Influence of Cooperativity on the Protein Folding Mechanism

A dissertation submitted to
Kent State University in partial
fulfillment of the requirements for the
degree of Doctor of Philosophy

by

Xianghong Qi

December 2008
Dissertation written by

Xianghong Qi

B.S., Hebei Normal University, 1994
M.S., Hebei Normal University, 1997
Ph.D., Kent State University, 2008

Approved by

________________________
Dr. John J Portman, Chair, Doctoral Dissertation Committee

________________________
Dr. Jonanthan Seilinger, Members, Doctoral Dissertation Committee

________________________
Dr. Elizabeth M Mann

________________________
Dr. Peter Tandy

________________________
Dr. Chun-che Tsai

Accepted by

________________________
Dr. Bryon D. Anderson, Chair, Department of Physics

________________________
Dr. John R.D. Stalvey, Dean, College of Arts and Sciences
3.4 Barrier crossing dynamics ........................................ 39
3.5 Generalized mobility matrix ...................................... 43
  3.5.1 Native density correlation function $C(t)$ .................. 43
  3.5.2 Monomer pair correlation: $G(t)$ ......................... 45
3.6 Model Parameters .................................................. 46
3.7 Conclusion .......................................................... 47

4 THE EFFECT OF COOPERATIVITY ON TWO-STATE PROTEIN FOLDING ........................................... 48
  4.1 Cooperativity ....................................................... 48
  4.2 Cooperativity effect on folding rate. ............................ 53
    4.2.1 Folding rates and prefactors .............................. 56
    4.2.2 Free energy barrier height ............................... 61
  4.3 Cooperativity effect on transition state structure. .......... 65
  4.4 Conclusion ........................................................ 72

5 SCALING OF FOLDING RATE WITH PROTEIN LENGTH ............ 74

6 CAPILLARITY OF FOLDING NUCLEUS GROWTH ..................... 80
  6.1 Capillarity approximation to protein folding nuclei .......... 80
  6.2 Characterizing the folded core and the interface ............ 85
  6.3 Growth of folding nucleus along the folding route .......... 87
  6.4 Growth pattern of the nucleus ............................... 90
  6.5 Polarized versus diffuse critical nucleus .................... 94
6.6 Conclusion ........................................................................... 96

7 STABILITY EFFECT ON PROTEIN FOLDING ......................... 100

7.1 Effect of temperature on protein kinetics .............................. 101

7.2 Movement of transition state .............................................. 101

7.2.1 Hammond shift ............................................................ 102

7.2.2 Sudden changes in Qts ................................................... 104

7.3 Effect on prefactor ............................................................. 108

7.4 Effect on chevron plot ....................................................... 110

7.5 Conclusion ........................................................................ 114

8 APPENDIX ............................................................................ 115

8.1 Glossary of terms ............................................................. 115

8.2 Folding path ................................................................. 117

8.3 2D-plot of native density for cooperativity and noncooperativity ... 119

8.4 Folding pattern .............................................................. 126

8.5 Φ-value analysis ............................................................. 131

8.6 Scaling exponent ............................................................. 133

8.7 Stability effects on temperature, transition state position, barrier height and prefactor ........................................... 135

8.8 Chevron plot ................................................................. 140
LIST OF FIGURES

1. The schematic illustration of the energy landscape for a minimally frustrated protein folding. Here, the landscape is inherently many-dimensional, so funnel schematic is only a projection. The depth represents the energy of a conformational state, the width represents the measure of a configurational entropy. .............................................. 4

2. The schematic illustration of free energy surface for fast, two-state proteins at folding temperature $T_f$. The native state (folded state) and globule state (unfolded state) are separated by a barrier at some value of reaction coordinate which corresponds to transition state ensemble (TSE). How to cross this free energy barrier gives the detail of dynamics of protein folding which will be reflected from the calculation of folding rate and structure characterization of TSE. .............................................. 9

3. The illustration of potential well and distribution of native contact pairs for $\lambda$–repressor. ................................................................. 31
Comparison between experimental and calculated folding rates. The correlation coefficient is $r = 0.8$, with $p$-value $p = 4.4 \times 10^{-7}$. (circle: $\alpha$—protein, square: $\beta$—protein, and diamond: $\alpha\beta$—protein.). Here, x-ray structured proteins are: 1pgb, 1coa, 1csp, 1c8c, 1pin, 1mb, 1enh, 1fkb, 1urn, 1shg, 1ten, *1div, *1fnf9, *1pgb16; NMR structured proteins are: 1o6x, 1imq, 1srl, 2ptl, 1aps, 2abd, *1a0n, *2pdd, *1mef, *1psf, *1wit, *1pks, *1vii, *1hdn. (Proteins not used in $\Phi$-value analysis are indicated by *).

5 Histogram of the inverse folding prefactor $\tau_0 = 1/k_0$ for the 28 two-state proteins. Ordinate is number of proteins.

6 Free energy profile as a function of global order parameter $Q$ given by Eq. 3.36: the red curve is the folding route with cooperativity, and the black curve is the route without cooperativity. The circles denote the critical points defining the folding route, and the curves are the steepest-descents paths.

7 Free energy barrier $F^\dagger/k_B T_f$ plotted against the absolute contact order, $ACO = 1/N_{con} \sum |i - j|$ where the sum is over the $N_{con}$ native contacts pairs. Red points correspond to cooperative folding routes, and black points correspond to non-cooperative routes. Symbols have the same meaning as Fig. 4. For the cooperative routes, the correlation coefficient is $r = 0.91$ with $p$-value $p = 1.6 \times 10^{-11}$; for non-cooperative routes, the correlation coefficient if $r = 0.41$ with $p$-value $p = 0.03$. 

vii
(color) Folding route for protein 1srl characterized locally by the normalized native density $\tilde{\rho}_i = (\rho_i - \rho_i(G))/(\rho_i(N) - \rho_i(G))$ and the global progress coordinate $Q = 1/N \sum \tilde{\rho}_i$. (a) The local and global structure along the folding route with cooperativity (left figure); (b) and without cooperativity (right figure). The degree of structural localization of each residue is reflected in the colors, linearly scaled between red ($\tilde{\rho}_i = 0$) to blue ($\tilde{\rho}_i = 1$).

(color) The change in the local structure of the transition state ensemble characterized by the cross-correlation coefficient $\Omega$ plotted against the change in the global order parameter $\Delta Q^*$ is shown in the center. Left and right panels give the typical examples of the native density profile for various values of overlap $\Omega$. In each example, the red line corresponds to the cooperative route, and, the black line corresponds to the non-cooperative route. The corresponding correlation coefficient between measured and predicted $\Phi$-value are: 1pin (0.77), 1lm (0.84), 1enh (0.87); 2ptl (0.0), 1imq (-0.1), 1o6x (-0.2).

The schematic illustration of chevron plot (the top one) for point mutations with $\Phi$ values of 0 and 1 respectively (the lower ones). The kinetics under various thermodynamics conditions produces characteristically V-shaped chevron plots of $\log k_f$ and $\log k_u$ versus denaturant concentration.
Barrier heights as a function of chain length. (a) $\beta \Delta F^\dagger$ vs. $N^{1/2}$, fitting line is $y = -7.0 + 0.88x$, and correlation coefficient is 0.89. (b) $\beta \Delta F^\dagger$ vs. $N^{2/3}$, fitting line is $y = -11.0 + 2.3x$, correlation coefficient is 0.9. (c) $\beta \Delta F^\dagger$ vs. $\ln N$, fitting line is $y = -24.8 + 7.8x$, correlation coefficient is 0.89. (d) correlation coefficient vs. power $P$.

Cutoff value of chainlength vs power $P$ and inset gives the $\tau_0 = 1/k_0$ vs. power $P$.

Illustration of folding nucleus: folded core, interfacial region, and unfolded halo. Growth of the nucleus can be characterized by fluxes entering the folded core and interfacial regions.

Scaling of the folded core with number of monomers. (a-b) correspond to $\lambda$–repressor (1lmb). (a) Residues with native density $\tilde{\rho}_i > 0.6$ (indicated by the dashed line) define a fiducial set of folded residues.

(b) Linear fit of $\log V_f$ vs. $\log N_f$ (dashed line) gives the exponent of $V_f \sim N_f^{3\nu}$. In this example, the fitting equation is $y = 5.6 + 0.97x$, so that $\nu = 0.32$ and $b^3 = 5.6a^3$, $a = 3.8\AA$ is the average distance between the $\alpha$ carbons. (c) Histogram of scaling exponent $\nu$ for 28 proteins; (d) Histogram of the packing fraction $\mu_f$ as given in Eq. 6.7 of the critical nucleus at $T_f$. 

ix
Illustration of growth of folding nucleus and interface along the folding route (increasing $Q$) for the $\lambda$-repressor protein (1lmb). Column 1 shows the three-dimensional folded structure with the fiducial set of folded residues colored blue and the unfolded residues colored red. In Column 2, the folded core (colored blue) is surrounded by the interfacial region (colored green). Column 3 is a projection of the indicator function $\tilde{n}(r)$ that defines the folded and interfacial regions in space. The values correspond to $max_z \tilde{n}(x, y, z)$, ranging from 1 (blue) to 0 (red) in steps of 0.01. Contour lines correspond to 0.1, 0.5, 0.7. Column 4 gives the corresponding $Q$ value for each row. The critical nucleus corresponds to $Q = 0.53$. The units for three plots are in Angstroms. This protein is belong to the Pattern C (balanced growth). The three-dimensional structure was produced by VMD.

Examples of three patterns of growth of the nucleus. Pattern A–C corresponds to (a–c), respectively. Solid line corresponds to the mean packing fraction of the folded core, $\mu_f$, while the dashed line corresponds to the mean packing fraction of the interface, $\mu_{int}$.

An example of a polarized critical nucleus (Q=0.45) for Src-SH3 (1sr1). Plots (a–c) correspond to the middle row of the diffuse nucleus shown in Fig. 15.
18 Inverse packing fraction of the interface and folded core. for 27 two-
state proteins. The gradual change color shows the continuous change
from polarized nuclei (red) to diffuse nuclei (cyan) .......................... 97
19 Temperature vs. stability for λ-repressor(1lmb). log $K_u$ is defined in
Eq. 7.2. ....................................................................................... 102
20 The transition state position $Q_{ts}$ vs. stability for λ-repressor(1lmb). 103
21 Antihammond behavior for protein 1o6x and 1fnf9. ...................... 104
22 $Q_{ts}$ and prefactor $k_0(1/\mu s)$ vs. stability for protein 1srl. .......... 105
23 Folding route of 1srl for varied stability. ...................................... 106
24 Φ values for each residue of protein 1srl. Φ values for each sec ondary
structure were plotted separately. ............................................. 107
25 Averaged Φ value vs. stability for 1srl ........................................ 108
26 An example of prefactor vs. stability (1enh). .............................. 109
27 Two-state chevron plot for 1wit (top) and chevron roll-over for 2pdd
(bottom). More positive stability values correspond to conditions un-
der which the native state is more favored thermodynamically. ....... 112
28 2D plots of native density for all of studied proteins. The horizontal
axis is reaction coordinate Q, the vertical axis is residue index. Here the
native density is normalized by native density at globule and native.
Red corresponds to fully unstructured, blue corresponds to fully or-
dered state, the gradual color change between red and blue correspond
to the structure change of each residue according to folding progress. 125
Scaling exponents for studied proteins. Here, the fitting line (dashed line) is the best linear fitting between logarithm of folded volume ($V_f$) and logarithm of the number of folded residues ($N_f$) in most folded process.
LIST OF TABLES

1 Two-state proteins I studied in this work. Kinetic data for proteins 1enh, 1vii, 1hdn were taken from Ref. 1; for the rest of proteins were taken from Ref. 2. ................................................................. 55

2 Correlation for theoretical and experimental $\Phi$ values for x-ray crystallography proteins. $R$ is correlation coefficient for the case without cooperativity effects. $R_{coop}$ is correlation coefficient for the case with cooperativity effects. ................................................................. 69

3 Correlation for theoretical and experimental $\Phi$ values for NMR structure proteins. $R$ is correlation coefficient for the case without cooperativity effects. $R_{coop}$ is correlation coefficient for the case with cooperativity effects. ................................................................. 71
ACKNOWLEDGMENTS

I am very glad to have this opportunity to express my thanks to all of people who have been part of my Ph.D life.

Firstly, I want to thank my advisor, John J Portman, for his constantly valuable guidance in research from June 2004 to Aug 2008. Especially, during writing dissertation, I really appreciate his great support and unbelievable patience. His deep passion to scientific research is really impressed. Without his great guidance, this dissertation work would not have been finished.

I also want to thank Dr. Kumar’s group. I spent two years in his lab and learned so much from Dr. Kumar, Dr. Wang (Tony), Dr. Choi.

Also our main office, Mrs. Kris, Mrs. Lauretta, Cindy and Kim, I even don’t remember how many times I asked for their kind help for past eight years. Of course, I still need thank my officemate, Swarendu Tripathy, we have been in one office for several years, we have so many wonderful and fruitful discussions.

Finally, I want to thank my families. I know only "thanks" can not express my feeling from bottom of heart. My daddy, mom and brothers, my mother-in-law and father-in-law, they gave me their deeply love and support from my oversea home. My special friend, life-last love, my dear husband, Yuezhong Yu, I don’t know how to thank him. I only know that I will do nothing without his dedicated support and love. In this whole life, I am so lucky having him beside me and we walked through
so many difficult times and share so many wonderful times together. I owe too much
to my lovely twin girls, Kathy Yu and Kelly Yu. How many weekends I should have
spent with them instead of working in the office.

I would like to thank Dr. Mann, Dr. Selinger, Dr. Tandy and Dr. Tsai for their
time and willingness to be my committee.
CHAPTER 1

INTRODUCTION

Proteins are linear polymers with monomers selected from 20 different amino acids joined by peptide bonds. They are the basic unit of life. It can also be said that life would be impossible without proteins. Every cell of your body constructs specific proteins to meet very diverse needs of human physiology. Modern molecular biology is centered around how the hereditary information is stored and passed on in the simple, one-dimensional (1D) sequence of DNA base pairs [3]. The connection between heredity and biological function is made through the transmission of this 1D information, through RNA, to the protein sequence of amino acids. The information contained in this sequence is sufficient to completely determine a protein’s geometrical 3D structure, at least for simpler proteins which are observed to reliably refold without chaperons or steric constraints [4]. Folding to a specific structure is typically a prerequisite for a protein to perform its biological function in the cell.

Therefore understanding the nature of the interactions that stabilize protein structure and govern protein folding mechanisms is a fundamental problem in molecular biology [5–10]. This subject attracts scientists from a wide range of disciplines [11] such as molecular biologists, chemists, physicists and mathematicians. Major advances have occurred in both theoretical [6, 12–19] and experimental approaches (such
as mass spectrometry, differential scanning calorimetry, dynamic light scattering, X-ray and neutron diffraction, NMR, etc) to investigate protein folding mechanism over last two decades. An understanding of this fundamental process is crucial to predict structure from sequence, design of proteins rationally, and to understand diseases associated with misfolded proteins. Here I give a brief review of protein folding problem.

1.1 Modern review of protein folding—energy landscape theory

In 1961, Anfinsen's classical experiments on renaturation of unfolded proteins [20] proved that the primary sequence of a protein is sufficient to determine its three dimensional structure. Soon after his experiments, Levinthal [21] first put forward the idea that proteins cannot fold by a random search of conformational space. Although the initial stages of folding must be nearly random, if the entire process was a random search, it would require too much time. For example, a 100 residue protein, if each residue is considered to have just 3 possible conformations, the total number of conformations of the protein is $3^{100}$. If conformational changes occur on a time scale of $10^{-13}$ seconds, i.e. the time required to sample all possible conformations would be $3^{100} \times 10^{-13}$ seconds which is about $10^{27}$ years. Even if a significant proportion of these conformations are sterically disallowed, the folding time would still be astronomical. However, proteins are known to fold on a time scale of sub-milliseconds to minutes. In order to solve this problem (later called that "Levinthal paradox"),
Levinthal postulated the notion of a definite protein folding pathway that would reduce the relevant conformational space. The search for such pathway then became a motive for researchers to characterize the protein folding problem. At the time, proteins with long-lived intermediates became the main focus of research.

Although the search for definite pathways dominated protein folding research for many years, the modern view sees this as only one way (and not the typical way!) to avoid Leventhal’s paradox. The energy landscape theory describes folding as occurring on a free energy surface in the shape of a rugged funnel [13, 17, 22–24] (see Fig. 1). Therefore, protein folding can be described as a progressive organization of an ensemble of partially folded structures through which a protein passes on its way to folded structure. In this view, folding doesn’t need to follow one specific pathway, but a multitude of pathways. The number of pathways progressively reduce when approaching to folded state (as shown in Fig. 1). The mechanism for folding is associated with an entropic bottleneck rather than a sequence of definite intermediates as in a single pathway. Bryngelson and Wolynes [22, 25] creatively developed the energy landscape theory by focusing on the difference between protein folding and random systems such as spin glasses. This approach gives new, quantitative insights into the interpretation of experiments and simulations of protein folding thermodynamics and kinetics. Importantly, the picture provides a unified framework to understand the diverse set of folding kinetics. I will review three aspects of the energy landscape theory of protein folding.

The first characteristic is statistical properties of energy landscape theory. The use
Figure 1: The schematic illustration of the energy landscape for a minimally frustrated protein folding. Here, the landscape is inherently many-dimensional, so funnel schematic is only a projection. The depth represents the energy of a conformational state, the width represents the measure of a configurational entropy.
of the statistics to describe protein folding phenomena is quite natural (even though each protein has a specific sequence, structure, and function essential to its biological activity). Proteins have a vast set of folded structures with an extraordinary diversity of conformational states of unfolded states. In order to describe such system with huge number of conformational states immediately requires a statistical method.

The second characteristic is that energy landscape theory employs a small number of order parameters. It is evident that protein folding occurs on an high-dimensional conformational space with intricate connectivities between configurational states. Nevertheless, Bryngelson and Wolynes [22, 23] argued that the kinetics of folding could be described by diffusion on a low-dimensional free energy surface, like a one-dimension free energy surface only if we find a proper variable to serve as a reaction coordinate which measures the degree of similarity to the folded state. We call this progressive coordinate the reaction coordinate or order parameter.

There are a variety of ways of measuring the similarity of a protein structure to the native state. Like the choice of Bryngelson and Wolynes [22, 23, 26], one can take the fraction of the amino acids residues which are in correct local configuration. Also the average energy of a state with a certain similarity to the native structure has a value that gets lower as the native state is approached. So it can be a good order parameter too. There are a variety of other possibilities. For example, the magnitude of the average conformational fluctuations of each amino acid clearly depends on the structure of protein. In general, the average fluctuation of amino acid is smaller in the native state and significantly larger in globule state (unfolded state). So this average
fluctuation also can be a good candidate for an order parameter.

Socci et al [27] proved the validity of the energy landscape ideas of Bryngelson and Wolynes [22, 23] in protein folding by using simulations of a 27-mer on a cubic lattice. Socci et al found that simulated folding kinetics can be captured by a simple one-dimension barrier crossing. Their results encourage us to think that for real proteins we can describe the folding as a low-dimensional diffusion process when the free energy profile can be calculated accurately.

The third characteristic is the principle of minimum frustration which is essential to energy landscape theory for protein folding. Bryngelson and Wolynes [22, 23] use the term principle of minimum frustration to describe the driving force towards the unique native structure. It is just this property of the Hamiltonian which distinguishes natural protein sequences from random heteropolymers. The principle of minimum frustration proposes that the nativelike structures of natural proteins sequences have a lower free energy on average than random nonnative configurations. This encourages fast folding and provides a slope in the free energy surface that guides the protein to native state. The protein landscape also has properties associated with frustration, trapped states, and glass transitions like random heteropolymers, but the landscapes also obey the principle of minimum frustration which serves as a driving force toward the native state. The resulting energy landscape of natural protein sequences resembles in shape a rugged funnel [22, 23]. The relative degree of frustration is measured by comparing a folding transition temperature ($T_f$) with a glass transition temperature ($T_g$), which characterizes the thermodynamics of trapping. Simulations
of simple models show that sequences with a high $T_f/T_g$ ratio fold faster than most sequences [28, 29] and with few intermediates. In addition, the quantitative minimum frustration principle has been successfully used to dramatically improve energy function for structure prediction [30, 31]. The quantitative theory of minimal frustration also provides an automated algorithm for designing sequences that can efficiently fold to nearly unique structures [32].

