Protein Folding & Dynamics Using Multi-scale Computational Methods

by

Taisong Zou

A Dissertation Presented in Partial Fulfillment
of the Requirement for the Degree
Doctor of Philosophy

Approved April 2014 by the
Graduate Supervisory Committee:

S. Banu Ozkan, Chair
Michael F. Thorpe
Neal W. Woodbury
Sara M. Vaiana
Giovanna Ghirlanda

ARIZONA STATE UNIVERSITY

May 2014

ABSTRACT

This thesis explores a wide array of topics related to the protein folding problem, ranging from the folding mechanism, *ab initio* structure prediction and protein design, to the mechanism of protein functional evolution, using multi-scale approaches.

To investigate the role of native topology on folding mechanism, the native topology is dissected into non-local and local contacts. The number of non-local contacts and non-local contact orders are both negatively correlated with folding rates, suggesting that the non-local contacts dominate the barrier-crossing process. However, local contact orders show positive correlation with folding rates, indicating the role of a diffusive search in the denatured basin. Additionally, the folding rate distribution of E. coli and Yeast proteomes are predicted from native topology. The distribution is fitted well by a diffusion-drift population model and also directly compared with experimentally measured half life. The results indicate that proteome folding kinetics is limited by protein half life.

The crucial role of local contacts in protein folding is further explored by the simulations of WW domains using Zipping and Assembly Method. The correct formation of N-terminal β-turn turns out important for the folding of WW domains. A classification model based on contact probabilities of five critical local contacts is constructed to predict the foldability of WW domains with 81% accuracy. By introducing mutations to stabilize those critical local contacts, a new protein design approach is developed to re-design the unfoldable WW domains and make them foldable.

After folding, proteins exhibit inherent conformational dynamics to be functional. Using molecular dynamics simulations in conjunction with Pertur-

bation Response Scanning, it is demonstrated that the divergence of functions can occur through the modification of conformational dynamics within existing fold for β-lactmases and GFP-like proteins: i) the modern TEM-1 lactamase shows a comparatively rigid active-site region, likely reflecting adaptation for efficient degradation of a specific substrate, while the resurrected ancient lactamases indicate enhanced active-site flexibility, which likely allows for the binding and subsequent degradation of different antibiotic molecules; ii) the chromophore and attached peptides of photocoversion-competent GFP-like protein exhibits higher flexibility than the photocoversion-incompetent one, consistent with the evolution of photocoversion capacity.

ACKNOWLEDGEMENTS

This doctoral thesis could not have been completed without the help and support from dozens of remarkable people around me. In particular I would like to gratefully and sincerely thank my PhD advisor, Dr. S. Banu Ozkan, for her guidance and support, and most importantly, her friendship during my graduate studies at Arizona State University. Her mentorship was invaluable to me on both an academic and a personal level. I will forever be thankful for having such a knowledgeable, enthusiastic, patient and supportive advisor to work with.

I would also like to thank Dr. Mike F. Thorpe. I started my PhD research with him. He has shared many insightful discussions about research and has given me the freedom to pursue the research of my interest.

My thesis committee guided me through all these years. Thank you to Dr. Giovanna Ghirlanda, Dr. Neal Woodbury and Dr. Sara M. Vaiana for the input, valuable discussions and accessibility. Dr. Ghirlanda has also been a terrific collaborator. She and Brian Woodrum not only provided the experimental data, but also educated me on protein synthesis and purification, as well as characterization. Additionally, I must thank Dr. Kingshuk Ghosh, Dr. Rebekka Wachter, Dr. Jose Manuel Sanchez-Ruiz, Dr. Silvia Cavagnero and Nicholas Williams, for being cooperative in many aspects. I could not have accomplished what I did without their support.

I acknowledge Dr. Adam DeGraff, Dr. Tyler Glembo, Dr. Z. Nevin Gerek and Ashini Bolia for their constructive suggestions and assistance. I also thank Avishek Kumar and Paul Campitelli for carefully reading the manuscript.

I would like to thank the Department of Physics at Arizona State University. It would not have been possible to complete my doctoral study without

the resources and financial assistantship. Thanks to the Extreme Science and Engineering Discovery Environment (XSEDE) and ASU Advanced Computing Center, for they made all computations possible.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Chapter 1

INTRODUCTION

Proteins are linear polypeptide chains composed of amino acids linked by peptide bonds. They are not only mechanical and structural parts of organisms, but also participating in almost every biological process in a living cell, such as enzymatic reactions, cell signaling and immune response [55, 191]. These innumerous roles are made possible only after the protein folds into a well-defined 3-D structure. The milestone work of Anfinsen and coworkers demonstrated that, without the assistance of any biochemical machinery, an enzyme with 124 residues including four disulfide bridges called ribonuclease A can correctly arrange all the pieces of amino acids and reach its globally native structure after completely unfolding in urea denaturant [10, 9]. This exciting discovery led to two powerful postulations: i) the structure of a native protein is thermodynamically stable, in which the protein reaches an unique structure with globally minimal free energy. This is referred to as the *thermodynamic hypothesis* of protein folding. ii) the native protein structure, and even consequent protein function, are encoded in the primary sequence of the protein [66]. Since then, understanding how the primary sequence guides the protein to a particular 3-D structure, known as the protein folding problem, has become a significant scientific endeavor for biologists, chemists, physicists and engineers.

Protein folding is a remarkable self-assembly process. First, protein structures are much more distinctive and complicated, compared to other biomolecular structures like a double stranded helix of DNA with repetitive units. Ac-

cording to the 2009 report of SCOP (Structural Classification of Proteins), a major database in the world of protein fold classification, 38221 PDB structures have been classified, corresponding to a total of 1195 folds [2]. Second, proteins can quickly fold into native states in the time scale of microseconds to seconds, while a random search for a native structure in the vast conformational space may require an astronomical amount of time [155]. This fact is known as the *Levinthal's Paradox*. An straightforward solution to resolve this Paradox is that protein must fold through specific folding pathways by a mechanism rather than a random search [154].

Several models have emerged to elaborate the folding mechanism. The *framework* model [138, 198, 199, 23] assumes a hierarchical assembly starting with the formation of local secondary structure elements, and then subsequent gluing of those substructures. The formation of local elements here is independent of the tertiary structure. The *hydrophobic-collapse* model [67, 68, 69, 71] claims that hydrophobic interactions provided by the expulsion of water are the dominant driving force of protein folding. The folding starts with a rapid hydrophobic collapse of nonpolar residues, followed by the formation of many secondary structures and native contacts in the tertiary structure. The hydrophobic core stabilizes the folding intermediates, which is referred to as a molten globule corresponding to a partially folded state. In the *diffusion-collision* model [131, 132, 133], microdomains (secondary structure motifs or hydrophobic clusters) form first and diffuse till they collide. Upon collision they may coalesce to form larger units. The diffusion process acts as the rate-limiting step in the folding. The *nucleation-condensation* model, which combined features of both the *hydrophobic-collapse* and *framework* models, proposed concurrent formation of secondary and tertiary native contacts. The

long-range and other native hydrophobic contacts stabilize a weak nucleus with marginal stability, which serves as a template for the rapid condensation of the remaining residues surrounding it. In most proteins, the combination of long-range tertiary interactions and secondary structures form a stable fold, known as the transition state of folding. The rate-limiting step is assumed to be the formation of the transition state. In fact, these models are not mutually exclusive and all of them have gained some experimental support. They just capture different aspects of protein folding, which also indicates the complexity of the protein folding problem. Moreover, this problem has become even more complex. The universality of *thermodynamic hypothesis* is questioned, since the structures of a few proteins, including insulin, α-lytic protease [228] and serpins [254], are dependent on the conditions of the experiment. These proteins are considered as instances of *kinetic control* folding [20]. Their energy surface contains multiple minima with sufficiently high energy barriers. The protein which starts folding from an unfolded state around one minimum would end up kinetically trapped there and cannot reach other minima.

To provide a more general description of protein folding, the recent view of of protein folding replaced the concept of pathway by the picture of funnel-shaped free energy landscape [39, 70, 72, 153, 181, 50, 66]. The free energy landscape represents the available conformational space of a protein. It is a high-dimensional surface in theory but often shown in 3-D, where x and y axes are conformational degrees of freedom (reaction coordinates) and z-axis represents the corresponding free energy. The shape of free energy landscape is like a funnel. The wide rim of the funnel represents the unfolded state with a vast conformational space (high entropy) and it narrows as one gets down to the bottom, where the near-native states are more compact (low entropy).

For those thermodynamically stable proteins, a single global minimum exists at the bottom of the funnel, representing the native folded state. The free energy landscape of protein is rugged (i.e., frustrated) to some extent, i.e., riddled with local minima where protein may be trapped transiently. A protein shapes its energy landscape with the energetic interactions among its residues, leading to different degrees of ruggedness on the landscape surface [181]. The artificially designed proteins with optimized stability tend to have a less rugged free energy landscape, suggesting that the ruggedness is the result of evolution [214]. The *principle of minimal frustration* of protein folding introduced in the late 1980s claims that proteins by evolution shape their free energy landscape to a surface with minimum energetic frustration, i.e., fewer non-native contacts and less energetic traps competing with the global energetic minima) [39, 181, 50]. The process of protein folding is like rolling a ball from the top of the funnel to the bottom. The funnel-shaped free energy landscape predict that the protein folding is a parallel event rather than a sequential event. There are multiple parallel pathways from a large ensemble of unfolded conformations to the native states at the bottom of the funnel. The protein exists as an ensemble of conformations at any stage of folding.

The funnel shape of the free energy landscape is largely determined by the entropy of a protein. It leads to a prediction that the topology of the native structure plays an important role in folding mechanism and folding kinetics [19, 182, 113]. Many topological metrics correlate with the folding speed of the proteins, whose folding timescale spans a range of more than eight orders [104, 110, 103, 266, 124, 119, 105, 157, 73] , after Plaxco *et al* observed the negative correlation between the logarithm of folding rates of two-state proteins and the contact order [195, 196]. In Chapter 3, I will discuss in detail

how the topology of a protein is related to the folding kinetics [268] and utilize its prediction to investigate the folding kinetics on the proteome level [269].

Besides the study of folding mechanism, considerable efforts have been made toward predicting 3-D structure from its 1-D sequence. The motivation of protein structure prediction arises from the demand to determine the structure and function rapidly and efficiently in the post-genomic era, as a complement or substitution of slow and expensive experiments. There are two general approaches of structure prediction, bioinformatics-based methods and physics-based methods. The progress and performance of various prediction methods are assessed and documented by CASP (Critical Assessment of Techniques for Protein Structure Prediction) competition every two years since 1994 [1]. Currently the most accurate bioinformatics-based method is homology modeling which assumes that similar sequences lead to similar structures. It aligns the sequence of unknown structure with the sequences of known structures, and use the structure of the sequence homology with high similarity as a folding template. Another bioinformatics-based method is fold recognition. It threads each amino acid of unknown sequence to a known structure template (fold) and evaluates the compatibility with the structure. In general, physics-based methods perform *ab initio* simulation with empirical energy function to calculate the native structure from an extended unfolded conformation on the basis of thermodynamic hypothesis, which suggests that the native structure is corresponding to the global minimum energy. They have two advantages over bioinformatics-based methods: i) they provide the physical picture of the folding event and reveals the underlying folding mechanism, such as folding pathways. ii) they do not require the knowledge of known structures. However, due to the limitation of computational power, physics-based methods have only

been applied to proteins with small sizes. In 1988, Duan and Kollman performed 1 ms molecular dynamics simulation of the 36-residue villin headpiece subdomain in explicit water [76]. The simulation achieved a compact structure with 4.5 Å root mean square deviation of C_α atoms (C_α-RMSD) to the NMR strucuture and revealed two folding pathways. In 2003, Pitera and Swope carried out 92ns replica-exchange molecular dynamics (MD) simulation of the 20-residue Trp-cage peptide in implicit solvent, yielding C_α-RMSD<1.0 Å to its NMR structure [193]. Vijay Pande initiated Folding@home in early 2000s, a distributed computing project using idle computing resources of personal computers owned by volunteers over the world for the simulation of protein folding [27]. Researchers of the Folding@home project reported 2.73 ms MD simulation of the Fip35 Hpin1 WW domain [83] and 1ms MD simulation on GPU of the 39-residue protein NTL9 [247]. In 2011, DE Shaw *et al* reported the successful folding of 12 structurally diverse proteins up to 1 ms using their specialized designed supercomputer for MD simulation called Anton [159]. The achieved C_α-RMSD to the experimental structures ranges between 0.5 Å to 4.8 Å.

These physics-based structure prediction are extremely computational expensive as they use the brutal MD simulation to search for the native structure from the extended initial structure. The computational cost can be significantly reduced if one incorporates certain folding mechanism with simulation. An algorithm called *Zipping and Assembly Method* (ZAM) implemented this idea and successfully predicted eight out of nine proteins (25 to 73 residues in length) with average 2.2 Å C_α-RMSD from the experimental structures [188]. The unsuccessful prediction of the last protein appeared to be caused by the flaw of the implicit solvent model [188]. Here is the general procedure of ZAM

[188, 100, 222]: i) segment the whole chain into overlapping small fragments with 8 residues (8-mers), which are simulated separately using REMD. ii) grow or zip the fragments that have metastable structures by adding a few new residues or assembling two such fragments together, with further REMD and iterations. iii) lock in place any stable residue-residue contacts with a harmonic spring, enforcing emerging putative physical folding routes, without the need to sample huge numbers of degrees of freedom at a time. In Chapter 4, I will use ZAM to predict protein structures and design new foldable sequences.

The accuracy and efficiency of structure prediction can be increased if one better understands the evolutionary information in sequences. The architect of evolutionary information encoded in protein sequences is extremely complicated due to the cooperative interactions among amino acids. These interactions could be pairwise interactions which are local or short-ranged, but could also involve other complex cooperative interactions in which residues are coupled in three-body or higher order ways. Most studies on evolutionary information focus on the single-body and two-body problems in the sequences, in terms of conservation and coevolution. The concepts of conservation and coevolution are two fundamental ideas behind protein evolution. Functional and structural restraints in evolution are expected to limit the amino acid substitution rate, resulting in similar amino acid composition at certain sites across homologous proteins [44]. Those conserved sites are usually involved in enzyme activities, ligand binding, protein protein interactions, or are buried in the cores [55]. Conserved proteins share similar or identical sequences across species. Highly conserved sequences are often crucial for the fundamental cellular function, stability and reproduction. While conservation highly determines the amino acid composition at a single site, coevolution affects the

pairwise correlation between two sites. Protein coevolution can occur on both intra- and intermolecular levels [163]. On the intramolecular level, once a residue changes, to stabilize a specific protein conformation, this mutation can be compensated by the change of a complementary residue; whereas on the intermolecular level such covelution usually occurs at two sites belongs to different proteins on the interface of the protein-protein interaction, in order to maintain high affinity and specificity.

