transition states at first and then rapidly descend into the native state, bypassing aggregation.

For two proteins, we observe two possible ordered aggregation conformations: a stable *closed* form dimer and an *open* aggregation state. The dimer is formed by domain swapping, where the two proteins exchange their RT-loops. We do not find any experimental evidence of the existence of this type of dimer for Src SH3 domain. However, our simulations suggest that it is a possible candidate oligomer state that may be found in future experiments. For the *open* form structure, the two proteins have their RT-loops packed together by swapping their two stands of the RT-loop. The *open* dimer structure allows further protein aggregation into fibrils. In our stimulations, we do not observe the aggregation scenario by propagational domain swapping [128, 129]. The closely packed RT-loops are stabilized by hydrogen bonds forming double β-sheets that face each other. From the molecular dynamics simulations we conclude that the core structure of Src SH3 amyloid fibrils is composed of RT-loop. Such a scenario may further be tested by experiments.

Our study reveals a general amyloidogenesis mechanism. Proteins, containing unstable β hairpins or loops, may be vulnerable to aggregation, and these unstable secondary structure elements can serve as the "building block" of amyloid fibrils [130, 137]. In the partially unfolded states, "building blocks" break apart from the rest a single protein. Then, these "building blocks" pack on top of one another by exchanging two complementary strands. During the aggregation process, amino acids in building blocks may reorient to make a stable connection, mainly by hydrogen bonding interactions. Due to the saturation and angular dependence of hydrogen bonds, only the exposed two ends can accept further deposition of building blocks from unbounded proteins and form an elongated double β-sheet structure. Thus, the resulting structure has the β-sheets parallel to the fibril axis and the β-strand perpendicular to the direction of fibril.

The presented model is similar in spirit to the β-nucleation model [138] to explain prion propagation. An important difference of our model from prion aggregation models is that in the case of prions aggregation, β structure is formed in the fragments

of the chain that are α-helical in the monomeric state of the protein, PrP^C, i.e. amyloid formation is accompanied by secondary structure transformations. However, in both cases the phenomenon of infectivity may be observed whereby existing fibrils may lower the kinetic barriers for monomeric proteins to join them forming larger fibrils.

Finally we note that our model may shed light on the observation that complex β-topologies are more often found in genomes than more simple meander-like ones [139, 140]. The former are unlikely to have contiguous fragments that are stable by themselves, while the latter, being simple topologies, do have such fragments. They are, therefore, more prone to aggregation/amyloid formation placing organisms that carry such proteins at evolutionary disadvantage.

4.3 Toward aggregation of α-helix-rich polypeptides: transition from α-helix to β-hairpin

A number of misfolded proteins and peptides aggregate into insoluble fibrils. The aggregation of some of these proteins into amyloid fibrils—amyloidosis—is related to fatal diseases [115, 116]. Recently, proteins not implicated in amyloid diseases have been found to form fibril structures *in vitro* under denaturing conditions [114, 123], suggesting that the fibril formation is a common feature of destabilized proteins [124]. Regardless of sequences and structures of proteins, the fibrils have similar core structures, mainly composed of β-sheets [117–119]. For example, the native Aβ peptide in Alzheimer's diseases [141,142] and prion proteins (PrP^C) in prion diseases [143,144] are α-helix-rich. The aggregation from α-rich proteins or peptides involves a conformational transition from α-helices to β-sheets. A similar transition has also been observed *in vitro* in some α-helical peptides [145–149] that aggregate into amyloid fibrils by means of changing the environment, such as varying the organic solvent condition [147], altering the pH [148], and controlling the redox state [149]. Moreover, similar α-β transitions also occur through the correct folding pathway in proteins with a nonhierarchical folding mechanism [150]. For example, β-Lactoglobulin, a

predominantly β-sheet protein [150], is observed to form non-native α-helical intermediates upon folding. Thus, understanding the α-β transition is important for both protein folding and protein aggregation.

The secondary structures of proteins, mainly α-helices and β-sheets, are determined by the amino acid sequence and stabilized by hydrogen bonds. However, under denaturing conditions, proteins with various organization of the secondary structure elements can aggregate into similar β-rich amyloid fibrils. We propose that the α-helix to β-hairpin transition is governed by sequence non-specific properties of proteins and peptides, i.e. the hydrogen bond network formed between backbones. It has been suggested that sequence non-specific hydrogen bond interaction among the backbones of proteins is an important factor for aggregation [14]. We hypothesize that the same type of interaction is also the driving force for the α-helix to β-sheet transition.

An overwhelming amount of computational simulations [151–159], and experimental [160–164] and theoretical [165–168] studies have been devoted to α-helix stability and the helix-coil transition. However, the possibility of β-hairpin formation in the folding pathway of peptides/proteins has not been addressed. The self-assembly of β-sheets by polyalanine segments, which usually form α-helices [152, 158], has been observed in silk-like multiblock copolymers [169]. Thus, we aim to identify the presence of metastable β-hairpin intermediate in the folding pathway of a simple polyalanine peptide, an α-helix in its native state [152, 158]. Due to the limitations of traditional molecular dynamics simulations, simplified models become crucial in studying protein folding and aggregation [8–11, 14, 158, 159, 170]. Discrete molecular dynamics [8, 9, 158], the combination of simple models and efficient dynamic simulation algorithms, can access the physical processes in the scale of milliseconds with a single simulation [171]. In contrast, the traditional all-atom molecular dynamics simulations can only resolve the time scale of several nanoseconds in one run or reach several microseconds combining a large number of runs [151]. In order to observe multiple transitions in a single simulation, we therefore employ the discrete molecular dynamics algorithm to study a polyalanine peptide.

4.3.1 Four-bead Model

The protein model, using three backbone beads and one side-chain bead to represent each residue [158, 170], has been developed to mimic protein backbone structure. Molecular dynamics studies in such a polypeptide system have shown a sharp helix-coil transition [158], which suggests that it is possible to study the transition from an α-helix to β-sheet in this model system. Thus, we use the four-bead model [158, 159, 170] to represent amino acids in the peptide. The amino acids are numbered from $k = 1$ (N-terminal) to $k = N$ (C-terminal), where N is the total number of residues. The kth amino acid is composed of nitrogen (N_k), prime carbon (C_k), alpha carbon ($C_{\alpha k}$), and beta carbon ($C_{\beta k}$) atoms (Fig. 4.10a). In Fig. 4.10a, the thick lines represent the covalent bonds and the thin lines denote effective bonds, mimicking the tetrahedral constraint of each amino acid and the planar constraint of the peptide bond. In our simulations, bonds are characterized by $\tilde{r}^{\rm AB}_{min} = D^{\rm AB}(1-\sigma)$ and $\tilde{r}^{\rm AB}_{max} = D^{\rm AB}(1+\sigma)$, where $D^{\rm AB}$ is the average distance between atoms A and B (listed in Table 4.1) and σ is chosen as 0.02.

Table 4.1: The parameters of bonds and hardcore radii used in our simulations.

Covalent bond,	$D^{\rm AB}$ (Å)	Effective bond,	$D^{\rm AB}$ (Å)	hardcore radius,	R (Å)
N_i, $C_{\alpha i}$	1.455	N_i, $C_{\beta i}$	2.442	C	1.50
$C_{\alpha i}$, $C_{\beta i}$	1.533	N_i, C_i	2.444	N	1.30
$C_{\alpha i}$, C_i	1.510	$C_{\beta i}$, C_i	2.494	C_α	1.85
C_i, N_{i+1}	1.325	$C_{\alpha i}$, N_{i+1}	2.406	C_β	2.20
		$C_{\alpha i}$, $C_{\alpha i+1}$	3.784		
		C_i, $C_{\alpha i+1}$	2.432		

First, we study the peptide with only backbone hydrogen bond interaction. The non-bonded atom pairs have either hardcore collision or hydrogen bond interactions. The hardcore radius R_A of four different types of atoms are listed in Table 4.1 ($r^{\rm AB}_{min} = R_A + R_B$). The protein backbone hydrogen bonds are formed between the

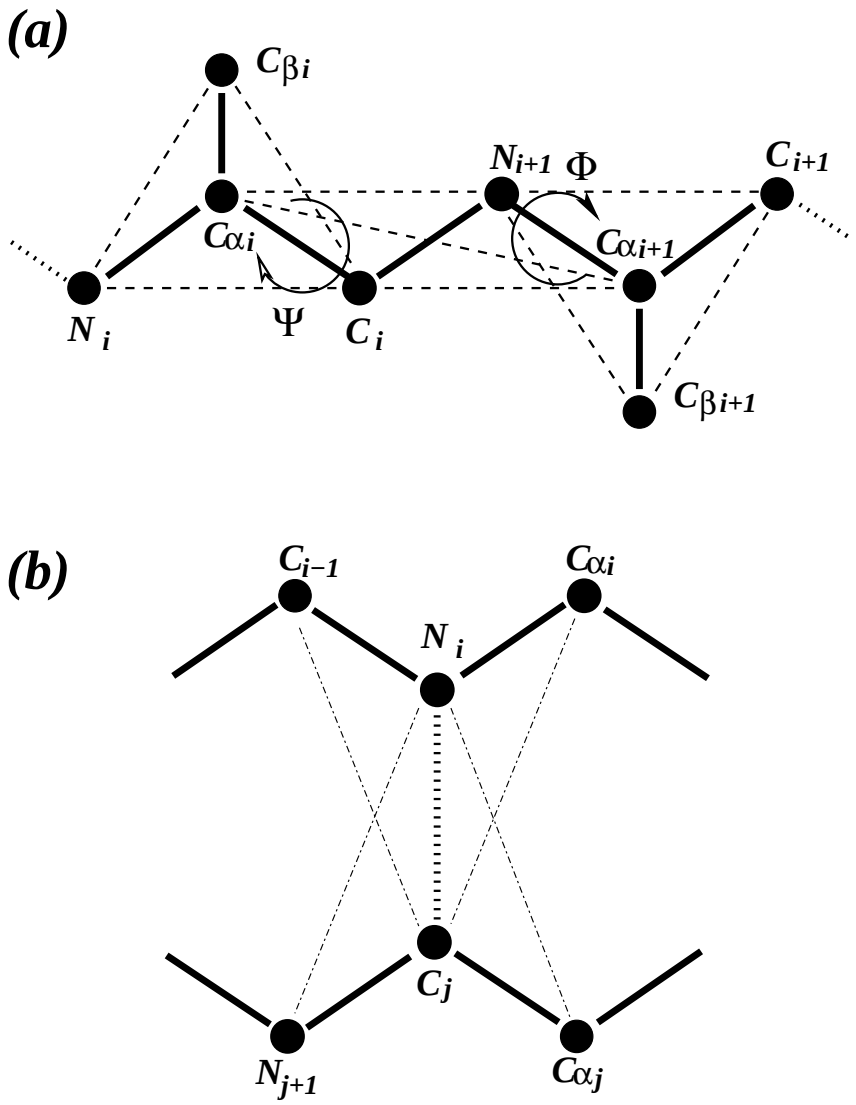


Figure 4.10: Schematic diagram of four-bead model

(a) A schematic diagram of the four-bead peptide model. The solid thick lines represent the covalent and the peptide bonds. The dashed thin lines denote the effective bonds which are assigned to mimic the tetrahedral constraint of each amino acid and the planar constraint of the peptide bond. (b) The schematic diagram of the hydrogen bond. The four thin dot-dashed lines connect the auxiliary pairs and the dashed line represents the hydrogen bond.

carbonyl oxygen and amide hydrogen. In the four-bead model, there is no carbonyl oxygen and amide hydrogen, but, the position of O and H can be determined by their

neighboring N, C, and C_α atoms. We model the hydrogen bond formed between the nitrogen N_i of the ith amino acid and the prime carbon C_j of the jth amino acid. Following the same notation as applied in Ref. [158, 159, 170], the separation along the sequence must satisfy the condition $|i-j| \geq 4$ to form a backbone hydrogen bond between N_i and C_j. It is well known that the hydrogen bond interaction has strong angular dependence i.e., the hydrogen bonded CO and NH groups are collinear with each other. The usual pairwise interaction can hardly model this multi-body interaction.

