UNIVERSITY OF CALIFORNIA, SAN DIEGO

Theoretical Studies of Biomolecular Self-Assembly near Equilibrium and Far from Equilibrium

A dissertation submitted for the degree Doctor of Philosophy

in

Chemistry

by

Chenghang Zong

Committee in charge:

Professor Peter G. Wolynes, Chair
Professor Partho Ghosh
Professor Katja Lindenberg
Professor J. Andrew McCammon
Professor José N. Onuchic

2007

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

Chair

University of California, San Diego

2007

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENT

I would like to take this opportunity to acknowledge those people who have been support to me through these years. First of all, I am deeply thankful to my parents and my wife for their love and caring, for their encouragement during the difficult times. Thanks to my current and previous fellow groupmates, particularly Tongye Shen, Michael Eastwood, Garegin Papoian, Michael Prentiss, Ting Lu, Patrick Weinkam, Joe Hegler and Sam Cho. I learned a great deal from the collaboration with Tongye Shen. Also I would like to thank my experimental collaborator, Corey Wilson, in the azurin-related work. Finally, I would like to express my appreciation to my advisor Peter Wolynes for his guidance during the course of this research. The financial and facility support is provided through NIH Grant No. CHE 95-30680 and CTBP. The publications which are included in this thesis are as follows: Chapter 2 is based on the material as it appears in Biochemistry 45: 6458-6466 (2006). Chapter 3 is based on the material as it appears in Proc. Natl. Acad. Sci. 104: 3159-3164 (2007). Chapter 4 is based on the material as it appears in J. Am. Chem. Soc. 128: 5168-5176 (2006). Chapter 5 is based on the material as it appears in Physical Biology, 3: 83-92 (2006).

VITA

1977	Born, Xuzhou, Jiangsu, China
2000	B.A. University of Science & Technology of China
2000–2003	M.S. University of California, San Diego
2003–2007	Ph.D., University of California, San Diego

PUBLICATIONS

C. Zong, C. J. Wilson, T. Shen, P. Wittung-Stafshede, S. L. Mayo, P. G. Wolynes, "Establishing the Entatic State in Folding Metallated Pseudomonas aeruginosa Azurin", Proc. Natl. Acad. Sci. 104: 3159-3164 (2007)

C. Zong, G. A. Papoian, J. Ulander, P. G. Wolynes, "The Role of Topology, Nonadditivity and Water Mediated Interactions in Predicting the Structures of alpha/beta Proteins", J. Am. Chem. Soc. 128: 5168-5176 (2006)

C. Zong, T. Lu, T. Shen, P. G. Wolynes, "Nonequilibrium Self-Assembly of Linear Fibers: Microscopic Treatment of Growth, Decay, Catastrophe, and Rescue" Physical Biology, 3: 83-92 (2006)

C. Zong, C. J. Wilson, T. Shen, P. G. Wolynes, P. Wittung-Stafshede, "Phi-Value Analysis of Apo-Azurin Folding: Comparison between Experiment and Theory" Biochemistry 45: 6458-6466 (2006)

M. C. Prentiss, C. Hardin, M. P. Eastwood, **C. Zong**, P. G. Wolynes, "Protein Structure Prediction: The Next Generation", J. Chem. Theory Comput. 2: 705-716 (2005)

T. Shen, **C. Zong**, D. Hamelberg, J. A. McCammon, P. G. Wolynes, "The folding energy landscape and phosphorylation: modeling the conformational switch of the NFAT regulatory domain", The FASEB Journal 19: 1389-1395 (2005)

P. Weinkam, **C. Zong**, P. G. Wolynes, "A funneled energy landscape for cytochrome c directly predicts the sequential folding route inferred from hydrogen exchange experiments", Proc. Natl. Acad. Sci. 102: 12401-12406 (2005)

T. Lu, T. Shen, **C. Zong**, J. Hasty, P. G. Wolynes, "Statistics of cellular signal transduction as a race to the nucleus by multiple random walkers in compartment/phosphorylation space", Proc. Natl. Acad. Sci. 103: 16752-16757 (2006)

FIELDS OF STUDY

Major Field: Physical Chemistry

ABSTRACT OF THE DISSERTATION

Theoretical Studies of Biomolecular Self-Assembly near Equilibrium and Far from Equilibrium

by

Chenghang Zong

Doctor of Philosophy in Chemistry

University of California, San Diego, 2007

Professor Peter G. Wolynes, Chair

The physical sciences have played a pre-eminent role in the advance of biology not only by providing advanced techniques, but also by providing simple concepts for navigating through the complexity of biological systems. One area where simple physics concepts help understanding complicated biological phenomena is the study of protein folding. By presenting the framework of a simple funneled energy landscape, folding is no longer a paradox from the physics point of view. In the following chapters, we present the investigations of both thermodynamics (predicting native structure) and kinetics (predicting ϕ-values) of protein folding on the basis of energy landscape theory.

On the other hand, the discovery of assembly using biological molecular machinery presents new challenges to statistical mechanics combining the aspects of complexity and far-from-nonequilibrium behavior. In the fifth chapter, a study of nonequilibrium dynamic assembly inspired by microtubule dynamics in cell is

presented. The theory provide a general scheme for studying nonequilibrium assembly in one dimension.

Chapter 2 is based on the material as it appears in Biochemistry 45: 6458-6466 (2006). The dissertation author was the primary investigator and author of this paper.

Chapter 3 is based on the material as it appears in Proc. Natl. Acad. Sci. 104: 3159-3164 (2007). The dissertation author was the primary investigator and author of this paper.

Chapter 4 is based on the material as it appears in J. Am. Chem. Soc. 128: 5168-5176 (2006). The dissertation author was the primary investigator and author of this paper.

Chapter 5 is based on the material as it appears in Physical Biology, 3: 83-92 (2006). The dissertation author was the primary investigator and author of this paper.

I

Introduction

The goal to understand protein folding mechanisms and to learn to predict structures from primary sequences, so as to provide a structural basis for functionality analysis has made the study of protein folding one of central battlefields in biophysics. The spontaneous self-assembly of protein into one specific three-dimensional structures within biological relevant time shows the uniqueness of evolved systems. The famous Levinthal paradox [1] raised a key kinetic question about folding since an astronomical number of conformations would seem to be available to proteins. Clearly, a correct thermodynamic picture is required before the kinetic problems can be attacked. In the late eighties, the emergence of the funneled energy landscape picture provides a fundamental statistical mechanical picture for folding proteins [2, 3, 4, 5]. The paradox about folding kinetics is well illuminated by this picture. Much of the emergent understanding of protein folding came from studies of the statistical mechanics of spin glasses and structural glasses [6, 7, 8]. Proteins are essentially a hetero-polymers with evolution selected sequences. The randomness in random heterogeneous polymers will cause conflicts between stabilization of different structural elements. This "frustration" will generate many energy wells with different depths, without forming a single dominant well on the energy landscape. For protein system, however, even though the heterogeneity in sequence still gives some ruggedness on the energy surface, one dominant well appears in the energy landscape. The elimination of the frustration sculpts a smooth slope toward the native state i.e. the landscape is a funnel.

The complex folding processes can be visualized on the surface of a high-dimensional energy landscape. The energy landscape is quantitatively described by the folding temperature T_F and the glass temperature T_G [2]. The folding temperature, T_F, is the temperature at which the free energy of the denatured and native states are equal. The glass temperature, T_G, is the temperature below which the system would be thermodynamically frozen into structurally diverse low energy states. The principle of minimal frustration, proposed by Bryngelson and Wolynes [2], requires that T_F is larger than T_G for a reliable fast folding. Another characteristic temperature T_A also enters [6] which describes the change in kinetics on a rugged energy landscape. At temperature T_A, the low energy states become into metastable states that can trap the system much longer than the time scale of the diffusive chain dynamics.

Theoretical analysis of folding kinetics using the funneled energy landscape and the comparison of those prediction with the experimental results quantitatively tests of this picture. In Chapter 2, we characterize the folding of *Pseudomonas aerugionsa* apo-azurin [9] with a variational free energy functional method, first introduced by Portman, Takada and Wolynes (PTW) [10, 11], based on a perfectly funneled landscape. The model utilizes an analogy between folding protein from random coils to the native state and the transition from the liquid phase to the solid phase. The folding/unfolding at the residue level is described by quantifying the fluctuations of each residue around its native position. This fluctuation amplitude is analogous to the Debye-Waller factor measured in experiments. A high dimensional free energy surface generated by the model is described by an analytical free energy functional having the resolution of residue level. Different position on this high dimensional free energy surface characterizes the structural ensemble with different probability distribution. The folding kinetics is described by the continuous change of the probability distribution of the structural ensemble on the folding routes. A fast numerical search procedure applied to this model allows us to study large systems with more than 100 residues. We compare experimental ϕ-values [12] with our theoretical calculations. The quantitative agreement between theory and experiment provides strong support for the funneled energy landscape concept.

Understanding how a protein's functions emerge along with the folding process challenges both theorists and experimentalists. The simplest testbeds for confronting this issue are provided by electron transfer proteins. The environment provided by the folded protein to the cofactor tunes the metal's electron transport capabilities as envisioned in the entatic hypothesis [13]. The entatic hypothesis states that through the polypeptide's folding induced rigidity, the protein fails to provide the expected geometry of ligating groups that would occur with freely mobile ligands in solution, thereby tuning the ligands redox characteristics. To see how the entatic state is achieved one must study how the folding landscape affects and in turn is affected by the metal. In Chapter 3, I develop the free energy functional method further in order to explicitly model how the coordination of the metal (which results in a so-called entatic or rack-induced state) modifies the folding the *Pseudomonas aerugionsa* azurin metallated with a zinc ion [14]. The free energy functional approach directly yields the proper non-linear free energy variations with temperature change for zinc-form azurin. The results agree quite well with corresponding laboratory experiments [15]. Furthermore, the routes found using the modified free energy functional provided a sufficient level of details to explicitly show how the ligated entatic state is formed during the process of folding of the backbone.

The minimal frustration principle and the funnel concept suggest that the main features of folding process can be predicted by correctly modeling the the stabilization energies of structural elements and the entropic costs of bringing the peptide into the scaffold. Associative Memory Hamiltonian (AMH) has been introduced by Friedrichs and Wolynes [16] and developed through the cumulative efforts of many coworkers in the Wolynes group [17, 18, 19, 20, 21, 22, 23, 24]. In Chapter 4, the details of the most complete version of this predictive model and specifically progress in developing this model for structure prediction for α/β proteins are described. A transferable optimization scheme based on the minimal frustration principle (increasing the ratio of T_F over T_G) is applied for the parameterization of the Hamiltonian.

The folding of α/β proteins involves most of the commonly known structural and dynamic complexities of the protein energy landscapes. Thus, the interplay among different structural components, taking into account cooperative interactions, is important in determining the success of protein structure prediction. We present further developments of our knowledge-based force field for α/β proteins and more realistic modeling of many-body interactions governing the folding of β-sheets. The model's innovations highlight both specific topological characteristics of secondary structures and the generic nonadditive interactions that are mediated by water. The studies also show how a coarse biasing of the protein morphology can be used to understand the role of heterogeneity in protein collapse.

After structural self-assembly of proteins, the next step for building biological functional molecular machines is the assembly of modular protein units. These assembly are essential cellular processes for functions and many of these processes consume energy. These systems will exist at far-from-equilibrium conditions when the energy is stored or released. How to treat the nonequilibrium nature of such large fluctuations is still one of difficult problems in nonequilibrium statistical mechanics. In Chapter 5, I describe a scheme derived from nonequilibrium variational principle to treat one-dimensional fiber assembly [25].

In cells, many of the large structures are constructed from fibers: actin, microtubules and intermediate filaments [27]. The fibers self-assemble from individual proteins in a far-from-equilibrium fashion. Nonequilibrium self-assembly results in a highly dynamic process at the subcellular level that can be regulated and tuned to carry out many of biological functions of the cell: growth, division and locomotion [27, 28, 29, 30, 31].

I construct and analyze a nonequilibrium model of the dynamic end of a biological fiber that possesses site-resolved resolution. The nonequilibrium variational principle we used is a classical version of Rayleigh-Ritz variational method in quantum mechanics [26]. The independent left and right trial states make the method suitable to treat dynamics described by a non-Hermitian operators. The steady states of this nonequilibrium system are solved using the variational method. The results are compared to exact numerical solutions for systems with modest size. As test, I apply this method also to model microtubule systems [32, 33, 34, 35, 36, 37, 38]. Using an effective reaction coordinate, we construct an effective potential from the steady state distribution. The stochastic transitions of the system can be analyzed in this representation. This picture provide a new perspective of the dynamic instability going beyond the usual picture based on a simple two-state switch. Predictions for macroscopic catastrophe, rescue, and dynamic instability in the steady states are made. We find that the length of the cap of the microtubule is small as argued in some experimental studies. Furthermore, our system can be looked as a typical automata system following the designated dynamics rules. The variational method provides new approaches for studying the dynamics of such automata.

II

Variational Free Energy Functional Method in Sculpting the Folding Energy Landscape

II.A Detailed Characterization of Free Energy Landscape

While the conceptual bottlenecks in understanding protein folding have been overcome in the framework of energy landscape, the high-dimensional energy surface presents major difficulties in quantitatively charactering both the thermodynamics and kinetics of protein folding. In this Chapter, I present a free energy functional that allows us to efficiently characterize the free energy profile analytically and following the folding process on the landscape. This variational scheme was first introduced by Portman, Takada and Wolynes [10, 11].

It is clear that the large number of degrees of freedom in protein allows the molecule to explore a diverse set of kinetic pathways. While superficially

TST (transition state theory) can be adapted to study the kinetics of protein folding, the large number of degrees of freedom and broader distribution of transient states in the protein folding make the detailed characterization of the complete free energy profile necessary. Sampling of the energy landscape by simulations provides one route to such a free energy profile. Yet, for studying many mutants and their folding at different concentrations of denaturant, performing simulations case by case becomes cumbersome. Furthermore, some mutants only make very delicate changes to the energy landscape, which can be hard to discriminate from the nearly inevitable statistical errors in sampling. Compared to most simulation schemes, the numerical calculations based on the variational principle offer improved efficiency and accuracy. Calibrating the variational calculation avoids the statistical disadvantages confronted in simulations. Small changes of energy landscape can be studied efficiently. With the variational method, multiple pathways can be clearly discerned. The high resolution of the variational calculation calibration also enables the discrimination of multiple transient states in one pathway as well as the detection of the shifting of transition states often observed in experiments.

II.B Variational Free Energy Functional

The PTW variational scheme starts with a simple model Hamiltonian of the protein system. Both the chain model of the backbone and contact interactions are included in the Hamiltonian expressed as $H = H_{chain} + H_{int}$. The first term $H_{chain} = \frac{3}{2a^2}\sum_{ij} r_i \Gamma_{ij} r_j + \frac{3}{2a^2} B \sum_i {r_i}^2$ models a stiff chain. Γ_{ij} defines the bond correlation matrix. The relation between Γ and the standard Rouse matrix R is given as follows: $\Gamma = (1-g)/(1+g) * R + g/(1-g^2) * R^2 - g^2/(1-g^2) * \Delta$. The parameter

g is the cosine value of the fixed free rotating angle θ between adjacent bonds and Δ accounts for the ends of the chain. The second term $H_{int} = \sum_{ij} \epsilon_{ij} u(r_{ij})$ represents the contact interactions, which are modeled by a pairwise potential $u(r_{ij})$ with an interaction strength coefficient ϵ_{ij}. The interactions strength coefficients based on the consensus analysis are used in the model. Here, we parameterize the ϵ_{ij} according to Miyazawa-Jernigan contact energies. The potential $u(r_{ij})$ is modeled as a sum of three Gaussian potentials representing short, intermediate and long range part, $u(r) = \sum_{k=s,i,l} \gamma_k \exp[-\frac{3}{2a^2}\alpha_k r^2]$. Three interaction components are parameterized to present a potential with the native distance at the minimum of the well. To construct the free energy surface using a variational procedure, a simple reference Hamiltonian is chosen as $H_0 = H_{chain} + \frac{3}{2a^2}\sum_i C_i(r_i - r_i^N)^2$. The parameter, C_i, describes the fluctuations of the ith residue. With a reasonable choice of the reference Hamiltonian, we can calculate the variational free energy as follows.

$$F[C] = -k_B T \log Z_0 + \langle H - H_0 \rangle_0$$

Z_0 is the partition function of the reference Hamiltonian and $\langle ... \rangle_0$ denotes the average with respect to the reference Hamiltonian. Using this relation, the calculation of the energy and the entropy is straightforward.

II.C ϕ-value Analysis and Chevron Plot

The ϕ-values are defined as the ratio $\Delta\Delta G^{\ddagger}/\Delta\Delta G$ (where $\Delta\Delta G^{\ddagger} = \Delta G^{\ddagger}_{mut} - \Delta G_{wt}$ and $\Delta\Delta G = \Delta G_{mut} - \Delta G_{wt}$; mut, mutant; wt, wild-type form of protein) if folding is a two-state process with a sharp transition state. In *in vitro* experiments of protein folding, the directly measured data are presented as relaxation curves corresponding to the rate of equilibration of folding and unfolding events at the particular condition (in terms of temperature, chemical-denaturant concentration and/or pH): $k_{obs} = k_u + k_f$. Theoretical relaxation curves that correspond to experimental k_{obs} are extracted from the folding pathways predicted by the PTW method, as we describe in detail below. The free-energy changes corresponding to the folding barrier, $\Delta G^{\ddagger}$, and the protein stability, ΔG, are then estimated via linear extrapolations of the logarithms of the observed relaxation rates to a common condition (such as zero denaturant concentration or a specific temperature). In our theoretical calculations, we altered the balance of folding and unfolding by changing the temperature. The Arrhenius rate coefficients k_u and k_f are expressed as follows, with $\Delta G^{\ddagger}_{u,f} = -(G^{\ddagger} - G_{F,U})$ and pre-factor A:

$$k_{u,f} = A\exp(\frac{\Delta G^{\ddagger}_{u,f}}{k_B T})$$

A contact map (input in the PTW variational calculations) for wild-type *P. aeruginosa* apo-azurin was constructed based on the distances between all heavy atoms in the side chains (pdb file 1E65). The contacts are classified into side-chain and backbone contacts according to the distances between side-chain atoms and their angular orientations with respect to each other. The energy unit ϵ_0 of Miyazawa-Jernigan contact energies is converted to kcal/mol with the estimated melting temperature [39] (using the room temperature as the melting temperature). Using this scaling, we calculated a relative folding temperature of 1.91 for apo-azurin. To match the experimentally determined thermodynamic stability of *P. aeruginosa* apo-azurin [12], the stability (G) at 298K was set to $10k_BT$ in our model. With this, the relative folding temperature of 1.91 corresponds to a value of 320K. Notably, this is only somewhat lower than the melting temperature reported in *in vitro* unfolding experiments of *P. aeruginosa* apo-azurin [40].

The free-energy profiles predicted from the PTW variational method for folding of wild-type apo-azurin appear two-state-like with one broad barrier; the profiles at the folding temperature (i.e., at T = 1.91) are shown in Figure II.1 (left panel). The finer details of the profiles arise from the irregular compensation of entropy loss by free energy gain. We identify two folding paths for apo-azurin at the folding temperature; the path with the lower barrier is the pathway we will focus on (since it is most probable). Although many saddle points are found in both pathways, there are no distinct or highly populated intermediate states found on either pathway. In the present treatment, therefore, we denote the point with the highest free energy as the folding-transition state and the folding rates are calculated as the relaxation times to cross this barrier. A structural interpretation of the folding-transition state, that is, the TSE for folding, can be made by examining the mean-square deviation (MSD) of each residue as predicted by the PTW variational algorithm. Here, the MSD of residue i is defined as how much residue i fluctuates around its mean position in the probability distribution belonging to the TSE. The MSD as a function of residue number is shown in Figure II.1 (right panel). for the globular (unfolded), native, and transition states on the two folding routes for wild-type apo-azurin. Similar information is visualized in the isodensity plots for the two TSEs (see Figure II.2). The construction of this type of plot has been described previously [41]. In both transition states, a nativelike structure appears in the core around residues 30, 50, 85, 95, and 110 with some additional nativelike structure proximal to the N-terminus in the pathway with the higher barrier.