From the above description of the energy landscape theory for protein folding, it is clear that the principle of minimum frustration and the funnel concept suggest that the main features of folding kinetics can be predicted by knowing the stabilization energies of elements of the native structure and the entropic costs of bring together parts of the scaffold. This design principle of the energy landscape demonstrates that the main driving force opposing the transition from a high-energy, high-entropy, disordered unfolded state to a low-energy, low-entropy ordered native state is the necessary loss of conformational entropy of the protein. Upon folding, the solvent averaged free energy of a chain configuration $\Delta E$ almost always decreases because most interactions are favorable, whereas acquisition of specific order results in the loss of chain entropy $\Delta S$. At temperature $T$, the free energy $\Delta F = \Delta E - T\Delta S$ has a barrier along the reaction coordinate due to incomplete cancellation of the entropy and energy loss. Therefore the free energy barrier separating folded and unfolded states is a consequence of an imbalance between entropy and energy for some value of reaction coordinate, which defines the transition state ensemble (TSE) during folding process. The structural characterization of TSE is the central focus of folding
mechanism since the evolution of TSE characterizes protein folding kinetic mechanism completely. If there is a single kinetic bottleneck, the free energy profile consists of an ensemble of two states separated by a free energy barrier as shown in Fig. 2. A large body of work shows that naturally occurring, fast folding, two-state proteins have a simple, relatively smooth funneled energy landscape. This is indicated by the observation of simple, single-exponential kinetics even at the lowest temperatures that can be investigated experimentally [33]. Furthermore, the energy landscape theory suggests that we can describe the folding as a low-dimensional diffusion process when the free energy profile can be calculated accurately. So, according to diffusion theory of folding on a minimally frustrated energy surface, the folding rate (inverse of folding time) should be well approximated by the Kramer’s escape rate formula

\[ k_f = k_0 \exp(-\beta \Delta F^\dagger). \]

Here, \( k_0 \) is the folding prefactor, which depends on the configurational diffusion coefficient and the geometric shape of the energy barrier. The microscopic dynamics are mostly characterized by the timescale expressed in \( k_0; \Delta F^\dagger \) is the barrier in the free energy profile, \( \beta = 1/k_B T \), and \( T \) is the temperature. In contrast to typical gas phase chemical reactions, specific decomposition of the folding rate into a prefactor and barrier is indirect because of the limited range of stability where proteins have two-state behavior.

In this thesis, I focus on naturally occurring fast two-state folding proteins. For this class of proteins, simple models have proved successful applied to describe and predict kinetics and the structure of the transition state ensemble (TSE). Furthermore, recent experiments have been devised to probe the early folding events and to
Figure 2: The schematic illustration of free energy surface for fast, two-state proteins at folding temperature $T_f$. The native state (folded state) and globule state (unfolded state) are separated by a barrier at some value of reaction coordinate which corresponds to transition state ensemble (TSE). How to cross this free energy barrier gives the detail of dynamics of protein folding which will be reflected from the calculation of folding rate and structure characterization of TSE.

Many theoretical and experimental studies focus on the structural characterization of transition state ensemble. One way to probe the structure of transition state ensemble experimentally is through $\Phi$–value analysis, which was developed by Fersht and coworkers [51]. In this genetic engineering technique, single site mutations serve as microscopic probes of the transition state. For every mutant, one can measure $\Phi$ value, which is an approximate measure of native structure content in the transition state ensemble. The $\Phi$ value is defined as the ratio of the change in free energy of TSE to the change in free energy of native state between the mutant and wild
sequence: $\Phi = \frac{\Delta \Delta G^f}{\Delta \Delta G^U_{U=N}}$. Usually, the $\Phi$ value is between 0 and 1, $\Phi \sim 0$ indicates that the mutated site is unstructured in the transition state ensemble as it is in the unfolded state ensemble; $\Phi \sim 1$ indicates that the mutated site is fully native-like in the transition state ensemble. $\Phi$ values have provided a very direct way to compare predictions of theoretical models to experimental measurements. We will discuss $\Phi$ analysis in detail in Chapter 3.

The globally funneled landscape guarantees that protein folding is robust. However, since the folding mechanism for a individual protein depends on many inherent and environmental factors, exploring how to describe the detailed mechanisms of folding is complicated. Nevertheless, the common features of folding mechanism still can be inferred from simple, perfect funnel models which in fact most naturally occurring proteins employ. Up to now, the most common and popular perfect funnel models are the Gō models which was first introduced in old lattice simulation work by Gō and coworkers [52]. One underlying assumption for this class of models is that only the interactions that are present in the native structure are taken into account, these are the so-called native interactions. The stabilization of partially folded conformations is approximated by which pair of residues in proximity in the native structure are brought together. In this approximation (using the pairwise additive potential), the $\Phi$ values for main components of transition state ensemble of small proteins such as protein A, CI2, and SH3 were accurately predicted [53]. This prediction was made by knowing the distribution of native contacts that is determined by the "topology" of the native structure. This success gives an important lesson for folding: a folded
structure determines the mechanism of folding. This idea also has been suggested from additional computation studies and experiments. For instance, the comparison of the $\Phi$ values of two homologous proteins, $\alpha$-spectrin SH3 [54] and C-src SH3 [55] shows that $\Phi$ values for both proteins are highly correlated, which was attributed to the similarity in their native folds. Onuchic et al [17, 56] found that the mean transition state ensemble structure is insensitive to the type of potential: homogeneous strength Gō-like potentials and inhomogeneous Gō-like potentials all have similar $\Phi$ values (correlated by about 75%) [57]. Also computer simulations of fragment B of protein A [58] from minimalist models leads to the similar conclusions: transition state ensemble heterogeneity is strongly determined by topology.

Further support for topology-determined folding mechanism is due to Plaxco, Simons and Baker [59] who discovered that a simple topological parameter, termed the relative contact order (RCO), correlated with the folding rates of two state folding proteins. After Plaxco’s remarkable findings, several other topology-based parameters have been proposed such as long range order (LRO) [60], total contact distance (TCD) [61], the number of sequence-distant native pairs ($Q_D$) [62]. These parameters all correlated well with the logarithm of folding rates. These studies strongly suggest that the topology is one of the main determinants of folding rates. But at same time, we have to remember that these simple parameters are not a complete description of topology, and are primarily phenomenological. We will discuss these parameters in detail in Chapter. 2.
While the above-mentioned intellectual observations support the topology of native structure determines the protein folding mechanism, predictions from simple models still have significant problems. In particular, the range of simulated folding rates is much smaller than experimental measurements and predicted barrier heights are much smaller than expected [63–65]. These models are based on native-centric \( C_\alpha \) model (G\( \ddot{\mathrm{o}} \)-like model) although they studied different set of proteins and used different modeling set-up. This may indicate the general properties of G\( \ddot{\mathrm{o}} \) or G\( \ddot{\mathrm{o}} \)-like models based on pairwise additive forces.

This mismatch in folding rate diversity and barrier heights is a central motivation for the work in this thesis. It has been pointed out that, many forms of G\( \ddot{\mathrm{o}} \) or G\( \ddot{\mathrm{o}} \)-like models fail to embody a sufficient high degree of "cooperativity" [66, 67] to mimic real two-state proteins [68, 69]. Here, cooperativity usually refers to a mechanism by which the presence of a structured region makes additional order more favorable. Indeed, as Eastwood and Wolynes pointed out [70] explicit cooperative interactions must be present in reality since simplified models introduced effective interactions which cannot be written as a sum of pair potential in general since many degrees of freedom have been integrated out in going from a detailed to a more coarsely grained description. Also, a local-nonlocal off-lattice model interaction scheme has been shown to increase the diversity in model folding rates with enhancement of cooperativity [71]. In addition, a recent study of a class of G\( \ddot{\mathrm{o}} \)-like chain model with three body interactions as a perturbation shows that such non-pairwise, many-body effects can significantly increase the range of model folding rates, and therefore
enhance the correlation between model and experimental rates [72]. These findings suggest that cooperativity and non-additive many-body interactions are likely to be a critical in accounting for the tremendous diversity among experimental folding rates [63].

In this thesis, I was motivated by the discrepancy between predicted folding rates and diverse experimental folding rates and also inconsistence between predicted barrier heights and expected ones. I will focus on exploring effect of cooperativity on protein folding mechanism based on improved variational theory which incorporated cooperativity by pairwise-additive way in the model. Explicitly, I was motivated through questions such as:
1. What is the effect of cooperativity on folding rates?
2. How can the barrier heights be influenced by the cooperativity?
3. How to describe the TSE structure appropriately under considering cooperativity?
4. What is the value of the prefactor and is it a uniform constant for two-state proteins?

We hope to get a better picture of protein folding after careful investigation about above questions.

1.2 Nucleation picture of protein folding

The mechanism for protein folding can also be understood through an analogy of first-order phase transition from unfolded ensemble to folded ensemble. A series of
theoretical studies [26, 73–75] and experiments [76, 77] suggested nucleation mechanism via the formation of a specific or non-specific folding nucleus. Here the nucleus means a minimal folded fragment that results in inevitable subsequent unidirectional folding from unfolded ensemble to folded ensemble. Correspondingly, the transition state ensemble is also termed as a critical folding nucleus. In itself this is not really different than how we have described the folding mechanism so far; yet the language of nucleation tends to emphasize distinct spatial properties of the folding mechanism.

The nature of folding nuclei, its evolution along the folding process, what controls the size of the folding nucleus, are still interesting and unresolved questions.

Especially, when the structure of transition state ensemble can be described through the site-directed mutagenesis experiment developed by Fersht group [77], the specificity of the transition state ensemble has been fully characterized. However, a complete description of transition state ensembles includes two independent aspects: specificity and spatial arrangement. So a clear description of spatial properties such as size or density of the transition state ensemble is still in demand and will serve as a necessary component to describe the folding mechanism in detail.

In fact, shortly after characterizing the transition state ensembles of CI2, Fersht proposed the *nucleation-condensation* mechanism to describe critical nucleus spatially [77]. According to the nucleation-condensation mechanism, the critical nucleus can be thought of as an expanded, partially ordered version of the native state ensemble. It is natural to think the extent of spatial arrangement of the native interactions among residues for individual states of this ensemble is different for different proteins.
For some proteins, native contacts formed in typical states of the transition state ensemble are spatially uncorrelated [22, 26], this kind of transition state ensemble is called the diffusive transition state ensemble. From Φ-values analysis, the diffusive critical nucleus show the intermediate Φ values spread across a large portion of the protein sequence. However, for some other proteins, all the stabilizing native interactions for typical configurations in the transition state ensemble cluster into a more “localized” nucleus with a well defined surface region. One usually refers to this kind of transition state ensemble as polarized transition state ensemble. Φ-value analysis show that one part of the structure has relatively high Φ values while the rest of the residues have low Φ values in the polarized transition state ensemble.

Theoretically, Wolynes describes a nucleus with capillarity-like order in which interface surrounding a relatively folded core is broadened by wetting of partially ordered residues [78]. This capillarity approximation of folding nuclei is based on the classical nucleation theory of first-order phase transition kinetics [26]. In this picture, the folding can be described as the growth of the folding nucleus. In this thesis, part of the work focuses on how to predict the evolution of the folding nucleus for 27 two-state fast folding proteins in terms of capillarity model. In Chapter. 6, I answer the following questions:

1. How does the critical folding nucleus scale with chain length?

2. What is the evolution of folding nucleus pattern for specific proteins given the sequence and native structure of a protein?

3. How to quantify the critical folding nucleus as the diffusive or polarized transition
1.3 Stability

The stability of a protein is defined as the difference in free energy between the folded and unfolded state, $\Delta F = F_U - F_N$. The stability determines the relative population of folded and unfolded conformations. It is often expressed as $\Delta F = F_U - F_N = RT \ln K$, where $K$ is the equilibrium constant. For two-state folding proteins, the equilibrium constant is the ratio of the folding rate and unfolding rate constant: $K = k_f/k_u$. It is important to study for many reasons. Understanding how folding kinetics depend on stability provides additional understanding of the basic thermodynamics of the process of folding. Also increased protein folding stability may be a multi-billion dollar value in food and drug processing, and in biotechnology and drug designing. In this thesis, I investigate how the changing stability can affect the folding mechanism. For instance, the effect on folding rate, free energy barrier heights, position of transition state ensemble, and behavior of chevron plots.

1.4 Organization of thesis

The thesis is organized as follows:

In Chapter 2, I give a detail review of some necessary background information about protein folding studies.

In Chapter 3, I describe the variational method developed by Portman, Takada and Wolynes [79–81]. This method is a microscopic theory of the free energy barriers and folding routes for minimally frustrated proteins. It presents microscopic calculations.
of transition state ensemble, activation free energies, and the dynamical prefactor.

Chapter 4, I present the variational model which incorporates cooperativity. By searching for minima and saddle points of free energy surface using an eigenvector-following algorithm, the average folding routes were determined for 28 small, two-state, fast folding proteins. Based on the average folding route, I calculated prefactors, free energy barrier heights and so the folding rates. Then I characterize the structure of transition state ensembles for all of studied proteins with native density.

In Chapter 5, I discuss the relationship between size of chain and folding rates.

In Chapter 6, I focus on the description of spatial properties of critical nuclei based on capillarity approximations. How the folding nuclei grow with the number of residues in the folded region is presented, also the balance between the density of folded core and interface region give a detail folding pattern picture. Lastly, the diffusiveness of critical folding nucleus is quantified by the volume of per particle in folded region and interface region.

In Chapter 7, I focus on investigating the effect of changing stability on the folding mechanism. I present the movement of the position of transition state ensemble with changes in stability. Also, the chevron rollover phenomenon is discussed for a few proteins. This may anticipate the presence downhill folding, which is currently somewhat controversial [82, 83].

Finally, I put some necessary results into Appendix and also give a glossary for reading convenience.
CHAPTER 2

BACKGROUND INFORMATION OF PROTEIN FOLDING

My aim in this chapter is to review the thermodynamic background to protein folding problems, with an overview of the current picture as I see it. Many detailed reviews in this area have appeared [4, 8, 13, 25, 84–90], and it is not my intention to cover the same ground as much detail as can be found there. Rather, I will try to provide sufficient basic background to allow understanding and critical appraisal of my thesis work.

2.1 Minimal protein folding models

In minimalist protein models, entire protein residues (or groups of residues) are represented at a simplified level by interacting beads connected by polymeric constraints. The earliest theoretical work of minimalist protein models was executed on 2D lattice models by Gō and coworkers [91]. They were able to construct a protein-like system that folds to a unique structure. Although their energy function was thought to be ad hoc since it encoded the native structure within it, they were able to study and observe a number of interesting properties of these models. At the time, there were no experiments to compare their results to and this approach was not pursued for some time. The potentials of similar forms are now commonly referred to as Gō or Gō-like potential and are still in use in studying a variety of questions [53, 58, 64, 92]. These models represent each amino acid by a single position corresponding to the
$C_\alpha$ atom in the real protein. The interaction are native-centric: favorable energy terms are assigned on the basis of a native contact set derived from the PDB structure. Usually, a pair of residues belong to the native contact if they satisfy the two conditions: they are not part of the backbone constraints; and the spatial distance between them should be within a certain cutoff value (A cutoff between 6 – 8Å is commonly used to define native contacts in a Gō model). In addition to the hard repulsion between beads which avoids unphysical conformations, an attractive interaction between every pair of amino acids which are in contact in the native structure is included. Simulation of Gō-models have the advantage that the polymeric backbone can be more faithfully represented, but are always subject to statistical errors which are particularly important in characterizing low probability states (such as the TSE). Analytical models avoid such statistical errors, but have systematic errors due to approximation required to make the models practical. Thus, both approaches are complementary to investigate the nature of the mechanism of protein folding.

Important insights have come from simulations of simplified representations of proteins in lattice and off-lattice models [6, 13, 13, 16, 16–19, 27, 85, 91, 93–118]. Such models provide simple, concrete examples that can be extremely helpful in clarifying both theoretical and experimental issues. Like lattice models, minimalist models can be used to study various aspects of protein folding such as the global properties of energy landscapes, conditions for folding; and the possible structure and organization of transition state. Removing the lattice constraint on degrees of freedom makes off-lattice model (simple continuum models) much easier to represent protein-like
structures, especially secondary structure. Also one can use a more natural potential to permit global movement of molecular dynamics which incorporate all length scales movement.

As mentioned earlier, a powerful experimental method for probing transition states is to measure the Φ values [51]. Partial or complete Φ values have been determined for many proteins including CI2 [77], SH3 [54, 55], a variant of the lambda repressor proteins [36], barnase [119], barstar [120], CheY [41] and protein L [121, 122]. The predictions from minimalist protein models give reasonable comparison with experimental measurements even up to the site-resolved level for several small, two-state, fast folding proteins [64, 92].

2.2 Topology determination mechanism

In recent years, much work established that the folding rates and structure of TSE are determined by native topology in native protein structure. One important question is how to understand the wide span of folding rates. The measured folding rates \( k_f \) (in units of \( s^{-1} \)) for proteins of about 100 residues have been found to span 6 orders of magnitude from microseconds for simple helical proteins [123, 124] to seconds for more complex topologies [125–127]. Another question is that how to predict the structure of TSE comparing to experimental probing (like Φ-values). If this idea is reasonable, the folding rates should show some kind of explicit relationship with some typical topological parameters accompanying with a certain extent prediction of structure of TSE.
Through observations of experimental folding rates, in 1998, Plaxco, Simons and Baker [59] discovered a simple parameter termed as the relative contact order (RCO), which was defined as the average distance of native contacts:

\[
RCO = \frac{1}{LN_{\text{con}}} \sum_{ij} |i - j| \delta_{ij}
\]  

(2.1)

If \(i, j\) are in native contact, \(\delta_{ij} = 1\), otherwise, \(\delta_{ij} = 0\). \(N_{\text{con}}\) is number of native contacts, \(L\) is chain length. \(|i - j|\) is the sequence distance between native pair \(i\) and \(j\).

RCO is a parameter that characterizes, to some extent, the “topology complexity” of native structure. For 12 small, two-state proteins, Plaxco’s findings showed that logarithm of folding rates are correlated well with RCO (correlation coefficient is 0.81). This findings supports that native topology is one of quantities which determines protein folding rates even though RCO is an very incomplete representation of folded state ensemble. After Plaxco’s findings [59], several other topology-based parameters have been proposed. They all show similar abilities to predict folding rates to a certain extent. For example, Gromiha and Selvaraj introduced LRO (long range order) [60], Zhou and Zhou [61] defined total contact distance (TCD), Rose group [128] proposed local secondary structure content, and the number of sequence-distant native pairs is proposed by Makarov and Plaxco [62].

So all of above mentioned intellectual observations support that the topology of native structure determines the protein folding rates. While these parameters show suggestive correlations with folding rates, they can not supply any information about structure of TSE. Thus, these simple parameters cannot fully explain similarities
and differences in different protein’s folding mechanisms. The true theoretical basis is still in demand. To a certain extent, these parameters are too simple to fully characterize the native topology complexity and thus the correlations between folding rates and these parameters only support the idea of topology determination of folding mechanism.

2.3 Topology-based simulations

Simulated Gō-models are based on a more complete description of the native-state topology than contact order. Such models can not only predict folding rates, but can also predict the structure of the TSE. Koga and Takada [64] performed folding simulations on 18 small two-state proteins with Gō-like potential. They found that a unified scaling law between folding rates and chain length, RCO at the folding temperature: $k_f \sim \exp(-cRCL^{2/3})$. Their findings is in harmony with the scaling properties $k_f \sim \exp(-cL^{2/3})$ predicted by Finkelstein and Badredtinov [129] and Wolynes [78] at folding temperature, which is based on a simple scaling law derived from nucleation in a first order transition. We will discuss the nucleation picture of folding in Chapter. 6. Another exciting aspect of this work is that they investigate the structure of TSE, too. Their findings are consistent with experimental Φ-value data for about half of proteins. Although their prediction about TSE structure is not perfect, it still provides a very instructive way to study TSE. Nevertheless, a striking difference between simulations and experiments was in the range of simulated folding rates: the measured folding rates varies over six orders of magnitude depending on
topology, while simulated folding rates only span two orders of magnitude.

Later, Chavez and Clementi [65] performed off-lattice simulations on 16 two-state proteins through Gō-like potential. The experimental folding rates of protein set almost span nine orders of magnitude. The raw data shows simulated folding rates span around four to five orders of magnitude, around half range of measured folding rates. Chavez and Clementi confront this problem with optimism. Through inhomogeneous rescaling of simulated folding rates, they bring the simulated folding rates to have the same range as experimental and also both rates are highly correlated each other (correlation coefficient is 0.94). But their rescaling procedure was questioned by Wallin and Chan [63].

Furthermore, Wallin and Chan [63] performed simulations on 13 two-state proteins with native-centric $C_\alpha$ model (Gō-like model). Their simulated folding rates and experimental folding rates correlated moderately with correlation coefficient 0.69. However, simulated rates span only approximately two orders of magnitude comparing to six orders of magnitude of measured folding rates.