4.3.2 Hydrogen Bond Interaction: Reaction Algorithm

We introduce the reaction algorithm to model the hydrogen bond interaction between N_i and C_j. Once N_i and C_j form a hydrogen bond, they change their type into N_i' and C_j' respectively and cannot form any other hydrogen bonds. Whether the "reaction" $N_i + C_j \rightleftharpoons N_i' + C_j'$ takes place or not is assessed when the distance between these atoms becomes equal to the hydrogen bond cutoff distance $D_{\mathrm{HB}} = 4.2\text{Å}$. The total potential energy change includes the potential energy gain ϵ_{HB} between N_i' and C_j', and the potential energy changes between the two atoms and their surrounding atoms due to the type changes. Once the kinetic energy is enough to overcome the total potential energy change, the forward reaction happens. Otherwise, the two atoms N_i and C_j do not change their types and undergo original hardcore collision. If the reaction is successful, the atoms change their atom types and interact with other atoms according to the interaction parameters related to their new types.

We implement the angular dependence of hydrogen bonds by assigning an auxiliary interaction between the atom pairs $N_i' - C_{\alpha j}$, $N_i' - N_{j+1}^{[\prime]}$, $C_j' - C_{\alpha i}$, and $C_j' - C_{i-1}^{[\prime]}$ (these four pairs are connected by thin lines in Fig. 4.10b; the bracket in the superscript indicates that the atom may or may not have its type changed due to hydrogen

bond formation) as

$$V = \begin{cases} \epsilon_{HB}, & d_{min} < d < d_0 \\ \epsilon_{HB}/2, & d_0 < d < d_1 \\ 0, & d_1 < d < d_{max} \\ +\infty, & \text{otherwise} \end{cases}, \tag{4.3}$$

where ϵ_{HB} is the potential energy gain between N_i and C_j, and the parameters d_0, d_1, d_{min}, and d_{max} are chosen to implement the hydrogen bond angular constraints (see Table 4.2). The other interactions involving the N_i and C_j atoms remain unchanged before and after the reaction. The new hardcore collision distance between N_i' and C_j' is assigned at 4.0Å. Thus, at the lowest energy state of a hydrogen bond, the distance of the four auxiliary pairs is within the distance range of $[d_1, d_{max}]$ and distance of N_i and C_j is within the hydrogen bond range [4.0Å, 4.2Å]: the CO and NH groups are aligned as approximately linear. Parameters d_{min} and d_0 are chosen to allow angular distortion with energy penalizations.

Table 4.2: The parameters of the auxiliary interactions

Pairs	d_{min}(Å)	d_0(Å)	d_1(Å)	d_{max}(Å)
N_i', $C_{\alpha j}$	4.46	4.66	4.82	5.56
N_i', $N_{j+1}^{[\prime]}$	4.47	4.62	4.78	5.41
C_j', $C_{\alpha i}$	4.40	4.56	4.72	5.39
C_j', $C_{i-1}^{[\prime]}$	4.44	4.62	4.79	5.39

When two atoms N_i and C_j approach each other at the hydrogen bond interaction cutoff distance $D_{\text{HB}} = 4.2$Å, we evaluate the total potential energy change by checking the four auxiliary interactions. The potential energy change can be $-\epsilon_{HB}$,$-\epsilon_{HB}/2$, 0, $\epsilon_{HB}/2$, ... $3\epsilon_{HB}$ and ∞, depending on the orientation of the N_i, C_j, and their neighbors. The larger the angular distortion, the higher the potential energy change. Once formed, the four auxiliary pairs will have a high probability of staying in the range of $[d_1, d_{max}]$ with the lowest energy, and thus the orientation of the hydrogen bond is *maintained*. The thermal fluctuations distort the orientation of

the hydrogen bond and large fluctuations may break the hydrogen bond. Once the two atoms N_i' and C_j' come again to the exact distance of D_{HB}, a reverse reaction may happen. We check the potential energy change due to the possible changes of types. The total potential energy change ranges between $-3\epsilon_{HB}$ to ϵ_{HB}, corresponding to different conformations of the hydrogen bond due to thermal fluctuations. Thus, a distorted hydrogen bond will be easier to break. A more realistic modeling of the angular dependence of the hydrogen bond is to increase the number of steps in Eq. 4.3 to make the interaction potential more continuous. However, the increase of the number of steps decreases the efficiency of the discrete molecular dynamics.

4.3.3 Polyalanine with hydrogen bond interaction only

We study the refolding thermodynamics of the 16-residue polyalanine with only backbone hydrogen bond interactions. We perform discrete molecular dynamics simulations at different temperatures T = 0.09, 0.10, 0.11, 0.12, 0.125, 0.13, 0.14, and 0.15 in units of ϵ_{HB}/k_B [1]. For each temperature, we perform ten separate molecular dynamics simulations starting from different random coil conformations (Fig. 4.11).

At high temperatures, the polyalanine remains at the random coil state and its average potential energy is close to zero. As we decrease the temperature, the peptide adopts a β-hairpin state (Fig. 4.11d). For a β-hairpin structure, the lowest potential energy conformations have the β-turn located near the center of the peptide (Fig. 4.12b), and the potential energy is equal to $-6\epsilon_{HB}$. If the turn is positioned differently along the peptide (Fig. 4.12c), the potential energy is higher than $-6\epsilon_{HB}$ due to the smaller number of hydrogen bonds that can be formed. Thus, the occurrence of additional β-hairpin types (Fig. 4.12c) is less probable. At temperature T = 0.13 (Fig. 4.11e), we observe a reversible random coil to β-hairpin transition and the probabilities of finding a random coil and a β-hairpin state are approximately equal (Fig. 4.11a,e), so the β-hairpin to random coil transition temperature is $T_{\beta-coil} \approx 0.13$. At temperature T = 0.13, we also detect rare fluctuations with potential energy lower than $-6\epsilon_{HB}$, corresponding to partially formed α-helix states.

[1] e.g. for ϵ_{HB} = 5kcal/mol [172], the temperature T=0.12 corresponds to $302K$.

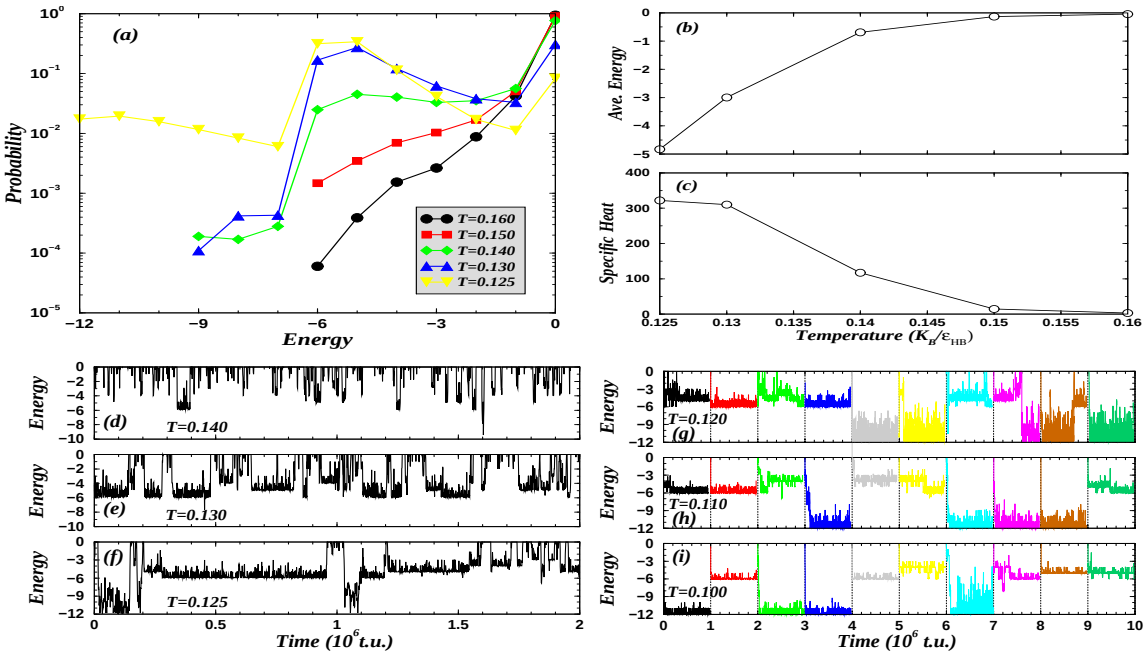


Figure 4.11: Thermodynamics of the polyalanine chain with backbone hydrogen bond interaction only.

(a) The probability distribution of potential energies, (b) the average energy, and (c) the specific heat at the temperatures where the peptide is not dynamically trapped. The typical energy trajectory at temperatures (d) 0.140, (e) 0.130 and (f) 0.125. At lower temperatures (g) 0.120, (h) 0.110, and (i) 0.100, the protein is easily trapped and each simulation results in either α-helix or β-hairpin states. For each temperature, the ten different potential energy trajectories of 10^6 time unites (t.u.) are separated by dashed lines and colored differently. At temperature $T = 0.120$, we observe transitions from the β-hairpin to the α-helix states (the $8th$ run) and from the α-helix to the β-hairpin states (the $9th$ run).

