RICE UNIVERSITY

Exploring the Folding Energy Landscape: Designed, Simplified, and α-helical Membrane Proteins

by

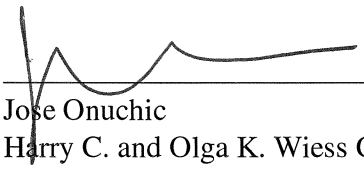
Ha Huynh Truong

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED, THESIS COMMITTEE:

Peter Wolynes, Chair
Bullard-Welch Foundation Professor of Science,Professor of Chemistry and Materials Science and NanoEngineering, Professor of Physics and Astronomy



Jose Onuchic
Harry C. and Olga K. Wiess Chair of Physics, Professor of Chemistry and Biosciences

Cecilia Clementi
Professor of Chemistry and Chemical and Biomolecular Engineering

Houston, Texas

April, 2016

ABSTRACT

Exploring the Folding Energy Landscape:
Designed, Simplified, and α-helical Membrane Proteins

by

Ha Huynh Truong

This thesis discusses our efforts in using the energy landscape theory and coarse-grained molecular dynamics protein folding models to explore the folding energy landscape of proteins. The Associative-memory, Water-mediated, Structure and Energy Model (AWSEM) is capable of performing de novo structure prediction on not only many natural globular proteins but also designed proteins such as Top7 and Takada. AWSEM also enables us to investigate the robustness of folding natural and designed protein sequences upon simplification of full sequences to the five-letter or two-letter code. More recent work, using AWSEM or structure-based (SB) model with the addition of an implicit membrane energy term, shows that the energy landscapes for folding α-helical membrane proteins are funneled once their native topology within the membrane is established, further proves that tertiary folding of α-helical membrane proteins is thermodynamically controlled. The first chapter is an overview of the energy landscape theory of protein folding, followed by subsequent three chapters which describe in details how the energy landscape theory can be used as a fundamental theoretical framework to elucidate the folding problems (folding and binding) for both globular (natural and designed) proteins and α-helical membrane proteins.

Acknowledgements

This acknowledgment section is for the many wonderful people that have influenced me, taught me valuable lessons and helped me through many tough days not only during the past five years of my PhD program but also throughout my ten-year journey as a foreign student in the U.S.

First of all, I want to thank my advisor, Professor Peter Wolynes. In the past five years, he has given me many scientific ideas and valuable suggestions. His passion for science and his deep scientific knowledge and intuition have continuously inspired me since our first meeting when I was a clueless first-year student and he just moved to Rice.

The Center for Theoretical Biological Physics has been a great place for me to do research. I enjoyed meeting and talking with all the CTBP members, especially the Wolynes lab members who I have had a great time working closely with: Dr. Nick Schafer and Dr. Bobby Kim, my two close collaborators who also have given me a lot of helps when I first joined the group and throughout many joint projects; Dr. Weihua Zheng for many great discussions about science, health, and life in general; Brian Sirovetz for being a great and supportive office-mate; Dr. Victor Tsai, Dr. Bin Zhang, and Dr. Davit Potoyan for many valuable suggestions; and others who have lent me their support in the past five years. Thanks to Tracy Hogan who has worked hard and has done a great job in making sure everything went smoothly during my PhD. Thank you to my committee members, Professor Jose Onuchic and Professor Cecilia Clementi for their time and valuable advises through the years.

I would not have pursued this PhD without the background and the many inspiring teachers/mentors Highline Community College and the University of Oregon provided me. Thank you especially to Professor Marina Guenza, my "Oregon mom," for introducing me to computational research, for her guidance and heartfelt advice.

To the many friends and support systems that I am so lucky to have; from Natoma, Kansas: the MacConnell family, especially papa Ben and momma Shannon who took me into their family and taught me a lot about American culture during my senior year

in high school as an exchange student; from Highline: Mama Moon, my ILSC family, and my DLVS family; from Oregon: my study partner - Kim Arias, my ISA team, and my fellow Vietnamese Ducks, especially Hong-Phuong Vo and Duy Minh Nguyen; thank you for the many years of amazing friendships. Thanks to the Orlowsky family, Genny, John and Morgan, for many fun memories and many pounds of crawfish that we have devoured. Thank you to the supportive team at the Rice Graduate Housing: Connie Real, Thuy Nguyen and my fellow RAs. I want to give a shout-out to my food-and-board-game buddies: Justin Chen, Likeleli Seitlheko, Carmen Young, thank you for the many hours of pigging/nerding out and for still being able to maintain our friendships despite all the table-flipping, friendship-ruining games we have played over the years. Also, to all of the many fantastic friends I have made in the past five years at Rice: Lydia Kisley, Jodie Lee, Zoe Pham, Xiaoyu Bai, Amanda Goodman, Char Wickremasinghe, and many others I could list. You all have been an essential aspect of my time here at Rice.

Finally, I would have not become the person I am today without the love and support from my family. My parents, Minh Cao Truong and Huong Xuan Thi Huynh, valued the importance of my education and have continuously supported me through the years even when things were difficult and they had to make many scarifies, and for that I am extremely grateful. Thank you to my older brother, Son Truong, for motivating me to learn from mistakes and to become a better person. Thank you to my younger brother, Duc Truong, who the past two years has not only been an awesome house mate (who introduced me to many cool video games) but also an amazing listener for me to share all the ups and downs of graduate school. To many of my aunts and uncles, their support and encouragements have been a big help for me to pursuit my education in the US. And last but not least, MyMy Nguyen, I don't know where I would be without her support.

Dedicated to my beloved parents, Minh Cao Truong and Huong Xuan Thi Huynh.

Without their love, supports and sacrifices, I would not have achieved what I have today.

Dành cho ba mẹ của con, Trương Cao Minh và Huỳnh Thị Xuân Hương.

Nếu không có tình thương, sự ủng hộ và hi sinh của ba mẹ thì con sẽ không thể đạt được những thành quả ngày hôm nay.

Contents

Illustrations

Tables

Chapter 1

Introduction

1.1 Motivation: The protein folding problem

Proteins are essential to all known forms of life. They make up most of the dry mass of the cell and they are the dominant structural and functional element of the cell. The fact that proteins fold into organized structures is a remarkable physical phenomenon. Despite a large number of possible configurations, a protein molecule is able to fold into an unique three dimensional structure on a biological timescale from a one dimensional sequence. Structure and function are closely linked in biology, thus, being able to obtain structures of proteins will allow us to gain functional insights and better understand issues of specificity important to systems biology and medicine. However, obtaining full three dimensional structures experimentally remains a challenge and is still relatively expensive. On the other hand, many computational approaches have been developed and have been successful in predicting structures of proteins. Using theory and molecular dynamics simulations, we are interested in finding answers for questions in the protein folding field such as: How does a protein go from the one dimensional sequence to an unique three dimensional structure? Can we predict structure of a protein from its sequence using molecular dynamics simulation? What is the folding mechanism? The Energy Landscape Theory and the Princi-

ple of Minimal Frustration have provided a theoretical framework for us to tackle questions in the folding of globular and membrane proteins.

1.2 Background

A successful and efficient search of the native state out of a large number of possible configurations is possible if the energy landscape of a protein is funneled. A two dimensional schematic funneled energy landscape is shown in Figure 1.1. Energy decreases and degree of foldedness increases as a protein goes from the top to the bottom of the funnel. At the top of the funnel is the denatured ensemble which consists of many states with very few intrachain contacts, extended structures, large structural entropy, and low stability. At the bottom of the funnel is the nearly unique native state and its related conformational substates, which are significantly energetically stable. As the configurations go from being completely extended to being more native-like, they must collapse and pass through the molten globule states, which contain many non-native local energy minima in which partially folded proteins can become trapped. For a protein to fold properly, the stability gap δE_s, defined as the energy difference between the native state and these molten globule states, has to be maximized to make the native structure much more stable than the partially folded structures and the energy variance, ΔE, has to be minimized for the protein to avoid being trapped in local minima.

Frustration arises from energetic conflicts, which are for the most part avoided in protein structure since the native interations are generally more stable than non-native inter-

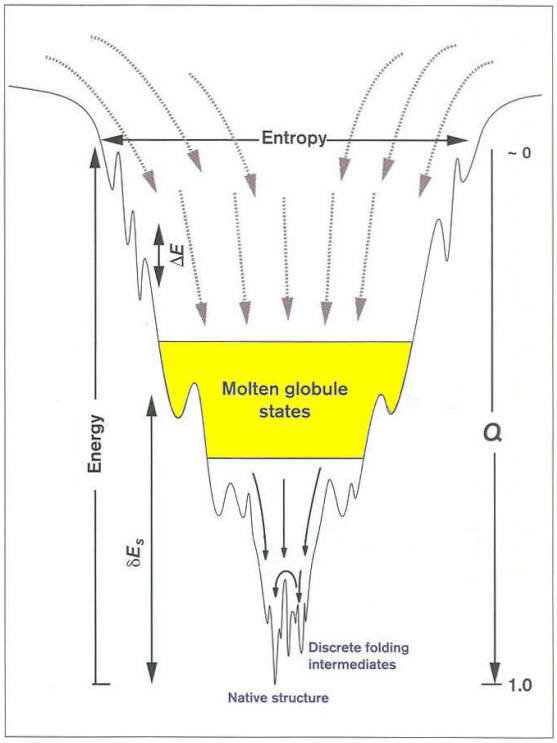


Figure 1.1 : Two dimensional schematic of the folding funnel. The width of the funnel represents entropy, and the depth represents energy. Denatured ensemble is at the top of the funnel. Collapsed candidate structures are the so called molten globule states. The nearly unique native state and its related comformational substates are at the bottom of the funnel. The faction of native contacts correctly made, Q, measures the degree of foldedness of the protein. This figure was adapted from Ref [1].

actions that might form in alternative conformations. This fact is known as the Principle of Minimal Frustration [7]. Previous studies [8, 9] show that highly frustrated contacts are often found at the protein surface, especially at functional sites or at parts of the protein that undergo conformational changes. The frustratometer [8] measures how favorable a particular native contact is relative to the set of all possible contacts in that location, normalized by the variance of that distribution. The frustratometer can be used as a tool to make useful prediction of sites that might have interesting functions or conformational changes just by

taking inputs from the sequence and structural information.

1.3 Models

There are two main models that are used in the studies which are presented in this thesis: the structure-based model (SBM) and the associative memory, water mediated, structure and energy model (AWSEM). Two main flavors of AWSEM, the "single memory" and predictive "fragment memory," are used. The structure-based model, single memory, and predictive AWSEM all share a coarse-grained backbone description wherein the position and orientation of each amino acid residue is dictated by the positions of its C_α, C_β, and O atoms (except Glycine, which lacks a C_β atom). Details for each model are further discussed in the following sections.

1.3.1 Structure-based model

The structure-based model is a perfectly funneled model, this means it takes into account only the native contacts which makes it an excellent model for us to investigate the topological frustration. Its Hamiltonian, shown in Equation 1.1 and 1.2, contains a backbone term ($V_{backbone}$) and a non-additive term (V_{na}), in which E_i is a pairwise-additive energy term and p is the non-additivity exponent. Values of p in the range of 2.0 to 3.0 have been shown to produce protein-like levels of cooperativity when global and local folding events are considered [10, 11].

$$V_{SBM} = V_{backbone} + V_{na} \tag{1.1}$$

$$V_{na} = -\frac{1}{2}\Sigma_i |E_i|^p \tag{1.2}$$

We employ a non-additive structure-based model ($p = 2.0$) to study the effect of topological frustration on the landscape of Top7, S6, and TakadaN in Chapter 2 and a modified pairwise-additive structure-based model ($p = 1.0$) to model folding of GlpG within the lipid bilayer or in detergent micelles in Chapter 4.

1.3.2 Associative memory, Water mediated, Structure and Energy Models

The associative memory, water mediated, structure and energy model (AWSEM) Hamiltonian, given in Equation 1.3, contains transferable and physically motivated terms (such as $V_{contact}$, V_{burial} and V_{HB}) that were optimized using the Energy Landscape Theory, and a bioinformatically-based term, V_{AM}. The model does not explicitly represent solvent molecules, so it is computationally efficient. The effects of water are modeled implicitly using the interaction terms in the Hamiltonian. The $V_{contact}$ term consists of a direct, pairwise-additive contact term, for residues that are close in space and can either attract or repel each other, as well as the non-pairwise additive water mediated interaction. The water mediated interaction is a sequence dependent pairwise contact interaction that switches smoothly between two different interaction weights depending upon the degree of burial of the interacting residues. This can be determined by counting the density of protein surrounding each residue. If the density of protein around both residues is low, the residues are given the water-mediated interaction weight. If the density of protein around either residue is high, the interaction is protein-mediated. This aspect of the model is fairly unique and is

illustrated in Figure 1.2. The details of the model are described further in Ref 2.

$$V_{AWSEM} = V_{backbone} + V_{contact} + V_{burial} + V_{HB} + V_{AM} \tag{1.3}$$

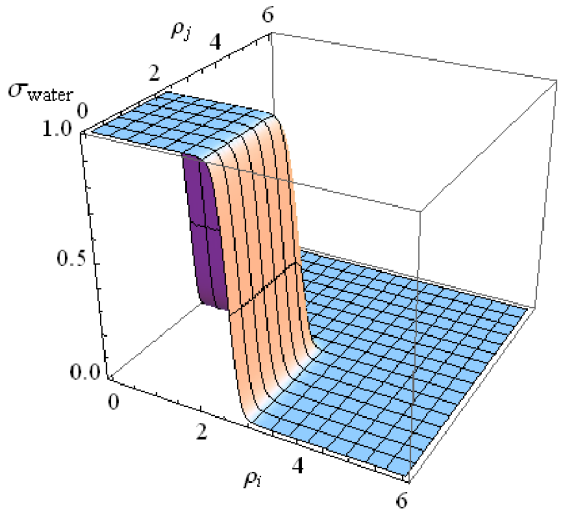


Figure 1.2 : The water mediated interaction switches smoothly between two interaction weights depending on the degree of burial of the interacting residues. This switching function is shown as a function of the protein density, ρ, of the two residues participating in the interaction. This figure was adapted from Ref 2.

The associative memory term, V_{AM}, is used to bias local-in-sequence configurations. There are two main flavors of AWSEM depending on the memories used, the single memory AWSEM and the predictive AWSEM, that are relevant to this thesis.

Single memory AWSEM

In single memory AWSEM, the experimentally determined structure is used as the only "memory" for the local in sequence associative memory interaction. We employ the single

memory AWSEM to construct free energy profiles and quantify the effects of tertiary energy frustration alone in designed versus natural globular proteins, as detailed in Chapter 2. The model is also used to predict the binding interfaces for membrane protein dimers in which the memory terms use local structural information about the monomers, as discussed in Chapter 3.

Predictive AWSEM

In predictive AWSEM or "homologs excluded" AWSEM, memories are obtained from the alignment of 9-residue fragments of the target sequence to a database of sequences corresponding to experimentally determined structures, and fragments from homologous sequences are excluded. AWSEM with a "homologs excluded" fragment library is used to predict protein structure via simulated annealing.

1.3.3 Implicit membrane model

The implicit membrane model is a density dependent residue-residue interaction potential which is used to capture the interaction between residues in the $30\mathring{A}$ modeled membrane plane. The potential was optimized using the Energy Landscape Theory on the hypothesis that the energy landscapes for folding α-helical membrane proteins are funneled once their native topology within the membrane is established. Low-density interactions are no longer water-mediated but membrane-mediated due to the optimization of the water-mediated tertiary interaction term using a database of α-helical membrane proteins. The implicit membrane potential term is added to the Hamiltonian of the structure-based model

or AWSEM to study folding of α-helical membrane proteins, shown in Chapter 3 and 4. Details of the implicit membrane model is described in Ref. 6.

1.4 Searching, sampling and analysis methods

1.4.1 Choice of order parameters

Root-mean-square deviation (RMSD) is often used as a measure of distance to the native state and is useful as a way of comparing structures within the native basin. Structures which have reasonably well formed secondary structure and significant topological similarity to the native state can still have a large RMSD if they have partial contact maps. Due to the sensitivity of RMSD, other reaction coordinates should also be considered. Reaction coordinates like Q are more useful when looking at structures comprehensively across the landscape. Q, the fraction of native contacts correctly formed in the simulated structure, measures the degree of foldedness of the protein. A formula for Q is given in Equation 1.4. Structures around $Q = 0.25$ are mostly unfolded and extended, whereas $Q = 0.4$ structures have reasonably well formed secondary structure and some topological similarity to the native state. The structure typically has overall correct topology at $Q = 0.55$ and is native-like with a RMSD of less than a few $\mathring{A}$ from the native structure at $Q = 0.7$. Q has been found to be a useful reaction coordinate for thermodynamics and kinetic analysis when the landscape is well funneled [12]. Despite that, there is no perfect set of order parameters and it is often useful to create new reaction coordinates that are specific to the problem that you are investigating. Examples of other useful reactions coordinates are shown in the result

section of the subsequent chapters.

$$Q = \frac{1}{N_p} \sum_i \sum_{j>i+2} \exp\left[\frac{-(r_{ij} - r_{ij}^{\mu})^2}{2\sigma_{ij}^2}\right] \tag{1.4}$$

1.4.2 Simulated annealing

The method we use for protein structure prediction is simulated annealing. Simulated annealing [13], the gradual cooling of a system from above to below its folding temperature, has proven to be a general method for searching for the native state in conformational space when the landscapes are funneled. Frustration in the native basin can still prevent the simulation from reaching the absolute lowest energy state despite the folding model being globally funneled [14]. If the landscape is sufficiently funneled, one nevertheless will still find a structure which is closely related to the true global minimum. The final structures obtained at the end of simulated annealing simulations are defined as "predicted structures." The similarity of the predicted structures to the native structure can be measured by looking at their global Q values and other parameters such as RMSD.

1.4.3 Umbrella sampling

Molecular dynamics combined with biased sampling techniques, such as umbrella sampling, can be used to obtain a global picture of the landscape. Simulations are biased by forces that constrain quantitative measures of the simulated molecule to measures of the known native structure at multiple temperatures. Biased simulations can be performed by

umbrella sampling along a predefined reaction coordinate. Q is often chosen as the biased reaction coordinate. A harmonic Q bias is given in Equation 1.5. Biased simulations then will be combined and unbiased to obtain multidimensional free energy profiles using the multi-state Bennett acceptance ratio (MBAR) method [15].

$$V_{Q-bias} = \frac{1}{2} k_{Q-bias} (Q - Q_0)^2 \tag{1.5}$$

1.4.4 Free energy landscape analysis

Free energy landscape analysis is useful because it gives an overview of which parts of the landscape will be sampled and which will not during prediction runs. Regions that are high in free energy will not be sampled very often at equilibrium; the system will spend most of its time in ensembles corresponding to low free energy regions. It is also possible to make estimates of how many distinct structures exist at various points along the reaction coordinate and from that, we can also characterize stable intermediates or the transition state ensemble and gain insight into the folding mechanism of the protein along its folding pathway(s). Many examples of the structural characterization based on free energy landscapes are shown in the subsequent chapters. Methods that allow for the calculation of free energy profiles can also be extended to calculate expectation values. One of the most interesting expectation values is that of energy versus degree of foldedness (Q). If the expectation value of the energy decreases as the configurations become more native, and it follows that the gap between the native basin and unfolded basin is large and

the energy variance is small, the landscape is said to be funneled. An example of this $E(Q)$ plot is shown in Chapter 2 Figure 2.5.

1.5 Overview

The work in Chapter 2 was published in the *Journal of Chemical Physics*in 2013 [16], in which we use structure prediction tools, frustration analysis and free energy profiles to illustrate the folding landscapes of Top7 and two other proteins designed by Takada. The role of topological frustration versus energetic frustration in these designed systems and how they differ from those found for natural proteins is discussed. We also study the robustness of folding upon simplification of sequences of these designed and natural proteins using fewer amino acid types. The quality of structure prediction using five-letter, two-letter code, and full sequences is compared.

Chapter 3 shows the use of AWSEM-Membrane in studying the energy landscapes for membrane protein oligomerization and in predicting the binding interfaces of various membrane protein dimers. Energy landscape analysis further shows that degeneracies in predicting structures of membrane protein monomers [6] are generally resolved in the folding of the higher order assemblies. We also study the reassembly of two cleaved bacteriorhodopsin fragments and demonstrate the important role of the retinal cofactor in the folding and function of the protein. The work described in Chapter 3 was published in 2015 in the *Journal of Chemical Physics* [17].

Chapter 4 presents work published in *Proceedings of the National Academy of Sciences*

USA in 2016 [18] and demonstrates the folding of GlpG, an intramembrane protease, within a lipid bilayer and in detergent micelles using perfectly funneled structure-based models with and without the presence of an implicit membrane energy term. Structural free energy landscape analysis reveals required backtracking that explains the negative ϕ-values observed in an experimental study [19]. Characterization of a near-native state shows functional motions which involve the unbinding of transmembrane helix 5 from the rest of the structure to expose GlpG's active site.

The work described in this thesis further demonstrated how the theoretical framework of the Energy Landscape Theory and the Principle of Minimal Frustration can elucidate the folding problems for designed proteins and α-helical membrane proteins. Optimized coarse-grained protein folding simulation models have proven to be extremely useful tools in not only predicting structures and studying protein-protein interactions but also identifying stable intermediates, and exploring pathway(s) and functional motions of proteins. Many of our results not only agree well with experimental observations but also provide explanations to puzzling experimental results and give new insights into the folding landscape of proteins. Many future possibilities exist for studying protein folding using optimized coarse-grained models. Further application and development of these models and collaborations with experimentalists will enable us to gain deeper understanding of the nature of life at its most basic level.

Chapter 2

Funneling and Frustration in the Energy Landscapes of some Designed and Simplified Proteins

2.1 Abstract

We explore the similarities and differences between the energy landscapes of proteins that have been selected by nature and those of some proteins designed by humans. Natural proteins have evolved to function as well as fold, and this is a source of energetic frustration. The sequence of Top7, on the other hand, was designed with architecture alone in mind using only native state stability as the optimization criterion. Its topology had not previously been observed in nature. Experimental studies show the folding kinetics of Top7 is more complex than the kinetics of folding of otherwise comparable naturally occurring proteins. In this paper, we use structure prediction tools, frustration analysis and free energy profiles to illustrate the folding landscapes of Top7 and two other proteins designed by Takada. We use both perfectly funneled (structure-based) and predictive (transferable) models to gain insight into the role of topological versus energetic frustration in these systems and show how they differ from those found for natural proteins. We also study how robust the folding of these designs would be to the simplification of the sequences using fewer amino acid types. Simplification using a five amino acid type code results in comparable quality of

structure prediction to the full sequence in some cases, while the two letter simplification scheme dramatically reduces the quality of structure prediction.

2.2 Introduction

There is considerable evidence that natural proteins have evolved to have minimally frustrated energy landscapes that are funneled towards the native state by native interactions that are stronger than alternative possibilities[7, 20, 21, 22, 23, 24]. Natural protein folding is thus under thermodynamic not kinetic control. Residual frustration exists, and localized frustrated regions have been shown to be correlated with functional regions of proteins such as binding sites [8] and regions that undergo partial local unfolding or reconfiguration during conformational changes necessary for allosteric regulation [9, 25]. The overall low degree of frustration however distinguishes natural proteins from random heteropolymers, which have many globally unrelated low energy states. What about proteins that have been designed rationally by people with the aid of computers? Such designed sequences have also undergone a selection process, but one guided by humans and their preconceptions rather than nature and its harsh functional constraints. With the hope of controlling protein structure and functions, many methods have emerged for designing a sequence that folds reliably to a target structure. Some important and successful protein designs have focused on either stabilizing the target folded state alone, as in the Baker group's system of Top7[26], or on funneling the global landscape, as in Takada's design[27]. Both design strategies can be consistent with the "principle of minimal frustration"[7] if the pre-conceived ideas

about the energetic force field employed in the design stage are good enough. In accord with the minimal frustration idea, robust protein design will generally also require destabilizing non-native states (explicit negative design)[28] as well as ensuring native stability. Disfavoring the vast number of non-native states remains a challenge for making protein design routinely successful.

Top7 was designed in 2003 in the Baker laboratory at the University of Washington by minimizing a model free energy of a targeted single folded monomeric structure that was specifically chosen to be unlike any that had previously been observed for a natural protein [26]. The design scheme started with a "sketch" of the topology and the initial sequences were generated by taking fragments from proteins with resolved structures such that the secondary structure agreed with the desired secondary structure elements of the design. They then iterated between Monte Carlo based sequence design and gradient based backbone optimization for multiple rounds, each time reoptimizing the lowest energy sequence/structure pairs found in the last round. The energy function used was a pairwise additive, implicit solvent fully atomistic model that contains hydrogen bonding and Lennard Jones terms, and gave special attention to tight packing of side chains. During the sequence optimization, most of the positions in the sequence were allowed to be mutated to any residue except for cysteine; only the surface residues of the β-strands were restricted to being polar residues. The resulting sequence had no significant homology to any known protein sequence. Despite having a novel topology and sequence, it was able to fold in the laboratory, and was found to be highly soluble and monomeric. The x-ray crystal structure

of the synthesized Top7 is very similar to the targeted goal, with a root mean square difference of 1.2Å. It was noted that the crystal structure was more ordered on the C-terminal half. It was also found to be unusually stable, being still apparently folded at 98°C. The equilibrium chemical denaturation showed a cooperative unfolding event with a midpoint around 6M Gu-HCl. At the time this was understood by some as demonstrating that extensive negative design and/or the explicit consideration of the kinetic process of folding was not necessary in order to design protein sequences that fold to unique structures.

In 2004, a study of the kinetics of several designed proteins, including Top7, was carried out [29]. Besides Top7, the other proteins in the study were designed using a similar procedure to that which produced Top7, but were all designed to fold into topologies of particular natural proteins. Most of the redesigned proteins were found to fold faster than their natural counterparts. Top7 also folds quickly compared to many natural proteins of its size, but it was unusual in that, unlike most natural proteins and unlike the redesigned proteins with natural topologies, its folding exhibits complex multiphase kinetics that are essentially denaturant concentration independent under a range of folding conditions. To explain the difference in folding rates between the natural and redesigned proteins, it was suggested that perhaps natural selection favors high barriers to unfolding in order to disfavor aggregation *in vivo*. Three possible sources for the unique behavior of Top7 were noted: highly populated intermediates with buried hydrophobic residues, a shift of the transition state towards the unfolded state, or an increase in internal friction. Further experimental characterization in 2007 [30] led to the conclusion that some non-native states of Top7 as

well as native-like fragments of Top7 were stable at equilibrium. The kinetics were also further resolved, leading to the conclusion that one of the slow rearrangments corresponds to a transition between two collapsed states. One possible reason for the presence of multiple collapsed states was suggested, namely that the optimization process lead to an expanded hydrophobic core. They also mention the possibility that the extreme regularity of Top7's β strands may make it easier for strand rearrangements to occur. Mutation studies helped to identify a subset of residues that are involved in a non-native intermediate as well as a different subset that was thought to be important to the transition state. In summary, this work demonstrated that not all protein sequences that can be crystallized have energy landscapes as smooth as those of most natural proteins.

Clearly then Top7 represents an interesting testing ground for protein folding theorists; some of the first serious simulation studies on this system were carried out in the Chan lab [31, 32]. Using several variations on an essentially native-centric model, they were able to observe a stable intermediate with a folded C-terminal fragment, consistent with the previous experimental work. They initially concluded that the nonnatural topology of Top7 was the dominant determining factor in its noncooperative folding, and speculated that perhaps some topologies were fundamentally uncooperative or that the artificial design procedure was not equal to that of natural evolution or selected for different traits. This initial study was followed up with a more thorough study of the simulated thermodynamics and kinetics of Top7 and S6 using a native-centric model that was perturbed by adding sequence dependent hydrophobic interactions. They found that despite only having

minor effects on the free energy profile, non-native hydrophobic interactions were absolutely essential to recreating something like the observed rollover in the folding arm of the chevron of Top7. In particular, they noted that 6 of the 7 residues mentioned as being important for non-native interactions in the experimental work of Baker were indeed found to make significant non-native interactions in their simulations, with the exception being V81. They concluded that the long stretch of hydrophobic residues in the C-terminal helix of Top7 is an important contributor to its strange folding behavior. They reiterate Baker's suggestion that the regularity of the β-strands might favor incorrect pairings, but note that this would not be captured in their essentially native-centric model.

By comparing results from a structure-based model and the Associative memory, Water mediated, Structure and Energy Model (AWSEM, an optimized predictive model), we have been able to investigate both topological and energetic factors to see about their relative impact using a fairly realistic energy function distinct from that used in the original design.

In early 2003, the Takada laboratory used a fully automated procedure inspired by energy landscape theory principles to design sequences for a target 3 helix bundle structure[27]. The focus in this case was on crafting the global landscape into a funnel shape by explicit negative design against the vast number of unfolded configurations. This computationally daunting task necessitated the use of a coarse-grained model that did not emphasize tight sidechain packing. The model is similar to AWSEM in that it uses a 3 atom per residue representation and explicit hydrogen bonds, but differs from AWSEM in its relatively simple hydrophobic interactions and context dependent electrostatic interactions. Like AWSEM,

Takada's model was originally developed for folding studies [33]. This allowed the Takada group to base the design procedure on a set of structures coming from folding simulations. These structures were generated before the final sequence was fixed. The unfolded structures, below a certain threshold value of the number of native contacts, were used as the denatured ensemble, and a truncated and relaxed version of the protein G-related albumin binding domain (PDB ID: 1PRB) was used as the targeted goal structure. The sequence that corresponds to the natural protein 1PRB will be referred to as TakadaN in this paper, to emphasize the structural similarity to the designs. A Monte-Carlo with simulated annealing search for optimal sequences was then performed with the Z score as the function to be optimized. The putative sequences were then tested with folding simulations. When it was found that these sequences did not fold in simulation, a variant on the Z score that in addition to accounting for the the gap between the unfolded state and target structure employs the gap between the intermediate states and the target structure to develop an objective function to search for new sequences. A subset of the sequences optimized with respect to the double Z score were found to fold quickly in simulation. Finally, the procedure was repeated with restrictions on the amino acid composition in order to ensure solubility, and three of the resulting sequences were chosen for experimental characterization. For the purposes of comparison, several sequences were generated for the target structure using only total energy of the native state as the objective function (while still using the same amino acid composition constraints), and the lowest of these was also experimentally characterized. All of the optimized sequences were found to have low sequence similar-

ity to the native sequence of the target structure. CD and NMR experiments led them to conclude that one of the double Z score optimized sequences (originally named DHB06, called TakadaZ in this study) had both stable secondary and tertiary structure. The energy optimized sequence (originally named DHBE, called TakadaE in this study) had similar amounts of secondary structure but a poorly resolved one-dimensional NMR spectrum. Diffusion measurements indicated that TakadaE was forming multimers in solution, which led to the conclusion that TakadaE may aggregate due to its lack of well defined tertiary structure. Finally, they noted that the other two Z score optimized sequences that were experimentally characterized showed problems, either with packing or large fluctuations from the native state, in all-atom simulations whereas TakadaZ did not. They then concluded that screening designed sequences coming from coarse-grained models with all-atom simulations may be a useful way of determining beforehand which of the sequences will likely be well behaved in the laboratory.