In contrast to Sect. 2.2, simulation studies provide more convincing evidence for the topology determination of folding mechanism from both folding rates prediction and TSE structure prediction. The mismatch between the range of simulated and experimental folding rates probably is a general properties for Gō-models as shown by Koga and Takada [64], Chavez and Clementi [65] and Wallin and Chan [63]. In my thesis, I will present our strategy to make up this discrepancy by incorporating cooperativity into our variational model in Chapter. 4
2.4 Survey of analytic protein folding models

At the same time, several theoretical models were also developed. Through the theoretical framework, the connection with experiments, simple models, and detailed all-atom simulations are further established, moving toward a fully quantitative theory for protein folding. For example, Shoemaker and Wolynes [130–132] formulated their free energy functional through a set of contact probabilities with considering the inhomogeneity of contact energies, chain entropy and cooperative contributions reflecting many-body character of some forces. Their results show a semi-quantitative agreement with experiments on hen lysozyme and α-lactalbumin [131], when they also apply this free energy functional method to a detailed investigation of transitions state ensemble of chymotrypsin inhibitor (CI2, pdb: 1coa) and λ–repressor [130, 132], the results qualitatively agree with the experimental data.

Based on Shoemaker and Wolynes’ free energy functional method [131, 132], Plotkin and Onuchic [4, 133] introduced a simple coarse-grained free energy functional in terms of a general distribution of contact probabilities. From the knowledge of this functional, all relevant thermodynamic functions can be calculated in general such as transition state entropies and energies, barrier heights and surface tensions. Moreover, the second derivative of functional can give the microscopic structure and dynamic behavior of transition state ensemble. This theoretical framework was tested against the lattice simulation. Finally this method elucidated the general properties of structural heterogeneity and native energetic heterogeneity on folding. Potentially this theory will be a useful guide in interpreting and predicting experimental results.
on many fast-folding proteins.

The work in this thesis is based on the model developed by Portman, Takada and Wolynes [79–81]. This is a variational approach to calculate the free energy profiles and characterize the transition state ensemble. With the help of a harmonic reference Hamiltonian, the free energy profile is formulated through the fluctuations of each residue about the average folded conformations. The critical points (minimum, maxima or saddle points) are used to map out the free energy surface and indicate the dominant folding route. Along the folding route, the transition state ensembles correspond to the saddle points in the multi-dimensional free energy surface. Through characterizing the structure of unstable growth mode near the saddle point, the microscopic dynamics involved in barrier crossing can be elucidated. Their results for λ-repressor agreed with the interpretation of kinetic data based on experimental analysis [36].

The above mentioned theoretical framework (Shoemaker and Wolynes [131, 132], Plotkin and Onuchic [4, 133], and Portman, Takada, Wolynes [79–81]) share a remarkable commonality: they are capable of characterizing partial ordered structure, or transition state ensemble structure.

There is another type of theoretical model: Ising-like models. These models began with the work of Zwanzig and his coworkers [134, 135]. They assumed that the residue can only be in two states: native (“correct”) or non-native (“incorrect”). The model itself had an analytic solution of the kinetic equations for motion along the free energy profile. Later, Muñoz et al. [136–139] improved the model and calculated free
energy profiles by using the number of native residues as the reaction coordinate. For 18 small, two-state proteins, their model basically produce the two-state behavior for most of proteins and also calculated folding rates correlated with experimental folding rates very well. However, the model failed in predicting structure of TSE indicated by Φ-analysis. Baker’s group [140] and Finkelstein’s group [141] presented somewhat similar models which are more successful in identifying the structure of transition state ensemble and provide a semiquantitative concordance with the experimental data. The folding rates are loosely correlated with experimental folding rates (correlation coefficient is around 0.6). Most recently, Henry and Eaton [142] developed a combinatorial approach which almost reproduced two-state behaviors for 25 proteins except α-helical proteins. Their calculated folding rates are correlated with experimental folding rates reasonably well.

For all of Ising-like models mentioned above, the precondition is that the residue only can be in complete folded state or complete unfolded state. This simplicity is a great point of this kind of model, however it is also an obstacle for investigating the detailed structure of TSE which is expected to involve partial order. In contrast, characterizing partially ordered structures is a great strength of the variational model which will be discussed in the next Chapter.
CHAPTER 3

REVIEW OF VARIATIONAL MODEL OF PROTEIN FOLDING

The site-resolved variational model developed by Portman, Takada, and Wolynes [80, 81] has proved reliable in predicting the structure of the transition state ensemble of individual proteins at the residue level of resolution [79, 143, 144]. These studies confirm that the simple model Hamiltonian and the scheme of approximation adopted in the model is adequate to describe the spatial order of the TSE. This variational model treats the folding and unfolding process as structural localization and delocalization transitions, similar to freezing and melting of an amorphous solid. The big advantage of this simple model is that it can characterize partially ordered structures. In this chapter, I will review this model in detail. The formalism presented here relies heavily John Portman’s Ph.D dissertation [145].

3.1 Stiff Chain Model

In my study, I model the polymer by the α carbons composing the polypeptide backbone. With this choice, the root mean square separation distance $|\mathbf{a}|$ is the typical distance between adjacent α carbons $|\mathbf{a}| \approx 3.8\text{Å}$. Also, chain stiffness is treated harmonically through the persistence length $l$. For simplicity, here the chain stiffness is assumed to be uniform so that the persistence length is related to the chain stiffness by $l \approx a/(1 - g)$. For proteins, a reasonable value for the chain stiffness is $g = 0.8$ which corresponds to the persistence length of polyalanine, $l = 5a \approx 20\text{Å}$ [146, 147].
In addition to stiffness, a confining potential controls the overall size of the polymer chain. The Hamiltonian for a collapsed chain can be written as:

\[ \beta H_{\text{chain}} = \frac{3}{2a^2} \sum_{ij} r_i \cdot \Gamma_{ij} \cdot r_j + \frac{3}{2a^2} B \sum_i r_i^2, \quad (3.1) \]

where \( r_i \) denotes monomer position \( \{r_i\} \) and the correlations of monomer positions are given by the quadratic coefficients:

\[ \langle r_i \cdot r_j \rangle/a^2 = [\Gamma]^{-1}_{ij}. \]

Polymeric connectivity and chain stiffness are modeled by the correlations of monomer positions [148]:

\[ \Gamma = \frac{1-g}{1+g} K^R + \frac{g}{1-g^2} [K^R]^2 - \frac{g^2}{1-g^2} \Delta, \quad (3.2) \]

where \( K^R \) is the Rouse matrix for a nearest-neighbor harmonic chain.

\[
K^R = \begin{bmatrix}
1 & -1 & \cdots & 0 \\
-1 & 2 & -1 & \vdots \\
\vdots & \ddots & \ddots & \ddots \\
0 & \cdots & -1 & 1 \\
\end{bmatrix},
\quad (3.3)
\]

and \( \Delta \) is accounts for the "boundaries" at the end of the chain:

\[
\Delta = \begin{bmatrix}
1 & -1 & \cdots & 0 \\
-1 & 1 & \vdots \\
\vdots & 1 & -1 \\
0 & \cdots & -1 & 1 \\
\end{bmatrix},
\quad (3.4)
\]
In the chain Hamiltonian expression (Eq. 3.1) the second term controls the degree of collapse of the chain through a confining potential where the parameter $B$ is conjugate to the radius of gyration of the chain. Then the chain Hamiltonian (Eq. 3.1) can be written as:

$$\beta H_{\text{chain}} = \frac{3}{2a^2} \sum_{ij} r_i \cdot [\Gamma^{(ch)}]_{ij} \cdot r_j$$  \hspace{1cm} (3.5)

with

$$\Gamma^{(ch)}_{ij} = \Gamma_{ij} + B\delta_{ij}$$  \hspace{1cm} (3.6)

denoting the inverse monomer correlations of the collapsed stiff chain.

The total Hamiltonian for this simple model is given by:

$$H = H_{\text{chain}} + H_{\text{int}}^N.$$  \hspace{1cm} (3.7)

Here, $H_{\text{chain}}$ is the backbone potential defining the polymeric correlations given in Eq. 3.5 and $H_{\text{int}}$ is the native interaction potential between distant monomers. The interaction is limited to the set of contacts found in native structure, named as native contact pairs. The native contact pair is defined to be the pairs of residues $(i + 4 \leq j)$ that have $\beta$ carbons ($\alpha$ carbons for glycine) distances within a 6.5Å cutoff in the folded structure. We also include in this set residue pairs that are likely to have hydrogen bonds (as determined by the DSSP algorithm [149]) but fall outside the cutoff. The strength of the interaction for this set depends on the residue identities of the pair. I use the magnitude of the Miyazawa-Jernigan energy parameters reported in Ref. [150] in units of $\epsilon_0 = k_B T_0$. Temperatures in this model are quoted in terms of the
reduced temperature $T = \frac{k_B T}{\epsilon_0}$. Now we can write $H_{int}^{Nat}$ as:

$$H_{int}^{Nat} = \sum_{ij} \epsilon_{ij} u(|r_i - r_j|)$$  \hspace{1cm} (3.8)

where, $\epsilon_{ij}$, the strength of the interaction, depends on the identity of the residues $i$ and $j$ [150]. The spatial dependence of the interactions between distant monomers, $u(r_{ij})$, consists of an attractive well and a repulsive core. For computational convenience, we approximate the interaction potential as the sum of three Gaussians over native pairs:

$$u(r) = \sum_{k=(s,i,l)} \gamma_k \exp\left[-\frac{3}{2d^2} \alpha_k r^2\right],$$  \hspace{1cm} (3.9)

where $(\alpha_s > \alpha_i > \alpha_l)$ are the ranges of the short-, intermediate-, and long-range interactions, respectively. The intermediate-range term is repulsive($\gamma_i > 0$) and the long-range term is attractive($\gamma_l < 0$); the intermediate- and long-ranged potential parameters are chosen so that the sum of these two terms gives a potential well at an appropriate distance for contacts in the native structure. The short-range term is repulsive($\gamma_s > 0$) and represents the hard core repulsion between residues (See the Fig. 3).

The partition function for the system with the total Hamiltonian (Eq. 3.7) $Z = \int \prod_i d\mathbf{r}_i \exp(-\beta H)$ cannot be solved exactly. Instead, we introduce a reference Hamiltonian $H_0$ that corresponds to a polymer in a non-uniform external field that constrains the monomers to lie near the locations in the native structure. This is to say, the reference Hamiltonian is taken to be a polymer with harmonic constraints that
localize each amino acid around its native position \( \{ r_i^N \} \),

\[
\beta H_0 = \beta H_{\text{chain}} + \frac{3}{2a^2} \sum_i C_i (r_i - r_i^N)^2. \tag{3.10}
\]

The strength of the harmonic constraints, \( C_i \), are conjugate to the fluctuations of each residue with respect to its native position. In protein X-ray crystallography, the magnitude of the fluctuation (Debye-Waller factors) are called temperature factors \( B_i = \langle (r_i - r_i^N)^2 \rangle \). The magnitude of these fluctuations can distinguish two stable phases of the protein: the globule (or unfolded) state corresponds to large fluctuations (weak constraints) and native (folded) state corresponds to small fluctuations (strong constraints). I will discuss the importance of \( C_i \) and \( B_i \) in next section.

The external constraints in \( H_0 \) influence both the correlations \( G_{ij} \) and average
positions \( \{ s_i \} \) of the monomers composing the reference chain:

\[
G_{ij} = \langle \delta r_i \cdot \delta r_j \rangle_0 / a^2 = [\Gamma^{(0)}]^{-1}_{ij} \tag{3.11}
\]

\[
s_i = \langle r_i \rangle_0 = \sum_j G_{ij} C_j r_j^N, \tag{3.12}
\]

where \( \delta r_i \) is the position of the \( i^{th} \) monomer relative to the average:

\[
\delta r_i = r_i - \langle r_i \rangle_0 = r_i - s_i. \tag{3.13}
\]

and \( \Gamma^{(0)} \) is the matrix of coefficients of the quadratic terms of \( H_0 \):

\[
\Gamma_0^{ij} = \Gamma_{ij}^{ch} + C_i \delta_{ij}. \tag{3.14}
\]

### 3.2 Variational free energy surface

We now consider the variational free energy surface. Let us denote the partition function corresponding to the reference Hamiltonian (Eq. 3.10), \( H_0 \), by \( Z_0 = \int \prod_i dr_i \exp(\beta H_0) \), and average with respect to \( H_0 \) by \( \langle \cdots \rangle_0 = \frac{\int \prod_i dr_i \cdots \exp(\beta H_0)}{Z_0} \).

We approximate the partition function with the help of the reference Hamiltonian as follows:

\[
Z = \int \prod_i dr_i \exp(\beta H) \tag{3.15}
\]

\[
= \int \prod_i dr_i \exp[-\beta(H - H_0)] \exp(\beta H_0) \tag{3.16}
\]

\[
\approx \int \prod_i dr_i \exp(-\beta H_0)[1 - \beta(H - H_0)] \tag{3.17}
\]

\[
\approx Z_0 - \beta Z_0 \langle H - H_0 \rangle_0 \tag{3.18}
\]

\[
\approx Z_0(1 - \beta \langle H - H_0 \rangle_0) \tag{3.19}
\]

\[
\approx Z_0 \exp(-\beta \langle H - H_0 \rangle_0) \tag{3.20}
\]
This gives the variational free energy surface parameterized by the constraint parameters \( \{ C_i \} \):

\[
F[\{ C \}] = -k_B T \log Z_0 + \langle H - H_0 \rangle_0.
\] (3.21)

Substituting the expression of \( H \) and \( H_0 \) gives the variational free energy \( F = E - TS \).

The energy and entropy can be expressed as functions of \( \{ C \} \):

\[
E = \sum_i \epsilon_{ij} \langle u(r_{ij}) \rangle_0
\] (3.22)

and

\[
S/k_B = \log Z_0 + \sum \langle C_i (r_i - r_i^N)^2 \rangle_0
\] (3.23)

Since \( H_0 \) is quadratic, the Boltzmann weight is a Gaussian distribution. Thus \( Z_0 \) and all of the averages are expressible in terms of correlations \( G_{ij} \) and average positions \( s_i \) of the monomers. One instructive way to calculate the averages is to introduce approximations to the density of monomer \( i \), \( \rho^1_i(r) = < \delta(r - r_i) >_0 \),

\[
\rho^1_i(r) = \left[ \frac{3}{2\pi a^2 G_{ii}} \right]^{3/2} \exp \left[ -\frac{3(r - s_i)^2}{2a^2 G_{ii}} \right],
\] (3.24)

and the pair density between \( i \) and \( j \), \( \rho^2_{ij}(r) = < \delta(r - (r_i - r_j)) >_0 \),

\[
\rho^2_{ij}(r) = \left[ \frac{3}{2\pi a^2 \delta G_{ij}} \right]^{3/2} \exp \left[ -\frac{(r - (s_i - s_j))^2}{2a^2 \delta G_{ij}} \right],
\] (3.25)

Where \( \delta G_{ij} = < (\delta r_i - \delta r_j)^2 >_0 / a^2 = G_{ii} + G_{jj} - 2G_{ij} \). These densities depend on the constraint parameters \( \{ C_i \} \) through \( G_{ij} \) and \( \{ s_i \} \). Averages over \( H_0 \) can be calculated through \( \rho^1_i(r) \) and \( \rho^2_{ij}(r) \). For example,

\[
<u(|r_i - r_j|) >_0 = \int d r \rho^2_{ij}(r) u(r).
\] (3.26)
In this way, the variational free energy can be viewed as a density functional with a particular approximation to the density that simultaneously incorporates the polymeric correlations and the monomeric fluctuations about the average positions. Therefore, after some manipulations, we can calculate energy and entropy in terms of the monomer correlations and mean positions.

The entropy can be written as:

\[
S[C] = \frac{3}{2} \log \det G - \frac{3}{2a^2} \sum_{ij} s_i \cdot \Gamma^{(ch)}_{ij} \cdot s_j + \frac{3}{2} \sum_i C_i G_{ii}. \tag{3.27}
\]

Here, the first term is the entropy of the chain due to polymeric fluctuations, the second term is the entropy of the loss of fixing each monomer to the average positions, and the last term is the entropy of the vibrations about the mean position (= \((3/2a^2) \sum C_i < \delta r_i^2 >_0\)). Similarly, the pair potential can be averaged over \(H_0\) to give the energy:

\[
E[C] = \sum_{ij} \epsilon_{ij} u_{ij}, \tag{3.28}
\]

here

\[
u_{ij} = \langle u(|r_{ij}|) \rangle_0 \tag{3.29}
\]

\[
= \sum_{k=(s,i,l)} \frac{\gamma_k}{(1 + \alpha_k \delta G_{ij})^{3/2}} \exp \left[ -\frac{3}{2a^2} \frac{\alpha_k (s_i - s_j)^2}{1 + \alpha_k \delta G_{ij}} \right]. \tag{3.30}
\]

Finally, we choose to measure the free energy relative to the unconstrained chain:

\[
\Delta F[C] = \Delta E[C] - T \Delta S[C], \tag{3.31}
\]

where, for example, \(\Delta F[C] = F[C] - F[C=0]\).
The above procedure transforms the initially difficult statistical mechanical problem of particles moving in 3D space to a free energy function of scalars for each residue, which can be evaluated numerically. The original $3 \times N$ degrees of freedom are the Cartesian coordinates representing individual microscopic configurations, whereas the reference Hamiltonian having $N$ constraint values each specify an ensemble of structures. Therefore, setting values for $\{C\}$ corresponds to selecting an ensemble of conformations specified by average positions $\{s_i\}$ and fluctuations $G_{ij}$.

The energy of a given pair is most stabilizing when the pair density is contained in the potential well, i.e., the mean separation between monomers is within the well and fluctuations are relatively small. However, accompanying this stabilization is the entropy loss of localizing the positions of the pair. We can understand this loss as follows. The free energy to form a single contact in an unstructured chain is a competition between stabilizing energy gain and the polymeric entropy loss to form a loop. In the current model, this is controlled by the constraints. The monomers are localized enough so that most of contacts are within the attractive well, but the price is the costs of entropy. For a network of contacts, the picture of this energy/entropy compensation remains. As folding proceeds, partially folded ensembles are described by an evolving network of contacts.

In the variational free energy surface, $F[C]$, the $N$ variational parameters, $C_i$, are conjugate to the Debye-Waller factors $\{B_i\}$, which are routinely found by fitting X-ray crystallography data to a model structure that allows for fluctuations. Each $C_i$ indicates how localized a given residue $i$ is around the native state positions. In
the globule(unfolded) state, the protein explores many conformations, and there is little well-defined structure. Here, the fluctuations of a residue about any particular position are relatively large. This will be indicated by smaller \( \{ \mathcal{C} \} \) values (weak constraints). In the folded state, the conformations are much more restricted and can be described as relatively small fluctuations about the localized positions of the average native structure. These small fluctuations are reflected by larger \( \{ \mathcal{C} \} \) values (strong constraints). Thus, we can represent the relative flexibility of a partially ordered chain by the normalized temperature factor:

\[
\bar{B}_i = (B_i - B_i^G)/(B_i^N - B_i^G)
\]  

where the superscripts \( G \) and \( N \) denote the fluctuations evaluated at the globule and native state, respectively. This is a natural local order parameter that can characterize the partially ordered residues.

The globule (unfolded) and native (folded) states correspond to minima in the free energy surface, \( F[\mathcal{C}] \). The path connecting these two minima which has lowest barrier is the preferred folding path. The top of the folding barrier is a saddle-point in \( F[\mathcal{C}] \), corresponding to the transition state ensemble along the folding path. The structure of transition state ensemble will describes the folding mechanism and control the folding kinetics to overcome the free energy barrier between the globule and native states.

Alternatively, the structure at the saddle-points of the variational free energy surface can be characterized by a Gaussian measure to the native structure. We refer
as native density:

\[
\rho_i = \left\langle \exp \left[ -\frac{3}{2a^2} \alpha^N (r_i^N - r_i^N)^2 \right] \right\rangle_0
= (1 + \alpha^N G_{ii})^{-3/2} \exp \left[ -\frac{3}{2a^2} \alpha^N (s_i - r_i^N)^2 \right].
\] (3.33)

This measure of the monomer density relative to the native position is the order parameter which is employed in the study of the dynamics of the barrier crossing (discussed in Sect. 3.4). The degree of native structure at the transition state can be characterized by the normalized native density:

\[
\bar{\rho}_i = \frac{\rho_i - \rho_i^G}{\rho_i^N - \rho_i^G}
\] (3.35)

here the superscripts G and N denote the fluctuations evaluated at the globule and native state, respectively.

The native density and temperature factors of each residue are local order parameters that can describe local structure. However, in order to study the folding transition from globule (unfolded) state to native (folded) state, we also need a global order parameter which reflects the progress towards the native state. One way to define the global order parameter is the average of the normalized native density,

\[
Q = \frac{1}{N_{chain}} \sum_i \bar{\rho}_i,
\] (3.36)

where \(N_{chain}\) is the chain length. This progress coordinate ranges from \(Q = 0\) (globule) to \(Q = 1\) (native).
3.3 Analysis of folding route

We calculate the transition states involved in the folding by searching for saddle-points in $F\{C_i\}$ using an eigenvector-following algorithm [151]. This algorithm is similar to Newton’s method for optimization, but involves diagonalizing the Hessian matrix, $\partial^2 F/\partial C_i \partial C_j$, at each iteration. In this routine, the point is updated by stepping in a direction to maximize along the eigenvector with lowest eigenvalue and minimize along all others. To find a minimum, a step is taken to minimize along all eigenvectors of the Hessian. In order to use this algorithm, we need to be able to differentiate the free energy with respect to $\{C_i\}$, $\partial \alpha F = \partial F / \partial C_\alpha$. These derivatives can be easily computed by the chain rule using the elementary derivatives $\partial \alpha G_{ij} = -G_{ia}G_{aj}$ and $\partial \alpha (\log \det G) = -G_{aa}$.