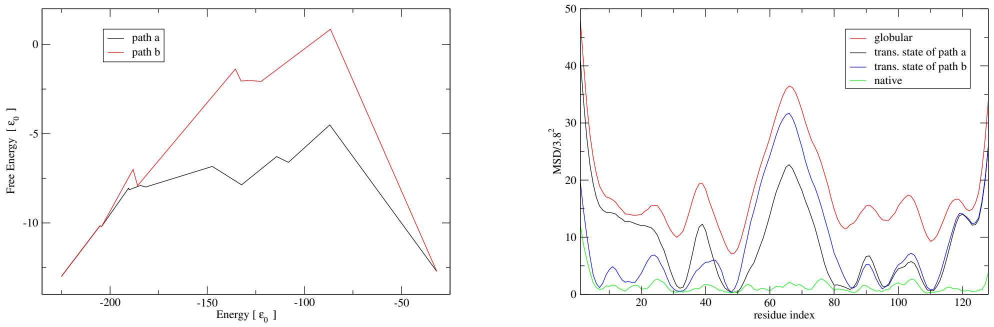

Figure II.1: Left panel: Variational free energy versus stabilization energy of the two folding pathways [high-energy barrier (red) and low-energy barrier (black)] for apo-azurin at T = 1.91 (i.e., at the folding temperature). The point with the highest free energy is denoted as the folding-transition state in each path. Right panel: MSD as predicted by the variational algorithm for different states on wild-type apo-azurin's folding pathways at T = 1.91. The MSD as a function of residue number is shown for the globular/unfolded state, the native state, as well as the high-barrier TSE and the low-barrier TSE of the two folding pathways identified in the left panel.

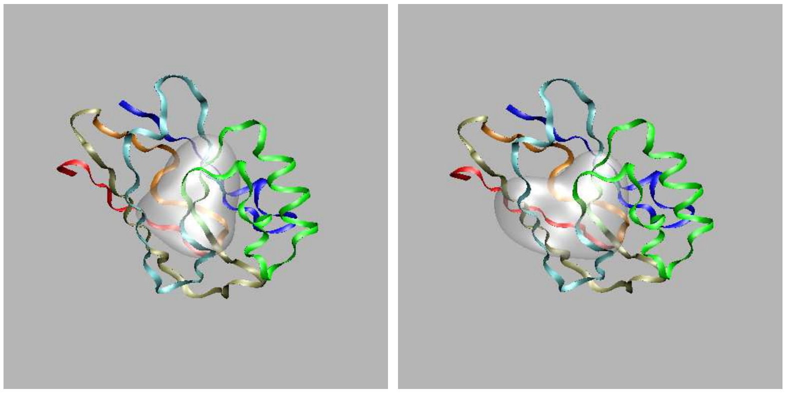

Figure II.2: The isodensity surface ($\rho = 0.005$) for the two TSE for apo-azurin folding at T = 1.91. The left model corresponds to the low-energy barrier TSE (i.e., the focus of this study), and the right model corresponds to the high-energy barrier TSE

The (un)folding barrier linearly depends on the temperature change at the first order approximation.

$$\Delta G^{\ddagger}_{u,f}(T) = \Delta G^{\ddagger}_{u,f}(T_0) + m_{u,f} \times (T - T_0)$$

Then $k_{u,f}(T)$ is given as follows.

$$\begin{aligned} k_{u,f}(T) &= A \exp\left[-\Delta G^{\ddagger}_{u,f}(T)/T\right] && \text{(II.1)} \\ &= A \exp\left[-(\Delta G^{\ddagger}_{u,f}(T_0) - m_{u,f} \times T_0)/T + m_{u,f}\right] && \text{(II.2)} \end{aligned}$$

Now the logarithms of the observed rate coefficients, $\log k_{obs}$, can be expressed as,

$$\log k_{obs}(T) = \log\left[\exp(a + b \times T^{-1}) + \exp(c + d \times T^{-1})\right] + \log A$$

with $a = m_u$, $b = -\Delta G^{\ddagger}_u(T_0) + m_u \times T_0$, $c = m_f$, and $d = -\Delta G^{\ddagger}_f(T_0) + m_f \times T_0$. The relaxation data for wild-type apo-azurin were used to prepare a theoretical chevron plot (i.e., a plot of the $\log k_{obs}(T)$ versus T data, Figure II.3) for wild-type apo-azurin that was then fitted with the above expression to assign values to the a, b, c, and d parameters. In the fitting procedure, parameters a and c will give opposite signs, corresponding to unfolding and folding, respectively.

To derive ϕ-values for specific positions, 16 apo-azurin variants with point-mutations were created based on the contact map of wild-type apo-azurin (Ile7Ala, Ile20Ala, Val22Ala, Val31Ala, Leu33Ala, His46Gly, Trp48Ala, Leu50Ala, Val60Gly, Ile81Ala, Val95Ala, Phe97Ala, Tyr108Ala, Phe110Ala, His117Gly, and Leu125Ala). In the theoretical description of the point-mutated variants, only the side-chain contacts of the residue in question are eliminated in each case. Backbone contacts, that is, the hydrogen bonds in secondary structures, are not altered in the mutated variants. The PTW variational calculation was performed on each azurin variant to ultimately obtain a set of chevron plots (Figure II.4). These were then fitted with the above expression to define the parameters a, b, c, and d for each variant. From the obtained values, the folding barrier ($\Delta G^{\ddagger}$) and the protein stability, $\Delta G = -(\Delta G^{\ddagger}_f - \Delta G^{\ddagger}_u)$, of each system were extracted. Next, the $\Delta G^{\ddagger}$

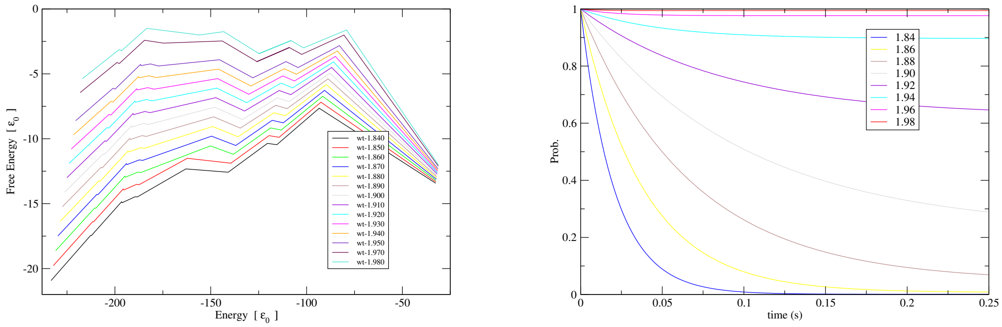

Figure II.3: Left Panel: Folding routes (variational free energy versus energy) at different temperatures for wild-type apo-azurin. Only the low-energy-barrier pathway is shown. The lines correspond to temperatures from 1.98 to 1.84 with 0.01 increments from top to bottom. Right Panel: Relaxation curves for wild-type apo-azurin of the probability to remain in the globular/unfolded state at various temperatures (with a zero probability value meaning full conversion to the folded state). The curves correspond to relaxation data at temperatures from 1.98 to 1.84 with 0.02 increments from the top to the bottom.

and ΔG values for wild-type and variants of apo-azurin were combined to yield the ϕ-value for each mutated position. We summarize the theoretically obtained stability and ϕ-value results in Table II.1 .

To assess the accuracy of the theoretical results as compared to *in vitro* experiments, we also collected experimental equilibrium- and kinetic-folding data on all the apo-azurin variants. In Figure II.5, we show the resulting chevron plots for folding and unfolding kinetics of the 16 apo-azurin variants. In accord with two-state kinetics, the constituting arms of a given chevron are linear with neither protein concentration dependence nor missing amplitudes.

The experimentally and theoretically derived ϕ-values are compared in Figure II.6 (Left Panel); the correlation coefficient between the two sets of data is 0.80. The ϕ-value of Leu125Ala has the largest deviation between experiment and theory. However, this mutation only shows a 3 kJ change in the experimental $\Delta G_U(H_2O)$ value, which may make the experimental ϕ-value calculation somewhat unreliable [42]. If we exclude the point for Leu125Ala in Figure II.6 (Left

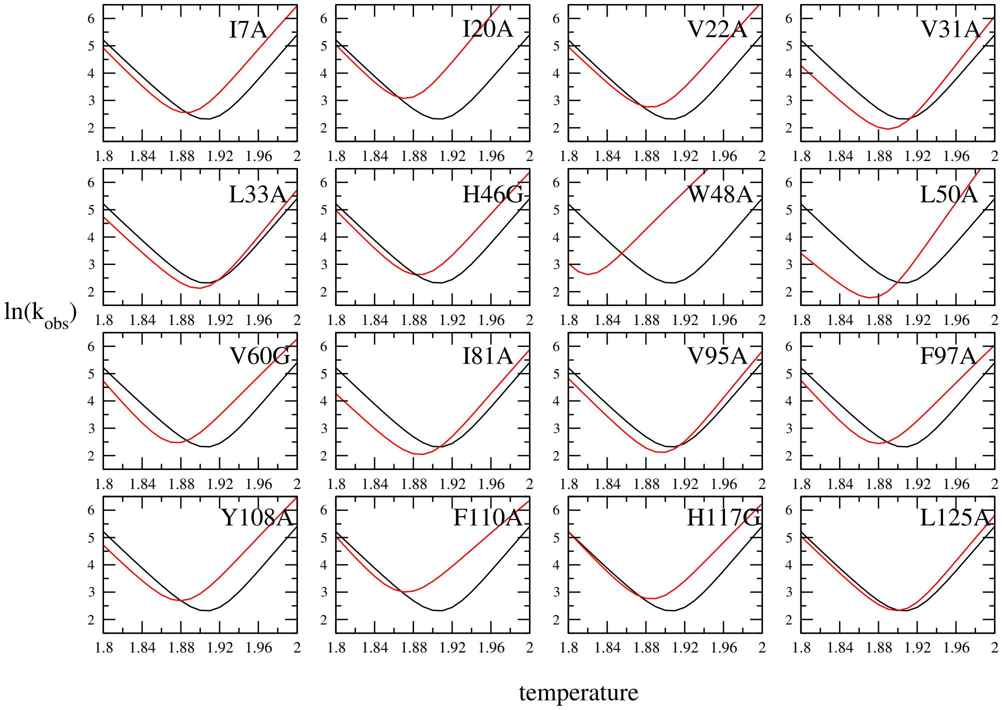

Figure II.4: Theoretical chevron plots ($\ln k_{obs}$ versus temperature) for the 16 apo-azurin variants studied herein. In each graph, the mutant data (red curve) is overlaid with the wild-type data (black curve).

Table II.1: mutation table and ϕ-values

wt and mutants	eliminated contacts	ΔG (k_BT)	ϕ-value
wt	-	8.82	-
I7A	15,16,17,31,33	7.15	0.18
I20A	29,48	6.28	0.07
V22A	29,99,125,127	6.95	0.14
V31A	48	7.91	1.04
L33A	84,87	8.40	1.14
H46G	9,10,35,46	6.96	0.13
W48A	20,31,84,95,110	1.56	0.33
L50A	81,97	7.09	0.67
V60G	111,113,118	5.99	0.17
I81A	97,101,108	7.19	0.58
V95A	48	8.34	0.81
F97A	29,50	6.25	0.18
Y108A	81,102,103,125	6.26	0.19
F110A	15,17,18	5.20	0.05
H117G	13,42,112	6.76	0.00
L125A	50	8.09	0.20

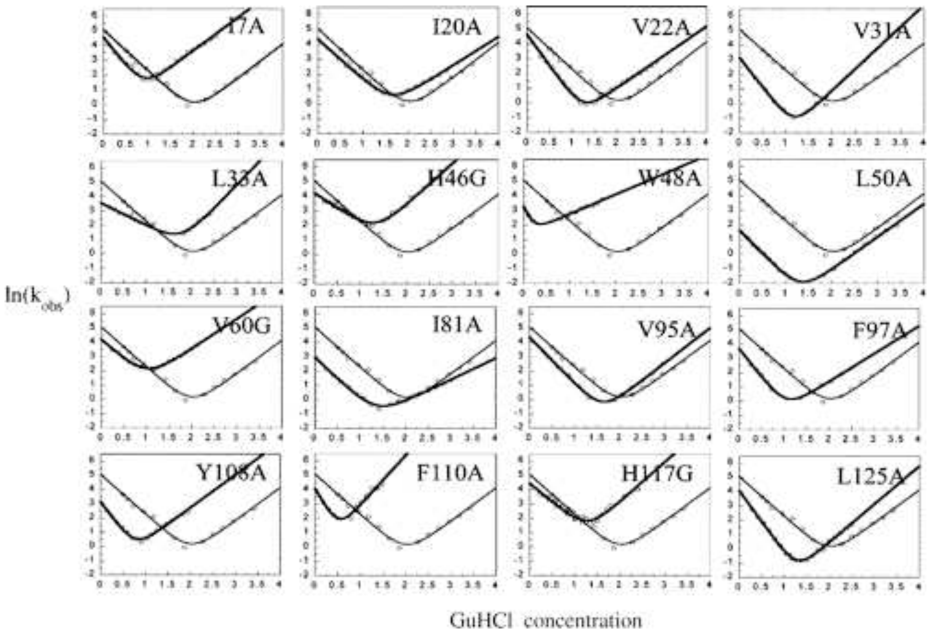

Figure II.5: Experimental chevron plots ($\ln k_{obs}$ versus GuHCl concentration) for the sixteen apo-azurin variants. In each graph, the mutant data (thick curve) is overlaid with the wild-type data (thin curve). The arrangement of panels is the same as in Figure II.4.

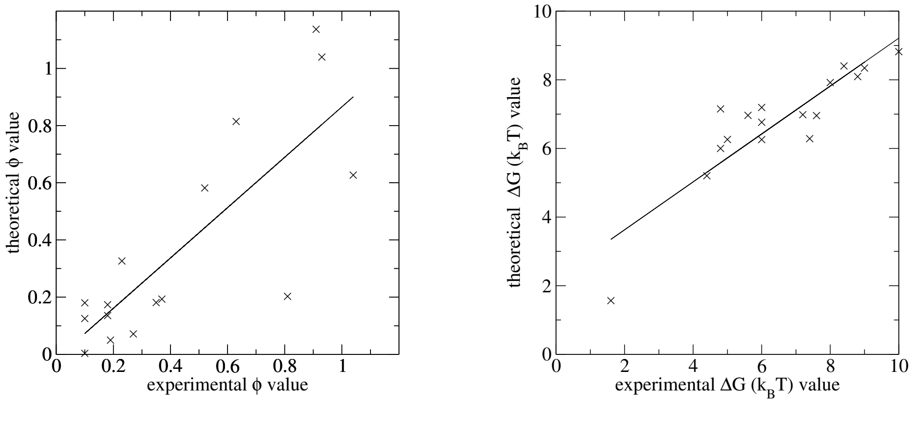

Figure II.6: Left Panel: Correlation between calculated and experimental ϕ-values at room temperature. Right Panel: Correlation between calculated and experimental stability values at room temperature.

Panel), the correlation coefficient increases to 0.90. The correlation between theoretical, $\Delta G(298K)$, and experimental, $\Delta G(H_2O)$, stability values for wild-type and mutant apo-azurins is also excellent; is 0.88 (Figure II.6, Right Panel). Taken together, the comparisons show that using the PTW variational method to calculate folding dynamics provides chevron plots and, thus, ϕ-values that are in fine agreement with data from *in vitro* protein-folding experiments. Furthermore, from the ϕ-values for the 16 positions, it emerges that apo-azurin has a localized TSE with almost nativelike interactions around Val31, Val33, Leu50, Ile81, and Val95 bringing parts of β-strands 3-6 together. The remaining positions, which cover all other secondary-structure elements in azurin (i.e., β-strands 1, 2, 7, and 8, as well as the α-helix), are not structured in apo-azurin's TSE for folding. This description of the TSE for folding of apo-azurin is in good agreement with the isodensity plot for the low-energy barrier TSE (Figure II.2, left).

Here, we have compared ϕ-values derived from theoretical experiments with ϕ-values calculated from *in vitro* chemical-denaturant measurements. Although temperature-jump experiments may seem more appropriate in providing experimental data that correspond to the theoretical analysis, temperature changes can induce nontrivial dynamic solvent effects [43, 44] , which can complicate straightforward analysis based on equilibrium free-energy landscape theory. Importantly, unfolding of apo-azurin when induced by temperature- and chemical-denaturant perturbations results in similar unfolded states with respect to far-UV CD and fluorescence characteristics (data not shown); moreover, both methods of perturbation are reversible and correspond to two-state transitions [45, 46]. From a technical perspective, chemical-denaturant jumps are less complicated to execute as compared to temperature jumps when complete chevron plots are desired.

II.D Conclusions

We used the PTW variational method to investigate the folding TSE for the β-sandwich protein *P. aeruginosa* azurin. We also prepared all azurin variants in the lab and tested their folding behavior *in vitro* via kinetic measurements. We found excellent agreement between theoretically and experimentally determined ϕ-values: apo-azurin's transition state is fixed with a set of nativelike interactions involving core residues from strands in both β-sheets. Detailed comparisons of theoretical and experimental data demonstrate that fine-tuning is needed in the theoretical description of point-mutated variants. Free-energy functional methods, such as the PTW scheme, allow one to readily calculate chevron plots/ϕ-values and, therefore, compare theoretical results directly with experimental measurements. For apo-azurin, this direct comparison shows that theory and experiment agree in a quantitative fashion. Combined theoretical/experimental studies to investigate how the presence of zinc induces a switch from a fixed to a moving TSE in azurin are in the following chapter.

II.E Acknowledgments

This chapter is based on the material as it appears in Biochemistry 45: 6458-6466 (2006), C. Zong; C. J. Wilson; T. Shen; P. G. Wolynes; P. Wittung-

Stafshede. The dissertation author was the primary investigator and author of this paper.

III

Establishing the Entatic State in Folding Metallated *Pseudomonas aeruginosa* Azurin

III.A Introduction of Entatic Hypothesis

The entatic state occurs in proteins when a group, metal or non-metal is forced into an unusual, energetically strained geometric or electronic state (rack-induced state) [47, 13, 48, 49]. Through the polypeptide's folding induced rigidity, the protein fails to provide the expected geometry of ligating groups that would occur with freely mobile ligands in solution, thereby tuning the ligands redox characteristics. In metalloproteins, the metal ions are typically bound to the protein through one or more lone pair donors — endogenous biological ligands (e.g., the imidazole moiety of histidine, the carbonyl oxygen of the main-chain or the side chain of an asparagine residue). In several cases the ligands are arranged such that an optimal geometry is precluded [47, 13, 48, 49]. The resulting entatic state in a given metalloprotein is determined by the entire rigid protein scaffold in concert with the hydrogen bonding network proximal to the coordination sphere [50, 51]. The particular geometry of the rack-induced state influences the electronic structure of the metal site. Moreover, the resulting forced electronic structure, at least in certain cases, becomes essential for the protein's biochemical function in electron transport [52]. We should remember the entatic hypothesis is in some respects still controversial. Results from some quantum calculations have suggested that the geometry of metal-ligand complexes identified as being rack-induced are not necessarily highly strained [53], while, other theoretical studies suggest that the rigidity of the protein may in fact be much more significant than initially thought [54].