The conservation and coevolution analysis routinely starts with multiple sequence alignment (MSA) which aligns homologous residues in columns among a set of query sequences sharing evolutionary linkages . The aligned residues in a column usually locate at a similar position in the structure and are descended from a common ancestral residue. A scoring metric is often used in the construction of MSA. The ideal MSA is corresponding to the global optimal of the scoring metric. Such scoring metric typically consist of two terms [78]. One term is to penalize the gaps (insertions and deletions) in the alignment. Since it is easier to extend a gap than to open one, the penalty for gap-open is often higher than that for gap-extension. The other term in the scoring metric is to score the aligned residues (substitutions). Most alignment approaches of proteins assume that the columns are statistically independent and thus the total substitution score over all columns can be written as the sum of the substitution score of each column separately. The conservation analysis usually takes two steps: estimate the amino acid frequencies at each position and calculate the conservation score for each position from those position-specific amino acid frequencies [190]. To correct for uneven and limited sampling in the space of natural sequences for the given alignment, different weights may be assigned to the sequences in the evaluation of amino acid frequencies.

Compared with the sequences with higher similarity to others (for example, sequences from closely related species), the divergent sequences are expected to impact more on amino acid frequencies and thus deserve a larger weight [190]. The coevolution between any pairwise sites depends on the joint probability of pairwise sites. It is related to the possibility of mutation at one site given the mutation of another site. In most cases the conservation analysis is required before the coevolution analysis, as the coevolution score has to be normalized based on the conservation score at related sites. Various algorithms have been proposed to perform coevoluton analysis, such as *correlation-based function* [102], *mutual information* (MI) [101], *statistically coupling analysis* (SCA) [162], *explicit likelihood of subset co-variation* (ELSC) [64], *Quartets and observed minus expected squared* (OMES) [150].

The implementation of conservation and coevolution analysis has enabled the prediction of critical sites, pathways and functional units (sectors) in proteins [162, 108]. Moreover, some artificial sequences designed in Ranganathan lab containing the conservation and coevolution information as in natural sequences turns out not only folded but also functional, while all the sequences lacking those information can not fold at all [227]. It suggests those two types of evolutionary information are necessary, and sometimes sufficient, for specifying a protein fold. However, the coevolution analysis using SCA for different families of proteins showed that [162] the map of amino acid interactions does not look like the contact map of native structures. Many direct packing residues are not coupled, while some distant sites linked through connected pathways with interacting residues bearing a high coevolution score. Therefore, the coevolution result of SCA cannot assist in structure prediction by providing accurate information about the residue pairs close in space.

A better inference method of residue coupling has been developed by Marks *et al* [167, 168] with high accuracy helping with structure prediction. It argues the coevolution is influenced by not only the structural restraints, but also phylogenetic restraints or statistical noises due to limited sampling. To reduce the influence of those confounding factors, they proposed an algorithm called *mean field direct coupling analysis*. This algorithm computes the amino acid coevolution in the same way as the MI approach but improves the calculation of the joint frequency at pairwise sites. Rather than focusing only on the two pairwise sites, it first seeks the global model $P(x_1, x_2, \cdots x_L)$ (x_i stands for the residue type x at site i) for the probability of occurrence of a particular amino acid sequence which maximizes the information entropy. Once the optimized $P(x_1, x_2, \cdots x_L)$ is obtained, the joint probability $P(x_i, x_j)$ of site i and j is used to replace the joint frequency in the MI equation and calculate the amino acid correlation termed as *Direct Information* (DI). The critical improvement of DI is that it optimizes the amino acid frequencies at all positions over the set of all possible sequences globally, while MI is a local statistical model where the joint frequency only depends on the amino acid composition at local positions i and j. The results showed that DI is an excellent predictor of residue proximity in native structures, much better than other methods of coevolution analysis [168]. More importantly, DI provides another way to predict protein structures from the sequences. Given the fact that the top-scoring DI residue pairs are mostly native contacts, they can be used as restraint for structure prediction. Using the distance constraints constructed from DI, they implemented *distance geometry algorithm*, a well-established method in the experimental structure determination by NMR, and successfully predicted the structures of 15 proteins with 48 to 258 residues and

43 transmembrane proteins with 217 to 961 residues [167, 117]. In Chapter 4, I perform ZAM simulation to understand how the evolutionary information helps specify a protein fold, and utilize the knowledge from ZAM and evolutionary information to develop a new method for protein design.

While protein structure prediction is to search the 3-D structure with the lowest energy for a fixed sequence, protein design is conceived as the inverse of protein-folding problem, to generate sequences adequate to a target structure or a specific function. Protein design is a rigorous test of our understanding of the principles underlying protein folding and can lead to the invention of novel drugs or enzymes. A common approach of computational protein design (CPD) consists of the search for optimal amino acid side chains given a fixed backbone topology (fold) [58, 214], similar to the folding recognition method used in the structure prediction for unkown sequence. Various types of amino acids are sampled at each position and the side chains are allowed to move within a set of low-energy confirmations called rotamers. A scoring function is used to evaluate the energy of each sequence with the optimal side chain configuration. Usually starting from a random sequence, either a stochastic algorithm like Monte Carlo search or deterministic method like dead end elimination is used to sample a large amount of possible amino acid sequences [214]. The amino acid sequences with low-energy configuration are expected to stabilize the given fold and thereby facilitate the desired activity. Using this strategy, Dahiyat and Mayo designed FSD-1 based on the backbone structure of a zinc finger domain [58]. A number of similar work followed. Using the fold of WW domain as scaffold, Kreamer-Pecore *et al* designed two artificial sequences, both of which adopted secondary structures of WW domain [143]. In particular, Kuhlman and Baker applied this approach to a protein topology

not observed in nature and yielded a stable 93-residue α/β protein called Top7 [147]. However, the designed proteins may lose coopertivity or be weakly cooperative, and the designed sequences with the best scores were not necessary to become a stable protein in experiment [58, 147]. Extra cares must be taken to ensure the cooperativity and stability of designed protein at desired pH and temperature condition [214]. In Chapter 4, I will demonstrate a novel approach of protein design based on the library of artificial sequences designed in Ranganathan lab [227]. In this approach, I perform ZAM simulation of those artificial sequences and identified a few critical contacts to folding. New proteins are then computationally designed by introducing mutations to the residues related to the critical contacts.

As the sequence of a protein dictates its structure, the structure of a protein is crucial for its function [191]. The most fundamental function of protein is binding [191]. The close relationship between protein structure and function is manifested in the *lock-and-key* hypothesis and *induced fit* model of protein binding, where the protein structure is fixed or locally adjusted when ligand binds [127]. However, protein is not static in cellular environment. After self-assembly to 3-D structures, they exhibit inherent structure-encoded dynamics that involves motions at different levels, from local atomic fluctuations and side-chain rotations to collective domain motions [91, 16]. This conformational dynamics in the unbound state may involve the transient motions toward the bound and functional conformations [178]. Moreover, rather than a single structure, protein exists as an ensemble of structures in the native equilibrium. Thus a emerging view about protein binding states that proteins in the unbound states include both bound and unbound configurations, and their population shifts from the unbound to the bound form upon binding

[56]. This is known as *population shift* mechanism. It highlights the importance of the protein dynamics to the function and also leads to a novel view about how protein function evolves: the divergence of new functions can occur within existing fold, since proteins are conformationally dynamic and functional promiscuous [240, 135]. In fact, the evolution of many modern proteins originates from a handful of common ancestors over millions to billions of years ago. It is believed that the evolutionary process starts from a limited repertoire of sequences which means limited structures, but ends up with large functional divergence nowadays [127]. Therefore the acquisition of new functions through modification of conformational dynamics within the existing fold may be a common way of function evolution [239, 240, 127]. Indeed, a small local structual modification such as a single mutation can cause a large change in conformational dynamics, even at quite distant residues due to allosteric regulation [130, 160, 18]. Furthermore, recent experimental and computational research have demonstrated that evolution may access new function or adapt to new environment through altering conformational dynamics [125, 98]. There is a challenge that the ancestral proteins no longer exist. However, the advances in phylogenetic analysis methods and the development of genomic databases have made it possible to reconstruct the ancestral sequences. A handful of ancestral proteins, including opsins [263, 264], GFP-like proteins [88, 243], steroid receptors[35, 186], β-lactamases [206] and others [94, 95, 128, 145, 237], have been synthesized in the laboratory. It opened a door for the characterization of their structures, stabilities, conformational dynamics and biochemical functions. Given the structure of ancestral β-lactamases and GFP-like proteins, I will probe their conformational dynamics using both MD simulations

and coarse-grained approaches, and discuss its relationship to functional divergence in Chapter 5.

In a short word, this thesis aims to utilize multiscale computational approaches to study protein structures, protein dynamics and the underlying folding mechanism. In Chapter 2, a brief introduction to the computational techniques is presented. Chapter 3 discusses the underlying folding mechanism indicated in the topology of native structures and also draw inference about biological properties of cells based on the prediction power of topological properties. A novel approach for protein design, using ZAM simulation in conjunction with a classification model, is addressed in Chapter 4. Chapter 5 provides insights about the roles of structural dynamics on protein evolution, by investigating β-lactamases and GFP-like proteins using both coarse-grained techniques and all atom MD simulation.

Chapter 2

COMPUTATIONAL METHODS FOR PROTEIN FOLDING STUDY

2.1 Introduction

Computational simulation is a powerful tool for solving scientific problems by performing numerical experiment on computers. It can reproduce the result of the bench experiment and elucidate how complex system works without synthesizing it. Moreover, it can capture the details of the system which may be invisible in bench experiment due to technique limitations. Besides, it can be used as a predictive tool which aids the experiment design and saves the cost. The most two popular approaches of computational simulation are molecular dynamics and Monte Carlo.

Molecular dynamics (MD) simulates the "real" motions of a set of particles in a system obeying classical mechanics over time. Under ergodicity hypothesis, the ensemble average is equivalent to the long-time average for a equilibrated system, thus one can estimate thermodynamic and conformational propteries of the system by taking the time average in the MD trajectory when the simulation is long enough to reach the equilibrium and properly sample the phase space.

To start a MD simulation, one first established an initial configuration of the system, in which the positions could be obtained from theoretical model or experimental results and the velocities are randomly generated from the Maxwell-Boltzmann distribution at the temperate of interest. Then the positions $\mathbf{r}$ and velocities $\mathbf{v}$ of particles are updated step by step based on Newton's

equations of motion. For a system with continuous force field (potential), the force acting on each particle is calculated by differentiating the force filed function $U(\mathbf{r})$. Although the force is changing over time due to the movement of particles, it is reasonable to be treated as constant during a small time period δt. Under this assumption, the integration of the derivative equations of motion can be approximated by a set of finite difference equations called integrator. A variety integrators have been developed, such as Verlet [246], leap-frog [114], velocity Verlet [234] and Beeman [28]. Using the integrator one derives the new configuration of the system at $t + \delta t$ from the the configuration at t. Thus MD is a deterministic approach where the future state of the system is predictable from the current state. The choice of time step δt depends on the characteristic time scale of motions present in the system. Only a small portion of the phase space can be sampled in a limited time if the time step is too small, while a large time step may cause instability of the system due to large errors of integration. Thus to maximize the efficiency of MD simulation and maintain stable dynamics, one usually chooses the time step which is approximately about one order of magnitude smaller than the highest frequency motion in the system [151]. As MD models the system on the classical level, it is not applicable for low temperatures at which quantum effects are more pronounced. Besides, if one is interested in detailed event, such as chemical reactions and chemical bonding of metal ions, quantum dynamics simulation is suitable because electronic motions and quantum effects are generally ignored in MD.

In contrast to MD, Monte Carto (MC) simulation is a stochastic approach which samples the system based on known probability distribution. Its another difference from MD is that there is no velocity component in MC. The new con-

figuration is generated by making a random change to the existing configuration iteratively. The simple MC simulation generates a large amount of configurations in the phase space with equal probablity and weight the contribution of each state to the ensemble average by a factor of $\exp(-U(\mathbf{r})/k_BT)$. However, this approach is inefficient since the high-enery states are energetically unfavorable in realty and their contributions to the ensemble average are extremely small. Thus a better way is to focus on sampling the states with high contributions, which is the essence of the Metropolis MC method. The Metropolis algorithm generates the configurations with a probablity $p = \exp(-U(\mathbf{r})/k_BT)$ and weight their contributions equally. This is achieved by using the acceptance ratio $\exp(-\Delta U(\mathbf{r})/k_BT)$ (ΔU is the energy of new configuration subtracted by that of its predecessor) to decide whether to retain or reject the new configuration. If the acceptance ratio is no less than 1 (the new configuration is not energetically unfavorable than its predecessor), the new configuration is accepted automatically. Otherwise, the acceptance ratio is compared with a random number from 0 to 1. The new configuration is accepted if it is larger than the random number, and rejected if not. This criterion ensures that the new configuration is accepted with probablity $\exp(-\Delta U(\mathbf{r})/k_BT)$. Once a new configuration is accepted, it is used as the starting point for the next iteration. Compared with MD, MC simulation is usually computational efficient since it does not require expensive numerical integration. It also often reach convergence rapidly especially for simple rigid molecules. However, for large flexible biomolecules such as proteins or DNAs, MD is more appropriate since the correct modeling of local motions requires small steps. Moreover, MD provides the kinetic and dynamic information of the system, which are omitted in MC.

2.2 Classical Molecular Dynamics Simulation

Classical molecular dynamics simulation is widely used to study the structure, dynamics and function of molecules. The improvement of computer hardwares, as well as the development of high-performing scalable MD engines like AMBER (Assisted Model Building and Energy Refinement) [212], CHARMM (Chemistry at HARvard Macromolecular Mechanics) [37], NAMD (NAno scale Molecular Dynamics) [192] and GROMACS (GROningen MAchine for Chemical Simulations) [197], have enables us to obtain longer simulation trajectories for biomolecules in a short period of time. The quantity of a MD simulation mostly depends on the accuracy of force field and the sufficiency of sampling. The all-atom empirical force field for biomolecular simulations typically consists of bonds, angles, torsional terms and non-bonded interactions, formed as

$$U = \sum_{bonds} K_r(r - r_{eq})^2 + \sum_{angles} K_\theta(\theta - \theta_{eq})^2 + \sum_{dihedrals} \frac{V_n}{2}[1 + cos(n\phi - \delta_n)] + \sum_{i<j} \sum_{j>i} \left\{ \epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right\} \tag{2.1}$$

The first three terms are bonded interactions where the bond stretching and bending are modeled in a harmonic expression and the dihedral energies are represented in a Flourier expansion function. The last term accounts for the non-bonded interactions where the van der Waals forces are modeled by the Lennard-Jones potential and the electrostatic interactions are treated by Coulomb potential for point charges. However, some atoms of the same molecule may get too close in space, leading the large values of non-bonded interactions. Therefore special treatment is required to prevent this. The

non-bonded interactions of atoms that bonded to each other (1-2 interactions) or atoms that bonded to the same atom (1-3 interactions) are usually neglected, and the non-bonded interactions of atoms at the end of a dihedral angle (1-4 interactions) are reduced by a factor. The parameters in the force field are derived from experiments and ab initio QM simulation. Although the general force filed function is similar, the parameterization varies a lot in different models. The most popular families of all-atom empirical force fields are AMBER[212], CHARMM[37] and OPLS (Optimized Potentials for Liquid Simulations)[129]. Unless specified, the AMBER 96 force field [189] with OBC GBSA model [184] is chosen in my studies, since it used to predict the structures of a set of single domain proteins with better accuracy than other combinations of AMBER forcefields with GBSA models in the previous studies [222, 223, 188].