The average folding route is characterized by a series of connected saddle-points and local minima in our variational theory. These average pathways are found as follows:

We first identify globule and native states by the local minima with the largest and smallest entropy, respectively. These are easy to identify, because the globule is the only stable minimum at high temperature and the native is the only one at low temperature. These minima can be used as the initial guesses for the optimization algorithm for incremental temperature changes until we have these minima at the same temperature. Using linear combinations of these two sets of constraints as initial guesses, we search for a saddle-point. From this saddle-point, we perturb the set of constraints $\{C_i\}$ along the unstable eigenvector and use the eigenvector following
algorithm with a small step size to find the closest minimum. This gives two local minima, one for each direction on the unstable eigenvector, connected by the saddle-point. This process is repeated until the globule and native state are connected by a series of local minima and saddle-points. We identify this connected sequence as the average folding route, characterizing the transition states and local minima that are important in the folding kinetics.

3.4 Barrier crossing dynamics

In this part, we review folding rate theory for minimally frustrated proteins by analyzing the microscopic dynamics of chain motions involved in barrier crossing of the protein folding model presented in Portman, Takada, Wolynes [79–81]. Multidimensional, non-Markovian Langevin dynamics are used to describe the barrier crossing, and identify the unstable mode as the reaction coordinate. For the folding rate calculation, we use a generalization of Kramer’s theory [152] that treats non-Markovian dynamics in many dimensions [153]. At the same time, the interpretation of a local reaction coordinate follows Langer’s theory of the nucleation rates [154]. The details of chain dynamics enter the formalism through a generalized friction matrix.

To calculate the barrier crossing rate, we study the relaxation rates for local order parameters that describe the phases of the system, and then study their motions. In particular, we consider the dynamics of the local native density as:

\[ \rho(r_i) = \exp\left[ -\frac{3}{2a^2} \alpha^N (r_i - r_i^N)^2 \right]. \] (3.37)

Following the interpretation of the order parameters given in Sect. 2.3, the average
native density of site i can be calculated as, $\rho_i[\{C\}] = \langle \rho(r_i) \rangle_0$, for a given set of variational constraints $\{C\}$, we will have:

$$\rho_i[\{C\}] = (1 + \alpha^N G_{ii})^{-3/2} \exp \left[ -\frac{3 \alpha^N (s_i - r_i^N)^2}{2d^2 \left( 1 + \alpha^N G_{ii} \right)} \right].$$

(3.38)

So, $\rho_i$ is a function of $\{C_i\}$ through the correlations, $G_{ij}$, and average position, $s_i$. The free energy can then be parameterized by $\{\rho_i\}$ with $\bar{F}[\{\rho\}] = F[\{C\}]$, where $\{\rho\}$ is evaluated at $\{C\}$. In particular, we denote a saddle-point of $F[\{C\}]$ by $\{C^*\}$ and the corresponding native density by $\rho_i^* = \rho_i[\{C^*\}]$.

We assume Generalized Langevin Equation for barrier crossing dynamics with memory for the native density, $\{\rho_i(t)\}$ is:

$$\partial_t \rho_i(t) = -\sum_j \int_0^t dt' \mu_{ij}(t-t') \frac{\partial \beta \bar{F}[\{\rho(t')\}]}{\partial \rho_j} + \xi_i(t),$$

(3.39)

where, $\beta = 1/k_B T$ is the inverse of temperature, and $\mu(t)$ is a generalized mobility matrix related to the random noise $\xi_i(t)$ through the correlations $\langle \xi_i(t)\xi_j(t') \rangle = \mu_{ij}(t-t')$. Here, we assume that $\mu(t)$ is a known function of time. We will discuss $\mu(t)$ in detail separately in next section.

To study the dynamics near the saddle-point, we expand $\bar{F}[\{\rho\}]$ about $\rho_i^*$ to second order:

$$\beta \bar{F}[\{\rho\}] \approx \text{const} + \frac{1}{2} \sum_{ij} \bar{\Gamma}^*_{ij} \delta \rho_i \delta \rho_j,$$

(3.40)

where, $\delta \rho_i = \rho_i - \rho_i^*$, and the Hessian matrix:

$$\bar{\Gamma}^*_{ij} = \frac{\partial^2 \beta \bar{F}[\{\rho^*\}]}{\partial \rho_i \partial \rho_j}$$

(3.41)

has one negative eigenvalue since it is evaluated at a saddle-point. In terms of derivatives with respect to the variational parameters, $\bar{\Gamma}^*$ is found by solving the chain
rule:
\[
\sum_{kl} \Gamma^{*}_{kl} \frac{\partial \rho^*_k}{\partial C_i} \frac{\partial \rho^*_l}{\partial C_j} = \frac{\partial^2 F[\{C\}]}{\partial C_i \partial C_j}
\]  (3.42)

where the derivatives with respect to \( C_i \) can be calculated as described in Sect. 3.2:

With this approximation, Eq. 3.39 can be written as:

\[
\partial_t (\delta \rho_i(t)) = -\sum_j \int_0^t dt' \mu(t - t') \Gamma^*_{ij} \delta \rho_j(t') + \xi_i(t). \quad (3.43)
\]

Accordingly, the native density correlation functions satisfy

\[
\partial_t C(t) = -\int_0^t dt' \mu(t - t') \Gamma^* C(t'). \quad (3.44)
\]

In Eq. 3.39, we have subtracted the stationary value \( \rho_i(t \rightarrow \infty) = \rho^*_i \), so that \( \delta \rho_i(t \rightarrow \infty) = 0 \) and \( C(t \rightarrow \infty) = 0 \). This long time solution is unstable, however, because it is a saddle-point of the free energy.

The unstable mode can be determined by the average equation of motion [155].

Denoting the Laplace transform of an arbitrary function of time by \( \hat{g}(\omega) = \int_0^\infty dt e^{-\omega t} g(t) \), the solution of the Eq. 3.39 averaged over the random noise is:

\[
\overline{\delta \rho_i(\omega)} = [\omega I + \hat{\mu}(\omega) \Gamma^*]^{-1} \delta \rho_j(0). \quad (3.45)
\]

By inverting the Laplace transform, \( \overline{\rho_i(t)} \) can be expressed as a sum of exponentials with time constants determined by the poles of \( [\omega I + \hat{\mu}(\omega) \Gamma^*]^{-1} \): The eigenfunctions \( \hat{\mu}(-\kappa) \Gamma^* \cdot \mathbf{u} = \kappa \mathbf{u} \) have time dependence \( \mathbf{u}(t) \sim e^{-\kappa t} \). Assuming that \( \hat{\mu}(\omega) \) is positive definite, there is one mode for which \( \kappa \) is negative, because \( \Gamma \) is the saddle-point Hessian of \( F[\rho] \):

\[
\mathbf{u}^*(t) = \mathbf{u}^* e^{\kappa^* t} \quad (3.46)
\]
where

\[
\dot{\mu}(-\kappa^*)\bar{\Gamma}^* \cdot \mathbf{u}_j^* = \kappa^* \mathbf{u}_j^* \quad (\kappa^* < 0).
\] (3.47)

Since motion along the unstable mode \(\mathbf{u}^*\) grows exponentially away from the saddle point, we identify the components of \(\mathbf{u}^*\) as the local reaction coordinate to surmount the barrier. Both \(\mathbf{u}^*\) and \(\kappa^*\) are essentially the unstable mode and curvature of \(\bar{F}[\rho]\), but renormalized by the dynamics of the barrier crossing incorporated in the mobility matrix. In Eq. 3.47, \(\kappa^*\) is the multi-dimensional analogue of the Grote-Hynes frequency (in the over-damped limit) [156].

The rate for barrier crossing corresponding to Eq. 3.39 is given by [153]:

\[
k = \frac{|\kappa^*|}{2\pi} \left| \frac{\det \bar{\Gamma}_{MS}}{\det \bar{\Gamma}^*} \right|^{1/2} \exp^{-\beta \Delta F^\dagger},
\] (3.48)

where \(\Delta F^\dagger\) is the barrier height and \(\bar{\Gamma}_{MS}\) and \(\bar{\Gamma}^*\) are the curvature matrices of \(F[\rho]\) evaluated at the metastable minimum and the saddle-point, respectively. Eq. 3.48 generalizes both the rate calculations presented by Langer [155] and Grote-Hynes [156]; it simultaneously accommodates both multi-dimensional diffusion (as in the Langer formula) as well as time dependent friction (as in the Grote-Hynes formula).

The ratio of determinants accounts for the entropic differences between the transition state and the metastable phase due to fluctuations of the order parameter. These contributions would already be included in an exact free energy so that the determinants should be absorbed into the exponential factor to most simply keep a consistent level of thermodynamic theory [152]. Consequently, the expression for the
rate becomes:

\[ k = \frac{|\kappa^*|}{2\pi} \exp^{-\beta \Delta F^*}, \]  

(3.49)

Here, \( \kappa^* \) is given by Eq. 3.47. As pointed out in Ref. 157, if the over-counting of entropy were not corrected, the ratio of the forward and backwards rate would not equal the equilibrium constant as determined by the starting free energy functional, \( k_{12}/k_{21} \neq \exp^{-\beta(F_2-F_1)} \).

3.5 Generalized mobility matrix

From the last section, we can see that barrier crossing dynamics is involved in the motion along unstable mode \( u^* \), which grows exponentially away from the saddle point. In order to solve Eq. 3.47, we need to specify the generalized mobility matrix \( \mu(t) \). In this section, I will describe how to determine \( \mu(t) \) by approximating the local dynamics with constrained polymer Hamiltonian \( H_0 \).

3.5.1 Native density correlation function \( C(t) \)

The native density function is defined as \( C_{ij}(t) = \langle \delta \rho_i(t) \delta \rho_j(0) \rangle \). For a general potential, the calculation of \( C_{ij}(t) \) is difficult. Here, we approximate the potential by the harmonically constrained polymer Hamiltonian \( H_0 \). This choice not only gives the reasonable interpretation of the order parameter as average over \( H_0 \), it also allows us to calculate the native density correlation function \( C_{ij}(t) \) directly from polymer dynamics of the partially folded chain.

Now we consider the native density correlation function \( C_{ij}(t) \) based on reference
Hamiltonian $H_0$:

$$C_{ij}(t) = \langle \delta \rho_i(t) \delta \rho_j(0) \rangle_0 = \langle \rho_i(t) \rho_j(0) \rangle_0 - \rho_i \rho_j. \quad (3.50)$$

the first term can be written as a function of the monomer pair correlation,

$$G_{ij}(t) = \langle \delta \mathbf{r}_i(t) \cdot \delta \mathbf{r}_j(0) \rangle_0/a^2. \quad (3.51)$$

Since $H_0$ is harmonic, the chain dynamics are contained in $G_{ij}(t)$ (discussed in next section). After a series of manipulation [81], the first term in Eq. 3.51 is:

$$\langle \rho_i(t) \rho_j(0) \rangle_0 = (\det M_{ij}(t))^{-3/2} \times \exp \left\{ -\frac{3\alpha^N}{2a^2} \mathbf{J}_{ij}^T \cdot M_{ij}(t)^{-1} \mathbf{J}_{ij} \right\}, \quad (3.52)$$

where

$$M_{ij}(t) = \begin{vmatrix} (1 + \alpha^N G_{ii}) & \alpha^N G_{ij}(t) \\ \alpha^N G_{ij}(t) & (1 + \alpha^N G_{jj}) \end{vmatrix}, \quad (3.53)$$

and $\mathbf{J}_{ij}$ is the two-component difference vector,

$$\mathbf{J}_{ij} = \begin{vmatrix} r_i^N - s_i \\ r_j^N - s_j \end{vmatrix}. \quad (3.54)$$

And the second term in Eq. 3.50 can be formed directly form Eq. 3.33.

We assume the native density correlations obey the same form of the generalized Langevin equation [158] as given in Eq. 3.43.

$$\partial_t C(t) = - \int_0^t dt' \mu^o(t - t') \Gamma^o C(t'), \quad (3.55)$$

where now $\Gamma^o$ are the inverse of the static correlations

$$[\Gamma^o]^{-1}_{ij} = \langle \delta \rho_i \delta \rho_j \rangle_0 = C_{ij}(0). \quad (3.56)$$
Since $C(t)$ is known (given in the Eq. 3.52 to Eq. 3.54), we can solve Eq. 3.55 for $\mu^\rho(t)$ by Laplace transforms giving,

$$\hat{\mu}^\rho(\omega) = C(0)\hat{\mathcal{C}}(\omega) - 1\hat{\mathcal{C}}(0) - \omega\hat{\mathcal{C}}(0).$$ (3.57)

Therefore, in order to get the mobility $\mu(t)$, we only need the Laplace transform of $C(t)$ which can be calculated numerically. Finally, we reduced the $\mu(t)$ to the native density correlations $C(t)$, which involves the chain dynamics governed by the constrained reference Hamiltonian, $H_0$.

3.5.2 Monomer pair correlation: $G(t)$

Now let us calculate the monomer pair correlation, $G(t)$, with the approximation to the chain potential $H_0$.

The monomer pair correlation is defined in Eq. 3.51 is assume to satisfy Smoluchowski equation

$$\partial_t G(t) = \sigma \tilde{\gamma}^{-1} \Gamma^{(0)} G(t),$$ (3.58)

where, $\Gamma^{(0)}$ is given in Eq. 3.14. And

$$\sigma = \frac{3k_B T}{a^2 \gamma_0} = \frac{3D_0}{a^2}$$ (3.59)

sets the time scale for the relaxation. Here, the dimensionless friction matrix $\tilde{\gamma}$ is defined through friction coefficient $\gamma$ and monomer diffusion coefficient $D$ according to

$$\beta D = \gamma^{-1}, \quad \gamma = \gamma_0 \tilde{\gamma}, \quad \gamma_0 = 6\pi a_{eff} \eta.$$ (3.60)

Here, $\gamma_0$ is defined by the Stokes law and $\eta$ is the solvent viscosity and $a_{eff}$ is typical monomer van der Waals radius. Since Eq. 3.58 is linear, it can be solved by
transforming to normal modes by a similarity transform

\[ Q^{-1} [\tilde{\gamma}^{-1} \Gamma^{(0)}] Q = \text{diag}\{\lambda_p\}, \]  

(3.61)

where the columns of \( Q \) are the right eigenvectors of \( [\tilde{\gamma}^{-1} \Gamma^{(0)}] \). Usually, \( Q \) can diagonalize \( \tilde{\gamma} \) and \( \Gamma^{(0)} \) separately, such as:

\[ Q^T \tilde{\gamma} Q = \text{diag}\{\nu_p\}, \quad Q^T \Gamma^{(0)} Q = \text{diag}\{\eta_p\}, \]  

(3.62)

with \( \lambda_p = \eta_p/\nu_p \). Here we take the friction to be diagonal in the monomer index, \( \tilde{\gamma}_{ij} = \delta_{ij} \), so that \( \nu_p = 1 \) and \( \eta_p = \lambda_p \).

The correlation between monomers \( i \) and \( j \) can be expanded onto the normal modes as,

\[ G_{ij}(t) = \sum_p \frac{1}{\eta_p} Q_{pi} Q^T_{jp} \exp(-\sigma \lambda_p t). \]  

(3.63)

This can be used in Eq. 3.52 to Eq. 3.54 and Eq. 3.57 to close the expression for the mobility matrix \( \mu(t) \).

3.6 Model Parameters

To apply the theory, the parameters that describe the interaction potential between residues and the polymer chain characteristics need to be specified. Some parameters are already pointed out in previous parts, like stiffness \( g \). We take \( g = 0.8 \), which corresponds to the persistence length for poly-L-alanine [146], \( l \sim 20 \text{Å} \). For native contacts, spatial distance is defined as 6.5Å and also sequence criterion is at least four monomers in sequence. For potential parameters, we find the intermediate- and long-ranged interaction parameters \((\gamma, \alpha; \gamma_l, \alpha_l) = (9, 0.54, 6.0, 0.27)\) give an effective
potential well that contains all native contacts distances and has a minimum at the most probable $C_\alpha - C_\alpha$ contact distance, $r^* = 1.6a$, with $u(r^*) = -1$ (see Fig. 3).

The short-ranged interaction represents the hard-core repulsion between residues and gives excluded volume; with the choice of values $(\gamma_s, \alpha_s) = (25.0, 3.0)$, the repulsion roughly balances the attractive energy in the globule state allowing us to study a folding transition that occurs directly from a random coil.

3.7 Conclusion

In this chapter, we mainly review the variational theory for protein folding free energy surface developed by Portman, Takada, Wolynes [79–81]. Within this theory, the sidechain and solvent degree of freedom were ignored, also non-native contacts and trapping effects were neglected, too. However, through this simple microscopic theory of folding rates for completely minimally frustrated proteins, we can analytically calculate structure of transition state ensembles, free energy barriers, and dynamical prefactors which involved the chain motions. It is very convenient to explore the folding events with the simplest model which can give a physically reasonable and direct picture of folding route for two-state proteins. Therefore normal systematic statistical errors arising from simulations are avoided. For example, the recent studies have shown that this variational model is reliable in predicting the structure of the transition state ensemble of individual proteins at the residue level of resolution [79, 143, 144].
CHAPTER 4

THE EFFECT OF COOPERATIVITY ON TWO-STATE PROTEIN FOLDING

4.1 Cooperativity

As discussed in the first three Chapters, simple models have been particularly useful for shaping our understanding of protein folding. However, it is also pretty clear that theory needs further quantitative development, and possibly even qualitative improvements, to compare directly with experimental measurements. One significant quantitative issue is that many simple theories predict a range of free energy barriers of folding that are much smaller than expected from experiments for two-state proteins. This small range of barriers directly results in a small range of folding rates. For instance, Koga and Takada [64], by using a Gō-like potential for 18 small single-domain proteins, obtained a correlation between contact order and the simulated folding rates, but the variation in folding rates covered only $\approx 1.5$ orders of magnitude. The barrier heights are within $\sim 5k_B T$ of each other. This is a far cry from the six orders of magnitude of variation in folding rates observed among real, small, single-domain proteins [127, 159]. More recently, Jewett et al [67] conducted an extensive lattice 27mer simulation study using Gō-like models. A correlation between contact order and simulated folding rates was found again, but the dispersion in folding rates spanned only 1 to 1.5 order of magnitude. Later, Wallin and Chan [63] simulated folding rates for 13 small proteins by applying a continuum Gō-like Cα model. Their
simulated rates spanned only approximately two orders of magnitude. We can see that simulated rates through Gō-like model are much less diverse than the corresponding experimental rates that cover more than six orders of magnitude.

In this thesis, I wish to investigate this significant discrepancy in folding rate diversity between experiments and the explicit chain model prediction. A possible candidate for the origins of the discrepancy is low cooperativity. Many of these Gō-like models have pairwise additive forces which may fail to mimic sufficiently high degrees of cooperativity [64, 67] in real two-state proteins [68, 69]. Indeed, consistent with this general assessment, a local-nonlocal coupling lattice model interaction scheme that enhances cooperativity has been shown to significantly increase the diversity in model folding rates [71].

Considering above analysis as a whole, these findings suggest that cooperativity and non-additive many-body interactions are likely to be critical in accounting for the tremendous diversity among experimental folding rates.

Then the question is what is the origin of cooperativity? We have to closely look at these simplified models. As in the Gō or Gō-like model, only native interactions are taken into account. As Eastwood and Wolynes [70] pointed out, when using reduced descriptions, we actually integrate out degree of freedom, such as the sidechain degree of freedom and the solvent degree of freedom. For example, when one native contact pair is formed, it should be easier to bring another nearby residue to form a contact pair since the entropic cost orienting the third residue has already been accounted for. Thus the interactions between residues is clearly not a two-body sum, but an
intrinsically multi-body interaction.

In the present analytic model, cooperativity is introduced through repulsive excluded volume interactions between residues in proximity to native contact pairs. This potential is effectively "neutral" because it primarily destabilizes partially ordered residues at the interface of the folding nucleus. The cooperativity term of the potential is pairwise additive in the space of all contacts, but it corresponds to an effective multibody potential when projected onto the set of native contacts. The particular form of cooperativity was developed so that the calculated barrier heights remain robust with respect to variations of excluded volume strength in the original variational model [79, 80].