Cupredoxins, a family of electron transfer metalloproteins, are believed to adopt such a rack-induced state by way of a distorted tetrahedral (type I) copper site. The geometry of the ligand set provided by the protein in this so-called entatic state is neither optimal for Cu^{1+} nor Cu^{2+}. As a result, redox interconversion does not result in dramatic structural changes. Consequently, the overall reorganization energy for the electron transfer, including the inner coordination sphere, of the type I copper site is relatively small [55, 56, 57], speeding the electron transfer process. The architecture of a typical type I copper sites involves four canonical ligands; specifically, a strongly coordinating thiolate of a cystine residue, the imidazole nitrogens of two histidines, and a weakly coordinating thioether sulfur on a methionine residue.

Pseudomonas aeruginosa azurin is a small (128 amino acid) cupredoxin (i.e., a blue copper protein) composed of eight β-strands arranged in a double-wound Greek key topology, which forms a rigid β-barrel [58]. Interestingly, the redox-active copper is coordinated to the protein via five ligands instead of four. In addition to the four canonical ligands (i.e., H46, C112, H117, and M121-to a lesser extent), a secondary weak-axial ligand — the main-chain carbonyl of G45 — completes the active site, resulting in a trigonal bipyramidal geometry rather than the canonical distorted tetrahedral arrangement often found.

Upon unfolding of metallated azurin, the copper remains bound to the denatured polypeptide in a trigonal coordination composed of the native ligands C112, H117, and possibly M121 [59]. In the denatured state the slow irreversible redox coupling between the C112 thiol and Cu^{2+} promotes sulfur oxidation. As a result, Cu^{2+} metallated azurin does not fold reversibly in the laboratory [60]; thus, a thorough investigation of how the metal center influences the protein's stability and folding dynamics is very difficult. Fortunately, Zn^{2+} can be exchanged for copper without significant change to the rigid structure of azurin [58, 61]. Because zinc is essentially redox inactive, a more detailed assessment of the metal's role in the folding landscape can be performed for this system. Moreover, the main properties of the entatic state at the least from the geometrical point of view still hold; the first coordination sphere as well as the intricate hydrogen bonding network that constitutes the second coordination sphere is largely unperturbed by the substitution and like copper its geometry is not optimal for zinc coordination.

Experiments on Zn^{2+} metallated azurin revealed a significant non-linear free energy relationship for the kinetics under both folding and unfolding conditions. The curvature in the so-called "Chevron plot" appears to result from transition-state movement. Recently, the protein engineering method (i.e., ϕ-value analysis) pioneered by Fersht [62, 63] was used to obtain snapshots of zinc-substituted azurin's dynamic folding nucleus with residue specific resolution. Analysis of several point mutated variants (typically involving hydrophobic-to-alanine transformation) of zinc metallated azurin, covering all of the secondary structure elements, revealed that the folding nucleus is spatially delocalized and gradually becomes more native-like around an epicenter situated on residue L50 [15]. The dramatic difference in kinetic folding behavior between apo-azurin, which has a fixed and rather polarized folding nucleus [12, 9], and the malleability exhibited by the zinc-form was rationalized in terms of changes on a common broad activation barrier. The present chapter studies the folding landscape for Zn^{2+} metallated azurin using a free energy functional scheme appropriately modified to treat metal coordination events to shed light on how the dynamic folding nucleus is involved in forming the so-called entatic state.

III.B The Theoretical Foundation

III.B.1 The Basis of the Variational Approach

Our study of the dynamic folding nucleus and free energy profile of zinc-metallated azurin uses a variational approach that explicitly incorporates the metal coordination reactions. The current approach starts with a functional developed by Portman, Takada, and Wolynes (PTW) [10, 11]. The PTW variational method is based on a coarse-grain free energy functional that only considers native contacts consistent with the dominance of native interactions required by the principle of minimal frustration [64, 65, 5]. The Hamiltonian for the polymeric assembly, as described in the previous chapter, is comprised of two terms, a residue centered contact interaction H_{int}, and a backbone scaffold term H_{chain} modeling a collapsed stiff chain of monomers each representing a residue in the protein's primary sequence (III.1).

$$\begin{aligned}
H &= H_{chain} + H_{int} \\
H_{chain} &= \frac{3}{2a^2}\sum_{ij} r_i \Gamma_{ij} r_j + \frac{3}{2a^2} B \sum_i {r_i}^2 \\
H_{int} &= \sum_{<ij>} \epsilon_{ij} u(r_{ij}) \\
u(r) &= \sum_{k=s,i,l} \gamma_k \exp[-\frac{3}{2a^2}\alpha_k r^2] \qquad \text{(III.1)}
\end{aligned}$$

Here a is a microscopic length taken to be the mean square distance between adjacent monomers in the chain, B is an energy term conjugate to the radius of gyration of the chain, r_i is the position of monomer i in the polymer chain, and the correlations between any two C_α positions are given by Γ^{-1}(25). The second term H_{int} in the energy functional contains a pairwise potential $u(r_{ij})$ with an interaction strength coefficient ϵ_{ij}. Again, we parameterized the ϵ_{ij} coefficients using Miyazawa-Jernigan contact energies. The interaction potential $u(r_{ij})$ is a sum of three Gaussian potentials representing short (s), intermediate (i) and long (l) range parts, where $\alpha_l < \alpha_i < \alpha_s$ are the long-, intermediate-, and short-range widths, respectively. The long-range term is attractive, while the intermediate- and short-range terms are repulsive (i.e., $r_l < 0$; $r_i > 0$; $r_s > 0$, respectively).

III.B.2 Modeling the Coordination Reaction

To model the metallated form of azurin, the cofactor was explicitly incorporated into the functional and the corresponding coordination event during the folding process was considered. First, the appropriate metal-ligand interactions were simply treated as contacting positions carrying electrostatic interactions during the folding event and as a separate step these ligands are allowed to undergo coordination reactions to the Zn^{2+}, which confer the appropriate binding stability. Separating, these two steps resembles the differentiation between forming contact pairs and inner-shell reorganization in inorganic solution reactions. For some metals there may be barriers for the coordination step, but these are small for Zn^{2+}. To describe the ligand-cofactor interactions, the C112 and H117 ligands were modeled as permanent constituents of the backbone connections, while cofactor interactions with residues G45 and H46 were allowed to form or break during the folding and unfolding process. The methionine at position 121 was classified as a weakly interacting ligand in the folded copper-metallated protein with an interaction distance of 3.2Å, whereas zinc-substituted azurin's interaction distance was approximated at 3.3Å [58, 61, 66, 67]. Considering that the resolution provided by x-ray crystallography for the Cu^{2+} and Zn^{2+} metallated azurin structures is presently limited to 1.5, the thioether's sulfur interactions with the cofactor are geometrically indistinguishable in practice. Moreover, the role of M121 as a coordinating residue in the unfolded state is still not settled [68, 59, 15, 69]. Accordingly, this particular residue was not explicitly modeled as a coordinating residue, only as a contacting residue. Furthermore, the limited resolution provided by our current model restricts our assessment to a given geometric structure; as a result, the detailed effects of changing the metal cofactor geometry on the folding landscape do not directly enter, but instead only the overall energetics of the coordination process enter the model. To treat electronic structure effects on the folding landscape explicitly would require extensive *ab initio* quantum mechanical calculation, or at the very least, a highly refined semi-empirical quantum treatment.

III.B.3 The Coordinating Stiff Chain

To model the C112 and H117 residues as constituents of the stiff chain, we introduce an additional term to the usual polymer backbone term H_{chain}. A fixed angle θ between adjacent bonds based on the molecular structure is assumed and explicitly modeled in the inverse of the monomer correlation Γ. The usual backbone scaffold term H_{chain} has a Γ matrix form as follows

$$\Gamma = \frac{1-g}{1+g}K^R + \frac{g}{1-g^2}[K^R]^2 - \frac{g^2}{1-g^2}\Delta$$

$$K^R = \begin{pmatrix} 1 & -1 & & \cdots & 0 \\ -1 & 2 & -1 & & \vdots \\ & \ddots & \ddots & \ddots & \\ \vdots & & -1 & 2 & -1 \\ 0 & \cdots & & -1 & 1 \end{pmatrix}$$

and

$$\Delta = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ -1 & 1 & \vdots & & \vdots \\ 0 & & \ddots & & 0 \\ \vdots & & \vdots & 1 & -1 \\ 0 & \cdots & 0 & -1 & 1 \end{pmatrix}$$

where $k_g = -cos\theta$, $(\pi/2 < \theta < \pi)$, and K^R is Γ in terms of a Rouse matrix, and Δ accounts for the polymer boundaries of the respective termini, based on the stiff chain model [70]. To account for the cofactor's interaction with the native ligand set, the correlation matrix is modified to be Γ_{holo}:

$$\Gamma_{holo} = \begin{pmatrix} \Gamma & 0 \\ 0 & C_{[129,129]} \end{pmatrix} + C_{[112,112]} - C_{[112,129]} - C_{[129,112]} + C_{[117,117]} - C_{[117,129]} - C_{[129,117]}$$

$C_{[112,129]}$ and $C_{[117,129]}$ describes the position correlations between the zinc ion and C_α atom of residue 112 and 117. The values are rescaled with the ligand length: $C_{[112,112]} = -C_{[112,129]} = -C_{[129,112]} = 0.626$, $C_{[117,117]} = -C_{[117,129]} = -C_{[129,117]} = 0.577$ and $C_{[129,129]} = C_{[112,112]} + C_{[117,117]}$. The resulting backbone scaffold is represented by

$$H_{chain} = \frac{3}{2a^2}\sum_{ij} r_i \Gamma_{holo,ij} r_j + \frac{3}{2a^2} B \sum_i {r_i}^2$$

III.B.4 Modeling the Non-covalent Ligand Interactions

Experimentally, one finds the cofactor-ligand interactions confer an additional 7 kcal mol^{-1} of stability to the folded protein [15]. The microscopic rates of the individual metal-ligand association reactions are significantly larger than the overall folding rate. This suggests the ligand cofactor interactions are most probably not rate-limiting during the folding process and can be treated as representing a quasi-equilibrium. To model the folding in the absence of the coordination reactions, the metal-ligand interactions with H46 imidazole and the carbonyl of G45 were first treated using a pairwise potential that would reflect only intramolecular charge-charge interactions within the protein. To approximate the electrostatics effects alone the weight of a given metal-ligand charged interaction was given by a strength coefficient ϵ_{ij} [1], based on the Miyazawa-Jernigan scale [39] with well depths set to 3 and 5 kcal mol^{-1} for glycine-Zn^{2+} and histidine-Zn^{2+}, respectively. These electrostatic well depths were chosen based on those for glycine or histidine interacting with singly positively charged residues, which we take to approximate the strength of the corresponding metal-ligand interactions, when there is no specific coordination.

To accurately fit the thermodynamics of the coordination in the context of the folded protein a different metal-ligand interaction $H_{int,coord}$ was used. When the residues become coordinated the contact interactions are increased in strength to have coefficients with well depths of 13 kcal mol^{-1} and 15 kcal mol^{-1} for glycine-Zn^{2+} and histidine-Zn^{2+} coordination, respectively. The ligation term when coordination occurs is written as

$$H_{int,coord} = \epsilon_{[45,metal]}u(r_{[45,metal]}]) + \epsilon_{[46,metal]}u(r_{[46,metal]}])$$

where $\epsilon_{[45,metal]}$ and $\epsilon_{[46,metal]}$ are termed the coordinate contribution of the histidine and glycine metal-ligand interactions, respectively. The difference between $\epsilon_{[45,metal]}$ and $\epsilon_{[46,metal]}$ reflect the expected difference between histidine nitrogen and carbonyl oxygen coordination. The overall magnitude of the binding results in a stability change at T = 1.91 due to the coordination event that is approximately 7 kcal mol^{-1}. Thus, the coordination strength fits the experimental thermodynamics. Notice that this is consistent with the entatic state hypothesis; the expected additional thermodynamic stability based solely on the coordination energies would be considerably higher (i.e., 28 kcal mol^{-1}) than the experimental value. This reflects the entropic cost of forming the coordination sphere in the context of the folded protein.

III.B.5 Approximating the Free Energy Surface

The free energy surface of the zinc-metallated protein, is obtained using a variational scheme based on a reference Hamiltonian H_0 as shown in previous chapter. The reference Hamiltonian constrains the biopolymer chain and the Zn^{2+} ion to fluctuate to varying extents about their location in the native state: $H_0 = H_{chain} + \frac{3}{2a^2}\sum_i C_i(r_i - r_i^N)^2$. Here C_i is a set of constraining variables that reflects the local Debye-Waller factors for main-chain motions, thereby monitoring the fluctuation of each residue about its native position r_i^N. The Feynman-Gibbs-Peierls-Bogoliubov variational principle is based on the reference Hamiltonian H_0 which yields variational free energy values as follows:

$$F[C] = -k_BT\log Z_0 + \langle H - H_0\rangle_0$$

Here, Z_0 is the partition function of the reference Hamiltonian and $< H - H_0 >_0$ denotes the average with respect to H0. Using this relation, energies and entropies were computed for the metallated wild-type and several variants as described by [41].

III.C Results and Discussion

III.C.1 The Folding Free Energy Landscape of Metallated Azurin: Qualitative Connection between Experiment and Theory

Figure III.1 exhibits the predicted folding free energy profile, when modified to incorporate ligation effects, as a function of a single reaction coordinate. Although, *a priori* the precise energetic consequences of the ligation events requires extensive quantum calculations, the available experimentally-measured stabilities of azurin with and without the zinc cofactor provides a reasonable parameterization of the energies [12, 15]. In Figure III.1 we show the free energy profile of zinc-metallated azurin first when the non-covalent ligands (i.e., G45 and H46) are treated as having electrostatic interactions H_{int} alone (filled circles) as well as the profile when the residues become coordinated $H_{int,coord}$ (open squares). Coordination confers an additional 7 kcal mol^{-1} of stability at T = 1.91. Very significant stabilization in the free energy profile arising from the coordination contribution already occur at the early transition state and native state ensembles. We see that, the entatic state forms concomitantly with the folding nucleus.

The predicted folding routes and the position of the folding barrier of Zn^{2+} substituted azurin are shown as a function of temperature in Figure III.2. This collection of folding profiles reveals a stark difference between the apo- and holo-azurin system (Figure III.2); specifically, for the Zn^{2+} form the position of the rate-limiting step in a given folding route varies as a function of temperature. In contrast to what is found for the apo-enzyme, the folding barrier for the Zn^{2+} metallated protein progressively moves towards the native structure as temperature increases, in good agreement with experimental observation. Interestingly, as the temperature increases, the ligation intermediate also becomes more stable relative

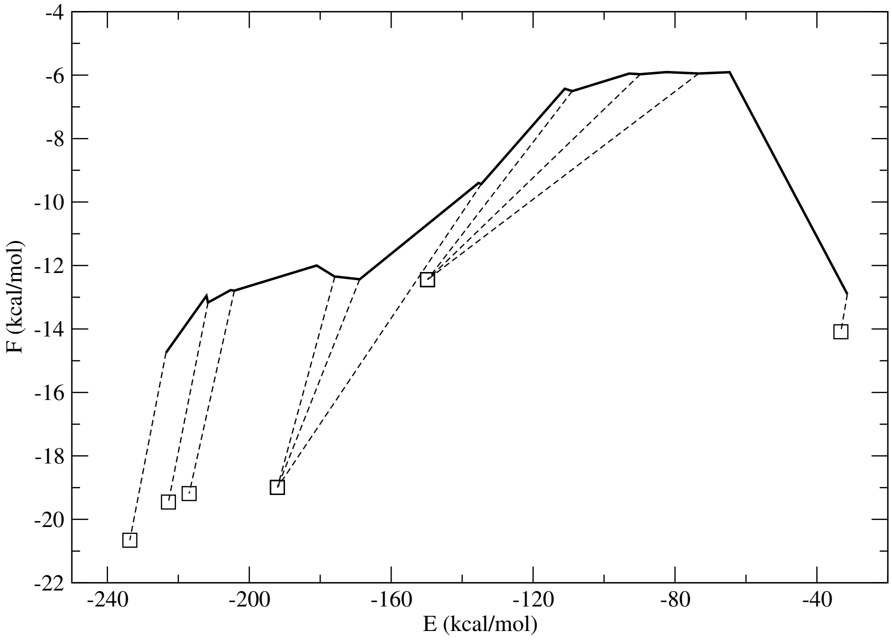

Figure III.1: The free energy profile of zinc metallated azurin at temperature T=1.91. The bold line represents the free energy profile when the metal-ligand interactions were simply treated as contacting positions carrying electrostatic interactions during the folding event. Dashed lines connect the corresponding positions of the free energy profile of the metallated enzyme treated with the coordinate contribution of the histidine and glycine metal-ligand interactions $H_{int,coord}$ (open squares).

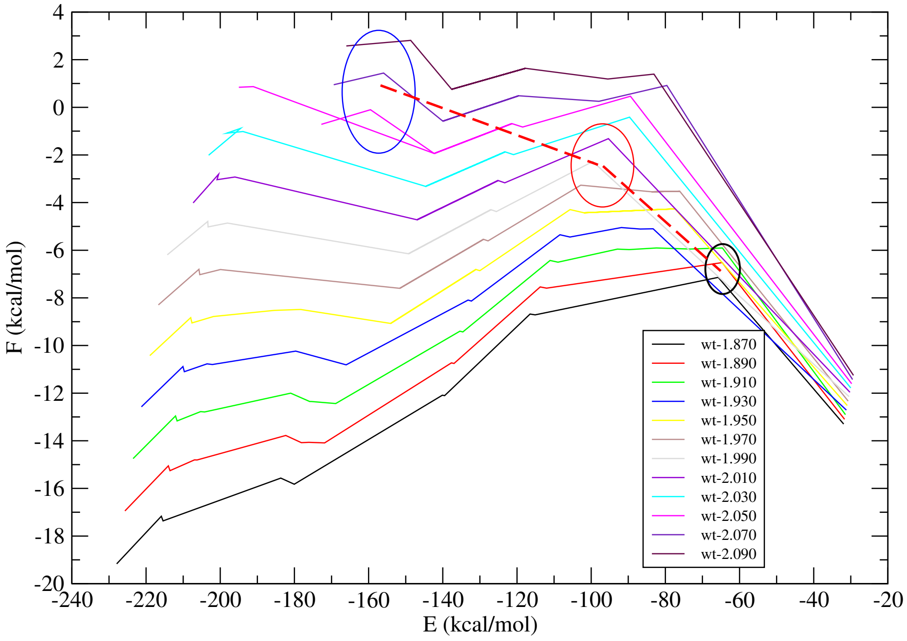

Figure III.2: The free energy profile for metallated-azurin as a function of temperature. The dashed-line follows the trajectory of the metallated folding nucleus as function of temperature. From right to left the corresponding temperatures for the folding barriers are ~ 1.86 (early ‡, black circle), ~ 1.96 (middle ‡, red circle), and ~ 2.06 (late ‡, blue circle).

to the metal-ligand interactions approximated by the electrostatics effects alone; thus, the two differently interacting conformational ensembles probably can coexist under some thermodynamic conditions (e.g., T $\sim$ 1.91) (Figure III.1). At higher temperature the ligation intermediate finally becomes more stable than the native state based only on the charge-charge interaction; thus, the formation of the entatic state makes a greater contribution to the folding reaction at higher temperatures.