If the problem of designing sequences to fold like proteins is understood as building in the signals necessary to fold starting from no sequence information, or from random sequences, then this immediately suggests another way of approaching the problem of determining what those signals are: gradually removing signals from natural protein sequences until folding fails. There are at least two reasonably controlled ways of accomplishing this. One interesting and practicable way is to gradually introduce more and more alanine mutations [34, 35]. These studies allow us to learn tremendous amounts of detail about which parts of the sequence are important for which aspects of folding, e.g., thermodynamics and

kinetics. Another equally interesting way to simplify sequences is to ask the question of how many amino acid types are necessary for a protein to fold on biological timescales. Homopolymers are unable to fold to a unique structure due to the degeneracy of collapsed conformations. If we were to introduce energetic heterogeneity to these collapsed conformations through a two-letter hydrophobic and polar code, we would begin to observe energetic discrimination among these states. However, a theoretical study has reported that two letter codes generally still give rise to many energetically low-lying non-native conformations [36]. A two letter hydrophobic/polar code can distinguish between any two states that have different degrees of segregation, but cannot go further. These theoretical considerations have been discussed further by Wolynes in ref. 37. Although folded helical proteins generated with a three letter code (Q,L,R) which undergo cooperative thermal denaturation have been reported [38], the Baker lab reported that a three letter code was insufficient in their attempts to simplify the sequence of the SH3 domain [39]. Rather, a five letter code (I, K, E, A, G), was required in order to build two variants of the SH3 domain in which approximately 70% of the sequence was simplified. One of the resulting variants folded at a rate similar to the native sequence, while the other variant folded even faster, suggesting that evolution may emphasize thermodynamic control. In 1999, Wang reported theoretical efforts to produce a simplified code based on the concept of mismatch between a reduced interaction matrix and the Miyazawa-Jernigan matrix [40]. This resulted in the same five letter code employed by Baker [39]. Sequences using the five letter code appeared to be kinetically foldable in their model studies. However, Chan [41] pointed out that 29% of

the residues of the simplest sequence studied by Baker [39] do not belong to the simplified IKEAG alphabet (there were, in fact, 14 amino acid types present in the sequence when all residue positions were counted). Later work by Wang [42, 43] indicated that the minimum number of amino acid types required for a protein to encode its structure might be as large as ten, which would be consistent with theoretical work by Levy [44] and Dill [45]. For highly symmetric structures, at least for small proteins, the minimum required number of letters might be lower [46].

In this study, we use AWSEM to study the effect of simplifying sequences of three designed proteins: Top7, TakadaZ, TakadaE, as well as the effect of simplification on the behavior of two natural controls: S6 and TakadaN. For the purposes of simplification, we have employed the five letter Miyazawa-Jernigan matrix scheme (MJ5) [40], and the two letter Blosum scoring scheme (BL2) [43].

2.3 Approaches

2.3.1 Systems investigated

Top7, the first protein to be designed to fold into a novel topology, has 92 residues and contains two α-helices packed on a five strand β-sheet with all anti-parallel strand pairings (see Figure 2.1A). The design process focused entirely on minimizing the free energy of the folded monomeric structure and did not use explicit negative design against possible alternative conformations nor consider the kinetic process of protein folding. Top7 is unusually stable compared to natural proteins, and exhibits complex, multi-phase kinetics in

its folding [29, 30], arising from the presence of several metastable intermediates. One intermediate state was found in simulation study to be more stable than either the folded or unfolded states[31]. Ribosomal protein S6 (PDB: 1RIS) was used as a comparison control system for Top7. S6 was chosen because of its similarity in length and secondary structure element composition to Top7 (Figure 2.1B), and also because it exhibits relatively simple, two-state kinetics as is quite common for natural proteins[47].

We have also studied two proteins designed by the Takada laboratory [27]. These two sequences were designed to fold into the structure of the truncated protein G-related albumin binding domain (PDB: 1PRB), which has the first unstructured N-terminal 6 residues cut out. This protein has a three-helix bundle topology, shown in Figure 2.1C. The two Takada sequences were formed using different automated computational approaches: the first used a sophisticated Z-score based criterion to build a globally funneled landscape employing a rather good coarse-grained energy function, whereas the second focused only on optimizing interactions within the target structure. The Z-score design variant is denoted as TakadaZ (originally DHB06), and the sequence designed by minimizing the energy of the target structure alone is denoted as TakadaE (originally DHBE) in this paper. In our studies, the truncated protein G-related albumin binding domain, which was the structural template for the design, was used as the control system. We denote it as TakadaN. We performed a short annealing simulation with the single memory AWSEM model (see Section 2.3.2) at low temperature, starting from the truncated 1PRB, and used the final structure as the target structure for the calculation of Q values. The target structure deviates only slightly from

the crystal structure, with a C_α RMSD of 2.4Åthat comes primarily from a tighter packing of the two terminal helices. This structure was also used for the Frustratometer analyses.

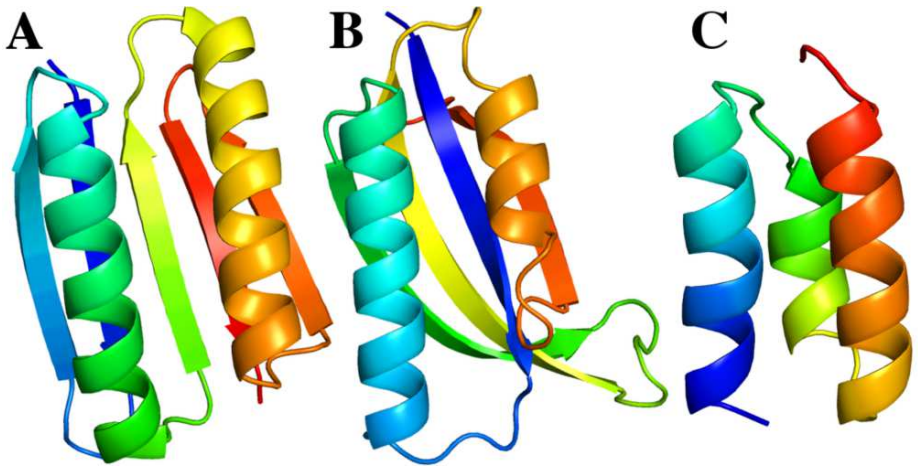


Figure 2.1 : The crystal structures of Top7 (A), and S6 (B), and the truncated and relaxed 1PRB structure, denoted TakadaN (C). Coloring of the structures is according to residue index starting at the N-terminal (blue) and going to the C-terminal (red). Structures were generated using PyMOL [3].

2.3.2 Models

Both the structure-based and predictive models are implemented in the molecular dynamics package, LAMMPS [48]. These models share a coarse-grained backbone description wherein the position and orientation of each amino acid residue are dictated by the positions of its C_α, C_β and O atoms (except Glycine, which lacks a C_β atom). The model does not explicitly represent solvent molecules, so it is relatively rapid to simulate. Instead, the effects of solvent are modeled implicitly using the interaction terms in the Hamiltonian. This predictive model contains water-mediated interactions that go beyond the usual hy-

drophobicity dominated contact models [49]. We believe it is likely these interactions are somewhat more realistic than those employed to make the original designs, which were already quite good.

We employ a non-additive structure-based model (SBM) to study the effect of topology on the landscapes of Top7, S6, and TakadaN. This model's Hamiltonian, shown in Equation 1.1, contains a backbone term ($V_{backbone}$) and a non-additive term (V_{na}), in which E_i is a pairwise-additive energy term and p is the non-additivity exponent as shown in Equation 1.2. For this study, a value of $p = 2.0$ was used. Values of p in the range of $2.0 - 3.0$ have been shown to produce protein-like levels of cooperativity when global and local folding events are considered [10, 11]. Complete details of this model are available in Ref. [50].

The Associative memory, Water mediated, Structure and Energy Model (AWSEM) was used to predict the structures and to study the role of non-native contacts (energetic frustration) on the landscape and folding free energy profiles of the designed and natural proteins mentioned above. The complete AWSEM Hamiltonian is given in Equation 1.3.

V_{AM} is a bioinformatically-based term, which depends on the fragment memories obtained from the alignment of 9-residue segments of the target sequence to a database of sequences corresponding to experimentally determined structures. Details of this model can be found in Ref. 2. A "single memory" model was also used for constructing free energy profiles so that effects of tertiary energy frustration alone could be quantified. In the single memory model, the fragments come directly from the experimentally determined structure

(PDB) and the secondary structure bias is taken from the STRIDE [51] assignment. In the "single memory" model, all interactions between residues that are close in sequence (less than 10 residues apart), including V_{AM}, are based on an experimentally determined structure. All other parts of the model are fully transferable, including $V_{contact}$, V_{burial}, and V_{HB}. AWSEM with a "homologues excluded" fragment library, together with simulated annealing simulation was used for structure prediction. This "fragment memory" model uses the JPRED prediction for its secondary structure bias [52]. The alignments coming from locally similar but globally unrelated structures introduces the possibility of frustration at the level of secondary structures, but in all probability overestimates this effect.

2.3.3 Simulation and analysis methods

The Frustratometer [8] has previously been used to measure and localize frustration in natural proteins by allowing us to computationally examine the changes in energy upon making mutations. The mutational frustration index, as described in Ref. 8, was used for all frustration calculations in this study. Roughly speaking, the mutational frustration index compares the stability of native interactions to a distribution of decoy interactions that are obtained by making mutations to the interacting residues themselves and the residues with which they are in contact. Frustration in general is the result of multiple competing interactions that cannot be simultaneous satisfied, and localizing frustration in the native structure of proteins can be useful in determining which parts of the protein are prone to local unfolding or misfolding. One way to represent localized frustration on a protein struc-

ture is to draw lines between residues in contact and color them according to their degree of frustration. In the resulting "frustratograms" (Figure 2.2), minimally frustrated contacts are shown in green, and highly frustrated contacts are shown in red. For most natural proteins, minimally frustrated linkages constitute a connected stable folding core for the molecule. The fraction of minimally frustrated and highly frustrated contacts were also calculated and are given in Table 2.1. Top7 is mostly minimally frustrated. Its very few highly frustrated contacts are between polar residues on the outside of the β sheet. The native structure of S6 has a highly frustrated region between the C-terminal unstructured coil and the β sheet. The coil region between the first β strand and first helix also make highly frustrated contacts with the twisted region of the second and third β strands. All three Takada sequences have a large fraction of minimally frustrated interactions. Notably, similar to what was found for Top7, TakadaE has an unusually high fraction of minimally frustrated contacts, while the fraction of highly and minimally frustrated contacts for the natural protein TakadaN and the designed TakadaZ are nearly identical. The largest cluster of highly frustrated contacts in TakadaN coincides with the putative albumin binding site at the N-terminal [53]. The frustratograms shown in Figure 2.2 and the frustration analysis in Table 2.1 were generated by the version of the Frustratometer that is implemented inside of AWSEM-MD, which is specifically designed to be consistent with the simulation Hamiltonian. Interactions within the range of fragment memory term V_{FM} are therefore excluded. The Frustratometer web server [54] includes these interactions.

Table 2.1 : Summary of the mutational frustration analysis for all sequence/structure pairs studied. The Columns of the table show the fraction of minimally, highly, and neutrally frustrated interactions present in the native structure (or putative native structure in the case of TakadaZ and TakadaE).

		Minimally	Highly	Neutrally
S6	FULL	0.41	0.13	0.45
	MJ5	0.40	0.14	0.46
	BL2	0.40	0.17	0.43
Top7	FULL	0.57	0.06	0.37
	MJ5	0.50	0.11	0.39
	BL2	0.55	0.06	0.39
TakadaN	FULL	0.49	0.06	0.45
	MJ5	0.48	0.08	0.44
	BL2	0.44	0.00	0.56
TakadaZ	FULL	0.49	0.05	0.46
	MJ5	0.55	0.07	0.38
	BL2	0.50	0.00	0.49
TakadaE	FULL	0.60	0.02	0.38
	MJ5	0.57	0.02	0.41
	BL2	0.53	0.09	0.38

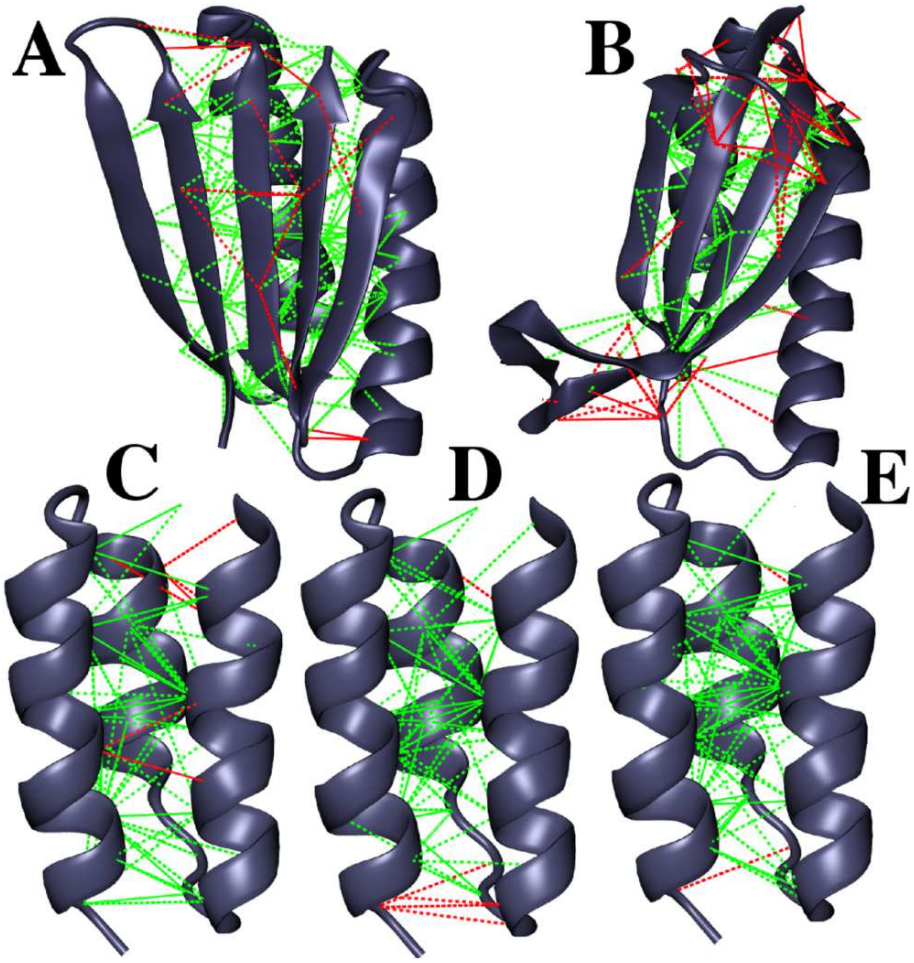


Figure 2.2 : The frustratograms show calculated mutational frustration of Top7 (A), S6 (B) and TakadaN, Z and E (C, D, E respectively). Minimally frustrated contacts are shown in green, and highly frustrated contacts are shown in red. Chapter1 were generated using VMD [4].

To survey the landscape of folding, we first employ simulated annealing simulations. These allow us to get an idea of how foldable a sequence is, and how robust the folding is to simplification of its sequence. These simulations were performed with a “homologs excluded” fragment library [2]. To generate a starting structure, a simulation starting from the native structure that was obtained from the Protein Data Bank[5] was first run at a

high temperature (well above the folding temperature), resulting in a random extended conformation. Starting from these extended conformations and this high temperature, the temperature was slowly brought down to below the folding temperature over the course of 1×10^7 steps, using a timestep of 2 fs. Coordinates of the system were saved every 1000 steps. For each saved snapshot, Q and radius of gyration values relative to the native structure were calculated. Q is the fraction of pairwise distances within 1Å of their distances in the native structure. The exact form of Q is given in Equation 1.4. Finally, structures were built from the last snapshot of each of these simulations, and the C_α RMSD was calculated for comparison to the experimentally determined structure in the cases of Top7 and S6, or the relaxed target structure in the cases of the three Takada sequences.

In order to sample along Q and calculate free energy profiles, we ran umbrella sampling simulations in which an harmonic bias (given in Equation 1.5) was added to the Hamiltonian. All free energy profiles and expectation values were calculated using the multi-state Bennett acceptance ratio (MBAR) method as implemented in the pyMBAR package [15]. Samples were collected for a range of temperatures near the empirically determined folding temperature.

2.4 Results

2.4.1 Top7 vs. S6

Since it has homogeneous interactions, the structure-based model allows us to evaluate the effects of topology alone on the folding of Top7. Figure 2.3 shows the two dimensional

free energy profiles of Top7 and S6 as a function of Q and radius of gyration computed using the non-additive structure-based model. There are two distinct low free energy regions, corresponding to the unfolded and folded states, in the free energy profiles of both Top7 and S6. Even for this single structure-based model corresponding to ideally funneled interactions, the free energy profile of Top7 is more complex than it is for the natural protein, S6, as reflected in the broad transition state region with multiple meta-stable states (Q values from 0.2 to 0.4). Structures at Top7's transition state have the C-terminal fragment preferentially formed, while the N-terminal fragment remains unfolded, whereas the narrower transition state region for S6 is more structurally homogeneous. The structure at the transition state of S6 has an overall correct topology, even though its secondary structures are incompletely formed. Compared to S6, which has two state folding kinetics in experiment [47], Top7 has intermediates with folded C-terminal fragments. These results are consistent with the previous simulation study [31]as well as with experiments on Top7. These results strengthen the hypothesis that topological frustration plays a dominant role in the complex folding kinetics of Top7.

The single memory model with transferable tertiary interactions was used to specifically study the role of tertiary energetic frustration on folding. The two dimensional free energy profiles $F(Q, rg)$ of S6 and Top7 (Figure 2.4 A and B respectively) are similar, though Top7 has a somewhat wider range of radii of gyration that are low in free energy. Both proteins still have an energetic bias towards native-like states as is shown in the $F(Q, E)$ plots (Figure 2.4 C and D respectively) with a low free energy basin that extends from low

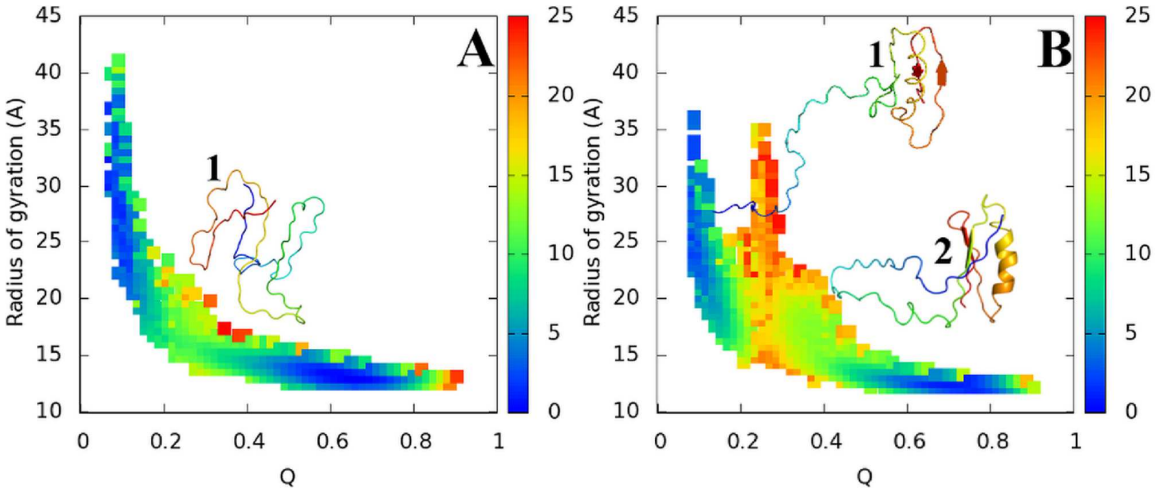


Figure 2.3 : Two dimensional free energy profiles of S6 (A) and Top7 (B) computed using the non-additive structure-based model. The free energy profile of S6 is less complex with two distinct low free energy regions corresponding to the unfolded and folded states. Structure A1, the representative structure at the transition state of S6, has an overall correct topology, though its secondary structures are incompletely formed. The free energy profile of Top7 is more complex, with multiple meta-stable states near the transition state. Structures (B1 and B2) at Top7's transition state (corresponding to Q values of 0.25 and 0.35) both have the C-terminal fragment preferentially formed, while the N-terminal fragment remains unfolded.

Q and high energy to moderately high Q and low energy. Figure 2.5 shows the energy of the tertiary interactions as a function of Q for both Top7 and S6. The tertiary energy is defined as the sum of $V_{contact}$ and V_{burial} in Equation 1.3. The tertiary energies of both proteins decrease as Q increases, and have approximately the same standard deviation, though Top7 has a slightly larger energy gap and is funneled to higher Q. The tertiary energy starts to flatten out at $Q = 0.5$ and $Q = 0.7$ for S6 and Top7, respectively.

Next, we used AWSEM with fragment memories to quantify the combined roles of secondary and tertiary frustration. Figure 2.8A shows the quality of structure prediction of

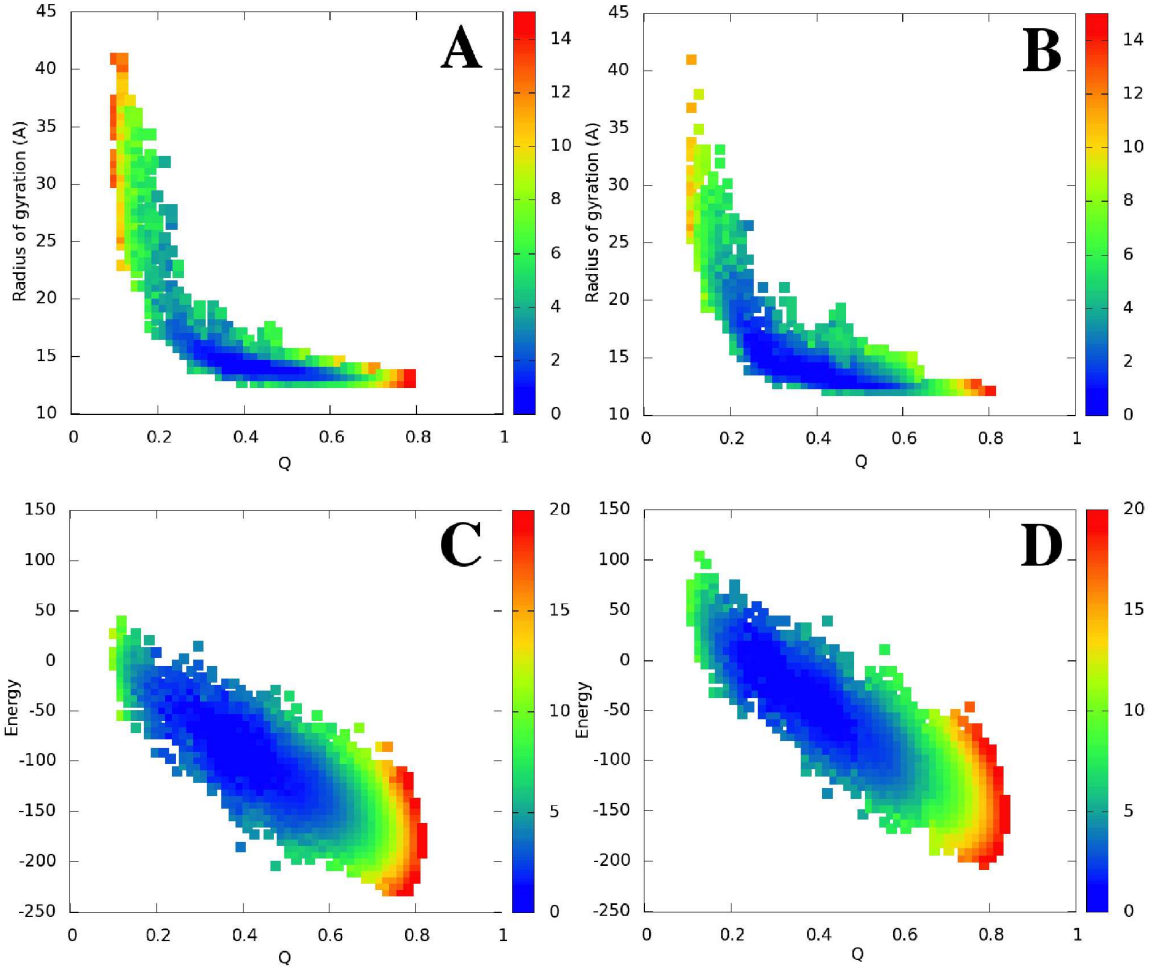


Figure 2.4 : The two dimensional free energy profile as a function of Q and the radius of gyration of S6 (A) and Top7 (B) are similar, though Top7 has a wider range of radii of gyration that are low in free energy. The two dimensional free energy profile as a function of Q and Energy of S6 (C) and Top7 (D) are also similar. All free energy profiles were calculated using the "single memory" model with transferable tertiary interactions.

Top7 and S6 over twenty simulated annealing runs. Top7 is better predicted, with overall better Q values than S6. The best predicted structure, with a Q value of 0.74 (Figure 2.6A), is the only well packed structure with the correct topology. Two of the energetically competitive structures are shown in Figure 2.6. We frequently observed swapping of the

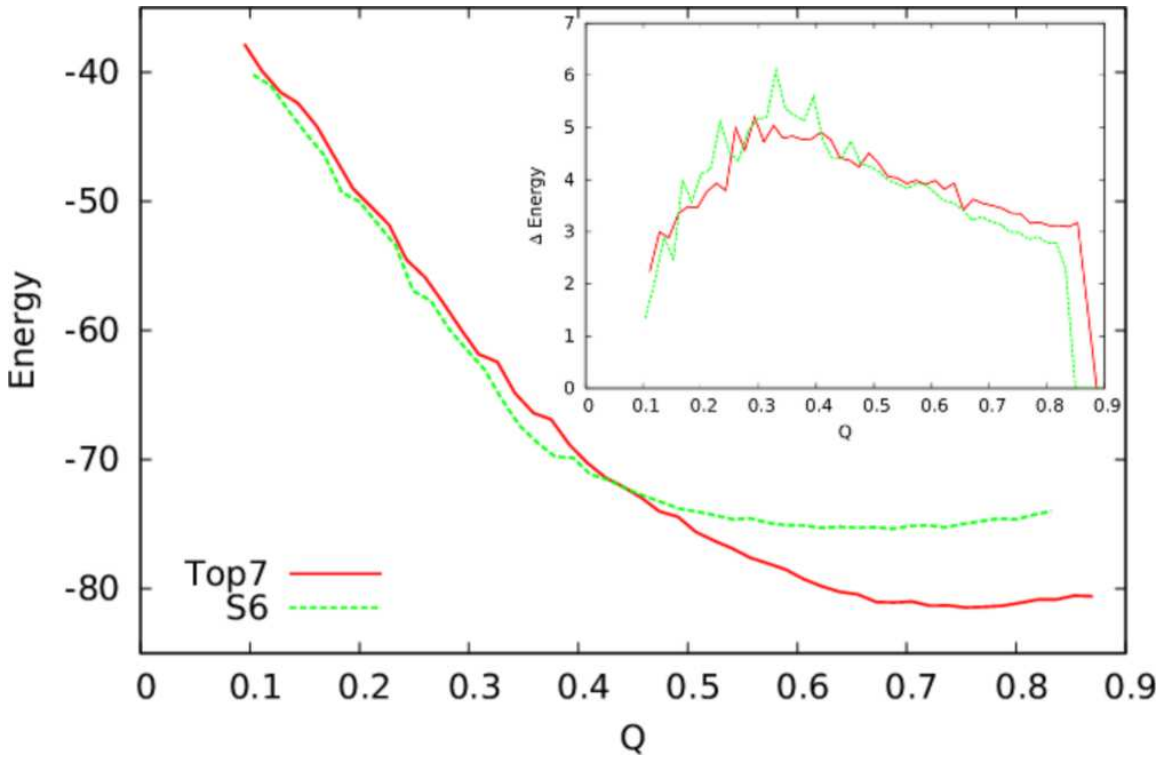


Figure 2.5 : Plot of the tertiary energy term as a function of Q for Top7 (red line) and S6 (green dashed line). Standard deviation of the tertiary energy term (Δ Energy) is shown in the top right corner.

fourth and fifth β strands, an example of which is shown in Figure 2.6B. The structure in Figure 2.6C is a pseudo mirror image of the native structure, which also has an incorrect wiring of the first and third β strands. These structures are energetically competitive in our model because they are compact and retain a full complement of hydrogen bonds as well as a well formed hydrophobic core. Circular dichroism experiments in the Baker laboratory suggested that fragments consisting of helices and subsets of Top7's β strands, including some subsets in which none of the β strands participate in native pairings, are stable in solution. All of this is consistent with Baker's observation, reiterated but unexplored by

Chan, that Top7 might be prone to misfolding via mispairing of its β strands.

The quality of S6's structure prediction is low in comparison to most natural proteins we have studied previously[2]; 19 out of 20 predicted structures have a Q value below 0.4 (Figure 2.7). Predicted structures of S6 have an extra helix in place of the second β strand as seen in two representative structures in Figure 2.7 A and B. Figure 2.7C shows a representative structure taken from an umbrella sampling simulation with a bias centered at $Q = 0.60$ using fragment memory AWSEM. This structure has overall correct topology, but its second β strand still has some helical character. This helical formation in predicted structures of S6 is apparently due to a discrepancy between JPRED's secondary structure prediction (which influence's AWSEM's Ramachandran potential and β hydrogen bonding term) and S6's actual secondary structure. As shown in Table 2.2, the secondary structure prediction of JPRED assigned the second β strand region to be coil.