As we lower the temperature to $T = 0.125$, we observe the occurrence of α-helical states (Fig. 4.11g and Fig. 4.12a). For the 16-residue polyalanine, the complete α-helix has four helix turns and the lowest energy is $-12\epsilon_{HB}$. At $T = 0.125$, the peptide can either adopt a random coil, an α-helix, or a β-hairpin state. The interconversion between an α-helix and a β-hairpin only takes place if the peptide first unfolds to a random coil state. This is mainly due to the drastic structural difference between

these two kinds of conformations and there is no direct pathway between them except via a random coil state. Thus, the α-helix to β-hairpin transition is coupled to the transition between the α-helix to random coil transition. The probability of an α-helix is smaller than that of a β-hairpin at temperature $T = 0.125$. We expect to observe more α-helix states at lower temperatures. However, at low temperatures the dynamics of hydrogen bond formation and disruption become slow and the polyalanine is easily trapped in the local minima of the free energy landscape, dominated by a metastable β-hairpin state. We find that the peptide remains in either an α-helix or a β-hairpin state during the simulation time of 10^6 time units [9] after a quick collapse from the random coil state (Fig. 4.11g,h,i).

We present the distribution of the torsion angles ϕ and ψ for different states in Fig. 4.12e,f,g. The distributions are in agreement with the Ramachandran plot [173] for secondary structures. We find that the distributions of the β-hairpin and random coil are similar. However, for the random coil state, the torsion angles for each amino acid are fully uncorrelated, while for the β-hairpin state, the torsion angles between the hydrogen bonded amino acids are highly correlated. For an α-helix, each residue forms two hydrogen bonds except those near the termini (Fig. 4.13a), thus, our peptide does not have excessive torsional freedom. On the other hand, for a β-hairpin strand approximately half of all amino acids do not form any hydrogen bonds (Fig. 4.13b), the peptide chain has a larger value of backbone entropy. Therefore, the β-hairpin has larger hydrogen bond energy and also a larger entropy than the α-helix. In order to illustrate the backbone flexibility for different states (α-helix, β-hairpin, random coil), we align various conformations with respect to a characteristic structure using C_α atoms for each of these states (Fig. 4.12h,i,j). We find that the β-hairpin (Fig. 4.12i) is more flexible than the α-helix (Fig. 4.12h) and, therefore, the β-hairpin has a higher backbone entropy [2]. Interestingly, the alignment for the random coil state exhibits a persistent overall topology (Fig. 4.12j), which is possibly due to the excluded volume effect of the residues and is consistent with the finding in Ref. [174]. The interplay between energy and entropy allows for the existence of

[2]We provide movies for the dynamic motions of an α-helix and a β-hairpin, and one instance of the α-helix to β-hairpin transition: http://www.unc.edu/~dokh/research/AB/home.html

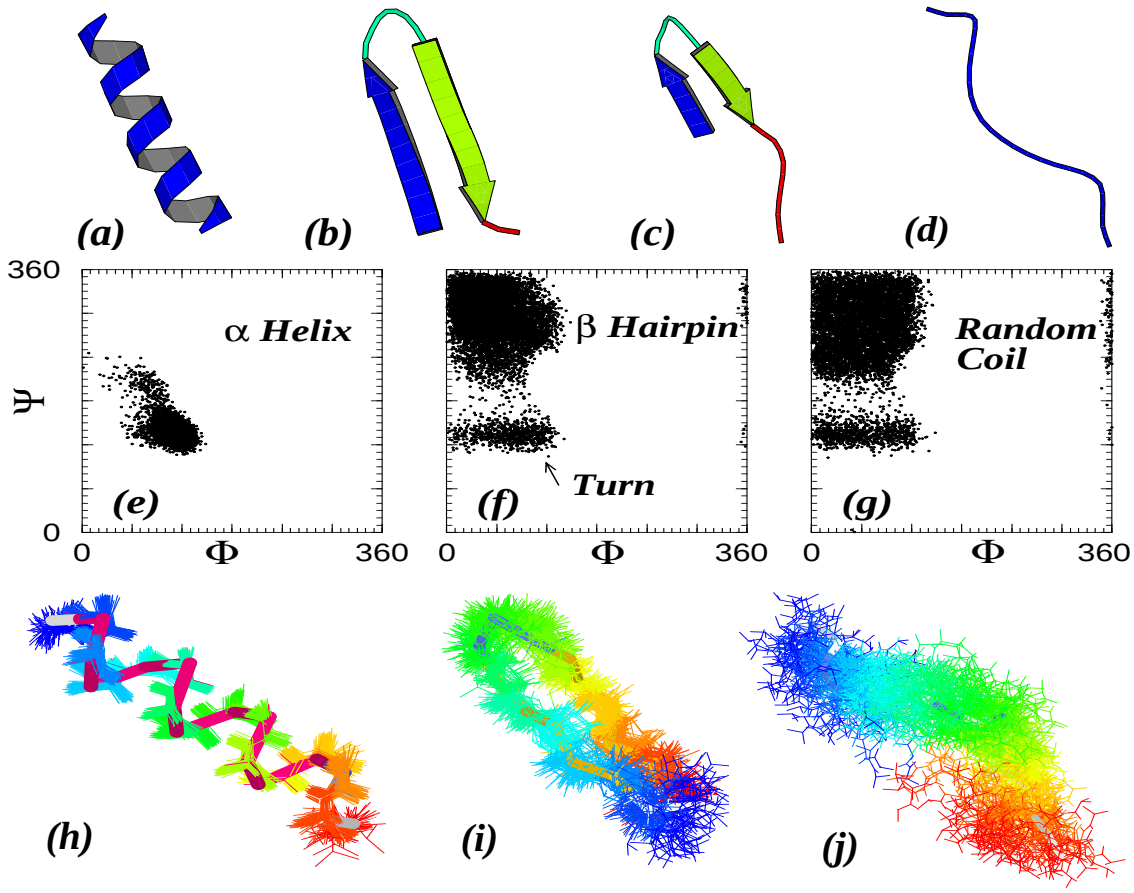


Figure 4.12: Characterization of different secondary structures.
Typical conformations of (a) an α-helix, (b) the β-hairpin with the β-turn located near the center, (c) an additional β-hairpin conformation with the β-turn positioned differently, and (d) a random coil. The distributions of torsion angles for (e) the α-helix, (f) the β-hairpin, and (g) the random coil states over equilibrated simulations. We align for each of the three states — (h) α-helix, (i) β-hairpin, and (j) random coil — various conformations with respect to a reference conformation using C_α atoms . The reference conformations are shown in backbone representation and the other conformations are displayed as wire-frames with the residues colored in the rainbow order from blue (N-terminal) to red (C-terminal).

the metastable intermediate state — the β-hairpin.

Due to the slow dynamics at low temperatures, we cannot accurately identify the

α-helix to β-hairpin transition temperature $T_{\alpha-\beta}$. The coexistence of these two states with the random coil state at $T = 0.125$ suggests that the transition temperature $T_{\alpha-\beta}$ is close to 0.125, which is in the vicinity of the β-hairpin to coil transition temperature $T_{\beta-coil} = 0.130$. Each of the four amino acids near the N- or C- termini of an α-helix has one free hydrogen bond donor or acceptor (Fig. 4.13a), similar to the β-hairpin strand where each amino acid on average has one free hydrogen bond donor or acceptor (Fig. 4.13b). Thus, the potential energy per residue of the α-helix terminal is equal to that of a β-hairpin strand. According to the similar constraints imposed by the hydrogen bonds, we hypothesize that the conformational entropy per residue for the β-hairpin strands and α-helix termini is also similar. Thus, the α-helix termini and β-hairpin strands have similar free energy per residue. Furthermore, the amino acids near the termini have larger free energies than those of the amino acids within the helix, which is consistent with the observation of large fluctuations of the termini even at low temperature. We observe that the polyalanine melting (i.e. the transition into random coil) always starts from the termini. The melting temperature of an α-helix $T_{\alpha-coil}$ is determined by the free energy per residue between the α-helix termini and the random coil, and the transition temperature from a β-hairpin to random coil $T_{\beta-coil}$ is determined by the free energy per residue between β-hairpin strands and the random coil. So, $T_{\alpha-coil}$ is close to $T_{\beta-coil}$. Since the α-helix to β-hairpin transition is coupled with the α-helix to coil transition (α-helix melting), $T_{\alpha-\beta} \approx T_{\beta-coil}$.

To further understand the contribution of backbone entropy to the α-helix to β-hairpin transition, we estimate the backbone entropy for different states. From the alignment of conformations in Fig. 4.12h,i,j, we find that all conformations fluctuate around the reference structure characteristic to the corresponding state. Assuming that (a) the fluctuations of the structures around the reference structure are Gaussian; and (b) the fluctuations of residues are uncorrelated, the conformational entropy can be approximated as S_x=$3N \ln\langle rmsd_r\rangle_x + S_0$, where $rmsd_r$ is the root mean square deviation from the reference structure. The average $\langle\rangle_x$ is taken over conformations out of the corresponding state $\{x\}$: α-helix, β-hairpin, and random coil. S_0 is a constant which can be determined by setting the α-helix as the reference