III.C.2 The Structural Interpretation of the Folding Dynamics: The Rise of the Entatic State

Figure III.3 shows the predicted mean square deviations (MSD) of each residue from its native location in the transition state ensemble both as a function of sequence and of temperature, based on our modified variational scheme. This plot provides a detailed structural interpretation of how the folding routes change with temperature. As the MSD of a residue becomes smaller, the more native-like that position becomes. This plot clearly shows that the folding nucleus becomes less diffuse (more native-like) with increasing temperature, which is consistent with the free energy folding routes shown in Figure III.2. Fixing our attention on the primary coordination sphere, residue C112 shows the smallest fluctuations throughout the dynamic transition, while H117 exhibits a progressive decrease in variability relative to its mean position as the temperature increases, finally assuming a near native-like fluctuation at T = 2.06. Interestingly, M121 (which was not explicitly modeled as a coordinating residue, but simply as a contacting position) demonstrates the most dramatic change in relative position early in the transition (i.e., from 1.86 to 1.96). Conversely, the non-covalent coordination ligands G45 and H46 simultaneously experience a marked change only later in the dynamic transition (i.e., 1.96 to 2.06) (Figure III.3).

How does the geometric entatic state develop relative to the formation of the complete scaffold? In the early transition state of the metallated protein, aggregation of the C-terminal region (residues 85-128 or β-strands 5, 6, 7, 8)

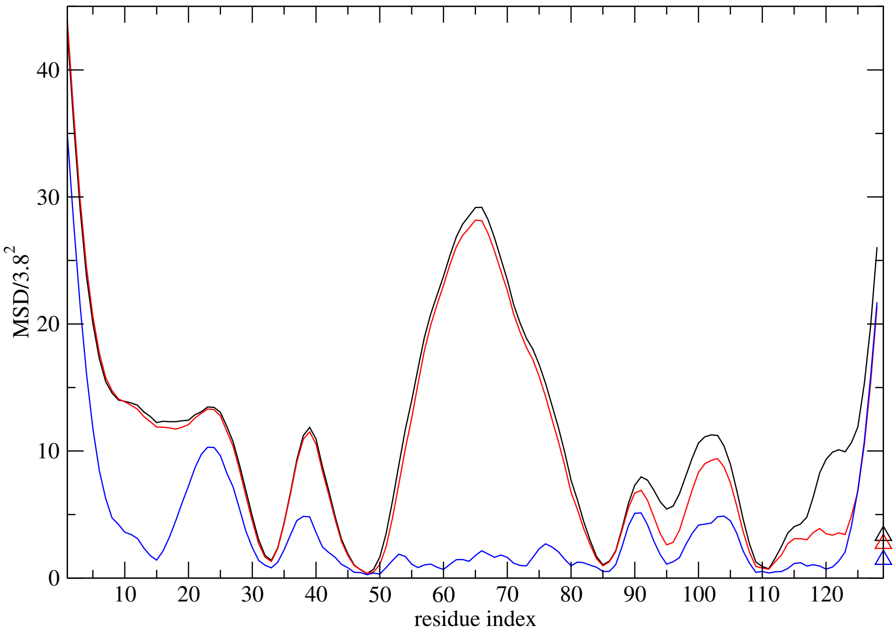

Figure III.3: The local fluctuations around the native structure of members of the transition state ensemble as measured by the mean square deviation of residues as function of temperature (i.e., T=1.86 (early ‡), 196 (middle ‡), and 2.06 (late ‡) represented as blue, red, and black, respectively) and residue sequence number. The fluctuation of a given residue constituting the fold barrier is given by the covariance matrix B, where $B_{ij} = a^{-2}(r_i - < r_i >)(r_j - < r_j >)$ and a is a scaling factor equal to 3.8 Å. The cofactor is represented by the set of triangles in the right lower corner.

provokes a more native-like geometry at the coordinating loop; in turn, the residues of the N-terminus (residues 1-85 or β-strands 1, 2, 3, 4, and the α-helix) experience a significant reduction in their fluctuations, completing the ligand set as well as the proper geometry of the entatic state (Figure III.3). A majority of the residues in the primary coordination sphere are formed very close to their final location very early in the moving transition state. This reflects a considerable degree of topological frustration in the system giving a large entropic penalty as a result of forming this early conformation of residues distant in sequence from each other. Concisely, the canonical loop forms and establishes native-like geometry for residues C112, H117, and possibly M121 but precedes the native interactions with ligands 45 and 46 that complete the entatic state. In our model, we have not explicitly included a term for non-native interactions or misligations; therefore, we do not explicitly show any possible energetic frustration around the coordination sphere that might result from these factors. However, the entropic factor caused by the stringent distance and geometry requirement by itself provides sufficient destabilization in accord with the entatic mechanism.

III.C.3 A Comparison of the Experimentally Inferred and Predicted Folding Dynamics

The calculated free energy profiles already provide a correct qualitative description of the folding event likewise they also give quantitative predictions. The apparent activation free energy determined by the natural logarithms of the observed (un)folding rates as a function of denaturating conditions often generate linear — or in our case, more interestingly, non-linear — extra-thermodynamic free energy relationships with stabilization free energies. These free energy relationships yield the so-called Chevron plots. Each Chevron plot provides an overview of the energetic consequences of mutations on the folding barrier as well as the relative position of the folding barrier along the reaction coordinate. Reconstruction of a non-linear (curved) Chevron plot is not trivial, requiring accurate prediction of absolute folding and unfolding rates. In the theoretical calculations, the balance of folding reaction is altered by changing the temperature, while in the laboratory the balance is changed using chemical denaturant.

In order to calculate the folding rate at a given temperature T, first one identifies the folding barrier position at E‡ on the free energy profile (Figure III.2). E‡ is the sum of the contact energies with the highest free energy. The sum of the contact energies is an order parameter paralleling the more commonly employed Q, which is appropriate for funneled landscapes. This choice of coordinates is sensible if non-native interactions are neglected. In the solvent denatured situation nonspecific collapse also probably contributes to E‡. Once the rate limiting step (i.e., the highest folding barrier) is identified, the corresponding free energy changes to the folding barrier $\Delta G^{\ddagger}_{u,f} = -(G^{\ddagger} - G_{u,f})$ can be calculated. The rate coefficients for folding k_f and unfolding k_u follow using the Arrhenius equation $k_{u,f} = A\exp(-\Delta G^{\ddagger}_{u,f}/k_BT)$. where A the pre-factor is be calculated microscopically [11]; we fit the parameter A in the present analysis. At last, the observed relaxation rate k_{obs} is the sum of k_f and k_u. To simplify the analysis, the rate coefficients at different temperatures were fit using a second-order polynomial in the exponent of Eqn. (III.2).

$$k_f(T) = k_f(T_1)\exp\left[a \times (1/T - 1/T_1) + b \times (1/T - 1/T_1)^2\right]$$

$$k_u(T) = k_u(T_2)\exp\left[c \times (1/T - 1/T_2) + d \times (1/T - 1/T_2)^2\right]$$

$$\begin{aligned}\log k_{obs}(T) \;&=\; \log\{k(T_1)\exp\left[a \times (1/T - 1/T_1) + b \times (1/T - 1/T_1)^2\right] + \\ &\quad k(T_2)\exp\left[c \times (1/T - 1/T_2) + d \times (1/T - 1/T_2)^2\right]\} \qquad \text{(III.2)}\end{aligned}$$

The parameters a and c give the linear dependence of folding and unfolding, respectively; while, the observed curvature of the folding and unfolding arms are reflected by the parameters b and d, respectively. The resulting in machina Chevron plots (Figure III.4) (i.e., for wild-type along with 14-point mutated variants) allow for a more thorough assessment of the transition state as a function of

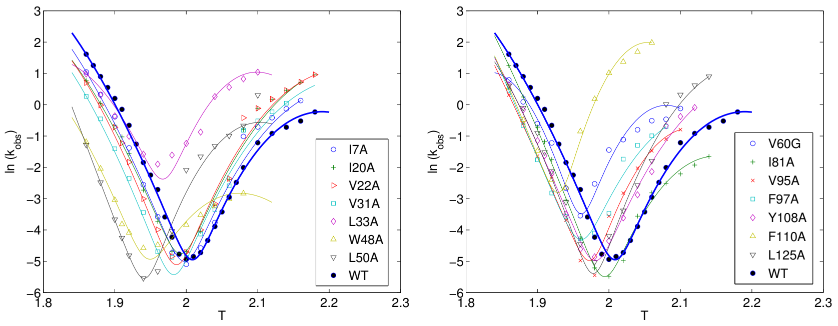

Figure III.4: Theoretical Chevron plots: $\ln k_{obs}$ versus temperature for 14 metallated-azurin variants provides an overview of the energetic consequences of mutations on the folding barrier along with the relative position of ‡.

temperature. In the fits the parameters a, b and c, d satisfy the stability requirement. So there are only two independent degrees of freedom in the fitting.

The calculated Chevrons allow one to compare the relative stability of the folded protein, $\Delta\Delta G_{N\_D}$ and the folding barrier, $\Delta\Delta G_{\ddagger\_D}$ for each variant compared to those of the wild-type. Combining the relative changes of the folding barrier and protein stability yields theoretical phi-values $\phi_T = \Delta\Delta G_{\ddagger\_D}/\Delta\Delta G_{N\_D}$, which can then be directly compared to experimentally determined phi-values ϕ_E (Figure III.5). A recent experimental study that employed ϕ-value analysis, as a function of discrete denaturant concentrations, already gave snapshots of the zinc-metallated azurin's dynamic folding nucleus with residue-specific resolution [15]. Figure III.5 provides a direct comparison of the theoretically and experimentally derived ϕ-values, at discrete temperatures and GuHCl concentrations with corresponding stabilities, respectively. This comparison clearly shows a solid correlation between the experimentally and theoretically derived ϕ-values at each condition (i.e., GuHCl concentration or temperature, respectively). Moreover, this correlation clearly shows that the present variational scheme is quite robust, and accurately predicts the zinc-metallated azurin's dynamic folding nucleus with residue-specific resolution that is at least on par with that provided by the experimental study.

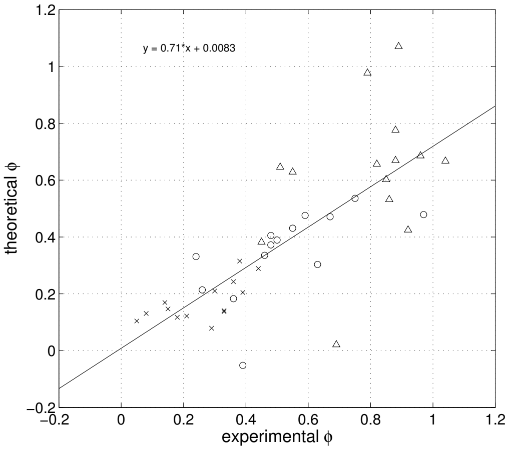

Figure III.5: A direct comparison of theoretical and experimental ϕ-values. cross label represents $\phi_{experimental}$ at 0M and $\phi_{theoretical}$ at T=1.86 (early ‡), circle label represents $\phi_{experimental}$ at 2M and $\phi_{theoretical}$ at T=1.96 (middle ‡), triangle label represents $\phi_{experimental}$ at 4M and $\phi_{theoretical}$ at T=2.06 (late ‡). The correlation coefficient between the calculated and experimental values is 0.77.

III.D The Effects of the Entatic State on the Dynamic Folding Nucleus

Although, the detailed electronic structure aspects — i.e., the quantum mechanical features — of forming the entatic state throughout the folding reaction (specifically with regard to the redox active copper site) can not be addressed explicitly using the model Hamiltonian we employ, our current approach would provide a crude prediction of the effects of tuning the reduction potential through metal substitution. Specifically, we can examine the redox phenomenon during folding by varying the relative coordinate contribution in the model Hamiltonian. Unfolded copper-metallated azurin has a reduction potential of ~0.5V, which can be ascribed to the electron-donating properties of the C112 thoilate moiety. As the protein folds, the progressive dehydration of the metal's milieu (i.e., hydrophobic encapsulation proximal to the active site) lowers the redox potential [55, 60]. Thus, as the metallated-protein folds the redox active copper becomes less susceptible to reduction. That is to say, the ligand interactions cooperatively change as the molecule becomes more native-like. Our calculations show the most dramatic changes due to ligation occur early in the resulting free energy profile.

III.E Conclusions

In this study, the folding dynamics of zinc-metallated *P. aeruginosa* azurin was investigated via a free energy functional, which models the coordination reaction explicitly. Both the qualitative form for the free energy profile and the quantitative predictions of the energetic consequences of mutations derived from our modified variational scheme agree very well with experimental observation [15]. The calculations show that the progressive movement of the folding barrier toward the native state reflects the effects of topological frustration in forming the geometric entatic state and results in a non-linear free energy relationship (i.e., a curved Chevron plot). The calculation clearly shows that at high temperature the activation energy required to break the bonds between the cofactor and respective ligands (i.e., resides G45 and H46) is much larger than the barrier to simply unfold the polypeptide. This additional rate limiting event results in a kinetic bottleneck which in turn changes the pattern of the overall free energy relationship for zinc metallated-azurin from that of the apo-protein. By combining theoretical modeling and experimental studies in the laboratory we can see how forming the entatic state is coupled to the dynamics of folding the metallated azurin at a level of detail that cannot be currently achieved by experiments alone.

III.F Acknowledgments

This chapter is based on the material as it appears in Proc. Natl. Acad. Sci. 104: 3159-3164 (2007), C. Zong; C. J. Wilson; T. Shen; P. Wittung-Stafshede; S. L. Mayo; P. G. Wolynes. The dissertation author was the primary investigator and author of this paper.

IV

Modeling α/β Proteins and Structure Prediction

The folding process is a great feat that evolution has achieved in biological systems. How proteins fold into organized structures based on their primary sequence has been a great mystery since the day Levinthal raised his puzzles. By characterizing the energy landscapes of proteins with principles of the statistical mechanics of disordered systems like spin glasses, a fundamental framework and flexible language for studying these complex and evolved systems has emerged over the last fifteen years or so. Experimental studies reveal that many aspects of folding dynamics can be quantitatively captured in the framework of a funneled energy landscape, as we illustrated in previous chapters.

Both fundamental molecular mechanisms underlying the folding and many biological processes related to folding process have been widely investigated. Even though the energy landscape theory has been widely accepted, ab-initio prediction the 3D protein structure from primary sequence remains a challenge. While successful predictions have been made for small size systems, there is still much to learn about folding mechanisms and much work to do for achieving universally reliable structure prediction, especially for large systems.

IV.A Introduction of the Associative Memory Hamiltonian

The starting point for our structure prediction potential development is the Associative Memory Hamiltonian (AMH) introduced by Friedrichs and Wolynes [16]. The AMH is intrinsically a coarse-grained model, where each residue is represented by carbon α, carbon β, and oxygen atoms. The Hamiltonian contains two major components: i) sequence-independent polymer physics terms to describe the backbone interactions, ii) sequence-dependent knowledge-based potentials optimized to achieve folding of a number of training proteins. The backbone interactions include chain-connectivity, excluded-volume, Ramachandran and chirality potentials. The sequence-dependent interactions involve only $C^{\alpha} - C^{\alpha}$, $C^{\alpha} - C^{\beta}$, and $C^{\beta} - C^{\beta}$ pairs. These interactions are grouped into three proximity classes according to the sequence distance between the interacting residues, as follows: short range ($3 \leq |i-j| < 5$), medium range ($5 \leq |i-j| \leq 8$), and long range ($|i-j| > 8$). For the short and medium classes, a pairwise interaction in the target protein is associated with a corresponding pairwise interaction in memory proteins. The associative part is then expressed as follows.

$$H_{AM} = -\sum_{\mu}^{n} \sum_{i<j} \gamma(P_i, P_j, P^{\mu}_{i'}, P^{\mu}_{j'}, (j-i)) \Theta(r_{ij} - r^{\mu}_{i'j'})$$

In order to characterize the geometric features of the backbone, terms for amino acid chirality, an excluded volume term, and a combination of harmonic terms, and SHAKE [71] constraints which maintain the planarity of the peptide bond, and appropriate bond lengths, and bond angles are included in $H_{backbone}$ as follows.

$$H_{backbone} = -(\lambda_{\phi\psi} V_{\phi\psi} + \lambda_{HB} V_{HB} + \lambda_{\chi} V_{\chi} + \lambda_{EV} V_{EV} + \lambda_{Harm} V_{Harm})$$

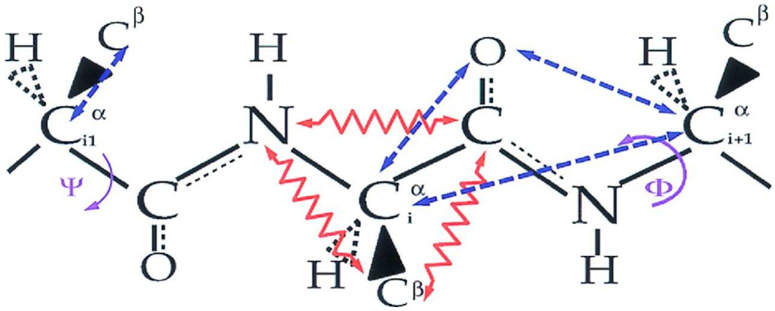

Figure IV.1: The atomic structure of protein backbone. Carbon α and carbon β are labeled as C^α and C^β. The two torsional degrees of freedom are labeled as Φ and Ψ angle. The SHAKE algorithm is used to constrained the distance between C^α, C^β and Oxygen atom as indicated by the blue lines. The planar geometry of the residue can be used the rest of heavy atoms in the backbone.

IV.B Water-Mediated Potential for Interactions of Large Sequence Distance

We have used the associative memory term for treating the interactions within close sequence distance (≤ 8). For the long range proximity class, simple square well potentials, unrelated to memory proteins [23, 24], are used. The terms of this function are partitioned into two wells, based on the physical distance. The first well covers the 4.5Å to 6.5Å interval, representing a simple contact between two residues. The second well covers the 6.5Å to 9.5Å interval, representing protein-mediated or water-mediated interactions. To determine whether an interaction is protein or water mediated, the local density around each pair of residues is computed from $\rho_i = \sum_k \theta^I_{ik}$. The water-mediation is switched on only when two residues are both exposed to water based on the criterion of density threshhold ρ_{trsh}. When the water-mediation is switched off, protein-mediated interactions are used. We use σ^{prot}_{ij} and σ^{wat}_{ij} to describe the above two types of interactions.

$$\sigma^{wat}_{ij} = H\left(\rho_i - \rho_{trsh}\right) H\left(\rho_j - \rho_{trsh}\right), \qquad \text{(IV.1)}$$

$$\sigma^{prot}_{ij} = 1 - \sigma^{wat}_{ij} \qquad \text{(IV.2)}$$

$$H\left(\rho_i - \rho_{trsh}\right) = 1/2\left(1 - \tanh\left(\kappa(\rho_i - \rho_{trsh})\right)\right) \qquad \text{(IV.3)}$$

κ is a parameter that describes the sharpness of the switching tanh functions (κ was set to 5.0). The σ switching functions are constructed in such a way, that when the local density ρ for each residue increases beyond a threshold value of ρ_{trsh}. The σ^{wat} switches smoothly from 1 to 0, whereas, σ^{prot} switches from 0 to 1.

To treat the well's boundary, we introduce θ_{ij}^{II} and summarize the interactions, H_{water}, as follows:

$$\begin{aligned} H_{water} &= -1/2 \sum_{i,j} \theta_{ij}^{II} \left(\sigma_{ij}^{wat} \gamma_{ij}^{wat} + \sigma_{ij}^{prot} \gamma_{ij}^{prot} \right), & \text{(IV.4)} \\ \theta_{ij}^{II} &= 1/4 \left(1 + \tanh\left(\kappa (r_{ij} - r_{min}^{II}) \right) \right) \left(1 + \tanh\left(\kappa (r_{max}^{II} - r_{ij}) \right) \right) & \text{(IV.5)} \end{aligned}$$

where r_{ij} is the distance between residues i and j, r_{min} and r_{max} indicate the endpoints of corresponding wells ($r_{min} = 4.5$Å and $r_{max} = 6.5$Å for the first well, $r_{min} = 6.5$Å and $r_{max} = 9.5$Å for the second well).

