UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Extending the theoretical framework of protein folding dynamics

Permalink

https://escholarship.org/uc/item/06g6657f

Author

Yang, Sichun

Publication Date

2006

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Extending the Theoretical Framework of Protein Folding Dynamics

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Physics (Biophysics)

by

Sichun Yang

Committee in charge:

Professor Herbert Levine, Chair
Professor José N. Onuchic, Co-Chair
Professor Terence T. Hwa
Professor Patricia A. Jennings
Professor J. Andrew McCammon

2006

Copyright
Sichun Yang, 2006
All rights reserved.

The dissertation of Sichun Yang is approved, and it is acceptable in quality and form for publication on microfilm:

_____

_____

_____

_____

Co-Chair

_____

Chair

University of California, San Diego

2006

TABLE OF CONTENTS

Signature Page . . . iii

Table of Contents . . . iv

List of Figures . . . vii

Acknowledgements . . . xiii

Vita and Publications . . . xiv

Abstract of the Dissertation . . . xv

1 Introduction . . . 1
1.1 The symmetrized Gō model . . . 3
1.2 An effective diffusive model . . . 4
1.3 Structural aspects of amyloid fibrils . . . 5
1.4 Overviews . . . 5

2 Protein Interaction and Self-assembly through Domain Swapping by using Symmetrized Gō Model . . . 6
2.1 Introduction . . . 7
2.2 Model and Methods . . . 9
2.2.1 Symmetrized Gō-type Hamiltonian . . . 9
2.2.2 The SH3 domain from Eps8 . . . 12
2.3 Results and Discussion . . . 12
2.3.1 Modeling a frustrated flexible n-src hinge region . . . 13
2.3.2 Monomer topology determines domain-swapped structures . 15
2.3.3 Mechanism of domain swapping . . . 15
2.4 Conclusions and Summary . . . 19

3 Protein Oligomerization through Domain Swapping: Role of Inter-molecular Interactions and Protein Concentration . . . 21
3.1 Introduction . . . 22
3.2 The model system . . . 25
3.3 Symmetrized structure-based (Gō) potential for self-assembly . . . 25
3.4 Biased replica exchange molecular dynamics simulations . . . 28
3.5 Reweighted histogram analysis method . . . 29
3.6 Results and Discussion . . . 30
3.6.1 Native-like monomer and domain-swapped dimer at equilibrium . . . 30
3.6.2 Large free energy barrier between monomer and domain-swapped dimer . . . 31

3.6.3 Second virial coefficient analysis characterizes intermolecular interactions and concentration-dependent protein oligomerization . . . 31
3.6.4 Connection with aggregation and amyloid formation . . . 33
3.6.5 Molecular self-assembly: Go beyond dimerization . . . 36
3.7 Conclusions . . . 38
3.8 Appendix: Derivation of the second virial coefficient . . . 39

4 Biological Implications: Structure of Infectious Prions Stabilized by Domain Swapping . . . 42
4.1 Introduction . . . 43
4.2 Hypothesis: A domain-swapped prion trimer is the minimal infectious unit . . . 44
4.3 Domain swapping: Motivation and Results . . . 44
4.4 Possible relevance to inherited diseases . . . 48
4.5 Proposed experimental tests . . . 49
4.6 Summary . . . 50
4.7 Appendix . . . 50

5 Effective Stochastic Dynamics on a Protein Folding Energy Landscape . 52
5.1 Introduction . . . 52
5.2 Theory and Methods . . . 54
5.2.1 The stochastic approach . . . 54
5.2.2 Langevin dynamics . . . 56
5.2.3 Reconstructing thermodynamics from the drifts and diffusion coefficients . . . 56
5.3 A simple test case: A two-state toy model . . . 57
5.4 Diffusion-drift motion on a funneled landscape . . . 64
5.5 Discussions and Conclusions . . . 67

6 Predicting the Folding Time from Replica Exchange Simulations . . . 71
6.1 Introduction . . . 71
6.2 Results . . . 73
6.2.1 Kinetics from replica exchange simulations . . . 73
6.2.2 Extracting reaction-coordinate-dependent drifts and diffusion coefficients . . . 75
6.2.3 Folding time prediction from Langevin simulations . . . 77
6.3 Discussion and Conclusions . . . 78
6.4 Theory and Methods . . . 79
6.4.1 Simulation details . . . 79
6.4.2 The stochastic approach . . . 80
6.4.3 Langevin dynamics . . . 81
6.4.4 Reconstructing thermodynamics from the drifts and diffusion coefficients . . . 82

7 Summary and Future Directions 83
7.1 The symmetrized Gō model 83
7.1.1 The successes 83
7.1.2 Limitations 84
7.2 The diffusive dynamics 84
7.2.1 The successes 84
7.2.2 Assumptions and simplifications 85
7.3 Future directions 86

Index 87

Bibliography 89

LIST OF FIGURES

2.1 Contact maps for Eps8 SH3 domain and the symmetrized Gō-type model. (**a**) The structure and contact map of Monomeric SH3 (MSH3, PDB ID 1I0C) of Eps8. (**b**) The structure and contact map of Dimeric SH3 (DSH3, PDB ID 1I07) of Eps8. (**c**) A schematic illustration for the symmetrized Gō-type model and the corresponding contact map for the two-chain system (see text for details). In this symmetrized Gō-type model contact map, the red and blue represents all the intra-chain contacts for chain A and B, respectively. The green are the native inter-chain contacts that are found in the domain-swapped structure of DSH3; the black are the non-native inter-chain contacts that are not found in the domain-swapped structure of DSH3. The structures are created with Molscript[64]. . . . 10

2.2 The flexibility of the n-src hinge loop. (**a**) The dihedral angle difference for C_α atoms between MSH3 and DSH3 shows that the large conformational change for domain swapping is located at the flexible n-src loop (residues 35-39). (**b**) Higher thermal B-factors for C_α atoms in the n-src and distal loops of DSH3 (data from PDB ID 1I07). 14

2.3 Symmetrized Gō-type model with flexible loops. Contact probability map for $Q_{interchain} > 100$ (above diagonal line) and symmetrized contact map (below diagonal line) for (**a**) the flexible n-src loop only, (**b**) the flexible (partly) RT- and n-src loops, and (**c**) the flexible n-src and distal loops. 16

2.4 Free energy landscapes for domain swapping, as a function of $Q_{intrachain}$ and $Q_{interchain}$ at five temperatures out of a eight-replica simulation with the interchain center-of-mass constraint $K = 0.1\varepsilon/r_0^2$ and $R_0 = 0.1r_0$. 17

2.5 Mechanism of domain swapping by symmetrized Gō potential. . . 19

3.1 The basic idea of the symmetrized structure-based (Gō) model is that all the native contact interactions existed in monomers should also exist in dimers or high order oligomers, *i.e.*, all the intramolecular interactions should also exist inter-molecularly. The intrachain and inter-chain interactions drive the formation of monomeric and oligomeric structures, respectively, from the ensemble of states of unfolded peptide chains. Meanwhile, a flexible hinge loop is very often involved and play a important role in transitions between monomeric and oligomeric configurations. 24

3.2 The native-like domain-swapped dimer is obtained by using the symmetrized structure-based model together with a flexible *n-src* hing loop, and the system undergoes the transitions between monomeric and domain-swapped dimeric conformations. Structures of (a) monomeric and (b) domain-swapped dimeric Eps8 SH3 domain. (c) The transition between native monomeric and domain-swapped dimeric configurations from a typical trajectory of the simulations at a temperature below the folding temperature, where red and blue represent the number of native contacts of chain A and chain B in MSH3, respectively; green represents the number of native contacts formed in the crystal DSH3. 26
3.3 Using the reweighted histogram analysis method (rWHAM), an unbiased potential of mean force or free energy profile (V_{eff}) is reached. As a comparison, two sets of biased REMD simulations are carried out with the harmonic constraints between two protein chains, $K = 0.08\varepsilon/r_o^2$ and $K = 0.10\varepsilon/r_o^2$, respectively. The dashed lines are plotted using the original WHAM without the reweighted; the solid lines, after using the rWHAM, show a nearly flat distribution at $R > 8r_0$. The plots are shown at a temperature $T = 0.94T_F$, where T_F is the folding temperature. 30
3.4 A large free energy barrier separates monomeric and domain-swapped dimeric conformations as shown from the effective interchain potential of mean force as a function of center-to-center distance R of two protein chains (Upper). The relative energy and entropy are also shown (Middle and Lower). The plots are shown at a temperature $T = 0.91T_F$, where T_F is the folding temperature. 32
3.5 Concentration-dependent protein oligomerization by using the second virial coefficient analysis. Above the folding temperature T_F, $\mathcal{B}_2$ is positive where proteins are unfolded. Below T_F, monomeric conformations are favored if protein concentration $c \ll c^*$, and dimeric conformations are favored if $c > c^*$ due to strong attractive interchain interactions. Temperature and concentration are in the units of T_F and $1/r_o^3$, respectively. 34
3.6 Direct simulations for two protein chains show how the population of monomeric and dimeric configurations change with thetemperature and center-to-center distance R. The result is consistent with that from the second virial coefficient analysisin Fig. 3.5. 35
3.7 A schematic illustration for molecular self-assembly by domain swapping. The open-end domain-swapped conformation can serve as templates and then domain-swap and form dimers, trimers, and even higher order oligomers. 36
3.8 Typical configurations of monomer (a), domain-swapped dimer (b), and domain-swapped trimer (c) are populated on the free energy surface (d) at a temperature below the folding temperature. 37

3.9 Oligomerization funnel: Unfolded proteins can fold into native monomeric or oligomeric conformations upon the competition of intra-molecular and inter-molecular interaction. On the other hand, monomer can cross a high free energy barrier to form oligomer under certain conditions, such as high protein concentration. 38

4.1 (A) Proposed β-helical structure of the truncated $PrP^{Sc}106$ prion trimer (BPT) taken from Ref. [38] (residues G89-F140). (B) Structural model of domain-swapped trimeric prion (DSTP) of the present paper. (C) The proposed hinge region expanded from the yellow loops in panels A and B. 44
4.2 MD simulations show that the nonswapped BPT (A) has less direct hydrogen bonds in the interface region than the domain-swapped DSTP (B). In both cases we have simulated residues G89-F140. Different colors correspond to three different 1 ns MD runs in each case initialized at different initial temperatures (see Appendix). . . 46
4.3 (A) Model for DSTP $PrP^{Sc}106$ (with residues G141-H176 deleted) and α-helices included. Residues 133-140 have been removed from the lower rung of the β-helices for each monomer, converted to random coil, and attached to residue 177 of the α-helices. (B) Number of direct hydrogen bonds for the DSTP structure with intact α-helices. The α-helices have no effect on the stabilization induced by the domain swapping. 47
4.4 The entanglement due to domain swapping introduces additional stability for the DSTP relative to the BPT as revealed by high temperature ($T = 500K$) simulations, where the radius of gyration of the BPT increases more rapidly than that of the DSTP, which reflects unbinding of the BPT. Principal component analysis at 300K reveals a similar unbinding tendency. 48

5.1 Schematic diagram of the diffusion-drift motion on a free energy landscape which is a function of an arbitrary reaction coordinate. The protein gets trapped in a local minimum, overcomes the barrier by diffusive motion, and finally folds downhill into the native state. The basic idea of the present paper is that we can determine, directly from a full simulation of the multi-dimensional microscopic dynamics, the drift and diffusion coefficients at each position along the reaction coordinate. We then simulate the Langevin dynamics with these drifts and diffusion coefficients, thereby characterizing the kinetic process. 53

5.2 (Top) The two-dimensional free energy surface of $F(X,Y) = [(X^2-1)^2 + (Y-1)^2] \times [1 + (Y+1)^2]$ (in units of k_BT) with two minima at $(X = -1, Y = 1)$ and $(X = 1, Y = 1)$, respectively. (Bottom) The free energy surface $F_{simulated}(X,Y) = -k_BT \log P(X,Y)$ computed from the two-dimensional Langevin simulations in Eq.(5.10 and 5.11) with constant $\Gamma = 1$ and $D = 1$ in both X and Y directions. The agreement between the simulated $F_{simulated}(X,Y)$ and theoretical $F(X,Y)$ validates our Langevin simulation protocol. 58

5.3 A typical example of the probability $P(X,t)$ as a Gaussian distribution centered at $X = -1.2$ at time $t = 0.002$ and $t = 0.005$. By measuring the shift of the center and growth of the width from the probability $P(X,t)$, we can determine the drifts and diffusion coefficients at each position along X. 59

5.4 A representative example for the centers X_c and the width square σ^2 as a function of time t at $X = -1.2$. For the two-step procedure for extracting the drifts and diffusion coefficients, due to the nonlinear dependency, we used a quadratic fitting with 4 data points at times $(0.002, 0.003, 0.004, 0.005)$ and then extrapolated drifts and diffusion coefficients at $t = 0$. 60

5.5 (Top) The one-dimensional free energy profile as a function of X, where $G(X)$ is computed from the kinetic approach from drifts and diffusion coefficients (Eq. 6.10) and $F(X) = -\log\left(\int e^{-F(X,Y)} dY\right)$ is the theoretical free energy profile by integrating out the Y dimension for $F(X,Y)$. (Middle and bottom) The position-dependent drifts $v(X)$ and diffusion coefficients $D(X)$ derived from the two-dimensional Langevin simulations. The bin size of $\delta X = 0.1$ was used at $t = 0$. 61

5.6 A typical one-dimensional Langevin simulation trajectory with simulated drifts $v(X)$ and diffusion coefficients $D(X)$ as shown in Fig. 5.5. A time step of $\delta t = 0.001$ was used. 63

5.7 The distribution of transition times by two-dimensional Langevin simulations (($\Gamma = 1$ and $D = 1$) and one-dimensional Langevin simulations with simulated drifts $v(X)$ and diffusion coefficients $D(X)$ (shown in Fig. 5.5). The distribution values have been normalized. Both simulations were run for 10^8 steps with a step size $\delta t = 0.001$. Both distributions are nearly Poisson, the standard deviations being approximately equal to the means. By using a single reaction coordinate X, the one-dimensional Langevin simulations robustly recover the two-dimensional transition-time distribution. 63

5.8 (Left) Cartoon representation of the SH3 domain (1PNJ.pdb). (Right) The reconstructed free energy profile and drifts $v(Q)$ and diffusion coefficients $D(Q)$ extracted from Gō-type simulations. $G(Q)$ is the free energy reproduced from $v(Q)$ and $D(Q)$ by using the kinetic approach (Eq. 6.10). $F(Q)$ is calculated from Gō-type simulations by $F(Q) = -\log P(Q)$ where $P(Q)$ is the probability distribution of Q. Q is the number of native contacts formed and $Q = 213$ for the native SH3 (using the Contacts of Structural Units software [113]). The Gō-type simulation was run at a temperature near the folding temperature. A total of 9×10^6 snapshots were used for the analysis. For the quadratic fitting, we picked four data points at times $(3, 4, 5, 6)$ (in units of the time interval between snapshots). The free energies are in units of k_BT. The bin size of $\delta Q = 3$ was used at $t = 0$. 66

5.9 A typical Langevin simulation trajectory for the number of native contacts Q with simulated $v(Q)$ and $D(Q)$ from the full Gō-type simulations. The time step $\delta t = 0.1$ was used for the Langevin simulations. 67

5.10 The folding time comparison between the Gō-type simulations and the Langevin simulations using Q as a single reaction coordinate, which remarkably reproduces the folding time within the same order of magnitude with the full Gō-type simulations. The Langevin simulations were run for 4×10^9 steps with a time step $\delta t = 0.1$. A total of 1,143 folding transitions from Langevin simulations and 277 folding events from full Gō-type simulations were used for comparison of folding time distribution. 68

6.1 The native structure for the second β-hairpin of the B domain of protein G. The parameter of $R = \sum_{i=1}^{5} \frac{r_i^o}{r_i}$ was used as a reaction coordinate to characterize the folding process, where r_i is defined as the distance between nearest C_α atoms (yellow dots) in the native structure (r_i^o). 73

6.2 A typical replica exchange simulation trajectory at a given target temperature. Shown is the replica index representing protein coordinates of simulation time. Clearly, the REMD simulations produce a large amount of short continuous trajectories (each segment of short trajectories is in order of picoseconds or tens of picoseconds), which provide *short-time* kinetic dynamics. The basic idea here is that the information of *short-time* dynamics can be derived from the replica exchange simulations and then used for further description of folding process. 74

6.3 The drifts and diffusion coefficients as a function of R at $T = 300\ K$. A linear fitting including data points at times $(0.4, 0.7)\ ps$ was used. The bin sizes of $\delta R = 0.05$ and $\delta t = 0.3\ ps$ were used to have a reasonable statistics. The error bars for $v(R)$ and $D(R)$ shown in brown were calculated according to Eqs. (6.5 and 6.6). The free energy profile $F(R)$ constructed from $v(R)$ and $D(R)$ are also shown by using Eq. 6.10. 76
6.4 A typical Langevin simulation trajectory describes the dynamic process for the single reaction coordinate R with determined drifts and diffusion coefficients. A time step $\delta t = 0.1\ ps$ was used for the Langevin simulation. 77
6.5 The simulated folding time is $(3.4 \pm 0.1)\ \mu s$ from 1,000 folding events by using Langevin simulations, in which the initial starting point is $R = 1.2$ and the system is considered to be folded when it reaches $R = 4.8$. And the experimental folding time is 6 μs. 78

ACKNOWLEDGEMENTS

I would sincerely like to thank my co-advisors Prof. Herbert Levine and Prof. José Onuchic for all the supports and consistent encouragement. I'm very grateful for fruitful discussions with Prof. Daniel Cox, Prof. Chris Dobson, Prof. Angel García, Prof. Peter Wolynes, Prof. Pat Jennings, and Dr. Michele Vendruscolo on a number of collaborative projects. Many thanks to Prof. Terence Hwa and Prof. Andy McCammon, who have given their valuable suggestions and comments. I have also enjoyed helpful discussions with Drs. Margaret Cheung, Shachi Gosavi, Peter McHale, Osamu Miyashita, Leslie Chavez, Xiaolin Chen, John Finke, Robert Kockeny, and former and current members of the Onuchic and Wolynes groups, the McCammon group, and CTBP. Thanks to Paul Whitford for careful reading of the abstract and the first and last chapters. Finally, I thank my wife and our parents, who have offered me unlimited freedom to develop myself and my career with their patience and trust.

Chapter 2 is adopted from Yang *et. al.* [132] as appeared in the *Proc. Natl. Acad. Sci. USA*, Sichun Yang, Samuel Cho, Yaakov Levy, Margaret Cheung, Herbert Levine, Peter Wolynes, and José Onuchic, 101(38):13786–13791 (2004). Chapter 3 is reprinted from Yang *et. al.* [133] in the *J. Mol. Biol.*, Sichun Yang, Herbert Levine and José Onuchic, 352:202–211 (2005). Chapter 4 is taken from Yang *et. al.* [134] in the *FASEB J.*, Sichun Yang, Herbert Levine, José Onuchic, and Daniel Cox, 19(13):1778–1782 (2005). Chapter 5 is largely reprinted from Yang *et. al.* [135], recently accepted in the *J. Chem. Phys.*, Sichun Yang, José Onuchic, and Herbert Levine. Chapter 6 is part of a manuscript in preparation with Herbert Levine and José Onuchic.

VITA

1998	B.S., Physics, Special Class Honors, Lanzhou University
2001	M.S., Physics, Peking University
2006	Ph.D., Physics (Biophysics), University of California, San Diego

PUBLICATIONS

Molecular modeling of the SH3 amyloid structure. S. Yang, et al. *in preparation*, 2006.

Folding time prediciton from all-atom simualtions. S. Yang, et al. *to be submitted*, 2006.

Effective stochastic dynamics on a protein folding energy landscape. S. Yang, J. N. Onuchic, and H. Levine. *J. Chem. Phys.*, in press, 2006.

Prion disease: exponential growth requires membrane binding. D. L. Cox, R. R. P. Singh, and S. Yang. *Biophys. J.*, 90(11):L77–79, 2006.

Structure of infectious prions: stabilization by domain swapping. S. Yang, H. Levine, J. N. Onuchic, and D. L. Cox. *FASEB J.*, 19(13):1778–1782, 2005.

Protein oligomerization through domain swapping: Role of inter-molecular interactions and protein concentration. S. Yang, H. Levine, and J. N. Onuchic. *J. Mol. Biol.*, 352:202–211, 2005.

Domain swapping is a consequence of minimal frustration. S. Yang, S. S. Cho, Y. Levy, M. S. Cheung, H. Levine, P. G. Wolynes, and J. N. Onuchic. *Proc. Natl. Acad. Sci. USA*, 101(38):13786–13791, 2004.

Exploration of unbound states by analytic continuation in the coupling constant method within relativistic mean field theory. S. Yang, S. G. Zhou, J. Meng, *Chin. Phys. Lett.* 18(2):196–198, 2001.

Deformed relativistic Hartree theory in coordinate space and in harmonic oscillator basis. S. G. Zhou, J. Meng, S. Yamaji, S. Yang, *Chin. Phys. Lett.* 17 (10):717–719, 2000.

ABSTRACT OF THE DISSERTATION

Extending the Theoretical Framework of Protein Folding Dynamics

by

Sichun Yang
Doctor of Philosophy in Physics (Biophysics)
University of California San Diego, 2006
Professor Herbert Levine, Chair

Proteins carry out most cell functions and thus it is important to understand protein dynamics, in practice, knowing how they fold and how they interact. Within the framework of energy landscape theory and the funnel concept, we used domain swapping as an example of protein interaction to demonstrate the physical principles that underlie protein interaction and aggregation. As a direct biological implication, a novel structural model for the infectious prion was proposed under the name of "domain-swapping trimeric prion". Along the line of course-grained modeling for protein interaction, we also utilized atomic simulations to study the detail of folding dynamics using β-hairpin as a model system. To reduce the high-dimensional folding problem, we introduced an "effective stochastic dynamics" approach to characterize the folding kinetics by using one or two functional reaction coordinates. This method provides a general framework to understand the dynamics of relevant biological systems.

1

Introduction

As one of the major players in the cell, proteins are not only the building blocks from which cells are built but they carry out nearly all cell functions [1, 95], Generally speaking, everything important in the cell is either a protein or results in proteins. Therefore, before we hope to know how life works, we must have a deep understanding of the structural and dynamic properties of proteins.

Proteins are by far one of the most structurally complex and functionally sophisticated molecules. A protein molecule is made from a long chain of amino acids, of which there are 20 types. Each protein folds into a particular three-dimensional shape, or conformation, determined by its amino acids sequence. Therefore, it is difficult to predict protein structure and folding dynamics in this extremely high-dimensional configuration space. The problem of deciphering the folding process of proteins given the primary sequence is known as the protein folding problem.

Understanding how a protein folds into its final folded conformation (native state) is one of major issues in biophysics. During the past decade, the field of protein folding has been theoretically and computationally focused. Energy landscape theory and the funnel concept have been shown to be very successful in helping us understand how a protein folds to its native state [91]. It is well established that proteins are minimally frustrated. The minimal frustration principle states that naturally occurring proteins have evolved to have sequences that efficiently folding to structurally organized ensemble of structures, with few traps arising from discordant energetic signals. As an application of the notion of minimal frustration funneled energy landscape models or Gō models, which include only interaction

that stabilizes the native structure, provide a clear picture of how a protein folds in general. Computer simulations of coarse-grained or all-atom models designed to implement the principle of minimal frustration have shown excellent agreement with experimental observations such as ϕ-value analysis of folding transition states.

However, there are still several possible questions related to the protein folding problem which need to be answered, including: (a) What is the nature of the dynamics or kinetics of protein folding starting from unfolded states? Analytic work of diffusive motion on the energy landscape was recognized by the originators of the landscape concept, but limited answer to this important question is beginning to emerge without considering the detailed characterization of local energy landscape environment. For example, it is still unclear how to predict the folding rate using realistic dynamics, in order to verify the diffusive model for the folding of proteins. (b) When the environment becomes crowded and crowding reduces the diffusional mobility of macromolecules, how do proteins interact and self-assemble? Can the same energy landscape principles associated with the folding of proteins into their monomeric conformations also describe how these proteins oligomerize into domain-swapped conformations? (Crowding can be more precisely termed the excluded volume effect, which emphasizes a physical nonspecific effect from steric repulsion. When crowding occurs, the effective concentration of each macromolecular species inside cells is greater than its actual concentration and this difference has kinetic and thermodynamic consequences for the properties of that macromolecule. This excluded volume effect for proteins come from two aspects: one is from crowding due to other macromolecules or other types of proteins and the other is from the same type of protein through self-assembly. Here we focus on the latter, the formation of homo-oligomers by self-association reactions such as domain swapping.)

In this dissertation, I attempted to provide answers to the two questions raised above. With these goals in mind, my work focused on the mechanism of domain swapping and the folding kinetics described as follow.

1.1 The symmetrized Gō model

Because proteins carry out their function while interacting with other cellular components, one may ask whether proteins have evolved to fold efficiently and whether proteins are optimized for selective binding. Domain swapping is an unconventional mechanism of oligomerization in which the structural elements, or "domains", of individual monomers are interchanged between identical partners by recruiting interactions that are crucial for stabilizing the protein in its monomeric form. The transition from monomers to a domain-swapped dimer must couple with at least partial unfolding to allow the swapping region to be exchanged. The energetic cost of breaking and replacing the swapping region interactions, which typically represent a large portion of the total contacts within the monomer, leads to an energetic barrier separating the monomeric and domain-swapped dimeric conformations. Domain swapping is an important protein assembly mechanism and has been suggested recently as a possible mechanism underlying protein aggregation and amyloid formation [72, 51, 112].

What determines the mechanism of domain-swapping proteins? To address this question, a "symmetrized Gō-type model" was introduced in this dissertation to study the mechanism of domain swapping and its role in amyloid formation. The basic idea of the symmetrized structure-based (Gō) model is that all the native interactions in monomers should also exist in dimers or higher-order oligomers, i.e., all the intra-molecular interactions should also exist inter-molecularly. The intra-chain and interchain interactions drive the formation of monomeric and oligomeric structures, respectively, from the ensemble of states of unfolded peptide chains. Technically, "biased" replica exchange simulations are introduced and then a re-weighting method is used in order to produce the correct potential of mean force for the dimerization process. We also characterize the role of protein concentration effects by using the second virial coefficient, a thermodynamic indicator of inter-molecular interactions. Furthermore, as a direct biological implication, we propose a structural candidate for the infectious prion, namely, the domain-swapped trimer prion.

1.2 An effective diffusive model

Computer simulations using reduced (e.g., Gō-like) and atomic molecular models have played a central role in the interpretation of experimental studies. Although reduced models have provided considerable insights into the overall mechanisms of protein folding, especially in helping to clarify the nature of folding funnels, intermediates, and transition states, they are limited in their ability to explore the detailed molecular aspects of folding. On the other hand, because of the large number of degrees of freedom in the protein folding problem and the rarity of folding events even in small model systems, all-atom simulations require considerable computational resources to obtain meaningful results. There are also some technical difficulties with kinetic studies using all-atom simulations as folding time-scales are often many orders of magnitude larger than the achievable simulation time.

A particularly powerful sampling technique for the calculation of free energy surfaces and other thermodynamic properties is replica exchange molecular dynamics (REMD) simulation. REMD has often been used to explore the folding/oligomerization energy landscape by improving barrier-crossing. It greatly enhances sampling compared to the conventional simulation methods, but it loses the *long-time* dynamics due to the randomness or discontinuity from frequent exchange between neighboring replicas. Therefore, in this sense no direct kinetic information can be obtained.

Can we use the power of replica exchange all-atom simulations to obtain not only the thermodynamics but also the kinetics of folding? To address the question, a new tool for folding kinetics, using stochastic reaction-coordinate dynamics, has been developed in this dissertation, in which the effective local drift velocities and diffusion coefficients are determined from microscopic simulation data. The resultant Langevin equation can then be used to directly simulate the folding process.

Our goal here is to apply this effective stochastic procedure to realistic computational simulations such as all-atom replica exchange simulations and make some realistic prediction on the folding kinetics. In this way we could fashion a lower-dimensional description of folding which is simpler and still quantitatively

accurate. It would be very valuable if we could get an estimate of the folding time distribution even when the longest obtainable simulation trajectories are shorter than the characteristic time scale on which folding events occur. Before we applied the approach to all-atom simulations and made some realistic predictions for folding time, we tested this approach by employing it to a toy two-state dynamical system and to a funnel-like structure-based (Gō-type) model (see Chapter 5). The folding time predictions agree very well with full simulation results in both reduced models. Therefore, from all-atom simulations for a β-hairpin as a simple model system, we derived the reaction-coordinate-dependent local drift velocities and diffusion coefficients provided by the *short-time* dynamics from all-atom REMD simulation data. The resultant folding time prediction for the β-hairpin peptide is consistent with experiments.

1.3 Structural aspects of amyloid fibrils

When a protein is under certain conditions, such as high concentration or low pH, it starts to misfold and forms amyloid fibrils. Although considerable progress has been made in the elucidation of amyloid fibril at a molecular level, little is known about their structure. Structural modeling on this aspect is in progress and will be communicated elsewhere.