Both Top7 and S6 have energetic biases toward native-like states in the fragment memory AWSEM model, as seen in the calculations of expectation values of the total, fragment memory and tertiary energy terms (Figure 2.8 B, C and D respectively). The formation of highly native-like states is somewhat disfavored in S6 due to the aforementioned non-native helix formation induced by the (incorrect) assumed secondary structure bias. Top7 has a large energy gap in both the tertiary and fragment memory energy terms. According to analysis using the frustratometer, Top7 has a higher fraction of minimally frustrated contacts and lower fraction of highly frustrated contacts than S6 (Table 2.1), consistent with

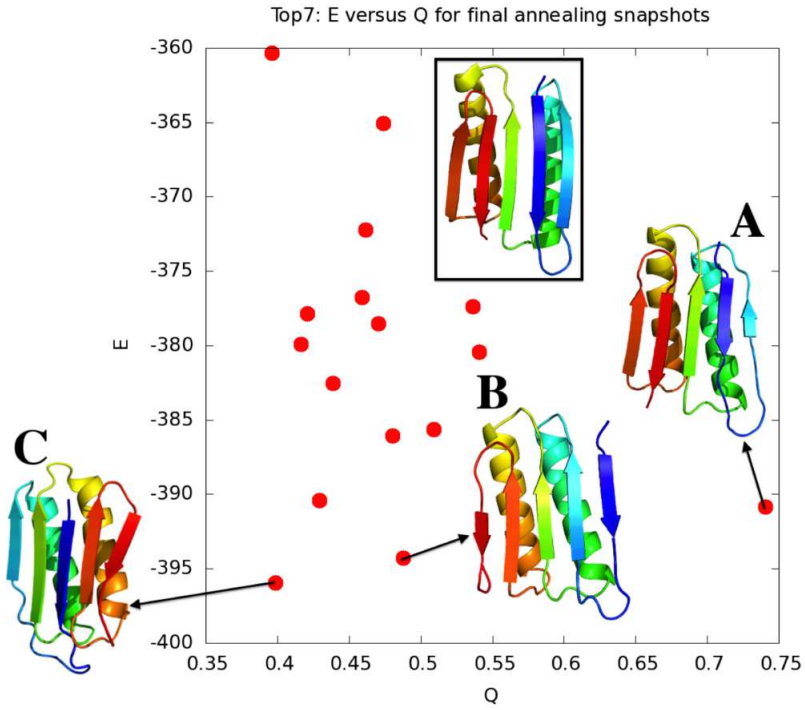


Figure 2.6 : Top7, plot of predicted structures as a function of Q and energy. The crystal structure of Top7 is shown in the rectangle. (A) is a predicted structure of Top7, which is in good agreement with the Top7 x-ray crystal structure, with a $Q = 0.74$ and a RMSD=2.09Å. (B) and (C) are competitive low energy predicted structures of Top7 that have lower Q values, $Q = 0.51$, RMSD=10.09 and $Q = 0.40$, RMSD= 9.46Å , respectively. These structures have all of their secondary structures formed but have incorrect wirings of the β strands.

Top7's unnaturally large hydrophobic core.

2.4.2 TakadaN vs. TakadaZ vs. TakadaE

Figure 2.9A shows the two dimensional free energy profile for TakadaN as a function of Q and the radius of gyration obtained using the structure-based model. Since the structures of TakadaN and the designs are essentially the same, these results would also apply to these

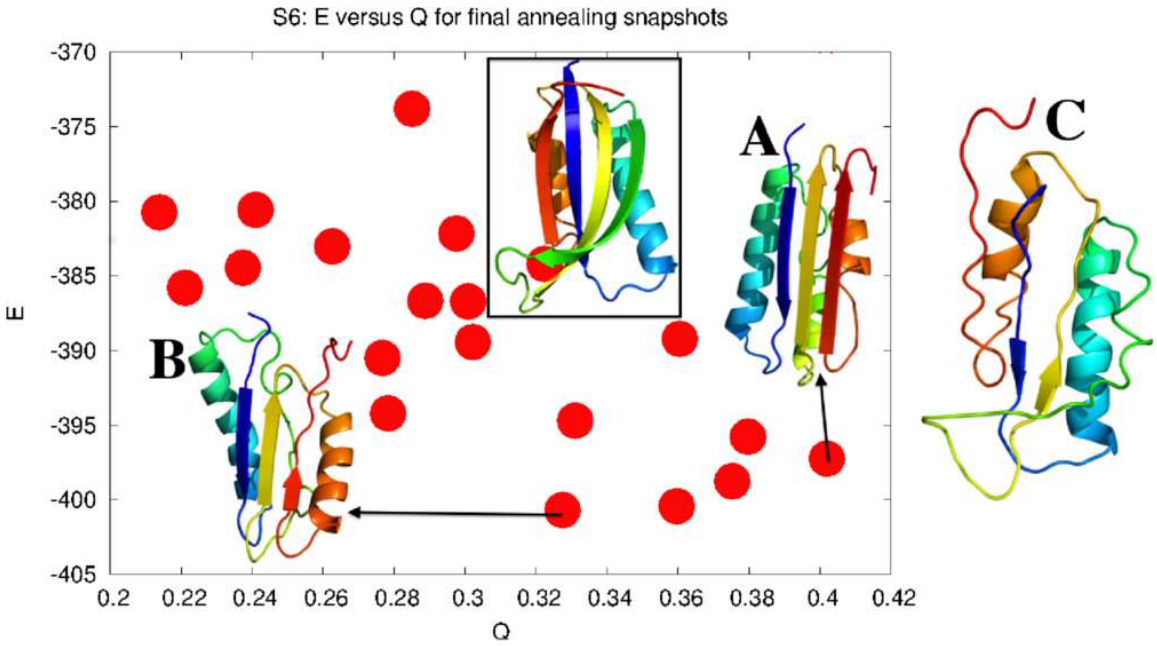


Figure 2.7 : S6, plot of predicted structures as a function of Q and energy. The crystal structure of S6 is shown in the rectangle. (A) is a predicted structure of S6 which has $Q = 0.40$ and a RMSD=9.18Å. (B) is a predicted structure which has the lowest energy with $Q = 0.33$, RMSD=9.91Å . (C) shows a representative structure taken from an umbrella sampling simulation with a biased centered at $Q = 0.6$.

systems. The unfolded and folded states are shown as two low free energy regions separated by a well defined transition state, indicating that the target structure is not topologically frustrated.

To see the effects of tertiary energetic frustration which might distinguish the artificial designs from the natural protein, the free energy profile as a function of Q and energy was also calculated using the AWSEM single memory model for the three Takada sequences at the same temperature and are shown in Figure 2.9 B, C and D. Free energy profiles of TakadaZ and TakadaE (Figure 2.9C and D) are more complex than the free energy profile

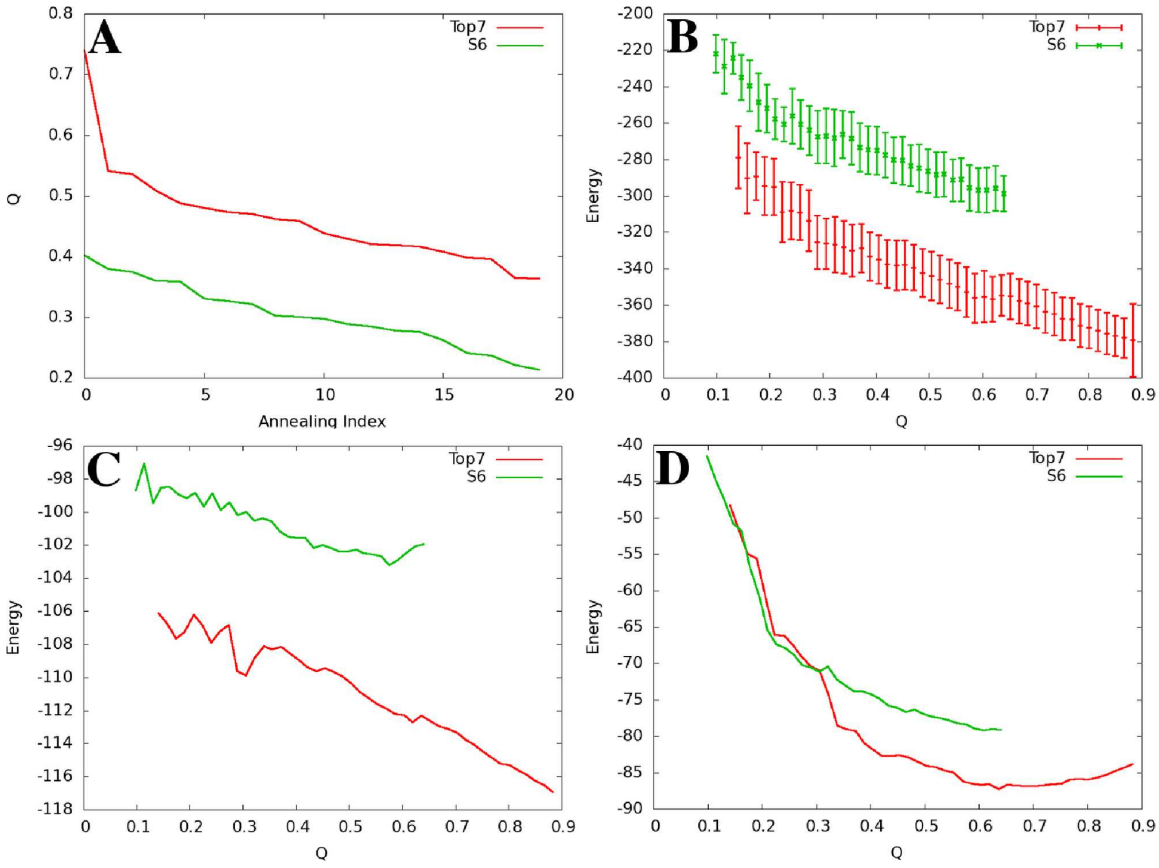


Figure 2.8 : Final Q versus annealing index of Top7 and S6 structure prediction. 20 random simulated annealing simulations were conducted and their final Q values were plotted in the order of decreasing Q from left to right (A). Note that "annealing index" does not refer to the actual order in which the simulations were carried out. Plot of expectation value of the energy term and its standard deviation (B), fragment memory energy (C) and tertiary energy (D) of Top7 and S6. Top7 is shown in red. S6 is shown in green.

of TakadaN.The free energy profile of TakadaN (Figure 2.9B) is funneled to a Q value of 0.75, which is the highest Q value among the three sequences having a common structure, which is consistent with it being the sequence that was used to obtain the relaxed structure. The tertiary energy as a function of Q is shown in Figure 2.10. TakadaE has a larger energy gap than both TakadaN and TakadaZ, and its energetic variance is also the largest among

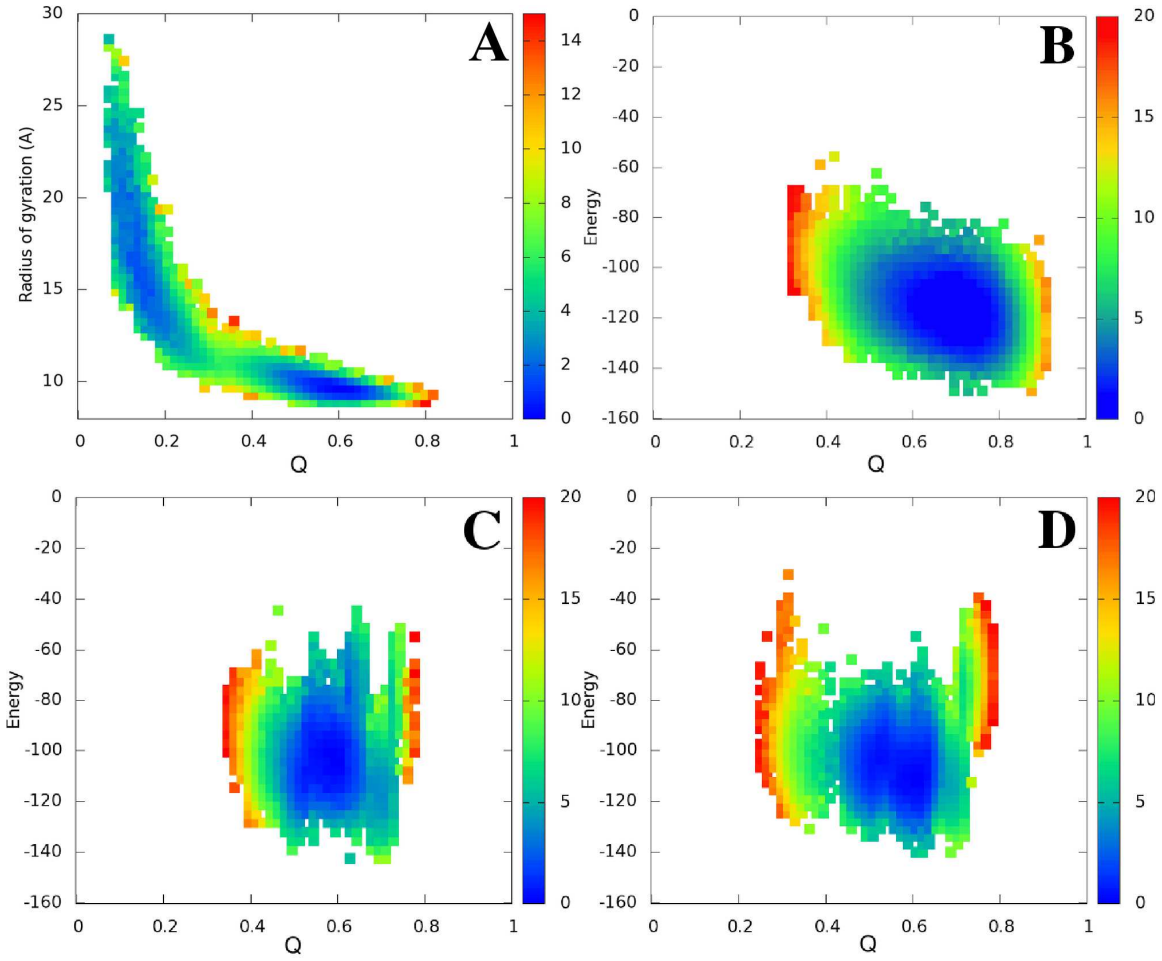


Figure 2.9 : The two dimensional free energy profile of TakadaN as a function of Q and the radius of gyration (A) was computed by using the non-additive structure-based model. The two dimensional free energy profiles of TakadaN (B), TakadaE (C) and TakadaZ (D) were computed by using the single memory AWSEM model. All free energy profiles were calculated at the same temperature.

the three sequences. For the AWSEM energy function, TakadaZ has a variance comparable to the variance of TakadaN, though its energy gap is slightly smaller. The natural protein, TakadaN, has its energy funneled smoothly to high Q value, and a small variance.

Next, predictions were performed with fragment memory AWSEM. TakadaN has the

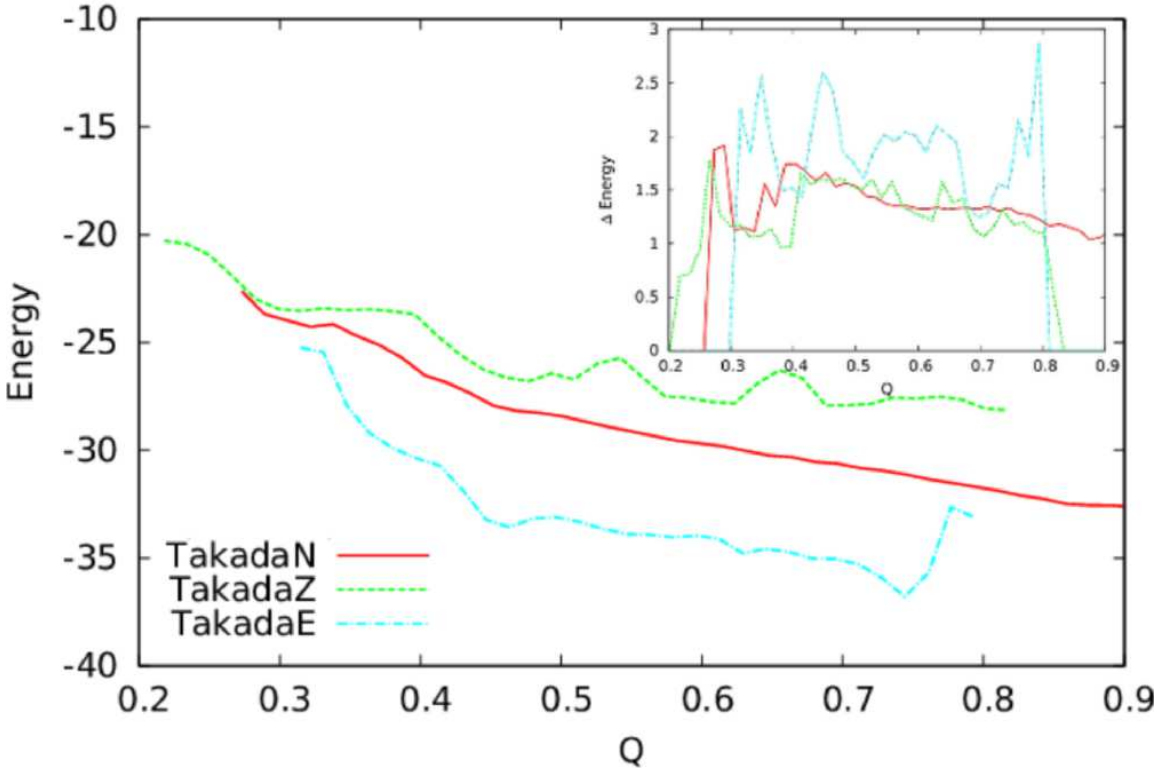


Figure 2.10 : Plot of the tertiary energy as a function of Q for TakadaN (red line), TakadaZ (green dashed line) and TakadaE (blue dot dashed line). Standard deviation of the tertiary energy (Δ Energy) is shown in the top right corner.

best predicted structures found by simulated annealing as shown in Figure 2.11A. Expectation values of the total, fragment memory and tertiary energy terms were calculated and plotted as shown in Figure 2.11 B,C and D. The thermal average of the total energy of all three Takada sequences are well funneled. The energy gap of TakadaE is comparable to TakadaN and is larger than the energy gap of TakadaZ, while the variance in the energies are similar for all three constructs. This explains why TakadaE is better predicted than TakadaZ with AWSEM model. Thus while the Z-score was optimized in the origi-

nal design with the original energy function, the Z-score is not so highly optimized when the AWSEM potential is used. JPRED's secondary structure prediction for the TakadaE sequence is also more similar to that of TakadaN's sequence than is TakadaZ's (as shown in Table 2.2), indicating that its secondary structure propensities match the target structure more closely. Figure 2.12 shows plots of final simulated structures as a function of Q and energy. TakadaZ and E have predicted structures that are scattered over a larger range of Q and energy. TakadaN has less scattered predicted structures, with 13/20 structures clustered at high Q and low energy. For each of these plots, a predicted structure which has the highest Q value and a predicted structure which has the lowest energy are shown. In all cases the lowest energy structures (as well as the highest Q structures) correspond to a correctly predicted overall fold; deviations come mostly in the form of partially formed secondary structures and differences in the details of helix-helix packing.

2.4.3 How robust is folding to sequence simplification?

Figure 2.13 shows the quality of structure prediction of the five proteins using various simplification schemes as compared to the structure prediction performed by simulated annealing on the full sequence. The exact sequences used and their corresponding JPRED secondary structure predictions are given in Tables 2.2. In all five cases, simplifying the sequences using only two amino acid types results in predicted structures with uniformly lower Q values when compared to the predictions using the full encoding or the five letter sequence codes. The simplified sequence using the five letter Miyazawa Jernigan matrix

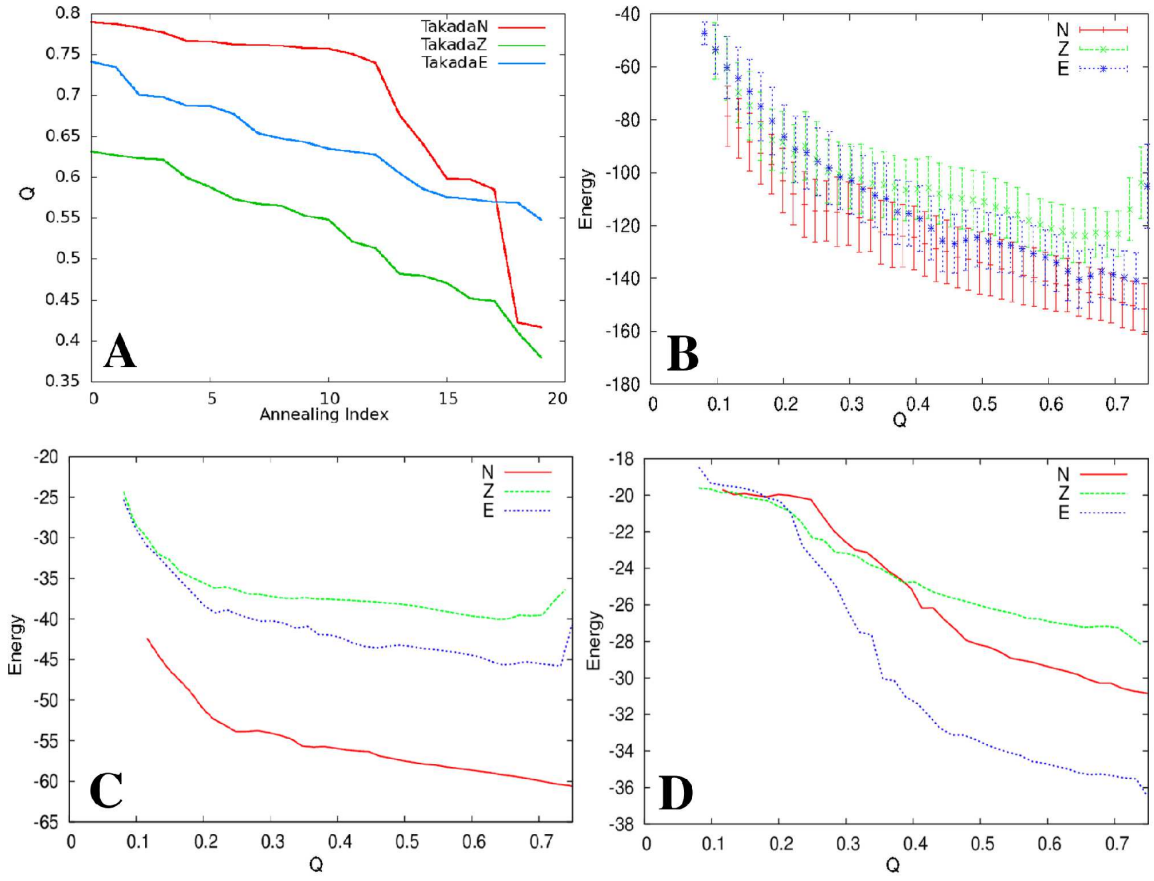


Figure 2.11 : Plot of Q versus annealing index of the three Takada sequences (A).Plots of the expectation value of the total energy (B), the fragment memory energy (C), and the tertiary energy (D). TakadaN is shown in red, Takada Z is shown in green and TakadaE is shown in blue.

scheme (MJ5) yields structure predictions comparable in quality to those for the full sequence in the case of S6 and TakadaE (Figure 2.13 A and B respectively). TakadaZ is actually better predicted when its sequence is reduced to five letters (Figure 2.13C). The quality of structure prediction is slightly reduced when the sequence is reduced to five letter level in the case of Top7, and is significantly reduced in the case of the natural protein TakadaN (Figure 2.13 D and E respectively).

It is typical for proteins to have ≈40% minimally frustrated contacts and 10% highly frustrated contacts (by the "mutational" measure) [8]. These statistics are consistent with the frustration patterns of the full sequence of S6 (41% minimally frustrated contacts / 13% highly frustrated contacts), while the full sequence of Top7 yields a structure with larger fraction of minimally frustrated contacts and fewer highly frustrated contacts (57% and 6% respectively) than is normal for natural proteins; see Table 2.1. The fraction of tertiary interactions that are minimally frustrated remains high and the fraction of highly frustrated interactions remains low for Top7 and S6 when the five letter simplified encoding is employed. The fraction of highly frustrated contacts in Top7 increases from 5% to 10%, contributing to a decrease in the quality of predicted structures when its sequences is simplified using the MJ5 scheme. There is no significant change in the frustration signals of S6 at the level of MJ5 simplification, which results in a comparable quality of structure prediction. TakadaE and TakadaN also show little change in the tertiary frustration signals when they are reduced to the MJ5 code, whereas TakadaZ actually has an increase in the fraction of minimally frustrated contacts.

The local structure frustration can also be important in determining the quality of structure prediction, as is illustrated in by the case of TakadaN. The expectation value of the total energy for its MJ5 simplified sequence has a local minimum along Q around $Q = 0.4$ and a global minimum at much higher Q (Figure 2.14). The origin of this trap can be seen in the expectation value of the fragment memory energy term, which has a wide global minimum between $Q = 0.25$ and $Q = 0.4$, indicating the presence of competing non-

native secondary structures. This makes it difficult for the reduced sequence to fold into the native structure; indeed, only 5/20 fixed-length simulated annealing simulations were able to reach the correct overall fold.

2.5 Discussion

Our study of Top7 indicates that Top7 has a good thermodynamic design as reflected in the local frustration and correct structure prediction. Nevertheless, the non-natural topology of Top7 by itself leads to multiple intermediates that are found using a non-additive structure-based model. These intermediates have folded C-terminal fragments, while the N-terminal fragment remains disordered. These results are consistent with previous experimental [29, 30] and simulation[31] studies.

The role of energetic frustration in the folding of Top7 was also examined by using both single memory and fragment memory versions of AWSEM. The average contact energy in the single memory model, and the average fragment memory, tertiary, and total energies in the fragment memory model all decrease up to high values of Q. However, we find in our simulated annealing simulations several non-native structures that are energetically competitive with the best predicted structure. Although the average energy appears to be well funneled in umbrella sampled data, the existence of these low-energy non-native structures in the simulated annealing simulations indicates kinetic complications with Top7 in our model. Furthermore, the predicted non-native structures are characterized by non-native β strand pairing. Baker and colleagues suggest that canonical nature of four of the five β

strands may be conducive to strand swapping. Also, the middle and slow phases observed in the Chevron plot of Top7 are reported to correspond to states in which no additional surface area is buried, but rather structural rearrangements between collapsed states[30]. The predicted non-native states of Top7 in our model are consistent with both of these ideas. The Z-score of Top7 according to AWSEM looks as though it should be sufficient to exclude possible alternative conformations by chance in the approximation that the molten globule is largely unstructured. Nevertheless, the highly regular and symmetric structure of Top7 apparently allows a small number of discrete competitor states to be significantly populated in solution. The simulated annealing results of AWSEM suggest these are a few structures that are competitive at a coarse-grained level that were not excluded by the elements of heuristic design that were employed to constrain the optimization of the Top7 sequence. These specific competitor structures in the coarse-grained simulations may have energetic packing issues when considered in full atomistic detail.

The truncated natural template of Takada's two designed sequences, TakadaN, has a funneled energy landscape and was found to have the highest quality of structure prediction using AWSEM. Unlike the nicely funneled energy profile for TakadaN using AWSEM, TakadaE and TakadaZ have complex features in their free energy profiles and there is scattered clustering of predicted structures from the simulated annealing runs. TakadaE is slightly better predicted than TakadaZ is, likely because of its larger energy gap and similar energetic variance using the fragment memory predictive model.

We have attempted to assess to what extent funneling and frustration in the energy land-

scape are changed by simplifying the sequences to five (MJ5) and to only two (BL2) amino acid types. Simplified sequences using the five letter Miyazawa-Jernigan matrix scheme produce predicted structures with comparable quality to predicted structures using the full sequences, except in the case of TakadaN, which when simplified now has an energetic trap at around $Q = 0.4$ as a result of competing secondary structures. With the exception of TakadaZ, predictions of full sequences are of better quality than their corresponding MJ5 simplified sequence. This result is consistent with the Frustratometer analysis described previously.

Simplifying to a two letter scheme generally gives lower quality results, as expected from the arguments laid out earlier in this work. These poorer results are partially the result of the sensitivity of the AWSEM model to the input JPRED secondary structure predictions, which influence both the Ramachandran potential and β hydrogen bonding terms the AWSEM potential. JPRED predictions for simplified sequences using the MJ5 mapping agree for the most part with those of the full sequences, consistent with the structure prediction results described previously. With the exceptions of TakadaE and Top7, the JPRED predictions for simplified sequences using the BL2 mapping are drastically different from the JPRED predictions based on the full sequence. Incorrect assignment of residues as being β is the most common anomaly, often resulting in deformed secondary structure in poorly predicted structures (structures not shown). JPRED predictions of BL2 simplified sequences for Top7 and TakadaE are notably more similar to those of the full sequence. Many of the predicted structures of the Top7 BL2 sequence have correctly formed sec-

ondary structure. Nevertheless, incorrect pairing of β strands is still frequently observed. This is expected as the energetic heterogeneity of the full sequence that potentially encodes the specificity of pairing has been completely lost in simplification.

Our results suggest that a five letter code may contain sufficient information for structure prediction of *de novo* designed sequences but may not be sufficient for natural proteins. TakadaN showed the most dramatic change in prediction quality upon MJ5 simplification, due to local in sequence frustration. It remains unclear how many flavors of amino acids are required to fold simplified sequences of natural proteins with as much accuracy as their native sequence. In contrast to some natural proteins, the three designed proteins examined in this work can be folded using a smaller number of amino acid types. Evolution has tuned natural sequences over millions of years to both fold and to function. Though *de novo* designed sequences are indeed proving to be intelligently designed, being able to fold into stable structures, they also seem to be less sensitive to simplification, perhaps implying that they are relying on less subtle signals than natural proteins.