IV.C Hydrogen Bond Potential and Modeling of β Sheets

The hydrogen bonding pattern between residues is described by the following potential:

$$\begin{aligned} \Theta^{HB}(ij) &= -\lambda_{HB}(|i-j|) \exp\left[\frac{-(r_{ij}^{ON} - < r^{ON} >)^2}{2\sigma_{NO}^2} - \frac{(r_{ij}^{OH} - < r^{OH} >)^2}{2\sigma_{HO}^2} \right] \times \\ & \quad 0.25 \times (1 + \tanh(r_{i-2,i+2}^{Ca} - r_c)) \times (1 + \tanh(r_{j-2,j+2}^{Ca} - r_c)) \qquad \text{(IV.6)} \end{aligned}$$

where r_{ij}^{ON} denotes the distance from the carbonyl oxygen on residue i to the nitrogen on residue j, and r_{ij}^{OH} denotes the distance from the oxygen on residue i to the bonded hydrogen on residue j. The geometric constraint is expressed by the two hyperbolic tangent terms in Eqn. (IV.6). r_c is the geometry parameter describing the minimum five-residue strand extension. When the length of five-residue segment is less than r_c, the middle residue of this segment can hardly form hydrogen bonds with other residues in the protein. Here, we choose r_c to be 12Å.

An energy function for β sheet formation was developed by C. Hardin *et al.* [22]. Here we adapt a similar formulation. Since β strands are usually quite extended in order to effectively form the hydrogen bonding network, we further added a constraint term to allow only small curvature of the strand in the β sheet formation. Furthermore, we set three sequence-separation based proximity classes for hydrogen bonding potentials: for the first class the sequence distance for a pair of interacting residues is less than 19; for the second class it is between 19 and 45; for the third class is larger than 45. The hydrogen bonding potentials include three terms to represent pairwise interactions, parallel nonadditivity, and antiparallel nonadditivity, respectively. When both pairwise and nonadditive interactions are present, the hydrogen bonds sometimes become too difficult to break, once they have formed. In order to avoid strong local collapse of β strands, we only turned on the two nonadditive terms (Λ_2 and Λ_3 terms) when the interacting residues are both predicted to be in a β strand from a secondary structure prediction server JPRED [72]. Here we hypothesize that these residues are the ones giving the most energetic stabilization to the β sheets. A similar conjecture about the stabilization of β strands has also been discussed for the folding mechanism of β hairpins [73].

$$\begin{aligned} V(ij)_{HB} &= -\Lambda_1(|j-i|)\Theta^{HB}(ij) - \Lambda_2(a_i, a_j, |j-i|)\Theta^{HB}(ij)\Theta^{HB}(ji) - \\ &\quad \Lambda_3(a_i, a_j, |j-i|)\Theta^{HB}(ij)\Theta^{HB}(j, i+2) \end{aligned} \tag{IV.7}$$

The Λ terms are given as follows:

$$\Lambda_1(|j-i|) = \lambda_1 \tag{IV.8}$$

$$\begin{aligned} &\Lambda_2(a_i, a_j, |j-i|) = \lambda_2 - \alpha_1 \ln P_{anti}(a_i) + \\ &\quad \alpha_1 \ln P_{anti}(a_j) + 0.5\alpha_2(|j-i|) \ln P_{HB}(a_i, a_j) - \\ &\quad 0.25\alpha_3(|j-i|) \left[\ln(P_{NHB}(a_{i+1}, a_{j-i}) + \ln P_{NHB}(a_{i-1}, a_{j+1}))\right] \end{aligned} \tag{IV.9}$$

$$\begin{aligned} &\Lambda_3(a_i, a_j, |j-i|) = \lambda_3 - \alpha_4 \ln P_{par}(a_{i+1}) - \\ &\quad \alpha_4 \ln P_{par}(a_j) - \alpha_5(|j-i|) \ln P_{par}(a_{i+1}, a_j) \end{aligned} \tag{IV.10}$$

where the Θ functions are shown in Eqn. (IV.6). The Λ_1 term describes the pairwise part of the stabilization in forming the hydrogen bonds. The Λ_2 term gives an additional stabilization to an anti-parallel β hydrogen bonding and the

Λ_3 term gives an additional stabilization to parallel β patterns. First, Λ_2 and Λ_3 depend on the types of amino acids labeled by a_i and a_j. A 20-letter code was used to present the residue preference for the parallel or anti-parallel formation. Moreover, the nonadditive terms Λ_2 and Λ_3 are only added when residue i and j were both predicted to be in β strands using the secondary structure prediction algorithm. The $|j-i|$ dependence of functions Λ_1, Λ_2, and Λ_3 indicates that the coefficients are set respectively for each proximity class. The detailed parameters of λ and α in the Λ term are described in the paper [24].

IV.D Spherical and Nonspherical Gyration Radius Potential

The spherical collapse potential is generally a harmonic potential to control the gyration radius of the protein. The native gyration radius is estimated by the formula $Rg^0(N) = 2.2N^{0.38}$. The harmonic potential is given by:

$$E_{radius} = \begin{cases} \lambda_{\text{radius}}(Rg - Rg^0)^2 & , \quad 0.75 < Rg/Rg^0 < 1.5 \\ \lambda_{\text{radius}}(1.5Rg^0 - Rg^0)^2 & , \quad Rg/Rg^0 \geq 1.5 \\ \lambda_{\text{radius}}(0.75Rg^0 - Rg^0)^2 & , \quad Rg/Rg^0 \leq 0.75 \end{cases}$$

We take $\lambda_{\text{radius}} = 10.0\epsilon$. For nonspherical collapse potential, the gyration on each axis is controlled. The expression and the cutoff are the same as for the spherical potential. The Rg^0_x, Rg^0_y, and Rg^0_z were chosen to reflect the shape of the native structures. In the case of T089, Rg^0_z is chosen to be $1.40 \times Rg^0$, Rg^0_x and Rg^0_y is chosen to be $Rg^0/1.40$. A rod-like shape is given by this setup of collapse potential.

IV.E Constrained Self-Consistent Optimization

A self-consistent optimization scheme was used to tune the various interaction strengths in the Hamiltonian. The optimization is based on the minimum frustration principle [74]. The energetic stabilization in the folding process is described by the energy gap, δE, between the molten globular states and the native-like states. At the folding temperature T_f, the energy gain, δE, is balanced by the loss of configurational entropy S_c. Thus, the folding temperature T_f is expressed as $\frac{\delta E}{S_c}$ [74]. Another important characteristic of the folding process is the ruggedness of energy landscape, described by the energy variance of molten globular states, $\sqrt{\Delta E^2}$. The ratio of this variance to the entropy of the molten globular states, $\frac{\sqrt{\Delta E^2}}{S_{mg}}$, provides an estimate of the polypeptide chain glass transition temperature T_g. Maximizing the ratio of the folding temperature over the glass transition temperature, $\frac{T_f}{T_g}$, provides a quantitative procedure to minimize the frustration presented in a knowledge-based Hamiltonian for a training set of proteins.

In this optimization scheme, additional constraints are imposed upon the mean and the variance of the molten globular structures for each proximity class. Thus, the optimization preserves the energy balance between different proximity classes. We used 14 α/β proteins to "train" the Hamiltonian. Decoy structures were self-consistently generated from samplings at high temperature, $1.2T_f$. The native-like ensemble of structures was generated from biasing sampling to the native region. A Lagrangian functional, containing the constraints on the mean and variance, was minimized for each proximity class [20, 24].

IV.F Simulated Annealing and Results Discussion

We carried out molecular dynamics simulations with temperature quenching (simulated annealing) to search for low energy conformations. Three α/β proteins, that were dissimilar to any of the training proteins, were used to test our current model. For each protein, 24 simulated annealing runs were carried out. Next, we define a critical assessment of structural similarity between the native structure and the predicted structures based on all pairwise residues distances: $Q = \frac{2}{(N-1)(N-2)} \sum_{i<j-2} \exp\left[-\frac{(r_{ij}-r_{ij}^N)^2}{2\sigma_{ij}^2}\right]$. A structure with $Q = 1.0$ corresponds

precisely to the native structure, while the conformations having Q values near 0.4 are typically characterized by 5Å RMSD fit to the native structure. We used yet another similarity measure to compare conformations, the Z score calculated with the combinatorial extension (CE) algorithm [75]. This score identifies general topological similarity disregarding the sequence information. In general, a Z-score of 3.5 indicates significant structural similarity, while strong structural similarity is achieved for Z scores larger than 4.0.

Three test proteins were CASP targets, with indices T089, T120, and T251. The crystal structure analysis for these proteins indicates diverse topologies. For instance, the β sheets in the test proteins are quite different in both their shapes and their locations. T089 is a single domain from protein 1E4F (a CASP4 target), taken from residues 86 to 166. In the T089 native structure, a long three-strand β sheet is formed around the α helix. The second test protein, T120, having 115 residues, is an N-terminal domain of human XRCC4DNA repair protein, 1FU1, (a CASP4 target). The native structure of this protein is comprised of two sandwich-like β sheets with two helices connecting them. The third test protein, T251, which contains 99 residues, was taken from protein, 1XG8 (a CASP6 target). This protein is comprised from three outer helices, in addition to a four-strand β sheet mainly located in the core of the protein. An interesting aspect of the T089 and T251 topology is the nonlocal nature of β sheets, with β strands separated far apart in sequence. This makes these proteins challenging targets for structure prediction.

The structure prediction results for three test proteins, evaluated using the Q score, are summarized in Fig. IV.2. The Q scores are plotted in the sorted order of numerical values. For each of the three test proteins, we reached conformations with Q score greater than 0.35 within 24 short annealing runs. Using the CE score measure, we found that about 10 annealing runs for each protein sampled structures with Z larger than 3.7, corresponding to rather native topologies. When the nonspherical collapse potential was added to constrain the overall topology of

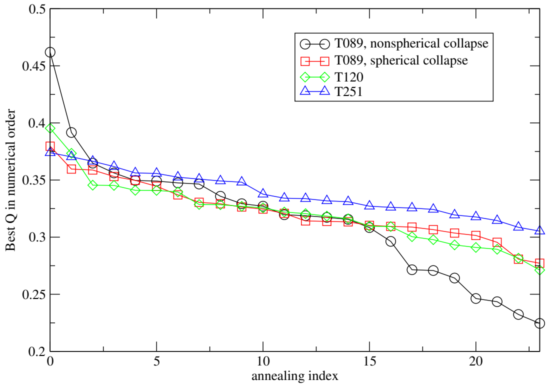

Figure IV.2: Best Q sampled in 24 annealing runs for proteins T089, T120 and T251. The spherical collapse potential is used for three proteins. The nonspherical collapse potential is used for the case of T089.

one of the targets, T089, the prediction results were significantly improved. The best Q score reached a high value of 0.45, exhibiting very strong similarity to the native structure. Samplings of the best predicted structures for three test proteins are presented in Fig. IV.3, IV.4, and IV.5. The native structures and the contact maps are also shown for comparison. Both structural drawings and the corresponding contact maps indicate that the predicted structures are very similar to the native structures, with some discrepancy in the packing of secondary structure elements.

Thermodynamically, the average energy decrease funnels toward the protein native state. On the other hand, the ruggedness of the energy landscape also critically affects the folding dynamics. The energy ruggedness leads to a glass transition at low temperatures needed to completely stabilize the native structure. To quantify the emergence of glassy behavior, while lowering the temperature, we evaluated Q-autocorrelation functions (Fig. IV.6). This plot provides a dynamic information on the ruggedness of the energy landscape at the given temperature scale. With decreasing temperature, the valleys of the energy landscape become too deep for the protein chain to overcome by simple thermal fluctuations, leading

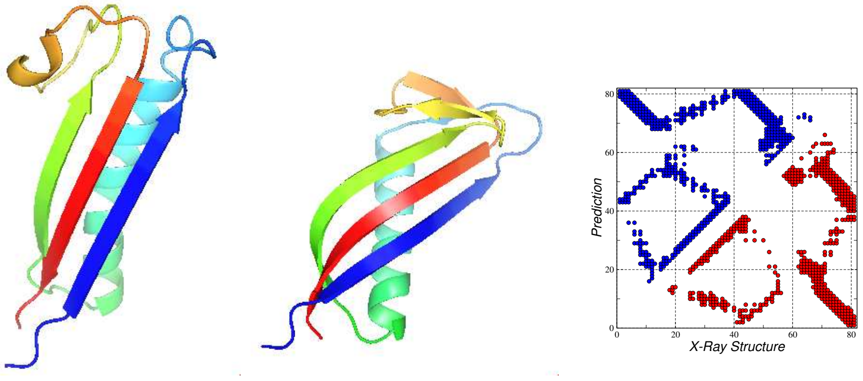

Figure IV.3: A predicted structure for protein T089 with the $Q = 0.46$ (CE: Z=4.1) (left) and the contact map (right). The prediction was generated with the nonspherical collapse potential.

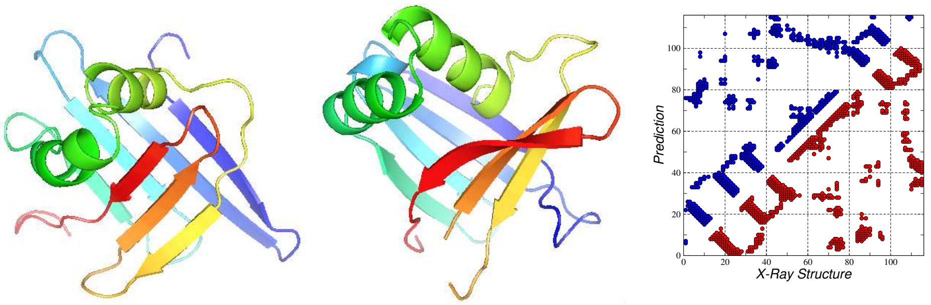

Figure IV.4: A predicted structure for protein T120 with the $Q = 0.39$ (CE: Z=4.7) (left), the native structure (middle) and the contact map (right).

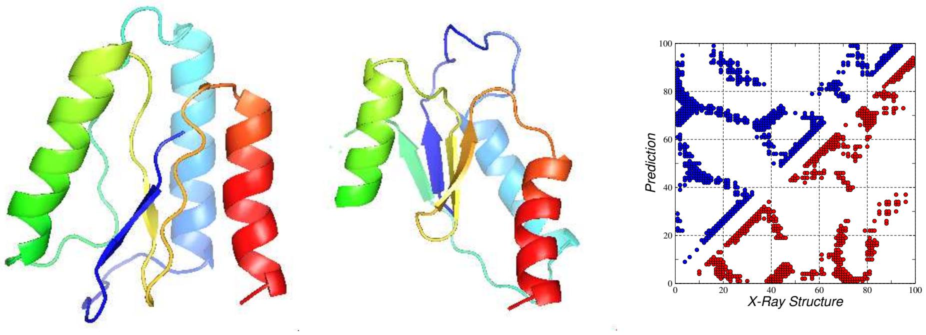

Figure IV.5: A predicted structure for protein T251 with the $Q = 0.37$ (CE: Z=3.6) (left), the native structure (middle) and the contact map (right).

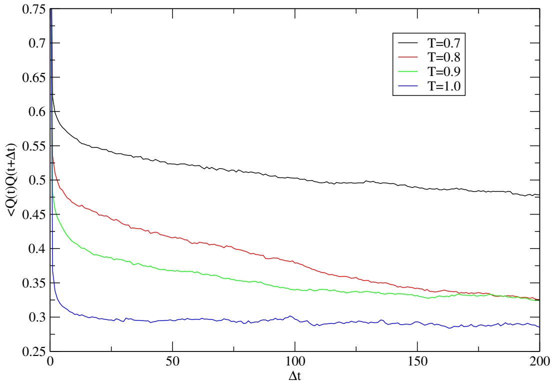

Figure IV.6: Q auto-correlation function at different temperatures for protein T089 (the collapse potential is spherical). The x-axis is time interval, in units of 450ps.

to trapping in low energy conformations. For protein T089, for example, when T is lowered below 0.9, the system no longer efficiently explores the configurational space on the simulation time scale. We found that the glass temperature for T089 with new developments is significantly lower than the results from the previous AMH study on this system [22]. We attribute this lowering to the water mediated interactions and the adjusted β potentials, that, in turn, help to decrease the energy ruggedness of the molten-globular states, resulting in a lower glass transition below $T = 0.9$. Reducing the energy ruggedness allows efficient sampling of native-like structures at lower temperatures, where the free energy favors more native structures.

In Fig IV.7, we provide a sequence of folding snapshots to illustrate the progression of conformations in the β sandwich-like protein T120. An initial collapsed conformation is shown in Fig IV.7(a). As the temperature is decreased, the helix starts to form with some alignment of β strands (Fig IV.7(b)). Partial formation of the N-terminal and C-terminal sheets were observed (Figs IV.7(c) and IV.7(d)). As temperature is further decreased, the full hydrogen bonding network is formed, producing a very native-like conformation.

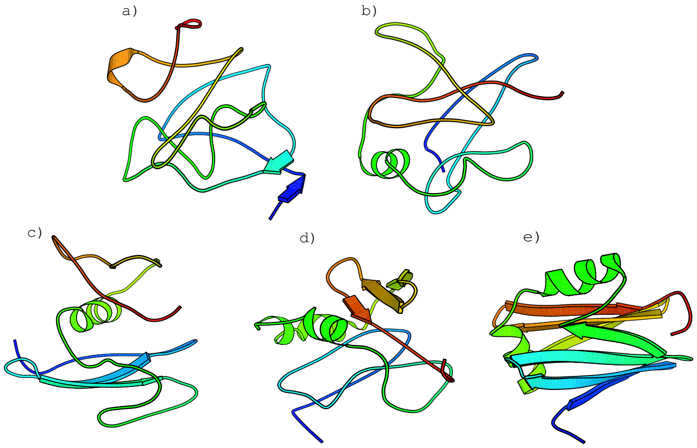

Figure IV.7: A sequence of snapshots was taken from a simulated annealing trajectory for protein T120. This trajectory eventually sampled a Q=0.39 structure. Snapshots (a-e) are sampled at T=1.75, 1.5, 1.48, 1.44 and 1.17 respectively.

IV.G Conclusion

In summary, our work demonstrates that correct modeling of cooperativity that is largely mediated by water is crucial for accurate structure predictions of β sheets in α/β proteins. First, many globular interactions are involved in forming β sheets. During the early events of protein folding, these tertiary interactions may occur prior to locking of hydrogen bonds between β strands. On the other hand, in the case of helices, the local hydrogen bonds appear prior to forming those globular interactions. Water mediated interactions play an important role for the recognition between β strands. They help to reduce the topological frustration, which in turn leads to more efficient sampling of structures having native-like packing. In α/β proteins, the early folding of α helices provides patches of hydrophobic surface to nucleate the alignment of β strands. This mechanism can be incorporated into a general capillary picture in protein folding. The exact timing of events between nucleation processes and the formation of secondary structures regulates the collapse of proteins. In the early stages of protein folding, the collapse can be nonspecific or specific. Our comparative analysis indicates that potentials that favor specific over non-specific collapse significantly improve structure prediction.

Water-mediated potentials may be combined with higher resolution models that include more details of the side-chains that take into account efficient packing of native-like protein structures.