Among the MD engines mentioned above, both AMBER and CHARMM implement the leap-frog algorithm as the default integrator (NAMD uses the verlocity Verlet integrator and GROMACS requires personal designation). The leap-frog integrator computes positions and velocities by [114]

$$\mathbf{r}(t+\delta t) = \mathbf{r}(t) + \mathbf{v}(t+\frac{1}{2}\delta t)\delta t \tag{2.2}$$

$$\mathbf{r}(t+\frac{1}{2}\delta t) = \mathbf{v}(t-\frac{1}{2}\delta t) + \mathbf{a}(t)\delta t \tag{2.3}$$

where $\mathbf{a}$ is the acceleration (the second derivative of potential). The highest-frequency motion in the MD simulation of the biomolecular system is usually bond stretching and one of the fastest bond stretching takes place in C-H bond. The C-H bond vibrates at the period of 10 femtoseconds (fs). In order to capture such motion with sufficient data points, the time step of MD simulation is generally selected to be 1 fs. However, this type of high-frequency

motions is usually not the interest of our study and a larger time step could be employed by freezing them. The freezing is accomplished using constraints, which prevent the related coordinates moving away from the equilibrium values. Mathematically, the constraints are introduced to the system through Lagrange multipliers, which is the essence of the two common algorithms for constraint MD named SHAKE [211, 238] and RATTLE [7]. The SHAKE algorithm and its variants were developed for the Verlet and leap-frog integrators, whereas the RATTLE algorithm was applicable for the velocity Verlet integrator [151]. In the simulations present in this thesis, the constraints on bond stretching are applied to all the bonds involving hydrogens and thus a larger time step of 2 fs is allowed.

The classical Newton's equations of motion preserve the total energy of the system as there is no damping term of friction. The resulting MD simulation is carried out in the microcanonical (NVE) ensemble, where the temperature and pressure of the system may vary significantly throughout the simulation. However, most experimental measurements are performed under constant temperature and/or pressure, thus it is desirable to modify the MD scheme and obtain the thermodynamic properties of the system in similar ensembles, such as the canonical (NVT) and isothermal-isobaric (NPT) ensembles.

The modified MD scheme to generate the ensemble of the system at constant temperature is called thermostat. The temperature of the system is associated with the kinetic energy, thus the simplest way to control the temperature is rescaling the velocities each time step by a factor of

$$\lambda = \sqrt{\frac{T_0}{T(t)}} \tag{2.4}$$

where T_0 is the desired temperature [260]. However, the fluctuation of temperature and kinetic energy becomes zero throughout the course of the simulation. To fix this problem, Berendsen thermostat couples the system with an external heat bath at the desired temperate and allows and exchange of heat between them at the rate of

$$\dot{T}(t) = \frac{T_0 - T(t)}{\tau} \tag{2.5}$$

where τ is coupling parameter which determines how tightly the bath and the system are coupled together [28]. The resulting scaling factor for velocities becomes

$$\lambda^2 = 1 + \frac{\delta t}{\tau}\left(\frac{T_0}{T} - 1\right) \tag{2.6}$$

Although Berendsen thermostat allows temperature fluctuation, it still cannot be mapped onto the rigorous canonical ensemble and capture the correct thermodynamic properties [174]. Instead the rigorous canonical ensemble can be generated by Anderson thermostat [6] and Nosé-Hoover thermostat [177, 116]. Anderson thermostat couples the system with the heat bath by imposing stochastic collision on randomly select particles. The probablity of collision in time interval $[t, t + \delta t]$ is given by $p(\nu, t)\delta t$ where $p(\nu, t)$ is in the Poisson form

$$p(t) = \nu \exp(-\nu t) \tag{2.7}$$

where ν is the frequency of stochastic collisions. At each time step, the velocities of those selected particles are reassigned according to the Boltzmann-Maxwell distribution at the desired temperature. As the presence of stochastic collisions, the kinetics of the system in the Anderson becomes unphysical and thus it is not suitable for obtain kinetic properties. Alternatively, Nosé-Hoover thermostat treats the heat bath as an additional degree of freedom s with ki-

netic energy $\frac{Q}{2}\dot{s}^2$ and potential energy $(3N+1)k_BT_0$. Q is the effective mass of s that controls the coupling and N is the total number of particles in the original system. The implement of Nosé-Hoover thermostat properly produces the canonical ensemble of the original system, whereas in fact the extended system (with the additional degree of freedom for heat bath) is mapped onto the microcanonical ensemble.

The modified MD scheme to maintain constant pressure of the system is call barostat. According to Clausius virial theorem, the pressure of the systems is calculated by

$$P = \frac{1}{\mathrm{V}}\left(Nk_BT + \frac{1}{3}\sum_i\sum_{j>i}\mathbf{r}_{ij}\cdot\nabla\mathbf{U}_{ij}\right) \tag{2.8}$$

From the thermodynamic point of view, the system can maintain constant pressure by exchanging its volume with the surroundings. Thus the control of pressure can be achieved through the change of volume, i.e., scaling the position of particles in the system. Most barostats are isotropic. They apply the same scaling factor for x, y and z directions and thus preserve the shape of system. The ideas of barosats are very similar to those of thermostats. For example, Berendsen barostat couples the system to an "pressure bath" at desired pressure and scales the position of particles [30]. Anderson barostat shares the similar approach with Nosé-Hoover thermostat rather than Anderson thermostat. It adds an extra degree of freedom to the original system, which can be considered as a piston [6].

2.3 Molecular Dynamics Simulation with Enhanced Sampling

Molecular dynamics can be used to search for conformations with lower energies. But it often fails in the search for global minimum, since the system

may be trapped in the local minima due to high energy barriers. Thus, in order to overcome high energy barriers and to reach the global minimum, new algorithms are proposed to enhance sampling. One way of enhanced sampling is to modify the potential surface, by adding a biased potential which in turn reduces (raises) the saddle points (wells) of the energy landscape, or guide the system toward certain targeted states. The selection of the biased potential is usually challenging as the energy barriers are generally unknown. In targeted MD (TMD), an artificial restraint potential is added through a Lagrange multiplier with a parameter λ ranging from 0 (initial state) to 1 (target state), to guide the system from the initial state toward the target state gradually in small steps [216]. Another popular protocol to modify the potential surface is umbrella sampling, which breaks down the reaction pathway into multiple non-interacting stages and introduces an extra potential in a quadratic form to enhance sampling in the regions of interest [241]. Then the actual free energy landscape can be recovered from the those stages in a self-consistency manner using weighting histogram analysis method (WHAM) [148]. The other way of enhanced sampling is to control the temperature of the system, which is the essence of global minimum hunting protocols like simulated annealing [139] and replica exchange (parallel tempering) [79]. The two protocols are commonly used in both MC and MD simulations as discussed below.

The simulated annealing protocol is based on an analogy with annealing in material science, where a metal is heated above the melting point and then cooled down under control to solidify into a defect-free crystalline structure or the global minimum state. It initially starts from a high temperature (or infinity) and gradually reduces the temperature as the simulation proceeds until T reaches zero or a targeted temperature. At each temperature stage,

the system is equilibrated following the general MC or MD procedures. In this way, the system is expected to explore a larger phase space at higher temperature more freely, cross energy barriers and then search for low-energy states in a narrowed regions at lower temperature. Theoretically, although not practically, simulated annealing is guaranteed to converge to the global minimum solution if the cooling is sufficiently slow and the equilibration is long enough.

The replica exchange protocol is initially developed for MC simulations and later introduced to MD simulations (REMD) by Sugita and Okamoto [233]. Replica exchange makes N replicas of the system in the canonical ensemble. Those replicas have identical force field and MD scheme but different temperatures. They evolve independently at their own temperatures and attempt to exchange with adjacent replicas after a certain time interval.

Let $\mathbf{x}_n^i$ stand for a state of the replica n at the temperature T_i where $\mathbf{x}_n^i = (\mathbf{r}^i, \mathbf{v}^i)_n$ for MD simulations, and $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N)$ represent a generalized state in the replica exchange ensemble consisting of all replicas. The weight factor of the state $\mathbf{x}_n^i$ in a single replica is given by the Boltzmann factor:

$$p_n^i = \exp\left(-\beta_i H(\mathbf{x}_n^i)\right) \tag{2.9}$$

where $\beta_i = 1/k_B T_i$ and $H(\mathbf{x}_n^i)$ is the Hamiltonian of the system which equals the sum of kinetic energy $K((v_n^i)$ and potential energy $U(\mathbf{x}_n^i)$. Since all the replicas are independent, the weight factor of the generalized state in the replica exchange ensemble equals the products of Boltzmann factors for each replica:

$$P(\mathbf{X}) = \exp\left(-\sum_{n=1}^{N} \beta_i H(\mathbf{x}_n^i)\right) \tag{2.10}$$

Suppose that the replica n at temperature T_i is attempted to exchange with the replica m at temperature T_j:

$$\mathbf{X} = (\cdots, \mathbf{x}_n^i, \mathbf{x}_m^j, \cdots) \to \mathbf{X}' = (\cdots, \mathbf{x}_n^{j\,\prime}, \mathbf{x}_m^{i\,\prime}, \cdots) \tag{2.11}$$

In MD simulations, one must consider both the positions and velocities of particles in the system and thus

$$\mathbf{x}_n^i = (\mathbf{r}_n^i, \mathbf{v}_n^i) \to \mathbf{x}_n^{j\,\prime} = (\mathbf{r}_n^{j\,\prime}, \mathbf{v}_n^{j\,\prime}) = (\mathbf{r}_n^i, \sqrt{\frac{\beta_i}{\beta_j}}\mathbf{v}_n^i) \tag{2.12}$$

Here the velocity is rescaled according to the temperature.

The acceptance ratio (exchange probablity) $P_{ex}(\mathbf{X} \to \mathbf{X}')$ must satisfy the detailed balance condition

$$P(\mathbf{X})P_{ex}(\mathbf{X} \to \mathbf{X}') = P(\mathbf{X}')P_{ex}(\mathbf{X}' \to \mathbf{X}) \tag{2.13}$$

which gives

$$\begin{aligned}
\frac{P_{ex}(\mathbf{X} \to \mathbf{X}')}{P_{ex}(\mathbf{X}' \to \mathbf{X})} =& \frac{P(\mathbf{X}')}{P(\mathbf{X})} \\
=& \exp\Bigg[\beta_i H(\mathbf{r}_n^i, \mathbf{v}_n^i) + \beta_j H(\mathbf{r}_m^j, \mathbf{v}_m^j) - \beta_j H(\mathbf{r}_n^i, \sqrt{\frac{\beta_i}{\beta_j}}\mathbf{v}_n^i) \\
& -\beta_i H(\mathbf{r}_m^j, \sqrt{\frac{\beta_j}{\beta_i}}\mathbf{v}_m^i)\Bigg] \\
=& \exp\left[\beta_i U(\mathbf{r}_n^i) + \beta_j U(\mathbf{r}_m^j) - \beta_j U(\mathbf{r}_n^i) - \beta_i U(\mathbf{r}_m^j)\right] \\
=& \exp\left[(\beta_i - \beta_j)(U(\mathbf{r}_n^i) - U(\mathbf{r}_m^j))\right]
\end{aligned} \tag{2.14}$$

It is noticeable that the acceptance ratio depends on the change of potential energy, not the total energy, as the rescaling of velocity cancels the terms of kinetic energy. A common choice of $P_{ex}(\mathbf{X} \to \mathbf{X}')$ to fulfill the requirement is the Metropolis criterion:

$$P_{ex}(\mathbf{X} \to \mathbf{X}') = \min\left\{1, \exp\left[(\beta_i - \beta_j)(U(\mathbf{r}_n^i) - U(\mathbf{r}_m^j))\right]\right\} \tag{2.15}$$

Replicas are allowed to attempt exchange with adjacent replicas periodically according to this probability. The low-energy states are expected to climb down the temperature ladder with high probablities.

For optimal performance, the following three requirement should be satisfied: i) the highest temperature must be high enough to prevent the replica from trapping at local minimum. ii) the acceptance ratio must be uniform (i.e., independent of temperature) between all pairs of adjacent replicas to ensure that each replica spends the same amount of time on average at different temperature [142, 233]. iii) the acceptance ratio should be adequate. If too high, the lower-temperature replica is not very different from the higher-temperature one; if too low, only very few exchange is successful which leads to little gain of the low-temperature replica. Since the acceptance ratio depends on the overlap of the sampled states in the adjacent replicas, from the entropy point of view, Kofke postulated that its average value is related with the entropy difference between the high-temperature and low-temperature replicas as [142]

$$\overline{P_{ex}} \sim \exp\left(-\Delta S/k\right) \tag{2.16}$$

If the temperature spacing between the two replicas is ΔT and the heat capacity is constant in this temperature interval, then the entropy difference can be written as

$$\Delta S = C_v \Delta T/T \tag{2.17}$$

and the average acceptance ratio becomes

$$\begin{aligned}\overline{P_{ex}} &\sim \exp\left(-\frac{C_v}{k}\cdot\frac{\Delta T}{T}\right)\\ &= \exp\left[\frac{C_v}{k}\log\left(1-\frac{\Delta T}{T}\right)\right]\\ &\approx \left(1-\frac{\Delta T}{T}\right)^{\frac{C_v}{k}}\\ &= \left(\frac{T-\Delta T}{T}\right)^{\frac{C_v}{k}}\end{aligned} \tag{2.18}$$

It clearly shows that the average acceptance ratio depends on the heat capacity of the system and the temperature ratio of the adjacent replicas. Thus for a system with constant heat capacity across temperatures, one can achieve uniform acceptance ratio by using a geometric progression of temperature ladder, i.e., $(T-\Delta T)/T = const$. For complicated systems, more elaborate schemes have been developed to maintain a target acceptance ratio by altering the system temperature iteratively [202, 218]. In particular, assuming that the density of states in each replica is normal distributed with mean $E(T)$ and variance $\sigma^2(T)$, Rathore *et al* showed that the acceptance ratio is governed by $\Delta E/\sigma_m$ and their functional dependence is not system-dependent [202]. Here $\Delta E = E(T_i) - E(T_{i-1})$ and $\sigma_m = \left[\sigma(T_i) + \sigma(T_{i-1})\right]/2$. Thus to reach the target acceptance ratio, one can allocate the temperature to obtain the corresponding $\Delta E/\sigma_m$ value. Practically, one can obtain the mean $E(T)$ and variance $\sigma^2(T)$ from a short simulation run with a few replicas. Then by fixing the lowest temperature, the following temperatures can be calculated by iteratively solving [202]

$$\left.\frac{\Delta E}{\sigma_m}\right|_{T_i} = \left.\frac{\Delta E}{\sigma}\right|_{target} \tag{2.19}$$

Based the deviation of thermodynamic properties obtained from the simulations with different acceptance values, Rathore *et al* also concluded that the

optimal acceptance ratio is 20% [202]. Interestingly, this empirical result is quite consistent with the theoretical work conducted by Kofke, which suggest the optimal acceptance ratio is 23% under the assumption of constant heat capacity of the simulated system [202].