Acknowledgments

We are grateful to Diego Ferreiro and Joachim Lätzer, who carried out preliminary calculations on Top7 several years ago, spearheading this more complete investigation. The project described was supported by Grant R01 GM44557 from the National Institute of General Medical Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of National Institute of General Medical Sci-

ences or the National Institutes of Health. Additional support was also provided by the D.R. Bullard-Welch Chair at Rice University. This work was also supported in part by the Data Analysis and Visualization Cyberinfrastructure funded by NSF under grant OCI-0959097 and the Shared University Grid at Rice funded by NSF under Grant EIA-0216467, and a partnership between Rice University, Sun Microsystems, and Sigma Solutions, Inc.

Table 2.2 : Sequences and secondary structure information. Beneath each sequence, its respective JPRED secondary structure prediction is given. “H”, “E” and “-” correspond to α-helix, β-strand and coil, respectively. The STRIDE assignment was converted into this representation by mapping all “AlphaHelix” assignments to “H” and all “Strand” assignments to “E”. All other types of STRIDE assignments were mapped to “-”.

Top7	
STRIDE	-EEEEEEEE----EEEEEEEE---HHHHHHHHHHHHHHHH---EEEEEEE---HHHHHHHHHHHHHHHHHH---EEEEEEE--EEEEEEE-
FULL	DIQVQVNIDDNGKNFDYTYTVTTESELQKVLNELMDYIKKQGAKRVRISITARTKKEAEKFAAILIKVFAELGYNDINVTFDGDTVTVEGQL --EEEEEE--------EEEEEE--HHHHHHHHHHHHHHHH----EEEEEEE---HHHHHHHHHHHHHHHHH-----EEEEE---EEEEE---
MJ5	EIKIKIKIEEKGKKIEIAIAIAAEKEIKKIIKEIIEIIKKKGAKKIKIKIAAKAKKEAEKIAAIIIKIIAEIGIKEIKIAIEGEAIAIEGKI --EEEEEE-----EEEEEEEEE-HHHHHHHHHHHHHHHH----EEEEEEE--HHHHHHHHHHHHHHHHHH---EEEEEE---EEEEE---
BL2	SISISISISSSSSSISISISISSSSSISSIISSIISIISSSSSSSISISISSSSSSSSSSISSIIISIISSISSISISISSSSSISISSSI -EEEEEEEE-----EEEEEEEE----HHHHHHHHHHEEE------EEEEEEE--------HHHHHHHHHH--EEEEEEEEEE---EEE----
S6	
STRIDE	-EEEEEEEEE-----HHHHHHHHHHHHHHHHHH--EEEEEEEEEEEEEEEEE--EEEEEEEEEEEEE----HHHHHHHHH----EEEEEEEE-----
FULL	MRRYEVNIVLNPNLDQSQLALEKEIIQRALENYGARVEKVEELGLRRLAYPIAKDPQGYFLWYQVEMPEDRVNDLARELRIRDNVRRVMVVKSQEPF ----EEEEEE-----HHHHHHHHHHHHHHHHH---EEEEE----------------EEEEEEEEE----HHHHHHHHH-----EEEEEEEEE----
MJ5	IKKIEIKIIIKGKIEKKKIAIEKEIIKKAIEKIGAKIEKIEEIGIKKIAIGIAKEGKGIIIIIKIEIGEEKIKEIAKEIKIKEKIKKIIIIKKKEGI ----EEEEEE-----HHHHHHHHHHHHHHHHH---EEEEE----------------EEEEEEEEE---HHHHHHHHH-----EEEEEEEEE----
BL2	ISSISISIIISSSISSSSISISSSIISSSISSSISSISISISISISSSSSSIIIIISISISSSSISSISSSISISSSISSIIIISSSSSSI ---EEEEEEEE------EEEEEEHHHH-HHHHH-----EEEEEEEEEEEEEEE-----EEEEEEEEE----EEEEE-EEEEE---HEEEEEE-----
TakadaN	
STRIDE	--HHHHHHHHHHH------HHHHHHHHH---HHHHHHHHHHHH----
FULL	LKNAKEDAIAELKKAGITSDFYFNAINKAKTVEEVNALKNEILKAHA --HHHHHHHHHHH----HHHHHHHH-----HHHHHHHHHHHH---
MJ5	IKKAKEEAIAEIKKAGIAKEIIKAIKKAKAIEEIKAIKKEIIKAAA --HHHHHHHHHHH----HHHHHHHH----HHHHHHHHHHHH--
BL2	ISSSSSSSISSISSSSISSSIIISSISSSSSISSISSISSSIISSSS --------EEEE-------EEEEEEE-----EEEEEEE---------
TakadaZ	
STRIDE	--HHHHHHHHHHH------HHHHHHHHH---HHHHHHHHHHH----
FULL	RGNDAKKAAARWKDRKFKFKAFIHRMDSFGAITEIHKAASAYAKKFG ----HHHHHHH----HHHHHHHH------HHHHHHHHHH-----
MJ5	KGKEAKKAAAKIKEKKIKIKAIIAKIEKIGAIAEIAKAAKAIAKKIG ----HHHHHHHHH-HHHHHHHHHHHHHHHHHHHHHHHHH---
BL2	SSSSSSSSSSISSSSISSIISSISSISSISSISSSSSSISSSIS -----------EE--EEEHHHHHHHH-----EEEE----------
Takada E	
STRIDE	--HHHHHHHHHHH------HHHHHHHHH---HHHHHHHHHHH----
FULL	AYKFAETFFEQWKKFGWQIKYFLEYMRRAGGAKKFYEMIRRWIKEGW --HHHHHHHHHHHH--HHHHHHHHHH-----HHHHHHHHHH---
MJ5	AIKIAEAIIEKIKKIGIKIKIIIEIIKKAGGAKKIIEIIKKIIKEGI --HHHHHHHHHHHHHEEEEEEHHHHHH----HHHHHHHHHH---
BL2	SISISSSIISSISSISISIIIISIISSSSSSSSIISIISSIISSSI -----HHHHHHHHHHHHHHHHHHEEEE------HHHHHHHHHH---

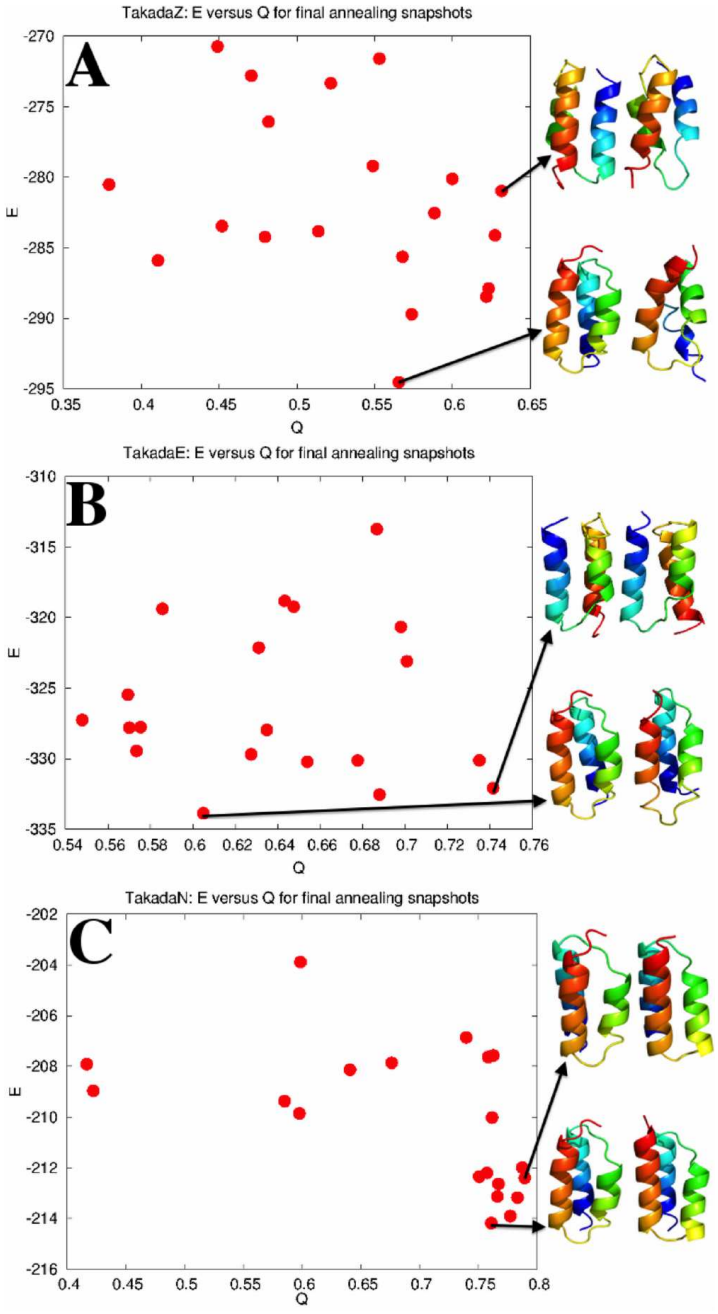


Figure 2.12 : Plots indicating where the final simulated annealed structures lie as a function Q and energy for TakadaZ (A), TakadaE (B) and TakadaN (C). Predicted structures are shown on the right along side a view of the target structure that has been rotated in order to highlight the differences. In all cases the lowest energy structures (as well as the highest Q structures) correspond to a correctly predicted overall fold; deviations come mostly in the form of partially formed secondary structures and differences in the details of helix-helix packing.

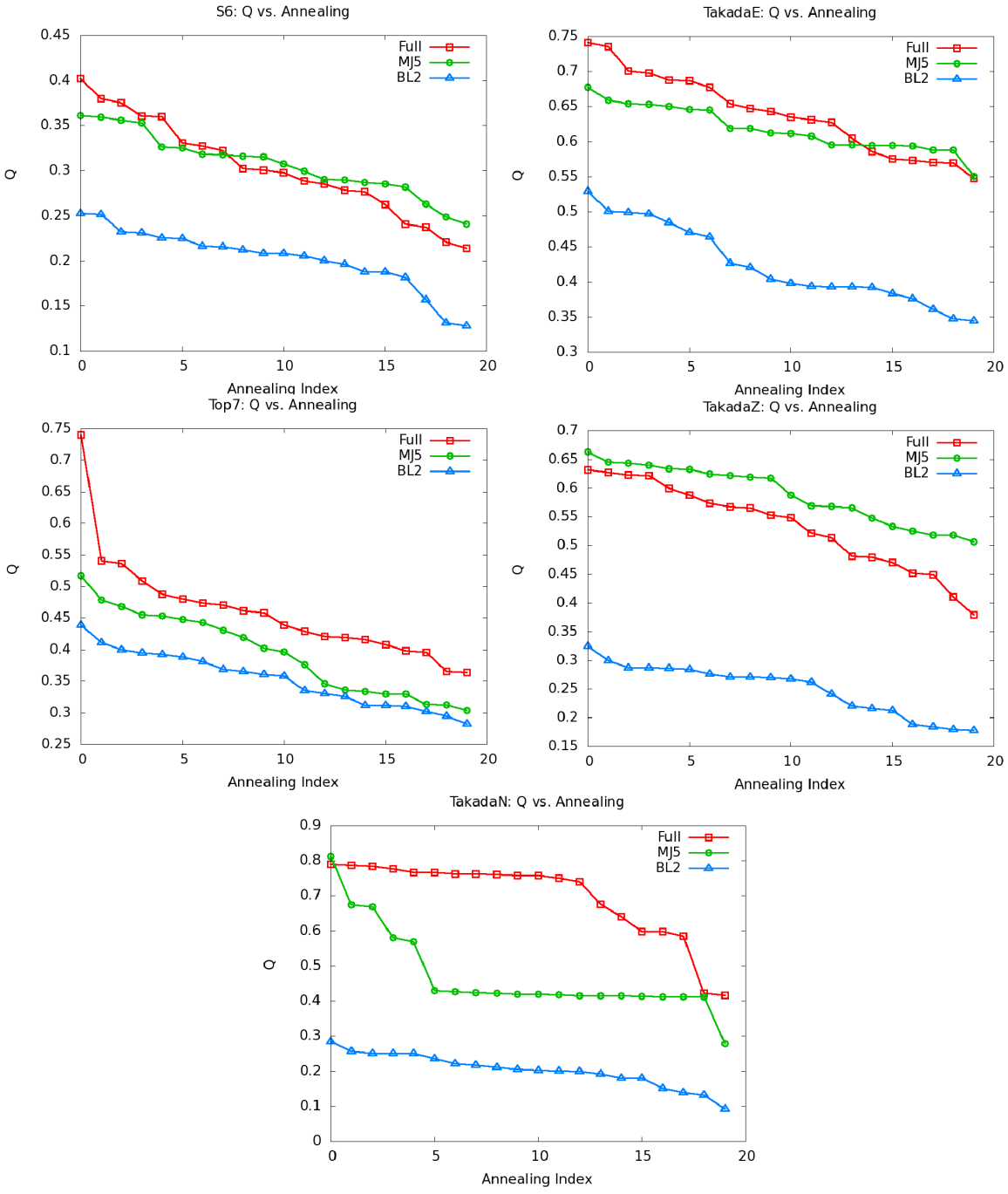


Figure 2.13 : Q versus annealing index is plotted with decreasing Q from left to right, summarizing the quality of structure prediction upon simplification of the sequences: S6 (A), TakadaE (B), Top7 (C), TakadaZ (D) and TakadaN (E). The full sequence is shown in red, the five letter simplified sequence is shown in green, and the two letter simplified sequence is shown in blue.

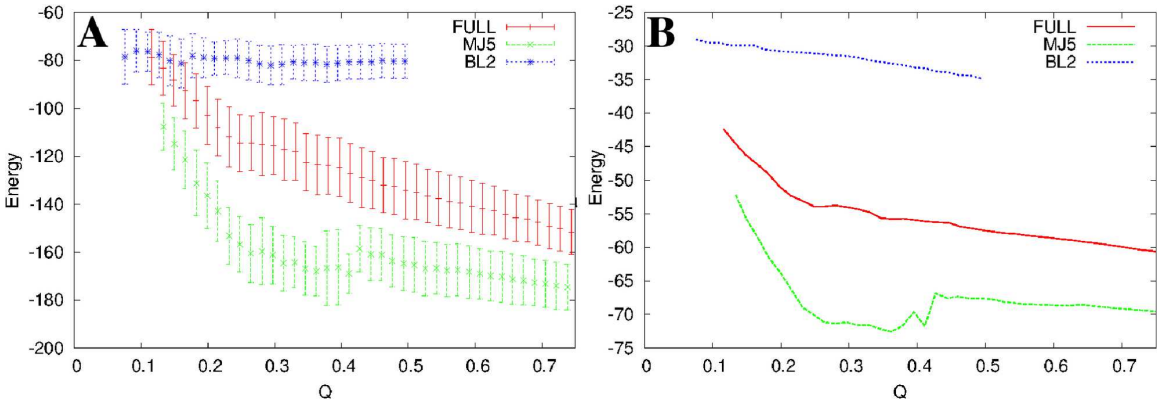


Figure 2.14 : Plots of the expectation value of the total energy (A) and fragment memory energy (B) of TakadaN at different levels of simplification. The full sequence is shown in red, the five letter simplification is shown in green and the two letter simplification is shown in blue.

Chapter 3

Predictive Energy Landscapes for Folding Membrane Protein Assemblies

3.1 Abstract

We study the energy landscapes for membrane protein oligomerization using AWSEM-Membrane, a coarse-grained molecular dynamics model previously optimized under the assumption that the energy landscapes for folding α-helical membrane protein monomers are funneled once their native topology within the membrane is established. In this study we show that the AWSEM-Membrane force field is able to sample near native binding interfaces of several oligomeric systems. By predicting candidate structures using simulated annealing, we further show that degeneracies in predicting structures of membrane protein monomers are generally resolved in the folding of the higher order assemblies as is the case in the assemblies of both nicotinic acetylcholine receptor and V-type Na^{+}-ATPase dimers. The physics of the phenomenon resembles domain swapping, which is consistent with the landscape following the principle of minimal frustration. We revisit also the classic Khorana study of the reconstitution of bacteriorhodopsin from its fragments, which is the close analogue of the early Anfinsen experiment on globular proteins. Here we show the retinal cofactor likely plays a major role in selecting the final functional assembly.

3.2 Introduction

Membranes in the cell are packed with proteins that make up roughly 50 percent of their volume. In such a crowded environment, it is no surprise that most membrane proteins form parts of larger oligomeric assemblies. As in the case of globular proteins, the energy landscape for folding membrane proteins and the landscape for forming complexes from them must be intimately connected. Ultimately, folding and assembly of proteins within a membrane must rely on the same forces, being modulated by the membrane environment. The fidelity of these processes is the result of aeons of evolution. The mechanistic consequences of the landscapes also must be similar since, apart from the higher local concentration in folding, it is impossible to distinguish the docking events that lead to an oligomeric assembly from the motions needed to organize a fully covalently connected single chain. For globular proteins, the intimate relation between folding and binding landscapes had been well documented a decade ago[55]. Nevertheless, the investigation of the globular protein binding landscapes led to the realization that, in the aqueous environment, water mediated hydrophilic interactions are needed to augment the well known hydrophobic forces if interfaces are to be predicted with accuracy[49, 56, 57, 58]. This experience leads us to inquire whether the coarse-grained force field models found successfully to predict the tertiary folds of membrane proteins in their milieu also will suffice to predict the structure of larger membrane protein assemblies. In this paper, we explore this issue by using AWSEM-Membrane simulations to carry out tertiary structure prediction on several membrane protein assemblies and by analyzing the major basins on the free energy land-

scapes of this model both for the complexes as a whole and for their constituent monomers in both the absence and presence of a binding partner.

We show here that a transferable, coarse-grained force field inferred using an energy landscape algorithm for folding monomeric membrane proteins also suffices to assemble larger complexes and can predict their tertiary structure at moderate resolution. This force field, called AWSEM-Membrane, uses an energy function of the same form as has been used for folding globular proteins: the Associative Memory Water-mediated Structure and Energy Model[2]. The parameters, however, were re-optimized using a database of individual membrane protein domains[59]. Very often, in fact, we find the tertiary structures predicted in the context of the multimeric assembly using AWSEM-Membrane are better than those that would be predicted when the monomer is studied in isolation. Degenerate free energy basins in the monomer free energy landscape often turn out to involve forming, internally, native quaternary contacts that have apparently evolved to put together the full multiunit protein assembly. Thus the near degenerate mis-predictions of the monomer actually correspond to what may be termed internally domain swapped structures.

We also examine the question of the mechanism and landscapes of folding versus assembly in the context of the classic investigation carried out by the Khorana group of the assembly of functional bacteriorhodopsin from cleaved fragments. In that classic study, correct reconstitution and proper assembly were tested spectroscopically by whether retinal binding forms a purple product which shows retinal has found its proper protein environment[60]. Here we show that the AWSEM-Membrane energy landscape predicts

the initial assembly involves a misfolded species but that, once constraints consistent with the retinal contacts are added, proper assembly follows. This suggests the reconstitution mechanism involves significant rearrangement after the initial fragment binding occurs. This is consistent with the relatively slow time scale of the reconstitution process[61].

3.3 Methods

The AWSEM-Membrane code was recently described in detail in Ref. 6. It is instantiated as the open source LAMMPS simulation code[48]. In some of the simulations carried out in this paper, we used the single memory (SM) AWSEM-Membrane model, in which the associative memory term is determined by the structure of the monomer in the experimentally determined structure, to elucidate the role of oligomerization in eliminating the non-native packing of the individual subunit and to predict the structure of dimer complexes of various proteins. In other simulations, we used the fragment memory (FM) AWSEM-Membrane model, in which local-in-sequence interactions are derived from a database of structures by performing homology searches using short fragments of the target sequence [2]. AWSEM-Membrane employs a coarse-grained backbone description wherein the position and orientation of each amino acid residue are dictated by the positions of its C_α, C_β and O atoms (except Glycine, which lacks a C_β atom). Full details of the functional form of the potential can be found in the SI of Ref. 2 and the re-optimized, membrane protein specific parameters can be found in the SI of Ref. 6.

We only focus on studying the second step of the membrane protein folding process,

which occurs after the protein conformations have already been restricted to have a proper topology within the membrane. The topology, by which we mean the "specification of the number of transmembrane helices and their in and/or out orientation across the membrane" [62, 63], were obtained directly from the three dimensional experimentally determined structure using the TMDET web server [64]. Similar results are expected if *a priori* predicted topologies[65] (which are generally quite good) would be used as input, as in our previous study [6].

In order to sample along Q_w, the fraction of pairwise distances within 1Å of their distances in the native structure (Equation 1.4), and to construct free energy profiles, we ran umbrella sampling simulations in which a harmonic bias (given in Equation 1.5) was added to the Hamiltonian. All free energy profiles and expectation values were calculated using the multi-state Bennett acceptance ratio (MBAR) method as implemented in the pyMBAR package [15]. Samples were collected for a range of temperatures near the empirically determined binding temperature of each system.

3.4 Results

AWSEM-Membrane is an implicit force field model built on the hypothesis that funneled energy landscapes drive the folding of α-helical membrane protein monomers within their native topological sector, which is that part of conformational space wherein all the transmembrane helices have adopted the same orientation with respect to the membrane as is seen in the final folded structure. The parameters of the coarse-grained force field have

been optimized using an algorithm based on the minimal frustration principle[7]. This algorithm developed by Goldstein et. al. [66] involves statistically optimizing a Z-score that is monotonically related to the ratio of the folding temperature over the glass transition temperature, T_f/T_g. The folding temperature is the temperature at which there are equal equilibrium populations of the low energy, low entropy native state and the high energy, high entropy denatured ensemble. The glass transition temperature is the temperature at which a sequence would be expected to become trapped in one of many degenerate low energy structures if that sequence has not been optimized by evolution to fold to a nearly unique native structure. The ratio of these two temperatures is a measure of the degree to which sequence evolution has lead to a bias guiding the protein from all parts of the conformational space towards the native state during conformational search. High values of T_f/T_g indicate a large bias and high specificity in the folded structure. Energy landscapes that have large biases lead to rapid folding and are said to be funneled[67, 20, 21, 22, 68].

The optimization of the AWSEM-Membrane model parameters was based on a non-redundant database of α-helical membrane protein monomer structures [6]. Following the earlier study of the folding landscapes of α-helical membrane monomers in isolation, we now in this paper examine whether the AWSEM-Membrane code can predict the binding interfaces of oligomeric systems. The calculations resolve the issue of whether degenerate tertiary packings that are found in the free energy landscapes of nicotinic acetylcholine receptor subdomain (2BG9) and V-type Na^+-ATPase subdomain (2BL2) monomers are the result of domain swapping[69, 55]. We also revisit the Khorana study of re-association

of fragments of bacteriorhodopsin [60] which historically plays the role for membrane proteins that Anfinsen's work did for globular proteins[70].

3.4.1 Prediction of the structures and binding interfaces of membrane protein complexes.

We carried out structure prediction studies aimed toward studying the binding interfaces of chain A and chain B of nicotinic acetylcholine receptor subdomain (2BG9)[71], V-type Na^+-ATPase (2BL2)[72], and bacteriorhodopsin (1BRR)[73] using simulated annealing of the AWSEM-Membrane force field, which is implemented in the open-source LAMMPS simulation package[48]. The molecular dynamics simulations begin with two folded monomers separated by approximately 80Å. A single, weak and non-specific spring potential was used to ensure that the centers of mass of the monomers would be brought together during the course of the annealing. These simulations start at such a high temperature that the flexible monomers are allowed to explore many possible internal conformations and binding interfaces. Following this, the thermostat's temperature was slowly reduced to a quenching temperature. We evaluate the quality of the structures using both Q_i, the fraction of native interface contacts formed, and the C_α root-mean-square deviation (RMSD).

Figure 3.1 shows the predicted quenched structures which have the highest final Q_i of such simulations for nicotinic acetylcholine receptor subdomain (2BG9) dimers, V-type Na^+-ATPase (2BL2) dimers, and bacteriorhodopsin (1BRR) dimers. These results show

that the single memory AWSEM-Membrane energy landscape gives accurate predictions for the interfaces of the nicotinic acetylcholine receptor subdomain (2BG9) and the V-type Na^+-ATPase (2BL2), with Q_i equal to 0.681 and 0.782 respectively. The predicted C_α RMSD values are both less than $3.5\mathring{A}$. The dimer interface of V-type Na^+-ATPase (2BL2) was better predicted, according to both its larger Q_i and its smaller overall C_α RMSD, than the nicotinic acetylcholine receptor subdomain dimer (2BG9) despite the larger size of the ATPase domain, likely due to its simpler free energy landscape (Figure 3.4) and larger interface. The bacteriorhodopsin dimer complex turns out to be harder computationally to sample during the simulation because it is a significantly larger system (460 residues) than either nicotinic acetylcholine receptor subdomain (2BG9) (182 residues) or V-type Na^+-ATPase (2BL2) (312 residues). Nonetheless, the predicted structure of the bacteriorhodopsin dimer shows that a significant fraction of native interface contacts are formed ($Q_i = 0.415$), and the overall structure is quite native like ($RMSD = 5.732\mathring{A}$). It should be noted that the experimentally determined structure of bacteriorhodopsin to which we are comparing has a retinal molecule situated in the core of each monomer. This cofactor is omitted from the AWSEM-Membrane prediction simulations just described. The absence of the cofactor allows distortions in helical packing which would likely be prohibited due to excluded volume when the retinal cofactor is present.

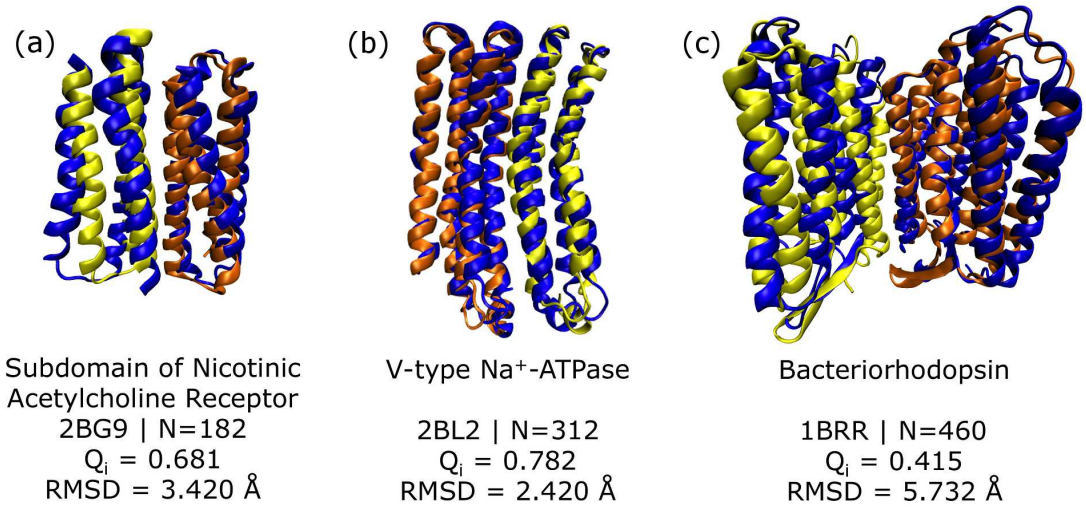


Figure 3.1 : Structures with the highest final Q_i values from ten simulated annealing runs using the AWSEM-Membrane force field for (a) nicotinic acetylcholine receptor subdomain, (b) V-type Na^+-ATPase, and (c) bacteriorhodopsin dimers. In all cases one of the chains, chain A, is colored in yellow, and the other, chain B, is colored in orange, and the experimental structure of the complex, obtained from the Protein Data Bank[5], is colored in blue. The names of the proteins, their PDB ID, and the number of residues are shown below each structure. The fraction of native interface contacts, Q_i and the C_α RMSD of the complex compared with the experimental structure indicate the quality of the AWSEM-Membrane predictions.

3.4.2 Non-native helical packings found for the monomers in isolation are disfavored in the presence of their binding partners.

In the previous study of individual monomers, the free energy landscape analysis for nicotinic acetylcholine receptor subdomain (2BG9) revealed the presence of two nearly degenerate free energy basins. These basins have similar contact maps, but only one corresponds to structures with the native helical packing, while the other basin centers on a structure that is a pseudo-mirror image of the native structure. The actual contacts that are made in

both basins are nearly the same [6]. Since both structural ensembles possess a high fraction of native contacts, they are essentially degenerate according to the contact energy term, $E_{contact}$. Free energy landscape analysis of V-type Na^{+}-ATPase subdomain (2BL2) also revealed the presence of two degenerate free energy basins for the AWSEM-Membrane force field. Again, one basin corresponds to the native helical packing but the other basin was characterized as having a non-native helical packing. These two structural ensembles also could not be distinguished by their contact energies. In the earlier study, these proteins were simulated as monomers, but in nature both proteins are part of larger multimeric assemblies [71, 72]. Is this degeneracy of alternative tertiary packings resolved when simulating a monomer in the presence of one of its binding partners? We now answer this question by carrying out free energy landscape analysis for both systems using two instantiations of the AWSEM-Membrane prediction scheme. In the first scheme, the local-in-sequence forces determined by the fragment memory term were chosen by performing a homology search of short sequence segments in a database of sequences with known structures and using these as the input associative memories. This approach mimics what must be done when using AWSEM models for fully *de novo* structure prediction. In the second scheme, only a single memory is used for the short range interactions such that only proper native secondary structure information is incorporated. Thus, this landscape has less secondary structure frustration than does the landscape that uses multiple inputs in the fragment memory term.

To be certain to sample a wide range of configurations efficiently, we used umbrella

sampling along the collective coordinate Q_w, a similarity measure based on the fraction of native pairwise distances, and use these sampled structures to construct two-dimensional free energy profiles $F(Q_i, RMSD)$ for the nicotinic acetylcholine receptor subdomain (2BG9) dimer and for the V-type Na^+-ATPase (2BL2) dimer. We employ the multi-state Bennett acceptance ratio (MBAR) method [15] to calculate free energy profiles and expectation values. Free energy is in k_BT, in which T was chosen to be below the folding temperature of the monomer but high enough to sample multiple bound configurations. The two order parameters used in the profile are Q_i, the fraction of native interface contacts, and RMSD, the C_α root-mean-square deviation. We also computed the expectation value of the potential energy (PE) and the contact energy ($E_{contact}$) for each system and display these also as two dimensional profiles with respect to Q_i and RMSD.