In our model, enhanced cooperativity is realized by the addition of repulsive interactions between nonnative contacts. In particular, I modify the interactions between native contacts through a repulsive potential between residues in close proximity to native contacts. $H_{int}^{coop}$ can be written as:

$$H_{int}^{coop} = \sum_{kl} \epsilon_0 u^{coop}(r_{kl})$$ (4.1)

and

$$u^{coop}(r_{kl})/\epsilon_0 = U(0) \exp(-\alpha_s r_{kl}^2).$$ (4.2)

Here, we define nonnative contact pairs in following way: for every native contact pair $(i, j)$ with $|i - j| \geq 12$, we include pairs within a window $[i \pm 4, j \pm 4]$ and eliminate duplicates or native contact pairs from the sum. $U(0)$ is the excluded volume strength at zero distance. The finite strength of the repulsion at $r = 0$ is
an artifact of the potential (and the finite native monomer density at short range). So there is some ambiguity in determining the appropriate value for $U(0)$. This is troubling because it was found that the calculated barrier height is sensitive to the value of $U(0)$, even though the structure of the transition state ensemble is relatively robust for most proteins. This sensitivity of barrier height on the excluded volume strength $U(0)$ indicates that the cooperativity in the model is relatively low. So increasing cooperativity by destabilizing partially ordered residues makes the barrier height less sensitive to $U(0)$.

Now we have two kinds of interactions, the Hamiltonian for the polymer chain can be rewritten as:

$$H = H_{\text{chain}} + H_{\text{Nat}}^{\text{int}} + H_{\text{coop}}^{\text{int}}.$$  \hspace{1cm} (4.3)

The energy becomes

$$E[\{C\}] = \sum_{ij}^{\text{Nat}} \epsilon_{ij} u_{ij} + \sum_{mn}^{\text{coop}} u_{mn}^{\text{coop}}.$$  \hspace{1cm} (4.4)

with:

$$u_{ij} = < u(|r_{ij}|) >_0$$  \hspace{1cm} (4.5)

$$= \sum_{k=(s,s,l)} \frac{\gamma_k}{(1 + \alpha_k \delta_G G_{ij})^{3/2}} \exp \left[ -\frac{3}{2} \frac{\alpha_k (s_i - s_j)^2}{1 + \alpha_k \delta_G G_{ij}} \right].$$  \hspace{1cm} (4.6)
\[ w_{\text{coop}} = < w^{\text{coop}}(|r_{mn}|) >_0 \]

\[ = \frac{\epsilon_0 U(0)}{(1 + \alpha_s \delta G_{mn})^{3/2}} \exp \left[ -\frac{3}{2a^2} \frac{\alpha_s (s_m - s_n)^2}{1 + \alpha_s \delta G_{mn}} \right]. \]  

\[ (4.7) \]

\[ (4.8) \]

With this form of cooperative interaction, the barrier height (\( \Delta F^\dagger/k_B T \)) is relatively insensitive to the excluded volume strength, \( U(0) \). This can be understood by considering the short-distance pair density of a partially ordered chain, \( n_{ij}(r) = \langle \delta(r - r_{ij}) \rangle_0 \).

Integrating over angles gives the radial pair density

\[ n_{ij}(r) \sim \frac{1}{\sqrt{a^2 \delta G_{ij} s_{ij}}} \frac{r}{s_{ij}} \sinh \left( \frac{3r s_{ij}}{a^2 \delta G_{ij}} \right) \exp \left( -\frac{3}{2} \frac{r^2 + s_{ij}^2}{a^2 \delta G_{ij}} \right). \]

\[ (4.9) \]

where the correlations \( G_{ij} = \langle \delta r_i \cdot \delta r_j \rangle_0 / a^2 \) and \( \delta G_{ij} = G_{ii} + G_{jj} - 2G_{ij} \) is the magnitude of the fluctuations about the relative mean separation \( s_{ij} = \sum_k (G_{ik} - G_{jk}) C_k r_k^N \).

The weight at short distances \( (r < r_0, r_0 \text{ is excluded volume interaction length scale}) \) is small when the pair is sufficiently delocalized \((r_0 \ll a \sqrt{\delta G_{ij}})\) or sufficiently localized \((a \sqrt{\delta G_{ij}} \ll s_{ij})\). Consequently, the sensitivity of the barrier height on the short distance repulsion is a consequence of the partially ordered residues in the transition state ensemble. Increasing the cooperativity by destabilizing partially ordered residues makes the barrier height less sensitive to \( U(0) \).

To this end, I choose the strength of the cooperativity interaction as the finite strength of the repulsion at \( r = 0, U(0) \). There is still some ambiguity in determining the appropriate value of \( U(0) \). The barrier height is relatively insensitive to the value of \( U(0) \) with \( \Delta F^\dagger \) varying less than about \( 1 - 2k_BT \) over a wide range of \( U(0) \) \((5 \leq U(0) \leq 60)\). I take \( U(0) = 50 \). The entropy is the same as given by Eq. 3.27.
and the free energy surface will follow Eq. 3.31 of a partially ordered chain. The critical points define an average folding route by connecting a series of minimum-saddle point-minimum from globule(unfolded) state to native(folded) state.

I will address two main issues: the first issue is how the folding rates are affected by the nonnative cooperativity interactions. The second issue addressed is the effect nonnative cooperativity interactions on the transition state ensemble structure at the residue level of resolution. For example, I investigate the specific structure change under considering cooperativity interaction comparing the case without cooperativity.

4.2 Cooperativity effect on folding rate.

The most accessible quantitative observables of two-state proteins are the folding, unfolding rates. As rates and free energies are the natural quantitative experimental measurements, relative or absolute prediction of these quantities is necessary for a direct connection to experiment and a true assessment of theoretical method. [89] According to this general principle, I will directly calculate free energy barrier heights and prefactors, and hence the folding rates for 28 two-state proteins (As shown in Table. 1). As we know, folding in small proteins is often well characterized as a transition between two well-defined structural populations, an unstructured globule ensemble and a structured folded ensemble. The transition rate between these local minima of the free energy is controlled by the dynamics of passing through an unstable transition region determined by saddle points in the free energy surface. Accordingly,
the rate is expected to follow Arrhenius form:

$$k_f = k_0 e^{-\beta \Delta F^\dagger},$$

(4.10)

where $\beta = 1/k_B T$ is the inverse temperature and $\Delta F^\dagger$ is the free energy difference between the unfolded and transition state ensembles. The exponential factor in Eq. (4.10) reflects the equilibrium population of the transition state ensemble relative to unfolded ensemble and the prefactor, $k_0$, is the timescale associated with the dynamics of crossing the free energy barrier. Successful identification of specific residues structured in the transition state ensemble by several different theoretical models [79–81, 130, 131, 139–141] and numerous simulation studies (see, Ref. 53, 64, 160 and references therein) have established that the topology of the native structure determines the folding mechanism of these proteins. In addition, two-state folding rates are well correlated with very simple measures of the native state topology such as contact order [1, 2, 59, 62]. While additive potentials often produce reasonable structural characterization of the transition state ensemble, the range of predicted folding rates and their relationship with contact order does not agree with experiment [64].

Experimental evidence supporting a specific decomposition of the Arrhenius rate (Eq. 4.10) into dynamic and thermodynamic components necessarily indirect [161]. While structural predictions from models with a strong native state bias (Go-models) are robust, the value of the barrier height (and corresponding absolute timescale $1/k_0$) is more sensitive to details of the model [142, 162]. The prefactor is commonly assumed to be roughly uniform for different proteins with a magnitude of $O(0.1–1\mu s^{-1})$ [161, 163–165], though prefactors as large as $O(100\mu s^{-1})$ have also
Table 1: Two-state proteins I studied in this work. Kinetic data for proteins 1enh, 1vii, 1hdn were taken from Ref. 1; for the rest of proteins were taken from Ref. 2.

<table>
<thead>
<tr>
<th>#</th>
<th>pdb</th>
<th>class</th>
<th>protein name (reference)</th>
<th>$k_f (s^{-1})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1enh</td>
<td>α</td>
<td>Engrailed homeodomain (1)</td>
<td>37153.5</td>
</tr>
<tr>
<td>2</td>
<td>1imq</td>
<td>α</td>
<td>Colicin E9 immunity protein (2)</td>
<td>1479.2</td>
</tr>
<tr>
<td>3</td>
<td>1lmb</td>
<td>α</td>
<td>lambda-repressor (2)</td>
<td>4910.4</td>
</tr>
<tr>
<td>4</td>
<td>1vii</td>
<td>α</td>
<td>Villin head piece (1)</td>
<td>21877.6</td>
</tr>
<tr>
<td>5</td>
<td>2abd</td>
<td>α</td>
<td>ACBP (2)</td>
<td>734.6</td>
</tr>
<tr>
<td>6</td>
<td>2pdd</td>
<td>α</td>
<td>E3/E1 binding domain of dihydrolipoyl acyltransferase (2)</td>
<td>18015.4</td>
</tr>
<tr>
<td>7</td>
<td>1a0n</td>
<td>β</td>
<td>SH3 domain (Fyn) (2)</td>
<td>90.0</td>
</tr>
<tr>
<td>8</td>
<td>1csp</td>
<td>β</td>
<td>CspB (Bacillus subtilis) (2)</td>
<td>664.694</td>
</tr>
<tr>
<td>9</td>
<td>1fnf9</td>
<td>β</td>
<td>Fibronectin 9th FN3 module (2)</td>
<td>0.4</td>
</tr>
<tr>
<td>10</td>
<td>1mef</td>
<td>β</td>
<td>CspA (2)</td>
<td>200.2</td>
</tr>
<tr>
<td>11</td>
<td>1pks</td>
<td>β</td>
<td>SH3 domain (PI3 kinase) (2)</td>
<td>0.3</td>
</tr>
<tr>
<td>12</td>
<td>1shg</td>
<td>β</td>
<td>SH3 domain (alpha-spectrin) (2)</td>
<td>4.1</td>
</tr>
<tr>
<td>13</td>
<td>1srl</td>
<td>β</td>
<td>SH3 domain (src) (2)</td>
<td>54.6</td>
</tr>
<tr>
<td>14</td>
<td>1ten</td>
<td>β</td>
<td>Tenascin (2)</td>
<td>3.0</td>
</tr>
<tr>
<td>15</td>
<td>1wit</td>
<td>β</td>
<td>Twitchin (2)</td>
<td>1.5</td>
</tr>
<tr>
<td>16</td>
<td>1pin</td>
<td>β</td>
<td>WW domain (2)</td>
<td>13346.6</td>
</tr>
<tr>
<td>17</td>
<td>1pgb16</td>
<td>β</td>
<td>beta-hairpin (41-56) of protein G (2)</td>
<td>162552.0</td>
</tr>
<tr>
<td>18</td>
<td>1psf</td>
<td>β</td>
<td>Photosystem I accessory protein (2)</td>
<td>24.5</td>
</tr>
<tr>
<td>19</td>
<td>1aps</td>
<td>αβ</td>
<td>Muscle Acp (2)</td>
<td>0.2</td>
</tr>
<tr>
<td>20</td>
<td>1c8c</td>
<td>αβ</td>
<td>DNA-binding protein sso7D (2)</td>
<td>1095.8</td>
</tr>
<tr>
<td>21</td>
<td>1coa</td>
<td>αβ</td>
<td>CI2 (2)</td>
<td>49.4</td>
</tr>
<tr>
<td>22</td>
<td>1div</td>
<td>αβ</td>
<td>N-terminal domain from L9 (2)</td>
<td>445.6</td>
</tr>
<tr>
<td>23</td>
<td>1fkb</td>
<td>αβ</td>
<td>FKBPI2 (2)</td>
<td>4.5</td>
</tr>
<tr>
<td>24</td>
<td>1hdn</td>
<td>αβ</td>
<td>Hpr (his-containing phosphocarrier protein) (1)</td>
<td>14.8</td>
</tr>
<tr>
<td>25</td>
<td>1o6x</td>
<td>αβ</td>
<td>Activation domain procarboxypeptidase A2 (2)</td>
<td>897.2</td>
</tr>
<tr>
<td>26</td>
<td>1pgb</td>
<td>αβ</td>
<td>Protein G (2)</td>
<td>403.2</td>
</tr>
<tr>
<td>27</td>
<td>1urn</td>
<td>αβ</td>
<td>Splicesomal protein U1A (2)</td>
<td>330.1</td>
</tr>
<tr>
<td>28</td>
<td>2ptl</td>
<td>αβ</td>
<td>Protein L (2)</td>
<td>60.3</td>
</tr>
</tbody>
</table>
been proposed recently [162, 166]. While the precise value of the prefactor plays a sub-dominant role in determining the absolute rate, accurate estimates give an important reference timescale essential, for example, to identify the fastest measured rates as downhill (or barrier-less) folding [161, 167, 168]. It is very necessary to check this assumption by direct calculation of prefactor.

4.2.1 Folding rates and prefactors

Following the folding routes for each protein, I calculated the folding rates and prefactors for each protein. Prefactors in absolute units depend on the time scale set by monomer relaxation rate \( \sigma = 3D_0/a^2 \), where \( D_0 \) is the monomer diffusion constant and \( |a| = 3.8 \text{Å} \) is monomer spacing along the chain. I take \( \sigma \) as a fitting parameter. As shown in Fig. 4, the predicted and measured rates are well correlated \((r = 0.8)\) with agreement within an order of magnitude for 80% proteins.

The best fit monomer relaxation time \( 1/\sigma = 30\text{ns} \) is on the order of the timescale of unfolding a helical segment [169]. With this microscopic timescale, the longest relaxation time of a chain of 100 monomers is approximately \( \tau_R \sim O(10 \mu s) \) which compares well [170] with the timescale for the fastest collapse kinetics measured in proteins and polypeptides [171–173]. On the other hand, \( 1/\sigma \) is an order of magnitude slower than estimates obtained from an effective diffusion coefficient inferred from loop closure experiments of small peptides \((\sim 1\text{ns})\) and two orders of magnitude slower than estimates from bare diffusion coefficients of the monomer \((\sim 100\text{ ps})\) [174]. The source of small effective diffusion coefficients associated with simple
Figure 4: Comparison between experimental and calculated folding rates. The correlation coefficient is $r = 0.8$, with $p$-value $p = 4.4 \times 10^{-7}$. (circle: $\alpha$–protein, square: $\beta$–protein, and diamond: $\alpha\beta$–protein.). Here, x-ray structured proteins are: 1pgb, 1coa, 1csp, 1c8c, 1pin, 1mb, 1enh, 1fkb, 1urn, 1shg, 1ten, *1div, *1fnf9, *1pgb16; NMR structured proteins are: 1o6x, 1imq, 1srl, 2ptl, 1aps, 2abd, *1a0n, *2pdd, *1mef, *1psf, *1wit, *1psf, *1vii, *1hdn.(Proteins not used in $\Phi$-value analysis are indicated by * ).
Gaussian models is not fully understood [175–177]. Nevertheless, results from recent experiments on small peptides under different solvent conditions indicate that intra-chain interactions (that can be interpreted kinetically as a kind of internal friction) induce local activation barriers that renormalize the effective monomer diffusion coefficient [178, 179]. Although controversial, internal friction may explain why the speed limit of protein folding is fixed at $\sim O(0.1–1.0 \mu s)$ [180].

**Figure 5:** Histogram of the inverse folding prefactor $\tau_0 = 1/k_0$ for the 28 two-state proteins. Ordinate is number of proteins.
Even though the prefactor of each protein is calculated individually, its value in absolute units ultimately depend on the calculated barrier heights through the microscopic timescale $\sigma$. Relative prefactors, on the other hand, are independent of this fitting parameter. As shown in Fig. 5, the distribution of prefactors is relatively uniform, varying within a factor of 5 for most proteins. Using the fitted value for $\sigma$, $\tau_0 = 1/k_0$ varies mainly between $1\mu$s to $5\mu$s with an average $\tau_0 = 4\mu$s. Given this narrow distribution, it is not surprising that a uniform prefactor of $\bar{k}_0 = (4\mu s)^{-1}$ gives essentially the same correlation to the measured and predicted rates (data not shown).

Thus, direct calculation of the barrier crossing dynamics gives solid evidence supporting the common assumption that the folding rate prefactor is largely independent of topology. Recent work by Henry and Eaton also suggests the prefactor is relatively uniform across two-state folding proteins based on analysis of folding rates from a different set of analytic models [142]. The value for the average prefactor $\bar{k}_0 \sim 10^5 s^{-1}$ agrees within an order of magnitude with estimates based on semi-empirical and theoretical models [139, 181, 182] as well as analysis of thermodynamic data from differential scanning calorimetry [183]. This value also is consistent with the fastest measured rates $\sim 1\mu$s, if the timescale for downhill folding is approximated by the Arrhenius rate with a vanishing barrier [161, 168].

A closer look at the two proteins (1lmb and 1pks) with exceptionally small calculated prefactors reveals that in each case the unstable mode becomes degenerate at a stability near the transition midpoint. The structure of the transition ensemble changes sharply, though continuously, as the temperature crosses the degenerate
point. In particular, the curvature of the unstable mode (and consequently the calculated prefactor) sharply vanishes in a cusp catastrophe [184]. Away from these isolated temperatures, the prefactors return to the range exhibited by the majority of the proteins studied. Several of the proteins studied have similar rapid changes of the transition state as a function of temperature, occurring at temperature sufficiently far away from the midpoint so that the prefactor is relatively unaffected near $T_f$. In this high dimensional model, catastrophes can be generally expected as local minima and saddle-points merge at isolated temperatures. The shape of the calculated prefactor versus temperature is thus determined by these degenerate points. For example, even for a route with a single transition state, the meta-stable unfolded or folded minimum disappear in a fold catastrophe at the limit of stability (spinodal) for both low and high temperatures [185, 186]. If there are no other catastrophes, the calculated prefactor obtains a maximum at an intermediate temperature and vanishes at the spinodals. Near the maximum the prefactor varies much more slowly with temperature than near the spinodal. This generic shape of the prefactor is interesting since it can account qualitatively for non-linear dependence of the rate with stability (chevron turnovers) and may help indicate kinetic signatures anticipating the onset of downhill folding.

Nevertheless, interpreting these results requires some care. The harmonic expansion of the free energy is not expected to accurately reflect the global curvature of the free energy over $\sim F^\dagger \pm k_B T$ when the local curvature is very small. For these cases, it is likely that the formalism should be modified away from strictly local curvatures to
get accurate estimates of the prefactor. Even if the renormalized prefactor is found to be relatively constant, the rapid change of the order parameter at the transition state that accompanies a catastrophe could alone account for chevron roll-over, similar to the transition state switching mechanism proposed by Oliveberg. [187] The subtle variation of the prefactor and free energy barrier height with stability changing is investigated in Chapter. 7.

4.2.2 Free energy barrier height

As shown in Section. 4.2.1, the calculated folding rates correlate reasonably well with the experimental folding rates. Recall that we define a global order parameter, \( Q \), to characterize the whole folding events along folding path based on local order parameters in Eq. 3.36. As illustrated in Fig. 6 (the protein 1srl), it is clear that the free energy barrier increases significantly due to increased cooperativity induced by the nonnative interactions. This is similar to the predicted barriers for all 28 two-state proteins (see Appendix. 8.2). Fig. 7 shows the free energy barriers for 28 two-state proteins as function of a simple topology parameter, absolute contact order (ACO) [2]. The range of calculated barrier heights span \( \sim 14k_B T_f \), in agreement with the measured range of rates given a uniform prefactor. In contrast, the range of barriers for the noncooperative routes spans only \( \sim 5k_B T_f \). Interestingly, this compressed range of barrier is just same as determined through coarse-grained Gō-model simulations. [63–65].

Fig. 7 also shows that the calculated barrier heights are highly correlated \( (r = 0.9) \)
Figure 6: Free energy profile as a function of global order parameter $Q$ given by Eq. 3.36: the red curve is the folding route with cooperativity, and the black curve is the route without cooperativity. The circles denote the critical points defining the folding route, and the curves are the steepest-descents paths.
with ACO when the cooperativity term is included in the model. The barrier heights calculated without cooperativity do not show significant correlation with ACO \((r = 0.41)\). This observation indicates that the relationship between native topology and folding rates is sensitive to the rigidity of the folding nucleus. This may in fact be a robust result, largely independent of the details of reasonable potentials that increase local cooperativity between native contacts \([72]\). Furthermore the low correlation between contact order and barrier heights of noncooperative routes is also reminiscent of results from Gö-model simulations \([64]\). Together, these results suggest that the cooperativity of typical Gö-model simulations based on two-body pair potential is too low \([67, 68, 72, 188]\).