Generally speaking, in canonical ensemble, the mean energy difference between two replicas and the variance are $\Delta E = C_v \Delta T$ and $\sigma^2 = C_v k_B T^2$ respectively. In order to achieve adequate acceptance ratio, certain overlap of the sampled states in the two replicas is necessary, which requires ΔE is comparable to σ or $\Delta E/\sigma \approx Const$. Thus $\Delta T \approx Const\sqrt{k_B T^2/C_v}$. As C_v is proportionally to the system size N, the temperature spacing is proportional to $1/\sqrt{N}$ and the number of replicas increases in the order of $\sqrt{N}$[93]. Therefore, the reduction of the system size for replica exchange is a way to bring down the computational cost. Many variants of replica exchange method have been developed based on this idea by exchanging part of the system. Liu *et al* developed replica exchange with solute tempering which avoids the evaluation of solvent-solvent interaction in the attempts of exchange Replica exchange with solute tempering: A method for sampling biological systems in explicit water. Okur *et al* proposed a hybrid solvent approach in which the simulations of each replica are performed with fully explicit solvent but the solvent molecules beyond the first solvation shell are replaced with a continuum model during the evaluation of acceptance ratio [180]. Cheng *et al* derived partial replica exchange molecular dynamics (PREMD) and local replica exchange molecular dynamics (LREMD) where the system is divided into the subsystem of interest and the remainder of the system [48]. Assuming weak coupling between them, PREMD generates replicas where the temperature of the subsystem of interest varies but that of the remainder remains the same in all

replicas. This approach reduces the energy difference between replicas and requires fewer replicas. For cases where the coupling is even weaker, LREMD creates replicas in which only the subsystem of interact is replicated and the remainder of the system share the identical coordinates in a mean-field manner. The exchange carries out only on the subsystem of interact and thus the actual system size in LREMD is significantly smaller than PREMD or REMD.

Another way to increase the efficiency of replica exchange method is to incorporate models at different levels in different replicas. Lwin and Luo proposed a dual replica exchange setup, with one group of replicas using high-resolution model and the other group of replicas using low-resolution model [164]. Besides the temperature exchange is attempted among replicas from the same group, the resolution exchange is also allowed between the replicas from different groups at the same temperature. Thus it takes the advantage of both high efficiency of the low-resolution models and high accuracy of the high-resolution model. Okur *et al*[179] and Roitberg *et al* [208] developed a framework called reservoir REMD (r-REMD), which couples replica exchange simulations to a reservoir of structures. It uses the same parameters as traditional REMD, but also allows the exchange between the high-temperature replica and the reservior. The structures in the reservior can be prepared using other efficient sampling techniques, such as the geometric based sampling technique called FRODAN [86] discussed later in this chapter. The introduction of reservoir can significantly speed up the convergence of the simulation.

2.4 Coarse-grained Models in Molecular Dynamics

2.4.1 Langiven Dynamics

Most biochemical process takes place in solvent and the solvent affects the behavior of the system. To taking into account this effect, the solvent has to be modeled explicitly or implicitly in the MD simulation. The explicit solvation model describes the solvent molecules like water molecules and salt ions explicitly. It captures the right physics but is usually computationally expensive due to the enormous degrees of freedom from the solvent. Moreover, in some cases, the behavior of solvent molecules is not the interest of simulation. One may ignore the solvent degrees of freedom and treat the solvent as a continuum surrounding the solute using an implicit solvation model, which is discussed further in the subsequent subsection. Generally, the implicit solvent model can achieve high computational efficiency by omitting the solvent sampling and enhancing the solute sampling. As the solvent molecules are not explicitly present, the solute molecules no longer experience the viscosity caused by the solvent. The lack of viscosity may be desirable as it increases the solute sampling efficiency. However, if one is interested in the realistic dynamics of the solutes, the viscosity of the solvent must be considered. This can be accomplished via Langevin dynamics, which models the solvent effect in terms of a viscous damping force (dissipation) and a random force. The related Lagenvin equation is formed as

$$\frac{d^2\mathbf{r}}{dt^2} = \mathbf{a}(t) - \gamma\frac{d\mathbf{r}}{dt} + \frac{\delta\mathbf{R}(t)}{m} \tag{2.20}$$

where $\gamma\frac{d\mathbf{r}}{dt}$ is the damping force and $\delta\mathbf{R}(t)$ is the random force satisfying

$$\langle\delta\mathbf{R}(t)\rangle = 0 \tag{2.21}$$

and

$$\langle \delta\mathbf{R}(t)\delta\mathbf{R}(t')\rangle = 2\gamma m k_B T \tag{2.22}$$

Here the random force follows a Gaussian distribution with zero mean and variance of $2\gamma m k_B T$. Equation (2.22) also indicates that the dissipation and fluctuation are not independent. The damping constant γ determines both the magnitude of the damping force and the variance of the random force. For a spherical particle with radius a in a solvent with viscosity η,

$$\gamma = \frac{6\pi\eta a}{m} \tag{2.23}$$

according to Stoke's law. As γ increases, the influence of solvent becomes stronger and the system moves towards the diffusive (Brownian) regime. A small γ constant should be used if the main purpose is to control temperature. In Langevin dynamics simulation with implicit solvent the γ constant often ranges from 2 to 20 ps^{-1}. Lagenvin equation is stochastic where the stochastic component is introduced by the random force. It rigorously converges to the correct canonical ensemble. A popular solution to Langevin equation is known as Brünger-Brooks-Karplus (BBK) integrator [38]. The Verlet form of BBK is given by

$$\mathbf{r}(t+\delta t) = \mathbf{r}(t) + \frac{1-\gamma\delta t/2}{1+\gamma\delta t/2}\left(\mathbf{r}(t) - \mathbf{r}(t-\delta t)\right) + \frac{1}{1+\gamma\delta t/2}\delta t^2\left[\mathbf{a}(t) + \sqrt{\frac{2\gamma k_B T}{m}}Z(t)\right] \tag{2.24}$$

$$\mathbf{v}(t+\delta t) = \frac{\mathbf{r}(t+\delta t) - \mathbf{r}(t-\delta t)}{2\delta t} \tag{2.25}$$

where $Z(t)$ is a set of Gaussian random variables with zero mean and one variance. Both the AMBER and NAMD packages chose the BBK integrator or its minor variants for Langevin dynamics [189, 192]

2.4.2 Implicit Solvation

As mentioned in the previous section, the implicit solvation model treats the solvent as a continuum with its mean-field characteristics surrounding the solute. It significantly provides efficient sampling of solute and saves the computational time as the total degrees of freedom in the system is drastically decreased. Moreover, as the solvation free energy is evaluated at each step based on the coordinates of the solute, the solvent is actually considered to be equilibrated instantaneously to any change in the system. This process, i.e., the equilibration of solvent, is instead very slow in the explicit solvent, especially when large conformational change of solute occurs, like protein folding and unfolding. Additionally, since the solute molecules no longer experience the viscosity caused by the solvent, the solute sampling also gets more efficient. The effect of viscosity can be recovered if necessary using Langiven dynamics approach as discussed.

The key of implicit solvation model is to calculate the solvation free energy ΔG_{solv} accurately and efficiently. The solvation free energy is the energy required to transfer the solute from vacuum to the solvent. It is usually divided into three components as:

$$\Delta G_{solv} = \Delta G_{cav} + \Delta G_{vdw} + \Delta G_{elec} \tag{2.26}$$

The first two terms are nonpolar solvation energy required to immerse the uncharged solute into the solvent. Specifically, ΔG_{cav} is the work of creating the cavity inside the solvent where the solute resides. ΔG_{vdw} is van der Waals attraction energy between solute and solvent. The third term ΔG_{elec} is polar solvation energy, associated with the electrostatic energy of turning on the charges of the solute in the solvent.

Precisely, the polar solvation energy equals to the difference of work of charging the solute in the solvent and that of charing the solute in the vacuum, given by

$$\Delta G_{elec} = \frac{1}{2} \int \rho^f(\mathbf{r}) \left[\psi_{sol}(\mathbf{r}) - \psi_{vac}(\mathbf{r})\right] \mathrm{d}^3\mathbf{r} \tag{2.27}$$

$\rho^f(\mathbf{r})$ denotes the free charge of the solute at position $\mathbf{r}$. The electrostatic potential has to be computed twice, in the solvent and in vacuum, named as $\psi_{sol}(\mathbf{r})$ and $\psi_{vac}(\mathbf{r})$. The most accurate way to obtain the electrostatics potential is through Poisson-Boltzmann equation (PBE) [90]

$$\nabla \cdot [\epsilon(\mathbf{r}) \nabla \psi(\mathbf{r})] = -4\pi\rho^f(\mathbf{r}) - 4\pi \sum_i c_i^\infty q_i \exp\left[\frac{-q_i\psi(\mathbf{r})}{k_B T}\right] \tag{2.28}$$

Here $\epsilon(\mathbf{r})$ is the dielectric constant. The unitless values of c_i is about 78 to 80 in water solvent, 2 to 20 within the protein surface and 1 in vacuum. and q_i denotes the charge and bulk concentration of the ith ion species respectively. The summation term is attributed to ions in the solvent, which vanishes in the case of $\psi_{vac}(\mathbf{r})$. It assumes that the ions surrounding the solute molecule follow the Boltzmann distribution. When $\psi(\mathbf{r}) < k_B T$, this term can be expanded in Taylor series up to the first order. The zero order term vanishes under the assumption of electroneutrality of the ionic solution, so that the PBE become linearized as

$$\nabla \cdot [\epsilon(\mathbf{r}) \nabla \psi(\mathbf{r})] = -4\pi\rho^f(\mathbf{r}) + 8\pi I(\mathbf{r})\psi(\mathbf{r}) \tag{2.29}$$

with the ionic strength

$$I(\mathbf{r}) = \frac{1}{2} \sum_i c_i^\infty q_i^2 \tag{2.30}$$

The PBE is well known as a formidable problem with analytic solution only available for systems with a simple geometry. For complicated systems like proteins, they require numerical methods to search for the convergent solution of electrostatic potential solution in a iterative way. Various numerical

methods have been developed to find the solution of the nonlinear and linear PBE accurately and efficiently, such as boundary element [267], finite element [21], and finite difference [115] as well as their scalable versions for sovling PBE on parallel platforms [245]. Particularly, a solver named APBS (Adaptive Poisson-Boltzmann Solver) based on the finite difference method is able to solve both nonlinear and linear PBE in parallel, from which the electrostatics of large biomolecules can be obtained [22].

Those numerical solvers provide accurate electrostatic potential, resulting in precise polar solvation energy. The Poisson-Boltzmann (PB) approach is widely used when the electrostatic interactions are believed to be critical, including predicting acid dissociation constant [136], studying the role of the electrostatics in the ion channels [209], ligand binding [219] and cotranslational folding around ribosomal surface [141]. However, they usually requires too much computational resources to be implemented in MD simulations. Instead, a common approach in MD simulations computes the polar solvation energy through Generalized Born approximation. It provides an analytic approximation of the polor solvation energy, given by a pairwise sum of Columbic potential over all charges in the solute and a correction term of Born solvation energy in the following form [230]:

$$\Delta G_{elec} = \left(\frac{1}{\epsilon_w} - \frac{1}{\epsilon_p}\right) \sum_{i<j} \frac{Q_i Q_j}{r_{ij}} + \frac{1}{2} \left(\frac{1}{\epsilon_w} - \frac{1}{\epsilon_p}\right) \sum_i \frac{Q_i^2}{B_i} \tag{2.31}$$

Here ϵ_w and ϵ_p denote the dielectric constants of solvent and solute. Q_i represents the charges of atom i of the solute. B_i is the effective Born radius of atom i, which characterizes the deepness of the atom burial inside the solute. It can be interpreted as the average distance to the solvent accessible surface of the solute. Thus it is not only dependent on the size of atom i itself, but

also the size and distance of other atoms. The equation equation (2.31) is a generalization of Born equation, which gives the solvation free energy of a nonpolarizable sphere with a point charge Q at it center, radius a and internal dielectric constant ϵ_p, surrounded by a medium with dielectric constant ϵ_w:

$$\Delta G_{Born} = \frac{Q^2}{2a}\left(\frac{1}{\epsilon_w} - \frac{1}{\epsilon_p}\right) \tag{2.32}$$

The two terms in equation equation (2.31) can be consolidated and it becomes [230]

$$\Delta G_{elec} = \frac{1}{2}\left(\frac{1}{\epsilon_w} - \frac{1}{\epsilon_p}\right)\sum_{i,j}\frac{Q_i Q_j}{f_{GB}\left(r_{ij}\right)} \tag{2.33}$$

with the function $f_{GB}\left(r_{ij}\right)$ defined as

$$f_{GB}\left(r_{ij}\right) = \left[r_{ij}^2 + B_i B_j \exp\left(-\frac{r_{ij}^2}{4B_i B_j}\right)\right]^{1/2} \tag{2.34}$$

At large distance r_{ij}, f_{GB} is asmpytotically close to the pairwise distance r_{ij}, indicating that the atom size is negligible if the two atoms are further apart, while at short distance the Born radii becomes dominant and especially f_{GB} equals effective Born radius if $i = j$. Furthermore, the salt screening effect due to the presence of ions in the solvent can be also incorporated via Debye-Hück theory. Thus the polar solvation energy can be obtained using the following expression which has been widely implemented in MD simulations [229]:

$$\Delta G_{elec} = \frac{1}{2}\left\{\frac{\exp\left[-\kappa f_{GB}\left(r_{ij}\right)\right]}{\epsilon_w} - \frac{1}{\epsilon_p}\right\}\sum_{i,j}\frac{Q_i Q_j}{f_{GB}\left(r_{ij}\right)} \tag{2.35}$$

where κ denotes the Debye-Hück screening parameter.

The effective Born radii are critical parameters in the GB approximation and the accuracy of the GB approach are fully relied on the determination of the effective Born radii [185]. The exact values of B_i's can be obtained by setting the polar solvation energy of atom i calculated via the PB approach equal

to that calculated using equation (2.32) with $a = B_i$. However, this would still require the numerical solution of PBE and results in no computational advantage over the PB approach. Instead by comparing the energy calculated usingequation (2.32) with that obtained through *Coulomb-field approximation*, one can derive the following expression of B_i [183]:

$$B_i = \left[\frac{1}{\alpha_i} - \frac{1}{4\pi}\int_{solute, r>\alpha_i} \frac{1}{r^4} \mathrm{d}^3\mathbf{r}\right]^{-1} \tag{2.36}$$

Various methods has been developed for the evaluation of the integral in the above expression which leads to different GB models, including STILL [230], HCT [112], OBC [183], ACE [215], etc.

The other component of solvation is the nonpolar solvation. It is dominated by the first solvation shell and its energetics is approximately proportional to the number of solvent molecules in the first solvation shell. Thus, the most popular estimation of the nonploar solvation energy is a term proportional to the total solvent-accessible surface area (SA) of the solute molecule like

$$\Delta G_{cav} + \Delta G_{vdw} = \gamma A \tag{2.37}$$

where γ is empirically determined surface tension coefficient and A is the solvent-accessible area. Although the γ value varies significantly from 5-7 $\mathrm{cal \cdot mol^{-1} \cdot \AA^{-2}}$ to 40-70 $\mathrm{cal \cdot mol^{-1} \cdot \AA^{-2}}$, depending on the associated force field and the polar solvation model, most studies suggested that the small γ value about 5 $\mathrm{cal \cdot mol^{-1} \cdot \AA^{-2}}$ yields the optimal results for protein simulations [47]. More precisely, the two contributions to the nonpolar solvation, i.e., cavity creation and van der Waals attraction, can be evaluated separately. The nonpolar solvation energy becomes a sum of three terms as [250]

$$\Delta G_{cav} + \Delta G_{vdw} = \gamma A + pV + \sum_i U_{vdw}(i) \tag{2.38}$$

Here the working of cavity creation ΔG_{cav} is expressed as a linear function of the solvent-accessible surface area A and the solvent-accessible volume V (SV). The van der Waals attraction energy ΔG_{vdw} is a sum of $U_{vdw}(i)$ over all solute atom i's, where $U_{vdw}(i)$ denotes the van der Waals interaction energy between solute atom i with the solvent. In principle, $U_{vdw}(i)$ is equal to the volume integral of the van der Waals term over the solvent region. Assuming homogeneous solvent density, in practice, $U_{vdw}(i)$ can be efficiently obtained from another effective Born radius C_i associated with the volume integral of a $1/r^6$ function, rather than the $1/r^4$ function used to compute B_i in the GB model equation (2.36) [92]. If the atom is spherical, C_i becomes equivalent to B_i. However, the comparison of different approaches to evaluate nonploar solvation energy showed that the improvement given by the additional van der waals term is negligible [152]. The simple and traditional estimation using a linear function of SA and/or SV remains a wise choice considering efficiency and accuracy together. The SA estimation of nonpolar solvation energy is often used in together with the Generailized Born model of polar solvation. The combination is called GBSA model. Since the sovalation energy in GPSA model can be evaluated quickly and the energy funcation is analytically differentiable, this model has been widely adopted in MD simulations.