The experimentally determined native binding interface of the first two chains (A and B) of the nicotinic acetylcholine receptor subdomain (2BG9) complex consists of two specific helix-helix interactions as shown in Figure 3.3(d): one of these interactions involves the association of the first helix of Chain A and the third helix of Chain B, (A1, B3), and the other involves the docking of the second helix of Chain A to the second helix of Chain B, (A2, B2). The free energy profile of the nicotinic acetylcholine receptor subdomain (2BG9), shown in Figure 3.2(a), has three low free energy basins. We characterize the contact maps of representative structures from each of these basins in Figure 3.3. Structures in basin 2 are the most native-like in structure. Structures in this basin have well-formed intra-monomer contacts and both native binding interface helix-helix interactions, (A1, B3)

and (A2, B2), as shown in Figure 3.3(b). Structures in basin 3 also contain the native binding interface contacts (A1, B3) and (A2, B2), but in this basin the intra-monomer contacts between helices (A2, A3) and (B2, B3) are disrupted (Figure 3.3(c)). Structures in basin 1, while successfully forming interface contacts between helices (A1, B3), lack stable native interface interactions between helices (A2, B2) and intra-monomer helix-helix interactions between (B2, B3) (Figure 3.3(a)). Structures in all three basins show some degree of over-collapse resulting from formation of non-native interface contacts. Nicotinic acetycholine receptor is a pentamer in its crystal structure, and this over-collapse will likely be resolved when folding the complete multimeric assembly. The observed over-collapse may also be a consequence of the generic cylindrical radius of gyration bias (R_g) employed in the AWSEM-Membrane code. This bias is similar to typical R_g bias used for globular proteins, but is applied only to coordinates in the membrane plane. This constraint was originally implemented in order to mimic the lateral pressure of the membrane lipid molecules on the protein that gives rise to the liquid crystalline-like ordering of helices in membrane proteins.

Figure 3.2(c) and Figure 3.2(d) show two dimensional profiles of the expectation values of the potential energy, PE, and of the contact energy, $E_{contact}$, for the nicotinic acetycholine receptor subdomain (2BG9) dimer, respectively. The full potential energy which includes both contact terms and associative memory terms appears to favor basins 1 and 2, while the contact energy landscape by itself dominantly favors basin 3 and moderately favors basin 2. Basin 3 is disfavored in the full potential energy landscape primarily be-

cause of the distortion of the intra-monomer structure of both chains in structures found in this basin. These configurations are energetically penalized by E_{SM}, the single memory term which favors proper secondary structure. Conversely, structures in basin 2 have fully native-like intra-monomer structure of both chains. Basin 1 is slightly less favored in the full potential energy landscape than is basin 2 due to the distortion observed in the intra-monomer structure of chain B as discussed above. Why are basins 1 and 3 favored when only the contact energy is considered? The over-collapse of both states allowed by the local distortion simply leads to a larger gross number of contacts formed when compared to basin 2, as is evident in the contact maps in Figure 3.3.

In Figure 3.4, we show the free energy profiles for the V-type Na^{+}-ATPase (2BL2) dimer. Here, the properly folded structure is strongly favored: the landscape has only one basin at low RMSD (between $2\mathring{A}$ and $3.5\mathring{A}$) and high Q_i ($Q_i \geq 0.7$), which corresponds to the native structure. The expectation values of the total potential energy, PE, and contact energy, $E_{contact}$, both show the native conformation to be the most stable.

Figure 3.5 shows the results of the free energy landscape analysis for the nicotinic acetylcholine receptor subdomain (2BG9) and the V-type Na^{+}-ATPase (2BL2) dimer complexes using fragment memory AWSEM-Membrane. The two dimensional free energy profile of nicotinic acetylcholine receptor subdomain (2BG9) dimer complex (Figure 3.5(a)) reveals two low free energy basins. Representative structures from the low RMSD basin are significantly native-like, forming native binding interface helix-helix interactions (A1,

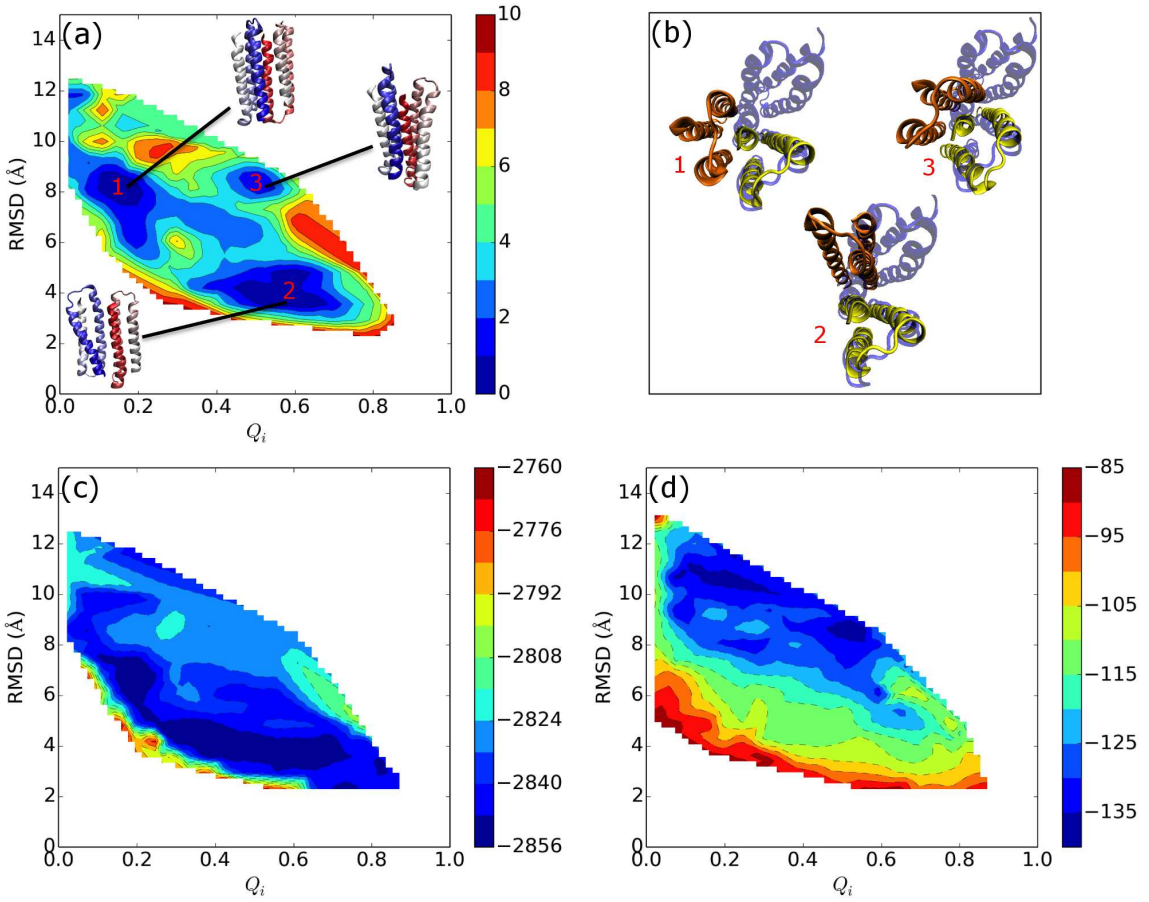


Figure 3.2 : (a). Free energy profile of the nicotinic acetylcholine receptor subdomain (2BG9) dimer complex obtained using single memory AWSEM-membrane. The free energy is plotted versus Q_i, the fraction of native interface contacts (x-axis) and the RMSD (y-axis). Representative structures are shown from the three free energy basins. (b): Top views of representative structures from each low free energy basin (chain A is colored in yellow, chain B is colored in orange), with the native structure (shown in transparent blue). Expectation values of (c) the total potential energy, PE, and (d) the contact energy, $E_{contact}$, are plotted versus the same order parameters.

B3) and (A2, B2), with moderate helix distortion. The representative structure from the high RMSD basin however exhibits a non-native association of the monomers such that the non-native helix-helix interactions (A1, B2) and (A3, B3) are formed. However, both sub-

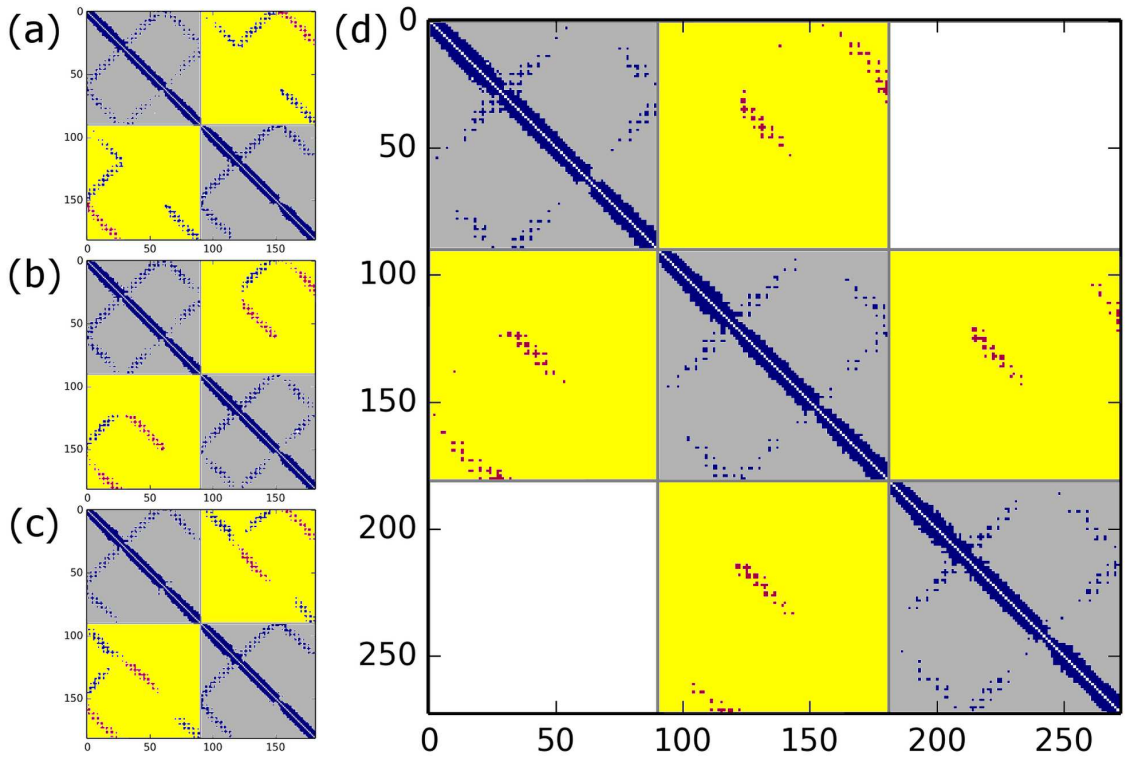


Figure 3.3 : Contact maps of representative structures obtained from basins of the free energy profile of nicotinic acetylcholine receptor subdomain (2BG9) dimer complex (contact maps (a), (b), and (c) correspond to structures 1, 2 and 3 in Figure 3.2, respectively) and (d) contact map of nicotinic acetylcholine receptor subdomain (2BG9) trimer including chain A, chain B and chain E of the full pentamer complex. Sections of the maps that show intra-monomer contacts are colored in gray. Sections of the maps that show inter-monomer contacts are colored in yellow. Inter-monomer contacts found in both the low free energy structures and in the trimer complex are colored in red.

units maintain a native helical packing in both basins, as shown in the free energy profile of chain 1 and chain 2 plotted versus Q_c, the fraction of intra-monomer native contacts, and RMSD for each chain respectively (the second and third panel of Figure 3.5(a)). There is only one low free energy basin at high Q_c (above 0.5) and low RMSD (below $5\mathring{A}$), which corresponds to the native monomeric helical packing.

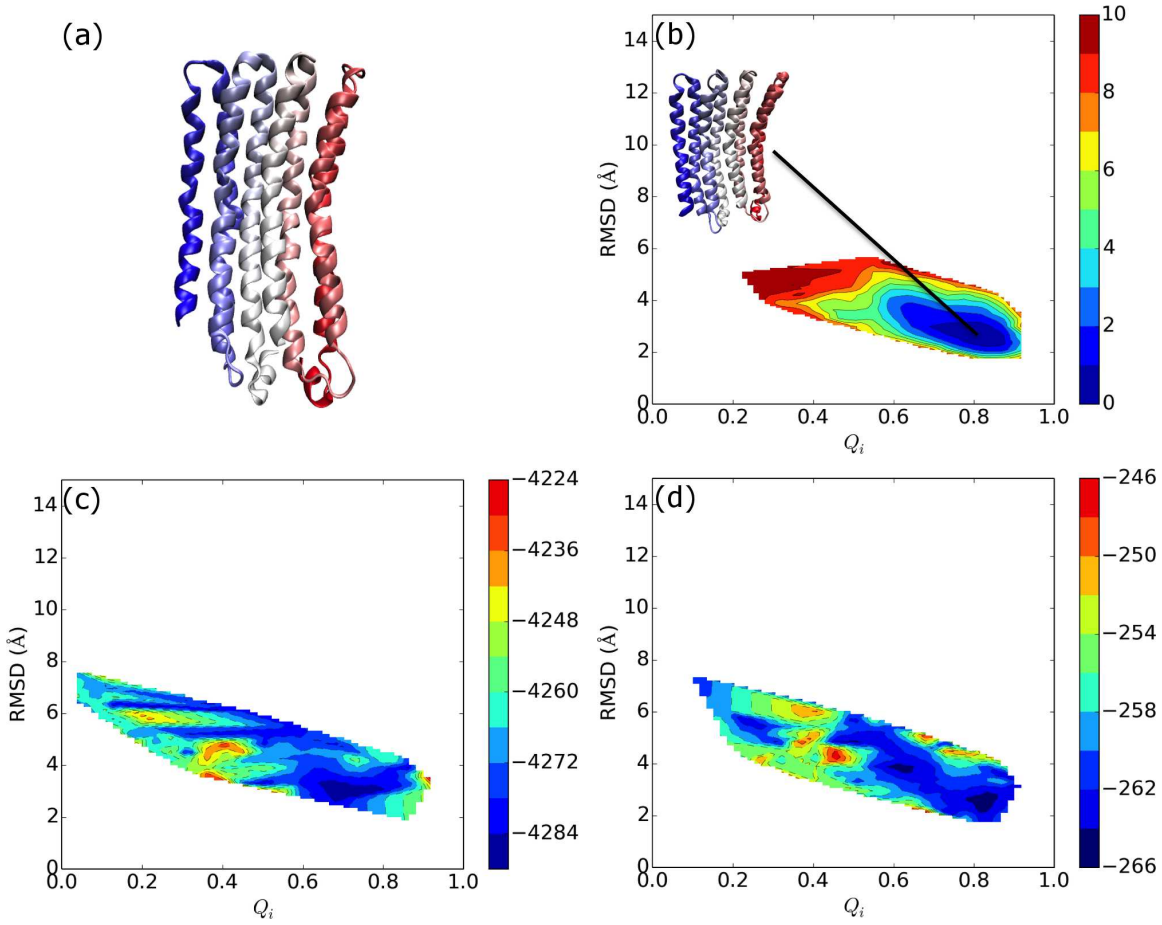


Figure 3.4 : (a): Native configuration of the V-type Na^+-ATPase (2BL2) dimer complex visualized using VMD [4]. (b): Free energy profile of the V-type Na^+-ATPase (2BL2) dimer complex obtained using single memory AWSEM-Membrane. The free energy is plotted versus Q_i, the fraction of native interface contacts (x-axis) and the RMSD (y-axis). A representative structure from the free energy basin is shown. The expectation values of (c) the total potential energy, PE, and (d) the contact energy, $E_{contact}$, are plotted versus the same order parameters.

The two dimensional free energy profile of the V-type Na^+-ATPase (2BL2) dimer complex (Figure 3.5(b)) exhibits two low free energy basins at high Q_i (from 0.5 to 0.65) and relatively low RMSD (at $\approx 6\mathring{A}$ and $\approx 5\mathring{A}$). These basins are separated by a low free energy barrier. Both of these two basins contain representative structures which are nearly native

having more than 55 percent of the native interface contacts formed ($Q_i \approx 0.55$). The difference in RMSD ($\approx 1\mathring{A}$) is mostly the result of a small helix distortion and the overall over-collapse of the structure. As for the nicotinic acetylcholine receptor subdomain (2BG9), we did not observe any stable non-native packings for the monomers. The absence of non-native contacts in the dimer contrasts with our previous study of the monomer energy landscape which did display non-native contacts. For the monomer both the native and a particular non-native helical packing were found to be nearly degenerate in free energy [6]. When simulated in the presence of another monomer, however, as shown in the second and third panels of Figure 3.5(b), only one low free energy basin is found in the free energy profile for folding each individual chain. This basin corresponds to a structure which has fully native helical packing. In other words, the non-native helical packing basins of the monomeric form of nicotinic acetylcholine receptor subdomain (2BG9) and V-type Na^+-ATPase (2BL2) are resolved when multiple chains are present and are allowed to interact in the simulation. In the case of the nicotinic acetylcholine receptor subdomain (2BG9) the native helical packing within each monomer is maintained whether the multimeric complex has a near-native binding interface or an alternative, non-native binding interface.

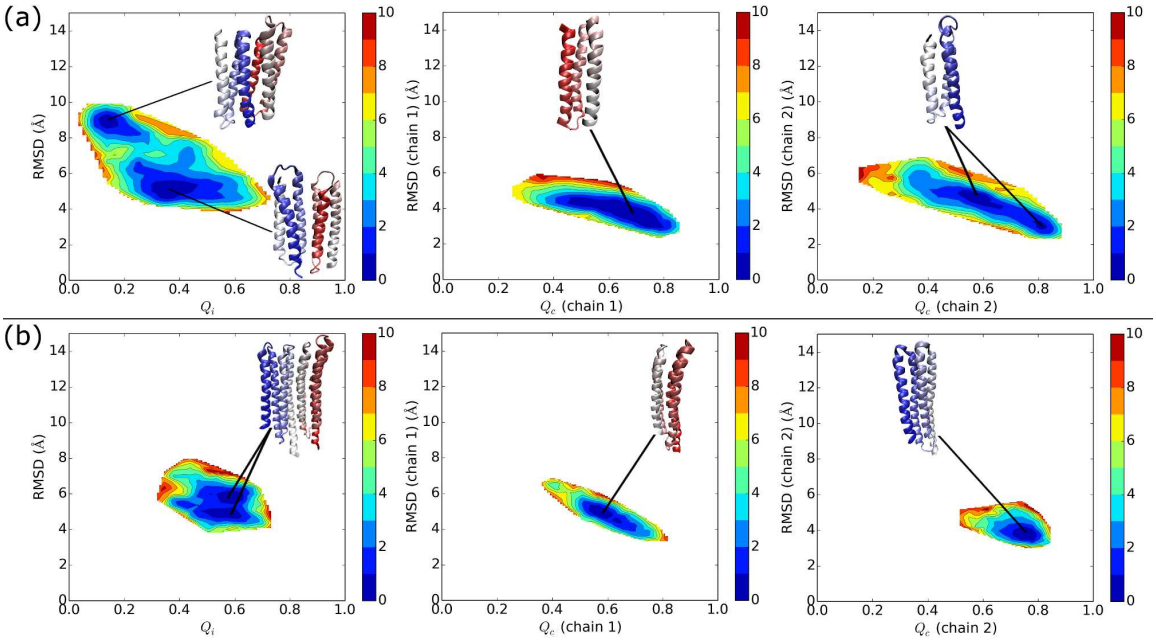


Figure 3.5 : Free energy profile for (a) the nicotinic acetylcholine receptor subdomain (2BG9) and (b) the V-type Na^+-ATPase (2BL2) dimer complexes obtained by using the fragment memory AWSEM-Membrane code. From left to right, the free energy is plotted versus Q_i, the fraction of native interface contacts (x-axis) and the RMSD aligned to the dimer complex (y-axis), and versus Q_c, the fraction of native contacts of each subunit (x-axis) and the RMSD of structures aligned to each subunit (y-axis), respectively. Representative structures from the low free energy basins are shown. Free energy profiles for the nicotinic acetylcholine receptor subdomain (2BG9) monomer and the V-type Na^+-ATPase (2BL2) monomer are shown in Figure 5 and Figure 6 of Ref. 6.

3.4.3 Re-association of fragments of Bacteriorhodopsin monomer and the role of cofactor and fragment rigidity.

One of the classic and indeed heroic early experimental studies of membrane protein folding was undertaken by Khorana's group in the 1980's. They showed that fragments of bacteriorhodopsin could reassociate in the presence of retinal to form a functional molecule [60]. We re-visit computationally their laboratory study of the re-association of the cleaved bac-

teriorhodopsin monomer from two of its fragments, C_2 consisting of the first and second helices of bacteriorhodopsin and C_1 consisting of the remaining five helices.

The experimentally determined structure of the bacteriorhodopsin monomer and simulated structures from two example structure predictions using single memory AWSEM-Membrane are shown in Figure 3.6. The experimentally determined structure of bacteriorhodopsin (Figure 3.6(a)) has a retinal molecule situated in its core. This cofactor supports a configuration in which the seven helices pack around it in an overall elliptic cylinder shape. When the retinal cofactor is omitted from the simulations, we observed over collapsed configurations and distortions in helical packing in both the intact and the cleaved bacteriorhodopsin systems. A predicted structure of the intact bacteriorhodopsin monomer, despite having more than 66 percent of the native contacts formed ($Q_c = 0.663$) and $RMSD = 6.459\mathring{A}$, is over collapsed. The first helix is buried and is surrounded by the other six helices as shown in Figure 3.6(b). The predicted structure of the cleaved bacteriorhodopsin monomer (Figure 3.6(c)) has less than 60 percent of the overall native contacts formed ($Q_c = 0.572$) and a still larger RMSD ($= 7.516\mathring{A}$). The binding interface of fragment C_1 and C_2 is also incorrectly predicted in the cleaved system. We observed non-native helix-helix interactions between the first helix and the sixth helix, (C_2-1, C_1-6), and between the second helix and the seventh helix, (C_2-2, C_1-7). Based on these observations, we infer that the retinal cofactor likely plays an important role in the reconstitution of cleaved bacteriorhodopsin and its effects must be taken into account in the simulations.

To mimic the effects of retinal, we applied three pairwise distance constraints to residue

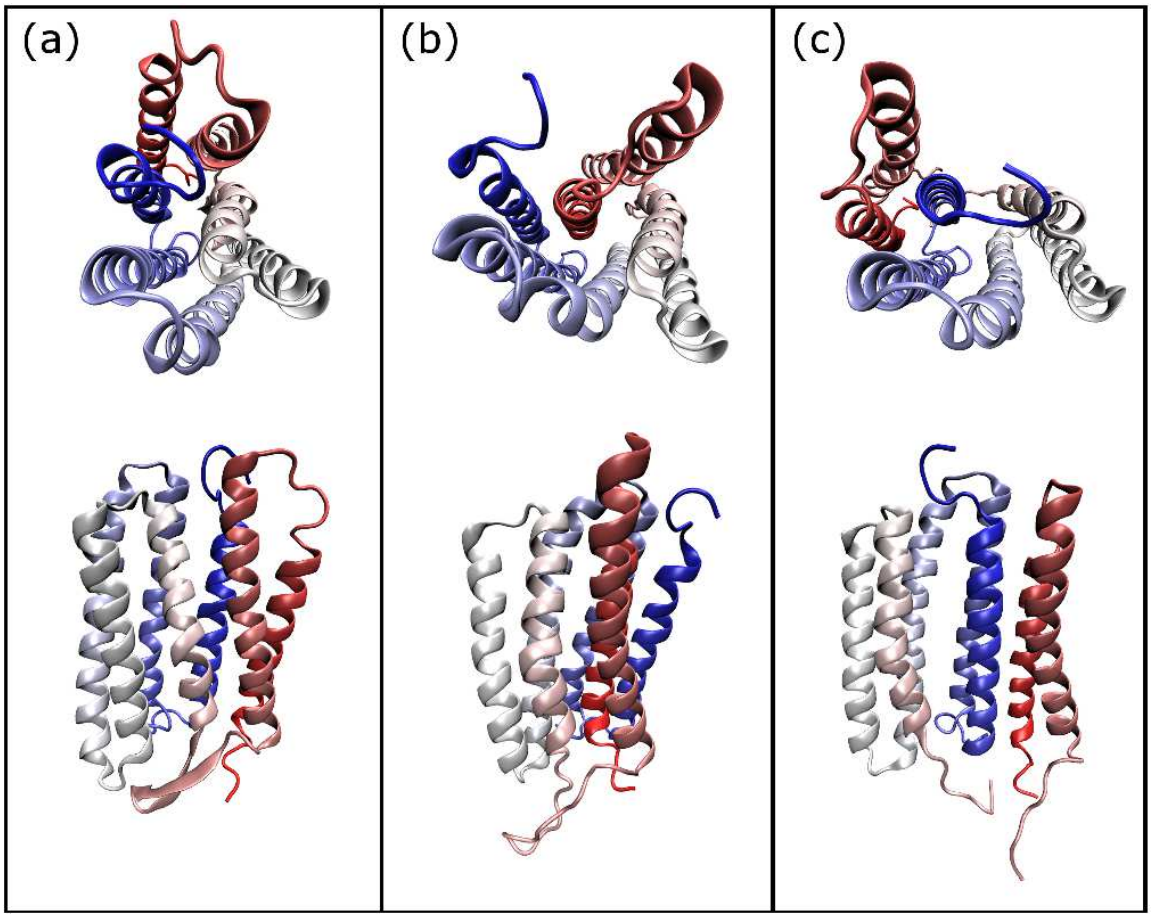


Figure 3.6 : The top view and side view of (a) the experimentally determined bacteriorhodopsin mononer structure (1BRR), (b) a predicted structure using single memory AWSEM-Membrane for the intact bacteriorhodopsin monomer, and (c) a predicted structure using single memory AWSEM-Membrane for the cleaved bacteriorhodopsin monomer. Figures were generated using VMD[4].

pairs in fragment C_1 alone that are in contact with the retinal molecule in the crystal structure. Note that there are no constraints that connect the two fragments together. These constraints, internal to fragment C_1, partially compensate for the lack of an explicit representation of the retinal molecule in our simulations. We also increased the strength of the memory term in order to rigidify the secondary structure where flexibility may also con-

tribute to over-collapse. Figure 3.7(a) summarizes the results from ten simulated annealing runs. The Q_i value of the final snapshot in the runs is plotted against the total potential energy (PE). The structure which has the highest Q_i ($Q_i = 0.764$) is a correctly bound structure in which the two fragments have re-associated to form the native structure (structure (1) in Figure 3.7(a)). This structure agrees well with the native crystal structure of the monomer and has a very high fraction of native contacts formed ($Q_c = 0.877$) and a very low RMSD for such a large system ($= 2.23\mathring{A}$). Three other energetically competitive structures were also observed in this set of simulated annealing runs. These are given the labels (2) and (3) in Figure 3.7. These structures all share helix-helix interactions between the second helix that belongs to the C_2 fragment and the fourth helix that belongs to C_1 fragment (C_2-2, C_1-4). Although not present in the monomer crystal structure, these strong helix-helix interactions are found on the binding interfaces in the complete bacteriorhodopsin trimer complex. Thus we can view these structures as resulting from a kind of domain swapping. To further investigate these (C_2-2, C_1-4) helix-helix interactions, we created a modeled domain swapped structure which consists of the first and second helices (fragment C_2) of chain A of the experimentally determined bacteriorhodopsin trimer complex and the last five helices (fragment C_1) of chain B of the experimentally determined bacteriorhodopsin trimer complex. This modeled domain swapped structure has the (C_2-2, C_1-4) interfacial helix-helix interactions mentioned above and is used as reference structure for calculating Q_i^*, the fractions of dimerization interfacial contacts formed, and $RMSD_{swapped}$. The (C_2-2, C_1-4) dimerization interfacial helix-helix interactions were ob-

served to some degree in six out of the ten simulated annealing simulations that we carried out, as shown in Figure 3.7(b). Five of the simulated annealing runs produced structures with $Q_i^* > 0.4$, indicating that half of the structures have 40 percent or more of the dimerization interfacial contacts formed.

An analysis of the contacts found in the representative predicted structures is shown in Figure 3.8. All of the native intra-fragment contacts in the two-helix fragment C_2 and five-helix fragment C_1 are present in all three structures (Figure 3.8(a), (b), and (c)). The native binding interface of the cleaved bacteriorhodopsin monomer involves helix-helix interactions between the second helix and third helix (C_2-2, C_1-3), between the first helix and seventh helix (C_2-1, C_1-7) and between the second helix and seventh helix (C_2-2, C_1-7). The native-like structure (1) has all of the native monomer interface contacts (C_2-2, C_1-3), (C_2-1, C_1-7) and (C_2-2, C_1-7) formed, as is shown in Figure 3.8(a). Therefore, this contact map looks very similar to the contact map of the bacteriorhodopsin monomer subunit (colored gray in Figure 3.8(d)). Structure (2) and structure (3) do not have the native monomer interface contacts formed, but instead both have the dimerization interface helix-helix interactions (C_2-2, C_1-4) formed, as we mentioned previously. These (C_2-2, C_1-4) helix-helix interactions are the same as the contacts that are made at the protein-protein interfaces of the larger bacteriorhodopsin trimer (colored in red and shown in the yellow region of the contact maps). The other, non-native, helix-helix interactions (C_2-1, C_1-4) and (C_2-1, C_1-5) found at the interface of structure (2), and the non-native, helix-helix interactions (C_2-1, C_1-4) and (C_2-2, C_1-5) of structure (3), are not shared with the

bacteriorhodopsin trimer.

Figure 3.9 shows the results of the free energy landscape analysis of the association of the cleaved bacteriorhodopsin complex using the single memory AWSEM-Membrane force field. The two dimensional free energy profile with respect to $RMSD_{swapped}$, aligned to the modeled domain swapped cleaved bacteriorhodopsin structure, and $RMSD_{native}$, aligned to the native cleaved bacteriorhodopsin structure, contains three low free energy basins (labeled 1, 2 and 3). The profile also exhibits two other somewhat energetically competitive basins with higher free energy ($\approx 3k_BT$) (labeled 4 and 5). Representative structures from both of the low $RMSD_{native}$, high $RMSD_{swapped}$ basin (basin 1) are significantly native-like. The high $RMSD_{native}$ and high $RMSD_{swapped}$ basin (basin 5) contains non-specifically bound structures. Structures in this basin have neither the binding interface contacts of the native structure nor the proper dimerization contacts. Representative structures from the three other low free energy basins (basins 2, 3 and 4) all have the (C_2-2, C_1-4) dimerization helix-helix interactions, and become more similar to the modeled reference dimer structure as $RMSD_{swapped}$ decreases. Figure 3.9 together with Figure 3.7 show that there are two dominant states of the cleaved bacteriorhodopsin monomer: the natively bound state and the state that contains dimerization contacts that would form in the higher order assembly. All other, nonspecifically bound states are higher in free energy.