As mentioned before, several simple topological parameters (such as RCO, LRO, TCD, described in Chapter 2) have been proposed. Logarithm of measured folding rates of small, two-state proteins correlate with these parameters as well. Logarithm of simulated folding rates from Gö-models \([63, 64]\) also show certain correlation with topological parameters, but span a much smaller range than measured rates due to a smaller range of free energy barriers if the prefactor is assumed as a uniform constant. Our results, based on an analytic model, give us not only similar correlation between barrier heights (and so for logarithm of folding rates according to \(k_f = k_0 \exp(-\beta \Delta F^{\dagger})\) and topological parameter ACO), but also give the correct range of folding rates.
Figure 7: Free energy barrier $F^+/k_B T_f$ plotted against the absolute contact order, $ACO = 1/N_{con} \sum |i - j|$ where the sum is over the $N_{con}$ native contacts pairs. Red points correspond to cooperative folding routes, and black points correspond to non-cooperative routes. Symbols have the same meaning as Fig. 4. For the cooperative routes, the correlation coefficient is $r = 0.91$ with p-value $p = 1.6 \times 10^{-11}$; for non-cooperative routes, the correlation coefficient if $r = 0.41$ with p-value $p = 0.03$. 
4.3 Cooperativity effect on transition state structure.

A complete theory of the folding mechanism must reliably predict structural properties of the transition state ensemble in addition to absolute folding rates. Recall that we define the degree of localization about the native positions $\rho_i$ in Chapter 3 to characterize the local order along the folding route:

$$\rho_i = \left\langle \exp \left[ -\frac{3}{2a^2} \alpha^N (r_i - r^N_i)^2 \right] \right\rangle_0.$$ 

For convenience, we normalized $\rho_i$ by native density in globule state $\rho_i^G$ and native state $\rho_i^N$: $\bar{\rho}_i = (\rho_i - \rho_i^G)/(\rho_i^N - \rho_i^G)$. As an example, I illustrated cooperativity effects on the structure of transition state ensemble in Fig. 8. Here I show the 2D-plot of normalized native density for Src tyrosine kinase SH3 domain (Protein Data Bank ID code:1srl) along the folding route (similar 2D plots for the other proteins are shown in the Appendix. 8.3). The color in Fig. 8 denotes the degree of order for each residue. Comparing Fig. 8(a) with cooperativity and Fig. 8(b) without cooperativity, it is clear that the coarse-grained structures of the transition state ensembles are basically similar for this protein, but the residues order more gradually in the noncooperative case. Still, even for the cooperative route, the interface has a finite width as the structural ensembles retain some partial ordering of the residues. The sharper interface of the cooperative case is just expected. The partially ordered structure is destabilized in favor of either ordered state or disordered state. So cooperativity narrows the interface. This change of the structure of the transition state ensemble is complicated to describe in general. Whether a particular intrefacial residue is excluded or incorporated into the folding nucleus is a subtle question, determined by the delicate balance between changes in entropy and energy due to localization. We
define a cross-correlation function $\Omega$ to characterize changes in local structural order.

$$\Omega = \hat{\rho}_{coop} \cdot \hat{\rho}_{noncoop},$$  \hspace{1cm} (4.11)

Where, $\hat{\rho}_{coop}$ and $\hat{\rho}_{noncoop}$ denote unit vectors with elements $\rho_i[\{C\}]$ for transition state ensembles with and without cooperativity, respectively. Fig. 9 shows the value $\Omega$ for each protein as well as a typical example of the overlap of native densities evaluated at $Q^\star$. For 80% proteins studied, the overlap between the transition state ensemble structures is greater than 60%. Nevertheless, the variation of $\Omega$ indicates this form of cooperativity does not effect every protein uniformly.

Figure 8: (color) Folding route for protein 1srl characterized locally by the normalized native density $\tilde{\rho}_i = (\rho_i - \rho_i(G))/(\rho_i(N) - \rho_i(G))$ and the global progress coordinate $Q = 1/N \sum \tilde{\rho}_i$, (a)The local and global structure along the folding route with cooperativity (left figure); (b) and without cooperativity (right figure). The degree of structural localization of each residue is reflected in the colors, linearly scaled between red ($\tilde{\rho}_i = 0$) to blue ($\tilde{\rho}_i = 1$).

Changes of the transition state ensemble can be also be characterized by the variation of the global order parameter $\Delta Q^\star$. As shown in Fig. 9, the majority of the proteins studied have $|\Delta Q^\star| \leq 0.1$. In terms of global order, the $\alpha$-helical proteins are not very sensitive to cooperativity, though the local structure of the transition
state ensemble can change significantly. For β and α/β proteins, some systematic errors in the calculated barrier height can be associated with relatively large changes in the global order. For proteins with $\Delta Q^* > 0.1$ (1urn, 1c8c, 1psf, 1csp), the model overestimates the barrier heights, while for proteins with $\Delta Q^* < 0.1$ (1pgb, 1a0n, 1coa, 1shg) the model underestimates the barrier height (See Fig. 4). This trend may be particular to form of cooperativity used in this model.

Experimentally, Alan Fersht pioneered the Φ-value analysis method by first applying it to the small bacterial protein barnase [51, 189] to study structure analysis of folding transition state ensembles. This method uses the free-energy perturbation of the transition state upon point mutation ($\Delta \Delta G^\dagger$) to estimate the extent of order of the mutation site in the TSE according to:

$$\Phi = \frac{\Delta \Delta G^\dagger}{\Delta \Delta G_{U-N}}, \quad (4.12)$$

where, $\Delta \Delta G_{U-N}$ is the accompanying change of the native-state stability.

Practically, rate constants are measured, so we can rewrite Eq. 4.12 as:

$$\Phi = \frac{\Delta \log k_f}{\Delta \log K_{U-N}} = \frac{\Delta \log k_f}{(\Delta \log k_u - \Delta \log k_f)}, \quad (4.13)$$

where $K_{U-N}$ is equilibrium constant, the ratio between the backwards rate ($k_u$) and forward rate constant ($k_f$), $K_{U-N} = k_u/k_f$. The mutation-induced changes of $\log k_f$ and $\log k_u$ are obtained directly from the shifts of the refolding and unfolding limbs of the chevron plot. The kinetics under various thermodynamics conditions produces characteristically V-shaped plots of $\log k_f$ and $\log k_u$ versus denaturant concentration, called chevron plots. As illustrated in Fig. 10, $\Phi \sim 1$ indicates that the mutated site
Figure 9: (color) The change in the local structure of the transition state ensemble characterized by the cross-correlation coefficient $\Omega$ plotted against the change in the global order parameter $\Delta Q^*$ is shown in the center. Left and right panels give the typical examples of the native density profile for various values of overlap $\Omega$. In each example, the red line corresponds to the cooperative route, and, the black line corresponds to the non-cooperative route. The corresponding correlation coefficient between measured and predicted $\Phi$-value are: 1pin (0.77), 1lm b (0.84), 1enh (0.87); 2ptl (0.0), 1imq (-0.1), 1o6x (-0.2).
experiences a fully native-like environment in the wild-type transition state ensemble, while $\Phi \sim 0$ indicates that the mutated site encounters as disordered an environment in the transition state ensemble as it is does in the unfolded state ensemble. For two-state proteins, there are only a few cases where $\Phi = 1$ which were observed. Instead the $\Phi$ values for two-state proteins are generally observed to be fractional [129, 190, 191].

Experimental $\Phi$-analysis can shed light on the local structure of the transition state ensemble. Many groups predict $\Phi$-values from detailed computer models to compare directly to experiments [8, 106, 192–197]. On a residue-based analysis, the correlation of the off-lattice predictions with experimental $\Phi$ values is only moderate, typically around 0.56. Recently, Plotkin et al. [72] raises the correlation to 0.78 after considering three-body interactions for Chymotrypsin inhibitor 2 (1coa). Koga and Takada [64] studied 18 small two-state proteins, they found that half of the cases have quite good correlations with experimental $\Phi$ values, but for a few cases, they did find there are large discrepancies between their simulations and experimental $\Phi$-values.

Table 2: Correlation for theoretical and experimental $\Phi$ values for x-ray crystallography proteins. $R$ is correlation coefficient for the case without cooperativity effects. $R^{coop}$ is correlation coefficient for the case with cooperativity effects.

<table>
<thead>
<tr>
<th>pdb</th>
<th>1pgb</th>
<th>1coa</th>
<th>1csp</th>
<th>1c8c</th>
<th>1pin</th>
<th>1lmb</th>
<th>1enh</th>
<th>1fkb</th>
<th>1urn</th>
<th>1shg</th>
<th>1ten</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^{coop}$</td>
<td>0.36</td>
<td>0.36</td>
<td>0.35</td>
<td>0.56</td>
<td>0.77</td>
<td>0.84</td>
<td>0.87</td>
<td>0.40</td>
<td>0.80</td>
<td>-0.68</td>
<td>-0.73</td>
</tr>
<tr>
<td>$R$</td>
<td>0.3</td>
<td>-0.2</td>
<td>0.54</td>
<td>0.31</td>
<td>0.48</td>
<td>0.03</td>
<td>0.31</td>
<td>0.65</td>
<td>0.48</td>
<td>0.0</td>
<td>-0.56</td>
</tr>
</tbody>
</table>

I calculated $\Phi$ values for some proteins in our set of proteins which have been investigated from experiment (see Appendix. 8.5). Following Garbuzyenskiy et al [162],
Figure 10: The schematic illustration of chevron plot (the top one) for point mutations with Φ values of 0 and 1 respectively (the lower ones). The kinetics under various thermodynamics conditions produces characteristically V-shaped chevron plots of $\log k_f$ and $\log k_u$ versus denaturant concentration.
Table 3: Correlation for theoretical and experimental Φ values for NMR structure proteins. \( R \) is correlation coefficient for the case without cooperativity effects. \( R^{coop} \) is correlation coefficient for the case with cooperativity effects.

<table>
<thead>
<tr>
<th>pdb</th>
<th>1o6x</th>
<th>1imq</th>
<th>1srl</th>
<th>2ptl</th>
<th>1aps</th>
<th>2abd</th>
<th>average</th>
</tr>
</thead>
<tbody>
<tr>
<td>( R^{coop} )</td>
<td>-0.2</td>
<td>-0.1</td>
<td>0.17</td>
<td>0.0</td>
<td>0.18</td>
<td>0.53</td>
<td>0.10</td>
</tr>
<tr>
<td>( R )</td>
<td>-0.13</td>
<td>-0.3</td>
<td>0.38</td>
<td>0.1</td>
<td>0.27</td>
<td>0.34</td>
<td>0.10</td>
</tr>
</tbody>
</table>

I make a distinction between contact maps obtained from native structures determined by x-ray crystallography and those from the first model of an NMR structure or minimized averaged NMR structure. Overall, the theory predicts Φ values for studied x-ray structures reasonably well. Eight of eleven proteins, the correlation coefficients between experimental and calculated Φ values are improved significantly. Still, there are exceptions. Of the 11 x-ray structures (see caption in Fig. 4), two proteins (1shg, 1ten) have large negative correlations. The average correlation coefficient for nine remaining proteins increases from 0.33 (for non-cooperative routes) to 0.6 (cooperative routes). The agreement with the NMR-determined structures is significantly worse with the average 0.1 for both non-cooperative routes and cooperative routes. In Fig. 9, I give three examples for x-ray structures and NMR structures respectively. Table 2 and Table 3 show our results for correlation between calculated Φ with experimental Φ for cooperative case and noncooperative case. My prediction is consistent with Garbuzynskiy’s result [162] in that models defined by proteins with X-ray structures tend to give a better predictions for Φ-values. We know the ensembles of NMR structures include many models. It is impossible to know which model in the PDB file will be the best for computation of Φ values. Perhaps, I should average all of calculated Φ values from all NMR models (Garbuzynskiy show that is
helpful, but this issue is not resolved at the moment). They found that $\Phi$ values of the internal residues are predicted equally from X-ray and NMR structures, while $\Phi$ values for the solvent-exposed residues are well predicted from X-ray structures, but poorly predicted from NMR ones. This may because the better refinement of internal residues in NMR structures (since internal residues have more NMR cross-peaks than the external one ) [162].

4.4 Conclusion

In this chapter, I investigate the effect of ”neutral” cooperativity on folding rates prediction and structure of transition state ensembles for a diverse set of two-state folding proteins. The coarse-grained variational model gives reliable folding rate predictions provided excluded volume terms that induce minor structural cooperativity are included in the interaction potential. In general, the cooperativity folding routes have sharper interfaces between folded and unfolded regions of the folding nucleus and higher free energy barriers. Also the calculated free energy barriers are strongly correlated with native topology as characterized by contact order. Increasing the rigidity of the folding nucleus changes the local structure of the the transition state ensemble nonuniformly across the set of proteins studied. Nevertheless, the calculated prefactor $k_0$ are found to be relatively uniform across the protein set, with variation in $1/k_0$ less that a factor of 5. This direct calculation justifies the common assumption that the prefactor is roughly the same for all small two-state folding proteins. Using the barrier heights obtained from the model and the best-fit monomer relaxation time
30ns, we find that $1/k_0 \sim 1 - 5\mu s$ (with average $1/k_0 \sim 4\mu s$). This model can be extended to study subtle aspects of folding such as the variation of the folding rate with stability or solvent viscosity and the onset of downhill folding. This approach was successful because the potential enhances cooperativity of the model since the repulsive potential between residues in proximity to native contacts is a convenient way to alleviate sensitivity on the excluded volume strength in the original model. Here our point of view is that the nature of interface of the folding nucleus is key in determining the behavior of folding rates and mechanisms, regardless of the specific form of cooperative interactions of the microscopic origins. If the qualitative results from this study can be extended beyond this variation model, it is likely to be limited to models that enhance cooperativity locally. Because these results are robust with respect to the exclude volume strength $U(0)$, the model lacks flexibility to explore a wide range of surface tensions. It will be interesting to see whether these conclusions hold when the interfacial surface tension is controlled directly through, for example, the formalism of density functional theory of first-order nucleation.
CHAPTER 5

SCALING OF FOLDING RATE WITH PROTEIN LENGTH

By theoretical work, Thirumalai [198], P.G.Wolynes [78], Finkelstein and Badred-dinov [129], Gütin and Shakhnovich [199] suggest that, for long chains, the folding barrier sublinearly scales with the length of the chain, $N$. That is,

$$\Delta F^\dagger \sim N^P, \quad (5.1)$$

with $0 \leq P < 1$. Different values of $P$ that have been suggested include $P = 0$ [199], $P = 1/2$ [198], $P = 2/3$ [78, 129]. The capillarity approximation to nucleation which gives $P = 2/3$, will be discussed in detail in Chapter 6.

Thirumalai [163, 182] recently attempted to determine the value of $P$ directly by fitting the measured folding rates of 57 proteins according to:

$$\ln k_f = b_1 N^P + b_2 \quad (5.2)$$

for different values of $P$. Here $b_1$ and $b_2$ are all constants. The correlations for $P = 0$, $P = 1/2$, and $P = 2/3$ were all similar. Thirumalai turned to a more subtle argument based on the value of the prefactor $k_0$ in the folding rate:

$$k_f = k_0 \exp(-\beta \Delta F^\dagger) \quad (5.3)$$

to at least exclude certain values of $P$. 

74
In particular, they assumed the scaling of the barrier given in Eq. 5.1 is true even for small \( N \), so that

\[ \beta \Delta F^\dagger = b_1 N^P. \]  

(5.4)

In this expression, the barrier vanishes at \( N = 0 \), so that prefactor, \( k_0 \), for a given value of \( P \) is given by the intercept of Eq. 5.2 as

\[ k_0 = \exp(b_2) \]  

(5.5)

By checking several power law fits between logarithm of folding rate \( \ln k_f \) and chainlength \( N \), they argue that some values of \( P \) can be ruled out some of fits by looking at whether the corresponding prefactor is reasonable. For instance, they first checked the power law fit between \( \ln k_f \sim \ln N \) (it should be noted that the case \( P \to 0 \) corresponds to correlation with \( \ln N \) because \( N^P = \exp(P \ln N) = 1 + P \ln N \) when \( p \to 0 \)). The fitting line for \( P \to 0 \) is \( \ln k_f = -5.5 \ln N + 25.8 \), which implies the prefactor \( k_0 = e^{28.5 s^{-1}} = (0.4 \text{ps})^{-1} \). This value is reasonable for small molecules, but is not appropriate to describe folding reactions. Then, they checked the power law fit between \( \ln k_f \sim N^{1/2} \), the fitting line is: \( \ln k_f = -1.1 N^{1/2} + 14.7 \), the prefactor is \( k_0 = e^{14.7 s^{-1}} = (0.4 \mu s)^{-1} \); The third scaling law they checked is between \( \ln k_f \sim N^{2/3} \), the fitting line is: \( \ln k_f = -0.36 N^{2/3} + 11.7 \), the prefactor is \( k_0 = e^{11.7 s^{-1}} = (8 \mu s)^{-1} \). The range \( 0.4 \mu s \sim 8 \mu s \) is consistent with an accepted value for the prefactor \( 1 \mu s \sim 10 \mu s \) and our previous estimation in Chapter 4. They conclude that \( 1/2 < P < 2/3 \) is acceptable since the corresponding prefactors are reasonable. They also comment that it is difficult to directly determine the precise value of scaling exponent \( P \) between 0
to 1 because the range of \( N \) is too small.

In this chapter, I reinvestigate this issue by looking at the barrier heights of 28 two-state proteins calculated from variational method as discussed in Chapter 3. From discussion in Chapter 4, I have already shown that the calculated folding rates correlated with experimental folding rates very well (correlation coefficient is 0.8). This indicates that our model can give us a relatively precise barrier heights. My calculations suggest that predicted barrier heights do have sublinear relationship with chainlength \( N \), but with a slightly different form with Thirumalai’s assumption (Eq. 5.4). Fig. 11 presents the sublinear relationship between predicted barrier heights and chainlength. The power law fitting equation between predicted barrier heights and chainlength can be written as:

\[
\Delta F^\dagger = a_1 N^P + a_2
\]

(5.6)

here, \( a_1 \) and \( a_2 \) are constants. Fig. 11 shows that barrier heights are well correlated with \( \ln N \) (Fig. 11a), \( N^{1/2} \) (Fig. 11b) and \( N^{2/3} \) (Fig. 11c), correlation coefficients are around 0.9. The fact is that the range of \( N \) of the measured rates is too small to use only the correlation coefficient of the fit to determine \( P \). All correlation coefficients are basically stay same for \( 0 < P < 1 \), as shown in Fig. 11d. Thus correlation coefficients are not enough to justify the scaling exponent \( P \).

More importantly, Fig. 11 shows that the barrier doesn’t vanish at \( N = 0 \) as Thirumalai and coworkers assume, but at a finite chain length, \( N^c_P \), for each \( P \). This changes the interpretation of the fitting parameters in Eq. 5.2.

From Fig. 11, the finite cutoff chainlength \( N^c_P \) can be calculated from the fitting
Figure 11: Barrier heights as a function of chain length. (a) $\beta \Delta F^\dagger / k_B T_f$ vs. $N^{1/2}$, fitting line is $y = -7.0 + 0.88x$, and correlation coefficient is 0.89. (b) $\beta \Delta F^\dagger / k_B T_f$ vs. $N^{2/3}$, fitting line is $y = -11.0 + 2.3x$, correlation coefficient is 0.9. (c) $\beta \Delta F^\dagger / k_B T_f$ vs. $\ln N$, fitting line is $y = -24.8 + 7.8x$, correlation coefficient is: 0.86. (d) correlation coefficient vs. power $P$. 
line between predicted barrier heights and chainlength. By extrapolating fitting line
to zero barrier height, \( N_c^P \) is given by:

\[
N_c^P = -a_2/a_1, \tag{5.7}
\]

For different scaling exponents \( P \), the cutoff chainlengths are shown in Fig. 12.
One can see that the values of \( N_c^P \) are not sensitive to the scaling exponent. For all
\( P \)'s between 0 and 1, \( N_c^P \) varies from 19 to 25.

![Graph showing cutoff value of chainlength vs power P and inset giving \( \tau_0 = 1/k_0 \) vs. power P.](image)

Now we consider the Thirumalai’s argument about the prefactor again using the
fact that the barriers vanish at $N_c^p$. By evaluating the fit at $N_c^p$, the prefactor can be estimated as

$$\ln k_0 = b_1 N_c^p + b_2 = b_1\left(-\frac{a_2}{a_1}\right) + b_2. \quad (5.8)$$

The inset of Fig. 12 shows the inverse of prefactor given by Eq. 5.8. It is clear that prefactors are relatively insensitive to the values of $P$, with $1/k_0$ ranging between $16 - 23\mu s$.