2.5 Geometry-based Simulation

Unlike MD simulation, the geometry-based simulation does not use any force field except geometric constraints, and thus can sample a large conformation space in a limited time. It is an efficient technique to sample the geometry-allowed conformation space for a given protein. Here I focus on a geometry-based technique for exploring all-atom pathways of protein from an

initial configuration to a target configuration, named as FRODAN (FRODA New) [86, 84], the newer version of FRODA (Framework Rigidity-Optimized Dynamics Algorithm) [257]. A short introduction of FRODAN is provided here following the work of Dr. Farewell [86, 84], which is suggested for further reading. First of all, the protein is decomposed into rigid units based on the covalent geometry. The assignment of rigid units is accomplished through the FIRST(Floppy Inclusions of Rigid Substructure Topology) software built on the Pebble Game algorithm [126]. Those rigid units become the only mobile entities in the system. A geometric constraint framework is constructed on the rigid units, in order to enforce various aspects of the structure in the allowable regions, including covalent bond length and angles, dihedral angles, hydrogen bonds and hydrophobic contacts. Once certain structural aspect moves into the disallowable region, a constraint energy penalty like a harmonic potential is applied to correct it. Thus the total constraint energy function of the system is consisted of a summation of quadratic functions, representing those geometric constraints.

To generate the pathway, FRODAN takes steps from the initial configuration toward the target configuration by gradually decreasing the root mean square deviation (RMSD) with respect to the target. For this purpose, a biased energy is added to the constraint energy function:

$$E_{rmsd} = \begin{cases} \frac{1}{2} k_{rmsd} \left(\text{RMSD} - C\right) & \text{RMSD} > C \\ 0 & \text{otherwise} \end{cases} \tag{2.39}$$

Here C is a controlling parameter which guides the pathway toward the target configuration. At each step, the parameter C decreases at a defined small step size δ (typically $\delta \leq 1\mathring{A}$) to insure the forwarding direction of the pathway. Next, the constraints energy function is minimized using conjugate gradient

algorithm to enforce both geometric and RMSD constraints. The configuration with minimum constraint energy is then subjective to the examination of non-overlaping constraints. If acceptable, it is aligned with the target configuration, providing the updated configuration and RMSD value for the next step. Such processes are iterated until RMSD value reach zero or a tolerance value. The series of acceptable configurations in together yields a geometric allowable pathway toward the target configuration. However, this resulting pathway is determinstic. Stochastic effect can also be incorporated through introducing random perturbation for both translational displacement and rotational motion before the enforcement of constraints. In certain cases, no acceptable configuration satisfying the constraints can be found at the step of energy minimization, as if the pathway is stopped by an obstacle. To tackle this problem, FRODAN utilizes a backtracking approach called "momentum steps" which allows the RMSD increases temporarily.

FRODAN has also been used to generated the unfolding pathways by targeting the N- and C-terminal residues to the pseudo residues placed on two sides of the protein with distance separation longer than the length of unfolded protein [61]. Hydrogen bond, salt bridge, and hydrophobic constraints are allowed to break in the pathway once the maximum load is exceeded. It can mimic the unfolding of proteins as shown in MD simulations and experiments.

As no actual force field present here except geometric constraints, FRODAN produces the pathways extremely faster than other simulation approach with force filed such as targeted MD [85]. Such advantage of FRODAN can benefit the sampling of configuration in other methods. For example, it has been used to assist the fragments assembly in protein structure prediction by treating those fragments as rigid units [100]. Recently, a structure refinement

method has been developed using the configurations provided in the unfolding trajectory as reservoir structures and the consensus contacts among those configuration as restraints in r-REMD, which leads to faster convergence of the simulation (unpublished: Avishek Kumar and Paul Campitelli).

2.6 Coarse-grained Network Models

2.6.1 Elastic Network Model

Elastic Network Model (ENM), also known as Anisotropic Network Model (ANM), is a simple tool to probe the large-scale cooperative vibrational motions of proteins around their equilibrium state [12]. The protein is dramatically simplified in the ENM where each residue is reduced to a single node, usually at the C_α atom and a harmonic spring is formed if the two residues are within a specified cutoff distance R_c. The dynamics of the network is governed by the potential in the form of [17]:

$$\begin{aligned} V &= \frac{\gamma}{2}\left[\sum_{i,j}^{N}\left(s_{ij}-s_{ij}^{0}\right)^{2} f\left(s_{ij}^{0}\right)\right] \\ &= \frac{\gamma}{2}\left\{\sum_{i,j}^{N}\left[\sqrt{(X_i-X_j)^2+(Y_i-Y_j)^2+(Z_i-Z_j)^2}-s_{ij}^{0}\right]^{2} f\left(s_{ij}^{0}\right)\right\} \end{aligned} \tag{2.40}$$

Here γ is the uniform sprint constant, N is the total number of residues in the protein, s_{ij} and s_{ij}^0 are the instantaneous and equilibrated distance between residues i and j, and X, Y, Z are the instantaneous Cartesian coordinates. The summation is carried out over residue pairs within the cutoff distance R_c through the Heaviside function $f\left(s_{ij}^0\right)$ with $f\left(s_{ij}^0\right) = -1$ if $s_{ij}^0 \leq R_c$ and 0 otherwise. The standard Normal Mode Analysis (NMA) is now performed to obtain the vibrational motions. The potential above is used to construct a $3N \times 3N$ Hessian matrix $\mathbf{H}$, which is the square matrix of second-order partial

derivatives respect to the coordinates of residues in the protein. The matrix can be viewed as a organization of $N \times N$ super-elements

$$\mathbf{H} = \begin{bmatrix} \mathbf{h}_{11} & \mathbf{h}_{12} & \cdots & \mathbf{h}_{1N} \\ \mathbf{h}_{21} & \mathbf{h}_{22} & \cdots & \mathbf{h}_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{h}_{N1} & \mathbf{h}_{N2} & \cdots & \mathbf{h}_{NN} \end{bmatrix} \tag{2.41}$$

in which the ijth super-element $\mathbf{h}_{ij}$ is of size 3×3 and defined as

$$\mathbf{h}_{ij} = \begin{bmatrix} \dfrac{\partial^2 V}{\partial X_i \partial X_j} & \dfrac{\partial^2 V}{\partial X_i \partial Y_j} & \dfrac{\partial^2 V}{\partial X_i \partial Z_j} \\ \dfrac{\partial^2 V}{\partial Y_i \partial X_j} & \dfrac{\partial^2 V}{\partial Y_i \partial Y_j} & \dfrac{\partial^2 V}{\partial Y_i \partial Z_j} \\ \dfrac{\partial^2 f}{\partial Z_i \partial X_j} & \dfrac{\partial^2 V}{\partial Z_i \partial Y_j} & \dfrac{\partial^2 V}{\partial Z_i \partial Z_j} \end{bmatrix} \tag{2.42}$$

Diagonalization of the Hessian matrix yields the $3N$ pairs of eigenvalues and eigenvectors (modes), which characterize the frequencies (eigenvalue equals to the square of frequency) and directions of the concerted motion of residues. Six of them are trivial with zero eigenvalues (and zero frequency) as they are related with the global translational and rotational motion of the protein. The rest $3N - 6$ modes are orthogonal and resonate independently which reflect the internal motion of the protein. In each mode, all the residues move on a straight line like harmonic oscillators at the same frequency and phase. Usually the low-frequency modes are of the most interest as they correspond to the functionally related motions observed in the experiment. Ignoring the 6 eigenvectors with zero eigenvalues, the psudo inverse of Hessian matrix can be formed as

$$\mathbf{H}^{-1} = \sum_{i=1}^{3N-6} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^T \tag{2.43}$$

where λ_i are the nonzero eigenvalues of $\mathbf{H}$ sorted in the ascending order and $\mathbf{u}_i$ are the corresponding eigenvectors. The mean square fluctuations of in-

dividual residues and the cross-correlations between different residues can be determined from $\mathbf{H}^{-1}$ [12, 49, 57]:

$$\begin{aligned}\left\langle \Delta R_i^2 \right\rangle =& \left\langle \Delta X_i^2 \right\rangle + \left\langle \Delta Y_i^2 \right\rangle + \left\langle \Delta Z_i^2 \right\rangle \\ =& \frac{k_B T}{\gamma} \left(\mathbf{H}^{-1}_{3i-2,3i-2} + \mathbf{H}^{-1}_{3i-1,3i-1} + \mathbf{H}^{-1}_{3i,3i} \right)\end{aligned} \tag{2.44}$$

and

$$\begin{aligned}\langle \Delta R_i \cdot \Delta R_j \rangle =& \langle \Delta X_i \Delta X_j \rangle + \langle \Delta Y_i \Delta Y_j \rangle + \langle \Delta Z_i \Delta Z_j \rangle \\ =& \frac{k_B T}{\gamma} \left(\mathbf{H}^{-1}_{3i-2,3j-2} + \mathbf{H}^{-1}_{3i-1,3j-1} + \mathbf{H}^{-1}_{3i,3j} \right)\end{aligned} \tag{2.45}$$

2.6.2 Perturbation Response Scanning

Perturbation Response Scanning is a tool to analyze the response of residues when the protein is perturbed around its equilibrium [13]. In PRS, the protein is reduced to the same elastic network as in ENM. PRS relies on sequentially applying externally random force (perturbation) on a single residue and record the linear responses (positional displacements) of the other residues. The first step is to find out how the positional displacement is related with external forces.

In ENM, as external forces are absent, each residue is under equilibiurm with balanced internal forces. The internal forces on any residues are summed up to zero in x, y and z directions. For a protein with N residues and M bonds, the equilibrium condition can be formed as [12]

$$[\mathbf{B}]_{3N \times M} [\mathbf{f}]_{M \times 1} = [\mathbf{0}]_{M \times 1} \tag{2.46}$$

Here $\mathbf{B}$ is the directional cosine matrix. $\mathbf{f}$ is a vector of internal forces related with the bond length, i.e, the distance between residue pairs.

PRS introduces external forces to the protein. Once the protein is subjective to external forces, the net force on each residue should equal to its

external force [12]:

$$[\mathbf{F}]_{M\times 1} = [\mathbf{B}]_{3N\times M}[\mathbf{f}]_{M\times 1} \tag{2.47}$$

where $\mathbf{F}$ is a vector of external forces. Under the action of those external forces, the protein structure may undergo conformational changes, introducing positional displacements $\mathbf{\Delta R}$ of residues and the bond deformations $\mathbf{\Delta r}$. The work done by external forces should be equal to that done by the internal forces, which gives [262]

$$[\mathbf{F^T}]_{1\times M}[\mathbf{\Delta R}]_{M\times 1} = [\mathbf{f}]_{1\times M}[\mathbf{\Delta r}]_{M\times 1} \tag{2.48}$$

When the conformation changes are relatively small compared with the protein size, the bond deformations $\mathbf{\Delta r}$ can be expressed in terms of a linear combination of positional displacements $\mathbf{\Delta R}$ [262],

$$[\mathbf{\Delta r}]_{M\times 1} = [\mathbf{A}]_{M\times 3N}[\mathbf{\Delta R}]_{3N\times 1} \tag{2.49}$$

Substituting equation (2.47) and equation (2.49) into the equation (2.48), one obtains

$$[\mathbf{f^T}]_{1\times M}[\mathbf{A}]_{M\times 3N}[\mathbf{\Delta R}]_{3N\times 1} = [\mathbf{f^T}]_{1\times M}[\mathbf{B^T}]_{M\times 3N}[\mathbf{\Delta R}]_{3N\times 1} \tag{2.50}$$

after rearragement. The equality equation (2.50) must be valid for any arbitrary $\mathbf{f}$ and $\mathbf{\Delta R}$, which leads to,

$$[\mathbf{A}]_{M\times 3N} = [\mathbf{B^T}]_{M\times 3N} \tag{2.51}$$

and

$$[\mathbf{\Delta r}]_{M\times 1} = [\mathbf{B^T}]_{M\times 3N}[\mathbf{\Delta R}]_{3N\times 1} \tag{2.52}$$

Moreover, in a linear approximation, the relationship of $\mathbf{f}$ and the bond deformation $\mathbf{\Delta r}$ can be stated as

$$[\mathbf{f}]_{M\times 1} = [\mathbf{K}]_{M\times M}[\mathbf{\Delta r}]_{M\times 1} \tag{2.53}$$

where $\mathbf{K}$ is a diagonal matrix whose ith element is the force constant of the ith bond. Left multipling $\mathbf{B}$ on both sides of equation (2.53), and then substituting $\mathbf{\Delta r}$ and $\mathbf{f}$ using equation (2.52) and equation (2.47), one reaches the relationship between the external forces and the induced positional displacements as [13]

$$([\mathbf{B}]_{3N\times M}[\mathbf{K}]_{M\times M}[\mathbf{B^T}]_{M\times 3N})[\mathbf{\Delta R}]_{3N\times 1} = [\mathbf{F}]_{M\times 1} \tag{2.54}$$

or

$$[\mathbf{\Delta R}]_{3N\times 1} = ([\mathbf{B}]_{3N\times M}[\mathbf{K}]_{M\times M}[\mathbf{B^T}]_{M\times 3N})^{-1}[\mathbf{F}]_{M\times 1} \tag{2.55}$$

Note that the $(\mathbf{BKB^T})^{-1}$ is equivalent to the inverse of Hessian as well as the covariance matrix $\mathbf{G}$ of atomic fluctuations. The covariance matrix $\mathbf{G}$ can be computed from the trajectory of the molecular dynamics simulation. With known $\mathbf{G}$, one could obtain the positional displacement of all residues under the perturbation of a random force by

$$[\mathbf{\Delta R}]_{3N\times 1} = [\mathbf{G}]_{3N\times 3N}[\mathbf{F}]_{3N\times 1} \tag{2.56}$$

2.7 Singular Value Decomposition/Principal Component Analysis

Singular Value Decomposition (SVD) is a multivariate statistical procedure to elucidate the underlying structure of data. It could be used to increase the signal-to-noise ratio and reduce the redundancy of data . Similar to Principal Component Analysis (PCA), SVD transforms data to new subspaces identified by orthonormal bases where the covariance of the data along different orthonormal bases is minimized. It is a powerful tool widely used from information science to biology [63, 31, 134, 258, 140].