3.5 Discussion

We were able to predict native-like binding interfaces for dimeric complexes with various topologies (three helix bundle to seven helix bundle) and sizes (up to 460 residues) using the AWSEM-Membrane force field. Using both the single memory and fragment memory flavors of the force field, we showed that oligomerization of the domains as occurs in the full *in vivo* assembly eliminates the non-native helical packing basins that were previously observed in the energy landscapes of the monomers of the nicotinic acetylcholine receptor subdomain (2BG9) and the V-type Na^{+}-ATPase (2BL2) monomers when they were simulated by themselves.

The retinal cofactor plays an important role in the folding process of bacteriorhodopsin. We found over-collapsed configurations and distortions in helical packing occur in both the intact and the cleaved bacteriorhodopsin simulations when the retinal cofactor is omitted from the simulations. We observed, however, proper association of the bacteriorhodopsin monomer fragments, the two-helix fragment, C_2, and the five-helix fragment, C_1, when the force field is augmented through the aid of three pairwise distance constraints to residue pairs that make heavy atom contact with the retinal cofactor in the experimental structure and the aid of the rigidified secondary structure. The key role of the cofactor is consistent with the observations made in Khorana's experiments in which the protein was only shown to refold into a functional form after retinal was added [60]. We also observed that simulated annealing of the cleaved bacteriorhodopsin fragments often resulted in structures that contain dimerization helix-helix interactions instead of the helix-helix interactions that

would lead to the proper monomer structure. In a certain sense, these structures are not misfolded but in fact can be viewed as resulting from a kind of internal domain swapping. This view would be consistent with the principle of minimal frustration applying in full force to membrane proteins much as it does for globular proteins. These domain swapped states are competitive in free energy terms with the native state when the constraints normally imposed by chain connectivity are relaxed by cleavage of the monomer into two fragments.

Acknowledgments

H.H.T. thanks Weihua Zheng for helpful discussions. The project described was supported by Grant R01 GM44557 from the National Institute of General Medical Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of National Institute of General Medical Sciences or the National Institutes of Health. Additional support was also provided by the D.R. Bullard-Welch Chair at Rice University, Grant C-0016. This work was also supported in part by the Data Analysis and Visualization Cyberinfrastructure funded by NSF under grant OCI-0959097.

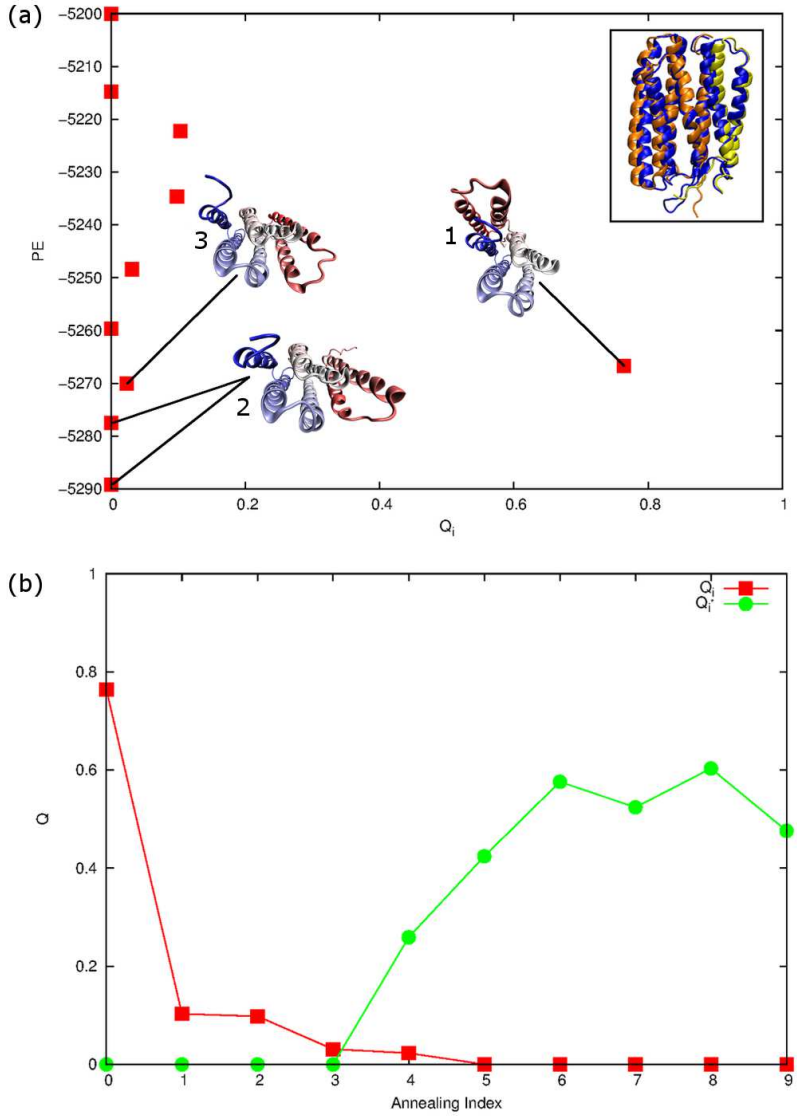


Figure 3.7 : (a): The final snapshots of simulated annealing runs are plotted as a function of the fraction of native interface contacts, Q_i, and the total potential energy, PE. Top view of the snapshots of the predicted structures are shown and colored according to residue index starting at the N-terminus (red) and going to the C-terminus (blue): Structure (1) is a correctly bound structure, which is shown superimposed on the native structure in the inset (yellow: fragment C_2, orange: fragment C_1, blue: native structure). Structures (2) and (3) are competitive low energy predicted structures. (b): Final Q versus annealing index of dimer interface predictions of the cleaved bacteriorhodopsin monomer. Ten independent simulated annealing simulations were conducted and their final Q_i values, the fractions of native interface contacts formed, were plotted in the order of decreasing Q_i from left to right. Q_i, the fraction of native interface contacts formed, and Q_i^*, the fraction of dimerization interfacial contacts formed, are plotted in red and green, respectively. Note that the "annealing index" does not refer to the actual order in which the simulated annealing simulations were carried out.

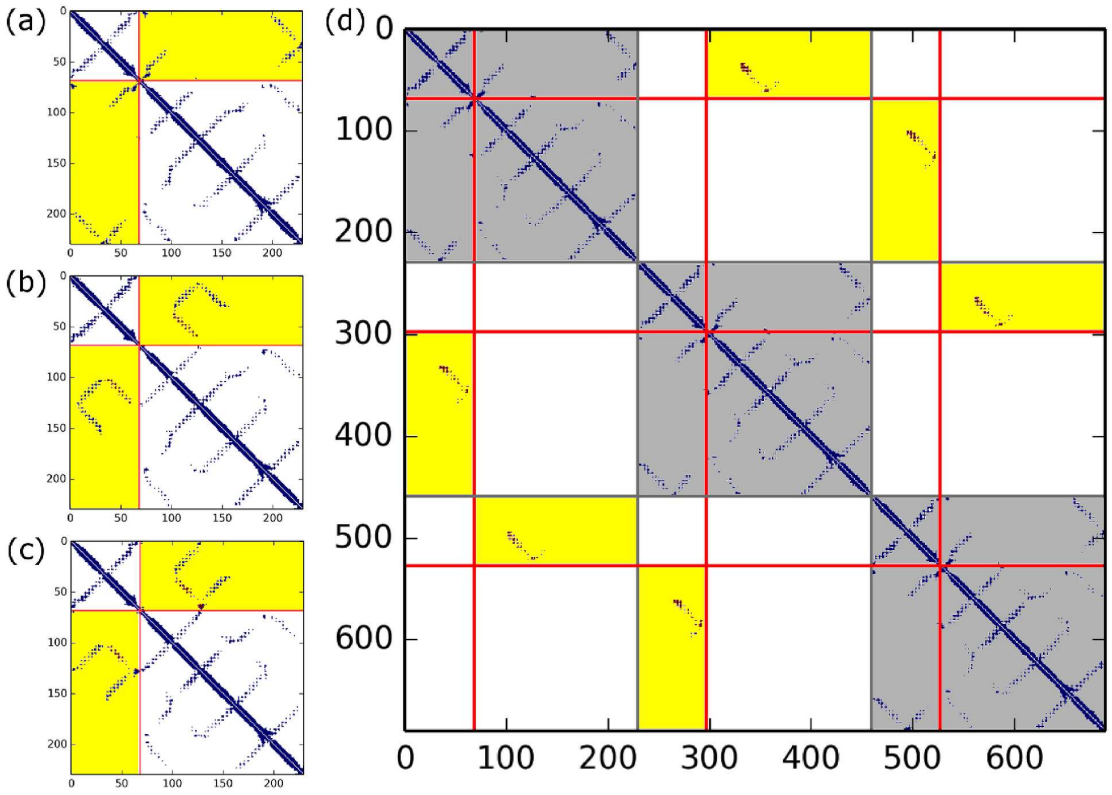


Figure 3.8 : Contact maps of representative structures obtained from simulated annealing simulations of the cleaved bacteriorhodopsin monomer (contact maps (a), (b), and (c) correspond to structure (1), (2), (3) in Figure 3.7, respectively) and the contact map of the bacteriorhodopsin trimer complex (d). Sections of the maps that show intra-monomer contacts are colored in gray. Sections of the maps that show inter-subunit contacts are colored in yellow, in which conserved contacts found in simulated annealing structures and the inter-subunit contacts in the trimer complex are colored in red. Red lines separate the C_1 fragment and C_2 fragment of the cleaved bacteriorhodopsin molecule.

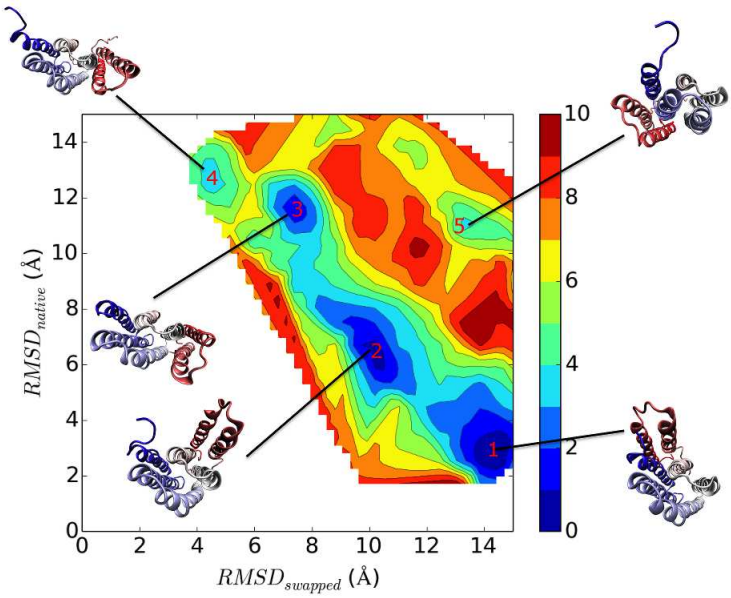


Figure 3.9 : Free energy profile of the cleaved bacteriorhodopsin assembly obtained using the single memory AWSEM-Membrane code. Free energy is in k_BT, in which T was chosen to be below the folding temperature of the monomer but high enough to sample multiple bound configurations. The free energy is plotted versus $RMSD_{swapped}$ and $RMSD_{native}$ (y-axis). Representative structures from the low free energy basins are shown.

Chapter 4

Topological constraints and modular structure in the folding and functional motions of GlpG, an intramembrane protease

4.1 Abstract

We investigate the folding of GlpG, an intramembrane protease, using perfectly funneled structure-based models that implicitly account for the absence or presence of the membrane. These two models are used to describe, respectively, folding in detergent micelles and folding within a bilayer with its accompanying topological constraints. Structural free-energy landscape analysis shows that while the presence of multiple folding pathways is an intrinsic property of GlpGs modular functional architecture, the large entropic cost of organizing helical bundles in the absence of the constraining bilayer leads to pathways that backtrack, i.e., local unfolding of previously folded substructures is required when moving from the unfolded to the folded state along the minimum free-energy pathway. This backtracking explains the experimental observation of thermodynamically destabilizing mutations that accelerate GlpGs folding in detergent micelles. In contrast, backtracking is absent from the model when folding is constrained within a bilayer, the environment in which GlpG has evolved to fold. We also characterize a near-native state with a highly mobile transmembrane helix 5 (TM5) that is significantly populated under folding conditions

when GlpG is embedded in a bilayer. Unbinding of TM5 from the rest of the structure exposes GlpGs active site, consistent with studies of the catalytic mechanism of GlpG that suggest that TM5 serves as a substrate gate to the active site.

4.2 Introduction

GlpG is a rhomboid protease that sits and functions in the cell membrane. Its homologues are found across all kingdoms of life. GlpG has been the subject of several biophysical experimental studies aimed towards understanding membrane protein folding and the relationships among protein structure, dynamics and function [74, 75, 76, 19, 77]. An extensive experimental ϕ-value analysis found ϕ-values significantly different from zero, indicative of structural changes during the rate-limiting step of folding, in transmembrane helices 1 through 5 (TM1-5) and the intervening loops [19]. Most of the non-zero ϕ-values, particularly in TM3-5 and in the large loop L1, were negative, meaning that although the corresponding mutation destabilizes the native state, it nonetheless accelerates folding. The preponderance of negative ϕ-values was puzzling and unprecedented, and at the time these effects were tentatively ascribed to non-native interactions in the transition state ensemble. In this work we show that, in fact, simple models with perfectly funneled energy landscapes that lack non-native interactions are able to explain the origin of these negative ϕ-values and how they arise when folding in detergent micelles rather than bilayers. lpha-helical membrane protein folding is thought to occur in two stages in vivo [78]. The first stage, setting up the proper topology of transmembrane helices, is handled by the translo-

con [79, 80]. In the present context, topology refers to specifying the directions in which a membrane proteins constituent transmembrane helices traverse the bilayer. The second stage, converting from properly inserted but dissociated helices into a functional folded structure, occurs spontaneously and is, in some ways, analogous to soluble protein folding. Yet we know, ranging from the hydrophobic effect [81, 82] to water-mediated [49] and screened electrostatic interactions [83], the solvent plays a role in determining what types of non-covalent interactions are stabilizing and destabilizing. While soluble proteins fold in polar and isotropic aqueous solutions, membrane proteins fold in largely apolar and anisotropic environments. These environmental differences complicate applying directly methods developed for studying soluble protein folding to the study of membrane protein folding. Nonetheless, experimentalists have been able to apply a variety of methods to study the kinetics and thermodynamics of membrane protein folding through the use of detergent micelles as a membrane-mimicking environment. Experiments that probe the folding mechanisms of membrane proteins have employed micelles composed of a mixture of anionic and nonionic detergents [19, 84, 85], which not only keep membrane proteins soluble but also, through use of mixed micelles, allow the equilibrium between folded and unfolded states to be tuned. Micelles predominantly composed of nonionic detergents, such as n-Dodecyl--D-maltopyranoside (DDM), preferentially stabilize a folded state that has been shown to be functional and is therefore likely to be structurally similar to the folded state in vivo. Micelles predominantly composed of anionic detergents, on the other hand, preferentially stabilize an unfolded state that contains significant amounts of secondary

structure. This ability to tune the equilibrium means that stopped-flow kinetic experiments can be combined with protein engineering techniques to determine folding mechanisms at the single residue level [19, 84, 86], in analogy to what has been done for soluble proteins [24, 87, 88]. Since carrying out these types of experiments in bilayers is still difficult, it is presently unknown how folding mechanisms determined in micelles compare to those in membranes. Confining proteins to a two-dimensional membrane is expected to constrain unfolded and partially folded ensembles to having structures with helices that are largely properly aligned and embedded in the membrane; such topological restrictions would be relaxed in a micellar environment. Theoretical [6, 17] and experimental [76, 19] work suggests that at least some membrane proteins can reversibly fold and unfold without the aid of the translocon or chaperones in vitro. It is therefore likely that membrane protein folding landscapes are funneled, much like globular protein landscapes [23, 7]. Structure-based models with perfectly funneled energy landscapes have proven useful for investigating the folding and binding of proteins [89, 90]. In this study, we employ a structure-based model to investigate folding of a membrane protein in two different situations: in the absence and presence of an implicit membrane energy term that biases conformations to have the correct topology with respect to the membrane. Simulations with the implicit membrane term are thus taken to model folding in a bilayer while simulations without the implicit membrane energy are taken to model folding in detergent micelles. Although this is an oversimplification, it captures the significantly increased topological freedom of membrane proteins in micellar environments compared to lipid bilayer. Figure 4.1 shows schematic representa-

tions of the corresponding denatured states of membrane proteins in bilayers and micelles. The same energy landscape that dictates folding routes also encodes functional motions. It has been suggested that the modularity in the structure of GlpG supports functional motions [74, 91]. The N-terminal domain, which contains transmembrane helices 1 and 2 (TM1-2) as well as the intervening L1 loop, functions as a structural scaffold [91] while the C-terminal domain with its four transmembrane helices (TM3-6) includes the catalytic site [91]. The C-terminal domain is apparently more flexible than the N-terminal domain; both the loop L5 [77] and the transmembrane helix TM5 [91] have been crystallized in multiple conformations. Due to this flexibility, it has been suggested that either L5 alone [77] or L5 and TM5 [91] may serve as a substrate gate for access to the catalytic site. Using free energy landscape analysis and perturbation methods along with structural analysis, we show there is a near-native state significantly populated under folding conditions and elucidate its connections to GlpGs folding mechanism and function.

4.3 Methods

4.3.1 Simulation and analysis methodology

We performed molecular dynamics simulations of a coarse-grained structure-based model [50] of GlpG based on the crystal structure with PDB ID 2xov [92]. We carried out two parallel sets of simulations: one with an implicit membrane present and one without a membrane. The implicit membrane model is described in Ref. 6 and the assignment of

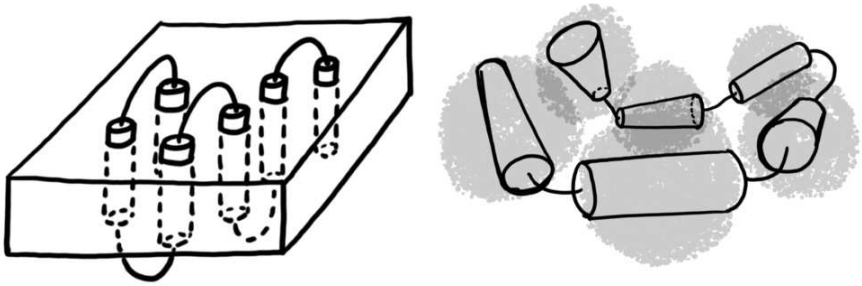


Figure 4.1 : Schematic diagrams of the unfolded state of -helical membrane proteins in bilayers (left) and detergent micelles (right). The transmembrane helices (cylinders) are connected by loops. Transmembrane helices are either embedded in a membrane (rectangular prism) or are surrounded by detergent micelles (transparent gray spheres). In this work we use an implicit membrane model to simulate folding within a bilayer and assume that folding in detergent micelles corresponds to folding without constraints on the alignment of helices. In both cases we assume that the unfolded state has near-native levels of secondary structure, as has been observed in experiments on the SDS denatured state of membrane proteins.

residues into the intramembrane and extramembrane residues is described in the Appendix (Figure 4.12). We sampled at multiple temperatures above and below the corresponding folding temperatures and used umbrella sampling at each of these temperatures to sample a wide range of folded, partially folded and unfolded structures. We then used the Multistate Bennett Acceptance Ratio (MBAR) method [15] to reconstruct unbiased free energy profiles, compute expectation values of structural order parameters, and perform perturbative calculations to test the effect of small changes to the Hamiltonian. We infer folding mechanisms by looking for low free energy routes between the unfolded and folded states in the unbiased free energy profiles and then performing analysis on structures sampled in the basins and saddle points along these routes. While the appropriateness of various reaction coordinates for describing protein folding kinetics is vigorously discussed [12, 93, 94],

here we take the pragmatic approach of comparing our inferred mechanisms to experimental data and find highly non-trivial agreement based on reaction coordinates that measure the degree of nativeness of different parts of the molecule. See the Appendix for a complete explanation of the methods.

4.3.2 Structure-based model of GlpG

The crystal structure used to define the stabilizing native interactions in our structure-based model is shown in Figure 4.2. GlpG has six transmembrane helices connected by five loops. The first loop, L1, is notable because it is large and contains several small interfacial helices. Our definition of the N- and C-terminal domains of GlpG was arrived at based on the analysis of our simulation results and is therefore not imposed on the model beforehand; these two domains are found to fold semi-independently (see the Results and Discussion section) using our structure-based model. Therefore, this definition arises as a direct consequence of the structure of GlpG given our way of defining its contact map. Structural bioinformatics studies have indicated that membrane proteins are stabilized by tight helix-helix interactions that are mediated by small and polar residues [95]. We therefore used a 6.5 C-C cutoff to define stabilizing native interactions, which is somewhat shorter than the cutoffs that have been applied to simulations of soluble proteins in the past. We have also selectively strengthened the local-in-sequence interactions in order to decouple secondary and tertiary structure formation. This modification of the model is motivated by the observation of native-like levels of secondary structure in the SDS unfolded state of GlpG [19].

See the Appendix for a precise description of the parameters used in the model.

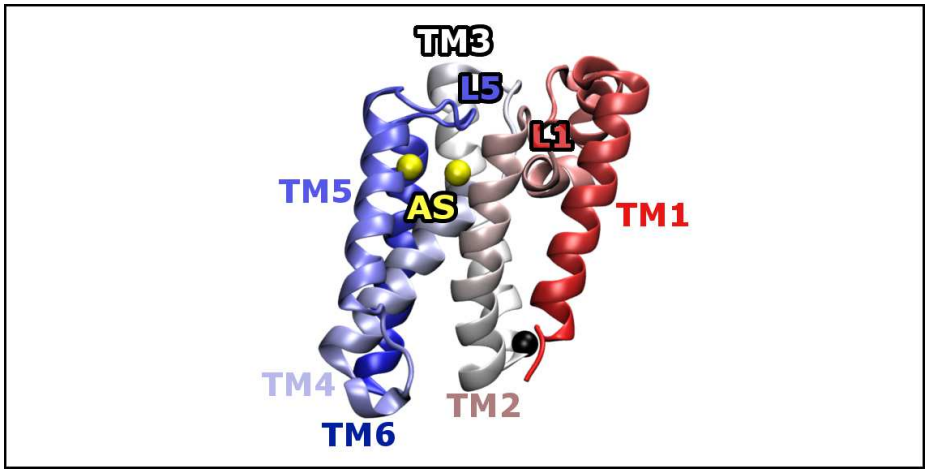


Figure 4.2 : Crystal structure of GlpG (PDB ID: 2xov). A black sphere demarcates the boundary between the N- and C-terminal domains. The catalytic dyad, shown in yellow and located on TM4 and TM6, is buried by TM5 and L5. The large loop L1 is made up of several interfacial helices whose axes run parallel to the membrane surface. The color of the backbone varies smoothly from red (N-terminal) to white and then to blue (C-terminal).

4.4 Results and Discussion

4.4.1 Unfolding always corresponds to loss of tertiary structure with retention of secondary structure but leads to a more expanded ensemble in the absence of the implicit membrane.

Experimental circular dichroism and tryptophan fluorescence measurements indicate that unfolding of GlpG in micelles corresponds to loss of tertiary structure but retention of native levels of secondary structure [19]. In the simulations, the expectation values of secondary and tertiary structure formation order parameters (see the Appendix for precise

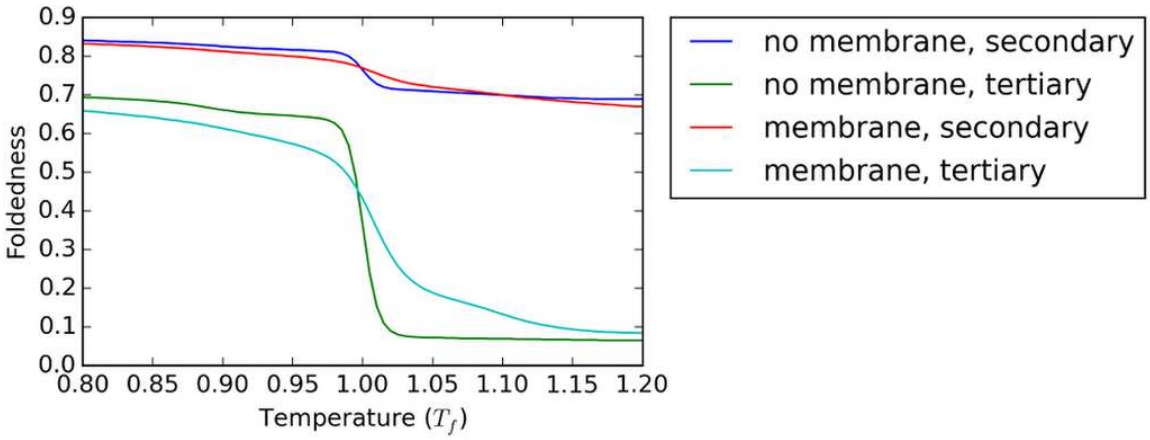


Figure 4.3 : Expectation value of tertiary and secondary structure formation order parameters both with and without the implicit membrane model. Temperature is normalized by the folding temperature of each model independently. Precise definitions of short- and long-range foldedness are given in the Order parameters section the Appendix.

descriptions) as a function of temperature indicate that likewise, both in the absence and the presence of the implicit membrane, unfolding corresponds to loss of tertiary structure and retention of secondary structure (Figure 4.3). When the implicit membrane is present, the unfolded structures largely retain native-like topologies with respect to the membrane (Figure 4.5 B), although excursions to the extramembrane regions are possible. The simulated unfolded ensemble thus resembles what is commonly understood to be the starting point for the second stage of membrane protein folding [78], which takes place once the helices have been inserted into the membrane by the translocon in their native orientations. The simulated unfolded ensemble in the absence of the bilayer is significantly more expanded (Figure 4.4B). Figure 4.6 and Figure 4.7 show a more detailed comparison of these two

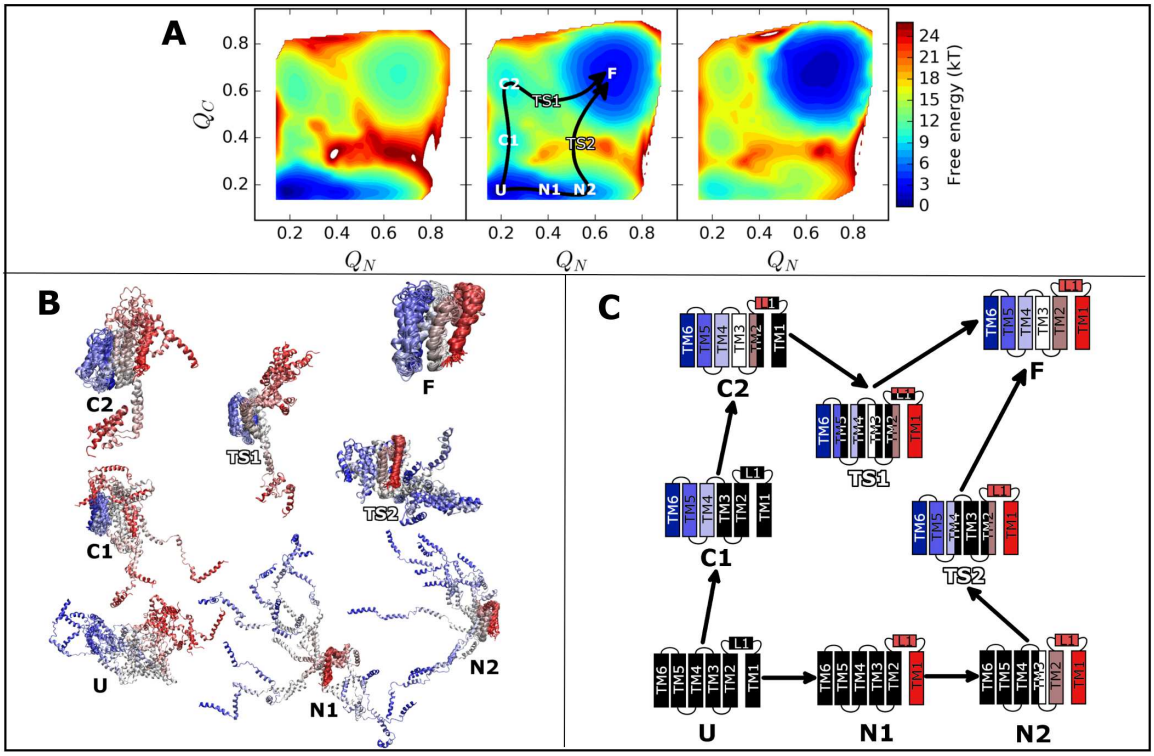


Figure 4.4 : Free energy analysis and strutural characterizations of GlpG without the implicit membrane. (A) Two dimensional free energy profiles above (left), at (middle) and below (right) the folding temperature with respect to Q_N and Q_C. Q_N and Q_C measure the degree of folding within the N- and C-terminal domains, respectively. Precise definitions are given in the Appendix. Key structural states are labeled, and the inferred folding pathways are indicated with arrows. Areas shown in white are high in free energy. (B) Structural ensembles made up of ten representative structures selected from low free energy basins and transition states; folded regions in each ensemble have been aligned for clarity. (C) Schematic representations of the structural ensembles. Transmembrane helices and the large loop L1 are shown as fully folded (full color), partially folded (half color), or unfolded (black). The colors used in B and C are the same as those established in Figure 4.2.

ensembles and experiment.