1.4 Overviews

As indicated above, the outline of this dissertation is as follows. Chapters 2 and 3 describe the symmetrized Gō model and domain swapping. Chapter 4 shows the biological implications by proposing a new structural candidate for the infectious prion protein. Chapters 5 and 6 introduce the new stochastic method for kinetics. In addition, this dissertation also addresses general conclusions and provides directions for future studies in Chapter 7. Although the models presented here may contain assumptions and limitations, this dissertation does provide a theoretical framework to understand protein folding dynamics.

2

Protein Interaction and Self-assembly through Domain Swapping by using Symmetrized Gō Model

The same energy landscape principles associated with the folding of proteins into their monomeric conformations should also describe how these proteins oligomerize into domain-swapped conformations. We test this hypothesis using a simplified model for the Eps8 SH3 domain protein, both of whose monomeric and domain-swapped structures have been solved. The model, which we call the symmetrized Gō-type model, incorporates only information regarding the monomeric conformation in an energy function for the dimer in order to predict the domain-swapped conformation. A striking preference for the correct domain-swapped structure was observed, indicating that overall monomer topology is a main determinant of the structure of domain-swapped dimers. Further, we explore the free energy surface for domain swapping by using our model to characterize the mechanism of oligomerization.

2.1 Introduction

It is now well established that proteins are minimally frustrated[12]. The minimal frustration principle states that naturally occurring proteins have been evolutionarily designed to have sequences that achieve efficient folding to a structurally organized ensemble of structures with few traps arising from discordant energetic signals[69, 11, 90, 23, 89, 31]. A recent computational survey has reported that proteins that associate via traditional binding modes also have funnel-like energy landscapes designed to have specific binding[94]. Domain swapping is an unconventional mechanism of oligomerization in which the structural elements, or "domains", of individual monomers are interchanged between identical partners by recruiting interactions that are crucial for stabilizing the protein in its monomeric form[3, 4]. When domain swapping occurs, non-covalent interactions between the swapping region and the rest of the monomer are broken and replaced by nearly the same interactions with the other monomer(s). Is domain swapping a sign of frustration? To answer this question, we are led to ask whether domain-swapping proteins have evolved such that the same energy landscape that determines their folding into monomeric conformations also determines which way(s) they will domain-swap into oligomeric conformations. Currently, about 60 proteins have been identified as domain-swapped complexes. The phenomenon has been proposed to play a role in the regulation of function, aggregation, and misfolding, but its biological implications are still unclear[107, 71, 87, 103, 24, 25, 102].

By avoiding frustrating conflicts between different energetic biases, proteins have evolved to have a funneled energy landscape that optimizes native structure-seeking interactions while selecting against interactions leading to traps[123, 86]. A funneled landscape reduces the highly intractable problem of configurational search through traps to a much smaller parallel search of the configuration space. Once the energetic frustrations associated with conflicting interactions have been minimized, the topology of protein becomes the key determinant of the folding mechanism, encoding the interplay between stabilizing free energies and chain entropy. Structure-based Gō-type minimalist models containing attractive native contacts and repulsive non-native contacts correspond to perfectly funneled energy landscapes. Such models have impressively reproduced the experimentally

observed folding mechanism of many single domain proteins[89, 19]. The quantitative form of minimal frustration of proteins has been successfully used to improve protein structure prediction energy functions[93] and for protein design[53]. Since proteins carry out their function while interacting with other cellular components, one may also ask whether proteins have evolved not only to fold efficiently but also may be optimized for selective binding.

What determines the mechanism of domain-swapping proteins? Is native topology a sufficient predictor? The monomeric and dimeric structures of a domain-swapping protein are expected to be comparable in the energies per molecule because the contacts are largely conserved between the structures[107]. The transition from monomers to a domain-swapped dimer must couple with at least partial unfolding to allow the swapping region to be exchanged. The energetic cost of breaking and replacing the swapping region interactions, which typically represent a large portion of the total contacts within the monomer, leads to a high-energy barrier separating the monomeric and domain-swapped dimeric conformations[3, 4, 102, 109]. Thus, it may be surprising that any protein would undergo such a large barrier transition.

Several previous studies have searched for distinguishing features common among domain-swapping proteins. However, there does not appear to be any sequence homology or secondary structure similarity that is found in all domain-swapping proteins[71]. Are there local signals for swapping? In some cases, proline residues are found in the hinge loop, suggesting their possible importance[6, 103], but there are many examples of domain-swapping proteins with no prolines in the hinge loop. Thus a unifying explanation for how proteins domain-swap has been elusive. In this present study, our goal is to determine whether the monomeric protein topology alone is sufficient for predicting how a protein will form complexes via the domain-swapping mechanism. To address this issue, we apply an extension of the usual Gō-type model, which we call the symmetrized Gō-type potential to Eps8 SH3 domain[57], a small domain-swapping protein. The model is based exclusively on the N native interactions that exist in the monomeric structure. In the symmetrized Gō-type potential for a domain-swapped dimer, the same native contacts that define the N intra-molecular non-covalent interactions in the native monomeric conformation also define 2N possible intra-molecular interactions and

2N possible inter-molecular non-covalent interactions, resulting in 4N total represented. It immediately follows that such a potential has the possibility of many topological traps and does not have a perfectly funneled energy landscape as does the traditional Gō-type model for monomeric proteins[89, 19]. There is frustration between the choice of whether to make any contact internally or with a corresponding partner's amino acid residue. The definition of the symmetrized Gō-type potential alone provides no information that would bias one swapping region over another. In principle, any region can swap, and nothing precludes the possibility of even multiple swapping regions, allowing for numerous traps. It is only if there is an inherent bias towards a single, intertwined conformation that we can say there exist, encoded in the monomer topology, instructions on how to domain-swap into an oligomer.

2.2 Model and Methods

2.2.1 Symmetrized Gō-type Hamiltonian

Here we introduce a symmetrized Gō-type potential for a two-chain protein. This model is based on the traditional Gō-type potential, a simplified C_α protein model. The potential is largely defined by attractive native contacts and repulsive non-native contacts, as well as the backbone geometry, to reflect a perfectly funneled energy landscape with no energetic frustration for the monomer so that it will fold into only its native conformation. In the symmetrized Gō-type potential, the native contacts in the monomeric structure serve to define both the *intra*-chain interactions and the *inter*-chain interactions. As illustrated by the cartoon (Fig. 2.1c), for each intra-chain interaction between residues i and j in chain A ($i^{'}$ and $j^{'}$ in chain B), the corresponding inter-chain interaction between residues i and $j^{'}$ (and $i^{'}$ and j) also exists.

The energy function of the symmetrized Gō-type potential for the two-chain

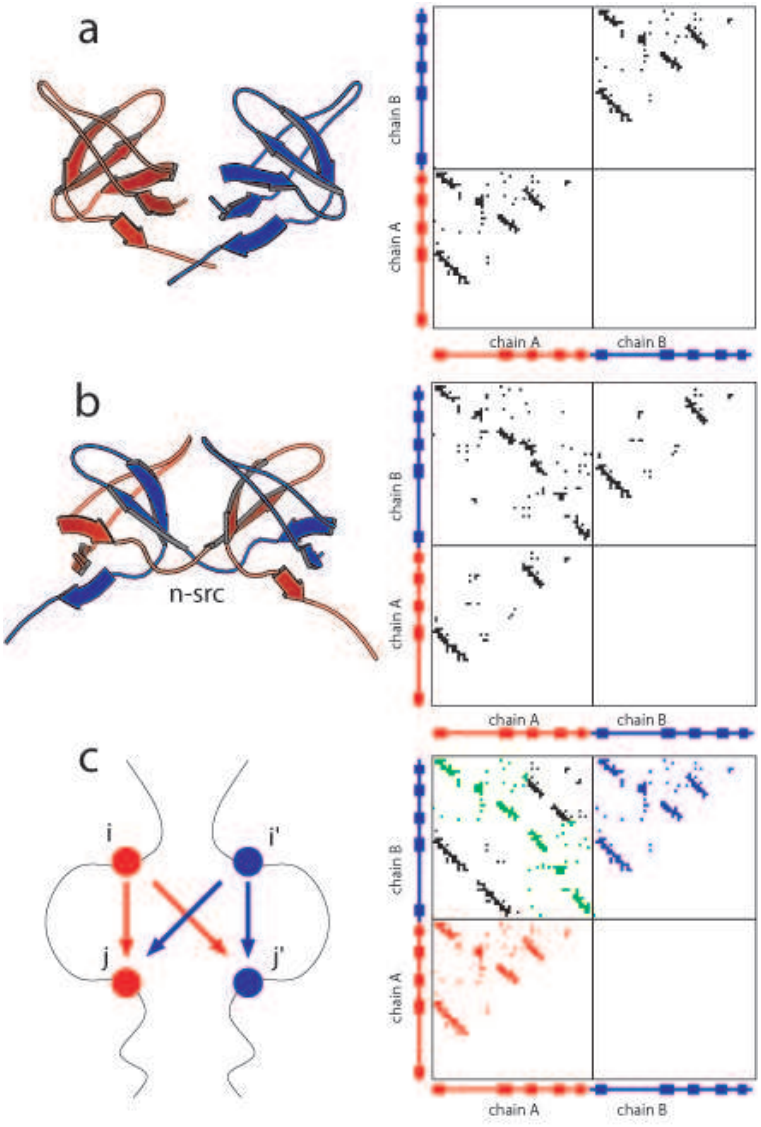

Figure 2.1 Contact maps for Eps8 SH3 domain and the symmetrized Gō-type model. (**a**) The structure and contact map of Monomeric SH3 (MSH3, PDB ID 1I0C) of Eps8. (**b**) The structure and contact map of Dimeric SH3 (DSH3, PDB ID 1I07) of Eps8. (**c**) A schematic illustration for the symmetrized Gō-type model and the corresponding contact map for the two-chain system (see text for details). In this symmetrized Gō-type model contact map, the red and blue represents all the intra-chain contacts for chain A and B, respectively. The green are the native inter-chain contacts that are found in the domain-swapped structure of DSH3; the black are the non-native inter-chain contacts that are not found in the domain-swapped structure of DSH3. The structures are created with Molscript[64].

(chain A and chain B) protein with configuration Γ is as follows:

$$
\begin{aligned}
E_{\text{total}}(\Gamma,\Gamma_0) &= E_{backbone} + E_{intrachain} + E_{interchain} \\
E_{backbone} &= \sum_{\text{bonds}} K_r(r-r_0)^2 + \sum_{\text{angles}} K_\theta(\theta-\theta_0)^2 + \sum_{\text{dihedrals}} K_\phi^{(n)}[1-\cos(n(\phi-\phi_0))] \\
E_{intrachain} &= \sum_{chainA}^{i<j-3} \left\{ \epsilon_1(i,j) \left[5\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - 6\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{10} \right] + \epsilon_2(i,j)\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} \right\} \\
&\quad + \sum_{chainB}^{i'<j'-3} \left\{ \epsilon_1(i',j') \left[5\left(\frac{\sigma_{ij}}{r_{i'j'}}\right)^{12} - 6\left(\frac{\sigma_{ij}}{r_{i'j'}}\right)^{10} \right] + \epsilon_2(i',j')\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} \right\} \\
E_{interchain} &= \sum_{A\to B}^{i<j'-3} \left\{ \epsilon_1'(i,j') \left[5\left(\frac{\sigma_{ij}}{r_{ij'}}\right)^{12} - 6\left(\frac{\sigma_{ij}}{r_{ij'}}\right)^{10} \right] + \epsilon_2(i,j')\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} \right\} \\
&\quad + \sum_{B\to A}^{i'<j-3} \left\{ \epsilon_1'(i',j) \left[5\left(\frac{\sigma_{ij}}{r_{i'j}}\right)^{12} - 6\left(\frac{\sigma_{ij}}{r_{i'j}}\right)^{10} \right] + \epsilon_2(i',j)\left(\frac{\sigma_{ij}}{r_{i'j}}\right)^{12} \right\} \quad (2.1)
\end{aligned}
$$

The local backbone interaction $E_{backbone}$ applies to both chain A and chain B; K_r, K_θ and K_ϕ are the force constants of the bond, angle, and dihedral angle, respectively. Note that r_0, θ_0 and ϕ_0 are the corresponding values taken from the native monomeric configuration Γ_0 only. The contact interactions, $E_{intrachain}$ and $E_{interchain}$, contain Lennard-Jones 10-12 terms for the non-local "native" intrachain and interchain contacts and short-range repulsive terms for "non-native" pairs.

As a starting point for the energy function, we chose $K_r = 100\varepsilon$, $K_\theta = 20\varepsilon$. $K_\phi^{(1)} = \varepsilon$ and $K_\phi^{(3)} = 0.5\varepsilon$. $\epsilon_1 = \epsilon_1' = \varepsilon$. $\epsilon_2 = 0.001\varepsilon$. σ_{ij} is equal to the distance between the pair of residues (i,j) in the native monomeric configuration and $\sigma_{ij} = 1$ (in units of $r_0 = 3.8\mathring{A}$) for all non-native residue pairs.

We performed molecular dynamics (MD) simulations with the replica exchange method by periodically exchanging configurations that are running at a sequence of temperatures[120, 140, 105]. A wide range of temperatures are chosen to ensure sufficient sampling from folded to unfolded states and provide reasonable overlap of energy histograms of neighboring pairs. All replica exchange simulations start from the native monomeric configurations. We imposed an interchain constraint $E_{constraint} = K(R-R_0)^2$ for the center-of-mass distance of the two chains R with a target distance R_0 to restrict the sampling of our simulations to a configuration space where domain swapping can be observed.

2.2.2 The SH3 domain from Eps8

For illustration, we chose an SH3 domain of the epidermal growth factor receptor pathway substrate (Eps8), which has been observed in both the monomeric and dimeric domain-swapped conformations[57]. The monomeric SH3 (MSH3) of Eps8 is a relatively small protein, formed by two orthogonal β sheets connected by a hinge named the n-src loop (Fig. 2.1a). The dimeric SH3 (DSH3) is formed by domain swapping, in which a part of secondary structure is exchanged between two protein chains. (Fig. 2.1b). The two structures have essentially identical secondary and tertiary structures except for the n-src loop, which has a closed form in the MSH3 and an open form in the domain-swapped DSH3.

To construct the symmetrized Gō-type potential for this protein, we determined the native contacts of MSH3 using the Contacts of Structural Units (CSU) software[113]. The list of the native contacts in the monomeric conformation was used as a template for all intrachain and interchain interactions as follows: For each intrachain interaction between residues i and j in chain A ($i^{'}$ and $j^{'}$ in chain B), we allowed interchain interactions by introducing the same interactions between residues i and $j^{'}$ ($i^{'}$ and j). The symmetrized contact map, illustrating the interactions represented in the symmetrized Gō-type potential for Eps8 SH3, are shown in Fig. 2.1c. Note that the map of interchain contacts includes both those that are found in the dimer x-ray structure (we call these "native interchain contacts") and those not observed in the dimer ("non-native interchain contacts").

2.3 Results and Discussion

Utilizing the symmetrized Gō-type potential, we are now able to determine how domain-swapping proteins undergo a transition from a monomeric to a domain-swapped conformation. The environment of a given residue is largely the same in monomeric or domain-swapped configurations, justifying our symmetrization of the intra-chain interactions to form the inter-chain interactions. A consequence of defining such a symmetrized energy function is that there exist competing energetic and topological states ("topological frustration"). There is no a priori guarantee that a single intertwined oligomer structure will be predicted by the symmetrized

Gō-type potential or that it will have the same domain-swapped structure found in nature. Unlike the intra-chain interactions, which yields a perfectly funneled energy landscape for the monomer, the inter-chain interaction leads to "ruggedness" for the dimer in that there exist favorable interactions that are not found in the domain-swapped structure. If a bias exists towards a single, stable domain-swapped conformation, it must be determined strictly from the energy landscape that folds the protein into a monomeric conformation.

2.3.1 Modeling a frustrated flexible n-src hinge region

There exists a very high energetic barrier for dimerization via domain swapping because it is involved in a large conformational change which occurs at the n-src loop (Fig. 2.1). In the "vanilla" version of the present model, all of the dihedrals are energetically biased to be those obtained from the monomeric conformation. The dihedral angle difference for C_α atoms between MSH3 and DSH3 clearly shows a large conformational change localized at the n-src loop where domain swapping occurs (Fig. 2.2a). The hinge region, the n-src loop, is a priori biased to remain in the monomeric conformation, not to convert to the dimer.

To quantify the significance of frustration in the hinge region of the dimer, we first performed replica exchange MD simulations. The simulations were carried out with the interchain center-of-mass constraint of $K = 1.0\varepsilon/r_0^2$ and $R_0 = 0.1r_0$. Even with the frustrated hinge, the native-like domain-swapped structures are rather well-represented and stable. The symmetrized Gō landscape favors the native interchain contacts rather than those not seen in the experimental dimer structure. Topological trap states (i.e. structures that appreciably involve contacts that are non-native but allowed by the symmetrized Gō model) exist, but are found to occur in low frequency.

To effectively account for the conformational change energetics at the hinge region, we first introduced a flexible n-src loop by turning off the dihedral terms in the n-src loop ($K_\phi^{(n)} = 0$). Therefore, this loop can rotate in any orientation and freely access the monomeric and dimeric configurations. This flexibility of the n-src loop is also consistent with the observed thermal B-factors in DSH3 (Fig. 2.2b). We found that allowing the n-src loop to be flexible stabilizes the domain-swapped

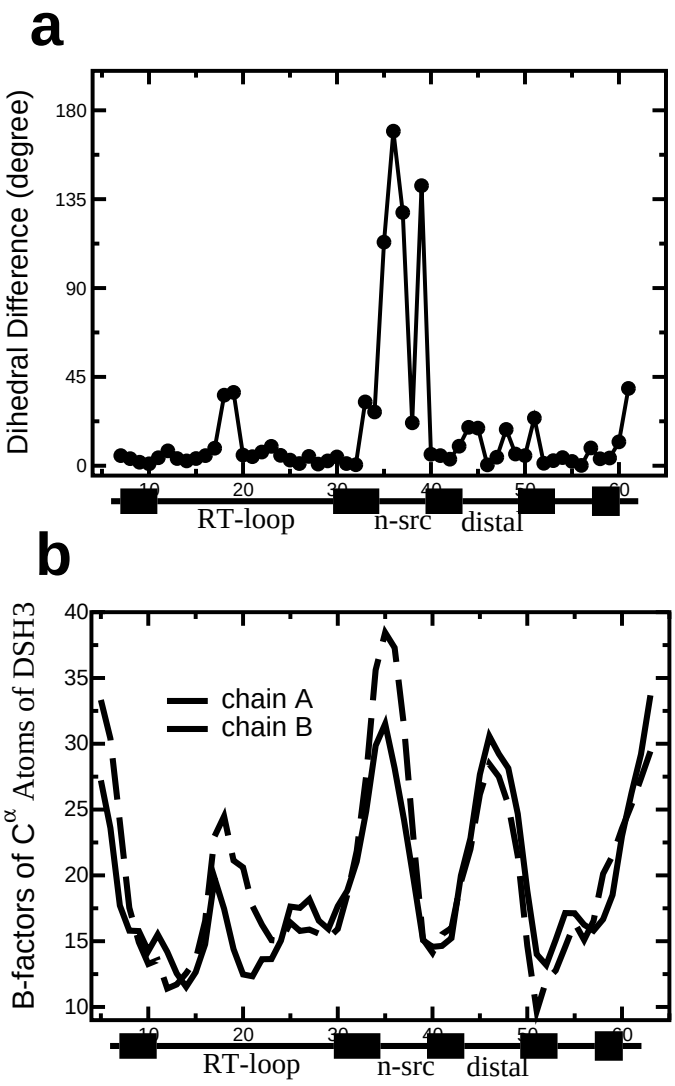

Figure 2.2 The flexibility of the n-src hinge loop. (**a**) The dihedral angle difference for C_α atoms between MSH3 and DSH3 shows that the large conformational change for domain swapping is located at the flexible n-src loop (residues 35-39). (**b**) Higher thermal B-factors for C_α atoms in the n-src and distal loops of DSH3 (data from PDB ID 1I07).

dimeric configuration by eliminating most of the trap states (as reflected by the low probability of forming non-native interchain contacts), as well as removing the bias that favors the monomeric configuration (Fig. 2.3a).

To determine whether the local properties of the loop determine the swapped structure, we also allowed the other loops (specifically the RT- and distal loops) to be flexible (Fig. 2.3b,c). The contact probability map for the case of flexible (partly) RT- and n-src loops shows that the native domain-swapped dimer is still the most dominant (Fig. 2.3b). The same is true for the case of flexible distal and n-src loops (Fig. 2.3c), but, when all loops are flexible more trap states were found. Nevertheless, Figure 2.3 shows that the domain-swapped conformation actually observed in the dimer crystal structure is clearly the most represented.

2.3.2 Monomer topology determines domain-swapped structures

Figure 2.3 also indicates that the monomer topology, rather than special features of the n-src loop, determines which of the domain-swapped dimeric structures will form. However, the yield of domain-swapped structures is, as we shall see, sensitive to the properties of the loop, which is consistent with the observed effects of mutations in the hinge region of p13suc1[103]. Mutations can affect the equilibrium and kinetics of swapping, yet the actual structure of the domain-swapped dimer is determined by the monomer topology.

2.3.3 Mechanism of domain swapping

The model with the flexible hinge loop has a lower energetic barrier from the one in which the hinge dihedral energy must be frustrated to swap. The lower barrier allows enhanced sampling of swapping transition state configurations, so the symmetrized Gō-type potential with the flexible n-src loop allows us to investigate the mechanism of domain swapping. Figure 2.4 shows the free energy surfaces at various temperatures as a function of $Q_{intrachain}$ and $Q_{interchain}$, the number of intrachain contacts and interchain contacts. These free energy surface maps are calculated using $F = -k_B T \log P(Q_{intrachain}, Q_{interchain})$, where $P(Q_{intrachain}, Q_{interchain})$ is the probability distribution of the intrachain and inter-

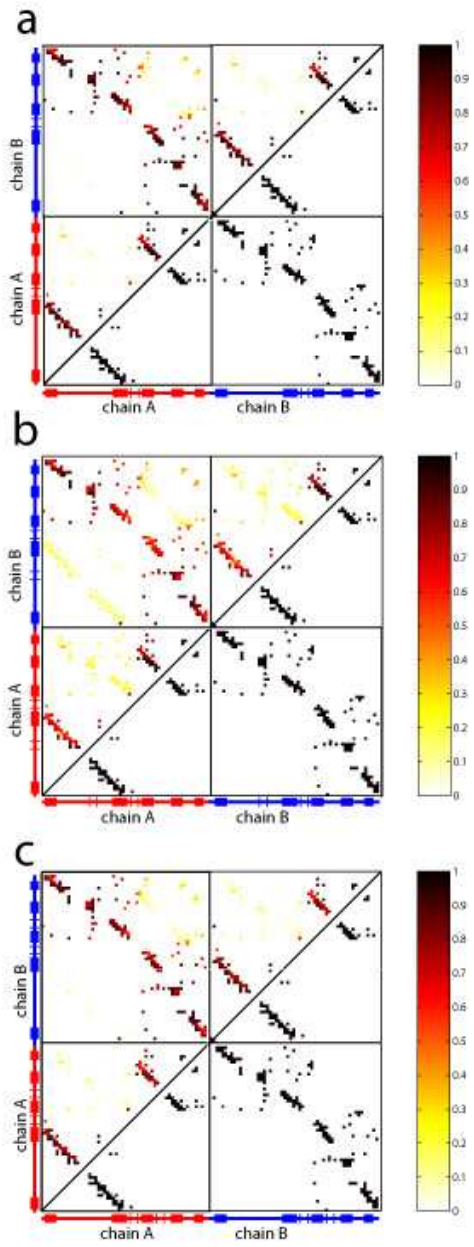

Figure 2.3 Symmetrized Gō-type model with flexible loops. Contact probability map for $Q_{interchain} > 100$ (above diagonal line) and symmetrized contact map (below diagonal line) for (**a**) the flexible n-src loop only, (**b**) the flexible (partly) RT- and n-src loops, and (**c**) the flexible n-src and distal loops.

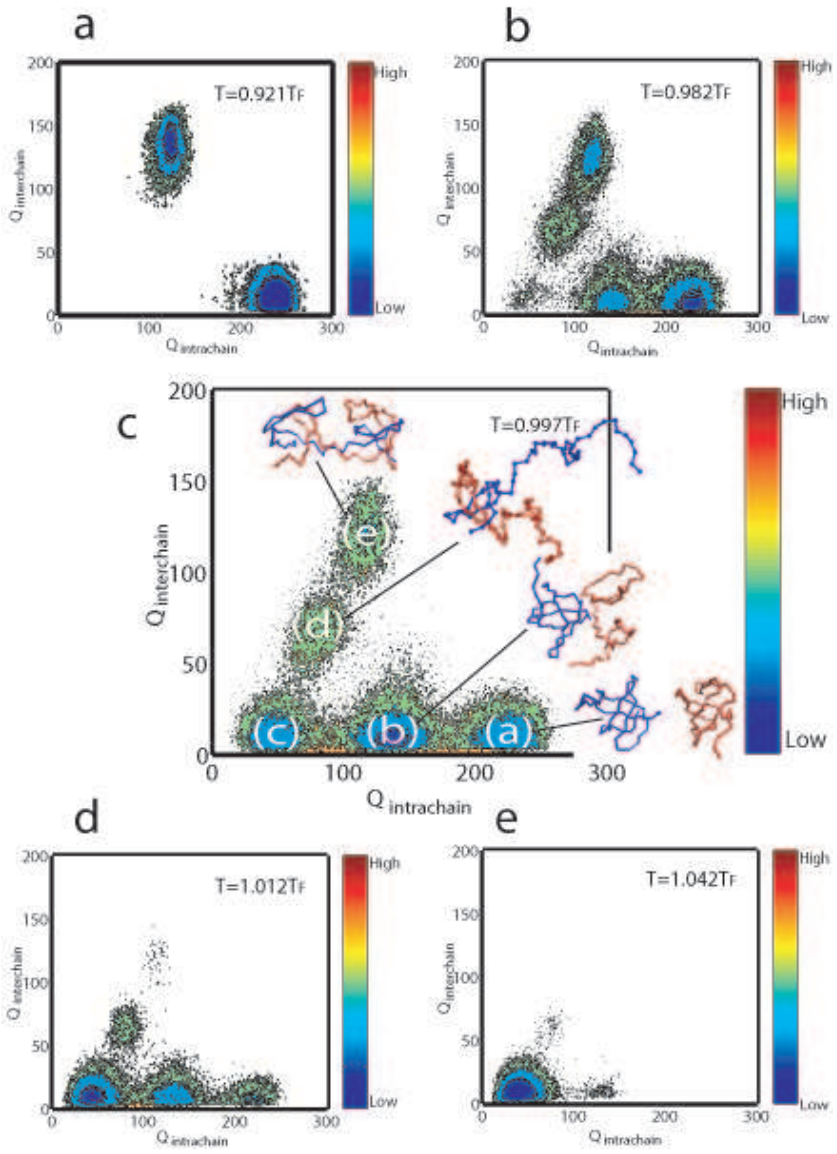

Figure 2.4 Free energy landscapes for domain swapping, as a function of $Q_{intrachain}$ and $Q_{interchain}$ at five temperatures out of a eight-replica simulation with the interchain center-of-mass constraint $K = 0.1\varepsilon/r_0^2$ and $R_0 = 0.1r_0$.

chain contact formation. At a very low temperature $T = 0.921T_F$ (Fig. 2.4a), where T_F is the folding temperature ($k_BT_F/\varepsilon = 1.324$), there exist two basins corresponding to the native-like monomeric and domain-swapped conformations. The replica exchange method enables us to find domain-swapped dimers even at low temperatures, while the regular constant temperature simulation yields monomeric conformations only. As the temperature increases, the two most probable ensembles of native-like structures tend to unfold (Fig. 2.4b). When T approaches T_F (Fig. 2.4c), five dominating free energy minima are observed with typical structures shown in Fig. 2.4c(a)-(e). These can be identified as native monomers (Fig. 2.4c(a)), partially unfolded monomers (Fig. 2.4c(b)), unfolded monomers (Fig. 2.4c(c)), partially folded or open-ended domain-swapped dimers (where approximately half of each monomer forms inter-chain contacts with its partner, and the rest remains unfolded) (Fig. 2.4c(d)), and domain-swapped dimers (Fig. 2.4c(e)). We infer that in this case the most favorable mechanism of domain swapping is: Native monomers $\rightleftharpoons$ Partially folded monomers $\rightleftharpoons$ Unfolded monomers $\rightleftharpoons$ Open-end domain-swapped dimers $\rightleftharpoons$ Domain-swapped dimers. At higher temperatures, unfolded structures become populated on the surface (Fig. 2.4d,e).

The importance of unfolded states is shown in Fig. 2.4c and Fig. 2.5. These unfolded states serve as free energy traps in the process of dimerization at the folding temperature. Thus, the native monomers unfold first and then form domain-swapped dimers. This result agrees with the experimental observation that domain swapping takes place from the unfolded state in many proteins, such as human prion and p13suc1[61, 109, 103, 104]. When proteins are unfolded, the competition between the intrachain and interchain interactions will lead proteins to monomeric or dimeric configurations, respectively.

In our simulations, the intermediate states, which resemble open-ended domain-swapped dimers (Fig. 2.4c(d)), are obligatory for monomers to access the domain-swapped form since the intermediate state is located midway between the unfolded state and the native-like domain-swapped dimeric state on the free energy surface shown in Fig. 2.4c. Under certain conditions such as an increase of protein concentration, these intermediates can be populated. As protein concentration is further increased, the interchain interactions will dominate over the intrachain one, and protein aggregates could be thus formed from the templates of the intermediates.