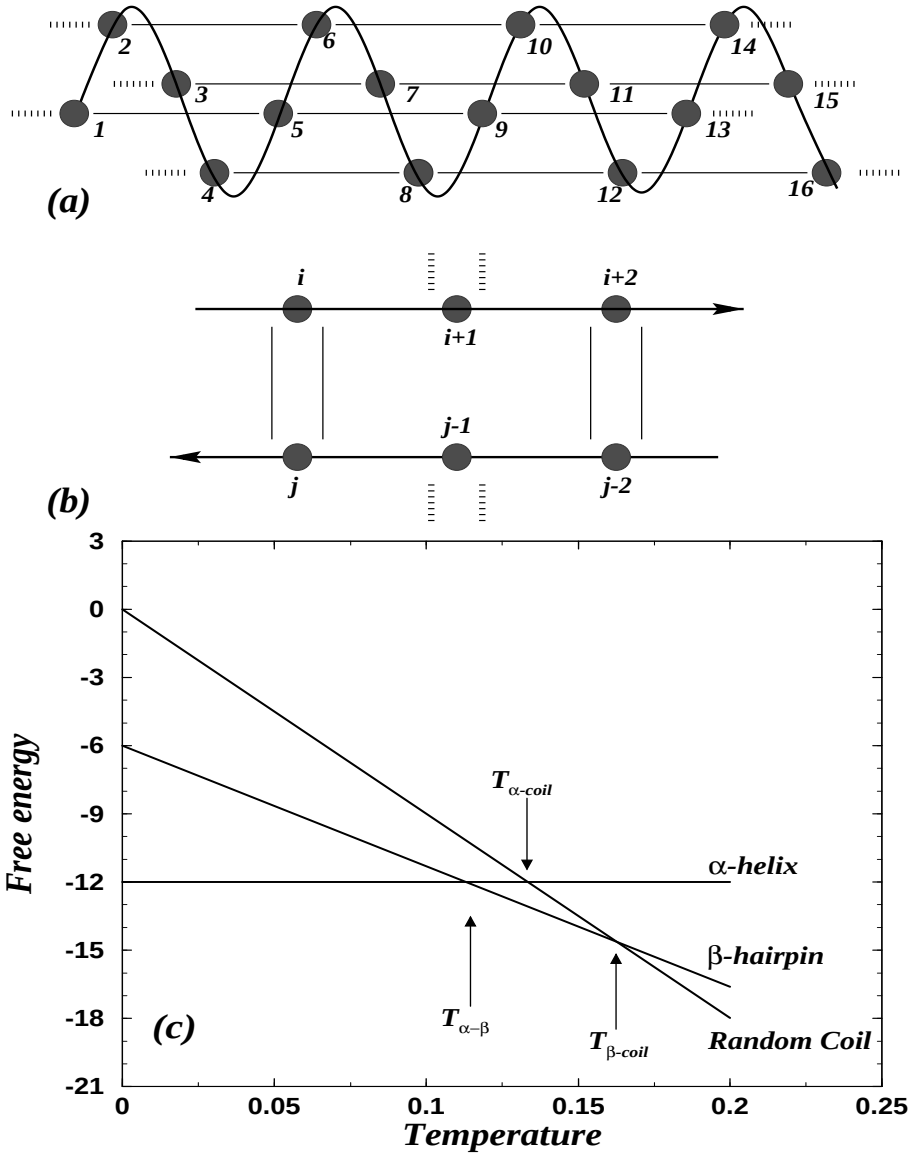


Figure 4.13: Free energy dependence of different secondary structures. The schematic diagram of the hydrogen bond pattern for (a) the α-helix and (b) the β-hairpin conformations, where the black beads represent each amino acid, the solid lines denote the backbone hydrogen bonds, and the dotted lines denote the free backbone hydrogen bond donor or acceptor. (c) The free energies for different states versus temperature using the estimated values of the backbone entropies.

state with S_α=0. Since the four atoms in each residue are constrained to fluctuate as one object, N represents the number of amino acids. To calculate the entropy we perform equilibrium simulations with hydrogen bonds intact (the α-helix and β-hairpin states) or without forming any hydrogen bonds (the random coil state). We calculate $\langle rmsd_r \rangle_x$ with respect to a selected conformation typical to the corresponding state. The values of estimated entropy for different states are listed in Table 4.3. The transition temperatures can be determined as $T_{\beta-coil} = 0.162$, $T_{\alpha-coil} = 0.133$, and $T_{\alpha-\beta} = 0.115$. These estimated transition temperatures agree with the values determined from simulations. The above assumptions might lead to underestimation of the backbone entropy for the random coil state, therefore, the estimated transition temperatures $T_{\beta-coil}$ and $T_{\alpha-coil}$ are higher than the determined values from simulations.

Table 4.3: The estimated values of the conformational entropy for different states. $S_0 = -3N \ln\langle rmsd_r \rangle_\alpha$=34.9 (N=16). The transition temperature between two different states can be obtained from the differences of potential energy and entropy, $\Delta E/\Delta S$ (see Fig. 4.13c).

x	$\langle rmsd_r \rangle_x$(Å)	Entropy, $S_x(k_B)$	potential energy, $E(\epsilon_{HB})$
α-helix (reference state)	0.483	0	-12
β-hairpin	1.460	53.0	-6
random coil	3.143	89.9	0

The existence of the metastable β-hairpin state has important implications for aggregation. A β-hairpin conformation that has the exposed hydrogen bond donors or acceptors is capable of further aggregation and can form amyloid fibrils [175]. Most real proteins do not aggregate by folding into the native state without long lifetime intermediates. Proper folding may be enforced by side-chain interactions in the evolutionarily selected sequence. We study a minimal model with hydrophobic side-chain interactions to uncover the propensities of the α-helix to β-hairpin transition for a hydrophobic-polar (HP) sequence which is designed to be an α-helix in the

native state.

4.3.4 Model peptide with hydrophobic-polar sequence

We also study the effect of side-chain interactions in a 16-residue model peptide chain, designed to be an α-helix in its native state. The peptide has the following sequence of hydrophobic (H) and polar (P) residues: PPHPPHHPPHPPHHPP [159, 176]. This sequence is derived from a peptide which is designed to be an α-helix in the experiment [177]. In our simulations, the interaction between hydrophobic side-chains (C_β atoms) is modeled as an attractive square well with the cutoff distance $D_{\rm HP} = 6.5$Å and the interaction strength ϵ_{HP}; the remaining side-chain interactions are hardcore collisions. The relative strength of the hydrophobic interactions with respect to the hydrogen bond interactions $\rho = \epsilon_{HP}/\epsilon_{HB}$ is the free parameter that can be tuned.

We perform molecular dynamics simulations of the HP peptide with various interaction ratios $\rho = \epsilon_{HP}/\epsilon_{HB}$, from 0.05 to 0.50 with a step of 0.05. For each ratio, we perform simulations at various temperatures. We find that the β-hairpin state becomes less stable as we increase ρ. For small ρ, the thermodynamic property of our HP peptide resembles that of the peptide without specific side-chain interactions. At a characteristic range of ρ values ($0.20 \leq \rho \leq 0.35$), the intermediate β-hairpin state disappears and the peptide folds cooperatively into the native α-helix state (Fig. 4.14).

For $\rho = 0.25$, the specific heat has a pronounced peak around $T_F = 0.128$ (Fig. 4.14), indicating a sharp transition specific to a two-state protein. At low temperatures, the peptide adopts the native α-helix structure (Fig. 4.14d,e). At temperature $T = 0.120$, we also observe some potential energy fluctuations corresponding to the partially unfolded α-helix with the unfolded N- and C-termini. In the vicinity of the transition temperature, the peptide adopts both the α-helix and the unfolded states. The hydrophobic interactions are formed with a higher probability even at high temperatures than the hydrogen bonds because of the larger interaction range ($D_{HP} = 6.5$Å $> D_{HB} = 4.2$Å) and the absence of angular depen-

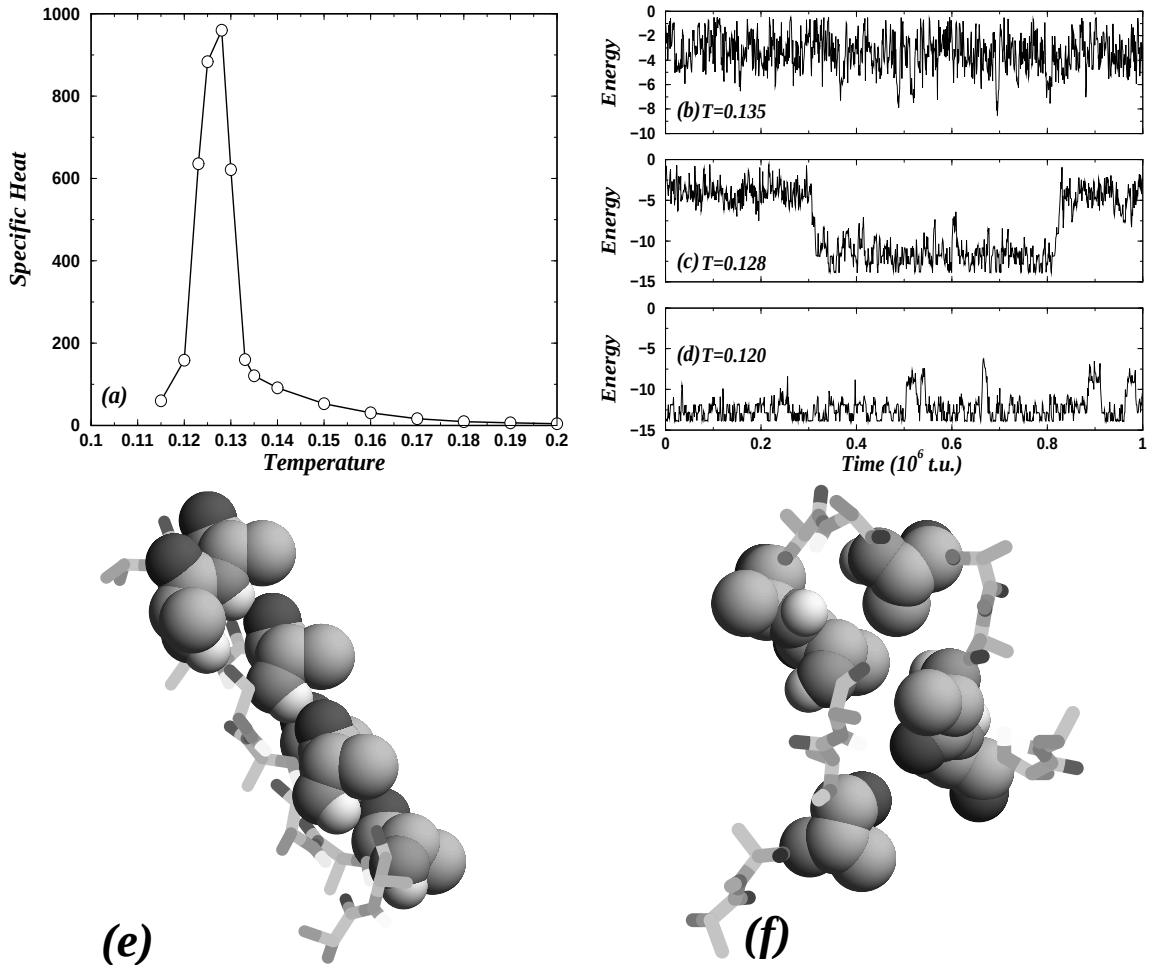


Figure 4.14: Thermodynamics of HP sequence.
(a) The specific heat for the hydrophobic polar peptide with $\rho = 0.25$, having a pronounced peak at $T_F = 0.128$. The typical potential energy trajectories at (b) $T = 0.135$, (c) $T = 0.128$, and (d) $T = 0.120$. The transition is between (e) the native state and (f) the molten globular states, where the space-filled amino acids are the hydrophobic atoms.

dence of the HP interactions. Snapshots of these unfolded states indicate that the HP peptide adopts "molten globular" [178] conformations (Fig. 4.14f) which have contacts formed between the hydrophobic residues. The average radius of gyration R_g for the molten globular state at temperature $T = 0.135$ is 7.35Å, compared to the unfolded state $R_g = 10.3$Å for $\rho = 0$ at the same temperature. We find that

the HP peptide first collapses into the "molten globular" state from the random coil state, which is similar to the coil-globular transition [178], at a higher temperature than T_F.