In order to more precisely compare the degree of compaction between the two simulated unfolded ensembles and to expose these unfolded ensembles to potential experimental falsification, we calculated the expected value of several intrahelical and interhelical dis-

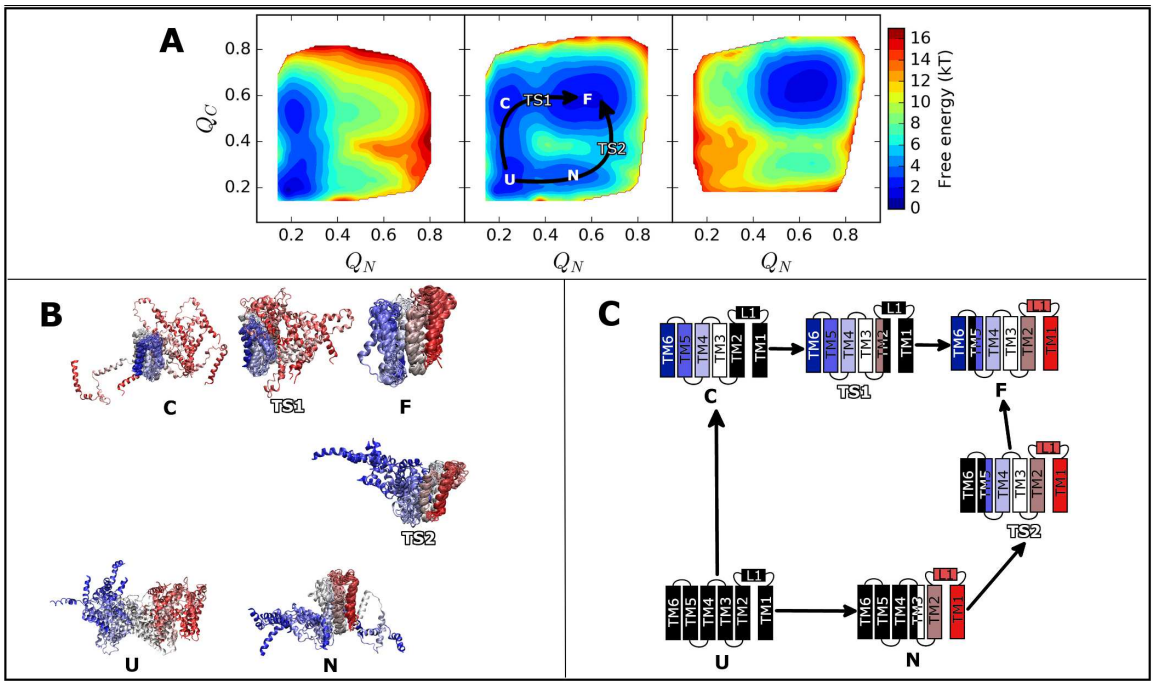


Figure 4.5 : Free energy analysis and strutural characterizations of GlpG with the implicit membrane. (A) Two dimensional free energy profiles above (left), at (middle) and below (right) the folding temperature with respect to Q_N and Q_C. Q_N and Q_C measure the degree of folding within the N- and C-terminal domains, respectively. Precise definitions are given in the Appendix. Key structural states are labeled, and the inferred folding pathways are indicated with arrows. Areas shown in white are high in free energy. (B) Structural ensembles made up of ten representative structures selected from low free energy basins and transition states; folded regions in each ensemble have been aligned for clarity. (C) Schematic representations of the structural ensembles. Transmembrane helices and the large loop L1 are shown as fully folded (full color), partially folded (half color), or unfolded (black). The colors used in B and C are the same as those established in Figure 4.2.

tances as a function of temperature. These distances were chosen in analogy to distances that were measured in the SDS unfolded state of bacteriorhodopsin using double electronelectron resonance (DEER) experiments [96]. To measure these distances during the simulations, we virtually labeled the C_α atoms of residues D116 (TM1), F146 (TM2), P195 (TM3), G199 (TM4), F245 (TM5), and A250 (TM6) on the periplasmic side of the pro-

tein and A93 (TM1), S171 (TM2 and TM3), P219 (TM4), L225 (TM5) and N271 (TM6) on the cytoplasmic side. Note that the cytoplasmic loop between TM2 and TM3 is very short, so we use S171 to represent the cytoplasmic ends of both TM2 and TM3. Based on the location of these atoms, we calculated 6 intrahelical distances (A93-D116, TM1; F146-S171, TM2; S171-P195, TM3; G199-P219, TM4; L225-F245, TM5; A250-N271, TM6) corresponding to the length of each helix and 12 interhelical distances (TM1-TM2, TM1-TM4, TM1-TM6, TM2-TM4, TM2-TM6, and TM4-TM6 on both the cytoplasmic and periplasmic sides).

The expected value of the intrahelical and interhelical distances above the folding temperature, i.e., in the unfolded ensembles, are plotted as a function of sequence separation in Figure 4.7. For comparison, and because these distances have not yet been measured in experiment for GlpG, experimental measurements of analogous distances in bacteriorhodopsin [96] are plotted alongside the simulation results for GlpG. Whether or not the implicit membrane is present, the intrahelical distances in GlpG show good agreement with those measured for bR due to the fact that, in all cases, the helices present in the folded state remain formed in the denatured state. The interhelical distances measured for the simulated ensemble of GlpG in the presence of the implicit membrane are also in approximate agreement with those measured in experiment for bR. However, whereas the distances in GlpG increase nearly monotonically with sequence separation, the measurements on bR indicate that the interhelical distances are nearly independent of sequence separation. The interhelical distances measured in the simulated ensemble of GlpG in the absence of the implicit

membrane are considerably higher than those in the presence of the implicit membrane but show the same increasing trend with sequence separation as the distances in the presence of the implicit membrane. Most of the distances measured in the absence of the implicit membrane exceed the stated maximum range of the DEER experiments that were used to measure the distances for bR [96]. Further experimental work on GlpG and computational study of unfolded ensembles of bR will be required in order to fully understand what constraints, if any, are imposed on the SDS unfolded ensembles of membrane proteins and how they might differ from protein to protein.

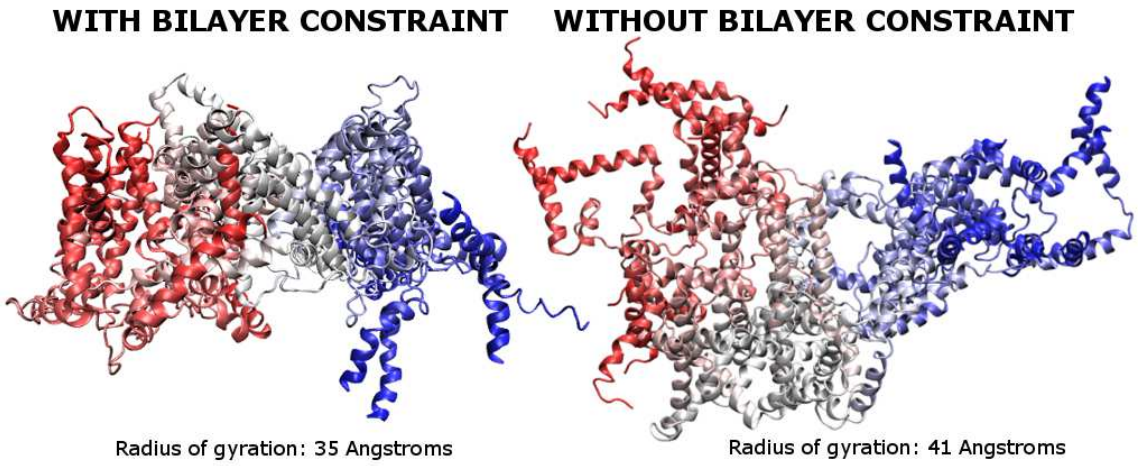


Figure 4.6 : Unfolded ensembles with (top) and without (bottom) the implicit membrane model.

4.4.2 Folding can be initiated in either the N- or C- terminal domain of GlpG.

Free energy profiles plotted as a function of Q_N and Q_C (see the Appendix for precise descriptions), which quantify how native-like the structures are for the N- and C-terminal

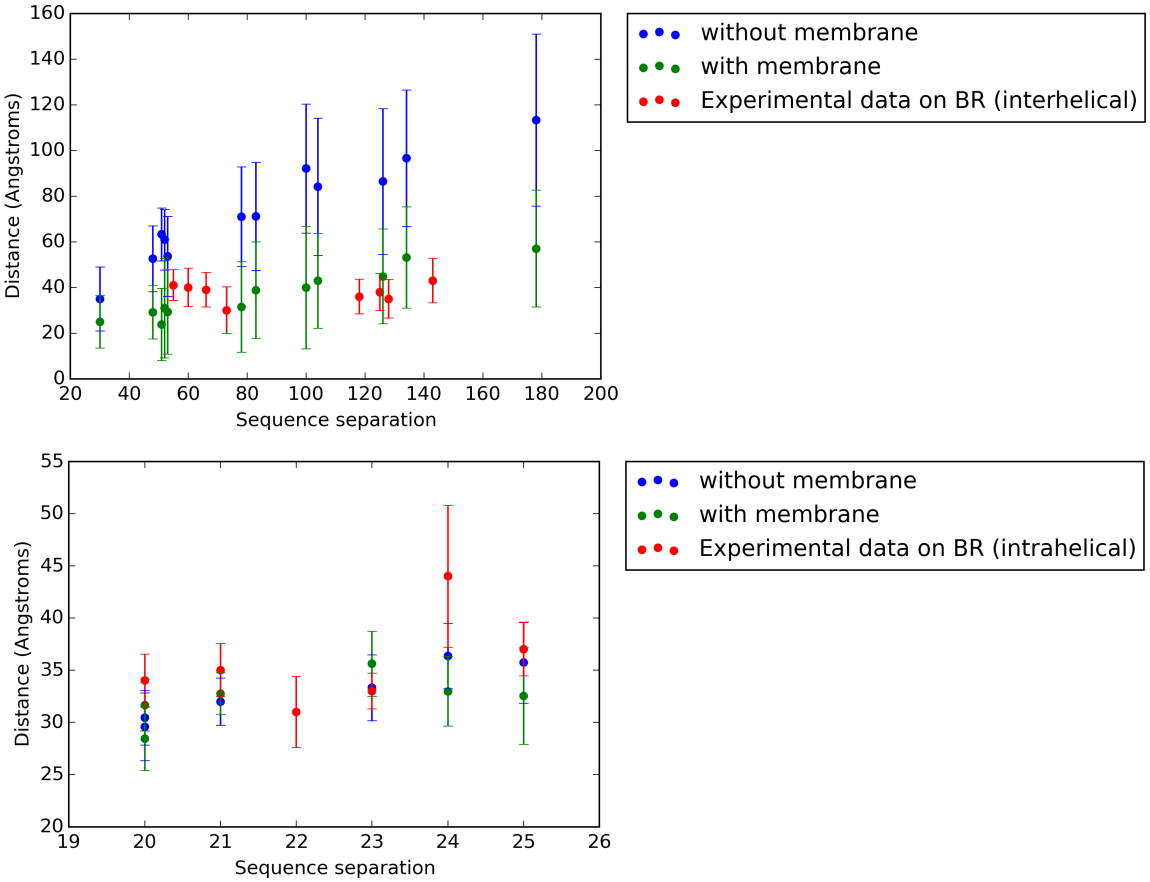


Figure 4.7 : Comparison of interhelical (top) and intrahelical (bottom) distances in the simulated denatured states of GlpG. The mean interhelical distances are plotted as a function of the sequence separation between the probed residues for both the model without (blue) and with (green) the implicit membrane present. Experimentally measured interhelical and intrahelical distances in the SDS denatured state of bR (red) are plotted for the sake of comparison. In all cases, standard deviations are indicated with error bars.

parts of the molecule, respectively, suggest that folding can be initiated by moving either along Q_N or Q_C, i.e., by forming native-like structure within either the N-terminal or C-terminal parts of the molecule (Figure 4.4 and Figure 4.5). This is true both in the absence and presence of the implicit membrane. Above but near to the folding temperature and in the presence of the implicit membrane, the molecule populates both the fully unfolded

state (U) and the C-terminal folded state (C) with TM3-6 folded. An orthogonal folding route towards the N-terminal folded state (N) is also present, though less favorable. In the absence of the implicit membrane and above the folding temperature, the molecule prefers the fully unfolded state (U) and a partially formed N-terminal structure with L1 folding onto TM1 (N1). Slightly higher in free energy in the same direction is another state with both TM1 and TM2, as well as the intervening L1, being well folded (N2). As in the case of the model with the implicit membrane, another folding route is available at higher free energy. There are also two intermediates along this route, the first with TM4-6 folded (C1) and a second which also includes folding TM2, TM3 and part of L1 onto the C-terminal part of the molecule (C2).

4.4.3 Optimal energy-entropy compensation for the modular structure results in a multistep folding pathway that backtracks during the rate-limiting step without the implicit membrane but does not backtrack in the membrane with its accompanying topological constraints.

After initiating folding through either the N- or C-terminal domains, GlpG must fold the other half of the molecule to arrive at the folded state. In the membrane, this occurs in a straightforward manner, with both pathways ($U \rightarrow C \rightarrow TS1 \rightarrow F$ and $U \rightarrow N \rightarrow TS2 \rightarrow F$) being approximately equal in free energy (Figure 4.5). Without the implicit membrane energy term to constrain the topology (Figure 4.4), however, folding becomes more complex. Although initiating folding via the N-terminal domain ($U \rightarrow N1 \rightarrow N2$)

is more favorable than initiating folding via the C-terminal ($U \rightarrow C1 \rightarrow C2$), starting along this route is ultimately not productive as the molecule later encounters a relatively high free energy barrier (TS2) associated with organizing the large and unconstrained C-terminal domain. Folding does not proceed by propagating the folding front through the interface between the N- and C-terminal domains because there are relatively few contacts on the interface. Instead, the high free energy barrier to folding is lowered somewhat through simultaneous organization of TM4-6 (a decrease in energy) at the same time as breaking the interface between L1/TM2 and TM3 (an increase in entropy), which was formed in N2. This is an example of backtracking, i.e., the required unfolding of natively folded substructures while proceeding from the unfolded state to the folded state. By making optimal use of energy-entropy compensation, GlpG is able to reduce the free energy barrier between a partially folded state and the completely folded state because there are multiple sites for nucleating folding. Once both domains are independently folded in TS2, a saddle point in the free energy surface is reached and folding can proceed downhill to the folded state (F). This effect is also operative when folding is initiated in the C-terminal direction ($U \rightarrow C1 \rightarrow C2$). Proceeding initially uphill in free energy, GlpG arrives at C2 where TM2-6 and parts of the loop L1 are folded. Since L1 is quite large, however, there exists a high entropic barrier to consolidate folding of TM1. Again, a compromise is made by simultaneously forming the interface TM1-TM2 and contacts within L1 along with releasing of L1 from its position docked against L3 and breaking the interface between TM2 and the C-terminal domain (TM3/L3/TM4). After this, folding can proceed down-

hill towards the folded state by re-forming the interface between TM2 and the C-terminal domain and re-inserting L1. Note that the presence of high-energy intermediates and multiple folding pathways are compatible with the apparent two-state behavior observed in the micelle-mediated folding experiments. Folding is cooperative in experiments and in our simulations, but free energy landscape analysis allows us to resolve high free energy intermediate states and multiple pathways that would not necessarily be apparent from the initial experimental data alone. With the simulation-derived structural model for the parallel pathways, it should be possible to design experiments that probe this aspect of GlpG folding. Of the two putative folding pathways, the latter one, initiated through folding the C-terminal domain, has the lower free energy transition state (TS1) and should be dominant. This differs from the inference, made without the aid of modeling and based on the distribution of experimentally measured classical ($0 < \phi < 1$) ϕ-values in GlpG, of a transition state ensemble with an unfolded C-terminal domain and N-terminal folding nucleus [19]. However, in that study, thermodynamically destabilizing mutations that accelerated folding and unfolding were found throughout TM3-5 and in L1. The resulting ϕ-values are negative. Destabilizing mutations that slow folding, leading to positive ϕ-values, were found largely on the interface TM1-TM2 but also in L1. Figure 4.8 shows the difference between the average contact maps of TS1 and C2 and its connection to the experimentally measured ϕ-values. Mutations that destabilize the interface between L1/TM2 and the C-terminal domain accelerate folding because formation of TS1 involves breaking those contacts. Mutations that destabilize the interface TM1-TM2 will slow folding because for-

mation of TS1 also involves forming that interface. Mutations that primarily effect contacts within the C-terminal domain result in near-zero ϕ-values, as those contacts are largely preserved in the $C2 \rightarrow TS1$ transition. Thus we see that the dominant mechanism predicted by simulations in the absence of the membrane ($U \rightarrow C1 \rightarrow C2 \rightarrow TS1 \rightarrow F$) provides a detailed structural explanation of the previously puzzling preponderance of negative ϕ-values measured in the C-terminal domain of GlpG. On a topologically unconstrained but perfectly funneled landscape, folding is complicated by GlpGs modular structure and the high entropic cost of organizing helical bundles from their unconstrained partially folded states. Non-native frustrated interactions need not be invoked to explain the presence of a large number of negative ϕ-values in GlpG.

A recent single molecule force spectroscopy study in bicelles and micelles also found evidence for structural modularity in GlpG unfolding [76]. They found that the unfolding of GlpG at high force was cooperative. They were also able to characterize two transiently populated metastable states. Their structural interpretation of the unfolding via intermediates closely corresponds to the reverse of one of our folding pathways ($F \rightarrow TS2 \rightarrow N \rightarrow U$) in the presence of the bilayer, while the structural decomposition of GlpG into domains given in their supplementary information corresponds more or less exactly to the reverse of one of our dominant folding pathways ($F \rightarrow C2 \rightarrow C1 \rightarrow U$) in the absence of the bilayer. These encouraging correspondences (see the Appendix for a more detailed discussion) suggest that further computational and experimental work should allow us to create a unified picture of SDS and force induced unfolding of GlpG in micelles and bilayers.

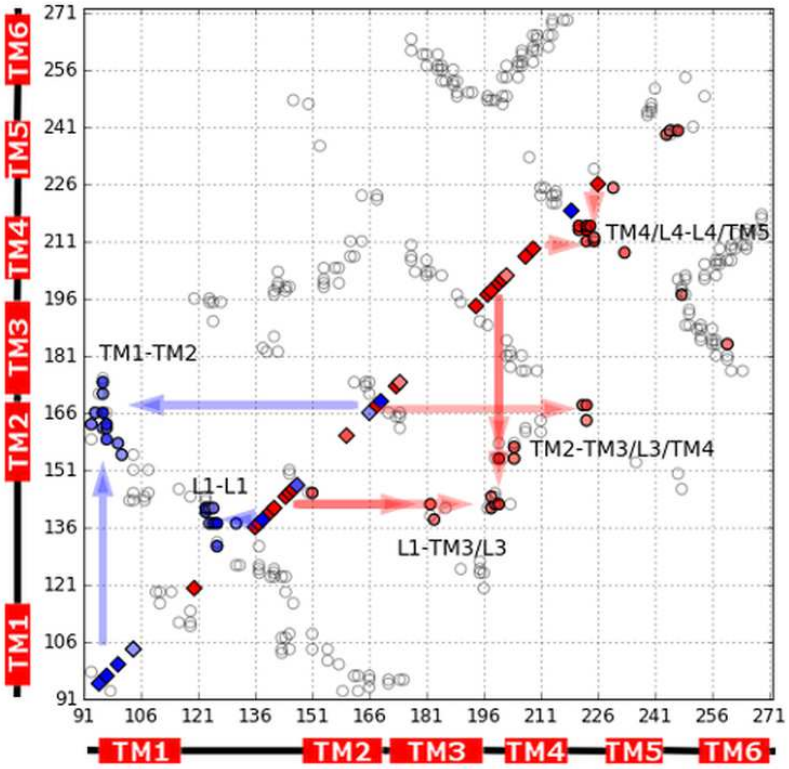


Figure 4.8 : Contact map of GlpG showing the C2 $\rightarrow$ TS1 structural transition. The axes are labeled with residue indices. Contacts that change their occupancy by more than 20% when going from C2 to TS1 are shown in blue (gained in TS1, upper diagonal) and red (lost in TS1, lower diagonal) filled circles. All other native contacts satisfying $|i-j|>4$ are shown as empty circles. Positive (blue) and negative (red) experimental -values satisfying $|\phi|>0.2$ are plotted along the diagonal as filled diamonds. Arrows illustrate the proposed connections between the experimental ϕ-values and the contacts that are either lost or gained in the simulated structural ensembles. Text labels indicate the interfaces that are either formed or broken during the transition. Note that the positive ϕ-value at position 219 (the only significantly positive -value in the C-terminal domain) is derived from a mutation that actually accelerates folding and unfolding, like those that lead to the negative ϕ-values, but is formally positive because the mutation slightly stabilizes (rather than destabilizes) the native state.

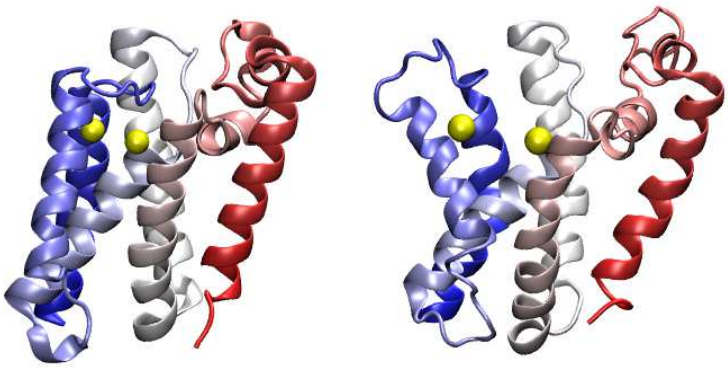


Figure 4.9 : A comparison of the closed (left, PDB ID: 2xov) and open (right, PDB ID:2nrf chain A) crystal structures of GlpG. The catalytic dyad (shown in yellow) is buried in the closed state and exposed in the open state. The largest differences between the open and closed states are in TM5 and L5. This observation led to the suggestion that TM5 serves as a gate for access to the catalytic dyad.

4.4.4 TM5 is loosely bound even under folding conditions.

GlpG is an intramembrane protease of the rhomboid serine protease class [97]. It cleaves specific transmembrane substrates using a catalytic dyad that is buried within the lipid bilayer [91]. Figure 4.9 shows two crystal structures of GlpG, one in a closed conformation, the one used to construct our structure-based energy landscape, and the other in an open conformation, where L5 and TM5 have bent away from the rest of the structure to expose partially the catalytic dyad. It has been suggested that TM5 functions as a substrate gate that opens for full-sized substrates to gain access to the catalytic site [91].

Preferential stabilization of the contacts within the N-terminal domain by 10% suffices to populate a near-native state (F^*) under folding conditions in the presence of the implicit

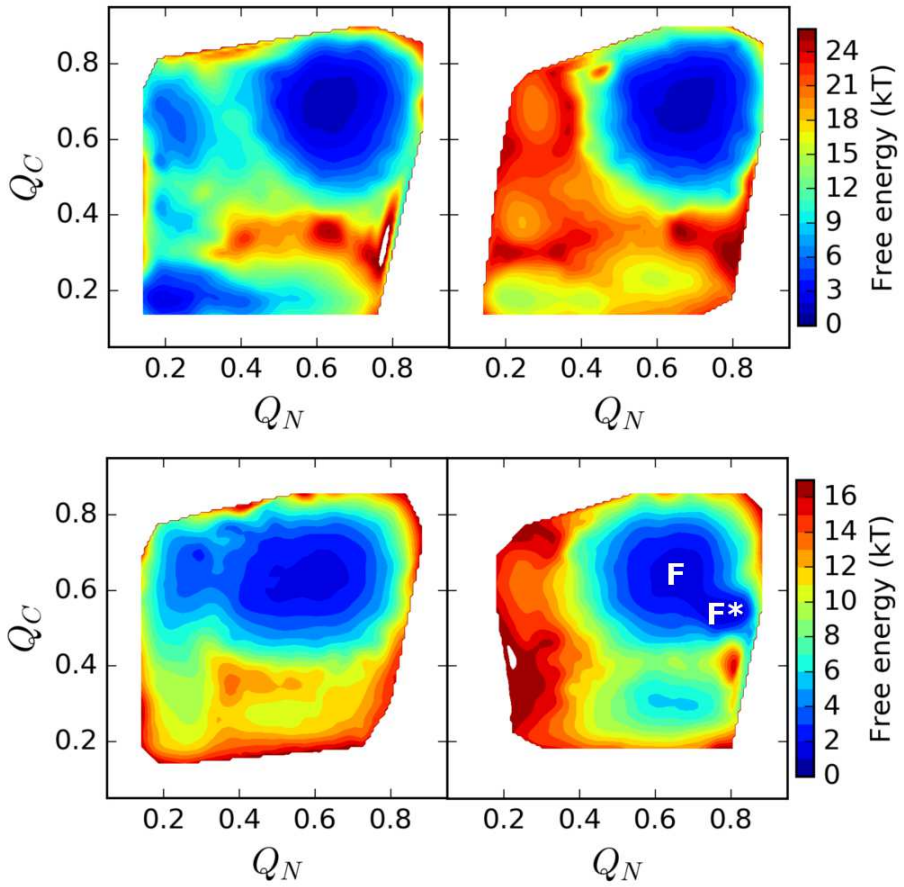


Figure 4.10 : Two dimensional free energy profiles of GlpG without (top) and with (bottom) the implicit membrane below the folding temperature, and with an N-terminal domain destabilized (left) and stabilized (right) by 10%. A near-native state (F^*) is highly populated and accessible from the folded state (F) when the N-terminal is stabilized and the implicit membrane is present.

membrane (Figure 4.10) according to our perturbation calculations. Structural analysis of this state revealed a heterogeneous ensemble of near-native conformations with a common feature: TM5 was unbound from TM4 and TM6, thereby exposing the catalytic dyad. In this state, deviations from the closed crystal structure occur most significantly in TM5 and the connecting loops L4 and L5 (Figure 4.11). Whether or not TM5 must undergo

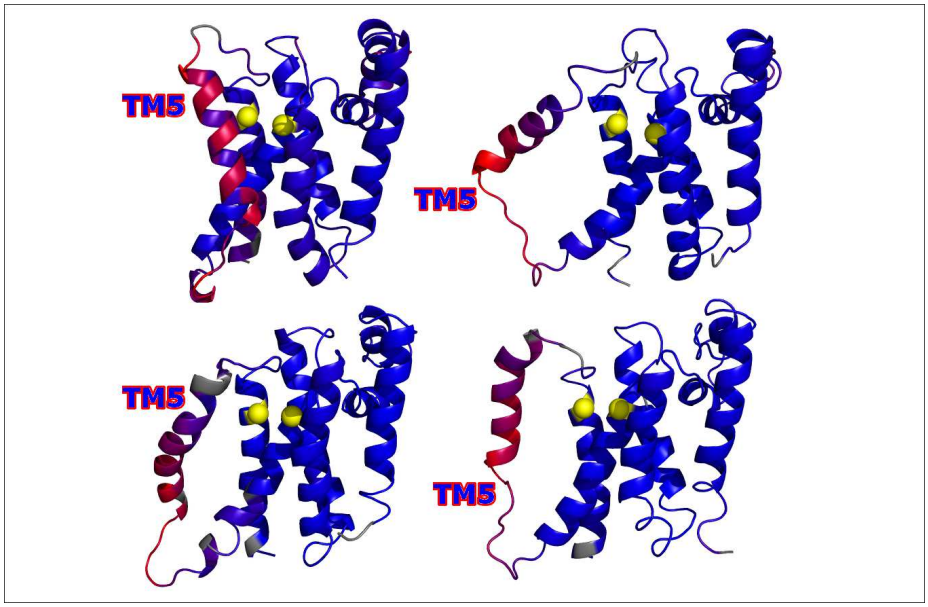


Figure 4.11 : Representative structures from a near-native state (F^*, Figure 5) sampled while simulating with the implicit membrane present. The structures were all aligned to the closed crystal structure (PDB ID: 2xov) and colored according to the individual residue RMSD values. Blue indicates low RMSD (high similarity to the crystal structure) and red indicates high RMSD. The catalytic dyad is shown using yellow spheres. High RMSD values are localized to the C-terminal half of the molecule and to TM5 in particular. Movement of TM5 exposes the catalytic dyad, thereby allowing substrate access. This state is highly populated under folding conditions when strengthening the contacts in the N-terminal half of the molecule by 10% relative to the contacts in the C-terminal half of the molecule.

significant conformational rearrangements in order for full-sized substrates to access the proteolytic site is a matter of some controversy [77, 91, 98]. Our model suggests that the conformation of TM5 is highly dynamic even under folding conditions, which is consistent with the experimental observation that tethering TM5 to TM2 eliminates enzymatic activity [74]. The fact that stabilizing the N-terminal part of the molecule increases the population of this state agrees with the experimental observation that destabilizing L1 reduces enzymatic activity [74, 91], highlighting the role of the N-terminal part of the molecule as a

structural scaffold. While TM5 is mobile in F^*, F^* differs, crucially, from TS2 in the implicit membrane (Figure 4.5) by TM6 remaining bound to TM4. The tight association between TM4 and TM6 is mediated by GXXXAXXG and GXXXGXXXA motifs, which stabilize the C-terminal domain and protect against unfolding during GlpGs functional motions.

4.5 Conclusions

Experiments that probe membrane protein folding on the single-residue [19, 99] and the single-molecule [76] levels begin to allow us to determine the mechanisms by which membrane proteins fold and function. Nevertheless, many details of these processes remain hidden to even the most sensitive experiments. Using mixed micelles provides powerful tools for investigating membrane protein biophysics due to its relative simplicity and general applicability, but the structure of the denatured state and its effect on folding mechanisms needs to be better understood. Thus far, studies of how residual structure in the denatured state affects folding have focused on soluble proteins and have employed atomistic simulations [100], NMR and other types of spectroscopy [101], or combinations of the two [102]. The question of residual structure is certainly no less important for membrane proteins, but the membrane environment poses challenges to both NMR and atomistic simulations. In this work, we used a coarse-grained energy landscape model to explore two limiting models of the folding of an intramembrane protease, GlpG: one limit in which the helices largely remain embedded in the membrane with their proper orientations, as is expected for the

denatured state in lipid bilayers, and another limit where no constraints are placed on the alignment of helices in the unfolded state, this being taken as a model for the SDS denatured state in micelles. Despite the simplicity of these models, on their basis, we have been able to propose a solution to the major puzzle in the experimental study of GlpGs folding mechanism, characterize a near-native state with potential functional significance, and show how these phenomena are related to GlpGs modular structure and topological constraints on the motions of partially folded states. The modular architecture of GlpG supports functional motions, including a highly mobile TM5, and leads to backtracking during the rate-limiting step of folding when the entropic cost of organizing helical bundles is high, as is the case in the absence of a bilayer. By providing a structurally detailed resolution of the ϕ-value puzzle, our analysis gives strong support to the notion that GlpG folding in mixed micelles proceeds by assembling helices with native levels of secondary structure from a state with few other constraints, as guided by a funneled, minimally frustrated landscape.