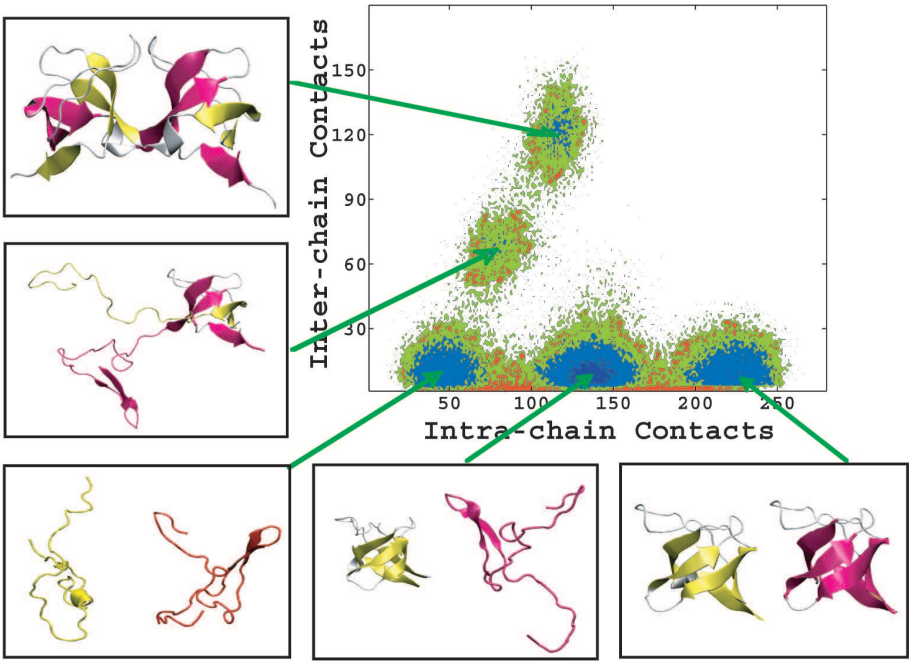

Figure 2.5: Mechanism of domain swapping by symmetrized Gō potential.

The partially folded intermediates thus serve as "templates" for self-assembled aggregates.

2.4 Conclusions and Summary

We have introduced and implemented the symmetrized Gō-type potential to study the mechanism of domain swapping, using Eps8 SH3 domain as a test case. The observation of the formation of the native-like domain-swapped dimer demonstrates that single domain proteins are evolutionarily designed not only for efficient folding, but can also lead to specific structures from oligomerization via domain swapping. The domain-swapped dimer actually observed in nature was found to be the most populated and the most stable in our simulations. The success of the symmetrized Gō-type model in predicting the domain-swapped structure indicates that the overall monomeric topology, rather than local signals in the hinge regions, determines where domain swapping can occur.

Acknowledgments

S.Y. thanks Leslie Chavez for critical reading of the manuscript. This work has been supported by the Center for Theoretical Biological Physics through National Science Foundation Grants PHY0216576 and PHY0225630. Computing resources were supported in part by W. M. Keck Foundation.

The chapter or portion thereof has been published in *Proc. Natl. Acad. Sci. USA*, Sichun Yang, Sam Cho, Koby Levy, Margaret Cheung, Herbert Levine, Peter Wolynes, and Jose Onuchic, 101(38):13786–13791, 2004 [132].

3

Protein Oligomerization through Domain Swapping: Role of Inter-molecular Interactions and Protein Concentration

Domain swapping has been shown to be an important mechanism controlling multi-protein assembly and has been recently suggested as a possible mechanism underlying protein aggregation. Understanding oligomerization *via* domain swapping is therefore of theoretical and practical importance. By using a symmetrized structure-based (Gō) model, we demonstrate that in the free-energy landscape of domain swapping, a large free-energy barrier separates monomeric and domain-swapped dimeric configurations. We investigate the effect of finite monomer concentration, by implementing a new semi-analytical method which involves computing the second virial coefficient, a thermodynamic indicator of inter-molecular interactions. This method, together with the symmetrized structure-based (Gō) model, minimizes the need for expensive many-protein simulations, providing a convenient framework to investigate concentration effect. Finally, we perform direct simulations of domain-swapped trimer formation, showing that this modeling approach can be used for higher-order oligomers.

3.1 Introduction

The conversion of monomeric proteins to oligomers by domain swapping has been proposed as one possible mechanism underlying the formation of early aggregates of amyloid fibrils [3, 49, 71, 66, 27, 110, 24, 132, 134]. For example, recent experiments on cystatin C show that prevention of domain swapping inhibits amyloid fibrils by 80% [88], and strongly support the idea that domain swapping has an important role in fibril formation. Domain swapping is a term used to describe a process wherein two or more protein chains exchange identical structural elements or "domains" [3]. The exchangeable domain is usually appended to a flexible hinge region and this hinge allows the placement of the swapped region in both intra-molecular and inter-molecular configurations. More than 40 proteins have been found to form domain-swapped configurations, such as the human prion [61, 67], the SH3 domain [58, 57], p13suc1 [56, 109, 103, 104], cystatins [117, 118, 51, 106, 50], and interleukin-10 and interferon γ [138, 79]. These domain-swapped proteins display diverse sequences and secondary structures of the swapped domains, and flexibility and diversity in the hinge regions. This diversity suggests that almost any protein may oligomerize *via* domain swapping [71, 87].

Often, domain swapping occurs at high protein concentration where a monomer becomes less favorable entropically and proteins have a tendency to aggregate. This may happen naturally in vivo given that crowding occurs in the cell [82, 29], which can trap certain proteins into a region with a high effective concentration. Thus the effect of concentration needs to be taken into account in the process of protein oligomerization and aggregation.

A straightforward computational approach to study concentration effects is to simulate multiple monomers in a large box. Clearly, this is computationally expensive. To overcome this difficulty and to reduce computational cost, we introduce a new semi-analytical technique that involves calculating the second virial coefficient for the monomer gas. The second virial coefficient has been used as a tool to understand protein-protein interactions in protein aggregation [35, 81, 8], as a particular case of colloidal assembly [131, 130, 124]. A negative second virial coefficient indicates a long-range attractive protein-protein interaction and a positive one indicates a repulsive interaction. Experiments also indicates the utility

of the second virial coefficient in helping to understand how certain conditions (such as temperature, protein concentration, ion concentration, and pH) promote or prevent aggregation [18, 47, 36].

Funnel-like energy landscapes have been successful in describing the configurational ensembles for a protein during folding toward its native basin [12, 69, 11, 91, 23, 86, 89, 19, 96]. Small/intermediate-size proteins are sufficiently well-designed that native interactions are commonly much more attractive than non-native interactions. Generalizing the concept of funnels for domain swapping, we previously used a symmetrized structure-based (Gō) model [132], to investigate how the competition between intrachain and interchain interactions leads proteins from unfolded states to monomeric or dimeric configurations, respectively (as illustrated in Fig.3.1). In this model, native contacts which exist in the monomer are also allowed between the same residues but now on different chains. We observed that the structure of the native-like domain-swapped dimer is determined by the overall monomeric topology and then characterized the mechanism of domain swapping on the free-energy landscape.

In this paper, we focus both on the concentration effect for dimerization and also on the formation of high-order oligomers, the latter with direct implications for amyloid formation. The following sections describe: the symmetrized structure-based (Gō) potential for protein self-assembly; the application of the biased replica exchange molecular dynamics (MD) simulations and reweighted histogram analysis method to the characterization of the domain-swapping free-energy landscape of the SH3 domain; and a second virial coefficient analysis introduced to investigate the effect of inter-molecular interactions and protein concentration and to estimate the critical concentration required for dimer formation. The results show that (1) there is a large free-energy barrier separating monomeric from domain-swapped dimeric conformations, and that (2) protein dimerization is highly dependent on temperature and concentration. We subsequently used the a generalized form of the symmetrized potential to study higher-order oligomer formation. The successful simulation of cyclic domain-swapped trimer formation indicates that we now have the tools in hand to investigate protein oligomerization through self-assembly.

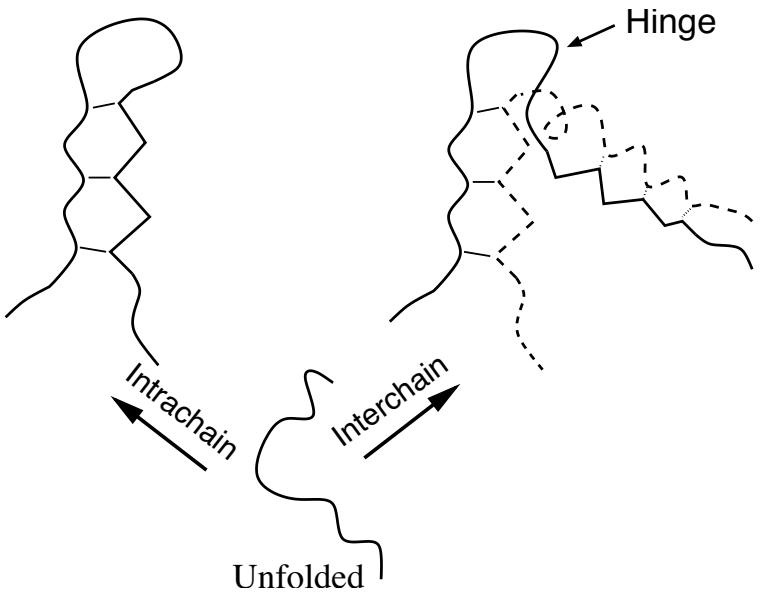

Figure 3.1 The basic idea of the symmetrized structure-based (Gō) model is that all the native contact interactions existed in monomers should also exist in dimers or high order oligomers, *i.e.*, all the intra-molecular interactions should also exist inter-molecularly. The intra-chain and inter-chain interactions drive the formation of monomeric and oligomeric structures, respectively, from the ensemble of states of unfolded peptide chains. Meanwhile, a flexible hinge loop is very often involved and play a important role in transitions between monomeric and oligomeric configurations.

3.2 The model system

We choose the SH3 domain of Eps8 (epidermal growth factor receptor pathway substrate 8) as our model protein because both the monomeric and dimeric domain-swapped structures have been solved and deposited into the PDB [58, 57]. The monomeric SH3 domain (MSH3, 1I0C.pdb) of Eps8 is formed by two orthogonal β sheets connected by a hinge called the n-src loop. The dimeric SH3 (DSH3, 1I07.pdb) is formed by domain swapping, in which an identical part of secondary structure is exchanged at the hinge region between two protein chains. The two structures have essentially identical secondary and tertiary structures except for the hinge loop, which has a closed form in the MSH3 and an open form in the domain-swapped DSH3 (Fig.3.2(a) and (b)).

3.3 Symmetrized structure-based (Gō) potential for self-assembly

The symmetrized structure-based model, which extends monomeric native interactions into dimers, has previously been introduced and implemented to study the mechanism of domain swapping [132]. Here we generalize this model to investigate protein self-assembly into higher-order oligomers, again through domain swapping. The native contacts in the monomer structure serve to define both the intrachain interactions and the interchain interactions (as shown in Fig.3.1); the transition from monomers to oligomers is then made by the dynamical choice of interchain interactions over intrachain interactions. In addition, the conformational change involved in the oligomerization is made possible by introducing a flexible hinge loop.

The symmetrized potential is applied to a simplified C_α model where the protein is reduced to a chain of interacting C_α atoms. The energy function for multiple

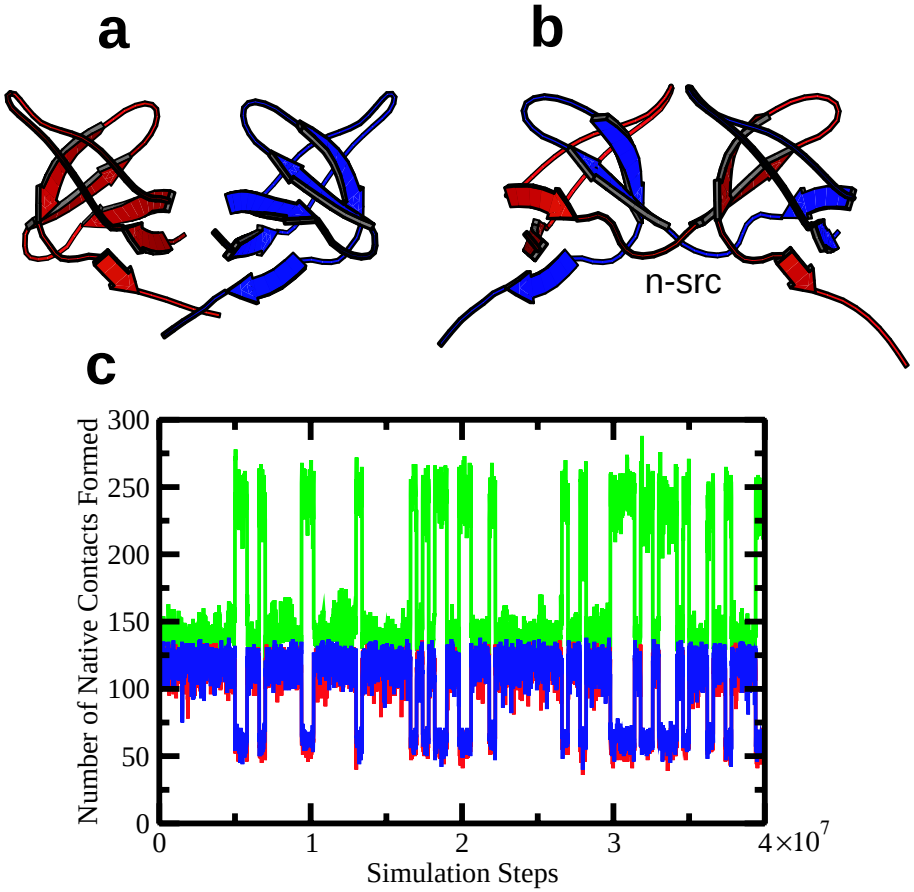

Figure 3.2 The native-like domain-swapped dimer is obtained by using the symmetrized structure-based model together with a flexible *n-src* hing loop, and the system undergoes the transitions between monomeric and domain-swapped dimeric conformations. Structures of (a) monomeric and (b) domain-swapped dimeric Eps8 SH3 domain. (c) The transition between native monomeric and domain-swapped dimeric configurations from a typical trajectory of the simulations at a temperature below the folding temperature, where red and blue represent the number of native contacts of chain A and chain B in MSH3, respectively; green represents the number of native contacts formed in the crystal DSH3.

protein chains with configuration Γ is as follows:

$$
\begin{aligned}
E(\Gamma,\Gamma_0) = \quad & E_{backbone} + \sum_{intrachain}^{i<j-3} P(i,j) + \sum_{interchain}^{i<j^{'}-3} P(i,j^{'}) + \sum_{non-native} R(i,j) \\
E_{backbone} = \quad & \sum_{\text{bonds}} K_r(r-r_0)^2 + \sum_{\text{angles}} K_\theta(\theta-\theta_0)^2 \\
& + \sum_{\text{dihedrals}} K_\phi^{(n)}(1-\cos(n\times(\phi-\phi_0))) \\
P(i,j) = \quad & \epsilon_1\left(5\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - 6\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{10}\right) \\
R(i,j) = \quad & \epsilon_2\left(\frac{\sigma}{r_{ij}}\right)^{12}
\end{aligned}
\tag{3.1}
$$

The local interaction $E_{backbone}$ applies to all protein chains. K_r, K_θ, and K_ϕ are the force constants of the bond, angle, and dihedral angle for adjacent C_α atoms, respectively, where $K_r = 100\varepsilon$, $K_\theta = 20\varepsilon$, $K_\phi^{(1)} = \varepsilon$, $K_\phi^{(3)} = 0.5\varepsilon$. r_0, θ_0 and ϕ_0 are the bond, angle, and dihedral values taken from the native monomeric configuration Γ_0 [19]. In what follows, all energies will be measured in units of $\varepsilon(= 0.5\ kcal/mol)$ and all distances in terms of $r_o(= 3.8\mathring{A})$.

To allow for domain swapping we must ensure a flexible switching capability for the *n-src* hinge region. Therefore, we set the force constants of the seven dihedrals encompassing the hinge region to zero ($K_\phi^{(1)} = K_\phi^{(3)} = 0$). When these dihedral terms in the energy function go to zero, the hinge can rotate into any configuration, specifically including both the open form of the dimer and the closed form of the monomer [132]. In our original paper on Gō-like models for dimerization, we showed that the SH3 domain native topology selected the appropriate hinge loop for swapping even when we allowed other loops in the protein to be flexible. In the interest of computer efficiency, we now limit this flexibility to the important swapping loop; by this restriction, we avoid sampling possible trap states. This is acceptable, as we are not interested here in showing that the protein topology is sufficient to control swapping (the topic of the previous paper); instead, we are interested in developing a new method for dealing with concentration and with moving toward aggregates larger than dimers.

$P(i,j)$ contains "native" intrachain and "native" interchain contact interactions of pairs of residues (i,j). Note that all the interchain contact interactions are

constructed with intrachain contacts within the monomer only. $\sigma_{i'j} = \sigma_{ij'} = \sigma_{ij}$ are equal to the native monomeric contact distance between the pair of residues (i, j) in the native monomeric configuration, and $\sigma = r_o$ for all non-native residue pairs. We choose $\epsilon_1 = \varepsilon$. A contact between two residues (i, j) is considered formed if $r_{ij} < \gamma\sigma_{ij}$ where $\gamma = 1.2$.

Excluding the "native" intrachain and interchain contact pairs, all other pairs of residues (i, j) are considered to be "non-native". These non-native pairs have a short-range repulsive interaction $R(i, j)$, with $\epsilon_2 = 0.001\varepsilon$, to model excluded volume effects.

3.4 Biased replica exchange molecular dynamics simulations

We performed replica exchange molecular dynamics simulations by periodically exchanging configurations between simulations running at a sequence of temperatures [120, 105, 32]. We followed the two-step algorithm for the replica exchange molecular dynamics (REMD) [120]: (1) A series of independent MD simulations are carried out simultaneously and independently for a certain number of MD steps with an optimal temperature distribution; (2) The pairs of configurations corresponding to neighboring temperatures are simultaneously exchanged with the probability determined by a Metropolis criterion that preserves detailed balance. We used the scheme suggested by Okamoto and coworkers [83]: exchanges are allowed only between replicas with neighboring temperatures. In our simulations, 10 replicas were run in parallel at temperatures between 0.89 and 1.07 T_F (where T_F is the folding temperature, $k_B T_F/\varepsilon = 1.32$) to ensure a reasonable overlap between the energy histograms of replicas with neighboring temperatures. One of the main advantages of using the replica exchange method in a Gō-like simplified model is that the rate of barrier crossing between the monomer and dimer structures at low temperatures is increased.

We also introduced a sampling bias by imposing a harmonic constraint $\Omega = K \times (R - R_0)^2$, where R is the center-to-center distance of two monomer chains and R_0 is the distance found in the native-like domain-swapped dimer. This bias

restricts sampling to regions in the configuration space where domain swapping can occur. The combination of this sampling bias with the REMD method were carried out in parallel with a modified AMBER package [14].

3.5 Reweighted histogram analysis method

To study the thermodynamics of domain swapping, a straightforward approach is the widely used weighted histogram analysis method (WHAM) [30, 65, 43, 111] in which one obtains the free energy by solving

$$F(R,T) = -k_B T \log \frac{\sum_k N_k(E,R) e^{-\beta E}}{\sum_k n_k e^{-f_k - \beta_k E}}, \tag{3.2}$$

where $\beta_k = 1/(k_B T_k)$ and $N_k(E,R)$ is the histogram (as a function of the inter-chain energy E and the center-to-center distance R) for the kth run with a target temperature of T_k from REMD simulations. n_k is the number of snapshots in the kth run $n_k = \sum_E N_k(E,R)$ and f_k is the free energy for the kth simulation at T_k.

As already mentioned, the harmonic constraint in our system biases the free energy landscape in favor of the interchain contacts. We can compensate for this bias by reweighting the histogram by

$$\tilde{N}_k(E,R) = \sum_{\Omega} N_k(E,E',R) e^{\beta_k \Omega}, \tag{3.3}$$

where $E' = E + \Omega$ and Ω is the constraint energy. $\tilde{N}_k(E,R)$ is the reweighted histogram at E and R, while $N_k(E,E',R)$ is the histogram at E, E', and R. One can then calculate the free energy at any temperature $T = 1/(k_B\beta)$ according to

$$F(R,T) = -k_B T \log \frac{\sum_k \tilde{N}_k(E,R) e^{-\beta E}}{\sum_k \tilde{n}_k e^{-f_k - \beta_k E}}, \tag{3.4}$$

where $\tilde{n}_k = \sum_E \tilde{N}_k(E,R)$ for the kth run. We call this procedure the reweighted histogram analysis method (rWHAM).

Figure 3.3 shows that rWHAM yields an inter-molecular force which vanishes smoothly at large inter-molecular distances, as obviously expected. In contrast, without reweighting WHAM yields an unphysical force at large separations where harmonic constraints dominate. Figure 3.3 also shows that the rWHAM results are insensitive to the value of K ($K = 0.08\varepsilon/r_o^2$ and $K = 0.10\varepsilon/r_o^2$), as required by

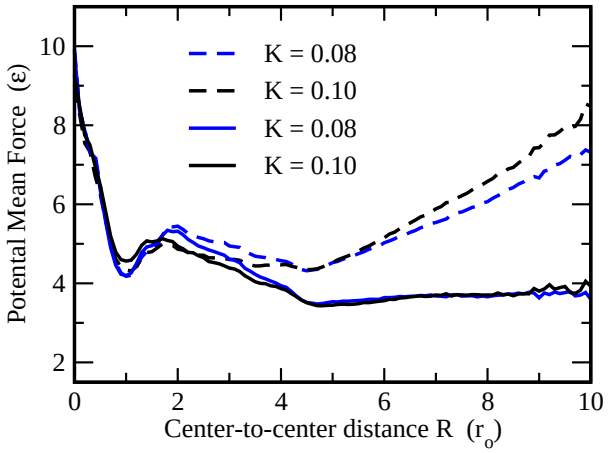

Figure 3.3 Using the reweighted histogram analysis method (rWHAM), an unbiased potential of mean force or free energy profile (V_{eff}) is reached. As a comparison, two sets of biased REMD simulations are carried out with the harmonic constraints between two protein chains, $K = 0.08\varepsilon/r_o^2$ and $K = 0.10\varepsilon/r_o^2$, respectively. The dashed lines are plotted using the original WHAM without the reweighted; the solid lines, after using the rWHAM, show a nearly flat distribution at $R > 8r_0$. The plots are shown at a temperature $T = 0.94T_F$, where T_F is the folding temperature.

consistency. The fact that the free energy profile is independent of the unphysical constraint (introduced just for sampling) is a indicator that we have obtained sufficient data for our reweighting procedure.

3.6 Results and Discussion

3.6.1 Native-like monomer and domain-swapped dimer at equilibrium

Assuming that all the native contact interactions which exist in monomers also exist in dimers (see Fig.3.1), we applied the symmetrized structure-based (Gō) model together with the flexible n-src hinge loop to our model protein, the SH3 domain from Eps8 (Fig.3.2(a) and (b)). The SH3 domain undergoes transitions between native-like monomeric and domain-swapped conformations at $T = 0.94T_F$, as shown by a typical REMD trajectory in Fig.3.2(c).

3.6.2 Large free energy barrier between monomer and domain-swapped dimer

Using the reweighted histogram analysis method (rWHAM) to remove the biased imposed by the harmonic constraints in the simulations (Eq.3.3 and 3.4), we are able to obtain the effective interchain potential of mean force, which is an orientationally averaged, but distance resolved, effective inter-molecular interaction. It depends only on the center-to-center distance R of two protein chains and is a statistical average over the remaining coordinates and the mutual orientations of the two protein chains (see Appendix). We now address the effects of inter-molecular interaction and protein concentration by studying the dimerization free-energy profile $F(R)$.

Figure 3.4 shows the potential of mean force ($V_{eff} = F(R) - F(\infty)$, see Appendix for details) as a function of the center-to-center distance R at a low temperature, where a large free energy barrier separates monomeric and domain-swapped dimeric conformations. There is no net intermolecular interaction between two protein chains when they are far away from each other; folded monomers move closer and form a free energy minimum around $R \sim 5r_o (= 19\mathring{A})$ due to long-range attractive intermolecular interactions; then monomeric proteins cross the free energy barrier and form domain-swapped dimers in the region of $R < 2r_o$.

3.6.3 Second virial coefficient analysis characterizes intermolecular interactions and concentration-dependent protein oligomerization

The second virial coefficient $\mathcal{B}_2$, as a thermodynamic parameter, can effectively measure the intermolecular interactions, the driving force of protein oligomerization. Experimentally, $\mathcal{B}_2$ can be determined by the static light-scattering method [35] and size exclusion chromatography [8]. Positive values of $\mathcal{B}_2$ indicate repulsive intermolecular interactions and negative values indicate attractive interactions [35]. From our simulations, it is straightforward to calculate $\mathcal{B}_2$ once the effective interchain interaction $V_{eff}(R)$ is known (see Appendix). Fig.3.5 shows a negative $\mathcal{B}_2$ below the folding temperature T_F, as expected for attractive interchain interac-

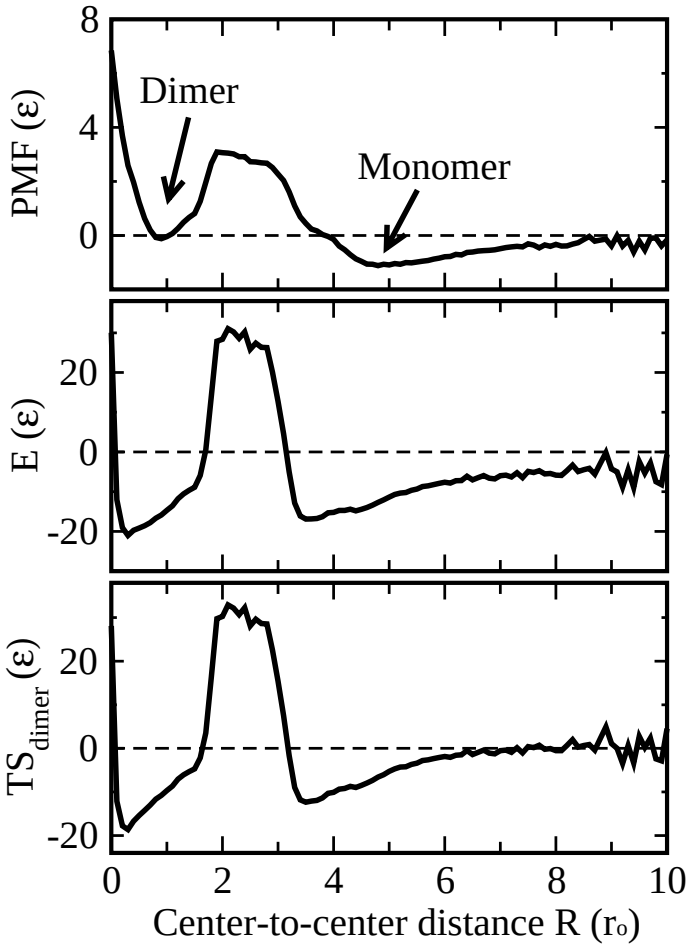

Figure 3.4 A large free energy barrier separates monomeric and domain-swapped dimeric conformations as shown from the effective interchain potential of mean force as a function of center-to-center distance R of two protein chains (Upper). The relative energy and entropy are also shown (Middle and Lower). The plots are shown at a temperature $T = 0.91T_F$, where T_F is the folding temperature.

tions where domain swapping occurs, and a positive $\mathcal{B}_2$ above T_F where proteins are usually unfolded.

To quantify the effect of interchain interactions, we define a critical quantity c^*,

$$c^* = -\frac{1}{\mathcal{B}_2}, \tag{3.5}$$

above which the free energy correction of two-body interactions $F_2 = Nk_BTcB_2$ are more important than with the ideal gas contribution $F_1 = -Nk_BT$ (see Appendix for details). As shown in the temperature-concentration plane in Fig.3.5, at low protein concentrations $c \ll c^*$, F_1 is dominant and F_2 can be ignored (an ideal gas behavior), and the protein system is dilute enough that weak interchain interactions favor monomeric configurations. At high concentrations $c > c^*$, the contribution from F_2 is large where attractive interchain interactions become strong, which favors oligomeric configurations.

One can use B_2 to estimate the total concentration, $c^* = 2[D] + [M]$, at which one forms reasonable amounts of dimeric conformations ($[M]$ and $[D]$ are the concentrations of monomers and dimers, respectively). In addition, one can further calculate the equilibrium reaction constant, $K = [D]/[M]^2 \sim 1/[M]$, assuming $[D] \sim [M]$ at the concentration c^*.

The second virial coefficient analysis has provided us with a clear picture of how protein concentration affects protein oligomerization. We can also show, by directly simulating the effect of confinement, how an effective high protein concentration promotes domain-swapped dimeric conformations. With a stronger constraint, $K = 0.4\varepsilon/r_o^2$ (as compared to the value $K = 0.1\varepsilon/r_o^2$ used for sampling purposes in previous sections), two protein chains are confined and the sampling is strongly restricted to achieve a fixed center-to-center distance. Figure 3.6 shows that dimeric conformations are favored at small center-to-center distances (i.e., effective high protein concentrations). These results are consistent with those obtained from the second virial coefficient analysis in Fig. 3.5.