Within the characteristic range of ρ values, the difference of the average potential energy between the β-hairpin state and the unfolded state is small, which is not enough to stabilize the β-hairpin state. For example, for $\rho = 0.25$ (Fig. 4.14) the average potential energy of the unfolded state is $-4\epsilon_{HB}$ at T_F, and the potential energy of the β-hairpin state $\approx -6\epsilon_{HB}$ is within the range of potential energy fluctuation of the unfolded state (Fig. 4.14c). However, the potential energy gap between the α-helix state and the unfolded state is $8\epsilon_{HB}$ (Fig. 4.14c), which is large enough to stabilize the α-helix state [28].

As we increase ρ, we rarely observe the helix-coil transition (e.g. for $\rho = 0.50$ the α-helix state is never reached from the unfolded state during the simulation of 10^7 time units). At low temperatures, our HP peptide is frozen in the molten globular states because of the strong hydrophobic interactions.

4.3.5 Discussion

A missing link in understanding the amyloidogenesis of α-helix-rich proteins to β-sheet-rich fibrils is the possible presence of a metastable β-hairpin intermediate state, prone to aggregation [175]. Our results suggest a generic framework that explains why this β-hairpin intermediate is favorable in terms of free energy. Although the potential energy of the β-hairpin state is higher than that of the α-helix state, the entropy of a β-hairpin is significantly larger than that of an α-helix due to fewer constraints imposed by hydrogen bonds. At high temperatures, the free energy of a β-hairpin can be smaller than that of an α-helix. Even though our simulations and discussions are focused on the β-hairpin (an anti-parallel two-stranded β-sheet) our analysis is appropriate for both parallel and anti-parallel two-stranded β-sheets that are entropicaly favorable with respect to α-helices.

Our simulations of temperature-driven α-helix to β-hairpin transition are consistent with recent experiments on the solvent-driven conformational transitions [179,

180]. By changing the solvents from one type that has a low ability to interact with the backbone peptide groups to another type that has a higher ability [179], the designed peptides are found to convert from α-helices to β-hairpins. Increasing the ability of the solvent to interact with the backbone, the energy gain to form a backbone hydrogen bond is effectively decreased. Instead of increasing the temperature, the decrease of the hydrogen bond's energy gain drives the conformational transition from α-helix to β-hairpin because of the dominating effect of backbone entropy.

Most proteins in physiological conditions do not aggregate. Proteins with evolutionarily selected sequences avoid aggregation by folding into the native state without metastable intermediate states. *In vitro* and *in vivo* experiments show that changes in environmental conditions lead to aggregation [124, 148]. The environmental change has a different effect on different types of interactions. Our simulations of HP sequences with various hydrophobic interaction strengths demonstrate that if the environmental changes effectively lead to the weakening of the relative side-chain interactions, the peptide or protein may misfold into a metastable β-hairpin intermediate.

Discrete molecular dynamics simulation methodology is a step in simplification of molecular modeling with respect to traditional molecular dynamics simulations. The principal drawback of the discrete molecular dynamics simulations is its difficulty to represent forces. Instead, system's dynamics is realized through ballistic collisions between particles. Interactions between particles are modeled by square-well potentials. Despite its simplicity, discrete molecular dynamics has been proved to be a powerful tool not only to study protein folding thermodynamics [8–11] and kinetics [10–12], but to identify the evasive protein transition state ensembles [10] and to witness aggregation of multiple proteins into amyloid fibrils [14]. The latter two goals have yet to be directly approached with traditional molecular dynamics simulations. In addition, the traditional all-atom molecular dynamics simulations are also a simplification of the quantum mechanics simulations, in which quantum interactions are replaced by approximate Newtonian interactions. The latter, in turn, are approximated by a large number of empirical parameters. The advantage of the discrete molecular dynamics simulations versus traditional molecular dynamics sim-

ulations is its ability to resolve larger time scales — 10^6 orders of magnitude. The traditional molecular mechanics simulations have similar advantage over quantum mechanics simulations. The traditional molecular dynamics simulations are based on several decades of improving and testing of model force field, while applications of discrete molecular dynamics simulations have been limited until recently to colloids and hard spheres. Despite of this we believe that modifying and improving parameters of discrete molecular dynamics simulations for proteins by testing them on simple systems such as the polyalanine chain studied here will eventually lead to models with quantitative predictive power.

Bibliography

[1] C. Levinthal, Are there pathways for protein folding?, *J. Chem. Phys.* **65**, 44 (1968).

[2] Anifsen, CB, Principles that goven the folding of protein chains, *Science* 1973; **181**:223-230

[3] E. I. Shakhnovich, Theoretical studies of protein-folding thermodynamics and kinetics, *Curr. Opin. Struct. Biol.* **7**, 29–40 (1997).

[4] H. Taketomi, Y. Ueda and N. Gō, Studies on protein folding, unfolding and fluctuations by computer simulations, *Int. J. Peptide Protein Res.* **7**, 445 (1975).

[5] Gō, N. & Abe, H., Noninteracting local-structure model of folding and unfolding transition in globular proteins. I. Formulation, *Biopolymers* **20**, 991–1011 (1981)

[6] C. D. Snow, N. Nguyen, V. S. Pande, M. Gruebele, Absolute comparison of simulated and experimental protein-folding dynamics. *Nature* **420**, 102–106 (2002).

[7] B. Zagrovic, C. D. Snow, M. R. Shirts and V. S. Pande, Simulation of folding of a small α-helical protein in atomistic detail using worldwide-distributed computing. *J. Mol. Biol.* **323**, 927–937 (2002).

[8] Zhou, Y. & Karplus, M., Folding thermodynamics of a three-helix-bundle protein, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 14429-14432 (1997).

[9] Dokholyan, N.V., Buldyrev, S. V., Stanley, H. E. & Shakhnovich, E. I., Molecular dynamics studies of folding of a protein-like model, *Folding & Design* **3**, 577-587 (1998).

[10] Ding F, Dokholyan NV, Buldyrev SV, Stanley HE, Shakhnovich EI. Direct molecular dynamics observation of protein folding transition state ensemble. *Biophys J* 2002;**83**:3525-3532.

[11] Borreguero JM, Dokholyan NV, Buldyrev SV, Stanley HE, Shakhnovich EI, Thermodynamic and folding kinetic analysis of the SH3 domain from discrete molecular dynamics. *J Mol Biol* 2002;**318**:863-876.

[12] Dokholyan, N.V.,Buldyrev, S. V., Stanley, H. E. & Shakhnovich, E. I., Identifying the protein folding nucleus using molecular dynamics, *J. Mol. Biol.* **296**, 1183-1188 (2000).

[13] S. D. Khare, F. Ding and N. V. Dokholyan, Folding of Cu,Zn superoxide dismutase and Familial Amyotrophic Lateral Sclerosis. *J. Mol. Biol*, **334**, 515–525 (2003).

[14] Ding F, Dokholyan NV, Buldyrev SV, Stanley HE, Shakhnovich EI. Molecular dynamics simulation of the SH3 domain aggregation suggests a generic amyloidogenesis mechanism. *J Mol Biol* 2002;**324**:851-857.

[15] Rapaport DC. *The art of molecular dynamics simulation*. Cambridge University Press 1997.

[16] Berendsen, H.J.C., Postma, J., van Gunsteren, W., DiNola, A., Haak, J., Molecular-Dynamics with coupling to an external bath *J. Chem. Phys* 1984, 81, 3684–3690.

[17] Gō, N., *Ann. Rev. Biophys. Bioeng.* **12**, 183-210 (1983)

[18] K. A. Dill, Theory for the folding and stability of globular proteins, *Biochemistry* **24**, 1501–1509 (1985).

[19] Chan, H. S. & Dill, K. A., Protein folding in the landscape perspective: Chevron plots and non-Arrhenius kinetics, *Proteins: Struc. Func. Genet.* **30**, 2-33 (1998)

[20] J. D. Bryngelson and P. G. Wolynes, Intermediates and barrier crossing in a random energy model (with applications to protein folding), *J. Phys. Chem.* **93**, 6902–6915 (1989).

[21] T. Creighton, *Proteins: structures and molecular properties, second edition* (W. H. Freeman and Co., New York, 1993).

[22] M. Karplus and E. I. Shakhnovich, Protein folding: theoretical studies of thermodynamics and dynamics, in *Protein Folding*, edited by T. Creighton (W. H. Freeman and Co., New York, 1994).

[23] O. B. Ptitsyn, The molten globule state, in *Protein Folding*, edited by T. Creighton (W. H. Freeman and Co., New York, 1994).

[24] Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I., Specific nucleus as the transition state for protein folding: evidence from the lattice model, *Biochemistry* **33**, 10026–10036 (1994)

[25] E. I. Shakhnovich, V. I. Abkevich and O. Ptitsyn, Conserved residues and the mechanism of protein folding, *Nature* **379**, 96–98 (1996).

[26] D. K. Klimov and D. Thirumalai, Criterion that determines the foldability of proteins, *Phys. Rev. Lett.* **76**, 4070–4073 (1996).

[27] Fersht, A. R., Nucleation mechanisms in protein folding, *Curr. Opinion Struc. Biol.* **7**, 3–9 (1997)

[28] Shakhnovich EI. Protein design: a perspective from simple tractable models. *Folding & Design* 1998;**3**:R45-R58.

[29] C. Micheletti, J. R. Banavar, A. Maritan and F. Seno, Protein structures and optimal folding emerging from a geometrical variational principle, *Phys. Rev. Lett.* **82**, 3372–3375 (1998).

[30] Du, R., Pande, V. S., Grosberg, A. Yu., Tanaka, T. & Shakhnovich, E. I., On the transition coordinate for protein folding, *J. Chem. Phys.* **108**, 334–350 (1998)

[31] Grantcharova, V. P., Riddle, D. S., Santiago, J. V. & Baker, D., Important role of hydrogen bonds in the structurally polarized transition state for folding of the src SH3 domain, *Nature Struct. Biol.* **5**, 714–720 (1998)

[32] Martinez, J. C., Pissabarro, M. T. & Serrano, L., Obligatory steps in protein folding and the conformational diversity of the transition state, *Nature Struct. Biol.* **5**, 721–729 (1998)

[33] N. V. Dokholyan, L. A. Mirny, and E. I. Shakhnovich, Understanding conserved amino acids in proteins, *Phys. Rev. Lett.* submitted (2000); preprint **cond-mat/0007084**.