IV.H Acknowledgments

This chapter is based on the material as it appears in J. Am. Chem. Soc. 128: 5168-5176 (2006), C. Zong; G. A. Papoian; J. Ulander; P. G. Wolynes. The dissertation author was the primary investigator and author of this paper.

V

Nonequilibrium Dynamic Self-Assembly of a One-Dimensional Fiber

V.A Fibers in the Cell and Introduction of the Model

After folding, proteins participate in biological functions at multiple scales of the architecture of the cell. The complexity inherited from proteins' sequence and structure heterogeneity enables proteins to assemble into molecular machines and accomplish a wide variety of biological functions. At the supramolecular level, one immediate task facing the cell is to provide the necessary mechanical support for its compartments. Fibers like actin, microtubule and intermediate filaments are the main structural elements in forming the large mechanical structures of cells [27, 31]. The fibers form by the assembly of many individual protein units. Among the well-studied examples of this supramolecular assembly are microtubules and actin filaments [76, 77, 28]. It is worth noticing that the assembly of the fibers requires chemical energy. The chemical-mechanical process of assembly will store chemical energy into the fiber structures for mechanical support and work. In general, to describe the nonequilibrium characteristics exhibited by such molecular machines presents a new aspect of statistical mechanics. In this Chapter, I focus on the nonequilibrium assembly/disassembly of microtubules. I present a nonequilibrium variational analysis that allows one to study the growth/decay dynamics and the transition between these two states.

A microtubule is a helical structure of multiple protofilaments and starts (the lateral bonds form a line of subunits with a pitch from the horizontal) with the $\alpha\beta$-tubulin heterodimer as its repeating unit. An actin filament is a double-stranded helical structure assembled from actin monomers. Rapid growth (polymerization) and decay (depolymerization) of the fiber ends are often observed. The transition from growth to decay, called a catastrophe, and the transition from decay to growth, called a rescue, are observed but with relatively low frequencies [34, 35, 78, 79, 32, 80, 81]. These dynamic aspects of microtubule assembly and the treadmill-like perpetual motion of actin filaments highlight the importance of nonequilibrium effects in cellular biology. Energy dissipation is critical for these processes: the adenosine triphosphate (ATP) or guanosine triphosphate (GTP), together called NTP in our notation, when associated with actin monomers or one of tubulin dimers is irreversibly hydrolyzed to adenosine diphosphate (ADP) or guanosine diphosphate (GDP), together called NDP, during the process of assembly.

Several models have been proposed to explain the nonequilibrium linear assembly of fibers. These models have not yielded a complete understanding of the behavior under all conditions that have been studied. The earliest approaches used general equilibrium assembly models, e.g., [82, 30], based on a single one-state picture [83, 84]. These models do not distinguish the NTP and NDP states of a subunit. Later some of the nonequilibrium features of the growth and decay were introduced. General two-state cap models have been studied by many researchers [36, 85, 86, 87, 88, 89]. However, these two-state models require a large set of parameters or introduce *ad hoc* hydrolysis rules or both. The cap has been commonly treated in two ways: one approach assumes that the hydrolysis is directly coupled with the growth, so only a single layer cap exists in microtubules. The other puts no *ad hoc* constraints on the hydrolysis; however, this type of model cannot readily explain the experimental observation that there is no apparent lag of hydrolysis at high concentration of tubulin-GTP in solution. Experiments have also shown that various lateral interactions may control the fiber dynamics [90, 91]. New developments based on lattice models with fewer hydrolysis rules have been presented in recent works [38, 92, 93].

Although simulation studies on detailed lattice models can reproduce many experimental observations, theoretical studies of simple nonequilibrium one-dimensional models derived from the microtubule system can give insights into nonequilibrium phenomena of self-assembly in general.

In this chapter, I will describe a general site-resolved model based on simple chemical processes. Multiple conformational states for each unit are explicitly invoked in this model. A related interesting site-resolved model of actin-cycle dynamics has recently been empirically derived [94]. However, a systematic way to construct the dynamics of these models has not been formulated. Here, our variational methods can provide such a systematic solution.

For microtubule systems, we simplify the description of the cylindrical structure by treating it as a one-dimensional fiber system. At first glance our model looks like a crude caricature of the sophisticated filament structure and the dynamics of microtubules. Nevertheless, this level of description is quite suitable for studies of dynamics as we can separately label different dynamic states for each site and list specific transitions between these states. Different lateral interactions of conformations can be treated as different states of each site in our linear system. For simplicity, we introduce only two states for each site in our present model. These two states can be thought of as the two values of a “spin” variable.

For nonequilibrium systems like these dynamical fibers, energy is consumed and equilibrium theories are inapplicable so we cannot rely on thermodynamics. Only dynamic rules are available at the coarse-grained level. While the usual methods of statistical thermodynamics are inapplicable, we can still apply a nonequilibrium action principle to approximate the steady states of this system since this method does not require the potential-like Hamiltonian. This approach also gives a simpler treatment for rare events than is provided by the more commonly used approaches based on the Fokker-Planck equation. The dynamic transitions between assembly and disassembly in microtubules have the typical features of rare events in chemical reaction systems. Furthermore, the variational scheme provides a simple presentation of an effective potential for the stationary state, which gives physical insight into the pathways for these dynamic events.

V.B Many-Body Master Equation for the Fiber End

For simplicity of modeling, we ignore the details of the helical structure of these fibers. We plan to discuss the cooperative effects related to lateral interactions in detail elsewhere. The rate constants are effectively changed with lateral cooperativity. It is quite straightforward to extend our approximations to treat such quasi-linear structures. Here, we study a one-dimensional lattice model [95]. Each occupied site can either have a monomer in an NDP state or in an NTP state labeled by $|0\rangle$ or $|1\rangle$ respectively. The state of the system can be expressed as a sequence $\{s_1, s_2, \ldots, s_N\}$ with $s_i = 0$ or 1. We focus on the dynamics of the fiber near one end, i.e., the system we study is in fact taken to be the last N sites of one fiber end. The rest of sites in the fiber are assumed to be in the NDP state. This latter assumption effectively serves as a boundary condition for the many-site Master equation. Hence in this analysis the structure of the system can be specified as a sequence of $(s_1, s_2, \ldots, s_N)$, for instance, (000101101...1), *etc.*

When new monomers arrive at the end or a monomer dissociates from the fiber's end, we shift the system accordingly and always keep track of the states of the last N sites.

The detailed dynamical rules for transitions of the fiber are listed in

Table V.1: elemental rules of the microscopic dynamics

process index	process	description	rate constant
c	$(1) \Rightarrow (0)$	NTP hydrolyzes to NDP	k_c
g	$(1) \Rightarrow (1,1)$	NTP monomer is added to NTP end	k_g
r	$(0) \Rightarrow (0,1)$	NTP monomer is added to NDP end	k_r
d	$(x,0) \Rightarrow (x)$	NDP end is removed, $x = 0, 1$	k_d
u	$(1) \Rightarrow (1,0)$	NDP monomer is added to NTP end	k_u
v	$(0) \Rightarrow (0,0)$	NDP monomer is added to NDP end	k_v
w	$(x,1) \Rightarrow (x)$	NTP end is removed, $x = 0, 1$	k_w

Table V.1. Although the last three processes (u, v and w) in Table V.1 are not of high probability, including them makes our model more general and allows the system to achieve equilibrium in special cases. For example, if k_r is set to be zero, we still can keep the equilibrium of the system when u or v are not zero. This also allows a check of the numerical studies and of our intuition.

The rate coefficients for the major processes are k_g, k_d, k_r, and k_c. Fiber growth is controlled microscopically by the process g called growth process. Fiber decay is controlled microscopically by the process d called decay process. The transition from decay to growth is microscopically controlled by the process r called rescue. These are microscopic events that should not be confused with their macroscopic results. The relative values of these rate coefficients control the state of the system. For convenience, we will set the k_c as our inverse time unit (unless we specify the value of k_c otherwise) and change the values of k_g, k_d, and k_r which are then dimensionless. As we demonstrate below, a combination of fast growing and fast decaying dynamics (large k_g and k_d) and rare rescue (small k_r) leads to the dynamic instability phenomenon as seen in real systems such as microtubules.

We use dynamic operators to describe the effect of these processes on the states of the fiber [96]. The change from the NDP state to the NTP state is controlled by creation operator $\hat{a}^+$. The change from the NTP state to the NDP state is controlled by annihilation operator $\hat{a}$. $|\Omega_0\rangle$ is denoted as the state of the fiber with every site in the NDP state. With these conventions a state vector of the fiber can be written as

$$|\phi_{s_1,s_2,\ldots,s_N}\rangle = \prod_{j=1}^{N}(\hat{a}_j^{+})^{s_j}|\Omega_0\rangle \tag{V.1}$$

Using these as basis vectors, we can write the state ensemble $|\psi\rangle$ as a linear combination

$$|\psi\rangle = \sum_{\{s\}=0,1} c_{s_1,s_2,\ldots,s_N}|\phi_{s_1,s_2,\ldots,s_N}\rangle \tag{V.2}$$

where $c_{s_1,s_2,\ldots,s_N}$ is the probability of finding the fiber at the $\phi_{s_1,s_2,\ldots,s_N}$ state.

The growth process and the decay process can both be expressed in the form of shift operators. The shift operators for the growth process are denoted as $\hat{b}^T$ and $\hat{b}^D$ for the association of an NTP monomer and an NDP monomer respectively. The shift operator for the decay process is denoted as $\hat{b}^S$. These shift operators can be explicitly written as functions of creation and annihilation operators.

$$\hat{b}^T|\phi_{s_1,s_2,\ldots,s_N}\rangle = \prod_{j=1}^{N-1}(\hat{a}_j^{+})^{s_{j+1}}\hat{a}_N^{+}|\Omega_0\rangle \tag{V.3}$$

$$\hat{b}^D|\phi_{s_1,s_2,\ldots,s_N}\rangle = \prod_{j=1}^{N-1}(\hat{a}_j^{+})^{s_{j+1}}|\Omega_0\rangle \tag{V.4}$$

$$\hat{b}^S|\phi_{s_1,s_2,\ldots,s_N}\rangle = \prod_{j=2}^{N}(\hat{a}_j^{+})^{s_{j-1}}|\Omega_0\rangle \tag{V.5}$$

The shift operators $\hat{b}^T$ ($\hat{b}^D$) represent the following joint dynamic events. When a monomer is added to the fiber, the operator $\hat{b}^T$ ($\hat{b}^D$) assigns the NTP (NDP) as the new state of the Nth site (the Nth site is the tip site of the fiber end). Simultaneously we shift the previous state of the ith site to the $(i-1)$th site $(i = 2, \ldots, N)$. The previous state of the first site will no longer be considered as part of the system. When dissociation of a unit from the tip occurs, the operator $\hat{b}^S$ will shift the previous state of the ith site to the $(i+1)$th site $(i = 1, \ldots, N-1)$. The state of the first site will now be assigned as the NDP state based on the boundary condition. The shifting operators intrinsically present the linear connectivity of the fiber and correlate the states between neighbour sites.

Given the above operator formulation, the evolution equation for the state of the last N sites of the fiber is written as

$$\frac{\partial}{\partial t}|\psi\rangle = -\hat{L}|\psi\rangle \tag{V.6}$$

where the rate operator $\hat{L}$ is:

$$\begin{aligned}\hat{L} = {} & k_c \sum_{i=1}^{N} (\hat{a}_i^+ \hat{a}_i - \hat{a}_i) + \\ & k_g(\hat{a}_N^+ \hat{a}_N - \hat{b}^T \hat{a}_N^+ \hat{a}_N) + k_r(\hat{a}_N \hat{a}_N^+ - \hat{b}^T \hat{a}_N \hat{a}_N^+) + k_d(\hat{a}_N \hat{a}_N^+ - \hat{b}^S \hat{a}_N \hat{a}_N^+) + \\ & k_u(\hat{a}_N^+ \hat{a}_N - \hat{b}^D \hat{a}_N^+ \hat{a}_N) + k_v(\hat{a}_N \hat{a}_N^+ - \hat{b}^D \hat{a}_N \hat{a}_N^+) + k_w(\hat{a}_N^+ \hat{a}_N - \hat{b}^S \hat{a}_N^+ \hat{a}_N)\end{aligned} \tag{V.7}$$

The operator form of the Master equation is given by Equation V.6. The first term in each bracket reduces the probability at one state while the second term increases by the same amount the probability at other states to keep the total probability conserved.

V.B.1 Exact Numerical Results for Finite Fiber End

In the following sections, we often set k_u, k_v and k_w to zero. As a result, the system does not satisfy detailed balance. Nevertheless we find only one stationary solution corresponding to the zero eigenvalue with the non-Hermitian and nondecomposable transfer matrix of the Master equation [97]. The other eigenvalues are transient modes whose amplitudes will decay to zero in the long time limit. For modest size systems we still can numerically diagonalize the matrix for a wide range of parameters.

For a system of size N, the number of the microscopic fiber states is 2^N, which is the linear size of the transfer matrix. The necessary computation resource increases exponentially with the system size. So it is difficult to obtain exact solutions for large systems. When the system has eight sites, the dimension of the transfer matrix is $2^8 \times 2^8$. The diagonalization of such a matrix is a simple

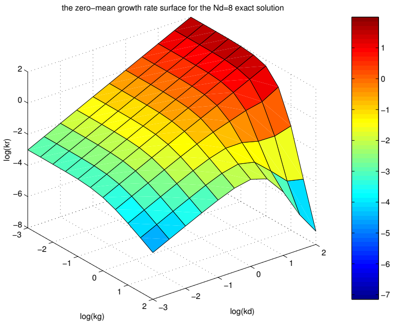

Figure V.1: The critical surface of vanishing mean velocity found using the exact solutions for a finite length of fiber end. The system size N is 8. The coordinates of the parameter space are the logarithms of k_g, k_r and k_d. The color bar indicates the value of $\log k_r$. k_c is set to be 1.0. Above the surface, the mean velocity is positive (growth) and below the surface, the mean velocity is negative (decay).

task. However, the diagonalization for a ten-site system is much more difficult. We present results for an eight-site system with the exact numerical method. We then use this exact solution to provide a test bed for approximation methods, which can then be used to solve larger systems.

We also calculate the mean growth velocity $\bar{v}_+$ (positive sign) or the decay velocity $\bar{v}_-$ (negative sign) by ensemble averaging the velocity using these exact solutions, where $\bar{v}_+ = (k_g \hat{a}_N^+ \hat{a}_N + k_r \hat{a}_N \hat{a}_N^+)|\psi\rangle$ and $\bar{v}_- = k_d \hat{a}_N \hat{a}_N^+ |\psi\rangle$. In Figure V.1 we plot the critical surface of zero mean velocity for a range of values of parameters. In this phase diagram, beyond the critical surface, the fiber grows on average; on the other side of the critical surface, the fiber decays on average. The critical surface shows how the rate constants k_g, k_r, and k_d influence the average growth or decay rate of the fiber. When k_g is large, the value of k_r is small on the critical surface. So a small change of the critical k_r can switch the sign of the mean velocity.

V.B.2 Mean Field Approximation of the Trial Function Ψ_A

The exact solution described above is limited by the size of the system. To treat a larger system, a more computationally efficient method is required. In the following two sections, we study a simple approach based on a variational principle for nonequilibrium systems [26]. Two trial functions are proposed for use with the variational method. The procedure for setting trial function Ψ_A is given in this section and the procedure for an improved trial function Ψ_B is given in next section.

A variational method for nonequilibrium systems can be used much like Rayleigh-Ritz variational method in quantum mechanics. In contrast to the Rayleigh-Ritz method for hermitian problems, not one but two trial vectors have to be constructed. For the nonequilibrium system, the right vector is related to the probability distribution of the system, while the left vector is an auxiliary vector needed for the variational procedure. In the trial function Ψ_A, first we introduce a set of parameters $\{\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_N\}$ to denote the NTP probabilities at each site. The auxiliary parameters are then given as $\{\alpha_1^L, \alpha_2^L, \alpha_3^L, \ldots, \alpha_N^L\}$. Using these parameters, we construct the left and right state vectors of our *ansatz* as

$$\begin{aligned} \langle \psi^L(\alpha^L)| &= \langle\Omega|(1 + \sum_{i=1}^{N} \alpha_i^L \hat{s}_i) \quad &\text{(V.8)} \\ |\psi(\alpha)\rangle &= \prod_{i=1}^{N} [\alpha_i \hat{s}_i + (1-\alpha_i)(1-\hat{s}_i)]|\Omega\rangle \quad &\text{(V.9)} \end{aligned}$$

Here $|\Omega\rangle = \sum_{\{s\}} |\phi_{s1,s2,\ldots,s_N}\rangle$ and $\hat{s}_i = \hat{a}_i^+ \hat{a}_i$ [98]. Note that the vector $|\psi\rangle$ characterizes the only probability distribution of the system based on the parameters $\{\boldsymbol{\alpha}\}$.

Given the rate operator described above, we write out the effective action as $\langle \psi^L, \hat{L}\psi(\alpha)\rangle$ [26]. In fact, the variation of this effective action yields the moment closure equations, *i.e.* $\sum_j \frac{\partial m_i(\boldsymbol{\alpha})}{\partial \alpha_j}\dot{\alpha}_j = V_i(\boldsymbol{\alpha})$, where $m_i(\boldsymbol{\alpha}) = \langle\Omega|\hat{s}_i|\psi(\boldsymbol{\alpha})\rangle$ and $V_i(\boldsymbol{\alpha}) = \left\langle \Omega|\hat{s}_i\hat{L}|\psi(\boldsymbol{\alpha})\right\rangle$. The dynamics of each moment of the distribution is listed as follows:

Site N:

$$-\dot{\alpha}_N = k_c\alpha_N - k_d(1-\alpha_N)\alpha_{N-1} - k_r(1-\alpha_N) + k_u\alpha_N + k_w\alpha_N(1-\alpha_{N-1})$$

Site N-1:

$$-\dot{\alpha}_{N-1} = k_c\alpha_{N-1} - (k_g + k_u)\alpha_N(1-\alpha_{N-1}) + (k_r + k_v)(1-\alpha_N)\alpha_{N-1} +$$

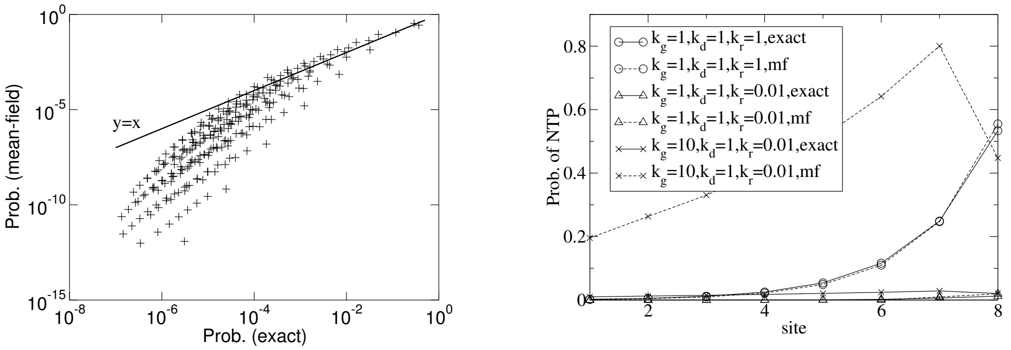

Figure V.2: Left Panel: The correlation of microscopic state probabilities between the exact solutions and the solutions of the trial function Ψ_A. There are total $2^8 = 256$ states for the eight-site system. The rate parameters are $k_c = k_g = k_d = k_r = 1.0$. Right Panel: The comparison of the NTP probability of each site between the exact solutions and the solutions of the trial function Ψ_A at several parameter sets is shown.

$$k_d(1-\alpha_N)(\alpha_{N-1}-\alpha_{N-2}) + k_w\alpha_N(\alpha_{N-1}-\alpha_{N-2})$$

Site $\neq$ 1,N-1,N:

$$\begin{aligned} -\dot{\alpha}_i &= k_c\alpha_i + (k_g+k_u)\alpha_N(\alpha_i-\alpha_{i+1}) + (k_r+k_v)(1-\alpha_N)(\alpha_i-\alpha_{i+1}) + \\ & \quad k_d(1-\alpha_N)(\alpha_i-\alpha_{i-1}) + k_w\alpha_N(\alpha_i-\alpha_{i-1}) \end{aligned}$$

Site 1:

$$\begin{aligned} -\dot{\alpha}_1 &= k_c\alpha_1 + (k_g+k_u)\alpha_N(\alpha_1-\alpha_2) + (k_r+k_v)(1-\alpha_N)(\alpha_1-\alpha_2) + \\ & \quad k_d(1-\alpha_N)\alpha_1 + k_w\alpha_N\alpha_1 \end{aligned}$$

The moment equation array has an asymmetric form. Site N and site $N-1$ are shown differently because these sites are directly coupled to the growth, decay, and rescue processes. Site 1 also has a special form due to the boundary condition.