3.6.4 Connection with aggregation and amyloid formation

A concentration-dependent nucleation-polymerization mechanism has been proposed for aggregation and amyloid formation of a variety of fibril-related pro-

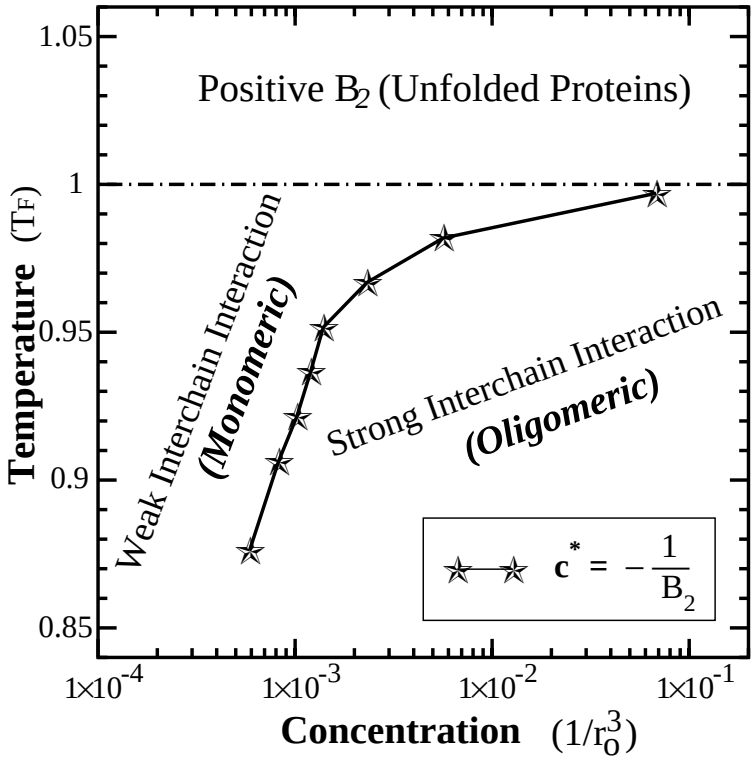

Figure 3.5 Concentration-dependent protein oligomerization by using the second virial coefficient analysis. Above the folding temperature T_F, $\mathcal{B}_2$ is positive where proteins are unfolded. Below T_F, monomeric conformations are favored if protein concentration $c \ll c^*$, and dimeric conformations are favored if $c > c^*$ due to strong attractive interchain interactions. Temperature and concentration are in the units of T_F and $1/r_o^3$, respectively.

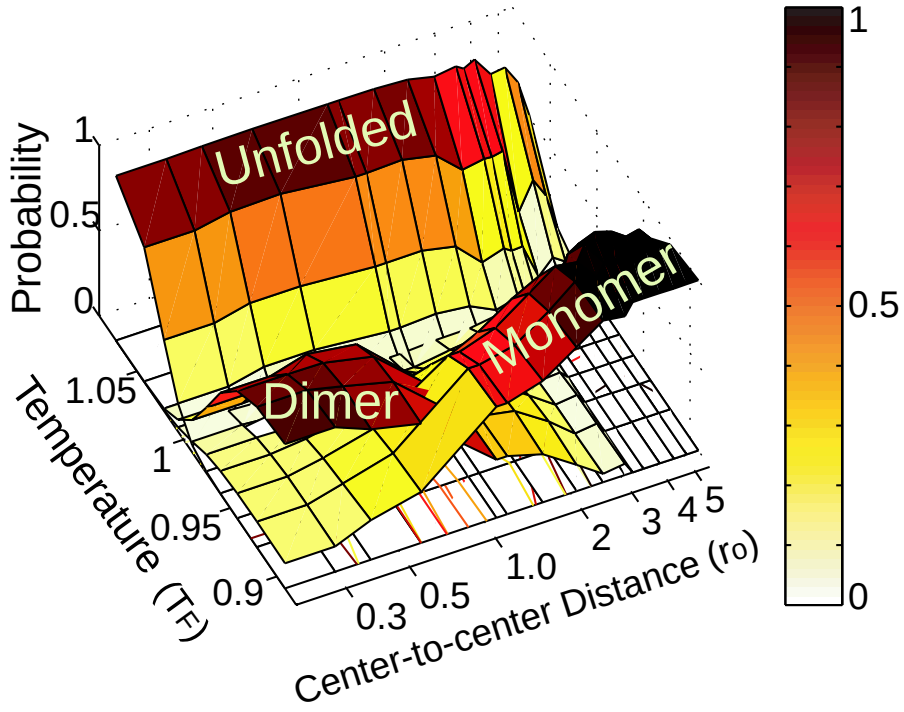

Figure 3.6 Direct simulations for two protein chains show how the population of monomeric and dimeric configurations change with thetemperature and center-to-center distance R. The result is consistent with that from the second virial coefficient analysisin Fig. 3.5.

teins [141, 52, 46, 74, 75]. A similar mechanism is also observed with other proteins [18, 47]. Our approach allows us to predict the critical concentration for oligomerization, presumably the first step in the formation of protofilaments and mature fibrils. For example, for the SH3 protein, the critical concentration is around $25\mu M$ (in unit of $10^{-3}/r_o^3$) at the temperature $T = 0.92T_F$. Moreover, the trend of the curvature of $c^* = -1/\mathcal{B}_2$ shown in Fig.3.5 can be tested and verified by experiment. Comparison with experiments should be done with care since our current model clearly over-estimates the concentration needed for oligomerization. The errors arises because we have included only short-range potentials in our model; corrections due to long-range interactions such as electrostatic may be quantitatively important. Also, interactions a few additional interfacial contacts are formed upon dimerization, as is evident from the dimer crystal structure; these will shift the critical concentration downward.

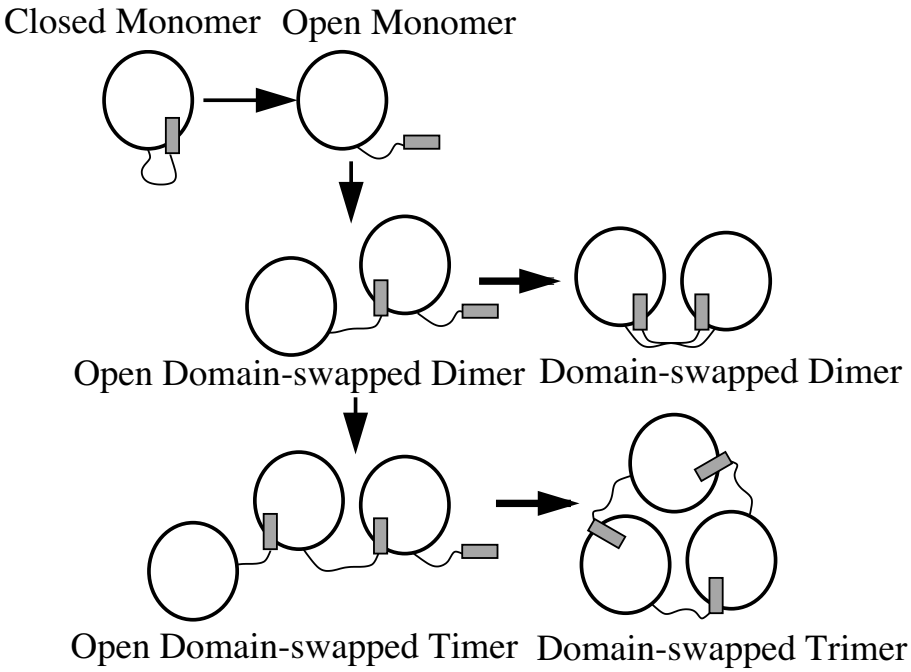

Figure 3.7 A schematic illustration for molecular self-assembly by domain swapping. The open-end domain-swapped conformation can serve as templates and then domain-swap and form dimers, trimers, and even higher order oligomers.

3.6.5 Molecular self-assembly: Go beyond dimerization

In our previous work [132] we focused on the following mechanism of domain swapping: native monomers in "closed" form $\rightleftharpoons$ partially folded monomers in "open" form $\rightleftharpoons$ unfolded monomers $\rightleftharpoons$ open-end domain-swapped dimers $\rightleftharpoons$ native-like domain-swapped dimers. The transition from "closed" to "open" monomers, and the observation of open-end domain-swapped dimers suggest that monomers could self-assemble into higher-order domain-swapped oligomers and serve as templates to form aggregates. This is illustrated in Fig. 3.7. Moreover, experiments have also shown the existence of several domain-swapped trimers [139, 73]. We test this hypothesis by extending the protocol used for two protein chains to the case of three chains of the Eps8 SH3 domain. Each pair of monomer chains has harmonic constraints ($\Omega = K \times (R - R_0)^2$ with $K = 0.10\varepsilon/r_o^2$ and $R = 0.1r_0$). We do indeed find that stable domain-swapped trimers are populated on the free energy surface as shown in Fig.3.8.

Many domain-swapping proteins, such as human cystatin [51], human prion [137,

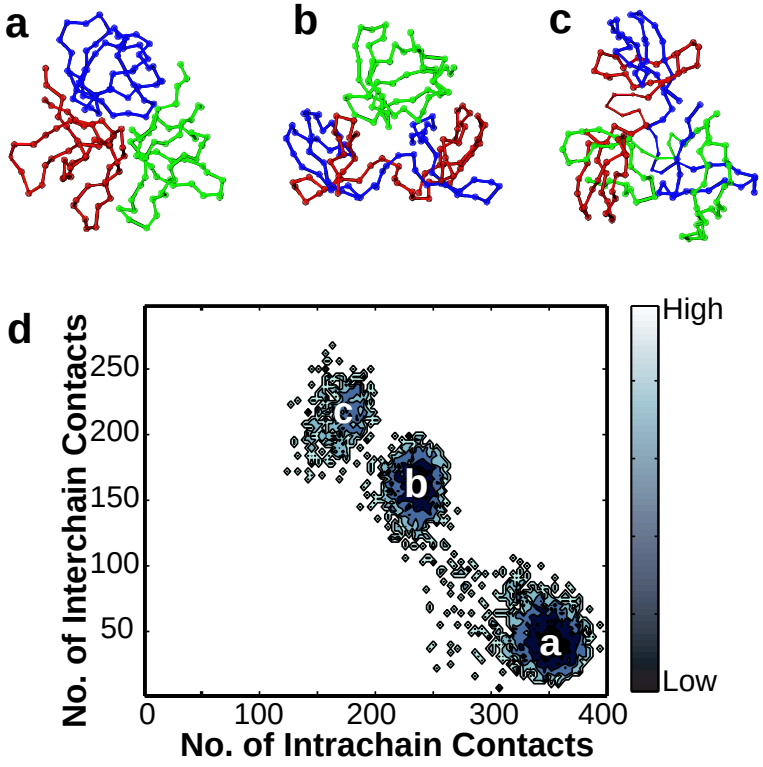

Figure 3.8 Typical configurations of monomer (a), domain-swapped dimer (b), and domain-swapped trimer (c) are populated on the free energy surface (d) at a temperature below the folding temperature.

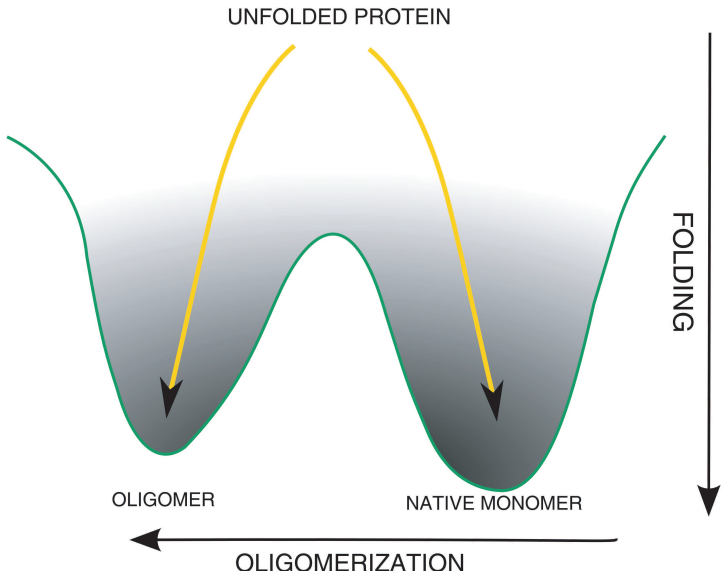

Figure 3.9 Oligomerization funnel: Unfolded proteins can fold into native monomeric or oligomeric conformations upon the competition of intra-molecular and inter-molecular interaction. On the other hand, monomer can cross a high free energy barrier to form oligomer under certain conditions, such as high protein concentration.

61], and p13suc1 [103], have been reported to form amyloid fibrils. Our simulations support the role of domain swapping in the early oligomeric aggregation of amyloid fibrils. It is very likely that domain swapping is a general mechanism of protein aggregation [51, 108, 20, 72, 112]. Domain swapping helps to understand how single domain proteins dimerize or trimerize and how these monomers can self-assemble into protofibrils through either propagation or nucleation.

3.7 Conclusions

We have used the symmetrized structure-based model and a flexible hinge region to study the mechanism of domain-swapped oligomerization. As shown in Fig.3.9, unfolded proteins can directly fold into native monomeric or oligomeric conformations, corresponding to a competition between intra-molecular and inter-molecular interactions. Below T_F, on the other hand, monomers might cross a high free energy barrier to form oligomers under certain conditions, such as high protein concentrations. The methods developed in this paper are generally appli-

cable as long as the monomer structure is known and could be used to investigate concentration effects for many proteins.

3.8 Appendix: Derivation of the second virial coefficient

For a non-ideal gas of monomeric proteins with a number of internal degrees of freedom, i.e., internal coordinates $\{\mathbf{r}_\alpha^{(i)}\}$ of monomer i and center-of-mass coordinates $\mathbf{R}_i = \sum_\alpha \mathbf{r}_\alpha^{(i)}$, the partition function of the entire protein system can be written as

$$Z = \prod_i \int d\mathbf{R}_i \delta(\mathbf{R}_i - \sum_\alpha \mathbf{r}_\alpha^{(i)}) \prod_\alpha d\mathbf{r}_\alpha^{(i)} e^{-\beta\mathcal{H}}, \tag{3.6}$$

where the full Hamiltonian $\mathcal{H}$ can break into two terms: a single-body term $\mathcal{H}_1$ and a two-body term $\mathcal{H}_2$,

$$\mathcal{H} \approx \sum_i \mathcal{H}_1(\{\mathbf{r}_\alpha^{(i)}\}) + \sum_{i\neq j} \mathcal{H}_2(\{\mathbf{r}_\alpha^{(i)}\}, \{\mathbf{r}_\alpha^{(j)}\}, \mathbf{R}_{ij}), \tag{3.7}$$

where $\mathbf{R}_{ij} = \mathbf{R}_i - \mathbf{R}_j$. If we define a single-particle partition function

$$Z_1 = \int \prod_\alpha d\mathbf{r}_\alpha^{(i)} \delta(\mathbf{R}_i - \sum_\alpha \mathbf{r}_\alpha^{(i)})\ e^{-\beta\mathcal{H}_1(\{\mathbf{r}_\alpha^{(i)}\})} \tag{3.8}$$

and a two-particle partition function

$$\begin{aligned} Z_2 \;=\; & \int d\mathbf{R}_{ij}\delta(\mathbf{R}_{ij} - (\mathbf{R}_i - \mathbf{R}_j)) \prod_\alpha d\mathbf{r}_\alpha^{(i)} d\mathbf{r}_\alpha^{(j)} \delta(\mathbf{R}_i - \sum_\alpha \mathbf{r}_\alpha^{(i)}) \delta(\mathbf{R}_j - \sum_\alpha \mathbf{r}_\alpha^{(j)}) \\ & \times e^{-\beta\mathcal{H}_1(\{\mathbf{r}_\alpha^{(i)}\})} e^{-\beta\mathcal{H}_1(\{\mathbf{r}_\alpha^{(j)}\})} (e^{-\beta\mathcal{H}_2(\{\mathbf{r}_\alpha^{(i)}\},\{\mathbf{r}_\alpha^{(j)}\},\mathbf{R}_{ij})} - 1) \end{aligned} \tag{3.9}$$

we then have that (with a second order approximation)

$$Z \approx (VZ_1)^N (1 + \frac{N(N-1)}{2V} \cdot \frac{Z_2}{Z_1^2}), \tag{3.10}$$

where V is the integral throughout the space ($V = \int d\mathbf{R}_i$). Consequently, the free energy is given by

$$\mathcal{F} \;=\; \mathcal{F}_1 + \mathcal{F}_2 \tag{3.11}$$

$$\approx\; -Nk_BT \ln(VZ_1) - \frac{1}{2} Nk_BT \cdot \frac{N-1}{V} \cdot \frac{Z_2}{Z_1^2}, \tag{3.12}$$

where $\mathcal{F}_1 = -Nk_BT\ln(VZ_1)$ is contributed by the ideal monomeric gas. The second term is from the pairwise interaction correction,

$$\begin{aligned}\mathcal{F}_2 &= -\frac{1}{2}Nk_BTc\frac{Z_2}{Z_1^2} \\ &= 2\pi Nk_BTc\int R^2dR(1-e^{-\beta V_{eff}(R,T)}),\end{aligned} \tag{3.13}$$

where $c = (N-1)/V \approx N/V$ is the protein monomer concentration. $V_{eff}(R,T)$ is an orientationally averaged, but distance resolved, effective pairwise interaction obtained by integrating out the remaining coordinates of internal degrees of freedom $\{\mathbf{r}_\alpha^{(i)}\}$ and $\{\mathbf{r}_\alpha^{(j)}\}$,

$$\beta V_{eff}(R,T) = -\ln\langle e^{-\beta\mathcal{H}_2(\{\mathbf{r}_\alpha^{(i)}\},\{\mathbf{r}_\alpha^{(j)}\},\mathbf{R}_{ij})}\delta(R-|\mathbf{R}_i-\mathbf{R}_j|)\rangle_{\{\mathbf{r}_\alpha^{(i)}\},\{\mathbf{r}_\alpha^{(j)}\}}, \tag{3.14}$$

where $R = |\mathbf{R}_{ij}|$ is the center-to-center distance between monomers (it does not matter which two, as they are all identical). Numerically, we first compute the free-energy profile $F(R)$ (where R is the center-to-center distance) by using the reweighted histogram method (Eq. 3.4) over all the snapshots from REMD simulations. This then determines the potential of mean force according to $V_{eff} = F(R) - F(\infty)$; the subtraction takes care of the arbitrary constant which appears in the WHAM free energy computation and reflects the fact that V_{eff} will go to zero as $R \to \infty$, as is clear from its definition in Eq.(3.14). Finally we can rewrite Eq. (3.13) as

$$\mathcal{F}_2 = Nk_BTcB_2, \tag{3.15}$$

where B_2 is the so-called second virial coefficient

$$B_2 = 2\pi\int R^2dR(1-e^{-\beta V_{eff}(R,T)}). \tag{3.16}$$

Acknowledgments

S.Y. thanks Shachi Gosavi and Leslie Chavez for critical reading of the manuscript; we also acknowledge Peter Wolynes, Jin Wang, Daniel Cox, Margaret Cheung, Koby Levy, Sam Cho, and Peter McHale for helpful comments and discussions. Computing resources were supported in part by W. M. Keck Foundation. This work was supported by the NSF-sponsored Center for Theoretical Biological Physics.

The chapter has been published in *J. Mol. Biol.*, Sichun Yang, Herbert Levine, and Jose Onuchic, 352:202–211, 2005 [133].

4

Biological Implications: Structure of Infectious Prions Stabilized by Domain Swapping

A candidate structure for the minimal prion infectious unit is a recently discovered protein oligomer modeled as a β-helical prion trimer (BPT); BPTs can stack to form cross-β fibrils and may provide insight into protein aggregates of other amyloid diseases. However, the BPT lacks a clear intermonomer binding mechanism. Here we propose an alternative domain-swapped trimeric prion (DSTP) model and show with molecular dynamics (MD) that the DSTP has more favorable intermonomer hydrogen bonding and proline dihedral strain energy than the BPT. This new structural proposal may be tested by lysine and N terminus fluorescent resonance energy transfer (FRET) either directly on recombinant prion protein amyloid aggregates or on synthetic constructs that contain the proline/lysine-rich hinge region critical for domains to swap. In addition, the domain swapping may provide (1) intrinsic entanglement, which can contribute to the remarkable temperature stability of the infectious prion structure and help explain the absence of PrP^{Sc} monomers, and (2) insight into why specific prolines are potentially relevant to three inherited forms of prion disease,.

4.1 Introduction

THE PROTEIN-ONLY MODEL for prion diseases remains the leading candidate to explain the infectious form of the disease [99], and this has been significantly strengthened by the successful production of disease from synthetic prion protein aggregates [68, 15]. Clearly, the understanding of prion disease pathology, propagation within the body, and therapeutic approaches will be greatly enhanced by a detailed knowledge of the structure of the minimal infectious protein unit.

The discovery of areal aggregates of infectious prion protein oligomers in purified brain extracts from mice has inspired a leading candidate model for the minimal infectious unit, a trimer in which large portions of the N terminus are converted to left-handed β-helices [129, 38]. We call this the β-helical prion trimer (BPT) model. Figure 4.1 (A) shows a top view of the BPT in the β-helical region. This BPT displays a 3-fold symmetry consistent with maps of the electron density difference between the PrP 27-30 oligomers and the "miniprion" $PrP^{Sc}106$ oligomers. The BPT can easily be stacked to form filaments, which can then be wound into amyloid fibrils. The model trimer also shares detailed structural features with the bacterial protein carbonic anhydrase trimer [70]. Inspired in part by this model, others have explored the possibility that the β-helix plays a role in other amyloid structures such as aggregates of polyglutamine peptides [119], β-amyloid peptides [42], and yeast prion-like proteins [59].

However, it is not clear in the BPT model what holds the trimer together. The carbonic anhydrase trimer, which was the template for the BPT model is in fact held together by intermonomeric covalent bonding of zinc ions [70], which are not present in PrP^{Sc}. It was suggested that a hydrogen bonding network is responsible for this binding [38], as in the case of the bacterial β-helical trimer serine acyltransferase (SAT) [101], but no effort was made to test this idea quantitatively. Moreover, the yellow loops of the BPT model (which are highly conserved [76]) shown in Fig. 4.1A and B and expanded in Fig. 4.1C contain two prolines, and prolines tend to disfavor this type of bending.

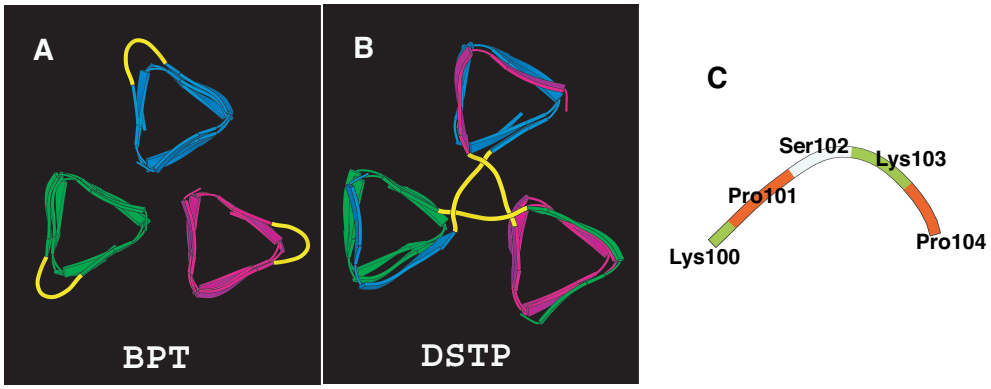

Figure 4.1 (A) Proposed β-helical structure of the truncated $PrP^{Sc}106$ prion trimer (BPT) taken from Ref. [38] (residues G89-F140). (B) Structural model of domain-swapped trimeric prion (DSTP) of the present paper. (C) The proposed hinge region expanded from the yellow loops in panels A and B.

4.2 Hypothesis: A domain-swapped prion trimer is the minimal infectious unit

We hypothesize that a stable oligomeric structure for the infectious prion is the domain-swapped trimeric prion (DSTP) of Fig. 4.1B, which can increase hydrogen bonding and reduce elastic energy relative to the BPT structure proposed earlier. Further, we conjecture that the DSTP structure may be relevant to (1) the unusual temperature stability of the prion protein and the lack of observation of PrP^{Sc} monomers (both due to the entanglement of the DSTP structure), (2) the kinetics of at least three inherited forms of human prion disease, and (3) the encoding of prion strains in conformation.

4.3 Domain swapping: Motivation and Results

Concern with stability of the BPT model led us to modify it by domain swapping [3, 4, 132], especially since prolines are present in the hinge or bridging regions of several proteins that domain swap [6]. For example, in p13suc1, the strain on two prolines in the hinge modulates the equilibrium between monomers and domain-swapped dimers [103, 102].

We constructed our DSTP model of Fig. 4.1B from the proposed BPT struc-

ture (we thank C. Govaerts for providing his coordinate file) by reorienting the monomers to place the loop regions near the trimer center and swapping residues 89-99 to a neighboring β-helix in a cyclical fashion, with the BPT loops serving as hinges. This DSTP model straightens these loops, adds more possible hydrogen bonds in the loop region, and preserves the putative monomer-monomer spacing and 3-fold symmetry of the electron density maps [129]. We ran all-atom, explicit solvent MD (see Appendix for details) on energy minimized DSTP and BPT structures using AMBER8 (http://amber.scripps.edu). As a control, we applied the same protocol to the serine acyltransferase (SAT) β-helical trimer, which has no strained prolines and for which there is direct crystallographic evidence for intermonomer hydrogen bond-mediated cohesion [101]. Wee also carried out two other simulation tests: (1) we examined the stability of the DSTP model with the inclusion of the C terminus α-helices proposed in Ref. [38], and (2) we carried out three distinct 1-ns runs for both the BPT and DSTP structures to make sure there is no special circumstance with regard to individual simulations.

We have concluded that the DSTP is significantly more stable than the BPT and therefore a more promising candidate structure. First, we have enumerated direct and water-bridged hydrogen bonds between monomers in the BPT model, our DSTP model, and the SAT structure (see Appendix), and carried this out for three separate MD runs on each structure. We show the direct hydrogen bond counts for three different 1-ns runs of the BPT and DSTP structures in Fig. 4.2. For the BPT, we find 2 ± 1, 5 ± 1, and 2 ± 1 direct hydrogen bonds for the three runs (averages over the last 100-ps of simulation time), and at the last simulation time 0, 0, and 1 water-mediated hydrogen bonds. For the DSTP, we find for all three runsi 9 ± 1 direct hydrogen bonds for the last 100-ps of simulation time and 2, 1, and 1 water-mediated hydrogen bonds in the last simulation snapshot. In the one BPT case with a relatively high number of direct hydrogen bonds, the N terminus from one monomer has contacted with the N terminus region of another. In general, the BPT hydrogen bonds form around Q89 only, while the DSTP hydrogen bonds form in the hinge region. Note that Fig. 4.2 shows direct hydrogen bonds formed in the monomer-monomer interface for the DSTP to generically increase with time while those from the BPT generically decrease with MD time. In contrast to the BPT, the simulated β-helical portion (F140-A237) of the SAT trimer has 11 direct and

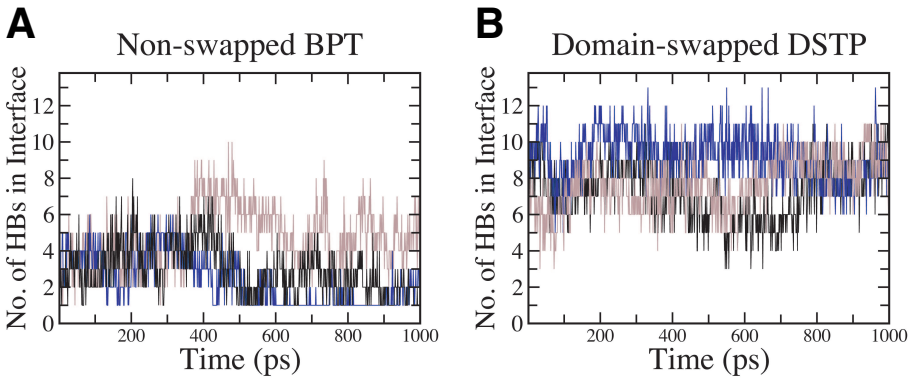

Figure 4.2 MD simulations show that the nonswapped BPT (A) has less direct hydrogen bonds in the interface region than the domain-swapped DSTP (B). In both cases we have simulated residues G89-F140. Different colors correspond to three different 1 ns MD runs in each case initialized at different initial temperatures (see Appendix).

3 water-bridged hydrogen bonds (after two stage equilibration with no restraints).

Second, we remark that the domain swapping also relaxes the P101 and P104 dihedral energy by 1.0 kcal/mol after 1-ns of MD. Hence, the domain swapping helps relieve elastic strain as anticipated.

Third, we have verified that the DSTP structure for the $PrP^{Sc}106$ can accommodate the C terminus α-helices without affecting stability. This is a nontrivial point, since a rigid rotation of the monomers within the BPT is not possible due to steric constraints. As shown in Fig. 4.3, we find that by "peeling off" one triangular edge (residues 133-140) of the lower (C terminus) end of each β-helix to form a random coil segment that we can link to the α-helices that begin at residue 177, with no resultant change in direct hydrogen bonding structure.