[34] Dinner, A. R. & Karplus, M., The thermodynamics and kinetics of protein folding: A lattice model analysis of multiple pathways with intermediates, *J. Chem. Phys.* **37**, 7976-7994 (1999)

[35] Bursulaya, B. D. & Brooks, C. L., Folding free energy surface of a three-stranded beta-sheet protein, J. Am. Chem. Soc. **121**, 9947-9951 (1999)

[36] L. A. Mirny and E. I. Shakhnovich, Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function, *J. Mol. Biol.* **291**, 177–196 (1999).

[37] N. V. Dokholyan and E. I. Shakhnovich, A model of hierarchical protein evolution, *J. Mol. Biol.* **312**, 289–307 (2001).

[38] Onuchic, J. N., Luthey-Schulten, Z. & Wolynes, P. G., Theory of protein folding: The energy landscape perspective, *Ann. Rev. Phys. Chem.* **48**, 545600 (1997)

[39] Onuchic, J. N., Nymeyer, H., Garcia, A. E., Chahine, J. & Socci, N. D., Insights into folding mechanisms and scenarios, *Adv. Prot. Chem.* **53**, 87152 (2000)

[40] Bryngelson, J. D., Onuchic, J. N., Socci, N. D. & Wolynes, P. G., Funnels, pathways, and the energy landscape of protein folding - A synthesis, *Proteins: Struc. Func. Genet.* **21**, 167195 (1995)

[41] Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I., Improved design of stable and fast-folding model proteins, *Folding & Design* **1**, 221230 (1996)

[42] S. E. Jackson, and A. R. Fersht, Folding of chymotrypsin inhibitor-2.1. Evidence for a two-state transition, *Biochemistry* **30**, 10428–10435 (1991).

[43] A. R. Viguera, J. C. Martinez, V. V. Filimonov, P. L. Mateo, and L. and Serrano, Thermodynamic and kinetic-analysis of the SH3 domain of Spectrin shows a 2-state folding transition, *Biochemistry* **33**, 10925–10933 (1994).

[44] V. S. Pande, A. Yu. Grosberg, D. S. Rokshar and T. Tanaka, Pathways for protein folding: is a new view needed?, *Curr. Opin. Struct. Biol.* **8**, 68–79 (1998).

[45] Galzitskaya, O. V. & Finkelstein, A. V., A theoretical search for folding/unfolding nuclei in three-dimensional protein structures, *Proc. Natl. Acad. Sci. USA* **96**, 11299-11304 (1999)

[46] J. P. K. Doye and D. J. Wales, "On potential energy surfaces and relaxation to the global minimum", *J Chem, Phys.* **105**, 8428-8445, 1996

[47] O. B. Ptitsyn, Stage mechanism of the self-organization of protein molecules, *Dokl. Acad. Nauk.*, **210**, 1213-1215, 1973

[48] P. S. Kim and R. L. Baldwin, Intermediates in the folding reactions of small proteins, *Annu. Rev. Biochem.*, **59**, 631-660 (1994).

[49] D. B. Wetlaufer, Nucleation, rapid folding, and globular intrachain regions in proteins, *Proc. Natl. Acad. Sci. USA*, **70**, 697-701, 1973

[50] D. B. Wetlaufer, Nucleation in protein folding - confusion of structure and process, *Trends Biochem. Sci.*, **15**, 414-415, (1990)

[51] O. Ptitsyn, How molten is the molten globule?, *Nature Struct. Biol.*, **3**, 488-490, 1996

[52] Alm, E. & Baker, D., Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures, *Proc. Natl. Acad. Sci. U. S. A.* **96**, 11305–11310 (1999)

[53] Matouschek, A., Kellis Jr., J. T., Serrano, L. & Fersht A. R., Mapping the transition state and pathway of protein by protein engineering, *Nature* **340**, 122-126 (1989)

[54] Matouschek, A., Kellis Jr, J. T., Serrano, L., Bycroft, M. & Fersht, A. R., Transient folding intermediates characterized by protein engineering, *Nature* **346**, 440-445 (1990)

[55] Ducruix, A. & Geige, R., Crystallization of nucleic acids and proteins, a practical approach, Oxford University Press, Oxford, (1991)

[56] Scopes, R. K., Protein Purification: Principles and Practice, Springer Verlag, New York, (1994)

[57] Drenth, J., Principles of protein x-ray crystallography, Springer-Verlag, New York, (1994)

[58] Rosenberg, I. A., Protein analysis and purification: benchtop techniques, Springer-Verlag, New York, (1996)

[59] Bryngelson, J. D., When is a potential accurate enough for structure prediction – Theory and application to a random heteropolymer model of protein-folding, *J. Chem. Phys.* **100**, 60386045 (1994)

[60] Pande, V. S., Grosberg, A. Yu. & Tanaka, T., Non-Randomness in Protein Sequences: Ev- idence for a Physically Driven Stage of Evolution?, *Proc. Natl. Acad. Sci. USA* **91**, 1297212975 (1994)

[61] Gutin, A. M. & Shakhnovich, E. I., Ground state of random copolymers and the discrete random energy model, *J. Chem. Phys.* **98**, 81748177 (1993)

[62] Li, A. & Daggett, V., Characterization of the transition state of protein unfolding by use of molecular dynamics: Chymotrypsin inhibitor 2, *Proc. Natl. Acad. Sci. USA* **91**, 10430-10434 (1994)

[63] Lazaridis, T. & Karplus, M., "New view" of protein folding reconciled with the old through multiple unfolding simulations, *Science* **278**, 19281931 (1997)

[64] Daggett, V., Li, A. J., Itzhaki, L. S., Otzen, D. E. & Fersht, A. R., Structure of the transition state for folding of a protein derived from experiment and simulation, *J. Mol. Biol.* **257**, 430440 (1996)

[65] Sheinerman, F. B. & III, C. L. Brooks, Molecular picture of folding of a small alpha/beta protein, *Proc. Natl. Acad. Sci. USA.* **95**, 15621567 (1998)

[66] Kolinski, A. & Skolnick, J., Monte-Carlo simulation of protein folding. Lattice model and interaction scheme., *Proteins: Struc. Func. Genet.* **18**, 338352 (1994)

[67] Skolnick, J. & Kolinski, A., Simulation of the folding of a globular protein, *Science* **250**, 11211125 (1990)

[68] Dill, K. A., Fiebig, K. M. & Chan, H. S., Cooperativity in protein-folding kinetics, *Proc. Natl. Acad. Sci. USA* **90**, 19421946 (1993)

[69] Shakhnovich, E. I. & Gutin, A. M., Formation of unique structure in polypeptide chains: Theoretical investigation with the aid of a replica approach, *Biophys. Chem.* **34**, 187199 (1989)

[70] Wang, W., Donini, O., Reyes, C. M. & Kollman, P. A., Biomolecular simulations: Recent developments in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions, *Ann. Rev. Biophys. Biophys. Struct.* **30**, 211243 (2001)

[71] Li, A. J. & Daggett, V., Molecular dynamics simulation of the unfolding of barnase: Characterization of the major intermediate, *J. Mol. Biol.* **275**, 677-694 (1998)

[72] Finkelstein, V., Can protein unfolding simulate protein folding? *Protein Eng.* **10**, 843-845 (1997)

[73] Dinner, A. R. & Karplus, M., Is protein unfolding the reverse of protein folding? A lattice simulation analysis, *J. Mol. Biol.* **292**, 403-419 (1999)

[74] Ladurner, A. G., Itzhaki, L. S., Daggett, V. & Fersht, A. R., Synergy between simulation and experiment in describing the energy landscape of protein folding, *Proc. Natl. Acad. Sci. USA* **95**, 84738478 (1998)

[75] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E., The Protein Data Bank, *Nucl. Acid Res.* **28**, 235242 (2000)

[76] Wilson, C. & Doniach, S., A computer model to dynamically simulate protein folding: Studies with Crambin, *Proteins* **6**, 193209 (1989)

[77] Teeter, M. M., Roe, S. M. & Heo, N. H., Atomic resolution (0.83Å) crystal structure of the hydrophobic protein Crambin at 130 K, *J. Mol. Biol.* **230**, 292311 (1993)

[78] Garrett, T. P., Clingeleffer, D. J., Guss, J. M., Rogers, S. J. & Freeman, H. C., The crystal structure of poplar apoplastocyanin at 1.8Åresolution. The geometry of the copper-binding site is created by the polypeptide., *J. Biol. Chem.* **259**, 28222834 (1984)

[79] Ding, F., Borreguero, J.M., Buldyrev, S.V., Stanley, H.E., and Dokholyan, N.V., "Mechanism for the alpha-helix to beta-hairpin transition", *Proteins: Struct. Func. and Gene.* **53** 220-228 (2003)

[80] Munoz, V & Eaton, W. A., A simple model for calculating the kinetics of protein folding from three-dimensional structures, *Proc. Natl. Acad. Sci. USA* **96**, 11311-11316 (1999)

[81] Guerois, R. & Serrano, L., The SH3-fold family: Experimental evidence and prediction of variations in the folding pathways, *J. Mol. Biol.* **304**, 967-982 (2000)

[82] Nymeyer, H., Socci, N. D. & Onuchic, J. N., Landscape approaches for determining the ensemble of folding transition states: Success and failure hinge on the degree of frustration, *Proc. Natl. Acad. Sci. USA* **97**, 634-639 (2000)

[83] Clementi, C., Nymeyer, H. & Onuchic, J. N., 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.* **278**, 937–953 (2000)

[84] Klimov, D. K. & Thirumalai, D., Multiple protein folding nuclei and transition state ensemble in two-state proteins, *Proteins: Struc. Func. Genet.* **43**, 465–475 (2001)

[85] Sali, A., Shakhnovich, E. & Karplus, M., How does a protein fold, *Nature* **369**, 248-251 (1994)

[86] Zhou, Y. & Karplus, M., Interpreting the folding kinetics of helical proteins, *Nature* **401**, 400–403 (1999)

[87] Dokholyan, N. V., Buldyrev, S. V., Stanley, H. E. & Shakhnovich, E. I., Identifying the protein folding nucleus using molecular dynamics, *J. Mol. Biol.* **296**, 1183–1188 (2000)

[88] Abkevich, V. I. & Shakhnovich, E. I., What can disulfide bonds tell us about protein energetics, function and folding: Simulations and bioinformatics analysis, *J. Mol. Biol.* **300**, 975–985 (2000)