Let $\mathbf{X}$ denotes a $m \times n$ matrix of interest which contains the information of n subjects characterized by m attributes. The row vector $\mathbf{a}_i$ with n dimension

represents the ith attribute in all n subjects, whereas the column vector $\mathbf{b}_j$ with m dimension represents the m attributes for the j subject. In general, SVD decomposes a given $m \times n$ matrix $\mathbf{X}$ into the product of three other matrices:

$$[\mathbf{X}]_{m\times n} = [\mathbf{U}]_{m\times m}[\mathbf{\Sigma}]_{m\times n}[\mathbf{V}]^T_{n\times n} \tag{2.57}$$

such that $\mathbf{U}$ and $\mathbf{V}$ have orthonormal columns and $\mathbf{\Sigma}$ is diagonal. The columns of $\mathbf{U}$, $\{\mathbf{u}_k\}$, are left-singular vectors. The columns of $\mathbf{V}$, $\{\mathbf{v}_k\}$, are right-singular vectors. In convention, the diagonal elements of $\mathbf{\Sigma}$ called singular values of $\mathbf{X}$ are sorted in descending order, i.e., $\Sigma = diag(\sigma_1, \sigma_2, \ldots, \sigma_n)$ $(\sigma_1 \geq \sigma_2 \ldots \geq \sigma_n)$. These diagonal elements represent the variance along the corresponding left-singular and right-singular vectors. Those vectors with large variance are interpreted to be important as they are most relevant to the main characteristics included in the matrix.

The left-singular vectors $\{\mathbf{u}_k\}$ can be considered as the eigenvectors spanning the new subject subspace. If one wishes to understand the relationship among the subjects, it is necessary to find out the new coordinates of the subjects in this left-singluar subspace. The original coordinates of subject j is given by the column vector $\mathbf{b}_j$. Referring to the definition in equation 2.57, the SVD equation for $\mathbf{b}_j$ is

$$\mathbf{b}_j = \sum_{k=1}^{r} v_{jk}\sigma_k\mathbf{u}_k \tag{2.58}$$

which is a linear combination of the left-singular vectors $\{\mathbf{u}_k\}$. r denotes the rank of matrix $\mathbf{X}$. According to equation 2.58, the jth row of $\mathbf{V\Sigma}$, designated as $\mathbf{b'}_j$ gives the coordinate of the subject j in the left-singlular subspace $\{\mathbf{u}_k\}$. If $r < m$, the attributes of subject can be captured with fewer variables by $\mathbf{b'}_j$ instead of $\mathbf{b}_j$. Thus SVD can be used for purpose of dimensional reduction.

In this subspace, the distance between two subjects j_1 and j_2 becomes

$$d_{j_1,j_2} = |\mathbf{x'}_{j_1} - \mathbf{x'}_{j_2}| = \sqrt{\sum_{k=1}^{r}(v_{j_1,k}\sigma_k - v_{j_2,k}\sigma_k)^2} \tag{2.59}$$

These distances in the subspace provide the basic measure for clustering the subjects. Additionally, the contribution of attributes i in the top left-singular vectors $\{\mathbf{u}_k\}$ is given by a weight

$$w_i = \sum_{k=1}^{r} \sigma_k |u_{ik}| \tag{2.60}$$

The weight indicates the significance of the attribute i in the use of distinguishing all subjects.

On the other hand, the right-singular vectors $\{\mathbf{v}_k\}$ can be viewed as the eigenvectors spanning the new attribute subspace. The new coordinates of the attributes in this rght-singluar subspace reveal the relationship among the attributes. The original coordinates of attribute i is given by the row vector $\mathbf{a}_j$, which can be expressed as a linear combination of the right-singular vectors $\{\mathbf{v}_k\}$:

$$\mathbf{a}_i = \sum_{k=1}^{r} u_{ik}\sigma_k \mathbf{v}_k \tag{2.61}$$

Thus the ith row of $\mathbf{U\Sigma}$, designated as $\mathbf{a'}_i$ gives the coordinate of the attribute i in the right-singlular subspace $\{\mathbf{v}_k\}$.The attributes can be grouped together based on their pairwise distance in this subspace, similar to the approach above used in clustering subjects in the left-singular subspace.

SVD are closely associated with PCA. It can be proven mathematically that the left-singular vectors of $\mathbf{X}$ are the principal components of $\mathbf{XX}^T$, and the right-singular vectors of $\mathbf{X}$ are also the principal components of $\mathbf{X}^T\mathbf{X}$. PCA is used in the analysis of MD trajectory to reveal to essential dynamics of biomolecules. After superposition the snapshot in MD trajectory to a

common reference structure, the covariance matrix of positional fluctuations is constructed:

$$\mathbf{C} = \left\langle (\mathbf{x}(t) - \langle \mathbf{x} \rangle) (\mathbf{x}(t) - \langle \mathbf{x} \rangle)^T \right\rangle \tag{2.62}$$

where $\mathbf{x}$ is the vector of coordinates and $\langle\rangle$ denotes the time average. PCA diagonalizes the covariance matrix via the orthonormal bases $\mathbf{T}$:

$$\mathbf{C} = \mathbf{T}\mathbf{\Lambda}\mathbf{T}^T. \tag{2.63}$$

Here the column vectors of $\mathbf{T}$ are eigenvectors (principal components) of $\mathbf{C}$, and the elements of the diagonal matrix $\mathbf{\Lambda}$ are eigenvalues which are usually sorted in descending order. For a $3N \times 3N$ matrix C of a system with N atoms, there are $3N - 6$ pairs of nonzero eigenvalues and eigenvectors as there are six degrees of freedom for the global translational and rotational motion of the biomolecule. The eigenvalues correspond the variance or the mean square fluctuation along the eigenvectors. Thus the eigenvectors with large eigenvalues are of the interest and they are more relevant with the functional motion.

Chapter 3

PROTEIN TOPLOGY IS A KEY DETERMINANT OF FOLDING KINETICS

As excerpted from:

Zou, T. and S. B. Ozkan, "Local and non-local native topologies reveal the underlying folding landscape of proteins", Physical Biology **8***, 6, 066011 (2011).*

and

Zou, T., Williams N., S. B. Ozkan and Ghosh K., "Proteome folding kinetics is constrained by protein half life". Biophysical Journal (submitted).

3.1 Introduction

Prior work indicates that the topology of the native structure of proteins is an important determinant of their folding mechanism[19, 182, 113]. Plaxco *et al.* first observed that the logarithm of in-water folding rates of two-state proteins is inversely correlated with a topological parameter named *contact order* (CO) or *relative contact order* (RCO) [195, 196]. Subsequent work explored more topology-related properties and folding rate has been found to correlate with many topological characteristics. Gromiha and Selvaraj defined *long-range order* [104, 110] from the content of non-local contacts (number of non-local contacts normalized by chain length) in the native structure and it shows a better correlation with the folding rate than CO. *Multiple contact index* (MCI) [103], which also emphasizes the non-local contacts, also has an inverse relationship with the folding rate. Zhou and Zhou introduced *total contact distance* (TCD) by incorporating CO and *long-range order* (LRO),

and established a good prediction of folding rate in all three structural classes [266]. Another topological parameter termed "*cliquishness*" or *clustering coefficient*, which measures interdependence of contacts (i.e., the extent to which two different residues contacting with the same third residue also contact with each other), is also a powerful indicator of folding speed [172, 15]. Moreover, the knowledge of secondary structures has been implemented to generate the topological parameters used for the prediction of folding rate, such as the *effective chain length* [124] and *secondary structure length* [119]. Based on the effect of chain topology, the methods developed predict the folding rate of a protein with various degrees of success [105, 157]. Besides, the topological properties like the *effective contact order* (ECO) [73] are also helpful for understanding the folding mechanism. ECO is the length of the loop that has to be closed in order to form a contact, given a set of previously formed contacts or contact clusters in the native fold. It has been used for exploring the folding routes and the kinetic impact of secondary structural motifs in folding [255, 256, 171].

The topological characteristics above are based mainly on the global topology of proteins. Besides studying the effect of global topology on folding kinetics, it is worthwhile to investigate the contribution from different components of topology and explore if they give more insight about the folding landscape. To this aim, we extracted different components of the native topology according to the sequence separation of contacts. We study the significance of these extracted components on folding kinetics by examining their contents of local and non-local contacts and the average sequence separation. We introduce the local and non-local contact order to characterize the average sequence separation of different components of the interaction network. Following the idea of

Gromiha and Selvaraj [104] and Harihar and Selvaraj [110], we also use short-range and long-range order to represent the content of local and non-local contacts. Our results show that by studying the native topology alone, not only can we get information about the barrier-crossing process (specifically for α and β proteins), but we also can understand the role of diffusive searches within the denatured ensemble on folding rates.

3.2 Methods

3.2.1 Training Data Set

The experimentally determined folding rates of 82 two-state proteins are collected as the basis for the present study [105, 157, 62], the largest data constructed to date. The PDB codes and experimental folding rates are listed in table 3.1. The structure classification of these proteins yields 25 α proteins, 27 β proteins and 30 α/β proteins. Their folding rates span over 6 orders of magnitude, from $\ln k_f$ = -1.47 for acylphosphatase (1APS) to $\ln k_f$ = 12.9 for albumin binding domain (1PRB).

3.2.2 Contact Network and Contact Order

The contact network is the simplest way to visualize the protein topology. It is constructed from the geometry of the native structures. Each node represents a residue (or an atom) and an edge is formed between two nodes if two residues (or two atoms) are within a specified cutoff distance.

The *contact order* (CO) is the average sequence separation of a protein, defined as [196]:

$$\mathrm{CO} = \frac{1}{N}\sum_{1}^{L-1} S_{ij} \tag{3.1}$$

Table 3.1: The list of 82 two-state proteins

PDB	Length	$\ln k_f$	Structure	PDB	Length	$\ln k_f$	Structure
1BA5	53	5.91	α	2HQI	72	0.18	α/β
1BDD	60	11.69	α	2PTL	62	4.10	α/β
1EBD	41	9.68	α	2VIK	126	6.80	α/β
1ENH	54	10.53	α	1DIV_n	56	6.60	α/β
1FEX	59	8.19	α	1DIV_c	93	0.00	α/β
1HRC	104	8.76	α	1LOP	164	6.60	α/β
1IDY	54	8.73	α	1O6X	81	6.63	α/β
1IMQ	86	7.28	α	1E65	128	4.91	α/β
1LMB	87	8.50	α	1K0S	151	7.44	α/β
1PRB	53	12.90	α	1SPR	103	8.74	α/β
1VII	36	11.51	α	1BF4	63	6.95	α/β
1YCC	109	9.62	α	1J5U	127	6.85	α/β
256B	106	12.20	α	2QJL	99	2.58	α/β
2A3D	73	12.70	α	1UBQ	76	7.33	α/β
2PDD	43	9.69	α	1C8C	64	6.95	β
1L2Y	20	12.40	α	1C9O	66	7.20	β
2ABD	86	6.55	α	1CSP	67	6.54	β
1L8W	291	1.61	α	1E0L	37	10.37	β
2BTH	45	11.78	α	1E0M	37	8.85	β
1SS1	60	11.48	α	1FMK	57	4.05	β
1YZA	106	8.40	α	1G6P	66	6.30	β
1W4E	45	10.22	α	1K8M	87	-0.71	β
1RYK	69	9.08	α	1K9Q	40	8.37	β
1AYI	86	7.20	α	1MJC	69	5.23	β
1W4J	51	12.25	α	1NYF	58	4.54	β
1APS	98	-1.47	α/β	1PIN	32	9.37	β
1AYE	81	6.80	α/β	1PKS	76	-1.06	β
1CIS	66	3.87	α/β	1PNJ	84	-1.10	β
1COA	64	3.87	α/β	1PSE	69	1.17	β
1FKB	107	1.45	α/β	1SHF	59	4.50	β
1HDN	85	2.69	α/β	1SHG	57	2.10	β
1HZ6	62	4.10	α/β	1SRL	56	4.04	β
1N88	96	3.00	α/β	1TEN	90	1.06	β
1PBA	81	6.80	α/β	1WIT	93	0.41	β
1PCA	96	6.80	α/β	2AIT	74	4.21	β
1POH	85	2.70	α/β	1FNF_9	90	-0.92	β
1RFA	78	7.00	α/β	1PGB_b	16	12.00	β
1RIS	97	5.90	α/β	1QTU	115	-0.36	β
1URN	96	5.76	α/β	1JO8	58	2.46	β
2ACY	98	0.84	α/β	2VKN	66	2.11	β
2CI2	65	3.87	α/β	1RLQ	56	4.04	β

where N is the number of pairwise contacts, L is the chain length and S_{ij} is the sequence separation of residues i and j. If the contact is based on residue separation, the corresponding contact order is called *residue-based* CO. In another version, the pairwise contacts are claimed if any two heavy atoms are within the cut-off distance, and we call the corresponding contact order *all-atom* CO.

To be distinguishable from other parameters of local and non-local contact networks below, we call the contact order of the original contact network the *global* CO which counts over all pairwise contacts of a protein.

3.2.3 Local and Non-local Network

In order to investigate the local topology of native structures, we extract local contact networks from the original contact network, which includes edges with sequence separation no larger than the upper bound S_{max}.

The local contact network is characterized by *local* CO and *short-range order* (SRO) in this study. *Local* CO measures the average sequence separation or compactness of a local contact network using a defined upper boundary for the allowable sequence separation S_{max} and is defined as:

$$local\ \text{CO} = \frac{1}{N_{local}} \sum_{2}^{S_{max}} S_{ij} \tag{3.2}$$

where N_{local} is the number of local contacts.

SRO represents the number of local contacts normalized by chain length L:

$$\text{SRO} = \frac{1}{L} \sum_{1}^{N} n_{ij} \text{ and } n_{ij} = \begin{cases} 1 & \text{if } |i-j| \leq S_{max} \\ 0 & \text{otherwise} \end{cases} \tag{3.3}$$

where N_{local} is the number of local contacts. Likewise, a non-local contact network is built by defining the lower bound of the allowable sequence sepa-

ration (S_{min}). The average sequence separation and the normalized number of non-local contacts can be measured by *non-local* CO and *long-range order* (LRO) [104, 110], respectively:

$$non-local\ \text{CO} = \frac{1}{N_{non\_local}} \sum_{S_{min}}^{L-1} S_{ij} \tag{3.4}$$

$$\text{LRO} = \frac{1}{L} \sum_{1}^{N} n_{ij} \text{ and } n_{ij} = \begin{cases} 1 & \text{if } |i-j| \geq S_{min} \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

where N_{non_local} is the number of non-local contacts.

3.3 Results and Discussion

We use local and non-local contact networks to represent different components of the native topology of a protein. Two simple parameters of local and non-local contact networks are studied here: (i) *contact order* (*local* CO or *non-local* CO) and (ii) *short-range order* (SRO) or *long-range order* (LRO). While LRO and SRO represent the number of local and non-local contacts, *local* CO and *non-local* CO measure the average sequence separation, or the average distance along the sequence of local and non-local contacts. *Relative contact order* (RCO) is another common parameter to quantize the average sequence separation of contacts. However, protein size is a determining factor of folding kinetics. CO shows better performance than RCO, which normalizes CO by the number of residues [62], thus we present CO analysis here. We will illustrate and discuss contact order based on the original all-atom version of CO [196] computed at a cut-off distance of 7.0 Å. All the results are consistent for both *residue-based* and *all-atom* CO, and insensitive to the value of cut-off distance.