Finally, high temperature MD provides direct evidence for BPT instability; at 500K, the DSTP remains compact and entangled, while the BPT monomer-monomer separation nearly doubles. Principal component analysis, which identifies the amplitude and character of the dominant fluctuations, shows similar BPT unbinding tendencies at 300K as in the 500K simulation, but no unbinding tendencies for the DSTP. Figure 4.4 shows the radius of gyration in the 500K runs as a function of simulation time. Within the 1-ns window of simulation, the BPT radius appears to keep growing, reaching about twice that of the apparently more stable

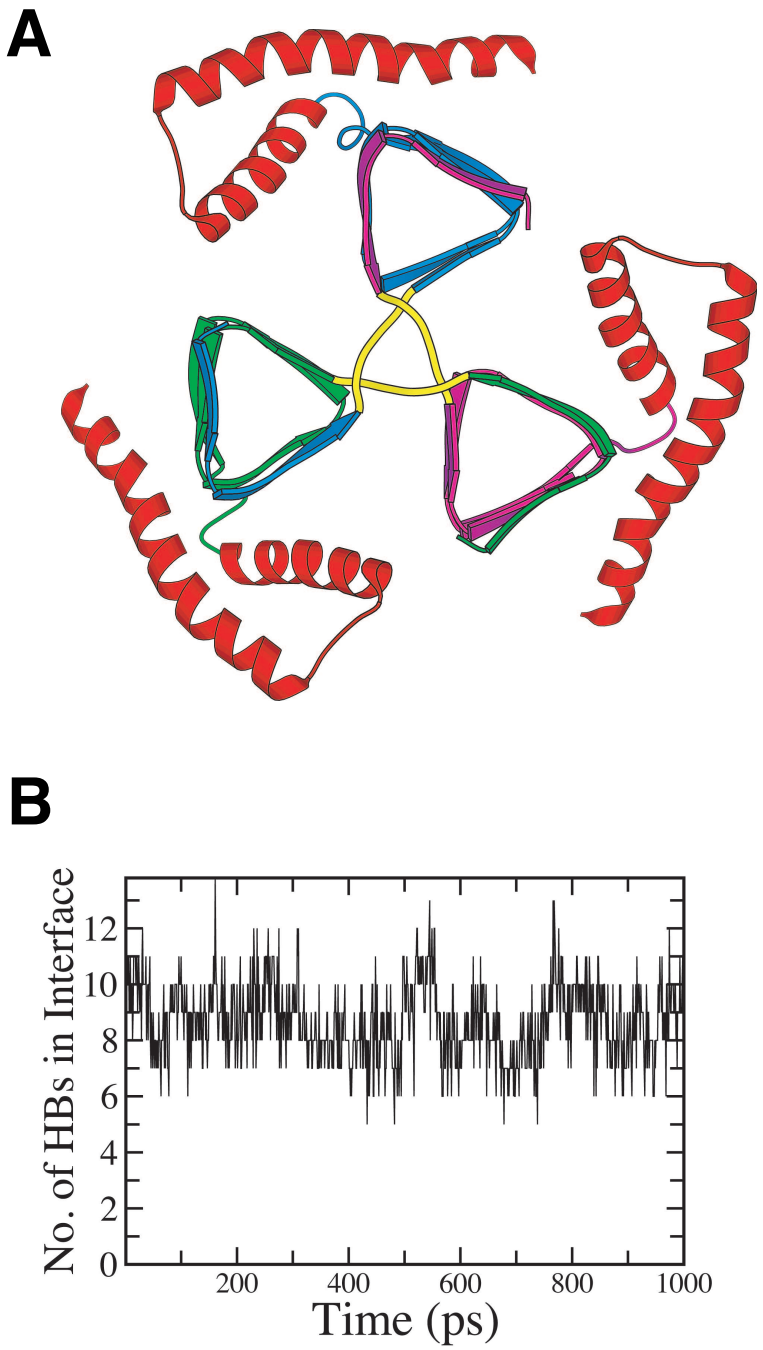

Figure 4.3 (A) Model for DSTP $PrP^{Sc}106$ (with residues G141-H176 deleted) and α-helices included. Residues 133-140 have been removed from the lower rung of the β-helices for each monomer, converted to random coil, and attached to residue 177 of the α-helices. (B) Number of direct hydrogen bonds for the DSTP structure with intact α-helices. The α-helices have no effect on the stabilization induced by the domain swapping.

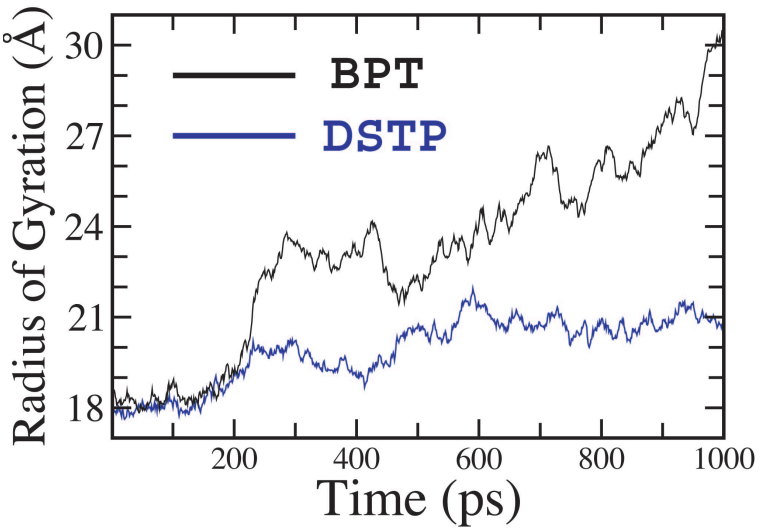

Figure 4.4 The entanglement due to domain swapping introduces additional stability for the DSTP relative to the BPT as revealed by high temperature ($T = 500K$) simulations, where the radius of gyration of the BPT increases more rapidly than that of the DSTP, which reflects unbinding of the BPT. Principal component analysis at 300K reveals a similar unbinding tendency.

DSTP structure. This stabilization through entanglement may be relevant to the unusual temperature stability of infectious prion proteins, although we cannot rule out that such stabilization is related to larger scale aggregates of prion oligomers. The entanglement may also be relevant to the inability to isolate monomeric forms of PrP^{Sc}, since the scrapies form is stabilized in this picture only in aggregate.

4.4 Possible relevance to inherited diseases

Finally, we discuss the possible relevance of the DSTP model to inherited diseases and strains. First, proline point mutations yield heritable Creutzfeldt-Jakob disease (P105T for humans, P104T for mice) or Gerstmann-Straussler-Scheinker disease (P102L, P105L for humans, P101L, P104L for mice; see chapter 14, Ref. [100]). Transgenic mice with P101I have even greater disease susceptibility whereas those with P101Y have somewhat less susceptibility than P101L (K. Nazor and G. Telling, private communication). We note that the centrality of these sites in the DSTP hinges affords possible burial of the relatively hydrophobic I, L, T, and Y side chains and a potential speed up of trimer formation through in-

creased hinge flexibility. Moreover, the insertion of relatively hydrophobic residues into the unstructured region of the normal PrP^C protein will raise the energy of this structure. Indeed, it is already known that the P102L mutation is marginally less stable than the wild-type protein (see p. 694, Ref. [100]). A decrease in the stability of the cellular protein form relative to wild-type monomers coupled to an increase in stability of the scrapies form (associated with more effective hydrophobic burial of the mutant residue) should lead to more rapid conversion kinetics via ramping of the overall free energy surface slope in favor of the scrapies form.

We note that a different domain swapping model for strains has been conjectured previously [21]; separately, domain-swapped prion dimers have been crystallized [61]. The latter result is consistent with prion oligomerization via intermolecular disulfide bonding as has been shown by redox induced prion fibrillization in vitro [67], although only intramolecular disulfide bonding has so far been observed *in vivo* [127, 80]. Other specific domain swapping mechanisms than those we have discussed here may play a critical role in amyloid fibril formation for prions as has been observed for other proteins [72, 51, 112]. For the PrP 27-30 scrapies form, the large loop (residues E145-R163) of Ref. [38] made from some of the deleted residues of the $PrP^{Sc}106$ structure could serve as a hinge region for alternative domain-swapped structures.

4.5 Proposed experimental tests

While the DSTP structure may ultimately be testable directly on recombinant infectious prion oligomers, as have been synthesized by Prusiner and collaborators recently [68], we first propose testing an analog system by FRET labeling. Poly-Q peptides are of considerable interest because of the polyglutamine diseases (especially Huntingtons). The "critical" length near 36 glutamine repeats for disease is suggestive of a left-handed β-helix structure, as discussed elsewhere [38, 119, 128]. The key experiment we propose is to construct synthetic peptides of the form $Q_{12}(KPSKPK)Q_m$, where $m = 24$ 36. The goal is to label the N terminus Q's in half the peptides with a donor, the K's with a different donor, and in the other half put in N terminus Q acceptors and K acceptors. If the argument about stress relaxation and hydrogen bond stabilization works for the prion, it should also work

here. There should be intense K-K FRET given intermonomer separations of the side chains by 4 Åif the domain swapping arises. There will also be weak N terminus-Q- N terminus-Q FRET. Presumably there could be an N terminus-Q- to- N terminus-Q FRET signal arising with time if the trimers are first formed and then they fibrillize, since the corresponding distances would get reduced by 60% case assuming stacking of the β-helices. If this works on the synthetic peptide it should be attempted on recombinant prion protein prior to in vitro aggregation as per Ref. [68]. We would anticipate little interfering contributions from other lysines in the prion protein sequence that are far from the hinge region.

4.6 Summary

We propose a domain-swapped trimer model for the minimal infectious prion oligomer, and note that the domain swapping stabilizes the oligomer relative to the earlier proposed model by reducing stress in proline containing loops, increasing intermonomer hydrogen bonding, and promoting entanglement. We conjecture that domain swapping may play a role in explaining the temperature stability of the scrapies form of the prion protein, the absence of PrP^Sc monomers, inherited prion disease phenotypes associated with mutating prolines in hinges of the DSTP. Finally, we propose experimental tests based on FRET labeling of principally lysine residues near the crucial pralines, which should produce intense FRET response if the domain swapping is present.

4.7 Appendix

The simulations were carried out with the AMBER8 molecular dynamics package [13] with the AMBER parm99 force field and explicit TIP3P waters, following these procedures. (1) Before we start a molecular dynamics simulation, we perform an energy minimization of 2,000 steps to partially relax the entire molecular system. (2) We perform a two-stage equilibration to further relax the protein and the surrounding solvent. In the first stage, we start the system from a low temperature of approximately 100K and gradually heat up to 300K over 20-ps of simulation time. We perform this stage of equilibration with the volume held constant. In

the second stage, we equilibrate the system using pressure and temperature control to adjust the density of water to experimental values. We have equilibrated the system for a total of 40-ps. During the first two steps, the backbone heavy atoms of the β-helical (and also the α-helical if applicable) portions are positionally restrained using a harmonic potential. The hinge loops and the regions linking the β-helical and α-helical domains are allowed to freely move. To achieve the three different MD runs for the BPT and DSTP structures, we start with different initial temperatures that seed a different run through a different initial thermal distribution of atom velocities. Finally, we carry out a 1-ns production run at constant pressure and temperature with no positional restraints. For the high temperature ($T = 500K$) simulations, the production runs are performed at constant volume with a reduced time step of integration to prevent any possible system blowup.

Hydrogen bonds are calculated based on the following criteria: (1) direct hydrogen bonds in interface are counted with a donor-acceptor distance of 3.5 Åand a hydrogen-donor-acceptor angle of 60^o, and (2) water-bridged hydrogen bonding in interface is counted if a water molecule couples by criterion (1) to an atom from each of two monomers.

Note added: we also observed that exponential growth requires membrane binding, where we find as well that the membrane is necessary for exponential growth of prion aggregates; without it, the kinetics is simply the quadratic-in-time growth characteristic of linear elongation as observed frequently in in vitro amyloid growth experiments with other proteins [22].

Acknowledgments

We thanks useful conversations with R. R. P. Singh and K. W. Plaxco; we also acknowledge fruitful discussions with P.G. Wolynes. This work was supported by the U.S. Army grant NP020132 and NSF Grants PHY0216576 and PHY0225630. Computing resources were supported by the San Diego Supercomputer Center and the NSF-sponsored Center for Theoretical Biological Physics.

This chapter has been published in *FASEB J.*, Sichun Yang, Herbert Levine, Jose Onuchic, and Daniel Cox, 19(13):1778–1782, 2005 [135].

5

Effective Stochastic Dynamics on a Protein Folding Energy Landscape

We present an approach to protein folding kinetics using stochastic reaction-coordinate dynamics, in which the effective drift velocities and diffusion coefficients are determined from microscopic simulation data. The resultant Langevin equation can then be used to directly simulate the folding process. Here, we test this approach by applying it to a toy two-state dynamical system and to a funnel-like structure-based (Gō-type) model. The folding time predictions agree very well with full simulation results. Therefore, we have in hand a fast numerical tool for calculating the folding kinetic properties, even when full simulations are not feasible. In addition, the local drift and diffusion coefficients provide an alternative way to compute the free energy profile in cases where only local sampling can be achieved.

5.1 Introduction

Protein folding is a complex dynamical process occurring in a high-dimensional conformational space [91]. It is usually not feasible to try to tackle this problem by using fully microscopic (all-atom) models, as the folding time-scales are often many orders of magnitude larger than the achievable simulation times. Many re-

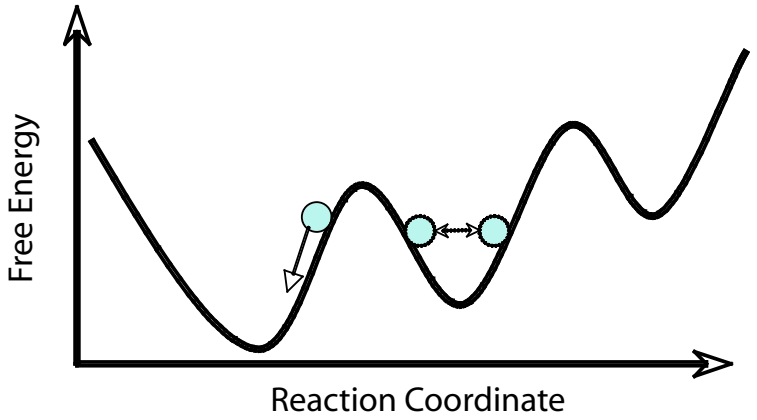

Figure 5.1 Schematic diagram of the diffusion-drift motion on a free energy landscape which is a function of an arbitrary reaction coordinate. The protein gets trapped in a local minimum, overcomes the barrier by diffusive motion, and finally folds downhill into the native state. The basic idea of the present paper is that we can determine, directly from a full simulation of the multi-dimensional microscopic dynamics, the drift and diffusion coefficients at each position along the reaction coordinate. We then simulate the Langevin dynamics with these drifts and diffusion coefficients, thereby characterizing the kinetic process.

searchers have approached this problem by focusing on the free energy "landscape" as a function of one or two reaction coordinates. For example, the by-now familiar funnel concept refers to the use of Q, the fraction of native contacts, as a good coordinate for labeling configurational ensembles encountered during typical folding events. This view has achieved great popularity, as it enables the creation of structure-based coarse-grained models which then allow for the determination of measurable features of the folding kinetics, such as residue sensitivities (ϕ-values).

It is clear, nonetheless, that one cannot directly determine true kinetic rates and folding time distributions from the landscape alone. We must also have dynamics on this landscape, which in practice means knowing the drift velocity and the diffusion coefficient [114] for the reaction coordinate motion as functions of position in the landscape (see Fig. 5.1). This idea was already recognized by the originators of the landscape concept; for example in the analytic work of Bryngelson and Wolynes [12], the folding time is determined both by the free energy barrier and a prefactor that depends on the ruggedness of the landscape, the latter entering via a constant diffusion coefficient. More recently, a diffusion approach has been used by several research groups [97, 85, 48, 126, 5].

Our goal here is to show that one can use short time "local" simulations of the full dynamics to derive values for the needed drift and diffusion coefficients. Although discovered independently, it turns out that our method is a specific realization of what Kevrekidis and co-workers have proposed under the name of "equation-free computational methods" [63]. The goal of this paper is to realize and validate this procedure for the specific case of protein folding. An application to a real folding example (the folding of the G-protein hairpin segment) will be presented elsewhere.

The outline of this work is as follows. First, we discuss how to extract the position-dependent drifts and diffusion coefficients along the energy landscape from simulation data. This allows us to simulate the folding dynamics by using a Langevin approach with the determined drifts and diffusion coefficients. To test the validity of this procedure, we first study a toy two-state system. We then apply the formulation to a funnel-like landscape, created by using a minimalist Gō-type model to characterize the kinetic process. The Gō-type simulation has been successful not only for protein folding [89, 19, 62, 17, 55] but also for protein oligomerization [132, 133]. Using the single reaction coordinate Q, we determine the drifts and diffusion coefficients at each Q from the full Gō simulation data. Then we carry out the one-dimensional Langevin simulations. The folding times computed in this manner agree very well with full simulation results.

As a byproduct of our calculations, we can construct a free energy profile directly from kinetic quantities such as drift and diffusion coefficients. Since the latter can be determined from local sampling, i.e. without knowledge of the entire landscape, this offers a possible improvement over traditional method.

5.2 Theory and Methods

5.2.1 The stochastic approach

We assume that a reaction coordinate Q captures the basic features of the folding process and that this coordinate moves diffusively over a free energy landscape. The Fokker-Planck equation, describes this stochastic motion superimposed with

deterministic drift [125, 34],

$$\frac{\partial P(Q,t)}{\partial t} = -\frac{\partial}{\partial Q}\left[v(Q)P - \frac{\partial}{\partial Q}D(Q)P\right], \tag{5.1}$$

where Q is an arbitrary reaction coordinate, $P(Q,t)$ is its probability density, $v(Q)$ is the drift velocity, and $D(Q)$ is the diffusion coefficient. Note that we explicitly allow these coefficients to depend on position Q, but imagine that the change is reasonably slow and gradual. The short-time solution of this equation with initial condition $P(Q) = \delta(Q_c)$ at $t = 0$ is given by [125],

$$P(Q,t) = \frac{1}{\sqrt{4\pi D(Q_c)t}} exp\left(-\frac{(Q - Q_c - v(Q_c)t)^2}{4D(Q_c)t}\right), \tag{5.2}$$

which is a Gaussian distribution whose center at Q_c moves at a velocity $v(Q_c)$ and whose width grows as the square root of time (standard deviation $\sigma = \sqrt{2D(Q_c)t}$).

The basic idea of the method is to use the above expression to determine the drift velocity v and diffusion coefficient D from microscopic simulation data. By microscopic, we have in mind the situation where there is a complete model for the folding process which is to be approximated by the aforementioned stochastic reaction dynamics were obtained from simulation data by the following procedure: (i) All the simulation time-points with the reaction coordinate equal to some fixed Q (within some bin size δQ) were identified. (ii) Starting from each of these points, we determine the distribution $P(Q,t_1)$ and $P(Q,t_2)$ at elapsed times t_1 and t_2. These two histograms were then fitted to Eq. 6.2 and centers $Q_c(t_1)$ and $Q_c(t_2)$ and standard deviations $\sigma(t_1)$ and $\sigma(t_2)$ extracted. (iii) The drift and diffusion coefficients were calculated from the shift of centers and growths of widths by [125, 63]

$$v(Q) = \frac{Q_c(t_2) - Q_c(t_1)}{\Delta t} \tag{5.3}$$

and

$$D(Q) = \frac{\sigma^2(t_2) - \sigma^2(t_1)}{2\Delta t}, \tag{5.4}$$

where $\Delta t = t_2 - t_1$. Theoretically, we should take the limit $\Delta t \to 0$ to obtain exact values for v and D from the two data points at t_1 and t_2. In practice, the values of t_1 and t_2 were chosen small enough so that the aforementioned short-timed approximation is valid. To test this procedure, we also used a quadratic fit to data from multiple time points between t_1 and t_2 and then extrapolated to zero

elapsed time to obtain the drift and diffusion coefficients. The overall process is then repeated to obtain these coefficients as functions of Q.

5.2.2 Langevin dynamics

Once the drift and diffusion coefficients are obtained, we study Langevin dynamics for Q [125],

$$\frac{dQ}{dt} = v(Q) + \eta(Q,t), \tag{5.5}$$

where v is the drift velocity and η is a Gaussian-type white noise, $\langle \eta(Q,0)\eta(Q,t)\rangle = 2D(Q)\delta(t)$ or

$$P[\eta] \propto exp\left(-\int^{\delta t} \frac{\eta^2(t)}{4D}\, dt\right) \overset{\delta t \to 0}{\approx} exp\left(-\frac{\eta^2 \delta t}{4D}\right). \tag{5.6}$$

If we define $\tilde{\eta} = \eta\sqrt{\delta t}$, then the Langevin equation can be numerically solved by

$$Q(t+\delta t) = Q(t) + v(Q)\delta t + \tilde{\eta}(Q)\sqrt{\delta t}, \tag{5.7}$$

where $\tilde{\eta}$ is a re-defined Gaussian random number distribution with zero mean and the standard deviation $\sigma_{\tilde{\eta}} = \sqrt{2D}$. We employ this numerical algorithms for the Langevin dynamics simulations once v and D are known.

5.2.3 Reconstructing thermodynamics from the drifts and diffusion coefficients

After obtaining the local kinetic quantities, it is possible to reconstruct the thermodynamics (i.e., the free energy profile itself) from the computed quantities v and D. Specifically, the free energy profile $G(Q)$ may be obtained from the equilibrium solution of the Fokker-Planck equation (Eq. 6.1) by [34, 63],

$$G(Q)/k_BT = -\int_{Q_{ref}}^{Q} \frac{v(Q')}{D(Q')} dQ' + \log D(Q) + const. \tag{5.8}$$

where k_B is the Boltzmann factor and T is the temperature of the system. The additive constant takes care of the arbitrary free energy of the reference state Q_{ref} and also preserves consistent units. This should of course agree with the free energy generated by measuring the equilibrium distribution from the reaction-coordinate simulations. At the same time, a putative free energy $F_R(Q)$ can be

directly calculated by $F_R(Q) = -k_B T \log P(Q)$, where $P(Q)$ is the equilibrium probability distribution generated from the full simulations. Later, we will discuss the possibility that Eq. 6.10 generates a more robust reconstruction in the presence of sampling noise.

5.3 A simple test case: A two-state toy model

The concept of stochastic motion on a folding energy landscape can be illustrated for two model systems for protein folding: a toy model and a more realistic structure-based model. In both cases, drifts and diffusion coefficients are first extracted from simulation data and then the folding dynamics and thermodynamics are derived via Langevin simulations. Given the relative simplicity of the full models, we can also carry out direct folding simulations; this would not be possible for a realistic all-atom model of even the smallest proteins. We can then compare folding time distributions. We can therefore ask: to what extent can a single reaction coordinate quantitatively capture the folding dynamics in the high dimensional conformational space.

As a first example, we consider a two-dimensional diffusion-drift dynamics on the free energy surface (Fig. 5.2) given by

$$F(X,Y) = \left[(X^2-1)^2 + (Y-1)^2\right] \times \left[1 + (Y+1)^2\right], \tag{5.9}$$

where X and Y are two independent coordinates. In the context of protein folding, the two minima at $(X=-1, Y=1)$ and $(X=1, Y=1)$ in Fig. 5.2 can be referred to as the "folded" and "unfolded" states, respectively. The "microscopic" Langevin dynamics on this free energy surface is given by

$$\frac{dX}{dt} = -\Gamma \frac{\partial F}{\partial X} + \eta, \tag{5.10}$$

$$\frac{dY}{dt} = -\Gamma \frac{\partial F}{\partial Y} + \eta, \tag{5.11}$$

where $\Gamma = 1/\gamma$ (Γ is the (constant) mobility and γ is the friction constant), the drifts along X and Y are $-\Gamma \frac{\partial F}{\partial X}$ and $-\Gamma \frac{\partial F}{\partial Y}$, respectively, and η is a Gaussian white noise with a homogeneous diffusion coefficient D, $\langle \eta(0)\eta(t)\rangle = 2D\delta(t)$.

We first performed two-dimensional Langevin simulations. We chose $\Gamma = 1$ and $D = 1$ and therefore the effective temperature is $T = 1$ by the Einstein

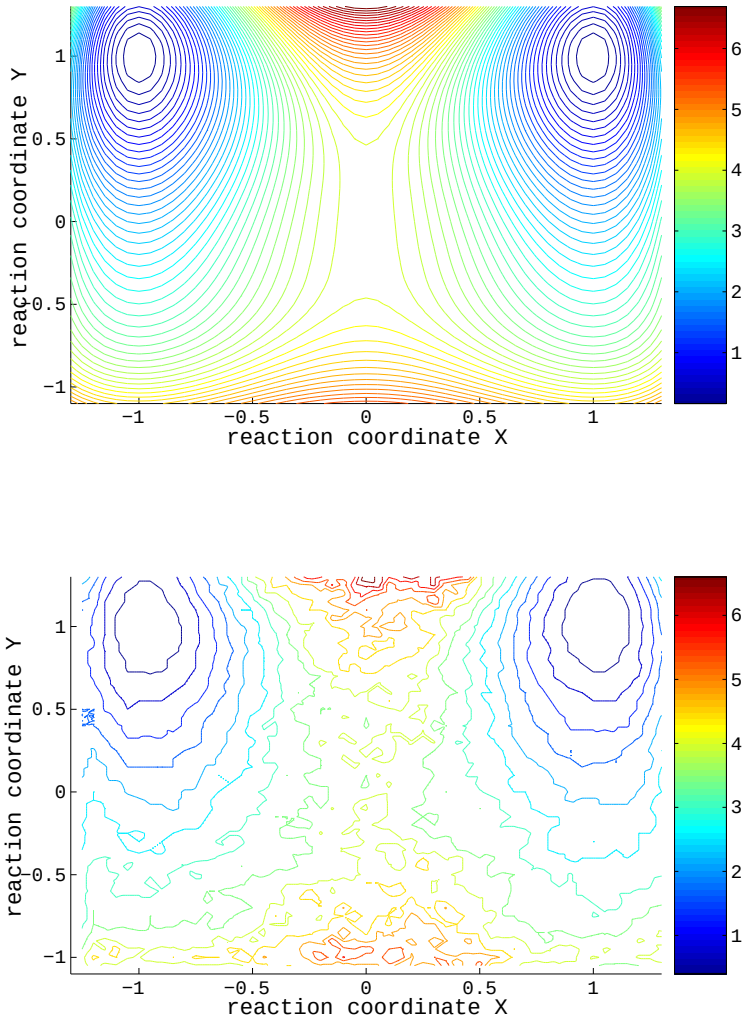

Figure 5.2 (Top) The two-dimensional free energy surface of $F(X,Y) = [(X^2-1)^2 + (Y-1)^2] \times [1 + (Y+1)^2]$ (in units of k_BT) with two minima at $(X=-1, Y=1)$ and $(X=1, Y=1)$, respectively. (Bottom) The free energy surface $F_{simulated}(X,Y) = -k_BT \log P(X,Y)$ computed from the two-dimensional Langevin simulations in Eq.(5.10 and 5.11) with constant $\Gamma = 1$ and $D = 1$ in both X and Y directions. The agreement between the simulated $F_{simulated}(X,Y)$ and theoretical $F(X,Y)$ validates our Langevin simulation protocol.

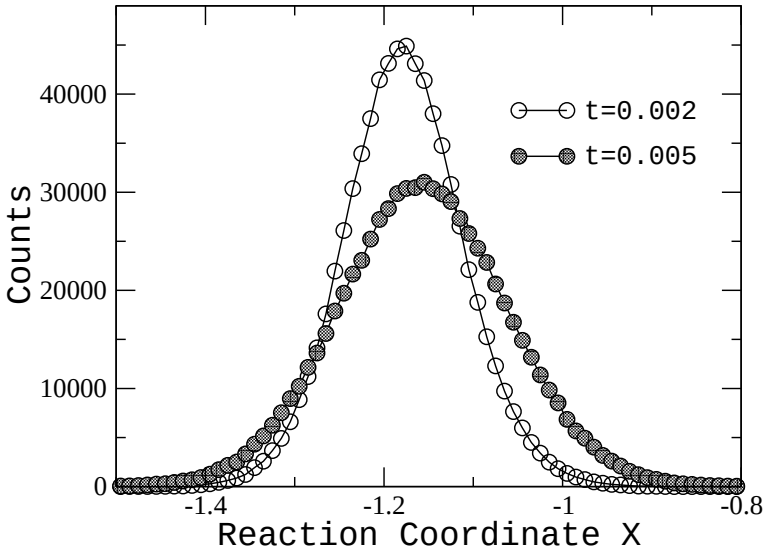

Figure 5.3 A typical example of the probability $P(X,t)$ as a Gaussian distribution centered at $X = -1.2$ at time $t = 0.002$ and $t = 0.005$. By measuring the shift of the center and growth of the width from the probability $P(X,t)$, we can determine the drifts and diffusion coefficients at each position along X.

relation $D/\Gamma = k_B T$. The units of temperature and energy have been chosen so that $k_B = 1$. The simulations were run for a total of 5×10^6 steps with a step size $\delta t = 0.001$, and the simulation time was picked to be much longer than the folding time so that complete thermal equilibrium was reached. One way to check this was to see the extent to which the Langevin simulations reproduced the model free energy surface (Eq. 5.9). From the two-dimensional simulation data, we computed a probability distribution $P(X,Y)$ and then evaluated the simulated free energy profile by $F_{simulated}(X,Y) = -k_B T \log P(X,Y)$. Figure 5.2 compares the simulated free energy surface and the theoretical free energy surface. The good agreement suggests that the Langevin simulation protocol we used is valid.