[89] Dokholyan, N. V., Buldyrev, S. V., Stanley, H. E. & Shakhnovich, E. I., Molecular dynamics studies of folding of a protein-like model, *Folding & Design* **3**, 577–587 (1998)

[90] Riddle, D. S., Grantcharova, V. P., Santiago, J. V., Alm, E., Ruczinski, I. & Baker, D., Experiment and theory highlight role of native state topology in SH3 folding, *Nature Struct. Biol.* **6**, 1016–1024 (1999)

[91] Jackson, S. E., How do small single-domain proteins fold?, *Folding & Design* **3**, R81–R91 (1998)

[92] Kabsch, W. Solution for the Best Rotation to Relate Two Sets of Vectors. *Acta Cryst.*, **A32**, 922-923 (1976)

[93] Vendruscolo, M., Paci, E., Dobson, C. & Karplus, M., Three key residues form a critical contact network in a protein folding transitional state, *Nature* **409**, 641–645 (2001)

[94] Grantcharova, V.P. & Baker, D. Folding dynamics of the src SH3 domain, *Biochemistry* **36**, 15685-15692 (1997)

[95] Itzhaki, L. S., Otzen, D. E. & Fersht, A. R., The structure of the transition-state for folding of chymotrypsin inhibitor-2 analyzed by protein engineering methods — evidence for a nucleation-condensation mechanism for protein-folding, *J. Mol. Biol.* **254**, 260–288 (1995)

[96] Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I., A protein engineering analysis of the transition state for protein folding: simulation in the lattice model, *Folding & Design* **3**, 183–194 (1998)

[97] Ozkan, S. B., Bahar, I. & Dill, K. A., Transition states and the meaning of -values in protein folding, *Nature Struct. Biol.* **8**, 729 - 819 (2001)

[98] Plaxco, K. W., Simons, K. T. & Baker, D., Contact order, transition state placement and the refolding rates of single domain proteins, *J. Mol. Biol.* **277**, 985–994 (1998)

[99] Munoz, V., Thompson, P., Hofrichter, J. & Eaton, W. A., Folding dynamics and mechanism of beta-hairpin formation, *Nature* **390**, 196–199 (1997)

[100] Larson, S. & Davidson, A., The identification of conserved interactions within the SH3 domain by alignment of sequences and structures, *Protein Science* **9**, 2170–2180 (2000)

[101] Grantcharova, V. P. & Baker, D., Circularization changes the folding transition state of the src SH3 domain, *J. Mol. Biol.* **306**, 555–563 (2001)

[102] Grantcharova, V. P., Riddle, D. S. & Baker, D, Long-range order in the src SH3 folding transition state, *Proc. Natl. Acad. Sci. U. S. A.* **97**, 7084-7089 (2000)

[103] Fersht, A.R., *Curr. Opinion Struc. Biol.* **5**, 79–84 (1995)

[104] Mirny, L.A. & Shakhnovich, E.I., *Ann. Rev. Biophys. Biophys. Struct.* **30**, 361-396 (2001)

[105] Shimada, J., Kussell, E.L. & Shakhnovich, E.I., *J. Mol. Biol.* **308**, 79-95 (2001)

[106] Vendruscolo, M., Dokholyan, N.V., Paci, E. and Karplus, M., "A small-world view of the amino acids that play a key role in protein folding." *Phys. Rev. E* **65**, 061910 (2002).

[107] Watts, D.J. & Strogatz, S.H., *Nature* **393**, 440-442 (1998)

[108] A.-L. Barabasi, R. Albert, Emergence of scaling in random networks, *Science* **286**, 509 (1999)

[109] H. Jeong *et al.*, The large-scale organization of metabolic networks, *Nature* **407**, 651 (2000)

[110] B. Bollobás, *Random graphs* (Academic Press, London, 1985)

[111] Dokholyan, N.V., Li, L., Ding, F., and Shakhnovich, E.I., "Topological determinants of protein folding." *Proc. Natl. Acad. Sci. USA* **99**, 8637-8641(2002).

[112] Neira , J.L. *et al.*, Towards the complete structural characterization of a protein folding pathway: the structures of the denatured, transition and native

states for the association/folding of two complementary fragments of cleaved chymotrypsin inhibitor 2. Direct evidence for a nucleation-condensation mechanism, *Folding & Design* **1**, 189-208 (1996)

[113] Thomas, P.J., Qu, H-H. and Pederson, P.L., *Thends Biochem. Sci.* **20**, 456-459 (1995)

[114] Guijarro, J.I., Sunde, M., Jones, J. A., Campbel, I. D. and Dobson, C. M., Amyloid fibril formation by an SH3 domain, *Proc. Natl. Acad. Sic. U.S.A.* **95**, 4224-4228 (1998).

[115] Tan, S.Y. & Pepys, M.B., Amyloidosis, *Histopathology* **25**, 403-414 (1994).

[116] Kelly, J.W., The alternative conformations of amyloidogenic proteins and their multi-step assembly pathways, *Curr. Opin. Strct. Biol.* **8**, 101-106 (1998).

[117] Sunde M, Serpell LC, Bartlam M, Fraser PE, Pepys MB, Blake CCF. The common core structure of amyloid fibrils by synchrotron X-ray diffraction. *J Mol Biol* 1997;**273**:729-739.

[118] Bonar, L., Cohen, A.S. & Skinner, M. , Characterization of the amyloid fibril as a cross-β Protein., *Proc. Soc. Expt. Biol. Med.* **131**, 1373-1375 (1967).

[119] Eanes, E.D. & Glenner, G.G., X-ray diffraction studies on amyloid filaments, *J. Histochem. Cytochem* **16**, 673-677 (1968).

[120] Sunde, M., Serpell, L.C. Bartlam, M., Fraser, P.E., Pepys, M.B. and Blake, C. C. F., The common core structure of amyloid fibrils by synchrotron X-ray diffraction, *J. Mol. Biol.* **273**, 729-739 (1997).

[121] Jimenez, J.L. et al., *EMBO Journal* **18**, 815-821 (1999)

[122] Zurdo, J., Guijarro, J.I., Jimenez, J.L., Saibil, H. R. and Dobson, C. M., Dependence on solution conditions of aggregation and amyloid formation by an SH3 domain, *J. Mol. Biol.* **311**, 325-340 (2001).

[123] Chiti, F., Webster, P., Taddei, N., Clark, A., Stefani, M., Ramponi, G. and Dobson, C. M., Designing conditions for in vitro formation of amyloid protofilaments and fibrils, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 3590-3594 (1999).

[124] Dobson, C.M., Protein misfolding, evolution and disease, *Trends Biochem. Sci.* **24**, 329-332 (1999).

[125] Hoshino, M., Katou, H, Hagihara, Y., Hasegawa, K., Naiki, H. & Goto Y., Mapping the core of the β_2-microglobulin amyloid fibril by H/D exchange, *Nature Struct. Biol.* **9**, 332-336 (2002).

[126] Tjernberg, L. O., Tjnberg, A., Bark, N., Shi, Y., Ruzsicska, B. P., Bu, Z. Thyberg, J. & Callaway, D. J. E., Assembling amyloid fibrils from designed structures containing a significant amyloid β-peptide fragment, *Biochem. J.* (2002).

[127] Bennett, M. J., Choe, S. & Eisenberg, D. S., Domain swapping - entangling alliances between proteins, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 3127-3131 (1994).

[128] Janowski, R., Kozak, M., Jankowska, E., Grzonka, Z., Grubb, A., Abahamson, M. and Jaskolski, M. Human cystatin C, an amyloidogenic protein, dimerizes through three-dimensional domain swapping, *Nature Struct. Biol.* **8**, 316-320 (2001).

[129] Liu, Y., Gotte, G., Libonati, M. and Eisenberg, D., A domain-swapped RNase A dimer with implications for amyloid formation, *Nature Struct. Biol.* **8**, 211-214 (2001).

[130] Sinha, N., Tsai C.J. & Nussinov, R., A proposed structural model for amyloid fibril elogation: domain swapping forms an interdigitaling β-structure polyer, *Protein Eng.* **14**, 93-103 (2001).

[131] Staniforth, A.A., Giannini, S., Higgins, L. D., Cornroy, J., Hounslow, A. M., Jerala, R., Craven, C.J. and Waltho, J. P., Three-dimensional domain swapping in the folded and molten-globule states of cystatins, an amyloid-forming structural superfamily, *EMBO J.* **20**, 4774-4781 (2001).

[132] Borreguero, J.M., Dokholyan, N. V., Buldyrev, S. V., Stanley, H. E. & Shakhnovich, E. I., Thermodynamic and folding kinetic analysis of the SH3 domain from discrete molecular dynamics *J. Mol. Biol.* **318**, 863-876 (2002).

[133] Singh A. P., Liu X. & Brutlag D. L., WWW tools for protein structure superposition and classification, *FASEB J.*, **13(7)**, A1542-A1542 *Suppl.* (1999).

[134] Guijarro, J.I., Marton, C.J., Plaxco, K.W., Campbell, I.D. & Dobson, C.M., Folding kinetics of the SH3 domain of PI3 Kinase by real-time NMR combined with optical spectroscopy, *J. Mol. Biol.* **276**, 657-667 (1998).

[135] Guo, C, Cheung, M.S. & Levin, H., Mechanisms of cooperativity underlying sequence-independent β-sheet formation, *arXiv*:cond-mat/0104065 (2001).

[136] Kraulis, P.J., Molscript - A program to produce both detailed and schematic plots of protein structures, *J. Appl. Cryst.* **24**, 946-950 (1991).

[137] Tsai, C. J., Maizel, J.V. & Nussinov, R., Anatomy of protein structures: Visualizing how a one-dimensional protein chain folds into a three-dimensional shape, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 12038-12043 (2000).

[138] Morrissey, M.P. & Shakhnovich, E.I., Evidence for the role of PrP^C helix 1 in the hydrophilic seeding of prion aggregates, *Proc. Natl. Acad. Sci. U.S.A.* **20**, 11293-11298 (1999).

[139] Chothia, C. & Finkelstein, A.V, The classification and origins of protein folding patterns, *Annu. Rev. Biochem.* **59**, 1007-1039 (1990).

[140] Gerstein, M. & Levitt, M., A structural census of the current population of protein sequences, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 11911-11916 (1997).

[141] Selkoe DJ. Amyloid beta-protein and the genetics of Alzheimer's disease. *J Biol Chem* 1996;**271**:18295-18298.

[142] Lansbury PT. A reductionist view of Alzheimer's disease. *Acc. Chem. Res.* 1996;**29**:317-321.

[143] Prusiner SB. Prion diseases and the BSE crisis. *Science* 1997;**278**:245-251.