In summary, I showed directly that the variational model gives barriers which scales sublinearly with $N$ for two-state folded proteins. My analysis shows that I can’t exclude any value of $P$ based on the corresponding value of the prefactor. Also my analysis suggests that the cutoff value of chainlength for scaling law is around 20 residues. Thirumalai’s most recent study suggests that the average size of the most probable nucleus for single domain proteins is to be between $15 - 30$ residues [200]. It is intriguing that the cutoff chainlength $N_c^p$ falls within this range. It seems reasonable that a typical nucleus could not be smaller than minimum protein length where a folding barrier exists, but this connection needs more careful investigation. In the end, I find the same problem that everybody faces when trying to use the measured rates to directly determine the scaling exponent $P$. Namely, the range of chainlength for which folding rates have been measured is too limited to provide a clear answer for the scaling exponent $P$. 
6.1 Capillarity approximation to protein folding nuclei

A full structural description of transition state ensembles in protein folding includes the specificity of the ordered residues composing the folding nucleus as well as spatial density. To our knowledge, the spatial properties of the folding nucleus and the interface of specific nuclei has yet to receive significant attention. Usually, the structure of the transition state ensemble is described through the degree of native-like order of specific residues due to the structural information inferred from site-directed mutagenesis folding experiments. Nevertheless, quantitative spatial properties such as size or density of the critical nucleus still remains unresolved. Indeed, shortly after characterizing the transition state ensemble of CI2, Fersht proposed a spatial description of the critical nucleus that supported the kinetic data. [77] The critical nucleus envisioned in this nucleation-condensation mechanism can be thought of as an expanded, partially ordered version of the native state ensemble with concomitant long-ranged tertiary and local secondary structure. Although diffuse nuclei appear to be the general rule, some nuclei are less diffuse than others. [181] Polarized nuclei have highly structured residues that are spatially clustered in the native structure with the rest of the residues having little definite order. [201–204] Such nuclei are similar to the capillarity approximation in homogeneous nucleation where a stable
phase droplet is separated from the metastable phase by a sharp interface [78, 129]. Exploiting this analogy, Wolynes describes a nucleus with capillarity-like order in which the interface surrounding a relatively folded core is broadened by wetting of partially ordered residues [78]. In this picture, folding can be described as a wave of order moving across the protein as the edge of the nucleus expands to ultimately consume the entire molecule [78]. This has been most thoroughly quantified using the Hammond shifts along with continuous Φ-value analysis for the spliceosomal protein U1A, [86, 205]. For U1A, fractional native order is established in most parts of the molecule before well-developed regions with high Φ values appear. This means that the interface region between ordered and disordered region is pretty broad and extended. However almost at the same time, it was found that barrier has a weaker scaling with chain length than what mean field theory suggested [129, 198, 199]. For example, Finkelstein and Badredtinov derived the scaling law $\Delta F^\dagger \sim N^{2/3}$ using a capillarity picture of the folding nucleus. Thirumalai [198] obtained different scaling law, $\Delta F^\dagger \sim N^{1/2}$, appealing to the scaling theory for glasses [206, 207]. Gutin and Shankhnovich suggested that $\Delta F^\dagger \sim \ln N$ at the temperature of fast folding but linear scaling with $N$ for slow folding events at low temperature [199].

In this Chapter, I reconsider the scaling of the barrier height with chain length within the capillarity picture of protein folding nucleation. The extended partially ordered interface of a capillarity-like ordered nucleus separates space into three regions: a folded core, a partially ordered interface region, and an unfolded halo (see Fig. 13). Growth of the nucleus is controlled by fluxes of residues passing through two moving
Figure 13: Illustration of folding nucleus: folded core, interfacial region, and unfolded halo. Growth of the nucleus can be characterized by fluxes entering the folded core and interfacial regions.

surfaces: one surface separates the folded core and interface, and the other surface separates the interface region and the unfolded halo. As the protein folds, the folded core grows by incorporating residues from the interface region, while the evolution of the interfacial region is determined by the net flux of residues entering the interface.

My analysis is based on folding routes calculated for 27 two-state proteins from a cooperative variational model described in Chapter 3. I note this model includes neutral cooperativity due to repulsive excluded volume interactions. This form of cooperativity has been shown to broaden the range of barrier heights allowing direct comparison between calculated and measured folding rates as shown in Chapter 4. Although this cooperativity tends to sharpen the interface between folded and unfolded regions, the interface from this model is not so sharp that each residue is
either completely folded or completely unfolded as assumed in some other analytic models [139–141]. In fact, an unbiased analysis of the spatial properties of the folding nucleus fundamentally depends the model’s ability to describe partial order.

The capillarity approximation of folding nuclei is based on classical nucleation theory of first order phase transition kinetics. [26, 78]. Within the capillarity approximation, the free energy of a nucleus with volume $V_f$ and surface area $A_f$ can be written as a sum of two terms

$$F = -\Delta f V_f + \gamma A_f,$$  \hspace{1cm} (6.1)

where $\Delta f$ denotes the bulk free energy difference per unit volume between the unfolded and folded ensembles, and $\gamma$ is the surface tension between the folded and unfolded regions.

A folded core with native-like density has a volume per monomer independent of its size. Relaxing this assumption, we take the number of residues in the folded core, $N_f$, to scale with its volume, $V_f$, according to

$$V_f = b^3 N_f^{3\nu}.$$  \hspace{1cm} (6.2)

Here, $\nu$ is the scaling exponent associated with the length scale of the folded core $R \sim b_0 N_f^\nu$, and $b^3$ is a geometry-dependent elementary volume proportional to the monomer volume, $b^3_0$. The free energy of a folded nucleus with $N_f$ residues then has the form: [78]

$$F(N_f) = -\Delta f b^3 N_f^{3\nu} + \gamma b^2 N_f^{2\nu}.$$  \hspace{1cm} (6.3)
Eq. 6.3 can be written as:

$$F(N_f/N) = -\Delta f b^3 (N_f/N)^{3\nu} N^{3\nu} + \gamma b^2 (N_f/N)^{2\nu} N^{2\nu}. \quad (6.4)$$

where $N$ is the number of monomers in the protein. At the folding transition temperature, $T_f$, free energy are equal for $N_f/N = 0$ and $N_f/N = 1$, that is: $F(0) = F(1)$, this gives: $\gamma = \Delta F b N^\nu$. In addition, at the maximum of the free energy, $\frac{\partial F}{\partial Q} = 0$ (here $Q = N_f/N$), this gives: $-3\Delta F b N^\nu Q^\nu + 2\gamma = 0$. Together all of arguments, it shows that the maximum of the free energy occurs at $N_f^\dagger = (2/3)^{1/\nu} N$, with an associated free energy barrier that scales as $\Delta F^\dagger \sim N^{2\nu}$. If we assume that the folded core has native-like packing, $\nu = 1/3$ and $b^3$ is the native-like volume per monomer, so that $N_f^\dagger = (2/3)^3 N$ and $\Delta F^\dagger \sim N^{2/3}$ [78, 129]. Simulations and alternative theoretical considerations also suggest that barrier height (logarithm of the folding time) scales sublinearly on chainlength, $\Delta F^\dagger \sim N^p$, with $0 \leq p \leq 1$ [64, 78, 129, 198] as discussed in Chapter 5. Direct analysis of folding rate data to determine the scaling exponent $p$ encounters the difficulty that the range of $N$ is too small to distinguish between different values of $p$ [2, 163, 164, 208, 209]. So, while it may be reasonable to expect that the scaling of the barrier height with chainlength is universal for sufficiently large proteins, the size of typical two-state proteins ($\sim 100$ amino acids) may well be too small to be governed by this generic behavior. In this case, both specificity and size of these smaller proteins should generally determine the properties of the critical nuclei. In this chapter, I assume that Eq. 6.2 is valid to describe the growth of the nucleus in all the two-state proteins, but the exponent $\nu$ and volume $b^3$ are allowed to be protein specific.
6.2 Characterizing the folded core and the interface

In the variational model considered in this chapter, partially ordered configurations are described by a variational Hamiltonian, $H_0$, corresponding to a stiff polymer chain inhomogeneously constrained to the native structure. Since this model is described in detail in Chapter 3, here I focus on how to define folded core, interface, and unfolded regions along the calculated folding route. This is not as straightforward as one might expect because the concept directly couples specificity of the nucleus with the spatial density.

I characterize the degree of structure of each residue by the extent of localization with respect to the native structure $\{r^N\}$, $\rho_i = \langle \exp(-\alpha^N(r_i - r_i^N)^2) \rangle_0$, with $\alpha^N = 0.1$. Here, the subscript denotes the average with respect to the Boltzmann weight with $\mathcal{H}_0$. Denoting the native density at the globule and native state by $\rho_i(G)$ and $\rho_i(N)$, respectively, I consider the normalized density as in Eq. 3.35,

$$\tilde{\rho}_i = \frac{\rho_i - \rho_i(G)}{\rho_i(N) - \rho_i(G)}$$

as a set of order parameters characterizing the folding of each residue. Progress along the folding route can be monitored by the global structural parameter $Q = 1/N \sum \tilde{\rho}_i$.

I use the normalized native densities to define a fiducial set of folded residues, $\{F\}$, with $\tilde{\rho}_i > 0.6$, as shown in Fig. 14a. Next, I define the spatial region of the folded core through the relative contribution of the density of the folded residues in $\{F\}$, $n_f(\mathbf{r}) = \sum_{\{F\}} \langle \delta(\mathbf{r} - \mathbf{r}_i) \rangle_0$, to the total density, $n(\mathbf{r}) = \sum_{i=1}^{N} \langle \delta(\mathbf{r} - \mathbf{r}_i) \rangle_0$. The
Figure 14: Scaling of the folded core with number of monomers. (a-b) correspond to λ-repressor (1lmb). (a) Residues with native density $\tilde{\rho}_i > 0.6$ (indicated by the dashed line) define a fiducial set of folded residues. (b) Linear fit of $\log V_f$ vs. $\log N_f$ (dashed line) gives the exponent of $V_f \sim N_f^{3\nu}$. In this example, the fitting equation is $y = 5.6 + 0.97x$, so that $\nu = 0.32$ and $b^3 = 5.6a^3$, $a = 3.8\AA$ is the average distance between the α carbons. (c) Histogram of scaling exponent $\nu$ for 28 proteins; (d) Histogram of the packing fraction $\mu_f$ as given in Eq. 6.7 of the critical nucleus at $T_f$. 
spatial extent of the folded core and interfacial regions in this analysis is determined by an indicator function

\[ \tilde{n}(\mathbf{r}) = \frac{n_f(\mathbf{r})}{n(\mathbf{r})}, \quad (6.6) \]

where \( 0 \leq \tilde{n}(\mathbf{r}) \leq 1 \). We define the folded core region, \( V_f \), as the points \( \{\mathbf{r}\} \) for which the density of the fiducial folded residues contributes at least 50% to the total density (\( \tilde{n}(\mathbf{r}) \geq 0.5 \)). The number of residues in this folded core region can be found by numerically integrating the density over the core region, \( N_f = \int_{V_f} n(\mathbf{r}) \, d\mathbf{r} \). The volume of the core region is given by \( V_f = \int_{V_f} \, d\mathbf{r} \).

Similarly, the interfacial region, \( V_{int} \), is defined as the points \( \{\mathbf{r}\} \) for which \( 0.1 \leq \tilde{n}(\mathbf{r}) < 0.5 \). The number of interfacial residues and volume of the interface is given by \( N_{int} = \int_{V_{int}} n(\mathbf{r}) \, d\mathbf{r} \), and \( V_{int} = \int_{V_{int}} \, d\mathbf{r} \), respectively.

The number of residues and the volume can be used to define a mean packing fraction of the folded core and partially ordered interface by

\[ \mu_f = \frac{N_f}{V_f} v_0 \quad \text{and} \quad \mu_{int} = \frac{N_{int}}{V_{int}} v_0, \quad (6.7) \]

respectively. Here, \( v_0 \) is the calculated volume per particle of the native structure at the folding transition temperature, \( T_f \). The growth of the nucleus can be characterized by the way the packing fractions \( \mu_f \) and \( \mu_{int} \) change along the folding route.

### 6.3 Growth of folding nucleus along the folding route

As illustrated in Fig. 14(a-b), the changes in \( N_f \) and \( V_f \) along a folding route can be fit to Eq. 6.2 to give an estimate of the scaling exponent \( \nu \) for each protein. Fig. 14(c) shows the distribution of predicted \( \nu \) from the folding routes obtained from
the variational model for 28 two-state proteins discussed in Chapter 3. The predicted scaling exponent $\nu$ ranges between $0.2 \sim 0.42$ with an average of $\nu = 0.33$. The mean exponent is very close to the the scaling associated with close-packed rigid objects, $\nu = 1/3$. For comparison, recent detailed statistical models indicate that the scaling exponent for the unfolded state of a protein is about $\nu = 0.59$ [210], whereas a wide variety of protein folded structures suggest that proteins with less than 300 amino acids have compact folded structures ($\nu = 0.3$) and larger proteins are less dense ($\nu = 0.4$). [211]

The mean packing fraction of the core scales with the number of monomers as:

$$
\mu_t = \frac{v_0}{b^3}N_t^{1-3\nu}.
$$

(6.8)

For the close packing value $\nu = 1/3$, $\mu_t$ is independent of the number of monomers. When $\nu > 1/3$, the core becomes less compact as monomers are added to the core. This is the familiar scaling from loosely packed or fractal objects. When $\nu < 1/3$, the core density increases as more monomers are incorporated into the core. This can be understood as the consolidation of structure in the folded core as folding progresses.

Although the spatial structure of the critical folding nucleus (transition state ensemble) is discussed in more detail later, it is instructive to consider the value of the mean packing fraction of the core here. Fig. 14d shows the distribution of packing fractions of the folded core evaluated at the maximum free energy barrier between folded and unfolded states at $T_f$. The packing fraction has a wide range from 0.2 to 1.0. Though some of the transition state nuclei have compact cores, the average packing fraction is only 0.59. This means that although the growth of a typical folded
core corresponds to rigidly packed objects, a typical transition state ensemble has a folded core with twice the volume as the volume of same number of monomers in the native state conformation \( (b^3 \approx 2v_0) \). That is, the monomers composing the nucleus are typically much less localized than in the native state.

Figure 15: Illustration of growth of folding nucleus and interface along the folding route (increasing \( Q \)) for the \( \lambda \)-repressor protein (1lmb). Column 1 shows the three-dimensional folded structure with the fiducial set of folded residues colored blue and the unfolded residues colored red. In Column 2, the folded core (colored blue) is surrounded by the interfacial region (colored green). Column 3 is a projection of the indicator function \( \tilde{n}(r) \) that defines the folded and interfacial regions in space. The values correspond to \( \max_z \tilde{n}(x, y, z) \), ranging from 1 (blue) to 0 (red) in steps of 0.01. Contour lines correspond to 0.1, 0.5, 0.7. Column 4 gives the corresponding \( Q \) value for each row. The critical nucleus corresponds to \( Q = 0.53 \). The units for three plots are in Angstroms. This protein is belong to the Pattern C (balanced growth). The three-dimensional structure was produced by VMD.

Fig. 15 shows a representative example of the growth of the folded core and the
interface region. Early in the folding, the folded nucleus is small and compact, surrounded by a partially folded interface. This small nucleus is partially ordered, occupying about twice the volume as the corresponding residues in the native state. Structural fluctuations giving nuclei corresponding to $Q < Q^\dagger$ are unstable with respect to the unfolded state due to relatively large surface free energy cost associated with small nuclei, whereas structural fluctuations with $Q > Q^\dagger$ tends to evolve to the folded state. As the nucleus grows, the volume of the nucleus evolves as interfacial regions are incorporated into the core, while unfolded residues become part of the partially ordered interface.

6.4 Growth pattern of the nucleus

The structural growth of the folding nucleus can be understood as the competition between growth of the folded core and the evolution of the interface. The flux of residues entering core through the interface region controls the growth of the core, and the net flux of residues entering interface region from the unfolded halo controls the growth of the interface (see Fig. 13). The evolution of the nucleus along the folded route can be monitored through changes in the number of residues, volume, and packing fraction as a function the global order parameter $Q$. The scaling relation given in Eq. 6.2, for example, is a parametric equation of $V_f(Q)$ and $N_f(Q)$. As noted below Eq. 6.9, the density of the folded core can increase, decrease, or stay the same as residues are incorporated into the core depending on the value of the scaling exponent $\nu$. Here we consider evolution of both the folded core as well as the interface
between folded and unfolded regions along the folding route to describe the growth of the nucleus.

We focus on the changes in the packing fraction $\mu_t(Q)$ and $\mu_{int}(Q)$ along the folding route to describe growth of the folding nucleus. That is, we consider the signs of

$$\dot{\mu}_t(Q) = \frac{d\mu_t}{dQ} \quad \text{and} \quad \dot{\mu}_{int}(Q) = \frac{d\mu_{int}}{dQ}$$

(6.9)

to identify different pattern of growth. From the two-state proteins used in this study, we can identify three distinct scenarios as illustrated in Fig. 16. (Similar plots for all the protein studies are given in the Appendix. 8.4.)

- **Pattern A** (*consolidation of core and interface*) As shown in Fig. 16a, proteins described by this growth pattern have $\dot{\mu}_t(Q) > 0$ and $\dot{\mu}_{int}(Q) > 0$. For these proteins, the number of folded residues increases faster than volume increases so that the folded core region becomes increasingly tighter packed along folding route. The consolidation of the core is closely related to the evolution of the interface region, shown as dashed lines in Fig. 16. The growth of the interface region determined by the density is similar, though this time both the number of residues and their occupying volume are decreasing along the folding route. Both the core and interface become more compact until the interface is almost completely depleted near the folded configuration.

- **Pattern B** (*consolidation of the core only*) As shown in Fig. 16b, proteins described by this growth pattern have $\dot{\mu}_t(Q) > 0$ and $\dot{\mu}_{int}(Q) \sim 0$. This means
Figure 16: Examples of three patterns of growth of the nucleus. Pattern A–C corresponds to (a–c), respectively. Solid line corresponds to the mean packing fraction of the folded core, $\mu_f$, while the dashed line corresponds to the mean packing fraction of the interface, $\mu_{\text{int}}$. 
that the folded core consolidates along the folding route as in Pattern A, but
the interface region evolves with relatively constant density. In the illustration
shown in Fig. 16b, the $N_{\text{int}}$ and $V_{\text{int}}$ remain relatively constant throughout much
the growth. Then the interface can be seen as a channel transferring residues
from the unfolded halo to the folded core with no net flux of residues through
the interfacial surface. For some other proteins in this class, a decrease in the
number of interfacial residues is balanced by a proportional decrease in the
interfacial volume (see the Supporting figures).

- **Pattern C** (*balanced growth*) As shown in Fig. 16c, proteins described by this
growth pattern have $\dot{\mu}_f(Q) \sim 0$ and $\dot{\mu}_{\text{int}}(Q) \sim 0$. For this set of proteins,
the cores do not consolidate until the end of folding. Rather, the evolution of
the number of residues and the volume keeps the density of both regions rela-
tively constant. (The same comments about balanced evolution of the interface
relevant to Pattern B applies here as well.)

Out of the 27 proteins studied, three proteins are difficult to classify by this scheme
(see Appendix. 8.4). The packing fractions of folded cores for these proteins show
clearly sharp variations between low and high values. Some of the ruggedness of these
curves is due to the rigid cutoff values defining the fiducial set of folded and interface
residues as well as the three spatial regions of the nucleus. Whether this accounts
for the anomalous calculated behavior for these three anomalous proteins is not yet
clear.
The growth mode of the nucleus for the 27 proteins considered in this paper (1pgb16 is too small to have a compact folded core) can be roughly classified as follows: Pattern A: 1pgb, 1a0n, 1pks, 1pin, 1psf, 1shg, 2ptl; Pattern B: 1c8c, 1coa, 1enh, 1fkb, 1hdn, 1vii, 1wit, 2pdd; Pattern C: 1aps, 1csp, 1mf9, 1imq, 1mef, 1o6x, 1sr1, 1ten, 1lmb; The growth patterns for three proteins were difficult to classify using this criterion: 1div, 1urn, 2abd.

6.5 Polarized versus diffuse critical nucleus

A folding mechanism is typically characterized by the structure of the critical nucleus. The spatial structure of the transition state ensemble, inferred from Φ-value analysis, has often been qualitatively summarized as either diffuse or polarized. [212] Intermediate Φ-values spread across a large portion of the protein sequence indicate a diffuse nucleus. In contrast, polarized transition states are inferred when only one part of the structure has relatively high Φ-values and the rest of the residues have low Φ-values. In addition to a bimodal distribution of Φ-values, the ordered residues in a polarized transition state ensemble are located in one region in the native configuration. Polarized and diffuse critical nuclei are sometimes called localized and delocalized transition state ensembles, respectively. [213] Of course, the critical nucleus of a given protein is expected to have structural properties somewhere between the two ideal limits. The second row of Fig. 15 gives an example a diffuse critical nucleus (1lmb). For comparison, Fig. 17 shows the corresponding plots for a protein with a polarized critical nucleus (1sr1). Comparing Fig. 15 and Fig. 17, it is clear that
the interface of 1lmb is much broader than the interface region of 1srl. Furthermore, the folded core of 1lmb is much more diffuse than the folded core of 1srl.

Figure 17: An example of a polarized critical nucleus (Q=0.45) for Src-SH3 (1srl). Plots (a–c) correspond to the middle row of the diffuse nucleus shown in Fig. 15.