The single reaction coordinate X should be able to capture the essential dynamics in this two-dimensional model system. To test this, we needed to compare transition times for the exact dynamics versus that of a one-dimensional effective dynamics for X. We projected the two-dimensional Langevin simulation data into a single-reaction-coordinate probability distribution $P(X,t)$, where as discussed in the previous section, t is the elapsed time since X was at some specific value. Figure 5.3 shows a representative example centered at the specific value $X = -1.2$. The distribution is Gaussian, confirming that for these times the short-time ap-

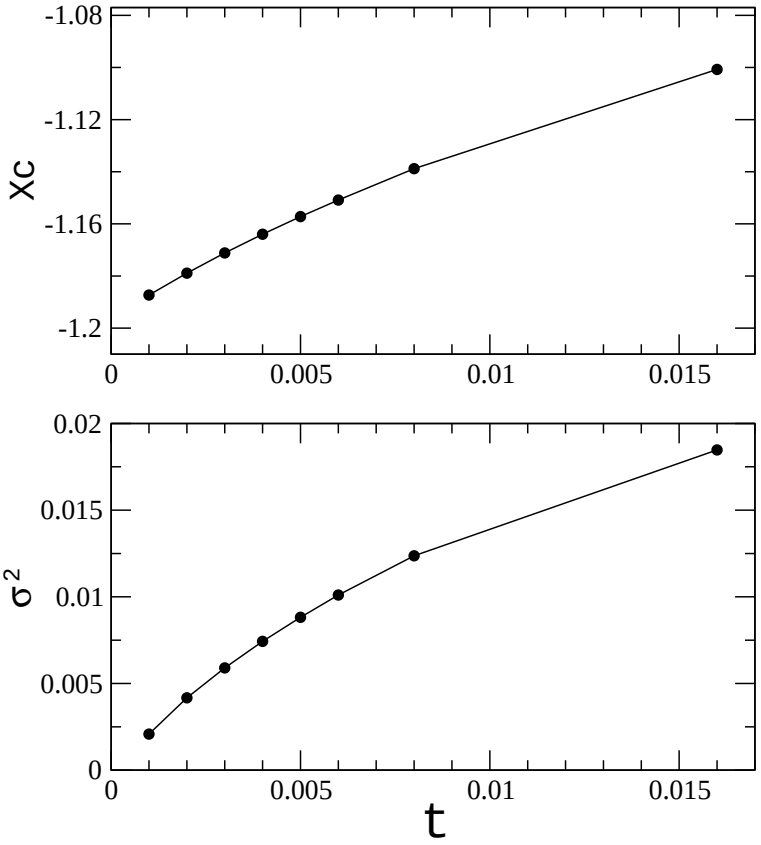

Figure 5.4 A representative example for the centers X_c and the width square σ^2 as a function of time t at $X = -1.2$. For the two-step procedure for extracting the drifts and diffusion coefficients, due to the nonlinear dependency, we used a quadratic fitting with 4 data points at times $(0.002, 0.003, 0.004, 0.005)$ and then extrapolated drifts and diffusion coefficients at $t = 0$.

proximation is valid. We then determined the local drifts $v(X)$ and diffusion coefficients $D(X)$ by measuring the temporal shift of centers and growth of widths according to Eq. 6.3 and 6.4. A linear fit for calculating v and D was used as discussed in Ref. [63]. Here we used a more sophisticated optimization procedure, first fitting multiple data points to a quadratic curve and then extrapolating the drift and diffusion coefficients to their value at $t = 0$. This extra care was taken due to the nonlinear dependence on time for both the centers (X_C) and the square of widths (σ^2) as shown in Fig. 5.4. For the quadratic fit, we picked four data points at times $(0.002, 0.003, 0.004, 0.005)$. In turns out (see Fig. 5.5 later) that this gives a highly accurate free energy reconstruction.

Figure 5.5 shows the computed drifts $v(X)$ and diffusion coefficients $D(X)$

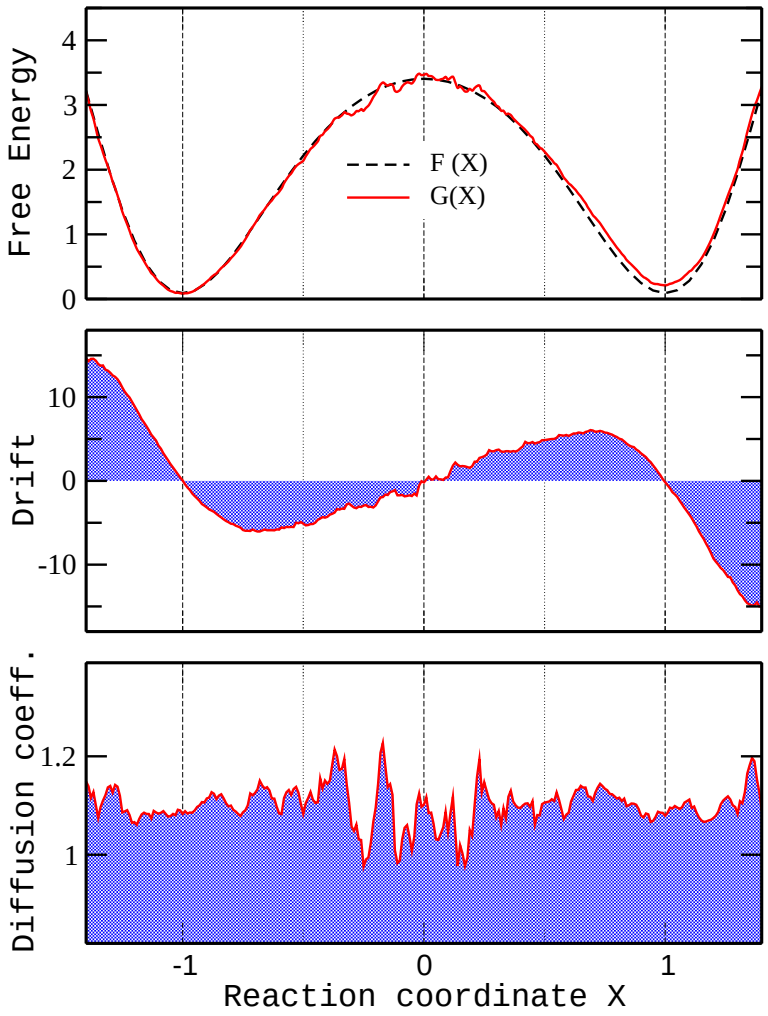

Figure 5.5 (Top) The one-dimensional free energy profile as a function of X, where $G(X)$ is computed from the kinetic approach from drifts and diffusion coefficients (Eq. 6.10) and $F(X) = -\log\left(\int e^{-F(X,Y)}dY\right)$ is the theoretical free energy profile by integrating out the Y dimension for $F(X,Y)$. (Middle and bottom) The position-dependent drifts $v(X)$ and diffusion coefficients $D(X)$ derived from the two-dimensional Langevin simulations. The bin size of $\delta X = 0.1$ was used at $t = 0$.

along the reaction coordinate X. At the "folded" and "unfolded" regions ($X = \pm 1$) as well as the transition barrier region ($X = 0$), the system moves very slowly with small drifts. In-between these regions, drift dominates over diffusion and the system goes through fast downhill folding. Since we started from constant diffusion coefficients in the two-dimensional case and since the motion in the two coordinates are uncoupled, we obtained near constant diffusion coefficients in this one-dimensional representation.

With the drift and diffusion coefficients in hand, we reconstructed the thermodynamics $G(X)$ from simulated $v(X)$ and $D(X)$ following the equilibrium solution of the Fokker-Planck equation (Eq. 6.10). As a comparison, we also calculated the exact theoretical one-dimensional free energy profile $F(X) = -\log\left(\int e^{-F(X,Y)}dY\right)$. We observed very good agreement between $G(X)$ and $F(X)$ as shown in Fig.5.5. This validates our procedure for extracting drifts $v(X)$ and diffusion coefficients $D(X)$ as well as the kinetic approach to reconstruct the thermodynamics from local v and D.

Next, we plugged in the kinetic quantities $v(X)$ and $D(X)$ into the Langevin dynamics, and carried out one-dimensional Langevin simulations using Eq. 6.9. Figure 5.6 shows a typical Langevin simulation trajectory as a function of X. The plot shows a two-state behavior of the system, consistent with the theoretical free energy potential.

We would now like to compare the folding time distributions obtained for the exact 2D model versus the reduced 1D dynamics. The folding time can be computed from one long or many short Langevin simulations; we chose a long simulation (10^8 steps with step size $\delta t = 0.001$). Figure 5.7 shows the folding time distribution from both two-dimensional Langevin simulations and one-dimensional Langevin simulations with simulated drifts $v(X)$ and diffusion coefficients $D(X)$. Two features are common to both of these. First, the distributions are nearly Poisson; the standard deviations are approximately equal to the means. These plots confirm that the system has exponential kinetics and two-state behavior. More importantly, the characteristic time or mean folding time curves in both cases are nearly identical. We conclude that the one-dimensional Langevin simulation, with X as a reaction coordinate, robustly recovers the two-dimensional folding time distribution within this model system.

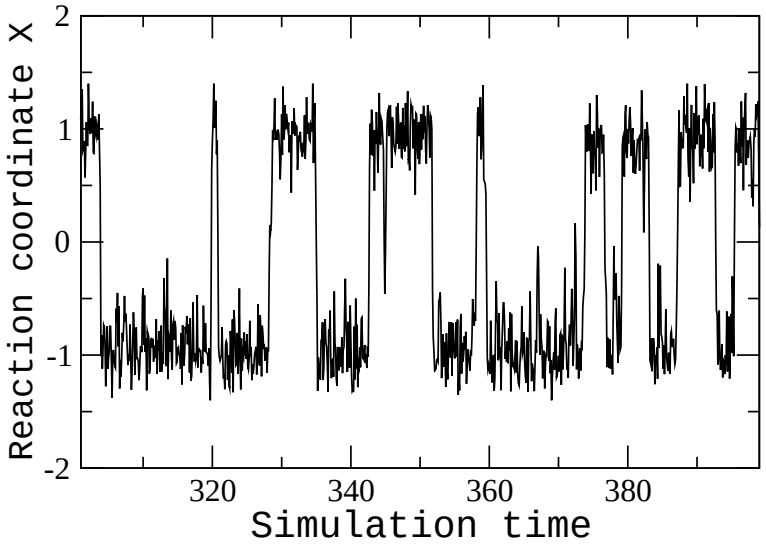

Figure 5.6 A typical one-dimensional Langevin simulation trajectory with simulated drifts $v(X)$ and diffusion coefficients $D(X)$ as shown in Fig. 5.5. A time step of $\delta t = 0.001$ was used.

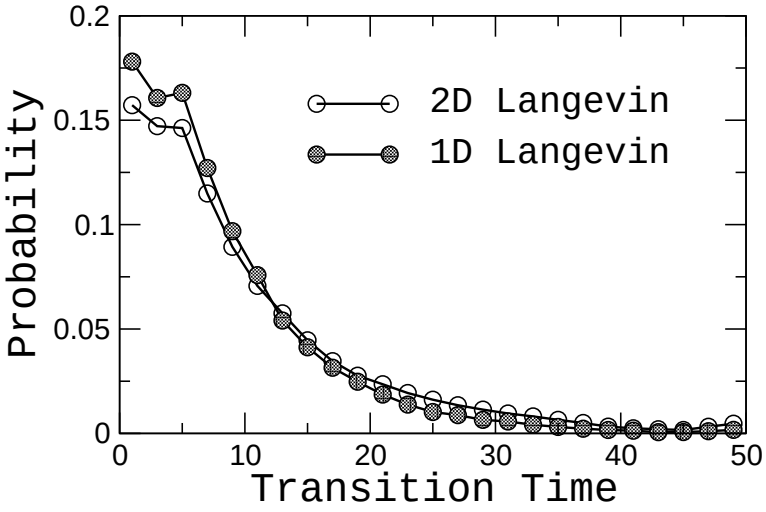

Figure 5.7 The distribution of transition times by two-dimensional Langevin simulations ($(\Gamma = 1$ and $D = 1)$ and one-dimensional Langevin simulations with simulated drifts $v(X)$ and diffusion coefficients $D(X)$ (shown in Fig. 5.5). The distribution values have been normalized. Both simulations were run for 10^8 steps with a step size $\delta t = 0.001$. Both distributions are nearly Poisson, the standard deviations being approximately equal to the means. By using a single reaction coordinate X, the one-dimensional Langevin simulations robustly recover the two-dimensional transition-time distribution.

In light of these encouraging results, it is reasonable to apply the diffusion-drift method to a more realistic model. In the next section we will describe protein folding dynamics with Gō-type simulations as stochastic dynamics on a Q landscape.

5.4 Diffusion-drift motion on a funneled landscape

As already discussed, protein folding is a complicated process where a protein folds into the native state through a high-dimensional configurational space [91]. However, the principle of minimal frustration suggests that even in this high-dimensional problem, it should be possible to find one or two coordinates that can capture the essence of the entire process. The key then, is to find the effective dynamics for these coordinates whose motion can quantitatively capture the folding kinetics.

To see how our procedure works for this problem, we used a structure-based (Gō-type) model to characterize the folding dynamics when the landscape is strongly funneled [69, 23, 89, 19, 96, 32, 16, 37]. Here, the number of native contacts, Q, was chosen as the single reaction coordinate. As a specific example, we used an SH3 domain that folds in a two-state manner [40]. A structure-based model of this protein was built from the NMR structure (1PNJ.pdb) as shown in Fig. 5.8. We used a slightly modified version of the Gō-type model to achieve a reasonable amount of sampling of the barrier-crossing. The energy function we used for the C_α Gō-type model is

$$\begin{aligned} E = & \sum_{\text{bonds}} K_r(r-r_0)^2 + \sum_{\text{angles}} K_\theta(\theta-\theta_0)^2 + \sum_{\text{dihedrals}} K_\phi^{(n)}(1-\cos(n(\phi-\phi_0))) \\ & + \sum_{contacts} \left\{ \epsilon_1 \left[5\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - 6\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{10} \right] + \epsilon_2 \left(\frac{\sigma}{r_{ij}}\right)^{12} \right\} \end{aligned} \tag{5.12}$$

where K_r, K_θ and K_ϕ are the force constants of the bond, angle, and dihedral angle for adjacent C_α atoms, respectively, and we chose $K_r = 100\varepsilon$, $K_\theta = 20\varepsilon$, $K_\phi^{(1)} = \varepsilon$, $K_\phi^{(3)} = 0.5\varepsilon$. The native experimental structure (r_0, θ_0 and ϕ_0) was taken from the NMR structure. We chose $\epsilon_1 = 0.5\varepsilon$ for native contacts and $\epsilon_2 = \varepsilon$ for non-native contacts. A reduced unit ε (= 1 kcal/mol) for energy and r_o (= 3.8

Å) for distance were also used. Further details and the simulation procedures can be found in previous publications [89, 19]. In this work, 9×10^6 configurational snapshots (separated by time intervals of 20 femtoseconds) were used for analysis. The simulation was run at a temperature near the folding temperature.

We followed the aforementioned procedure for extracting local drift and diffusion coefficients directly from the Gō-type simulation data. For the quadratic fitting of the center position and Gaussian width versus elapsed time, four data points at times $(3, 4, 5, 6)$ (in units of the time interval between snapshots) were used. The computed Q-dependent drift $v(Q)$ and diffusion coefficient $D(Q)$ are shown in Fig. 5.8.

Several interesting properties may be noted from the diffusive motion shown in Fig. 5.8. Most importantly, the configurational diffusion coefficient is not constant in the one-dimensional configuration space. Instead, it first increases and then decreases as a protein moves toward the folded state. The transition state region does not have the largest D and the folded state has a fairly large D, about 6 times larger than in the unfolded state. Both $v(Q)$ and $D(Q)$ depend simply on the local moves allowed to the system and the local landscape. The small diffusion in the unfolded region presumably reflects the relative difficulty of increasing Q by random fluctuations in the underlying model dynamics if we are far away from the native structure.

Using the computed $v(Q)$ and $D(Q)$, we reconstructed the free energy profile $G(Q)$ by using (Eq. 6.10). Figure 5.8 shows this profile, which indicates the protein has a two-state folding behavior, as expected. $G(Q)$ agrees reasonably well with $F(Q) = -k_B T \log P(Q)$ where $P(Q)$ is the probability taken directly from Gō-type simulation data.

Next, a Langevin simulation was be carried out for the single reaction coordinate Q using Eq. 6.9 as before. We used a time step $\delta t = 0.1$ for the one-dimensional Langevin simulation. Figure 5.9 gives a typical Langevin simulation trajectory, which clearly demonstrates the two-state folding transition as expected from the two-state behavior shown in Fig. 5.8. Finally, we can determine the folding time distribution as shown in Fig. 5.10. The dashed curve is the distribution for the simulated folding time computed from the diffusion-drift approach and the solid curve is directly from Gō-type simulations. Again, both curves show a near

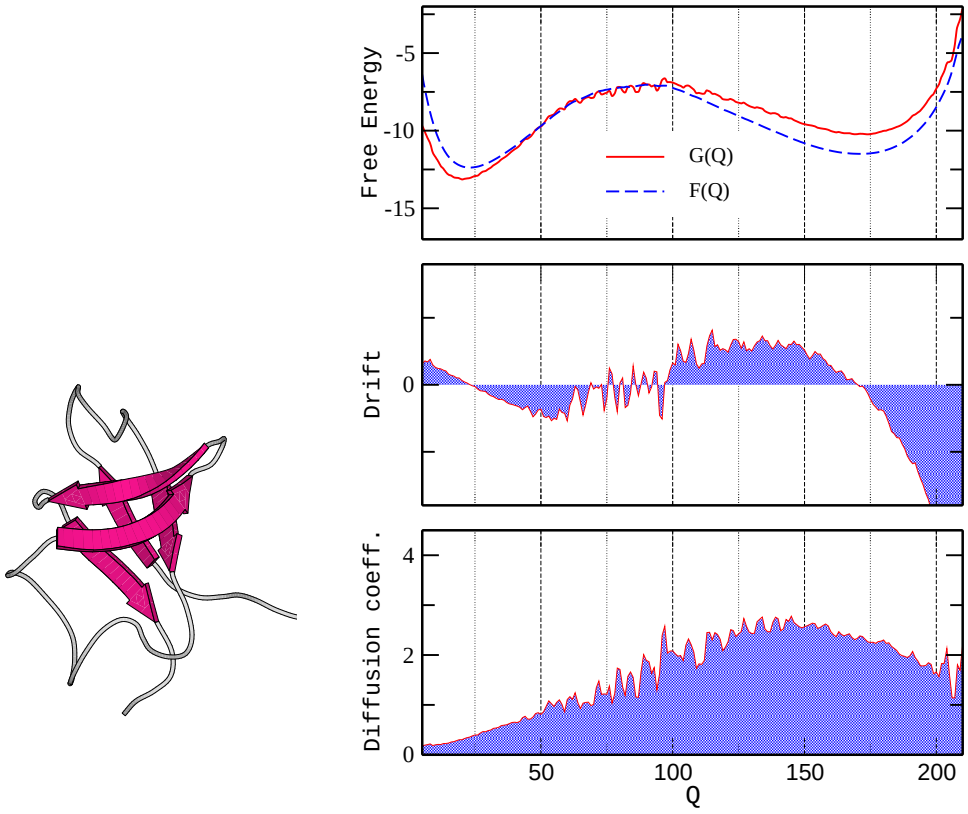

Figure 5.8 (Left) Cartoon representation of the SH3 domain (1PNJ.pdb). (Right) The reconstructed free energy profile and drifts $v(Q)$ and diffusion coefficients $D(Q)$ extracted from Gō-type simulations. $G(Q)$ is the free energy reproduced from $v(Q)$ and $D(Q)$ by using the kinetic approach (Eq. 6.10). $F(Q)$ is calculated from Gō-type simulations by $F(Q) = -\log P(Q)$ where $P(Q)$ is the probability distribution of Q. Q is the number of native contacts formed and $Q = 213$ for the native SH3 (using the Contacts of Structural Units software [113]). The Gō-type simulation was run at a temperature near the folding temperature. A total of 9×10^6 snapshots were used for the analysis. For the quadratic fitting, we picked four data points at times $(3, 4, 5, 6)$ (in units of the time interval between snapshots). The free energies are in units of $k_B T$. The bin size of $\delta Q = 3$ was used at $t = 0$.

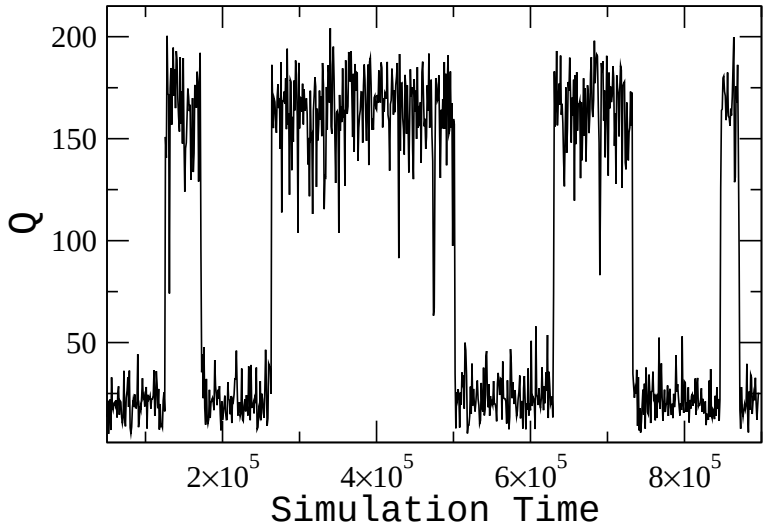

Figure 5.9 A typical Langevin simulation trajectory for the number of native contacts Q with simulated $v(Q)$ and $D(Q)$ from the full Gō-type simulations. The time step $\delta t = 0.1$ was used for the Langevin simulations.

Poisson distribution, as demonstrated in various protein models [114, 44, 78]. The difference is quite small if one considers the complexity of the folding process compared to the simplicity of our effective dynamics: after all, the protein folds in an extremely high-dimensional configuration space (with about 3^N degrees of freedom in the Gō-type model, where N is the number of residues).

5.5 Discussions and Conclusions

We have shown that one can derive effective stochastic equations for one reaction coordinate, whose computed kinetics allows one to adequately recover folding dynamics from an underlying multi-dimensional model. This was demonstrated for both a toy model and a Gō-type model. This relatively simple procedure appears capable of both mimicking the qualitative features of the folding and, if we look again at the two folding time distribution curves shown in Fig. 5.10, getting within a factor of 2 of the quantitatively correct kinetics.

As far as the basic method is concerned, the approximation we make is essentially an adiabatic one. We are assuming that the rate of progress in the folding direction is rather slow and hence there is time during the barrier crossing to forget about the detailed microscopic configuration at any one time-step. We can then

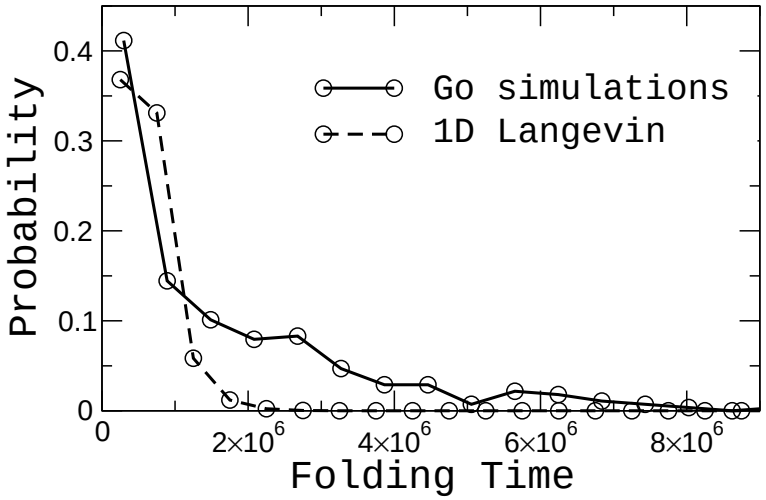

Figure 5.10 The folding time comparison between the Gō-type simulations and the Langevin simulations using Q as a single reaction coordinate, which remarkably reproduces the folding time within the same order of magnitude with the full Gō-type simulations. The Langevin simulations were run for 4×10^9 steps with a time step $\delta t = 0.1$. A total of 1,143 folding transitions from Langevin simulations and 277 folding events from full Gō-type simulations were used for comparison of folding time distribution.

average over all such configurations and derive the probability distribution of the macroscopic variable Q at the next time point. As already mentioned, this idea is specific realization of the "equation-free methodology" proposed by Kevrekides and co-workers. The important factor at work here is that the elucidation in recent years of how proteins fold with minimal frustration lends support to the idea that such a reduction might be semi-quantitatively valid.

Of course, our approach is not perfect or exact. A closer look at the folding curve for the Gō-type model seems to indicate deviations from pure two-state exponential kinetics, possibly due to local traps. One must be slightly careful, here, as we only have about 300 folding trajectories for the full model and hence the measured distribution is somewhat noisy. Nevertheless, to the extent that these slow events are real, they do not seem to have been captured by the one dimensional reduction. One can imagine more complex reductions, keeping for example two reaction coordinates, but this has not yet been attempted. We do expect on general grounds that our calculation will always give a sort of lower bound on the folding time distribution, as the breakdown of ergodicity in the

exploration of the microscopic degrees of freedom (due e.g. to local traps not seen in the 1D reduction) should always make folding proceed more slowly.

Of course, our real goal is to apply this procedure to more realistic computational simulations such as all-atom replica exchange simulations [120, 105, 32]. This would enable us to get an estimate of the folding time distribution even when the longest obtainable simulation trajectories are shorter than the characteristic time scale on which folding events occur. Work along this direction is in progress and will be communicated in future publications.

In passing we wish to mention that the kinetic approach therefore provides us with an alternative way to build a free energy profile besides the conventional WHAM algorithm [30, 65, 43, 111, 115, 133]. As long as sufficient sampling of local conformation space is available for extracting drift and diffusion coefficients along the reaction coordinate, we can determine by integration a global free energy profile. Simulation convergence, as required by the WHAM, is no longer a necessary condition for obtaining thermodynamics. This is especially important for all-atom simulations of large protein systems. Again, this will be discussed elsewhere.

Finally, there is no reason in principle why the concept of stochastic dynamics for several macroscopic degrees of freedom needs to be limited in its application to just the folding problem. For example, we hope to investigate functional motions by this approach, perhaps by using a few normal modes as coordinates [84]. We hope that this type of approach, namely using the full dynamics to build (instead of merely postulating) accurate phenomenological models, can get us beyond the obvious limitations of all-atom simulations for answering questions on relevant biological timescales.

Acknowledgments

S.Y. thanks Peter McHale, Shachi Gosavi, and Alex Schug for critical reading of the manuscript; we also acknowledge David Kessler, Angel García, Peter Wolynes, and Attila Szabo for comments and discussions. This work was supported by the NSF-sponsored Center for Theoretical Biological Physics through grants PHY0216576 and PHY0225630 and an additional NSF grant MCB-0084797.

This chapter has been accepted for publication in *J. Chem. Phys.*, Sichun Yang,

Jose Onuchic, and Herbert Levine [135].

6

Predicting the Folding Time from Replica Exchange Simulations

We present an approach to predicting the folding time from replica exchange simulations using stochastic reaction-coordinate dynamics in which effective drift velocities and diffusion coefficients are determined from the short-time dynamics of replica exchange simulations, for the second β-hairpin of the B domain of protein G. The resultant Langevin equation can then be used to directly simulate the folding kinetics. The folding time prediction agrees well with experimental measurement. Therefore, we have in hand a fast numerical tool for calculating the folding kinetic properties from replica exchange simulations.

6.1 Introduction

Protein folding is a complex dynamical process occurring in a high-dimensional configurational space [91]. It is usually not feasible to tackle this problem by using atomic computer simulations, as folding time-scales are often many orders of magnitude larger than the achievable simulation time [28]. As a particular powerful tool to enhance sampling, replica exchange molecular dynamics (REMD) simulations are often used to explore the folding/oligomerization energy landscape by improving barrier-crossing [120, 105, 140, 32, 132, 2, 133], in which multiple replicas of protein conformations are simulated in parallel over a given temperature range. The exchange is done by periodically swapping protein conformations by

using a Metropolis criterion to ensure efficient sampling for barrier crossing. Although REMD enhances sampling compared with conventional simulation methods, it loses the *long-time* (e.g., typical folding times of microseconds or higher) dynamics due to the randomness or discontinuity from frequent exchange between neighboring replicas. Therefore, in this sense no direct kinetic information can be obtained.