[144] Harrison PM, Bamborough P, Daggett V, Prusiner SB, Cohen FE. The prion folding problem. *Curr Opin Struct Biol* 1997;**7**:53-59.

[145] Takahashi Y, Ueno A, Mihara H. Design of a Peptide Undergoing α-β Structural Transition and Amyloid Fibrillogenesis by the Introduction of a Hydrophobic Defect. *Chem Eur J* 1998:**4**:2475-2484.

[146] Fezoui Y, Hartley DM, Walsh DM, Selkoe DJ, Osterhout JJ, Teplow DB. A *de novo* designed helix-turn-helix peptide forms nontoxic amyloid fibrils. *Nature Struct Biol* 2000;**7**:1095-1099.

[147] Ono S, Kameda N, Yoshimura T, Shimasaki C, Tsukurimichi E, Mihara H, Nishino N. Supper-secondary structure with amphiphilic β-strands probed by phenylalanine. *Chem Lett* 1995:965-966.

[148] Cerpa R, Cohen FE, Kuntz ID. Conformational switching in designed peptides: The helix/sheet transition. *Folding & Design* 1996;**1**:91-101.

[149] Schenck HL, Dado GP, Gellman SH. Redox-triggered secondary structure changes in the aggregated states of a designed methionine-rich peptide. *J Am Chem Soc* 1996;**118**:12487-12494.

[150] Hamada D, Segawa S, Goto Y. Non-native alpha-helical intermediate in the refolding of beta-lactoglobulin, a predominantly beta-sheet protein. *Nature Struct Biol* 1996;**3**:868-873.

[151] Ferrara P, Apostolakis J, Caflisch A. Thermodynamics and kinetics of folding of two model peptides investigated by molecular dynamics simulations. *J Phys Chem* 2000;**104**:5000-5010.

[152] Hansmann UHE, Okamoto Y. Finite-size scaling of helix-coil transitions in poly-alamine studied by multicanonical simulations. *J Chem Phys* 1999;**110**:1267-1276.

[153] Sung SS, Wu XW. Molecular dynamics simulations of synthetic peptide folding. *Proteins* 1996;**25**:202-214.

[154] Takano M, Yamato T, Higo J, Suyama A, Nagayama K. Molecular dynamics of a 15-residue poly(L-alamine) in water: helix formation and energetics. *J Am Chem Soc* 1999;**121**:605-612.

[155] Bertsch RA, Vaidehi N, Chan SI, Goddard WA. Kinetic steps for α-helix formation. *Proteins* 1998;**33**:343-357.

[156] Hirst JD, Brooks CL. Molecular dynamics simulations of isolated helices of myoglobin. *Biochemistry* 1995;**34**:7614-7621.

[157] Daggett V, Levitt M. Molecuar dynamics simulations of helix denaturation. *J Mol Biol* 1992;**223**:1121-1138.

[158] Simth AV, Hall CK. α-Helix Formation: Discontinuous Molecuar Dynamics on an Intermediate-Resolution Protein Model. *Proteins* 2001;**4**:344-360.

[159] Smith AV, Hall CK. Protein Refolding Versus Aggregation: Computer Simulations on an Intermediate-resolution Protein Model. *J Mol Biol* 2001;**312**:187-202.

[160] Eaton WA, Munoz V, Thompson PA, Chan CK, Hofrichter J. Submillisecond kinetics of protein folding. *Curr Opin Struct Biol* 1997;**7**:10-14.

[161] Callender RH, Dyer RB, Gilmanshin R, Woodruff WH. Fast events in protein folding: the time evolution of primary processes. *Annu Rev Phys Chem* 1998;**49**:173-202.

[162] Ballew RM, Sabelko J, Gruebele M. Direct observation of fast protein filding: the initial collapse of apomuoglobin. *Proc Natl Acad Sci USA* 1996;**93**:5759-5764.

[163] Shastry MCR, Roder H. Evidence for barrier-limited protein filding kinetics on the microsecond time scale. *Nature Struct Biol* 1998;**5**:385-392.

[164] Pascher T, Chesick JP, Winkler JR, Gray HB. Protein folding triggered by electron transfer. *Science* 1996;**271**:1558-1560.

[165] Schellman JA. The factors affecting the stability of hydrogen-bonded polypeptide structures in solution. *J Phys Chem* 1958;**62**:1485-1494.

[166] Zimm BH, Bragg JK. Theory of the phase transition between helix and random coil in polypeptide chains. *J Chem Phys* 1959;**31**:526-535.

[167] Lifson S, Roig A. On the theory of helix-coil transition in polypeptides. *J Chem Phys* 1961;**34**:1963-1974.

[168] Munoz V, Serrano L. Elucidating the folding problem of helical peptides using empirical parameters. *Nature Struct Biol* 1994;**1**:399-409.

[169] Rathore O, Sogah DY. Self-Assembly of β-sheets into Nanostructures by Poly(alanine) Segments Incorporated in Multiblock Copolymers Inspired by Spider Silk. *J Am Chem Soc* 2001;**123**:5231-5239.

[170] Takada S, Luthey-Schulten Z, Wolynes PG. Folding dynamics with nonadditive forces: a simulation study of a designed helical protein and a random heteropolymer. *J Chem Phys* 1999;**110**:11616-11629.

[171] Zhou Y, Karplus M. Folding of a model three-helix bundle protein: A thermodynamic and kinetic analysis. *J Mol Biol* 1999;**293**:917-951.

[172] Honig B, Yang AS. Free-energy balance in protein-folding. *Adv. Protein Chem.* 1995;**46**:27-58.

[173] Ramakrishnan C, Ramachandran GN. *Biophys J* 1965;**5**:909.

[174] Pappu RV, Srinivasan R, Rose GD. The flory isolated-pair hypothesis is not valid for polypeptide chains: implications for protein folding. *Proc Natl Acad Sci USA* 2000;**97**:12565-12570.

[175] Richardson JS, Richardson DC. Natural β-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc Natl Acad Sci USA* 2002;**99**:2754-2759.

[176] Guo Z, Thirumalai D. Kinetics and thermodynamics of folding of a *de novo* designed four-helix bundle protein. *J Mol Biol* 1996;**263**:323-343.

[177] Ho SP, DeGrado WF. Design of a 4-helix bundle protein: synthesis of peptides which self-associate into a helical protein. *J Am Chem Soc* 1987;**109**:6751-6758.

[178] Grosberg AY, Khokhlov AR. *Statistical Physics of Macromolecules*. AIP Press 1994.

[179] Awasthi SK, Shankaramma SC, Paghothama S, Balaram P. Solvent-induced β-hairpin to helix conformational transition in a designed peptide. *Biopolymers* 2001;**58**:465-476.

[180] Sha YL, Li YL, Wang Q, Fan KQ, Liu DS, Lai LH, Tang YQ. CD evidence of a peptide ongoing alpha/beta/random transition in different solutions. *Protein Pep Lett* 1999;**6**:137-140.

Curriculum Vitæ
Feng Ding

Boston University, Physics Department
Center For Polymer Studies
590 Commonwealth Avenue
Boston, Massachusetts 02215 USA

Telephone: 617/353-8936
Facsimile: 617/353-9393 or 617/353-3783
E-mail: fding@polyme.bu.edu
WWW: http://polymer.bu.edu/~fding

EDUCATION

- Ph.D., 2003 - Physics; Boston University, Boston, MA
- B.S. honors, 1997 - Physics; Nanjing University, China

EMPLOYMENT

- Research Assistant, Physics Department, Boston University. Spring 2001 - present.
- Teaching Assistant, Physics Department, Boston University. Fall 1997 - Spring 2001.

HONORS, AWARDS

- 1997: "Exellent Graduate", Nanjing University, 1997
- 1993-1997: Recipient of People's Stipend, Nanjing University

REFERENCES

- H. Eugene Stanley: University Professor of Physics, Professor of Physiology, Director of Center for Polymer Studies; Department of Physics, Boston University, 590 Commonwealth Avenue, Boston, MA 02215. **Tel:** (617) 353 2617, **FAX:** (617) 353 3783,

  Email: hes@bu.edu.
- Eugene I. Shakhnovich: Professor of Chemistry, Department of Chemistry, Harvard University, Cambridge, MA
- Nikolay V. Dokholyan: Assistant Professor, Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, NC

List of Publications

I. Condensed Matter Physics:

[1] X.Y. Lei, H. Li, F. Ding, W. Zhang and N.B. Ming, "Novel application of a perturbed photonic crystal: High-quality filter.", *Appl. Phys. Lett.* **71**, 2889-2891 (1997)

II. Protein folding and aggregation:

[2] N.V. Dokholyan, L. Li , F. Ding and E.I. Shakhnovich, "Topological determinants of protein folding", *P. Natl. Acad. Sci. USA* **99**, 8637-8641 (2002)

[3] F. Ding, N.V. Dokholyan, S.V. Buldyrev, H.E. Stanley and E.I. Shakhnovich, "Molecular dynamics simulation of C-Src SH3 aggregation suggests a generic amyloidogenesis mechanism", *J. Mol. Biol.* **324**, 851-857 (2002)

[4] F. Ding, N.V. Dokholyan, S.V. Buldyrev, H.E. Stanley, and E.I. Shakhnovich, "Direct molecular dynamics observation of protein folding transition state ensemble.", *Biophys. J.* **83**, 3525-3532 (2002)

[5] N.V. Dokholyan, J.M. Borreguero, S.V. Buldyrev, F. Ding, H.E. Stanley, and E.I. Shakhnovich, "Identifying the importance of amino acids for protein folding from crystal structures." Methods in Enzymology, Vol. 374: Macromolecular crystallography D. Editors: C.W. Carter Jr. and R.M. Sweet (2003)

[6] F. Ding, J.M. Borreguero, S.V. Buldyrev, H.E. Stanley, and N.V. Dokholyan, "Mechanism for the alpha-helix to beta-hairpin transition", *Proteins: Struct. Func. and Gene.* **53**, 220-228 (2003)

[7] S. Khare, F. Ding and N.V. Dokholyan, "Folding of Cu,Zn superoxide dismutase and Familial Amyotrophic Lateral Sclerosis", *J. Mol. Biol.* **334**, 515-525 (2003)

[8] J.M. Borreguero, F. Ding, S.V. Buldyrev, H.E. Stanley, and N.V. Dokholyan, "Multiple Folding Pathways of the SH3 domain", *J. Mol. Biol.*, submitted (2003)

[9] S. Peng, F. Ding, B. Urbanc, S.V. Buldyrev, L. Cruz, H.E. Stanley and N.V. Dokholyan, "Discrete molecular dynamics simulations of peptide aggregation", *Phys. Rev. E*, submitted (2003)