In Figure V.2, we compare the results from the exact solutions with the results of the first trial function. For a range of parameters of k_g, k_d, and k_r, the two solutions agree well with each other. However, for the parameters with relative large k_g and k_d but small k_r, we have found that this trial function breaks down when compared to the exact solution. This breaking down, much as other mean-field treatments, fails because the correlation between the sites becomes important and large fluctuations emerge in the dynamics. The large fluctuation signals the onset of two different long-lived dynamic states: one of which is fast growing and the other fast decaying. In the above comparisons, we use the tip site's NTP probability to judge whether two solutions agree or not. We choose this property because the tip site responds quickly to the fluctuation and influences other sites of the fiber. We have set the criterion for "agreement" as follows: we calculate the relative difference of the tip site's NTP probability between the solutions of Ψ_A and the exact solutions. If the relative difference is less than 5%, we regard the solution of Ψ_A as agreeing with the exact solution. In Figure V.3, k_r is scanned for two-dimensional parameter space ($\log k_g$, $\log k_d$). The critical value of k_r is identified by the minimum k_r giving the consistent solutions by the above criterion. We summarize these critical values in Figure V.3.

V.B.3 The Tip-Site Dependent Trial Function Ψ_B

In the above formulation, the coupling between different sites is included in the shift operators in the processes of growth, rescue, and decay. However, we found that the NTP state probabilities of a growing fiber can be quiet different from those of a decaying fiber in the exact solution as shown in Figure V.2(b). It is important to introduce the bistable character of growth and decay dynamic to improve the trial functions. We also know that the tip site (site N) directly influences the ongoing dynamics. Thus we must pay special attention to the Nth site's state in the dynamics in the improved *ansatz* Ψ_B.

First, we denote the NTP probability of the Nth site by the parameter γ. Thus the Nth site has the probability of γ to be in the NTP state and the probability of $1 - \gamma$ to be in the NDP state. More importantly, the fiber will generally access different processes for different states of the Nth site. Thus we

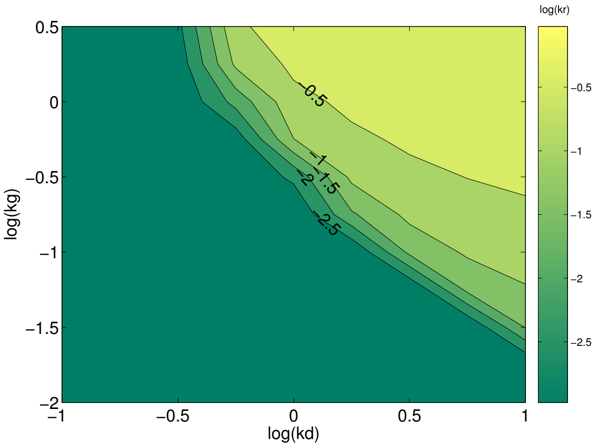

Figure V.3: The comparison between the solutions of the trial function Ψ_A with the exact solutions for the eight-site system. The x axis and the y axis are $\log k_d$ and $\log k_g$ respectively. For each point $(\log k_d, \log k_g)$ on the surface, we scan $\log k_r$. The color indicates the critical value of $\log k_r$, above which the trial function Ψ_A works well and below which Ψ_A fails by the judgment defined above: the relative difference of the tip site's NTP probability between the solutions of Ψ_A and the exact solutions less than 5%. When the critical value of $\log k_r$ is less than -3.0, we only show the value of -3.0 in the figure.

should have two sets of NTP probabilities $\{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$ for the sites $1, 2, \ldots, N-1$ based on the state of the Nth site. When the Nth site is in the NTP state, we denote the NTP state probabilities of sites $(1, 2, \ldots, N-1)$ as $(\alpha_1, \alpha_2, \ldots, \alpha_{N-1})$. When the Nth site is in the NDP state, we denote the NTP state probabilities of sites $(1, 2, \ldots, N-1)$ as $(\beta_1, \beta_2, \ldots, \beta_{N-1})$. We can see that the two-state dynamic features are present in this formulation. Compared to the total of N degrees of freedom in the trial function Ψ_A, the trial function Ψ_B has $2N-1$ independent state variables. It is much like a configurational interaction (CI) wave function in quantum chemistry.

Using the same variational strategy described in the previous section, the *ansatz* Ψ_B is now written as follows:

$$\langle\psi^L(\alpha^L)|L = \langle\Omega|(1 + \gamma^L \hat{s}_N + \sum_{i=1}^{N-1} \alpha_i^L \hat{s}_i \delta_{s_N,0} + \sum_{i=1}^{N-1} \beta_i^L \hat{s}_i \delta_{s_N,1}) \qquad \text{(V.10)}$$

$$
\begin{aligned}
L|\psi(\alpha)\rangle \;\; &= \;\; (1-\gamma)\prod_{i=1}^{N-1}[\alpha_i\hat{s}_i+(1-\alpha_i)(1-\hat{s}_i)]\delta_{s_N,0}|\Omega\rangle + \\
&\quad\;\; \gamma\prod_{i=1}^{N-1}[\beta_i\hat{s}_i+(1-\beta_i)(1-\hat{s}_i)]\delta_{s_N,1}|\Omega\rangle \qquad \text{(V.11)}
\end{aligned}
$$

Again the left eigenvector yields a moment closure of $\{\alpha_i\}$, $\{\beta_i\}$, and γ. The following equations (V.12) are derived from the variations of $\{\alpha_i^L\}$, $\{\beta_i^L\}$, and γ^L. Note that $\alpha_0 = 0$, $\alpha_N = 0$, $\beta_0 = 0$, and $\beta_N = 1$.

$$
\begin{aligned}
-(1-\gamma)\dot{\alpha}_i \;\; &= \;\; k_c(\alpha_i-\gamma\beta_i)+k_d(1-\gamma)(1-\alpha_{N-1})(\alpha_i-\alpha_{i-1})+k_u\gamma(\alpha_i-\beta_{i+1})+ \\
&\quad\;\; k_v(1-\gamma)(\alpha_i-\alpha_{i+1})+k_w\gamma(1-\beta_{N-1})(\alpha_i-\beta_{i-1}) \qquad \text{(V.12)} \\
\gamma\dot{\beta}_i \;\; &= \;\; k_c\gamma\beta_i+k_g\gamma(\beta_i-\beta_{i+1})+k_d(1-\gamma)\alpha_{N-1}(\beta_i-\alpha_{i-1})+ \\
&\quad\;\; k_r(1-\gamma)(\beta_i-\alpha_{i+1})+k_w\gamma\beta_{N-1}(\beta_i-\beta_{i-1}) \qquad \text{(V.13)} \\
\dot{\gamma} &= -k_c\gamma+k_d(1-\gamma)\alpha_{N-1}+k_r(1-\gamma)-k_u\gamma-k_w\gamma(1-\beta_{N-1}) \qquad \text{(V.14)}
\end{aligned}
$$

In Figure V.4, we compare the solutions of Ψ_B with the exact solutions for the parameter sets where the trial function Ψ_A failed. The solutions of the trial function Ψ_B agree well with the exact solutions.

Furthermore, we notice that the NTP and the NDP states at site N are followed by different NTP state probabilities for the sites $(1,2,\ldots,N-1)$. When the Nth site is in the NTP state, sites $(1,2,\ldots,N-1)$ have high probabilities of being in the NTP state. When the Nth site is in the NDP state, sites $(1,2,\ldots,N-1)$ have very small probabilities of being in the NTP state. So the two sets of state probability distributions are coupled with different states of the Nth site.

We also scanned the solutions for the trial functions Ψ_B in the same parameter region as shown in Figure V.3 and compared these with the exact solutions. The trial function Ψ_B works much better than the trial function Ψ_A. The solutions of Ψ_B satisfies the same criterion for Ψ_A in the whole scanned region including the regions for which Ψ_A failed.

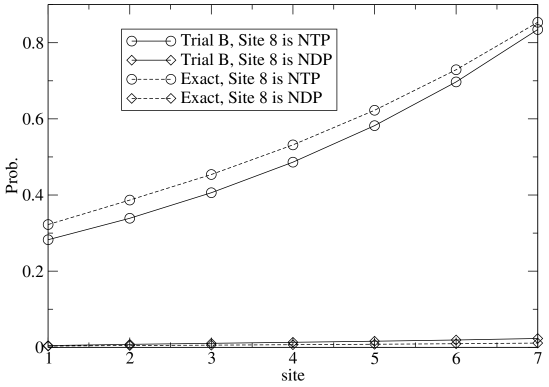

Figure V.4: The comparison between the solution of the trial function Ψ_B and the exact results. The rate parameters are $k_c = 1.0$, $k_g = 10.0$, $k_d = 1.0$, and $k_r = 0.01$ respectively. Bold lines are the solution from trial function Ψ_B; dash lines are the solution from the exact method. The NTP probabilities of sites $(1, 2, \ldots, N-1)$ when site N is NDP are labeled with diamond; the corresponding NTP probabilities when site N is NTP are labeled with circle. The NDP probability of site N is 0.979 in the exact solution and 0.968 in the solution of the trial function Ψ_B. It is clear that when the Nth site is NDP, the rest sites of the fiber have very small probabilities of being NTP.

V.C Dynamical Barriers and Transition Rates

Considering the large number of microscopic states and the complicated transitions between the steady states, one approach is to use an effective reaction coordinate to project the dynamics of the many-site states. The choice of reaction coordinate for such a complex dynamic system can be tricky. The parameter γ in the trial function Ψ_B provides one choice. However, it is not a very good indicator of the state of the chain. This can be easily seen: for example, the fiber state $(11\ldots1110)$ is most likely in the growth state because all sites except the very tip are in the NTP state and the zero at the tip can be dropped easily. We need a collective reaction coordinate to describe the overall features of the fiber's state. An obvious choice is to count the number of sites in the NTP state. With such a description method much detailed information about the site positions is ignored, but the counting in any event is simple and straightforward for display. In this section, we will therefore use the number of sites in the NTP state as the reaction coordinate. Note that the density of states will be not be uniform along this reaction coordinate.

The probability of every microscopic fiber state can be calculated from the stationary results of the trial function Ψ_B. We simply multiply the state probability of each site in the fiber to give the probability of the corresponding fiber state. The fiber states with same number of NTPs belong to the same point on our reaction coordinate. The probabilities of these fiber states are summed to give the probability of the reaction coordinate. After the probability distribution is given, we can define an effective potential through the negative logarithm of these probabilities.

Before showing the detailed results, we again discuss the issue of the boundary condition. It is important to use a sufficiently large N so as not to bring in strong boundary effects. In Figure V.5, we show the change of the effective potentials as we increase the system size N. When the system size increases, the

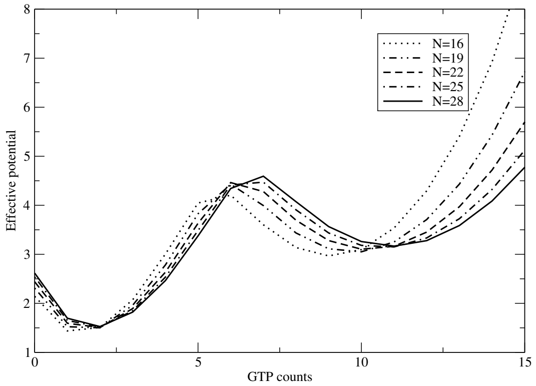

Figure V.5: The boundary effect on the shape of the effective potential. The rate parameters are $k_c = 1.0$, $k_g = 20.0$, $k_d = 1.0$, and $k_r = 0.1$ respectively. The x coordinate is the NTP count. NTP count expresses how many sites are in NTP state in our system, which is exactly our reaction coordinates of effective potentials. The y coordinate is the effective potential. The trend of the curves indicates the convergence of the effective potential.

effective potential converges to an ideal potential with no boundary effects. Based on this comparison, a reasonable system size N can be chosen.

Now we start to investigate a 26-site system and the boundary condition has already been tested in each case. We study how k_g and k_d influence the effective potential in Figure V.6. The rate k_r is taken to be one or two orders less than the growth rate or decay rate as indicated in experiments.

In Figure V.6 (Left Panel), we keep k_c, k_g, and k_r as constants and change the value of k_d. We see that only when k_d is small enough can two basins exist at the same time. We term these the NDP-populated basin and the NTP-populated basin. In Figure V.6 (Right Panel), we keep k_c, k_d, and k_r as constants and change the value of k_g. We now see that when we decrease k_g, the potential starts to tilt toward the origin. So the fiber state with all sites in NDP states becomes dominant. We also note that when k_g is larger than k_d (see the left panel of Figure V.6), there are many sites of the fiber now in the NTP state; when k_d is larger than k_g (see the right panel of Figure V.6), there are only a small number of sites in the NTP

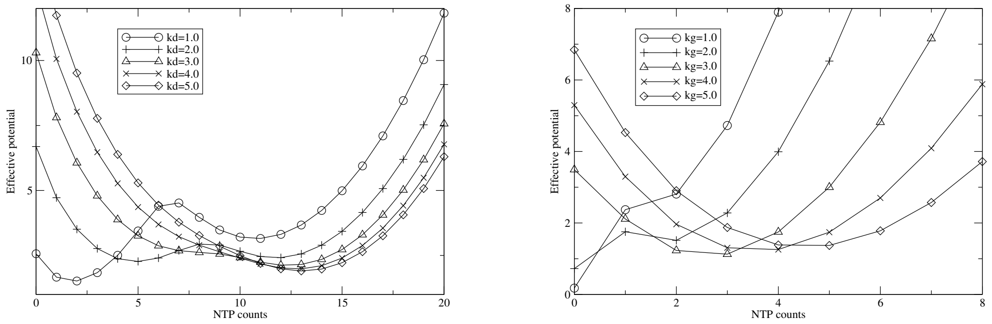

Figure V.6: The effective potentials of a 26-site system in different parameter sets are shown as functions of the NTP count. Left: k_d varies and $k_c = 1.0$, $k_g = 20.0$, $k_r = 0.1$. Right: k_g varies and $k_c = 1.0$, $k_d = 20.0$, $k_r = 0.1$.

state.

We can also calculate the mean cap length from the ensemble of microscopic states. Here, the cap length describes the number of the terminal consecutive NTPs. In Figure V.7, the cap length is plotted as a function of the effective reaction coordinate. A sharp increment of the cap length indicates the appearance of the NTP cap and indicates the transition from decay to growth of the fibers. Furthermore, we observed that at the most probable state of the NTP-populated basin, not all the sites in the NTP state belong to the cap. Instead, many sites that are disconnected from the cap region still have NTPs bound. Apparently these states are entropically favored over the fiber state with all NTPs in the cap.

After we project the fiber states into the simple effective reaction coordinate, the complex dynamics of the fiber can be approximated by the dynamics of this effective potential. We first present the discrete effective potential numerically. Then, we approximate the effective diffusion coefficient as

$$D_i = 0.5 * [k_c * i + k_g * \frac{\binom{N-2}{i-1}}{\binom{N}{i}} + k_r * \frac{\binom{N-2}{i}}{\binom{N}{i}}] \qquad \text{(V.15)}$$

The subscript i describes the position on the effective reaction coordinate and $\binom{N}{M} = \frac{N!}{M! \times (N-M)!}$. This effective diffusion coefficient is due to the chemical reaction

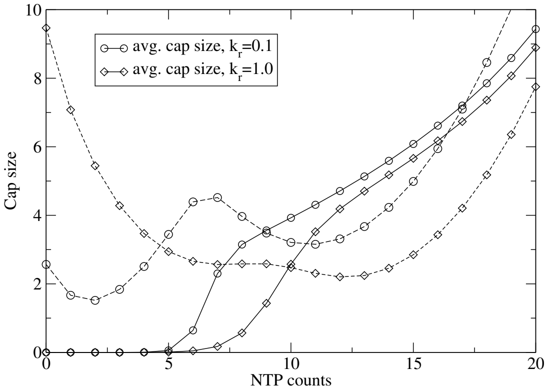

Figure V.7: The mean cap length for two sets of parameters. The rate parameters are $k_c = 1.0$, $k_g = 20.0$, and $k_d = 1.0$ respectively. k_r is indicated in the figure. The system size is 26. The mean cap length is plotted as the solid lines. For comparison, the corresponding effective potentials are plotted with dash lines.

noise. We see that the diffusion coefficient does not depend on the k_d, because the decay of NDP at the tip does not change the amount of NTP in the system. Moreover, the diffusion coefficients are not homogeneous, as shown in Figure V.8.

The effective potential may present only one basin. The dynamics on this type of effective potential is simply a relaxation process. However, rare "thermal" jumps to high positions on the potential can cause the fiber to behave differently. Further discussion of fiber dynamics under a single-basin potential is detailed in the next section.

For the effective potential with dual basins, transitions between these two states become the essential dynamic feature of the system. The jumping rate from the NDP-populated basin to the NTP-populated basin corresponds to the macroscopic rescue rate and the jumping rate of the macroscopic reverse process corresponds to the macroscopic catastrophe occurrence rate. In this case, a one-dimensional Smoluchowski equation [97] can be solved numerically to estimate the jumping rates between the NDP-populated basin and NTP-populated basin. We estimate the jumping rate from one basin to the other assuming it is a Poisson

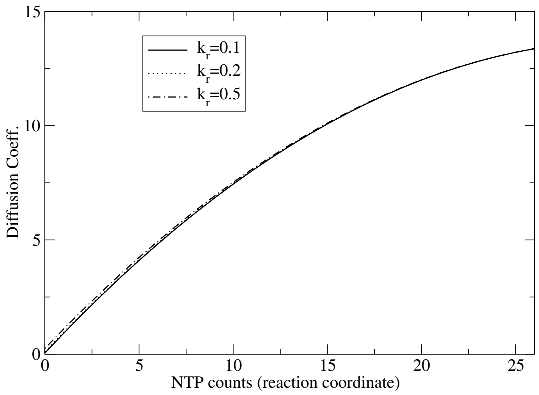

Figure V.8: The diffusion coefficient defined by equation V.15 is shown as a function of the NTP counts. The rate parameters: $k_c = 1.0$, $k_g = 20.0$, and $k_d = 1.0$.

Table V.2: Compare the jump rate with the slowest mode in exact solution

rate parameters	the modulus of slowest mode	jumping rate
$k_c = 0.1, k_g = 2.0, k_d = 0.1, k_r = 0.05$	0.074	0.050
$k_c = 0.1, k_g = 1.0, k_d = 0.1, k_r = 0.02$	0.047	0.024

process.