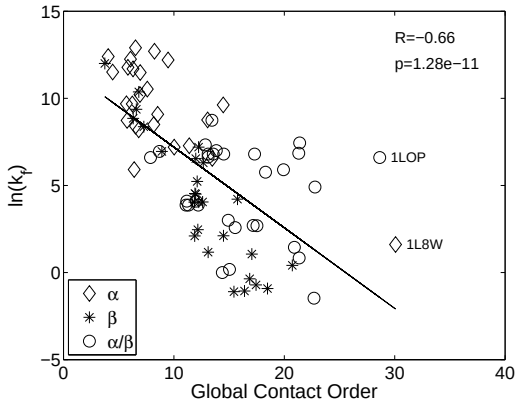


Figure 3.1: The correlation between global CO and $\ln k_f$ for 82 two-state proteins $R = -0.66, p = 1.28 \times 10^{-11}$. There are two proteins (1LOP and 1L8W) that are outliers in the plot, having high global CO but folding very fast. If the two outlier are excluded, the correlation improves with $R = -0.72$ and $p = 7.10 \times 10^{-14}$.

3.3.1 Non-local Native Contacts Dominating the Barrier-crossing Step

As expected, *global* CO has a strong negative correlation with folding rates (figure 3.1, $R = -0.66, p = 1.28 \times 10^{-11}$), showing that proteins with predominantly local interactions should fold more rapidly [196]. The dependence of folding rate on *global* CO can be explained by the loss of configurational entropy upon closing unstructured loops in the native-like transition state [87]. There are two outliers (1L8W: α protein; 1LOP: α/β protein) in the plot, having high global CO and folding very fast. If we focus on different classes of proteins separately, the correlation holds quite well for β proteins, but it turns out to be ambiguous for α proteins, due in part to the outlier 1L8W. There is no correlation for proteins without 1L8W but negative correlation for the case with 1L8W. Thus the role of topology on the folding kinetics of α proteins needs further investigation as more experimental data about α proteins becomes available. Many experiments demonstrate that the formation of helices, hairpins and other local structures is orders of magnitudes faster than

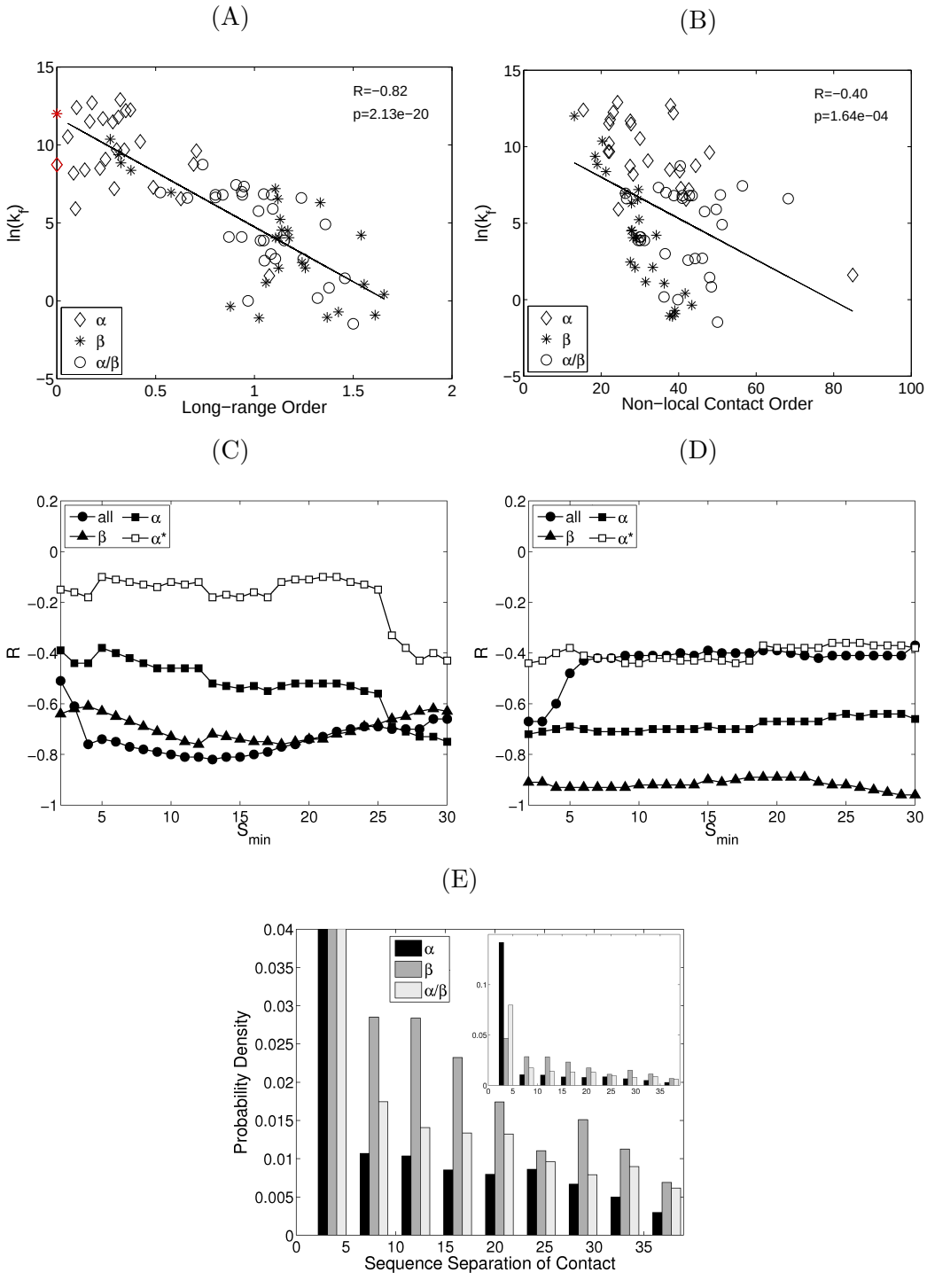
the rate-limiting speed [107, 236, 32, 81, 80]. Moreover, there is almost no energy barrier of folding for these isolated structural elements [89, 42]. Thus the sampling of these local structures happens very rapidly in the denatured ensemble. Since non-local contacts take longer to form, they should account for barrier-crossing. Makarov *et al* showed that folding rate positively correlates with the probability of the formation of non-local native contacts [166]. Here we investigate the non-local contacts by taking into account contacts with sequence separation larger than a minimum value, S_{min}. Interestingly, both the *non-local* CO and LRO consistently show negative correlations with folding rate (figure 3.2(A)-(B)). This demonstrates that an increase in the sequence separation or the number of non-local contacts slows down the folding process. The negative correlation between those parameters and folding rates remains constant over a large range of S_{min}, and is still valid even when S_{min} is 30. It indicates that the negative correlation between *global* CO and folding rates is mainly caused by the non-local contacts, i.e., non-local contacts dominate the barrier-crossing step. Furthermore, LRO shows a better correlation with folding rate than *non-local* CO, suggesting that the effect of the number of non-local contacts on the folding speed is more significant than the average sequence separation of non-local contacts.

It is also interesting to investigate the correlation coefficient profile (R) for different structural classes of proteins separately (figure 3.2(C)-(D)). The correlation for β proteins is better than that for α proteins overall. Besides, the correlation for α proteins is mainly ascribed to 1L8W, as explained when discussing *global* CO above. In figure 3.2(D), the correlation for β proteins remains constant over a large range of S_{min}, whereas for all proteins the correlation drops as S_{min} decreases. It is probably due to the fact that β proteins

are characterized by numerous distant contacts along the sequence, compared with other proteins (figure 3.2(E)), and thus their *non-local* COs are less sensitive to the threshold S_{min}.

3.3.2 Local Native Contacts Related with Conformational Entropy in the Denatured Ensemble

We also study the *local* CO and SRO where local contacts are defined as those contacts with separation less than an upper boundary, S_{max}. Surprisingly, *local* CO also shows a correlation with folding rates, but the slope is positive (figure 3.3(A)-(B)). This suggests that proteins with the higher local CO should fold faster, i.e., large sequence separation of local contacts helps folding. The correlation coefficient peaks around the upper boundary of allowable sequence separation of 6 ($S_{max} = 6$) for all proteins ($R = 0.62, p = 3.54 \times 10^{-10}$). This is interesting because $S_{max} = 6$ is a reasonable value when considering the maximum allowable sequence separation that includes the local motifs of helical and β-hairpin turns. When analyzing different classes of protein separately, the positive correlation is also valid for β proteins, but not for α proteins alone (figure 3.3B), as observed for the *global* CO and *non-local* CO above. Furthermore, in order to justify the behavior of local contacts, we extract a subset of 40 proteins ($\sim$half of the population of the original data set) with similar *non-local* CO, i.e. proteins with similar non-local topology but different local topology . Still, there is a positive correlation between *local* CO and folding rates for the subset (figure 3.3(C)-(D). $R = 0.74, p = 3.92 \times 10^{-8}$ at $S_{max} = 6$). We also validate the positive correlation using different subset sizes consisting of proteins with similar *non-local* CO.


(A)
(B)
R=−0.82
p=2.13e−20
R=−0.40
p=1.64e−04
ln(k_f)
Long−range Order
Non−local Contact Order
α
β
α/β
(C)
(D)
all
α*
R
S_min
(E)
Probability Density
Sequence Separation of Contact

The positive slope in figure 3.3A and (C) seems counterintuitive to that in figure 3.1, but actually it may indicate different information about folding. Earlier study on transition states and intermediate states indicates that folds start locally, and the native secondary structure bias is apparent before the formation of tertiary contacts [123, 24, 144]. A lot of experiments and simulations also confirm the presence of residual structures in the denatured state [226, 74, 224, 173, 225, 158, 51, 34, 207, 200, 201, 259, 52]. In particular, a specific investigation of tertiary interactions by simulation showed that most of the non-local native contacts in the denatured state are lost [259]. The existence of the local structure bias before transition states and local native contacts in the denatured ensemble brings out a possible explanation for the positive correlation in figure 3.3(A)-(C). For illustration purpose, let us divide local contacts into two subgroups: very-local ($(i,\ i+4)$ or shorter) and mid-local ($(i, i+5)$ to $(i, i+8)$). Proteins with a large *local* CO that are rich in mid-local contacts would have a high chance to sample the mid-local contacts in the denatured state. To explore this further, we select two proteins along the fitting line of figure 3.3(A): 1FNF_9 and 1E0M (figure 3.4(A)), which fold slowest and fastest among β proteins. Their histograms of sequence separation

---

Figure 3.2 *(preceding page)*: (A) The correlation between *long-range order* (LRO) and $\ln k_f$ when $S_{min} = 13$ ($R = -0.82, p = 2.13 \times 10^{-20}$). Two proteins (1PGB and 1IDY) are colored read and excluded from the fitting since they have LRO $= 0$ when $S_{min} = 13$. (B) The correlation between *non-local* CO and $\ln k_f$ when $S_{min} = 13$ ($R = -0.40, p = 1.64 \times 10^{-4}$). (C)-(D) The correlation coefficient R of $\ln k_f$ vs. LRO (C) and *non-local* CO (D) at different values of the threshold S_{min} for all 82 proteins (solid circle) and different classes of proteins separately (β proteins: solid triangle; α proteins: solid square; α proteins except 1L8W: open square) at different values of the threshold S_{min}. (E) Histogram of sequence separation of contacts for different classes of proteins.

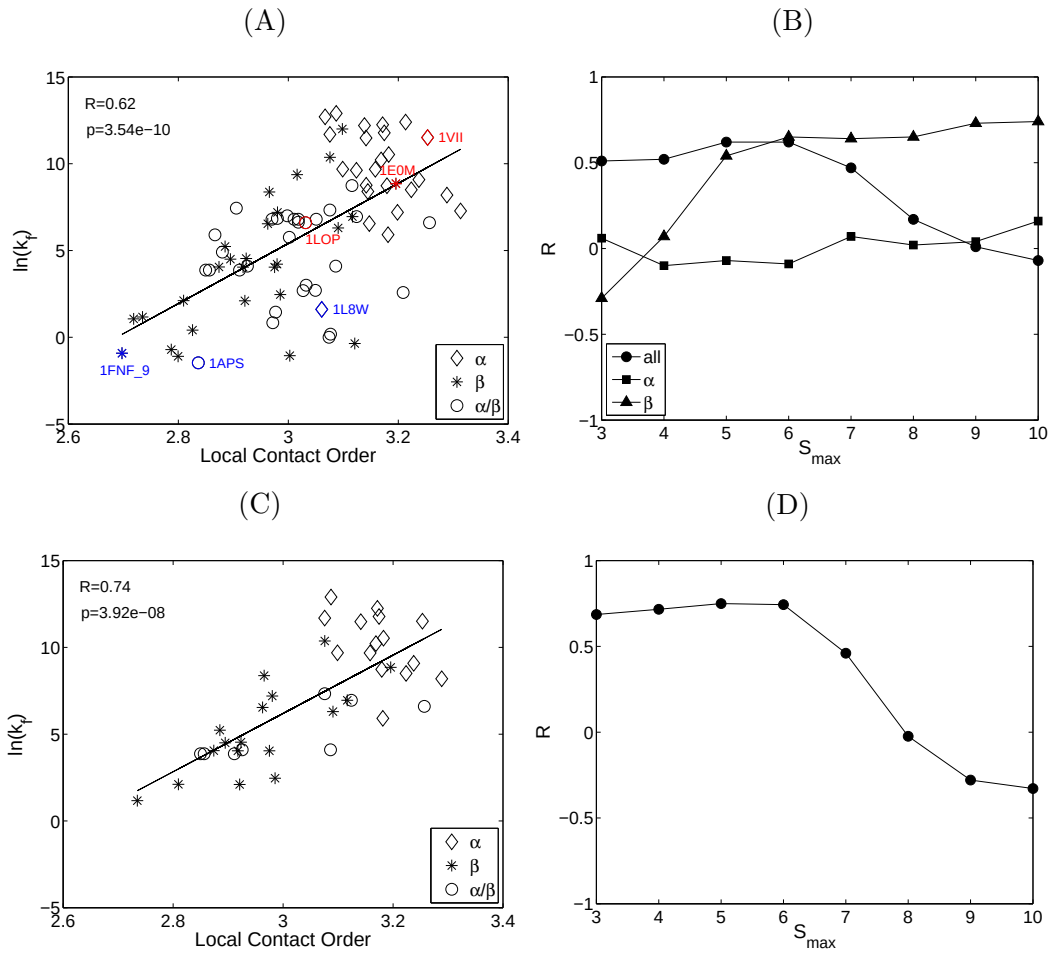


Figure 3.3: *Local* CO shows positive correlation of $\ln k_f$. (A) The correlation of local CO and $\ln k_f$ for 82 two-state proteins at $S_{max} = 6$ ($R = 0.62, p = 3.54 \times 10^{-10}$). (B) The correlation coefficient R of $\ln k_f$ vs. *local* CO for all 82 proteins (solid circle) at different values of the threshold S_{max}, along with only β proteins (solid triangle) and only α proteins: (solid square). (C) The correlation of *local* CO and $\ln k_f$ for 40 proteins with similar *non-local* CO at $S_{max} = 6$. (D) The correlation coefficient R of $\ln k_f$ vs. *local* CO for the 40 proteins changes with the threshold S_{max}.