To overcome this difficulty, we here present an approach to folding time prediction in which one can extract *short-time* (in order of picoseconds) dynamics from replica exchange simulations and then use the effective stochastic dynamics namely determining the drift velocity and diffusion coefficient on the landscape to recover the folding dynamics or further determine the dominated and/or competing folding pathways (not investigated here). In the effective stochastic model, which has been described in a previous paper [135] as a specific realization of the "equation-free computational methods" [63], we considered a diffusive motion superimposed with deterministic drift influenced (reflected?) by the local folding landscape. This stochastic or diffusive model has been successful in describing folding dynamics both analytically [114, 97, 85, 48, 126] and computationally [135]. The goal here is to establish a theoretical framework to investigate the folding kinetics combined with the power of REMD simulations, for the case of the β-hairpin which was considered as a simple model system by many research groups [7, 86, 26, 92, 60, 77, 41, 33, 140, 136, 121, 2, 10].

The outline of this work is as follows. First, we started with two sets of atomic replica exchange simulations covering from folded to unfolded states. Then we showed how to extract the reaction-coordinate-dependent drifts and diffusion coefficients from these simulation data. With the determined drifts and diffusion coefficients, we can simulate the folding dynamics very fast by using a reduced one-dimensional Langevin approach. Finally, the folding time can be predicted from Langevin simulations and eventually compared with experimental measurement [86].

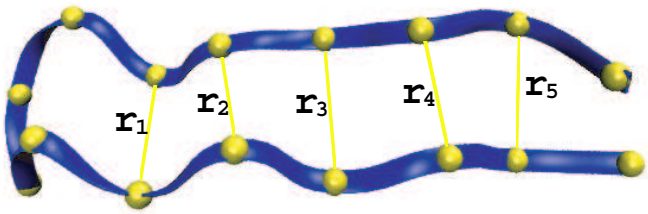

Figure 6.1 The native structure for the second β-hairpin of the B domain of protein G. The parameter of $R = \sum_{i=1}^{5} \frac{r_i^o}{r_i}$ was used as a reaction coordinate to characterize the folding process, where r_i is defined as the distance between nearest C_α atoms (yellow dots) in the native structure (r_i^o).

6.2 Results

The β-hairpin structure is taken from the C-terminal fragment of the B domain of protein G [7, 86] (see Fig. 6.1). To explore the folding configurational space, we chose

$$R = \sum_{i=1}^{5} \frac{r_i^o}{r_i}$$

as a reaction coordinate, where r_i is defined as the distance between nearest C_α atoms in the native structure (r_i^o). The system with explicit solvent is simulated by two sets of replica exchange simulations: one is starting from the native state and the other is from a random structure with RMSD=4.47Å (see Simulation details). These two sets of REMD simulations were carried out to sample the entire configurational space from unfolded states ($R \sim 1$) to folded ($R \sim 5$).

6.2.1 Kinetics from replica exchange simulations

Replica exchange simulations has been widely used as a powerful sampling method to explore the folding energy landscapes because efficient barrier-crossing can be achieved at high temperatures [120, 105, 140, 32, 2]. The exchange between replicas is done by periodically swapping protein coordinates whenever Metropolis criterion is ensured to achieve a canonical distribution. Figure 6.2 shows a small time sequence of a typical replica exchange trajectory at a given target temperature, where protein coordinates represented by the replica index are frequently swapped. In order to have efficient sampling to cross transition barriers, the ex-

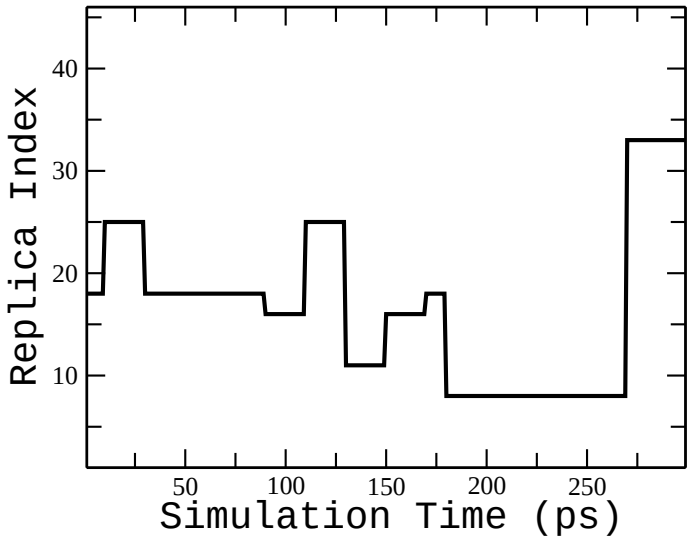

Figure 6.2 A typical replica exchange simulation trajectory at a given target temperature. Shown is the replica index representing protein coordinates of simulation time. Clearly, the REMD simulations produce a large amount of short continuous trajectories (each segment of short trajectories is in order of picoseconds or tens of picoseconds), which provide *short-time* kinetic dynamics. The basic idea here is that the information of *short-time* dynamics can be derived from the replica exchange simulations and then used for further description of folding process.

change must be attempted very often in a time scale of picoseconds. However, the time scale over which a REMD trajectory evolves is too short (compared with the folding time of microseconds) to obtain good kinetic data by conventional method (e.g., statistical average over many microsecond trajectories). We present here a method to circumvent this difficulty. The method extracts kinetic data from short (picoseconds or tens of picoseconds) trajectories generated by REMD simulations as exampled in Fig. 6.2. In this way, we reduce the high-dimensional folding problem to 1D. These kinetic parameters are then used to reconstruct the full (microseconds or higher) dynamics *via* a 1D Langevin simulation, which is much faster than the full atomic simulation. We thus use this method to predict a folding time which is in good agreement with experiment.

6.2.2 Extracting reaction-coordinate-dependent drifts and diffusion coefficients

It is clear that one cannot directly determine true kinetic rates and folding time distribution from the landscape alone. We must also have dynamics on this landscape, which in practice means knowing the drift velocity and diffusion coefficient on the landscape [63, 135]. We assume that the reaction coordinate R captures the basic feature of diffusive motion on the folding energy landscape. From the *short-time* dynamics provided by replica exchange simulations, one can explicitly determine drifts and diffusion coefficients as functions of R by measuring the temporal shifts of centers and growth of widths for the Gaussian distribution at some fixed R according to Eqs. (6.3 and 6.4) (see details in Appendix and applications for simplified models in Ref. [135]). Figure 6.3 shows these computed drifts $v(R)$ and diffusion coefficients $D(R)$ at each position throughout the folding landscape.

Several interesting properties may be noted from the stochastic motion shown in Fig. 6.3. First, the configurational diffusion coefficient in not constant in the R-space. Instead, it first increases gradually as the protein moves toward the folded state, similar to what we observed in a simplified Gō-like model [135]. Second, the folded state in this case ($R \sim 4.8$) has a fairly small D, which about is the same as the unfolded states. The small diffusion in both the folded and unfolded regions presumably reflects the relative difficulty of changing R by random fluctuations due to the configurational entropy penalty.

With these effective kinetic parameters in hand, one can reconstruct the thermodynamics $F(R)$ from computed kinetic quantities $v(R)$ and $D(R)$, following the equilibrium solution of the Fokker-Planck equation (Eq. 6.10, see details in Appendix). The reconstructed free energy profile by this kinetic approach is shown in Fig. 6.3. The system first crosses some small barriers of size similar to thermal fluctuations and a relative high barrier at $R \sim 3.8$ (around 3 k_BT) and then climbs a higher barrier at $R \sim 4.5$ (around 12 k_BT) and eventually reaches the folded state.

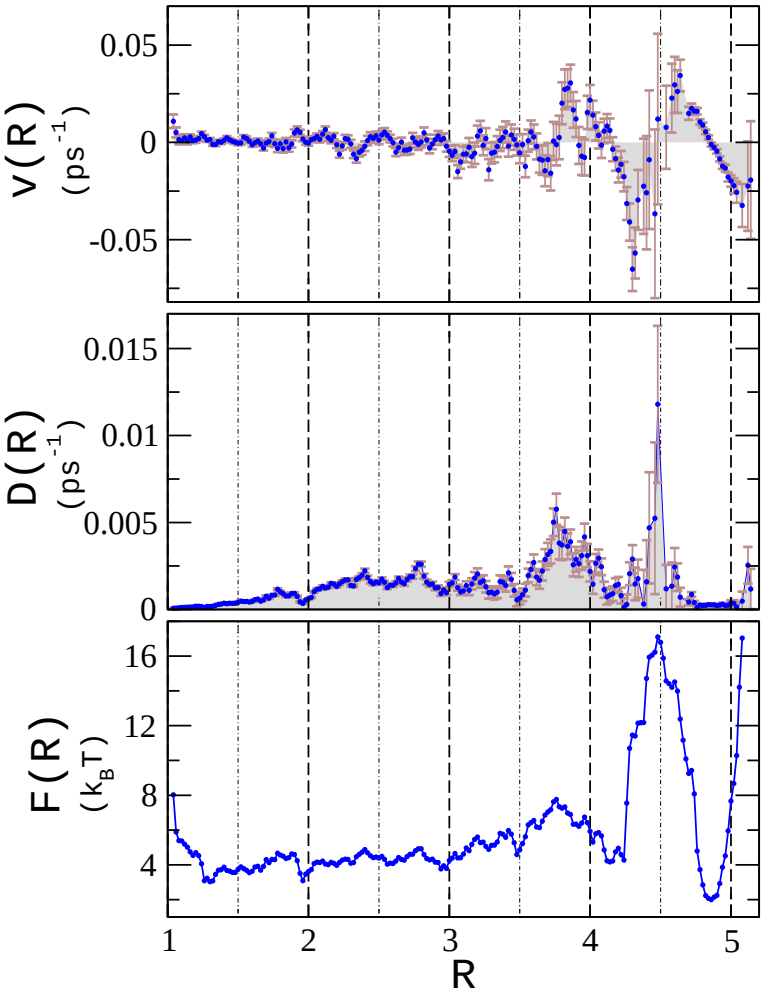

Figure 6.3 The drifts and diffusion coefficients as a function of R at $T = 300\ K$. A linear fitting including data points at times $(0.4, 0.7)\ ps$ was used. The bin sizes of $\delta R = 0.05$ and $\delta t = 0.3\ ps$ were used to have a reasonable statistics. The error bars for $v(R)$ and $D(R)$ shown in brown were calculated according to Eqs. (6.5 and 6.6). The free energy profile $F(R)$ constructed from $v(R)$ and $D(R)$ are also shown by using Eq. 6.10.

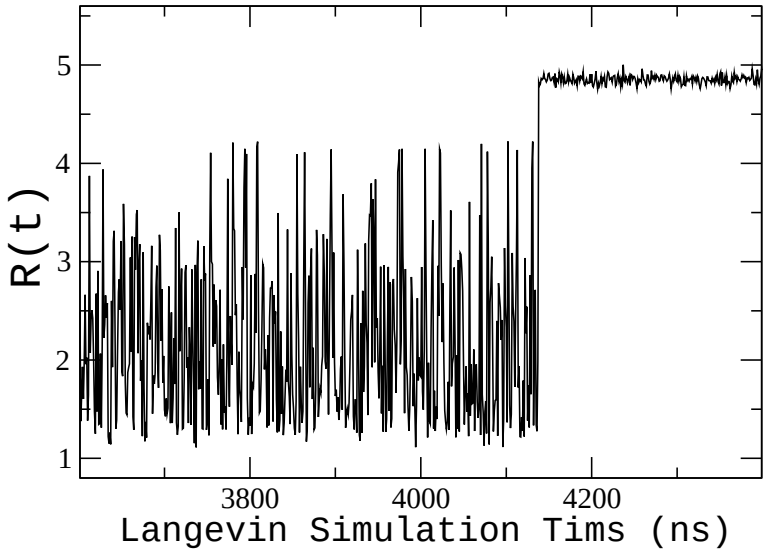

Figure 6.4 A typical Langevin simulation trajectory describes the dynamic process for the single reaction coordinate R with determined drifts and diffusion coefficients. A time step $\delta t = 0.1\ ps$ was used for the Langevin simulation.

6.2.3 Folding time prediction from Langevin simulations

Once effective drifts and diffusion coefficients are determined from atomic REMD simulations, a Langevin dynamics can be performed for the single reaction coordinate R according to Eqs. 6.7 and 6.9, to adequately recover the folding dynamics from an underlying multi-dimensional process. Figure 6.4 shows a typical 1D Langevin simulation trajectory $R(t)$, where a time step $\delta t = 0.1\ ps$ was used and the system initially started from a unfolded state at $R = 1.2$.

We now can examine the folding time distribution from Langevin simulation. We run 1,000 independent Langevin simulations, which all start from an unfolded state $R = 1.2$. A folding event is considered to have occurred once the system reaches $R = 4.8$. The folding time is defined here as the first passage time it takes from the starting point to the folded state. Figure 6.5 shows the histogram of folding times corresponding to these Langevin folding simulations. The distribution is nearly Poisson; the standard deviation is approximately equal to the mean. The plot also shows that the system has a near single exponential kinetics, corresponding to one dominant barrier on the landscape. We fit the folding time distribution to a single exponential function and obtained a relaxation time constant 3 μs. Compared with the experimental folding time of 6 μs [86], the difference is quite

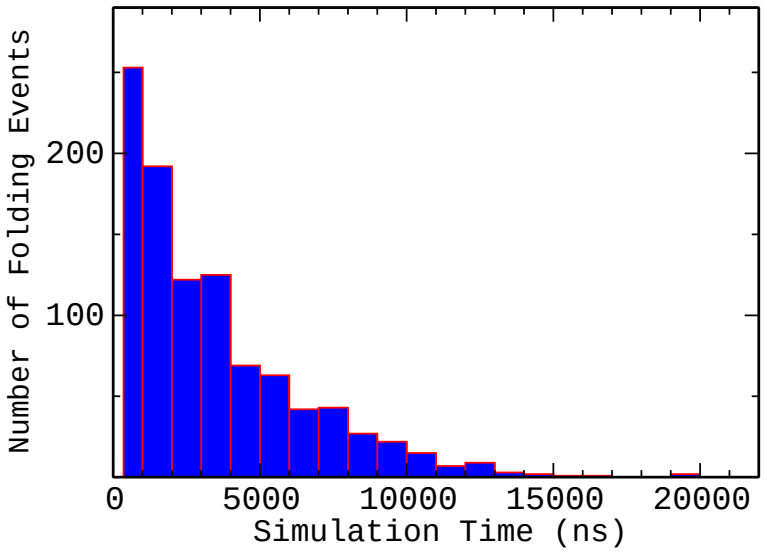

Figure 6.5 The simulated folding time is $(3.4 \pm 0.1)\ \mu s$ from 1,000 folding events by using Langevin simulations, in which the initial starting point is $R = 1.2$ and the system is considered to be folded when it reaches $R = 4.8$. And the experimental folding time is 6 μs.

small if one considers the complexity of the folding process compared to the simplicity of our effective dynamics for the single reaction coordinate: after all, the system folds in an extremely high-dimensional configurational space (with about 3^n degrees of freedom, where n is the number of atoms).

6.3 Discussion and Conclusions

We have shown that one can extract kinetics from replica exchange simulations and then use effective stochastic dynamics for one reaction coordinate to recover folding dynamics from an underlying multi-dimensional model. This allows us to have a detailed kinetics analysis. The simple idea appears capable to getting the β-hairpin folding time within a factor of 2 of the experimental kinetics. We see that one dimensional reduction by using the reaction coordinate captures the basic feature of folding dynamics. Of course, one can imagine more complex reductions, keeping for example two reaction coordinates, but this has not yet been attempted. The approach can be further used to investigate the hairpin folding mechanism such as previous proposed models under debate: the hydrogen-bond zipping model and the hydrophobic collapse model [86, 26, 9], which is not our goal here.

The folding time prediction can also be improved by more sampling around the major barrier region ($R \sim 4.5$) in which we could obtain drifts and diffusion coefficients with a better accuracy. This also will help more accurate determination of the thermodynamics (i.e., the free energy profile of R) which is majorally determined by integration of drifts and diffusion coefficients in this case. The prediction can also be improved by introducing as second-order correction for the region around barrier since the slope is rather steep. As a first-order approximation, our methods give a reasonable agreement with experimental folding time value. Overall, the agreement between simulated and theoretical folding time is better than could be expected, considering that we are using many simplifications.

Additional tests need to be done in order to verify that the one reaction coordinate approach is indeed giving a correct prediction. First, we assume that the reaction coordinate R can capture the basic features of the folding process. The physical justification of this approximation has to be evaluated. This can be addressed by checking whether there are extra major barriers in a higher-dimensional space (e.g., 2-dimensional). Further tests, to check whether or not the diffusive dynamics along R gives correct predictions, include (1) Can the folding dynamics near the major free energy barrier be considered diffusive? (2) From the simulated free energy profile, one can estimate that the early folding process (e.g., from $R = 1$ to $R = 3$) is very fast and on the order of nanoseconds. Direct kinetic simulations should be used to check this. (3) The non-Markovian correction might help improve the predictions, since there might be non-zero time correlations in R, especially at sharp transition regions.

6.4 Theory and Methods

6.4.1 Simulation details

The native structure is taken from the β-hairpin of the C-terminal fragment (residues 41 - 56) of protein G B1 domain [7, 86], blocked with the normal ACE and NME groups at the N- and C- termini. The peptide is solvated with 2,770 water molecules (TIP3P) and neutralized with three counter Na+ ions, resulting in a total 8,569 atoms. All structures were prepared and simulated by using the GROMACS

3.2 package [116] with the OPLS-AA/L all-atom force field [54]. Temperature coupling was performed by Nose-Hoover algorithms, and constant volume was used with a cubic box size of 43.34 Å. All simulations were performed by using the REMD method [120, 105, 140, 32]. Exchanges were attempted every 500 integration steps with a integration step of 2 fs. Protein configurations were saved every 0.1 ps. Two sets of REMD simulations were carried out starting with the native structure and a random conformation (RMSD=4.47 Å), respectively. The former has 42 replicas at temperatures between 275 K and 551 K and was simulated for a total of 20 ns for each replica; the latter has 46 replicas between 274 K and 571 K and was simulated for 30 ns for each replica. Thus total simulation time is 2.22 μs.

6.4.2 The stochastic approach

We assume that a reaction coordinate R captures the basic features of the folding process and that this coordinate moves diffusively over a free energy landscape. The Fokker-Planck equation, describes this stochastic motion superimposed with deterministic drift [125, 34],

$$\frac{\partial P(R,t)}{\partial t} = -\frac{\partial}{\partial R}\left[v(R)P - \frac{\partial}{\partial R}D(R)P\right], \tag{6.1}$$

where $P(R,t)$ is the probability density, $v(R)$ is the drift velocity, and $D(R)$ is the diffusion coefficient. Note that we explicitly allow these coefficients to depend on position R, but imagine that the change is reasonably slow and gradual. The short-time solution of this equation with initial condition $P(R) = \delta(R_c)$ at $t = 0$ is given by [125],

$$P(R,t) = \frac{1}{\sqrt{4\pi D(R_c)t}} exp\left(-\frac{(R - R_c - v(R_c)t)^2}{4D(R_c)t}\right), \tag{6.2}$$

which is a Gaussian distribution whose center at R_c moves at a velocity $v(R_c)$ and whose width grows as the square root of time (standard deviation $\sigma = \sqrt{2D(R_c)t}$).

The basic idea of the method is to use the above expression to determine the drift velocity v and diffusion coefficient D from microscopic simulation data. By microscopic, we have in mind the situation where there is a complete model for the folding process which is to be approximated by the aforementioned stochastic

reaction dynamics were obtained from simulation data by the following procedure: (i) All the simulation time-points with the reaction coordinate equal to some fixed R (within some bin size $\delta R = 0.0$) were identified. (ii) Starting from each of these points, we determine the distribution $P(R, t_1)$ and $P(R, t_2)$ at elapsed times t_1 and t_2. These two histograms were then fitted to Eq. 6.2 and centers $R_c(t_1)$ and $R_c(t_2)$ and standard deviations $\sigma(t_1)$ and $\sigma(t_2)$ extracted. (iii) The drift and diffusion coefficients were calculated from the shift of centers and growths of widths by [125, 63]

$$v(R) = \frac{R_c(t_2) - R_c(t_1)}{\Delta t} \tag{6.3}$$

and

$$D(R) = \frac{\sigma^2(t_2) - \sigma^2(t_1)}{2\Delta t}, \tag{6.4}$$

where $\Delta t = t_2 - t_1$. Theoretically, we should take the limit $\Delta t \to 0$ to obtain exact values for v and D from the two data points at t_1 and t_2. In practice, the values of t_1 and t_2 were chosen small enough so that the aforementioned short-timed approximation is valid. To test this procedure, we used a linear fit to data from multiple time points between t_1 and t_2 to obtain the drift and diffusion coefficients. The overall process is then repeated to obtain these coefficients as functions of R.

The uncertainties in the drifts $v(R)$ and diffusion coefficients $D(R)$ were obtained by [122]

$$\delta v(R) = \frac{1}{\Delta t}\sqrt{\frac{\sigma^2(t_1)}{N_1} + \frac{\sigma^2(t_2)}{N_2}} \tag{6.5}$$

and

$$\delta D(R) = \frac{1}{\Delta t}\sqrt{\frac{\sigma^4(t_1)}{2(N_1 - 1)} + \frac{\sigma^4(t_2)}{2(N_2 - 1)}}, \tag{6.6}$$

where N_1 and N_2 are numbers of data points used for the Gaussian fitting at t_1 and t_2, respectively.

6.4.3 Langevin dynamics

Once the drift and diffusion coefficients are obtained, we study Langevin dynamics for R [125],

$$\frac{dR}{dt} = v(R) + \eta(R, t), \tag{6.7}$$

where v is the drift velocity and η is a Gaussian-type white noise, $\langle \eta(R,0)\eta(R,t)\rangle = 2D(R)\delta(t)$ or

$$P[\eta] \propto exp\left(-\int^{\delta t} \frac{\eta^2(t)}{4D}\, dt\right) \overset{\delta t \to 0}{\approx} exp\left(-\frac{\eta^2 \delta t}{4D}\right). \tag{6.8}$$

If we define $\tilde{\eta} = \eta\sqrt{\delta t}$, then the Langevin equation can be numerically solved by

$$R(t+\delta t) = R(t) + v(R)\delta t + \tilde{\eta}(R)\sqrt{\delta t}, \tag{6.9}$$

where $\tilde{\eta}$ is a re-defined Gaussian random number distribution with zero mean and the standard deviation $\sigma_{\tilde{\eta}} = \sqrt{2D}$. We employ this numerical algorithms for the Langevin dynamics simulations once v and D are known.

6.4.4 Reconstructing thermodynamics from the drifts and diffusion coefficients

After obtaining the local kinetic quantities, it is possible to reconstruct the thermodynamics (i.e., the free energy profile itself) from the computed quantities v and D. Specifically, the free energy profile $F(R)$ may be obtained from the equilibrium solution of the Fokker-Planck equation (Eq. 6.1) by [34, 63],

$$F(R)/k_BT = -\int_{R_{ref}}^{R} \frac{v(R')}{D(R')} dR' + \log D(R) + const. \tag{6.10}$$

where k_B is the Boltzmann factor and T is the temperature of the system. The additive constant takes care of the arbitrary free energy of the reference state R_{ref} and also preserves consistent units.

Acknowledgments

Computing is supported by the San Diego Supercomputer Center. This work was supported by the NSF-sponsored Center for Theoretical Biological Physics.

The chapter is being prepared for publication with Herbert Levine, Angel Garcia, and Jose Onuchic.

7

Summary and Future Directions

In this dissertation I've focused on two major topics: One aims to investigate the underlying principles that control domain swapping. The other aims to predict the folding kinetics by using all-atom simulations. To summarize, I'm going to describe the success of our models, the difficulties we have experienced during the method development process, the limitations of the methods, and future improvements.

7.1 The symmetrized Gō model

7.1.1 The successes

We have introduced and implemented a symmetrized Gō model to study the molecular assembly process via domain swapping. The model clearly shows that the monomeric topology determines the domain-swapped conformation. We also observed the importance of flexibility of hinge regions. It would be much more difficult to form the correct domain-swapped conformations without the flexible hinge, although it is not the main factor that determines the correct domain-swapped structure. Our results show that the symmetrized Gō model is still able to predict the correct structure when more connecting loops are flexible.

7.1.2 Limitations

Although our model can predict the correct domain-swapped structure, it is not perfect. The model is an extension of the regular Gō model, which only considers the monomeric native contacts. Therefore, it is not able to capture the extra contacts that exist only in the domain-swapping conformations. This could be improved by selecting the lowest energy states among all possible orientations when domain swapping occurs. The selection could be done by an extensive conformational search with atomic details.

7.2 The diffusive dynamics

This study has shown that one can derive effective stochastic equations with only one reaction coordinate, whose computed kinetics allows one to adequately recover folding dynamics from an underlying multi-dimensional model. This was demonstrated for the cases of a toy model, a Gō-type model, and a realistic model using all-atom simulations for β-hairpin.

This newly developed method will be very important for all-atom simulations of large protein systems, especially when the obtainable simulation trajectories are much shorter than the characteristic time scale on which folding events occur. By using effective kinetic quantities determined from local sampling, we can determine by integration a global kinetic information as well as free energy profiles.

7.2.1 The successes

In the reduced lower-dimensional description, each protein conformation is represented as a single point in a low dimensional space formed by a given reaction coordinate. The folding rate through this space is derived from the drift and configurational diffusion coefficients, which take into account for protein topology, the choice of reaction coordinates used, the projection of protein motion into only one dimension, and the integration over other degrees of freedom. Even after these reductions, the results using this effective diffusive approach agree reasonably well with either full simulations or experiments.

7.2.2 Assumptions and simplifications

Although this effective stochastic approach provides powerful predictions, it contains many assumptions and simplifications, which are described as follow:

- The accuracy of the description of effective stochastic dynamics depends on the reaction coordinate(s) used to describe the system. To pick a physical and meaningful reaction coordinate is a non-trivial question. For example, RMSD is a "good" measure of similarity for nearly identical structures, but once the shape of two proteins begins to diverge, RMSD looses its effectiveness. We assume that the reaction coordinate Q can capture the basic feature of the folding process. The physical justification of this approximation involves the fundamental problems of statistical mechanics and is still the subject of many discussions. It is assumed that the progress along the coordinate of Q is rather slow and hence other faster degrees of freedom are integrated out. In recent years it has been shown that Q, the number of native contacts in a Gō model, lends support to the idea that such a reduction might be valid. In a system with more frustration or ruggedness of the folding landscape, the selection of a reaction coordinate has to be done carefully.

- We simplify the folding process by projection into only one coordinate. Of course, in more complex folding system, other reaction coordinates may be necessary for an accurate prediction of folding rates since it is clear that a simple diffusive model cannot be correct for the system with many local traps. For example, if there are parallel or multiple pathways, one might need two or more reaction coordinates to distinguish these pathways.

- An important assumption of the diffusion equation is that the process is Markovian. It has been recognized that, in principle, any isolated physical system can be described as a Markov process by including all microscopic degrees of freedom. The reduction to a much smaller number of macroscopic variables by projection is often used to find those variables that are needed to make the description (approximately) Markovian. The assumption for a Markov process here is equivalent to the "zero" or "negligible" correlation time for the stochastic noise (and equivalently for Q). This assumption can

be checked by calculating the normalized auto-correlation function of Q

$$C_Q(\tau) = \frac{\langle Q(t)Q(t+\tau)\rangle - \langle Q(t)\rangle^2}{\langle Q^2(t)\rangle - \langle Q(t)\rangle^2}. \tag{7.1}$$

If the correlation time is short (on the time scales that are accessible within MD simulations), the evolution of the system into the near future depends only on its properties at the current time (and the current position) and not on any of its history. In this case, a Markov description can be appropriately correct for the folding process as a diffusive motion. For a rugged landscape, the passage over a sharp barrier, where the memory time of the fluctuating force acting on the reaction coordinate is comparable to the time of barrier-crossing, is non-Markovian. The dynamics that determine the folding time are often associated with "high-frequency" motion near sharp barriers [39]. We may consider extension of the Langevin equation (Eq. 6.7) by introducing the non-Markovian generalized Langevin equation (GLE) [39, 45, 98, 97]. Nonetheless, the non-Markovian extension can be considered as a second-order approximation. In other words, our current Markovian description gives a lower bound on the folding time distribution.

In addition, an important questions that needs to be addressed is whether the sampling, especially near the free-energy barrier, is sufficient to provide statistically accurate initial conditions for extracting the local kinetic quantities. As a matter of fact, since protein folding is an extremely high-dimensional process, it becomes non-trivial to make sure sufficient sampling is achieved. We are doing our best to sample the transition regions by implementing the replica exchange technique.

7.3 Future directions

Although extensive research has been focused on domain swapping, in both experimental and computational studies, it is still unclear about the role of domain swapping in protein aggregation and amyloid formation. Direct experiments on high-resolution fibril structures would be ideal, but one may see this goal is not achievable in the near future. So future theoretical and computational studies on these aspects would definitely provide valuable insights on the dynamics of conversion from native states to β-rich aggregated states.

For the folding dynamics, the identification of dominant kinetic paths would be achievable with the diffusive approach with two or more reaction coordinates, if there exist multiple pathways on the folding energy landscape. Of course, it would require much more sampling compared to the one-dimensional reduction. Furthermore, a careful examination of the diffusive model should be done in a more complete model to take into account the assumptions and simplifications used in this dissertation.