We then can compare the effective potential-derived jumping rate in the eight-site system with the modulus of the slowest transient mode from the exact solution. The rate coefficients k_c, k_g, k_r, and k_d are given in the first column of Table V.2. The double basin potential is presented for both parameter sets. We note that the modulus of the slowest transient mode is not exactly equivalent to the jumping rates between two basins. However, they should be on the same time scale because the jumping between two basins is the major process of the non-zero transient modes. Table V.2 shows that both rates are at the scale of 10^{-2}. From this comparison, the effective potential treatment of nonequilibrium states seems able to capture the characteristic time-scale of the dynamics in our model.

V.D Relation between the Catastrophe/Rescue Rate and the Growth Rate

In microtubule systems, dynamic instability has been observed. The switching between the two dynamic states occurs on the time scale of minutes. Both the growth and the decay are much faster than this switching frequency. Experiments [33, 35] also show that the decay rate is generally 10 to 100 fold faster than the growth rate. The hydrolysis rate has been estimated to be on the same time scale as the growth rate [99, 100, 101, 102]. The microscopic rescue rate is estimated to be smaller than the hydrolysis rate by a factor of 10 to 100 fold. We may set the dimensionless rate constants consistent with these estimated ratios and set the hydrolysis rate k_c to be 1.0. The decay rate k_d is set to be 20.0 and the microscopic rescue rate k_r is set to be 0.1. Finally the catastrophe rate and rescue rate are estimated based on the effective potential.

We have already plotted the effective potentials with this set of the parameters in Figure V.6 (Right Panel). We observe that when k_g is 1.0, the effective potential has only one minimum at zero NTP counts. When k_g increases to 2.0, the second minimum appears. When k_g becomes larger than 2.0, the new basin becomes the only one on the effective potential. This shift of the minimum followed by increasing k_g value shows the dynamic transition between decay and growth. The critical value of k_g is between 1.0 and 2.0.

When k_g increases to 3.0, the single minimum is located in the growth region as shown in Figure V.6 (Right Panel). We can calculate the mean growth velocity for each position of the reaction coordinate. Based on the sign of mean growth velocity, we thus separate the coordinate space into growth and decay regions. The jumps from the positive growth minimum to reach the negative mean growth velocity region can be regarded as catastrophe events. The catastrophe involves climbing the barrier of the effective potential. In Figure V.9, we plot the jumping rate to reach the negative mean velocity region versus k_g. The catastrophe

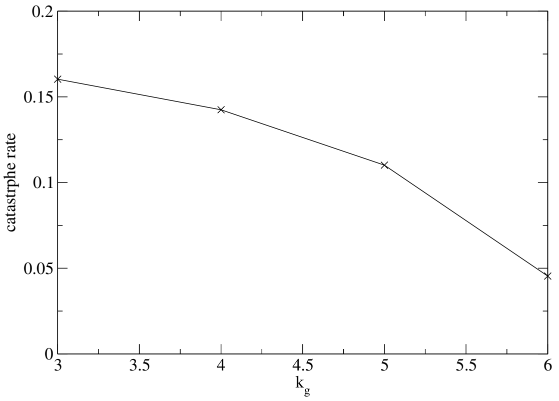

Figure V.9: The catastrophe rate as a function of k_g. Again the system size is 26.

rate is approximately linearly proportional to the growth rate. This result agrees with experimental observations on microtubule systems [99].

V.E Conclusion and Outlook

We present a discrete linear model to study the nonequilibrium assembly of biological fibers. The model's focus is on the the first N sites of the growing dynamic fiber. By tracing the paths forming the microscopic structures of this dynamic region, the model captures the macroscopic kinetics of the system.

In nonequilibrium fiber systems, we cannot use equilibrium statistical thermodynamics. Only dynamic rules are available. The evolution of the classical system can be expressed in an operator formulation in analogy to quantum field theory. However, the 'Hamiltonian' $\hat{L}$ is non-Hermitian for our system. We applied the variational method for non-Hermitian dissipative systems introduced by Eyink [26] to study this system using different trial functions. The current study mainly focused on the stationary solutions. A dynamic analysis based on the minimization of so-called effective action is an interesting prospect for future study.

An "effective potential" is used to connect the steady state probability distributions with the rate of rare events. This projection of state distributions into simple potentials can be quite useful for studying far-from-equilibrium systems in general.

Using reasonable ratios between rate coefficients according to microtubule experiments, the model captures some phenomenological features of microtubule assembly. The present model may be adapted to study other one-dimensional fiber assembly processes and to study assembly in the more complicated environment in which interactions with other fibers must be included. One may introduce multiple species to study the regulation of the heterogeneous assembly of the cytoskeleton. In particular a complete model will need to introduce other microtubule associated proteins (MAPs) implicitly or explicitly to mimic real cells. Many improvements are needed to extend the studies to the cytoskeleton, but the mathematical methods we have introduced should make such an extension practical. Other quasi-linear systems that include lateral interactions are also being investigated.

V.F Acknowledgments

This chapter is based on the material as it appears in Physical Biology, 3: 83-92 (2006), C. Zong; T. Lu; T. Shen; P. G. Wolynes. The dissertation author was the primary investigator and author of this paper.

VI

Final Remarks

To understand biological processes at molecular level, we need to characterize not only structures of individual molecules, but also the dynamics at supramolecular level. Accurately predicting protein structures from primary sequences was the first goal, that has led us into the postgenomic era. While the conceptual bottlenecks have been overcome, in order to routinely be successful, we still will need better models and faster algorithms capable of dealing with energy calculation of diverse conformations of heterogeneous biomolecules. Beyond the level of individual molecules, the next step in postgenomic theory is to characterize how biomolecules interact with each other, to characterize which processes are functional and which only contribute to the noise. Even though biomolecules usually exist at equilibrium, biological functions are often achieved when the systems are perturbed from equilibrium into nonequilibrium. Studying nonequilibrium structure formation will be another major task in biophysics.

In this thesis, I first discussed protein folding problems and near equilibrium energy landscape theory. This framework essentially present a picture of the thermodynamic states for proteins that follows from the funneled energy landscape. Such a global landscape may be characterized quantitatively by T_F and T_G whose relation encodes the minimal frustration principle, requiring essentially that $T_F > T_G$ for natural proteins. This principle was integrated into the polymer model used in Chapter 2 and 3. Free energy profiles were then found from this landscape via variational free energy functional.

In chapter 4, I presented studies using energy landscape ideas for predicting protein structures. A coarse-grained model implicitly modeling water is introduced to treat solvation and water mediated interactions. Topological preference is introduced in β-strand modeling. With this coarse-grained model, samplings of energy landscape can be performed efficiently. Additional sources of cooperativity may be needed to achieve a universally predictive model. Many-body terms may allow the native interactions to be much more favored compared to wrong interactions than pairwise models allow. The extra stabilization from cooperative many-body interactions does not have to be very big compared to pairwise energy in order to more precisely form the native state. The correct cooperativity will not only stabilize the native structure, but also should destabilize many of the wrong traps.

Besides complexity intrinsic to information-carrying macromolecules, the other characteristic of biological systems is that they are often far from equilibrium. Chemical energy is consumed to realize biological functions, leading to the nonequilibrium state. On the other hand, fluctuations should not be ignored at the scale of biomolecular machines. Sometimes fluctuation are critical to functionality. In chapter 5, I presented a study using a nonequilibrium variational principle to study cellular assembly processes, focusing on the nonequilibrium growth/decay of one-dimensional fiber. A dynamic network of many such fibers provides the mechanical support for the cellular compartments. The extension of this approach to the network level is a task in the future.

Bibliography

[1] C. Levinthal, *Mossbauer Spectroscopy in Biological System*, edited by P. Debrunner, J. Tsibris, and E. Munck; University of Illinois Press, Urbana, IL, 1969.

[2] J. D. Bryngelson and P. G. Wolynes, Proc. Natl. Acad. Sci. USA **84**, 7524 (1987).

[3] P. E. Leopold, M. Montal, and J. N. Onuchic, PNAS **89**, 8721 (1992).

[4] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Proteins **21**, 167 (1995).

[5] J. N. Onuchic, Z. L. Schulten, and P. G. Wolynes, Annual Review of Physical Chemistry **48**, 545 (1997).

[6] T. R. Kirkpatrick and P. G. Wolynes, Phys. Rev. B. **36**, 8552 (1987).

[7] T. R. Kirkpatrick and D. Thirumalai, Phys. Rev. B. **36**, 5388 (1987).

[8] M. Mezard and G. Parisi, J. Chem. Phys. **111**, 1076 (1999).

[9] C. H. Zong, C. J. Wilson, T. Shen, P. G. Wolynes, and P. Wittung-Stafshede, biochemistry **45**, 6458 (2006).

[10] J. J. Portman, S. Takada, and P. G. Wolynes, J. Chem. Phys. **114**, 5069 (2001).

[11] J. J. Portman, S. Takada, and P. G. Wolynes, J. Chem. Phys. **114**, 5082 (2001).

[12] C. J. Wilson and P. Wittung-Stafshede, PNAS **102**, 3984 (2004).

[13] B. L. Vallee and R. J. P. Williams, PNAS **59**, 498 (1968).

[14] C. H. Zong et al., PNAS **104**, 3159 (2007).

[15] C. J. Wilson and P. Wittung-Stafshede, Biochemistry **44**, 10054 (2005).

[16] M. S. Friedrichs and P. G. Wolynes, Science **246**, 371 (1989).

[17] M. S. Friedrichs and P. G. Wolynes, Tet. Comp. Meth. **3**, 175 (1990).

[18] M. S. Friedrichs, R. A. Goldstein, and P. G. Wolynes, J. Mol. Biol. **222**, 1013 (1991).

[19] M. Sasai and P. G. Wolynes, Phys. Rev. A **46**, 7979 (1992).

[20] C. Hardin, M. P. Eastwood, Z. Luthey-Schulten, and P. G. Wolynes, Proc. Natl. Acad. Sci. USA **97**, 14235 (2000).

[21] M. P. Eastwood, C. Hardin, Z. Luthey-Schulten, and P. G. Wolynes, IBM J. Res. & Dev. **45**, 475 (2001).

[22] C. Hardin, M. P. Eastwood, M. C. Prentiss, Z. Luthey-Schulten, and P. G. Wolynes, Proc. Natl. Acad. Sci. USA **100**, 1679 (2003).

[23] G. A. Papoian, J. Ulander, M. P. Eastwood, and P. G. Wolynes, Proc. Natl. Acad. Sci. USA **101**, 3352 (2004).

[24] C. H. Zong, G. A. Papoian, J. Ulander, and P. G. Wolynes, JACS **128**, 5168 (2006).

[25] C. H. Zong, T. Lu, T. Shen, and P. G. Wolynes, Physical Biology **3**, 83 (2006).

[26] G. L. Eyink, Phys. Rev. E **54**, 3419 (1996).

[27] T. D. Pollard and W. C. Earnshaw, *Cell biology*, W.B. Saunders, New York, 2002.

[28] J. Howard, *Mechanics of Motor Proteins and the Cytoskeleton*, Sinauer Assoc, Sunderland, Mass., 2001.

[29] A. D. Bershadsky and J. M. Vasiliev, *Cytoskeleton*, Cellular organelles, Plenum Press, New York, 1988.

[30] T. L. Hill, *Linear aggregation theory in cell biology*, Springer-Verlag, New York, 1987.

[31] A. Mogilner and G. Oster, Current Biology **13**, 721 (2003).

[32] H. Erickson and E. T. O'brien, Annu. Rev. Biophys. Biomol. Struct. **21**, 145 (1992).

[33] T. Mitchison and M. Kirschner, Nature **312**, 232 (1984).

[34] T. Mitchison and M. Kirschner, Nature **312**, 237 (1984).

[35] T. Horio and H. Hotani, Science **321**, 605 (1986).

[36] T. L. Hill, Proc. Natl. Acad. Sci. USA **81**, 6728 (1984).

[37] D. K. Fygenson, E. Braun, and A. Libchaber, Phys. Rev. E **50**, 1579 (1994).

[38] V. Vanburen, D. Odde, and L. Cassimeris, Proc. Natl. Acad. Sci. USA **99**, 6035 (2002).

[39] S. Miyazawa and R. L. Jernigan, J. Mol. Biol. **256**, 623 (1996).

[40] H. F. Engeseth and D. R. McMillin, Biochemistry **25**, 2448 (1986).

[41] T. Shen, C. P. Hofmann, M. Oliveberg, and P. G. Wolynes, Biochemistry **44**, 6433 (2005).

[42] I. E. Sanchez and T. Kiefhaber, J. Mol. Biol. **334**, 1077 (2003).

[43] N. D. Socci, J. N. Onuchic, and P. G. Wolynes, J. Chem. Phys. **104**, 5860 (1996).

[44] C. L. Lee, G. Stell, and J. Wang, J. Chem. Phys. **118**, 959 (2003).

[45] I. Pozdnyakova and P. Wittung-Stafshede, Biochemistry **40**, 13728 (2001).

[46] I. Pozdnyakova, J. Guidry, and P. Wittung-Stafshede, Biophys. J. **82**, 2645 (2002).

[47] R. J. P. Williams, European Journal of Biochemistry **234**, 363 (1995).

[48] B. G. Malmstrom, European Journal of Biochemistry **223**, 711 (1994).

[49] H. B. Gray, B. G. Malmstrom, and R. J. P. Williams, Journal of Biological Inorganic Chemistry **5**, 551 (2000).

[50] C. Dennison, Coordination Chemistry Reviews **249**, 3025 (2005).

[51] S. Yanagisawa, M. J. Banfield, and C. Dennison, Biochemistry **45**, 8812 (2006).

[52] E. I. Solomon, R. K. Szilagyi, S. D. George, and L. Basumallick, Chemical Reviews **104**, 419 (2004).

[53] U. Ryde, M. H. M. Olsson, K. Pierloot, and B. O. Roos, Journal of Molecular Biology **261**, 586 (1996).

[54] U. Ryde and M. H. M. Olsson, International Journal of Quantum Chemistry **81**, 335 (2001).

[55] J. R. Winkler, P. WittungStafshede, J. Leckner, B. G. Malmstrom, and H. B. Gray, PNAS **94**, 4246 (1997).

[56] A. J. DiBilio et al., JACS **119**, 9921 (1997).

[57] L. K. S. nad T. Pascher, J. R. Winkler, and H. B. Gray, JACS **98**, 1102 (1998).

[58] E. T. Adman, Advances in Protein Chemistry **42**, 145 (1991).

[59] J. Marks, I. Pozdnyakova, J. Guidry, and P. Wittung-Stafshede, Journal of Biological Inorganic Chemistry **9**, 281 (2004).

[60] J. Leckner, P. Wittung, N. Bonander, B. G. Karlsson, and B. G. Malmstrom, Journal of Biological Inorganic Chemistry **9**, 368 (1997).

[61] H. Nar et al., European Journal of Biochemistry **20**, 1123 (1992).

[62] A. M. amd J. T. Kellis, L. Serrano, M. Bycroft, and A. R. Fersht, Nature **346**, 440 (1990).

[63] A. Matouschek and A. R. Fersht, Methods in Enzymology **202**, 82 (1991).

[64] H. Frauenfelder, S. G. Sligar, and P. G. Wolynes, Science **254**, 1598 (1987).

[65] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Proteins-Structure Function and Genetics **21**, 167 (1995).

[66] H. Nar, A. Messerschmidt, R. Huber, M. Vandekamp, and G. W. Canters, Journal of Molecular Biology **221**, 765 (1991).

[67] B. R. Crane, A. J. D. Bilio, J. R. Winkler, and H. B. Gray, JACS **123**, 11623 (2001).

[68] S. DeBeer et al., Inorganica Chimica Acta **297**, 278 (2000).

[69] I. Pozdnyakova, J. Guidry, and P. Wittung-Stafshede, Journal of Biological Inorganic Chemistry **6**, 182 (2001).

[70] M. Bixon and R. Zwanzig, Journal of Chemical Physics **68**, 1896 (1978).

[71] J. Ryckaert, G. Ciccotti, and H. Berendsen, J. Comput. Phys. **23**, 327 (1977).

[72] A. J. Cuff, E. M. Clamp, S. A. Siddiqui, M. Finlay, and G. J. Barton, Bioinformatics **14**, 892 (1998).

[73] V. Munoz, P. Thompson, J. Hofrichter, and W. Eaton, Nature **390**, 196 (1997).

[74] R. A. Goldstein, Z. Luthey-Schulten, and P. G. Wolynes, Proc. Natl. Acad. Sci. USA **89**, 4918 (1992).

[75] I. Shindyalov and P. Bourne, Protein Eng. **11**, 739 (1998).

[76] D. Pantaloni, C. L. Clainche, and M.-F. Carlier, Science **5521**, 1502 (2001).

[77] J. Howard and A. Hyman, Nature **422**, 753 (2003).

[78] L. Cassimeris, N. K. Preyer, and E. D. Salmon, J. Cell Biol. **107**, 2223 (1988).

[79] P. J. Sammak and G. G. Borisy, Nature **332**, 724 (1988).

[80] D. K. Fygenson, E. Braun, and A. Libchaber, Phys. Rev. E **50**, 1579 (1994).

[81] A. Desai and T. J. Mitchison, Annu. Rev. Cell Dev. Biol. **13**, 83 (1997).

[82] F. Oosawa and S. Asakura, *Thermodynamics of the polymerization of protein*, Academic Press, London, 1975.

[83] A. Mogilner and G. Oster, Eur. Biophys. J. **28**, 235 (1999).

[84] G. S. van Doorn, C. Tanase, B. M. Mulder, and M. Dogterom, Eur. Biophys. J. **29**, 2 (2000).

[85] S. Martin, M. J. Schilstra, and P. M. Bayley, Biophys. J. **65**, 578 (1993).

[86] M. Dogterom and S. Leibler, Phys. Rev. Lett. **70**, 1347 (1993).

[87] H. Flyvbjerg, T. E. Holy, and S. Leibler, Phys. Rev. Lett. **73**, 2372 (1994).

[88] H. Flyvbjerg, T. E. Holy, , and S. Leibler, Phys. Rev. E **54**, 5538 (1996).

[89] D. J. Bicout and R. J. Rubin, Phys. Rev. E **59**, 913 (1999).

[90] E. B. Stukalin and A. B. Kolomeisky, J. Chem. Phys. **122**, 104903 (2005).

[91] H. Wang and E. Nogales, Nature **435**, 911 (2005).

[92] M. I. Molodtsov, E. A. Ermakova, E. E. Shnol, E. L. Grishchuk, and J. R. M. F. I. Ataullakhanov, Biophys. J. **88**, 3167 (2005).

[93] V. Vanburen, D. Odde, and L. Cassimeris, Biophys. J. (2005).

[94] M. Bindschadler, E. A. Osborn, C. F. Dewey, and J. L. McGrath, Biophys. J. **86**, 2720 (2004).

[95] V. Privman, *Nonequilibrium statistical mechanics in one dimension*, Cambridge, New York, 1997.

[96] D. C. Mattis and M. L. Glasser, Rev. of Mod. Phys. **70**, 979 (1998).

[97] N. G. van Kampen, *Stochastic process in physics and chemistry*, pages 100–108, North-Holland, New York, NY, revised edition, 1992.

[98] M. Doi, J. Phys. A: Math. Gen. **9**, 1465 (1976).

[99] R. Walker et al., J. Cell. Biol. **107**, 1437 (1988).

[100] E. T. O'brien, W. A. Voter, and H. P. Erickson, Biochemistry **26**, 4148 (1987).

[101] R. J. Stewart, K. W. Farrell, and L. Wilson, Biochemistry **29**, 6489 (1990).

[102] A. Vandecandelaere, M. Brune, M. R. Webb, S. R. Martin, and P. M. Bayley, Biochemistry **38**, 8179 (1999).