Characterizing a capillarity-like ordered nucleus as either diffuse or polarized is a statement of the sharpness of the interface as well as the compactness of the core. For convenience, we monitor both regions by the normalized volume per monomer (inverse packing fraction): $1/\mu_f$ and $1/\mu_{int}$. The results for the two-state proteins considered in this work are shown in Fig. 18. Nuclei with small values of $1/\mu_f$ and $1/\mu_{int}$ are more polarized, possessing relatively compact cores and sharp interfaces (similar to those envisioned in the strict capillarity approximation). Diffuse nuclei, on the other hand, have extended regions of partial order that correspond to larger values of $1/\mu_f$ and/or
1/\mu_{int}. We note that relatively polarized nuclei can have cores that are still loosely packed compared to the native state density (e.g., 1pgb). Furthermore, relatively diffuse nuclei can have tightly packed cores but extended interfaces (e.g., 2abd, 1imq, 1fkb).

Our analysis suggests polarized critical nuclei consistent with classification inferred by experimental \(\Phi\)—value analysis (such as 1csp [204], 1srl [202], 1shg [201], 1pin [203] 2ptl [122], and 1pgb [214]). Our model also is consistent with several proteins classified as having diffuse critical nuclei (such as 1lmb [36], 2abd [215], 1imq [216] and 1fkb [217]). This favorable comparison for a wide variety of proteins is a reassuring assessment of the model. At the same time, we realize this comparison is necessarily qualitative and subjective.

With this caveat in mind, there are several proteins for which the model appears at odds with the characterization of the critical nucleus inferred from \(\Phi\)-values (such as CI2 [77], 1aps. [218], U1A [205], and 1pgb [214]). In these exceptional cases, \(\Phi\)-value distributions indicate that the critical nucleus is rather diffuse but our model predicts more polarized nuclei. This tendency may indicate that the model is too cooperative for these proteins, since high cooperativity is expected have sharp interfaces and more polarized transition state ensembles.

6.6 Conclusion

In this chapter, we directly characterize folding in terms of the capillarity-like growth of the folding nucleus. The nature of the partially folded interfacial region
Figure 18: Inverse packing fraction of the interface and folded core. for 27 two-state proteins. The gradual change color shows the continuous change from polarized nuclei (red) to diffuse nuclei (cyan).
between the folded core and unfolded halo is the central focus of characterizing the
growth patterns of the nucleus. We find that the growth of the nucleus can be
classified into three different patterns: (A) the core and interface both condense along
the folding route; (B) the core condenses at the expense of the interfacial region; and
(C) the growth of the core is balanced by the monomers entering the interfacial region
from the unfolded halo. The picture of the core as close packing of rigid monomers
appears to be valid on average, though the size of the effective monomers is larger than
one would expect for a native-like, compact core. This analysis clarifies that diffuse
nuclei inferred by the distribution of intermediate Φ-values, for example, can arise
from either a diffuse folded core, a broad interfacial regions, or both. The predictions
from our calculations can be tested from the analysis of the evolution of Φ-values as a
function of the movement of the transition state ensemble (β†) pioneered by Oliveberg
and coworkers. [143, 205]

The variational model considered here includes a uniform “neutral,” excluded vol-
ume type cooperativity developed to account of general trends in the absolute folding
rates of two-state proteins. The exceptional qualitative discrepancies of the the po-
larized versus diffuse characterization of the critical nucleus (such as CI2, 1aps, 1pgb,
and U1A) permit an opportunity to assess the form and strength of the cooperativity
of this model. The spatial density of the critical nucleus can be used as an inde-
pendent criterion to check the value of the cooperativity obtained by simultaneously
fitting Φ-values and barrier height by the parameterization of the cooperativity for
each protein. There are some indications that one should consider variations in the
strength of the cooperativity for different proteins (though, admittedly this is very closely tied to the specific form of the cooperativity in the model). For example, Eftehadi and Plotkin recently found that the strength of cooperativity from three-body interactions can be tuned for each protein to bring simulations of \( \Phi \)-values into better agreement with experimental measurement [72]. The generally good qualitative agreement between our calculations and experimental inferences about the spatial extent of folding nuclei suggest that tuning the excluded volume strength for each protein would not greatly improve the results presented here for the majority of the proteins studied.
CHAPTER 7

STABILITY EFFECT ON PROTEIN FOLDING

Many factors affect protein behavior in the cell, such as viscosity, salt concentration, molecular crowding, stability, temperature, pressure, PH value. In such a complicated environment, it is at present impossible to study protein folding problem accounting for all of factors simultaneously. Currently, the best way to learn about folding is to isolate such factors and study the physical consequences of environment conditions one at a time. Research has focused on characterizing protein dynamics as a function of viscosity [220–222], temperature [223–225], pressure [226–228], pH [229, 230], stability [189, 231]. In this Chapter, I mainly focus on stability, the free energy difference between globule state and native state. Stability characterizes the populations of folded and unfolded states. In previous Chapters, I studied folding dynamics at the folding transition temperature, which means that globule state and native state have same free energy. That is, at equilibrium, the protein is equally likely to be found in the unfolded state ensemble as the folded state ensemble. In this Chapter, I present some initial studies on how the variational model works under different environmental conditions that affect stability. Though not complete, I find some interesting effects on folding: Hammond shift, catastrophe in the free energy surface, kinetic rollover in Chevron plots, and possible signature of downhill folding are demonstrated.
7.1 Effect of temperature on protein kinetics

In our model, stability is determined by the temperature. For example, zero stability corresponds to folding transition temperature $T_f$. An example of temperature dependence of stability is shown in Fig. 19 (The whole set of data are shown in Appendix. 8.7). The temperature dependence of stability is easy to understand. At high temperatures, the globule state is favored over the folded state since the globule state is less constrained and has more conformational entropy. At lower temperatures, the native state is favored over the globule state since the native state is the low energy ensemble. Thus lower temperatures present folding environmental conditions, whereas high temperatures present unfolding conditions. If we denote the unfolding rate by $k_u$ and the folding rate by $k_f$, the stability is reflected in the equilibrium constant

$$K_u = k_f / k_u = \exp[-\beta(F_N - F_U)],$$

(7.1)

where $F_G$ and $F_N$ is free energy at globule and native state respectively. Stability is often expressed as

$$\log K_u = \beta(F_U - F_N) = \Delta F_{UN}.$$  

(7.2)

7.2 Movement of transition state

Intensive studies on Cl2, barnase and N-terminal domain of L9 and tendamistat show that Hammond behavior is a general phenomenon in protein folding [232–237].
Figure 19: Temperature vs. stability for λ−repressor(1lmb). \( \log K_u \) is defined in Eq. 7.2.

In particular, mutational studies and denaturant-induced, temperature-induced folding/unfolding experiments revealed that the position of the transition state may significantly change with protein stability. These results were interpreted in terms of the Hammond postulate [238], which is well established in classical organic chemistry. It states that the transition state moves towards the unstable state as stability is changed.

7.2.1 Hammond shift

The stability can be changed by mutations of the sequence, variation of the temperature, or solvent conditions. In my study, I change stability directly by adjusting temperature. As an example, Fig. 20 show the order parameter change of transition
state with stability for $\lambda$–repressor(1lmb). Similar data for all of other proteins are shown in in Appendix. 8.7. It is clearly noticeable that order parameter Qts changes smoothly with stability for most proteins, such as: 1pks, 1lmb, 2pdd, 1c8c, 1pin, 2abd, 1mef, 2ptl, 1pgb, 1ten, 1urn, 1enh, 1a0n, 1vii, 1coa, 1wit, 1fkb, 1aps (although not completely perfect). Generally, Qts increases when the stability reduced, making the native state more favorable. This is entirely in accord with Hammond postulate. At high temperature, the native state is destabilized and for unfolding becomes the transition state structurally more native-like.

Nevertheless, I do observe exceptions to this general rule. Two proteins, 1o6x and 1fnf9 show anti-Hammond behavior. Here, the transition state moves towards unstructured global state when the folded protein is destabilized. According to Jencks's
review [239], Hammond behavior results from movement parallel to the reaction coordinate. However, energy surfaces are highly multidimensional, so movement can occur perpendicular to the reaction coordinate as well. The energy surface around a transition state is saddle shaped with energy going through a maximum along the reaction coordinate and a minimum in directions perpendicular to it. If the transition state is destabilized in one of these perpendicular directions, then the transition state can move towards the stable state. This leads to anti-Hammond behavior [235].

![Graph](image)

Figure 21: Antihammond behavior for protein 1o6x and 1fnf9.

7.2.2 Sudden changes in Qts

In a couple of proteins, we find that the position of transition state ensemble is not a smooth function of stability. Proteins, 1srl and 1shg, show sudden deviation of Qts according to changes in stability. Basically, they still follow Hammond behaviors, but the origin for order parameter Qts movement with changes in stability are totally different. For 1srl, around $\log K_u \approx 5k_B T$, there is a sudden, but continuous drop from 0.25 to 0.15 as $\log K_u$ decreases. This point is a unusual point, not only because
quick drop of order parameter $Q_{ts}$, but also we can see this point is very special in
prefactor, too. Fig. 22 shows that around $\log K_u \approx 5k_B T$, prefactor is very small
and finally goes to zero. Mathematically, this phenomenon is called a catastrophe,
where a critical point becomes degenerate [184]. In particular, here the curvature of
the unstable mode vanishes in a so called cusp catastrophe controlled by the kernel
$x^4 + bx^2 + cx$ [184]. Since the prefactor characterizes recrossing of the barrier along
the unstable mode of the transition state ensemble in our harmonic approximation, it
is proportional to the second derivative of free energy at the barrier top. If prefactor
goes to zero, this indicates that the free energy surface is getting flat and broad. I
present the folding route of 1srl for certain range of stability in Fig. 23. It is quite
clear to see this scenario from Fig. 23.

Moreover, we studied how $\Phi$-values change with stability for 1srl. We mutate
each residue by Alanine and calculate the $\Phi$ value for each residue. My results are
presented in Fig. 24 for each secondary structure segment. From the Fig. 24, it is
clear that $\Phi$ value for each residue has a smooth drop, but by different amounts. This
Figure 23: Folding route of 1srl for varied stability.
leads to the average $\Phi$ value, which is presented in Fig. 25: having a continuous, smooth change around $5k_B T$ stability. Both individual $\Phi$'s and averaged $\Phi$ tell us that structure of each residue, and so for total protein, are becoming ordered around $5k_B T$ when stability reduced. This is consistent with Hammond behavior predicted from order parameter $Q_{ts}$. This example may help interpret experiments which probe the movement of TSE pioneered by Oliveberg [205].

Figure 24: $\Phi$ values for each residue of protein 1srl. $\Phi$ values for each secondary structure were plotted separately.
7.3 Effect on prefactor

The folding rate prefactor describes a microscopic time-scale of folding. It is strongly temperature-dependent or stability-dependent, and closely related the shape of the barrier top. Within the two-state fold stability range, prefactor show some universal properties. In my study, the behavior of prefactor as stability varies for different proteins is shown in Appendix. 8.7. Overall, we find that the prefactor generically is non-monotonic with a maximum at intermediate stabilities. A typical example is shown for 1enh in Fig. 26. The prefactor decreases as the stability approaches the limits of two state behavior. This indicates that the existence of a denatured native spinodal at low temperature and a denatured globule spinodal at high temperature. As in the nucleation of a liquid droplet from a vapor, the microscopic
Figure 26: An example of prefactor vs. stability (1enh).
origin of the barrier is nontrivial [5, 191, 219, 240]. This is another catastrophe, this
time called a fold catastrophe, controlled by the kernel $x^3 + ax$. If the stabilities of
the two stable states (folded state and unfolded state) can be sufficiently changed,
one can imagine that a protein may enter a regime in which the barrier disappears.
In this regime, there is no obvious separation of time scales [25] in the transition time
and the equilibration times within the metastable basins. In the laboratory, adjust-
ing thermodynamic conditions so that the biases of the system toward the folded or
unfolded is strong enough, energy landscape can give a single minimum [241]. The
physiological condition for the true spinodal points is usually hard to find, though
spinodal points exist in theoretical models [242].

7.4 Effect on chevron plot

A chevron plot is a plot of natural log of the observed rate constant, $k = k_f + k_u$,
versus denaturant concentration. It is a way of representing protein folding kinetic
data in the presence of varying concentrations of denaturant that disrupts the pro-
tein’s native tertiary structure. One of the defining characteristics of apparent two-
state folding is a chevron plot whose folding and unfolding arms are linear [243].
However, two-state behavior is not a universal property of natural proteins [66, 188].
Non-linearities are usually observed in the either both the arms or one of them and
are termed chevron roll-overs. The origin for such an observations is not totally clear.
Many interpretations including on-pathway intermediates [244], dead-time limita-
tions, transition state movements(Hammond effect) [205], aggregation artifacts [245],
downhill folding [246] and salt-induced Debye-Huckel effects [247] have been proposed to explain this behavior.

In my study, I constructed the chevron plots as functions of stability($\Delta F_{UN}$), U means unfolded state and N means folded state. Here, $\Delta F_{UN}$ was taken as a rudimentary correspondent to the changes in denaturant concentration in experimental measurement. Here, I present two typical examples for normal chevron plot and chevron roll-overs, as shown in Fig. 27. Part of the rest of chevron plots are shown in Appendix. 8.8.

From all of data (Appendix. 8.8), we can see that most proteins show two-state behavior in certain range of stability, both folding and unfolding arms show linear relationships between log of observed folding rates($k_f + k_u$) and stabilities. These proteins are 2ptl, 1aps, 1div, 1fnf9, 1shg, 2abd, 1srl, 1pin, 1mef, 1pgb. A few proteins show both chevron roll-overs in both folding and unfolding arms, such as 1enh, 1a0n. Also some proteins show that there is at least one chevron rollover which lies in either folding arm or unfolding arm, such as 1urn, 1pks, 1c8c show rollover in unfolded arms, 1imq, 1lmb, 1ten, 1fkb, 1o6x, 1vii, 1coa show rollover in folding arms.

For those proteins with rollovers in folding arms, rollovers always happen around the strong native conditions (large positive stability), but are two-state when the native bias is weakened [248, 249]. This can be understood as the barrier vanishes under strongly favored native conditions. Thus at certain stability, there might be a transition between two-state kinetics to non-two-state kinetics (downhill folding).
Figure 27: Two-state chevron plot for 1wit (top) and chevron roll-over for 2pdd (bottom). More positive stability values correspond to conditions under which the native state is more favored thermodynamically.
and finally got a unimodal conformational distribution [25]. As Bryngelson et al proposed [25] that there is only one significant populated state in the barrierless regime. Also the rate of the folding and unfolding are determined by the local features of the free energy landscape as the protein diffuse toward to folded state or unfolded state. Intrinsic barrierless folding has been claimed for the folding process for one protein BBL experimentally [82], although those results have been called into questions [83]. As illustrated in Fig. 23, it does show the trends of barrierless folding and unfolding at strong native conditions and unfolded conditions in free energy landscape.

Perhaps, chevron rollover at both low and high temperatures could provide kinetic signatures of downhill folding and unfolding. As well as from experimental studies for barnase [250, 251], ribonuclease A [252], hen lysozyme [253], and U1A [187, 254]. These all strongly suggest that the folding and unfolding under strong native conditions and unfolding conditions are qualitatively different from those under milder conditions. For our protein set, the drastic chevron rollovers under strong native conditions and unfolded conditions have not been totally confirmed experimentally. This suggests that such conditions, which probably coincide with downhill folding [25, 249, 255] may not be realizable in the laboratory. This is not surprising since native stability can be arbitrarily high in the model (Δ$F_{FN}$ can be arbitrarily large), but in reality, the native stability is limited by solvent conditions, for example, the freezing and boiling point of water. So it is interesting to explore when to expect this behavior, such as corresponding to the very strongly folding conditions for some proteins so that we have a clear kinetic signature of downhill folding [256].
7.5 Conclusion

Finally, as a remark, our studies in this part shed a light on the stability effects on folding temperatures, transition state positions, prefactors and chevron plots. We captured two-state behaviors for a certain range of stabilities range, also for most proteins, we captured the chevron rollovers at both folding arms and unfolding arms or either one under strong native conditions or unfolded conditions. Although the mechanism behind chevron rollovers are still not very clear, the results at least suggest that the local features of free energy landscape is might be a source under strong native conditions and unfolded conditions. This will need more delicate experiments and further theoretical work to resolve this issue.
CHAPTER 8

APPENDIX

8.1 Glossary of terms

- *Contact order*—a measure of the mean sequence length separating interacting residues in the native structure.

- *Energy landscape*—the network of all conformational states of a protein, with an internal free-energy associated with each conformation, and with the connectivity of the network specified or assumed implicitly.

- *Φ*-values— a measure of the partition of residues in the folding transition state ensemble. A particular residue is mutated to one interacting similarly with its neighbors, and the effect on folding rates and stability treated as a small perturbation. The ratio of the change in log rate to the change in stability is the Φ-value for that residue.

- *Folded state*— thermodynamic state dominated by the correct native conformation (the native state).

- *Unfolded state*—the high entropy state where protein function is lost, attributed to the loss of the 3D folded structure. The unfolded state may be globular or coil depending where parameters of the system lie on the phase diagram.
• *Misfolded state*—a protein configuration dominated by low-energy, non-native interactions

• *Folding temperature*—the temperature at which the native state is in thermodynamic equilibrium with the unfolded state, i.e. at the temperature the free-energy typically has a double well structure of nearly equal depths for a first-order folding transition.

• *Free-energy profile*—the free energy $E - TS$ as a function of an order parameter or parameters, which are usually taken to measure structural similarity to the native conformation. Serves as a reaction surface for folding.

• *Gō-model*—In the Gō-model, only the native interaction are favorable and all non-native interactions do not contribute to the Hamiltonian [257, 258].

• *Topology*—is defined by its set of native contacts, i.e. pairs of amino acid residues that are in close spatial proximity in the native structure. *Minimal frustration principle*—the interactions between residues that do not form contacts in the native structure are less important than those that do [22].

• *Folding nucleus*—Structured part of molecule in the transition state (on the top of the free energy barrier). Same as critical nuclei.
8.2 Folding path

In this section, we show the folding path for all of two-state proteins we studied here. The x-coordinate is order parameter $Q$, the y-coordinate is free energy $F$ in unit of $k_B T_f$. 
8.3 2D-plot of native density for cooperativity and noncooperativity

In this section, we list all of 2D-plots of native density for cooperativity and noncooperativity cases. The x-coordinate is $Q$, the y-coordinate is residue index.

1a0n

1aps

1c8c

1coa
Figure 28: 2D plots of native density for all of studied proteins. The horizontal axis is reaction coordinate Q, the vertical axis is residue index. Here the native density is normalized by native density at globule and native. Red corresponds to fully unstructured, blue corresponds to fully ordered state, the gradual color change between red and blue correspond to the structure change of each residue according to folding progress.
8.4 Folding pattern

In this section, we show the folding patterns for all of studied proteins. Here $\mu$ is packing fraction for folded core (solid line) and interface (dashed line); $n$ is the number of residues in folded core (solid line) and interface region (dashed line); $v$ is the volume of folded core (solid line) and interface (dashed line). $\mu$, $n$ and $v$ are all normalized by the corresponding $\mu$, $n$, and $v$ at native state. The protein names are shown in first column.
Pattern A:

1a0n 127
Pattern B:

1coa

1enh

1fkb

1hdn

1vii

1wit

2pdd
Pattern C:

1aps

1csp

1fnf9

1imq

1mef

1o6x

1srl

1ten
Exceptionals:

1div

1urn

2abd
8.5 Φ-value analysis

In this section, we will show the experimental Φ-values and calculated Φ-values for some proteins which have experimental Φ analysis. The x-coordinate is residue index and y-coordinate is Φ values. The red curves correspond to experimental Φ values, the green curves correspond to calculated Φ values.
8.6 Scaling exponent

In this section, we present the fitting of \( \ln V_f \) vs. \( \ln N_f \) for all of studied two-state proteins. The scaling exponent will be given by one third of fitting line slope.
Figure 29: Scaling exponents for studied proteins. Here, the fitting line (dashed line) is the best linear fitting between logarithm of folded volume ($V_f$) and logarithm of the number of folded residues ($N_f$) in most folded process.
8.7 Stability effects on temperature, transition state position, barrier height and prefactor

In this section, we show the stability effects on temperature, transition state position, barrier height (black curve: folding barrier; red curve: unfolding barrier.) and prefactor. For protein 1vii, it has two routes. The changing of transition state position (Qts) and prefactor ($k_0$) according to stability ($\log K_u$) are denoted by the red and black curves for two routes. The corresponding unfolding and folding barrier heights are denoted by blue (red) and black (green) curves.
In this section, we show the results for some proteins with chevron plots.

![Diagram of Chevron Plots](image)

*1a0n, 1c8c, 1div, 1enh, 1fnf9, 1imq, 1lmb, 1o6x, 1pin, 1pks, 1shg, 1srl, 2abd, 2pdd*
Bibliography