In summary, we have developed a new tool to study protein folding kinetics which should be generalizable and applicable to other biologically relevant systems. This dissertation describes the symmetrized Gō model and domain swapping, and shows biological implications by proposing a new structural candidate for the infectious prion protein. In addition, we introduces a newly developed effective stochastic method for folding kinetics, addresses general conclusions and provides directions for future studies. We believe that this dissertation provided and extended theoretical framework to understand protein folding dynamics.

Index

Biased replica exchange molecular dynamics, 28

Conformational change, 13

Diffusion coefficient, 55
Domain swapping, 12
Domain-swapped trimeric prion, 44
Drift velocity, 55
DSTP, 44

Flexibility of a hinge loop, 13
Funneled landscape, 64

Kinetics, 56

Langevin dynamics, 56

Mechanism of domain swapping, 15

non-Markovian, 86

Oligomerization funnel, 38

Protein aggregation, 38
Protein concentration, 33
Protein oligomerization, 25

Replica exchange molecular dynamics, 28
Reweighted histogram analysis method, 29

Second virial coefficient analysis, 31
Self-assembly, 25
Structure-based (Gō-type) model, 64
Symmetrized Gō model, 9

The diffusion equation, 54
Thermodynamics, 56

Bibliography

[1] B. Alberts, A. Johanson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. Garland, New York, 4rd edition, 2002.

[2] M. Andrec, A. K. Felts, E. Gallicchio, and R. M. Levy. Protein folding pathways from replica exchange simulations and a kinetic network model. *Proc. Natl. Acad. Sci. USA*, 102(19):6801–6806, 2005.

[3] M. Bennett, S. Choe, and D. Eisenberg. Domain swapping: entangling alliances between proteins. *Proc. Natl. Acad. Sci. USA*, 91(8):3127–3131, 1994.

[4] M. J. Bennett, M. P. Schlunegger, and D. Eisenberg. 3D Domain swapping: A mechanism for oligomer assembly. *Protein Sci.*, 4(12):2455–2468, 1995.

[5] A. Berezhkovskii and A. Szabo. One-dimensional reaction coordinates for diffusive activated rate processes in many dimensions. *J. Chem. Phys.*, 122(1):014503, 2005.

[6] M. Bergdoll, M.-H. Remy, C. Cagnon, J.-M. Masson, and P. Dumas. Proline-dependent oligomerization with arm exchange. *Structure*, 5(3):391–401, 1997.

[7] F. Blanco, G. Rivas, and L. Serrano. A short linear peptide that folds into a native stable beta-hairpin in aqueous-solution. *Nat. Struct. Biol.*, 1(9):584–590, 1994.

[8] J. Bloustine, Berejnov, V., and S. Fraden. Measurements of protein-protein interactions by size exclusion chromatography. *Biophys. J.*, 85(4):2619–2623, 2003.

[9] P. G. Bolhuis. Transition-path sampling of beta-hairpin folding. *Proc. Natl. Acad. Sci. USA*, 100(21):12129–12134, 2003.

[10] P. G. Bolhuis. Kinetic pathways of beta-hairpin (un)folding in explicit solvent. *Biophys. J.*, 88(1):50–61, 2005.

[11] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes. Funnels, pathways and the energy landscape of protein folding. *Proteins: Struct. Funct. Genet.*, 21:167–195, 1995.

[12] J. D. Bryngelson and P. G. Wolynes. Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. USA*, 84:7524–7528, 1987.

[13] D. Case, T. Darden, I. T.E. Cheatham, C. Simmerling, J. Wang, R. Duke, R. Luo, K. Merz, B. Wang, D. Pearlman, M. Crowley, S. Brozell, V. Tsui, H. Gohlke, J. Mongan, V. Hornak, G. Cui, P. Beroza, C. Schafmeister, J. Caldwell, W. Ross, and P. Kollman. *AMBER 8*. University of California, San Francisco, 2004.

[14] D. A. Case, D. Pearlman, J. W. Caldwell, I. T.E. Cheatam, W. S. Ross, C. L. Simmerling, T. A. Darden, K. M. Merz, R. V. Stanton, A. L. Cheng, J. J. Vincent, M. Crowley, V. Tsui, R. J. Radmer, Y. Duan, J. Pitera, I. Massova, G. L. Seibel, , U. C. Singh, P. Weiner, and P. Kollman. *AMBER 6*. University of California, San Francisco, 1999.

[15] J. Castilla, P. Saa, C. Hetz, and C. Soto. In vitro generation of infectious scrapie prions. *Cell*, 121:195–206, 2005.

[16] L. Chavez, J. Onuchic, and C. Clementi. Quantifying the roughness on the free energy landscape: Entropic bottlenecks and protein folding rates. *J. Am. Chem. Soc.*, 126(27):8426–8432, 2004.

[17] M. S. Cheung, A. E. García, and J. N. Onuchic. Protein folding mediated by solvation: Water expulsion and formation of the hydrophobic core occur after the structural collapse. *Proc. Natl. Acad. Sci. USA*, 99:685–690, 2002.

[18] E. Y. Chi, S. Krishnan, B. S. Kendrick, B. S. Chang, J. F. Carpenter, and T. W. Randolph. Roles of conformational stability and colloidal stability in the aggregation of recombinant human granulocyte colony-stimulating factor. *Protein Sci.*, 12(5):903–913, 2003.

[19] C. Clementi, H. Nymeyer, and J. N. Onuchic. Topological and energetic factors: what determines the structural details of the transition state ensemble and "on route" intermediates for protein folding? an investigation for small globular proteins. *J. Mol. Biol.*, 298:937–953, 2000.

[20] F. E. Cohen. Pathologic conformations of prion proteins. *Annu. Rev. Biochem.*, 67(1):793–819, 1998.

[21] F. E. Cohen. Pathologic conformations of prion proteins. *Annu. Rev. Biochem.*, 67(1):793–819, 1998.

[22] D. L. Cox, R. R. P. Singh, and S. Yang. Prion Disease: Exponential Growth Requires Membrane Binding. *Biophys. J.*, 90(11):L77–79, 2006.

[23] K. A. Dill and H. S. Chan. From levinthal to pathways to funnels: the "new view" of protein folding kinetics. *Nat. Struct. Biol.*, 4:10–19, 1997.

[24] R. Dima and D. Thirumalai. Exploring protein aggregation and self-propagation using lattice models: Phase diagram and kinetics. *Protein Sci.*, 11(5):1036–1049, 2002.

[25] F. Ding, N. V. Dokholyan, S. V. Buldyrev, H. E. Stanley, and E. I. Shakhnovich. Molecular dynamics simulation of the SH3 domain aggregation suggests a generic amyloidogenesis mechanism. *J. Mol. Biol.*, 324(4):851–857, 2002.

[26] A. R. Dinner, T. Lazaridis, and M. Karplus. Understanding beta-hairpin formation. *Proc. Natl. Acad. Sci. USA*, 96:9068–9073, 1999.

[27] C. M. Dobson. Protein folding and misfolding. *Nature*, 426:884–890, 2003.

[28] Y. Duan and P. Kollman. Pathways to a protein-folding intermediate observed in a 1-microsecond simulation in aqueous-solution. *Science*, 282:740–744, 1998.

[29] R. J. Ellis. Macromolecular crowding: an important but neglected aspect of the intracellular environment. *Curr. Opin. Struct. Biol.*, 11:114–119, 2001.

[30] A. M. Ferrenberg and R. H. Swendsen. New monte carlo technique for studying phase transitions. *Phys. Rev. Lett.*, 61:2635–2638, 1988.

[31] H. Frauenfelder, S. G. Sligar, and P. G. Wolynes. The energy landscapes and motions of proteins. *Science*, 254:1598–1603, 1991.

[32] A. E. Garcia and J. N. Onuchic. Folding a protein in a computer: An atomic description of the folding/unfolding of protein A. *Proc. Natl. Acad. Sci. USA*, 100(24):13898–13903, 2003.

[33] A. E. García and K. Y. Sanbonmatsu. Exploring the energy landscape of a β-hairpin in explicit solvent. *Proteins*, 42(3):345–354, 2001.

[34] C. W. Gardiner. *Handbook of Stochastic Methods*. Springer Verlag, New York, 3rd ed. edition, 2004.

[35] A. George and W. Wilson. Predicting protein crystallization from a dilute solution property. *Acta Cryst.*, D50:361–365, 1994.

[36] L. M. Gloss, B. Robert Simler, and C. Robert Matthews. Rough energy landscapes in protein folding: dimeric E. Coli Trp repressor folds through three parallel channels. *J. Mol. Biol.*, 312:1121–1134, 2001.

[37] S. Gosavi, L. L. Chavez, P. A. Jennings, and J. N. Onuchic. Topological frustration and the folding of interleukin-1[beta]. *J. Mol. Biol.*, 357:986–996, 2006.

[38] C. Govaerts, H. Wille, S. B. Prusiner, and F. E. Cohen. Evidence for assembly of prions with left-handed beta-helices into trimers. *Proc. Natl. Acad. Sci. USA*, 101(22):8342–8347, 2004.

[39] R. F. Grote and J. T. Hynes. The stable states picture of chemical reactions. II. Rate constants for condensed and gas phase reaction models. *J. Chem. Phys.*, 73(6):2715–2732, 15 Sept. 1980.

[40] J. Guijarro, C. J. Morton, K. W. Plaxco, I. D. Campbell, and C. M. Dobson. Folding kinetics of the sh3 domain of pi3–kinase by real time nmr combined with optical spectroscopy. *J. Mol. Biol.*, 276:657–667, 1998.

[41] C. Guo, H. Levine, and D. A. Kessler. How does a beta -hairpin fold/unfold? competition between topology and heterogeneity in a solvable model. *Proc. Natl. Acad. Sci. USA*, 97(20):10775–10779, 2000.

[42] J. T. Guo, R. Wetzel, and Y. Wu. Molecular modeling of the core of aβ amyloid fibrils. *Proteins*, 57:357–364, 2004.

[43] Z. Guo and C. L. Brooks. Thermodynamics of protein folding: A statistical mechanical study of a small all β protein. *Biopolymers*, 42:745–757, 1997.

[44] Z. Y. Guo and D. Thirumalai. Kinetics of protein folding—nucleation mechanism, time scales, and pathways. *Biopolymers*, 36(1):83–102, July 1995.

[45] P. Hanggi and F. Mojtabai. Thermally activated escape rate in presence of long-time memory. *Phys. Rev. A*, 26(2):1168–1170, 1982.

[46] J. D. Harper and J. Lansbury P.T. Models of amyloid seeding in alzheimer's disease and scrapie: Mechanistic truths and physiological consequences of the time-dependent solubility of amyloid proteins. *Annu. Rev. Biochem.*, 66:385–407, 1997.

[47] J. G. Ho, Middelberg, A. P.J., Ramage, Paul, Kocher, and H. P. The likelihood of aggregation during protein renaturation can be assessed using the second virial coefficient. *Protein Sci.*, 12(4):708–716, 2003.

[48] G. Hummer and I. G. Kevrekidis. Coarse molecular dynamics of a peptide fragment: Free energy, kinetics, and long-time dynamics computations. *J. Chem. Phys.*, 118(23):10762–10773, 2003.

[49] R. Jaenicke and H. Lilie. Folding and association of oligomeric and multimeric proteins. *Adv. Protein Chem.*, 53:329–362, 2000.

[50] R. Janowski, M. Abrahamson, A. Grubb, and M. Jaskolski. Domain swapping in-truncated human cystatin. *J. Mol. Biol.*, 341(1):151–160, 2004.

[51] R. Janowski, M. Kozak, E. Jankowska, Z. Grzonka, A. Grubb, M. Abrahamson, and M. Jaskolski. Human cystatin, an amyloidogenic protein, dimerizes through three-dimensional domain swapping. *Nat. Struc. Biol.*, 8:316–320, 2001.

[52] J. T. Jarrett and P. T. Lansbury Jr. Amyloid fibril formation requires a chemically discriminating nucleation event: Studies of an amyloidogenic sequence from the bacterial protein osmb. *Biochemistry*, 31:12345–12352, 1992.

[53] W. Jin, O. Kambara, H. Sasakawa, A. Tamura, and S. Takada. De novo design of foldable proteins with smooth folding funnel: automated negative design and experimental verification. *Structure*, 11:581–590, 2003.

[54] G. Kaminski, R. Friesner, J. Tirado-Rives, and W. Jorgensen. Evaluation and reparametrization of the opls-aa force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B*, 105(28):6474–6487, 2001.

[55] J. Karanicolas and C. L. Brooks. The origins of asymmetry in the folding transition states of protein L and protein G. *Protein Sci*, 11(10):2351–2361, 2002.

[56] N. Khazanovich, K. Bateman, M. Chernaia, M. Michalak, and M. James. Crystal structure of the yeast cell-cycle control protein, p13suc1, in a strand-exchanged dimer. *Structure*, 4:299–309, 1996.

[57] K. V. Kishan, M. E. Newcomer, T. H. Rhodes, and S. D. Guilliot. Effect of pH and salt bridges on structural assembly: molecular structures of the monomer and intertwined dimer of the Eps8 SH3 domain. *Protein Sci.*, 10(5):1046–1055, 2001.

[58] K. V. Kishan, G. Scita, W. Wong, P. Di Fiore, and M. Newcomer. The SH3 domain of Eps8 exists as a novel intertwined dimer. *Nat. Struct. Biol.*, 4:739–743, 1997.

[59] A. Kishimoto, K. Hasegawa, H. Suzuki, H. Taguchi, K. Namba, and M. Yoshida. β-helix is a likely core structure of yeast prion sup35 amyloid fibers. *Biochemical and Biophysical Research Communications*, 315:739–745, 2004.

[60] D. K. Klimov and D. Thirumalai. Mechanisms and kinetics of beta-hairpin formation. *Proc. Natl. Acad. USA*, 97:2544–2549, 2000.

[61] K. J. Knaus, M. Morillas, W. Swietnicki, M. Malone, W. K. Surewicz, and V. C. Yee. Crystal structure of the human prion protein reveals a mechanism for oligomerization. *Nat. Struct. Biol.*, 8(9):770–774, 2001.

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

[63] D. I. Kopelevich, A. Z. Panagiotopoulos, and I. G. Kevrekidis. Coarse-grained kinetic computations for rare events: Application to micelle formation. *J. Chem. Phys.*, 122:044908, 2005.

[64] P. J. Kraulis. Molscript: A program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.*, 24:946–950, 1991.

[65] S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comp. Chem.*, 13(8):1011–1021, 1992.

[66] H. A. Lashuel, D. Hartley, B. M. Petre, T. Walz, and P. T. Lansbury. Neurodegenerative disease: amyloid pores from pathogenic mutations. *Nature*, 418:291–291, 2002.

[67] S. Lee and D. Eisenberg. Seeded conversion of recombinant prion protein to a disulfide-bonded oligomer by a reduction-oxidation process. *Nat. Struct. Biol.*, 10(9):725–730, 2003.

[68] G. Legname, I. V. Baskakov, H.-O. B. Nguyen, D. Riesner, F. E. Cohen, S. J. DeArmond, and S. B. Prusiner. Synthetic mammalian prions. *Science*, 305(5684):673–676, 2004.

[69] P. E. Leopold, M. Montal, and J. N. Onuchic. Protein folding funnels: kinetic pathways through compact conformational space. *Proc. Natl. Acad. Sci. USA*, 89:8721–8725, 1992.

[70] S. Lindskog. Structure and mechanism of carbonic anhydrase. *Pharmacol. Ther.*, 74:1–20, 1997.

[71] Y. Liu and D. Eisenberg. 3D domain swapping: As domains continue to swap. *Protein Sci.*, 11(6):1285–1299, 2002.

[72] Y. Liu, G. Gotte, M. Libonati, and D. Eisenberg. A domain-swapped RNase A dimer with implications for amyloid formation. *Nat. Struct. Biol.*, 8:211–214, 2001.

[73] Y. Liu, G. Gotte, M. Libonati, and D. Eisenberg. Structures of the two 3D domain-swapped RNase A trimers. *Protein Sci.*, 11(2):371–380, 2002.

[74] A. Lomakin, D. S. Chung, G. B. Benedek, D. A. Kirschner, and D. B. Teplow. On the nucleation and growth of amyloid β-protein fibrils: Detection of nuclei and quantitation of rate constants. *Proc. Natl. Acad. Sci. USA*, 93:1125–1129, 1996.

[75] A. Lomakin, D. B. Teplow, D. A. Kirschner, and G. B. Benedek. Kinetic theory of fibrillogenesis of amyloid β-protein. *Proc. Natl. Acad. Sci. USA*, 94:7942–7947, 1997.

[76] D. A. Lysek and K. Wuthrich. Prion Protein Interaction with the C-Terminal SH3 Domain of Grb2 Studied Using NMR and Optical Spectroscopy. *Biochemistry*, 43(32):10393–10399, 2004.

[77] B. Y. Ma and R. Nussinov. Molecular dynamics simulations of a beta-hairpin fragment of protein g: Balance between side-chain and backbone forces. *J. Mol. Biol.*, 296:1091–1104, 2000.

[78] N. J. Marianayagam, N. L. Fawzi, and T. Head-Gordon. Protein folding by distributed computing and the denatured state ensemble. *Proc. Natl. Acad. Sci. USA*, 102(46):16684–16689, 2005.

[79] N. J. Marianayagam, M. Sunde, and J. M. Matthews. The power of two: protein dimerization in biology. *Trends Biochem. Sci.*, 29:618–625, 2004.

[80] B. C. H. May, C. Govaerts, S. B. Prusiner, and F. E. Cohen. Prions: so many fibers, so little infectivity. *Trends in Biochemical Sciences*, 29:162–165, 2004.

[81] A. P. Minton. Molecular crowding: Analysis of effects of high concentrations of inert cosolutes on biochemical equilibria and rates in terms of volume exclusion. *Methods Enzymol.*, 295:127–149, 1998.

[82] A. P. Minton. Implications of macromolecular crowding for protein assembly. *Curr. Opinion Struct. Biol.*, 10:34–39, 2000.

[83] A. Mitsutake, Y. Sugita, and Y. Okamoto. Generalized-ensemble algorithms for molecular simulations of biopolymers. *Biopolymers*, 60(2):96–123, 2001.

[84] O. Miyashita, J. N. Onuchic, and P. G. Wolynes. Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins. *Proc. Natl. Acad. Sci. USA*, 100(22):12570–12575, 2003.

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

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

[87] M. E. Newcomer. Protein folding and three-dimensional domain swapping: a strained relationship? *Curr. Opinion Struct. Biol.*, 12(1):48–53, 2002.

[88] M. Nilsson, X. Wang, S. Rodziewicz-Motowidlo, R. Janowski, V. Lindstrom, P. Onnerfjord, G. Westermark, Z. Grzonka, M. Jaskolski, and A. Grubb. Prevention of domain swapping inhibits dimerization and amyloid fibril formation of Cystatin C. *J. Biol. Chem.*, 279(23):24236–24245, 2004.

[89] H. Nymeyer, A. E. García, and J. N. Onuchic. Folding funnels and frustration in off-lattice minimalist models. *Proc. Natl. Acad. Sci. USA*, 95:5921–5928, 1998.

[90] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes. Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem.*, 48:545–600, 1997.

[91] J. N. Onuchic and P. G. Wolynes. Theory of protein folding. *Curr. Opinion Struct. Biol.*, 14(1):70–75, 2004.

[92] V. S. Pande and D. S. Rokhsar. Molecular dynamics simulations of unfolding and refolding of a beta-hairpin fragment of protein G. *Proc. Natl. Acad. Sci. USA*, 96:9062–9067, 1999.

[93] G. A. Papoian, J. Ulander, M. P. Eastwood, Z. Luthey-Schulten, and P. G. Wolynes. From The Cover: Water in protein structure prediction. *Proc. Natl. Acad. Sci. USA*, 101(10):3352–3357, 2004.

[94] G. A. Papoian, J. Ulander, and P. G. Wolynes. Role of water mediated interactions in protein-protein recognition landscapes. *J. Am. Chem. Soc.*, 125(30):9170–9178, 2003.

[95] R. Phillips and S. R. Quake. The biological frontier of physics. *Phys. Today*, 59(5):38–43, May 2006.

[96] S. S. Plotkin and J. N. Onuchic. Understanding protein folding with energy landscape theory. Part I: Basic concepts. *Q. Rev. Biophys.*, 35:111–67, 2002.

[97] S. S. Plotkin and P. G. Wolynes. Non-markovian configurational diffusion and reaction coordinates in protein folding. *Phys. Rev. Lett.*, 80:5015–5018, 1998.

[98] E. Pollak, S. C. Tucker, and B. J. Berne. Variational transition-state theory for reaction rates in dissipative systems. *Phys. Rev. Lett.*, 65(12):1399–1402, 1990.

[99] S. B. Prusiner. *Prion Biology and Diseases*. Cold Spring Harbor Press, Cold Spring Harbor, New York, 2004.

[100] S. B. Prusiner. *Prion Biology and Diseases*. Cold Spring Harbor Press, Cold Spring Harbor, New York, 2004.

[101] V. E. Pye, A. P. Tingey, R. L. Robson, and P. C. E. Moody. The Structure and Mechanism of Serine Acetyltransferase from Escherichia coli. *J. Biol. Chem.*, 279(39):40729–40736, 2004.

[102] F. Rousseau, J. W. H. Schymkowitz, and L. S. Itzhaki. The unfolding story of three-dimensional domain swapping. *Structure*, 11:243–251, 2003.

[103] F. Rousseau, J. W. H. Schymkowitz, H. R. Wilkinson, and L. S. Itzhaki. Three-dimensional domain swapping in p13suc1 occurs in the unfolded state and is controlled by conserved proline residues. *Proc. Natl. Acad. Sci. USA*, 98(10):5596–5601, 2001.

[104] F. Rousseau, J. W. H. Schymkowitz, H. R. Wilkinson, and L. S. Itzhaki. Intermediates control domain swapping during folding of p13suc1. *J. Biol. Chem.*, 279(9):8368–8377, 2004.

[105] K. Sanbonmatsu and A. E. García. Structure of met-enkephalin in explicit aqueous solution using replica exchange molecular dynamics. *Proteins*, 46:225–234, 2002.

[106] A. Sanders, C. Jeremy Craven, L. D. Higgins, S. Giannini, M. J. Conroy, A. M. Hounslow, J. P. Waltho, and R. A. Staniforth. Cystatin forms a tetramer through structural rearrangement of domain-swapped dimers prior to amyloidogenesis. *J. Mol. Biol.*, 336:165–178, 2004.

[107] M. P. Schlunegger, M. J. Bennett, and D. Eisenberg. Oligomer formation by 3D domain swapping: A model for protein assembly and misassembly. *Adv. Protein Chem.*, 50:61–122, 1997.

[108] M. P. Schlunegger, M. J. Bennett, and D. Eisenberg. Oligomer formation by 3D domain swapping: a model for protein assembly and misassembly. *Adv. Protein Chem.*, 50:61–122, 1997.

[109] J. Schymkowitz, F. Rousseau, L. Irvine, and L. L. Itzhaki. The folding pathway of the cell-cycle regulatory protein p13suc1: clues for the mechanism of domain swapping. *Structure*, 8:89–100, 2000.

[110] D. J. Selkoe. Folding proteins in fatal ways. *Nature*, 426:900–904, 2003.

[111] J. E. Shea and C. L. Brooks. From folding surfaces to folding proteins: a review and assessment of simulation studies of protein folding and unfolding. *Annu. Rev. Phys. Chem.*, 52:499–535, 2001.

[112] N. Sinha, C.-J. Tsai, and R. Nussinov. A proposed structural model for amyloid fibril elongation: domain swapping forms an interdigitating beta-structure polymer. *Protein Eng.*, 14(2):93–103, 2001.

[113] V. Sobolev, A. Sorokine, J. Prilusky, E. Abola, and M. Edelman. Automated analysis of interatomic contacts in proteins. *Bioinformatics*, 15:327–332, 1999.

[114] N. D. Socci, J. N. Onuchic, and P. G. Wolynes. Diffusive dynamics of the reaction coordinate for protein folding funnels. *J. Chem. Phys.*, 104:5860–5868, 1996.

[115] M. Souaille and B. Roux. Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations. *Comp. Phys. Comm.*, 135:40–57, 2001.

[116] D. V. D. Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen. Gromacs: Fast, flexible, and free. 26(16):1701–1718, 2005.

[117] R. A. Staniforth, J. L. E. Dean, Q. Zhong, E. Zerovnik, A. R. Clarke, and J. P. Waltho. The major transition state in folding need not involve the immobilization of side chains. *Proc. Natl. Acad. Sci. USA*, 97:5790–5795, 2000.

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

[119] M. Stork, A. Giese, H. A. Kretzschmar, and P. Tavan. Molecular Dynamics Simulations Indicate a Possible Role of Parallel beta-Helices in Seeded Aggregation of Poly-Gln. *Biophys. J.*, 88(4):2442–2451, 2005.

[120] Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, 314(1-2):141–151, 1999.

[121] W. Swope, J. Pitera, F. Suits, M. Pitman, M. Eleftheriou, B. Fitch, R. Germain, A. Rayshubski, T. Ward, Y. Zhestkov, and R. Zhou. Describing protein folding kinetics by molecular dynamics simulations. 2. example applications to alanine dipeptide and a beta-hairpin peptide. *J. Phys. Chem. B*, 108(21):6582–6594, 2004.

[122] J. R. Taylor. *An Introduction to Error Analysis*. University Science Books, Sausalito, CA, second edition, 1997.

[123] F. A. Tezcan, W. M. Findley, B. R. Crane, S. A. Ross, J. G. Lyubovitsky, H. B. Gray, and J. R. Winkler. Using deeply trapped intermediates to map the cytochrome c folding landscape. *Proc. Natl. Acad. Sci. USA*, 99(13):8626–8630, 2002.

[124] V. N. Uversky and A. L. Fink. Conformational constraints for amyloid fibrillation: the importance of being unfolded. *Biochim. Biophys. Acta.*, 1698:131–153, 2004.

[125] N. G. van Kampen. *Stochastic Processes in Physics and Chemistry*. North-Holland, Amsterdam, 2004.

[126] J. Wang, K. Zhang, H. Lu, and E. Wang. Quantifying kinetic paths of protein folding. *Biophys. J.*, 89:1612 – 1620, 2005.

[127] E. Welker, L. D. Raymond, H. A. Scheraga, and B. Caughey. Intramolecular versus intermolecular disulfide bonds in prion proteins. *J. Biol. Chem.*, 277(36):33477–33481, 2002.

[128] R. Wetzel. Ideas of order for amyloid fibril structure. *Structure*, 10:1031–1036, 2002.

[129] H. Wille, M. D. Michelitsch, V. Guenebaut, S. Supattapone, A. Serban, F. E. Cohen, D. A. Agard, and S. B. Prusiner. Structural studies of the scrapie prion protein by electron crystallography. *Proc. Natl. Acad. Sci. USA*, 99(6):3563–3568, 2002.

[130] S. Xu, B. Bevis, and M. F. Arnsdorf. The assembly of amyloidogenic yeast sup35 as assessed by scanning (atomic) force microscopy: an analogy to linear colloidal aggregation? *Biophys. J.*, 81:446–454, 2001.

[131] S. Xu and B. Lin. The mechanism of oxidation-induced low-density lipoprotein aggregation: an analogy to colloidal aggregation and beyond? *Biophys. J.*, 81:2403–2413, 2001.

[132] S. Yang, S. S. Cho, Y. Levy, M. S. Cheung, H. Levine, P. G. Wolynes, and J. N. Onuchic. Domain swapping is a consequence of minimal frustration. *Proc. Natl. Acad. Sci. USA*, 101(38):13786–13791, 2004.

[133] S. Yang, H. Levine, and J. N. Onuchic. Protein oligomerization through domain swapping: Role of inter-molecular interactions and protein concentration. *J. Mol. Biol.*, 352:202–211, 2005.

[134] S. Yang, H. Levine, J. N. Onuchic, and D. L. Cox. Structure of infectious prions: stabilization by domain swapping. *FASEB J.*, 19(13):1778–1782, 2005.

[135] S. Yang, J. N. Onuchic, and H. Levine. Effective stochastic dynamics on a protein folding energy landscape. *J. Chem. Phys.*, in press, 2006.

[136] B. Zagrovic, E. J. Sorin, and V. Pande. beta-hairpin folding simulations in atomistic detail using an implicit solvent model. *J. Mol. Biol.*, 313:151–169, 2001.

[137] R. Zahn, A. Liu, T. Luhrs, R. Riek, C. von Schroetter, F. Lopez Garcia, M. Billeter, L. Calzolai, G. Wider, and K. Wuthrich. NMR solution structure of the human prion protein. *Proc. Natl. Acad. Sci. USA*, 97(1):145–150, 2000.

[138] A. Zdanov, C. Schalk-Hihi, A. Gustchina, M. Tsang, J. Weatherbee, and A. Wlodawer. Crystal structure of interleukin-10 reveals the functional dimer with an unexpected topological similarity to interferon γ. *Structure*, 3:591–601, 1995.

[139] I. Zegers, J. Deswarte, and L. Wyns. Trimeric domain-swapped barnase. *Proc. Natl. Acad. Sci. USA*, 96(3):818–822, 1999.

[140] R. Zhou, B. J. Berne, and R. Germain. The free energy landscape for beta hairpin folding in explicit water. *Proc. Natl. Acad. Sci. USA*, 98(26):14931–14936, 2001.

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