4.6 Acknowledgements

We thank Sin Urban for ongoing constructive discussions about GlpG. K.L.L. and N.P.S. acknowledge support from the Novo Nordisk Foundation. K.L.L., N.P.S. and D.E.O. were supported by the Danish Research Council (DFF-4090-00220) and the Carlsberg Foundation (CF14-0287). H.H.T and P.G.W. were supported by the National Institute of General Medical Sciences (Grant No. R01 GM44557) and the D.R. Bullard-Welch Chair at Rice University (Grant No. C-0016). Computational resources were supported in part by the

Data Analysis and Visualization Cyberinfrastructure funded by the NSF under Grant OCI-0959097.

4.7 Appendix

4.7.1 Simulation methodology

Simulations were performed using the AWSEM-MD [2] simulation package, which is implemented in the LAMMPS molecular dynamics simulation package [48]. We employed a modified version of the structure-based model described in Ref 50. The "p-value, which determines the degree non-additivity, was set to 1, yielding a pairwise additive model. The local-in-sequence contacts ($3 \leq |i-j| \leq 5$) were given a strength of 1, whereas long-range contacts ($|i-j| > 5$) were given a strength of 0.5. For soluble proteins, typical values for short and long range interactions are 0.25 and 0.5, respectively. A 6.5 Angstrom cutoff between C_β atoms (C_α for glycine) was used to define native contacts based on the crystal structure with PDB ID 2xov. The implicit membrane model used is described in Ref. 6 and the assignment of residues for GlpG is explained below. We performed two sets of simulations, with and without the implicit membrane present. For each model, umbrella sampling simulations using the potential given in Equation 1.5 were performed at 4 temperatures separated evenly by 25 K intervals. The temperature range in the case of the simulations with the implicit membrane was 150 K-225 K in the case of the simulations with the implicit membrane and 135 K-210 K in the case of the simulations without the implicit membrane. The folding temperature was determined empirically using the peak

of the heat capacity curves, and all temperatures referenced within the body of the paper were normalized to units of the folding temperature of each model independently. Twenty umbrella sampling simulations were run at each temperature with bias centers ranging from $Q_0 = 0.00$ to $Q_0 = 0.95$, spaced evenly. Each simulation was run for 20,000,000 timesteps of 2 fs each. Structures and energies were saved every 1,000 steps.

4.7.2 Order parameters

Several order parameters were calculated for all structures. Global Q (used for umbrella sampling), Q_N and Q_C as well as the secondary and tertiary structure foldedness parameters were calculated using Equation 1.4, varying only the pairs of residues that were summed over and the corresponding normalization constant such that all order parameters had a maximum range of 0 to 1. For global Q, all unique pairs of residues are included. For Q_N and Q_C, the sum runs over all unique pairs within the N- and C-terminal domains, residue IDs 91 to 171 and 172 to 271, respectively. The short-range foldedness is calculated by summing over all unique pairs satisfying $3 \leq |i - j| \leq 8$, the long-range foldedness by those pairs of residues satisfying $|i - j| \geq 9$. Intra- and inter-helical distances were calculated using the C_α atoms of select residues as described below.

4.7.3 Visualization of structures

Visualization of structures was performed using VMD [4] and pymol [3]. Representative structures were picked based on the range of Q_N and Q_C of the low free energy basins and transition states found in the free energy profiles. For each state, ten structures were

visualized, chosen evenly from throughout all samples that belong to that state, and aligned according to which parts of the molecule are folded in that state.

4.7.4 Implicit membrane energy term and topological assignment

The implicit membrane force field is a function of the z coordinates of the C_α atoms. Residues were assigned to be either intramembrane or extramembrane based on the z-coordinate of their C_β atom: $|z| < 15$ Angstroms, intramembrane, otherwise extramembrane. The proper topology of GlpG within the membrane was obtained directly from the three dimensional experimentally determined structure using the TMDET web server [64]. Residues are assigned to be in periplasmic, transmembrane, or cytoplasmic regions in the simulation, in which periplasmic and cytoplasmic environments are treated equally. Residues 135-143 (those residues in L1 that are below the membrane plane) were re-assigned to be in the transmembrane region. Proper topology of GlpG used in the implicit membrane model is shown in Figure 4.12.

4.7.5 Comparison to single molecule force spectroscopy study in bicelles

We noted a close structural correspondence between states described in Ref. 76 and those that we found during our simulations. Their structural interpretation of the unfolding via intermediates given in the main text (in their notation) ($N \rightarrow I1 \rightarrow I2 \rightarrow U$) closely corresponds to the reverse of one of our folding pathways ($F \rightarrow TS2 \rightarrow N \rightarrow U$) in the presence of the bilayer. Note that, in their notation, N refers to “native”, while in our nota-

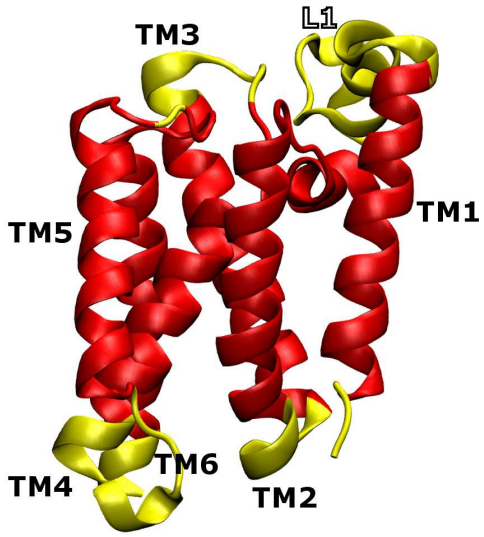


Figure 4.12 : Proper topology of the native structure of GlpG used in the implicit membrane model. Residues in the transmembrane region are colored in red. Periplasmic and cytoplasmic residues are colored in yellow. L1 is large and contains two interfacial helices, of which residues 137-143 were assigned to be in the transmembrane region.

tion N refers to "N-terminal". Using mutational perturbations and by examining unfolding rip lengths, they infer a unidirectional unfolding pathway that starts at the C-terminal and proceeds roughly two helices at a time. I1 therefore corresponds to unfolding of TM5 and TM6. In our TS2, TM6 is unfolded and TM4 and TM5 are in the process of being folded onto the N-terminal domain. I2 corresponds to the unfolding of two more helices, leaving only TM1 and TM2 folded. The N-terminal folded domain in our simulations (N) in the presence of the bilayer consists of a folded TM1 and TM2 and a partially folded TM3. The structural decomposition of GlpG into domains given in their supplementary information (into N, M and C domains) corresponds more or less exactly to the reverse of one of our dominant folding pathways ($F \rightarrow C2 \rightarrow C1 \rightarrow U$) in the absence of the bilayer. The N

domain consists of TM1 and L1, which is the unfolded part of the molecule in our C2. The M domain consists of TM2 and TM3, which are the two helices that unfold when going from C2 to C1 in our analysis. Finally, the C domain consists of TM4-6, which is the minimal folding unit for the C-terminal domain in our simulations and makes up the folded region in C1.

Bibliography

[1] P. Wolynes, Z. Luthey-Schulten, and J. Onuchic, "Fast-folding eriments and the topography of protein folding energy landscapes," *Chemistry & biology*, vol. 3, no. 6, pp. 425–432, 1996.

[2] A. Davtyan, N. P. Schafer, W. Zheng, C. Clementi, P. G. Wolynes, and G. A. Papoian, "Awsem-md: Protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing," *J. Phys. Chem. B*, vol. 116, no. 29, pp. 8494–8503, 2012.

[3] W. L. DeLano, "The pymol molecular graphics system," 2002.

[4] W. Humphrey, A. Dalke, and K. Schulten, "Vmd: visual molecular dynamics," *Journal of molecular graphics*, vol. 14, no. 1, pp. 33–38, 1996.

[5] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 235–242, 2000.

[6] B. L. Kim, N. P. Schafer, and P. G. Wolynes, "Predictive energy landscapes for folding α-helical transmembrane proteins," *Proceedings of the National Academy of Sciences*, vol. 111, no. 30, pp. 11031–11036, 2014.

[7] J. D. Bryngelson and P. G. Wolynes, "Spin glasses and the statistical mechanics of protein folding," *Proc. Natl. Acad. Sci. U.S.A*, vol. 84, no. 21, pp. 7524–7528, 1987.

[8] D. U. Ferreiro, J. A. Hegler, E. A. Komives, and P. G. Wolynes, "Localizing frustration in native proteins and protein assemblies," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 104, no. 50, pp. 19819–19824, 2007.

[9] D. U. Ferreiro, J. A. Hegler, E. A. Komives, and P. G. Wolynes, "On the role of frustration in the energy landscapes of allosteric proteins," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108, no. 9, pp. 3499–3503, 2011.

[10] P. O. Craig, J. Lätzer, P. Weinkam, R. M. Hoffman, D. U. Ferreiro, E. A. Komives, and P. G. Wolynes, "Prediction of native-state hydrogen exchange from perfectly funneled energy landscapes," *J. Am. Chem. Soc.*, vol. 133, no. 43, pp. 17463–17472, 2011.

[11] N. P. Schafer, R. M. Hoffman, A. Burger, P. O. Craig, E. A. Komives, and P. G. Wolynes, "Discrete kinetic models from funneled energy landscape simulations," *PloS one*, vol. 7, no. 12, p. e50635, 2012.

[12] S. S. Cho, Y. Levy, and P. G. Wolynes, "P versus q: Structural reaction coordinates capture protein folding on smooth landscapes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 3, pp. 586–591, 2006.

[13] W. H. Press, *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.

[14] C. Hardin, M. P. Eastwood, M. Prentiss, Z. Luthey-Schulten, and P. G. Wolynes, "Folding funnels: the key to robust protein structure prediction," *Journal of computational chemistry*, vol. 23, no. 1, pp. 138–146, 2002.

[15] M. R. Shirts and J. D. Chodera, "Statistically optimal analysis of samples from multiple equilibrium states," *J. Chem. Phys.*, vol. 129, no. 12, 2008.

[16] H. H. Truong, B. L. Kim, N. P. Schafer, and P. G. Wolynes, "Funneling and frustration in the energy landscapes of some designed and simplified proteins," *The Journal of chemical physics*, vol. 139, no. 12, p. 121908, 2013.

[17] H. H. Truong, B. L. Kim, N. P. Schafer, and P. G. Wolynes, "Predictive energy landscapes for folding membrane protein assemblies," *The Journal of chemical physics*, vol. 143, no. 24, p. 243101, 2015.

[18] N. P. Schafer, H. H. Truong, D. E. Otzen, K. Lindorff-Larsen, and P. G. Wolynes, "Topological constraints and modular structure in the folding and functional motions of glpg, an intramembrane protease," *Proceedings of the National Academy of Sciences*, vol. 113, no. 8, pp. 2098–2103, 2016.

[19] W. Paslawski, O. K. Lillelund, J. V. Kristensen, N. P. Schafer, R. P. Baker, S. Urban, and D. E. Otzen, "Cooperative folding of a polytopic α-helical membrane protein

involves a compact n-terminal nucleus and nonnative loops," *Proceedings of the National Academy of Sciences*, vol. 112, no. 26, pp. 7978–7983, 2015.

[20] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, "Funnels, pathways, and the energy landscape of protein folding: a synthesis," *Proteins*, vol. 21, no. 3, pp. 167–195, 1995.

[21] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes, "Theory of protein folding: the energy landscape perspective," *Annu. Rev. Phys. Chem.*, vol. 48, no. 1, pp. 545–600, 1997.

[22] P. G. Wolynes, "Energy landscapes and solved protein–folding problems," *Phil. Trans. R. Soc. A*, vol. 363, no. 1827, pp. 453–467, 2005.

[23] J. N. Onuchic and P. G. Wolynes, "Theory of protein folding," *Curr. Opin. Struct. Biol.*, vol. 14, no. 1, pp. 70–75, 2004.

[24] M. Oliveberg and P. G. Wolynes, "The experimental survey of protein-folding energy landscapes," *Q. Rev. Biophys.*, vol. 38, no. 03, pp. 245–288, 2005.

[25] O. Miyashita, J. N. Onuchic, and P. G. Wolynes, "Nonlinear elasticity, protein-quakes, and the energy landscapes of functional transitions in proteins," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 100, no. 22, pp. 12570–12575, 2003.

[26] B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, and D. Baker, "De-

sign of a novel globular protein fold with atomic-level accuracy," *Science*, vol. 302, no. 5649, pp. 1364–1368, 2003.

[27] W. Jin, O. Kambara, H. Sasakawa, A. Tamura, and S. Takada, "De novo design of foldable proteins with smooth folding funnel: Automated negative design and experimental verification," *Structure*, vol. 11, no. 5, pp. 581–590, 2003.

[28] N. Koga, R. Tatsumi-Koga, G. Liu, R. Xiao, T. B. Acton, G. T. Montelione, and D. Baker, "Principles for designing ideal protein structures," *Nature*, vol. 491, no. 7423, pp. 222–227, 2012.

[29] M. Scalley-Kim and D. Baker, "Characterization of the folding energy landscapes of computer generated proteins suggests high folding free energy barriers and cooperativity may be consequences of natural selection," *J. Mol. Biol.*, vol. 338, no. 3, pp. 573–583, 2004.

[30] A. L. Watters, P. Deka, C. Corrent, D. Callender, G. Varani, T. Sosnick, and D. Baker, "The highly cooperative folding of small naturally occurring proteins is likely the result of natural selection," *Cell*, vol. 128, no. 3, pp. 613–624, 2007.

[31] Z. Zhang and H. S. Chan, "Native topology of the designed protein top7 is not conducive to cooperative folding," *Biophys. J.*, vol. 96, no. 3, pp. L25–L27, 2009.

[32] Z. Zhang and H. S. Chan, "Competition between native topology and nonnative interactions in simple and complex folding kinetics of natural and designed proteins," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 107, no. 7, pp. 2920–2925, 2010.

[33] Y. Fujitsuka, S. Takada, Z. A. Luthey-Schulten, and P. G. Wolynes, "Optimizing physical energy functions for protein folding," *Proteins: Struct., Funct., Bioinf.*, vol. 54, no. 1, pp. 88–103, 2004.

[34] Y. Kuroda and P. S. Kim, "Folding of bovine pancreatic trypsin inhibitor (bpti) variants in which almost half the residues are alanine," *J. Mol. Biol.*, vol. 298, no. 3, pp. 493–501, 2000.

[35] M. M. Islam, S. Sohya, K. Noguchi, M. Yohda, and Y. Kuroda, "Crystal structure of an extensively simplified variant of bovine pancreatic trypsin inhibitor in which over one-third of the residues are alanines," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 40, pp. 15334–15339, 2008.

[36] K. Yue, K. M. Fiebig, P. D. Thomas, H. S. Chan, E. I. Shakhnovich, and K. A. Dill, "A test of lattice protein folding algorithms," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 92, no. 1, pp. 325–329, 1995.

[37] P. G. Wolynes, "As simple as can be?," *Nat. Struct. Mol. Biol.*, vol. 4, no. 11, pp. 871–874, 1997.

[38] A. R. Davidson, K. J. Lumb, and R. T. Sauer, "Cooperatively folded proteins in random sequence libraries," *Nat. Struct. Biol.*, vol. 2, no. 10, pp. 856–864, 1995.

[39] D. S. Riddle, J. V. Santiago, S. T. Bray-Hall, N. Doshi, V. P. Grantcharova, Q. Yi, and D. Baker, "Functional rapidly folding proteins from simplified amino acid sequences," *Nat. Struct. Biol.*, vol. 4, no. 10, pp. 805–809, 1997.

[40] J. Wang and W. Wang, "A computational approach to simplifying the protein folding alphabet," *Nat. Struct. Mol. Biol.*, vol. 6, no. 11, pp. 1033–1038, 1999.

[41] H. S. Chan, "Folding alphabets.," *Nat. Struct. Biol.*, vol. 6, no. 11, p. 994, 1999.

[42] K. Fan and W. Wang, "What is the minimum number of letters required to fold a protein?," *J. Mol. Biol.*, vol. 328, no. 4, pp. 921–926, 2003.

[43] T. Li, K. Fan, J. Wang, and W. Wang, "Reduction of protein sequence complexity by residue grouping," *Protein Eng.*, vol. 16, no. 5, pp. 323–330, 2003.

[44] L. R. Murphy, A. Wallqvist, and R. M. Levy, "Simplified amino acid alphabets for protein fold recognition and implications for folding," *Protein Eng.*, vol. 13, no. 3, pp. 149–152, 2000.

[45] P. D. Thomas and K. A. Dill, "An iterative method for extracting energy-like quantities from protein structures," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 93, no. 21, pp. 11628–11633, 1996.

[46] P. G. Wolynes, "Symmetry and the energy landscapes of biomolecules," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 93, no. 25, p. 14249, 1996.

[47] E. Haglund, M. O. Lindberg, and M. Oliveberg, "Changes of protein folding pathways by circular permutation," *J. Biol. Chem.*, vol. 283, no. 41, pp. 27904–27915, 2008.

[48] S. Plimpton, P. Crozier, and A. Thompson, "Lammps-large-scale atomic/molecular massively parallel simulator," *Sandia National Laboratories*, 2007.

[49] G. A. Papoian, J. Ulander, and P. G. Wolynes, "Role of water mediated interactions in protein-protein recognition landscapes," *J. Am. Chem. Soc.*, vol. 125, no. 30, pp. 9170–9178, 2003.

[50] M. P. Eastwood and P. G. Wolynes, "Role of explicitly cooperative interactions in protein folding funnels: A simulation study," *J. Chem. Phys.*, vol. 114, p. 4702, 2001.

[51] M. Heinig and D. Frishman, "Stride: a web server for secondary structure assignment from known atomic coordinates of proteins," *Nucleic Acids Res.*, vol. 32, no. suppl 2, pp. W500–W502, 2004.

[52] J. A. Cuff, M. E. Clamp, A. S. Siddiqui, M. Finlay, and G. J. Barton, "Jpred: a consensus secondary structure prediction server.," *Bioinformatics*, vol. 14, no. 10, pp. 892–893, 1998.

[53] M. U. Johansson, M. de Château, M. Wikström, S. Forsén, T. Drakenberg, and L. Björck, "Solution structure of the albumin-binding ga module: a versatile bacterial protein domain," *J. Mol. Biol.*, vol. 266, no. 5, pp. 859–865, 1997.

[54] M. Jenik, R. G. Parra, L. G. Radusky, A. Turjanski, P. G. Wolynes, and D. U. Ferreiro, "Protein frustratometer: a tool to localize energetic frustration in protein molecules," *Nucleic Acids Res.*, vol. 40, no. W1, pp. W348–W351, 2012.

[55] Y. Levy, S. S. Cho, J. N. Onuchic, and P. G. Wolynes, "A survey of flexible protein binding mechanisms and their transition states using native topology based energy landscapes," *Journal of molecular biology*, vol. 346, no. 4, pp. 1121–1145, 2005.

[56] G. A. Papoian, J. Ulander, M. P. Eastwood, Z. Luthey-Schulten, and P. G. Wolynes, "Water in protein structure prediction," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 10, pp. 3352–3357, 2004.

[57] C. Zong, G. A. Papoian, J. Ulander, and P. G. Wolynes, "Role of topology, nonadditivity, and water-mediated interactions in predicting the structures of α/β proteins," *Journal of the American Chemical Society*, vol. 128, no. 15, pp. 5168–5176, 2006.

[58] W. Zheng, N. P. Schafer, A. Davtyan, G. A. Papoian, and P. G. Wolynes, "Predictive energy landscapes for protein–protein association," *Proceedings of the National Academy of Sciences*, vol. 109, no. 47, pp. 19244–19249, 2012.

[59] C. A. Schramm, B. T. Hannigan, J. E. Donald, C. Keasar, J. G. Saven, W. F. DeGrado, and I. Samish, "Knowledge-based potential for positioning membrane-associated structures and assessing residue-specific energetic contributions," *Structure*, vol. 20, no. 5, pp. 924–935, 2012.

[60] K. Huang, H. Bayley, M.-J. Liao, E. London, and H. Khorana, "Refolding of an integral membrane protein. denaturation, renaturation, and reconstitution of intact bacteriorhodopsin and two proteolytic fragments.," *Journal of Biological Chemistry*, vol. 256, no. 8, pp. 3802–3809, 1981.

[61] J.-L. Popot, S.-E. Gerchman, and D. M. Engelman, "Refolding of bacteriorhodopsin in lipid bilayers: a thermodynamically controlled two-stage process," *Journal of molecular biology*, vol. 198, no. 4, pp. 655–676, 1987.

[62] G. Von Heijne, "Membrane protein structure prediction: hydrophobicity analysis and the positive-inside rule," *Journal of molecular biology*, vol. 225, no. 2, pp. 487–494, 1992.

[63] G. von Heijne, "Membrane-protein topology," *Nat. Rev. Mol. Cell Biol.*, vol. 7, pp. 909–918, Dec. 2006.

[64] G. E. Tusnády, Z. Dosztányi, and I. Simon, "TMDET: web server for detecting transmembrane regions of proteins by using their 3D coordinates," *Bioinformatics*, vol. 21, pp. 1276–1277, 1 Apr. 2005.

[65] T. Nugent and D. T. Jones, "Transmembrane protein topology prediction using support vector machines," *BMC bioinformatics*, vol. 10, no. 1, p. 159, 2009.

[66] R. A. Goldstein, Z. A. Luthey-Schulten, and P. G. Wolynes, "Protein tertiary structure recognition using optimized hamiltonians with local interactions.," *Proceedings of the National Academy of Sciences*, vol. 89, no. 19, pp. 9029–9033, 1992.

[67] P. E. Leopold, M. Montal, and J. N. Onuchic, "Protein folding funnels: a kinetic approach to the sequence-structure relationship.," *Proceedings of the National Academy of Sciences*, vol. 89, no. 18, pp. 8721–8725, 1992.

[68] N. P. Schafer, B. L. Kim, W. Zheng, and P. G. Wolynes, "Learning to fold proteins using energy landscape theory," *Israel journal of chemistry*, vol. 54, no. 8-9, pp. 1311–1337, 2014.

[69] S. Yang, S. S. Cho, Y. Levy, M. S. Cheung, H. Levine, P. G. Wolynes, and J. N. Onuchic, "Domain swapping is a consequence of minimal frustration," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 38, pp. 13786–13791, 2004.

[70] C. B. Anfinsen *et al.*, "Principles that govern the folding of protein chains," *Science*, vol. 181, no. 4096, pp. 223–230, 1973.

[71] N. Unwin, "Refined structure of the nicotinic acetylcholine receptor at 4å resolution," *Journal of molecular biology*, vol. 346, no. 4, pp. 967–989, 2005.

[72] T. Murata, I. Yamato, Y. Kakinuma, A. G. Leslie, and J. E. Walker, "Structure of the rotor of the v-type na+-atpase from enterococcus hirae," *Science*, vol. 308, no. 5722, pp. 654–659, 2005.

[73] L.-O. Essen, R. Siegert, W. D. Lehmann, and D. Oesterhelt, "Lipid patches in membrane protein oligomers: crystal structure of the bacteriorhodopsin-lipid complex," *Proceedings of the National Academy of Sciences*, vol. 95, no. 20, pp. 11673–11678, 1998.

[74] R. P. Baker, K. Young, L. Feng, Y. Shi, and S. Urban, "Enzymatic analysis of a rhomboid intramembrane protease implicates transmembrane helix 5 as the lateral

substrate gate," *Proceedings of the National Academy of Sciences*, vol. 104, no. 20, pp. 8257–8262, 2007.

[75] R. P. Baker and S. Urban, "Architectural and thermodynamic principles underlying intramembrane protease function," *Nature chemical biology*, vol. 8, no. 9, pp. 759–768, 2012.

[76] D. Min, R. E. Jefferson, J. U. Bowie, and T.-Y. Yoon, "Mapping the energy landscape for second-stage folding of a single membrane protein," *Nature chemical biology*, 2015.

[77] S. Zoll, S. Stanchev, J. Began, J. Škerle, M. Lepšík, L. Peclinovská, P. Majer, and K. Strisovsky, "Substrate binding and specificity of rhomboid intramembrane protease revealed by substrate–peptide complex structures," *The EMBO journal*, p. e201489367, 2014.

[78] J.-L. Popot and D. M. Engelman, "Membrane protein folding and oligomerization: the two-stage model," *Biochemistry*, vol. 29, no. 17, pp. 4031–4037, 1990.

[79] M. Pohlschröder, W. A. Prinz, E. Hartmann, and J. Beckwith, "Protein translocation in the three domains of life: variations on a theme," *Cell*, vol. 91, no. 5, pp. 563–566, 1997.

[80] B. Zhang and T. F. Miller, "Long-timescale dynamics and regulation of sec-facilitated protein translocation," *Cell reports*, vol. 2, no. 4, pp. 927–937, 2012.

[81] K. A. Dill, "Dominant forces in protein folding," *Biochemistry*, vol. 29, no. 31, pp. 7133–7155, 1990.

[82] W. Kauzmann, "Some factors in the interpretation of protein denaturation," *Advances in protein chemistry*, vol. 14, pp. 1–63, 1959.

[83] B. Honig and A. Nicholls, "Classical electrostatics in biology and chemistry," *Science*, vol. 268, no. 5214, pp. 1144–1149, 1995.

[84] J. P. Schlebach, N. B. Woodall, J. U. Bowie, and C. Park, "Bacteriorhodopsin folds through a poorly organized transition state," *Journal of the American Chemical Society*, vol. 136, no. 47, pp. 16574–16581, 2014.

[85] P. Curnow, N. D. Di Bartolo, K. M. Moreton, O. O. Ajoje, N. P. Saggese, and P. J. Booth, "Stable folding core in the folding transition state of an α-helical integral membrane protein," *Proceedings of the National Academy of Sciences*, vol. 108, no. 34, pp. 14133–14138, 2011.

[86] D. E. Otzen, "Mapping the folding pathway of the transmembrane protein dsbb by protein engineering," *Protein Engineering Design and Selection*, p. gzq079, 2010.

[87] A. R. Fersht and S. Sato, "ϕ-value analysis and the nature of protein-folding transition states," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 21, pp. 7976–7981, 2004.

[88] L. S. Itzhaki, D. E. Otzen, and A. R. Fersht, "The structure of the transition state

for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding," *Journal of molecular biology*, vol. 254, no. 2, pp. 260–288, 1995.

[89] N. Go, "Theoretical studies of protein folding," *Annual review of biophysics and bioengineering*, vol. 12, no. 1, pp. 183–210, 1983.

[90] Y. Levy, P. G. Wolynes, and J. N. Onuchic, "Protein topology determines binding mechanism," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 2, pp. 511–516, 2004.

[91] Z. Wu, N. Yan, L. Feng, A. Oberstein, H. Yan, R. P. Baker, L. Gu, P. D. Jeffrey, S. Urban, and Y. Shi, "Structural analysis of a rhomboid family intramembrane protease reveals a gating mechanism for substrate entry," *Nature structural & molecular biology*, vol. 13, no. 12, pp. 1084–1091, 2006.

[92] K. R. Vinothkumar, K. Strisovsky, A. Andreeva, Y. Christova, S. Verhelst, and M. Freeman, "The structural basis for catalysis and substrate specificity of a rhomboid protease," *The EMBO journal*, vol. 29, no. 22, pp. 3797–3809, 2010.

[93] R. Du, V. S. Pande, A. Y. Grosberg, T. Tanaka, and E. S. Shakhnovich, "On the transition coordinate for protein folding," *The Journal of chemical physics*, vol. 108, no. 1, pp. 334–350, 1998.

[94] W. Zheng and R. B. Best, "Reduction of all-atom protein folding dynamics to one-

dimensional diffusion," *The Journal of Physical Chemistry B*, vol. 119, no. 49, pp. 15247–15255, 2015.

[95] M. Eilers, A. B. Patel, W. Liu, and S. O. Smith, "Comparison of helix interactions in membrane and soluble α-bundle proteins," *Biophysical journal*, vol. 82, no. 5, pp. 2720–2736, 2002.

[96] V. Krishnamani, B. G. Hegde, R. Langen, and J. K. Lanyi, "Secondary and tertiary structure of bacteriorhodopsin in the sds denatured state," *Biochemistry*, vol. 51, no. 6, pp. 1051–1060, 2012.

[97] K. R. Vinothkumar and M. Freeman, "Intramembrane proteolysis by rhomboids: catalytic mechanisms and regulatory principles," *Current opinion in structural biology*, vol. 23, no. 6, pp. 851–858, 2013.

[98] Y. Wang, Y. Zhang, and Y. Ha, "Crystal structure of a rhomboid family intramembrane protease," *Nature*, vol. 444, no. 7116, pp. 179–180, 2006.

[99] H. Hong, T. M. Blois, Z. Cao, and J. U. Bowie, "Method to measure strong protein–protein interactions in lipid bilayers using a steric trap," *Proceedings of the National Academy of Sciences*, vol. 107, no. 46, pp. 19802–19807, 2010.

[100] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, "How fast-folding proteins fold," *Science*, vol. 334, no. 6055, pp. 517–520, 2011.

[101] U. Mayor, J. G. Grossmann, N. W. Foster, S. M. Freund, and A. R. Fersht, "The

denatured state of engrailed homeodomain under denaturing and native conditions," *Journal of molecular biology*, vol. 333, no. 5, pp. 977–991, 2003.

[102] K. Lindorff-Larsen, S. Kristjansdottir, K. Teilum, W. Fieber, C. M. Dobson, F. M. Poulsen, and M. Vendruscolo, "Determination of an ensemble of structures representing the denatured state of the bovine acyl-coenzyme a binding protein," *Journal of the American Chemical Society*, vol. 126, no. 10, pp. 3291–3299, 2004.