confirm that 1E0M (i.e. the fast folding β protein) contains more mid-local contacts than 1FNF_9, as we expected in figure 3.4(B). 1FNF_9 contains 4.01% mid-local and 49.50% non-local contacts, while the percentages of mid-local and non-local contacts for 1E0M are 13.86% and 24.75% respectively. The presence of mid-local contacts has two impacts. First, it can reduce the time of diffusive searching in the denatured basin. The formation of local contacts causes the loss of configurational entropy in the denatured state and decreases the size of the denatured ensemble [97]. The effect is made stronger by those mid-local contacts with large sequence separation. The smaller denatured ensemble needs less time to be explored. Second, the mid-local contacts can help the formation of non-local native contacts. When a mid-local contact forms, it serves as a constraint. A small additional search will bring the non-local contacts into spatial proximity and they are then able to form with a lower entropic penalty, as has been shown in Zipping and Assembly method (ZAM) [188, 100, 222]. Indeed, ZAM uses zipping and assembly mechanism which a microscopic recipe for folding. ZAM works by: (i) breaking the full protein chain into small fragments (initially 8-mers), which are simulated separately using replica exchange molecular dynamics (REMD), (ii) then growing or zipping the fragments having metastable structures by adding a few new residues or assembling two such fragments together, with further REMD and iterations, (iii) locking in place any stable residue-residue contacts with a harmonic spring, enforcing emerging putative physical folding routes, without the need to sample huge numbers of degrees of freedom at a time. The existence of mid-local contacts is very crucial in efficient folding for ZAM. The formation of those non-local contacts would be unfavorable if there were only very-local contacts, because in the absence of mid-local contacts it would involve a large

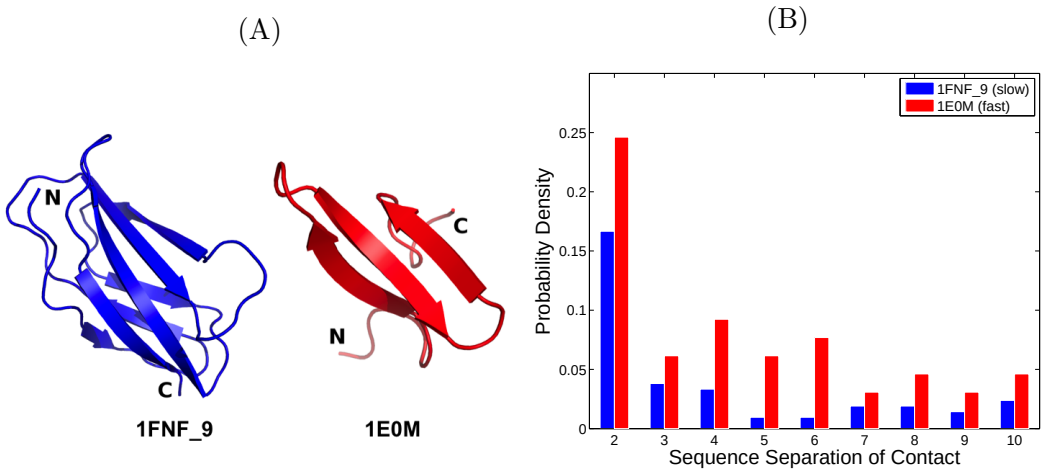


Figure 3.4: Comparison between two β proteins: 1FNF_9 (slow folding) and 1E0M (fast folding). (A) Structures of 1FNF_9 and 1E0M. (B) The histogram of local sequence separation shows that fast folding 1E0M contains more mid-local contacts than slow folding 1FNF_9. Y-axis represents the probability density of contacts with certain sequence separations among all nonbonded contacts of the protein.

conformational search as observed for 1FNF_9 (figure 3.4). Moreover, the two outlier proteins (1LOP and 1L8W) on the *global* CO plot (figure 3.1) do not deviate significantly in the *local* CO plot. Further analysis of their topology indicates that 1LOP is rich in mid-local contacts and this might be a reason why it folds relatively faster than many α/β proteins with *lower global CO*. 1L8W folds much slower than other α proteins and it is too poor in mid-local contacts . Briefly, the formation of mid-local contacts can decrease the number of possible conformations in the denatured ensemble and acts as a bridge to the appearance of non-local native contacts. This may be the reason why proteins with larger local CO fold faster. SRO also shows positive correlation with folding rate. It means that the increase of the number of local contacts (thus, decrease in non-local contacts) speeds up folding, which is in agreement with the conclusion from LRO.

In summary, the study of non-local and local topology provides a better understanding of the effect of topology on folding kinetics. Especially, non-local CO and local CO show opposite trends with folding rates. We suggest that the non-local topology may be more related to the barrier-crossing process, while the local topology of the native structure may dominate the entropy of the denatured ensemble.

3.4 Application: Predicting Proteome Folding Kinetics

The prediction power of protein topology can be used to get insights about the folding kinetics on the proteome scale, where the folding speed of most proteins has not been measured by experiment. Here, we use a slight variant of CO that captures the details of native topology, benchmarked against the largest set of (116 proteins) experimentally measured folding kinetics [210]. For a given protein, we predict folding speeds for different domains, assuming each domain folds independently. Since the domain with the slowest folding speed is rate limiting, we use the folding speed of the slowest folding domain to be the folding speed of the protein. In order to compute folding time from the native structures, we selected proteins from the Yeast and E. coli proteome for which both structures and abundance values are known. For the Yeast proteome we used domain assignment from Yeast resource center (YRC) database [75]. Next we performed a BLAST search of the corresponding sequences to identify the best possible match for their structures. We collected only those proteins that simultaneously satisfy a minimum of 80% sequence coverage and 50% identity match. In order to maximize the coverage of proteins from the proteome, we cross-referenced these proteins against the most comprehensive list of protein abundance values, integrated list from PaxDB database [252]. This method

yields a total of 755 Yeast proteins. For E. coli proteome, we followed a similar approach but used the dataset collected by OBrien *et al* [53]. The original dataset reported in OBrien *et al* [53] categorizes proteins based on a single abundance scale. We cross-referenced the combined list against the integrated list of abundance from PaxDb [252] yielding a total of 848 E. coli proteins. In summary, our datasets provide the largest fraction of proteomes (in E. coli and Yeast) for which both the abundance and structural informations are now available.

Copy number weighted folding rate ($\ln k_f$) distributions in E. coli and Yeast show a broad range of folding speeds, from microseconds to minutes figure 3.5(A). The average folding time for E. coli ($\tau_f = 1/\langle \ln k_f \rangle$) is found to be $\approx$ 100 milliseconds, and remains unaltered when protein expression level is ignored. The average folding time for Yeast is 170 milliseconds and 60 milliseconds for copy number weighted and unweighted distribution, respectively. Recent work, grounded in the hypothesis of global selection against toxic effect of misfolding explaining observed correlation between abundance and evolution rate [221], predicts highly abundant proteins are more stable [220]. Given this link between stability-abundance and possible interdependence between stability and folding kinetics [253], it is natural to expect a possible relation between abundance and folding kinetics as well. However, based on the results stated above, we do not see any noticeable effect of abundance on the proteome folding kinetics. It is also interesting to note, folding speed distribution in E. coli and Yeast are very similar, indicating an universal behavior in the folding kinetics.

The universal distribution figure 3.5(A) of the folding speed, irrespective of the details of the species, is well explained by a diffusion drift model of

(A) (B)

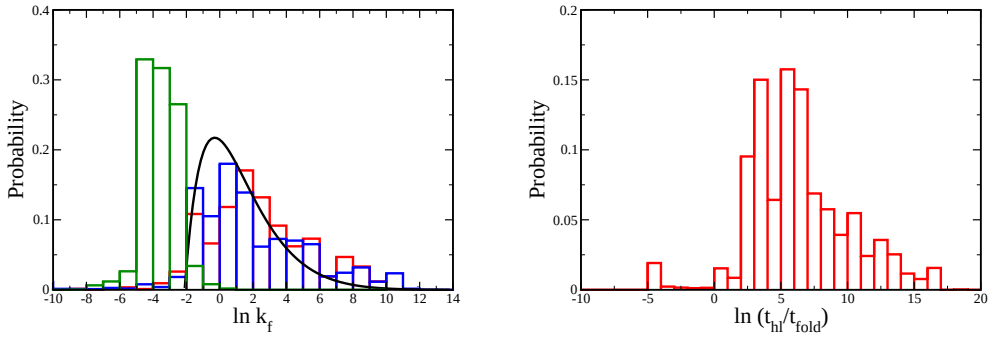


Figure 3.5: (A) Copy number weighted folding rate ($\ln k_f$) distribution for E. coli (in red) and Yeast (in blue). The distribution of degradation rates for proteins in Yeast [29] is shown in green. Both folding rates and degradation rates are presented in the unit of s^{-1}. The predicted folding rate distribution using a diffusion-drift model (equation (3.6)) with the boundary condition of slowest folding time limit of eight seconds is shown in black. (B) Distribution for the ratio of experimentally measured half life (τ_{hl}) [17] and predicted folding time (τ_{fold}).

mutations altering folding free energy barrier ($\Delta G^{\dagger}$). The model is very similar to what has been used to explain stability distribution assuming random mutations, with a drift, alter stability subject to the constraint of fitness arising due to degradation or misfolding [265]. Here we use similar idea where mutations alter the free energy barrier of folding. The model imposes two boundary conditions, $P(\Delta G^{\dagger}_{min}) = P(\Delta G^{\dagger}_{max}) = 0$, at the two extremities of the free energy barrier, $\Delta G^{\dagger}_{min}$ and $\Delta G^{\dagger}_{max}$. On one hand it is simply impossible to make proteins that fold faster than the speed limit of folding, setting the lower limit of the barrier $\Delta G^{\dagger}_{min}$. On the other hand, extremely slow folding proteins - if not folded at birth - even if highly stable, are prone to misfolding or degradation due to prolonged residence in their unfolded state. This sets a selection pressure against slow folding proteins with extremely high barriers ($\Delta G^{\dagger}_{max}$). Thus, the model assumes flat fitness landscape for $\Delta G^{\dagger} < \Delta G^{\dagger}_{max}$,

with a severe drop in fitness for $\Delta G^{\dagger} > \Delta G^{\dagger}_{max}$. Following Zeldovich *et al* [265] the predicted distribution for the free energy barrier can be formed as [269]:

$$P(\Delta G^{\dagger}) = A \exp\left(\frac{h\Delta G^{\dagger}}{h^2 + D}\right) \sin\left(\pi \frac{\Delta G^{\dagger} - \Delta G^{\dagger}_{min}}{\Delta G^{\dagger}_{max} - \Delta G^{\dagger}_{min}}\right) \tag{3.6}$$

where, A is a normalization constant, h and D are the average and variance, respectively, of the distribution of barrier height changes upon mutation. Formally, $h = \langle \Delta\Delta G^{\dagger} \rangle$ and $h^2 + D = \langle (\Delta\Delta G^{\dagger})^2 \rangle$; $\Delta\Delta G^{\dagger} = \Delta\Delta G^{\dagger}_{mutant} - \Delta\Delta G^{\dagger}_{wt}$. From the dataset of 858 mutations across 24 different proteins [175], we find $h = 0.6 k_B T$ and $h^2 + D = 1.12 (k_B T)^2$. The lower limit of the barrier is assumed to be zero ($\Delta G^{\dagger}_{min} = 0$), consistent with barrier less folding proteins that define the speed limit of folding [107, 97, 210]. Using equation (3.6) and speed-barrier height relation $k_f = k_0 \exp(\Delta G^{\dagger}/k_B T)$, we predict the folding speed distribution of the proteome. We use $k_0 \sim 1 \mu s^{-1}$, consistent with several estimates of folding speed limit [107, 97, 210, 146, 120]. The lower speed limit - hence the maximum barrier height was determined by fitting fitting the distribution. Thus, we just use a single variable to fit the distribution. For Yeast proteome, we find the best fit value of the lower speed limit to be approximately eight seconds. Remarkably, this value is in the neighborhood of the fastest degradation times observed in Yeast [29]. This argument quantitatively supports the hypothesis that the proteome degradation imposes limits on the folding time distribution. For comparison, we also plot the degradation rate scale for Yeast in figure 3.5 (in green). It is interesting to note, using equation 8 from [265] and the values of h, D, $\Delta G^{\dagger}_{max} - \Delta G^{\dagger}_{min}$, obtained by matching the speed distribution, impose an upper limit on the number of mutations per portion of the genome encoding essentials genes per replication to be $\sim$5.4. This value is indeed very close to 5.7(≈ 6) predicted by Zeldovich *et*

al from the consideration of the stability distribution and matches well with experiments [265].

Next, we directly compare experimentally measured degradation time [29] and predicted folding times for each individual proteins in our list figure 3.5(A).We select proteins from our list - used to predict the folding time in the Yeast proteome - for which degradation times are known [29]. We find less than 3% of the proteome (13 out of 520 proteins in our list) has a folding time slower than the degradation time.The overwhelming number of proteins with a faster folding time than the degradation, further supports the hypothesis that the lower limit of protein folding speed is indeed constrained by protein degradation. Although 3% is a minor fraction, one can further reason these possible exceptions. It is likely that some of the slow folding proteins actually undergo co-translational folding, a process adopted by almost 30% of the proteome in E. coli [53]. Similarly, chaperones can play an important role to facilitate folding [169, 36, 54]. Third, it is possible that the kinetics of the slowest folding domains are altered due to possible interdependence between multiple domains [26], an aspect not included in our model. Although our prediction of folding speed is based on models benchmarked strictly against in-vitro folding data, recent work shows little difference between folding speed measured in-vitro and in-vivo [106]. It is worth noting that main conclusions do not change if different metrics, for example CO or chain length [235], are used to predict the folding time.

3.5 Conclusion

We have systematically analyzed the relationship between protein topology and folding kinetics. The topology of native structure carries more information

about folding than just predicting folding rate. We find that both the increase in the number of non-local contacts and in the average sequence separation of non-local contacts slow down the folding speed. We also observe that the average contact order of local and non-local contacts show opposite correlation with folding rates because they might be related with different regimes of the folding energy landscape. We propose that local topology is more related to the configurational entropy of the denatured state and non-local native contacts become less unfavorable in the presence of certain mid-local contacts, as also explained in the zipping mechanism [188, 187]. Therefore, on one hand, local contact order positively correlates with the experimental folding rates. On the other hand, non-local topology dominates the barrier-crossing step and results in the negative correlation of the folding rate on the global contact order. By dividing the native topology into local and non-local contact networks, it is possible to gain more insights about the folding landscape.

We have also predicted the folding time distribution, weighted by protein expression level, for E. coli and Yeast proteome, based on the topology of proteins. The folding time distributions of E. coli and Yeast proteome weighted by protein expression levels present a universal picture that the lower speed limit (≈ 8 s) for folding is determined by protein degradation time scale. This finding is supported by: i) a diffusion drift model of mutations altering folding free energy barrier that captures the obtained distributions, and ii) a direct comparison between predicted folding time and experimentally measured halflife at the individual protein level.

Chapter 4

A PHYSICS-BASED APPROACH TO UNDERSTAND PROTEIN FOLDING

4.1 Introduction

It is the basis of protein science that the amino acid sequence of a protein governs its structure and its function. To further understand the role of the evolutionary information on protein fold, Ranganathan group designed artificial sequences based on a computationally and experimentally facile model system, the WW domain [227]. WW domain is a small protein module with about 35-40 amino acids that present in a variety of proteins involved in signaling, regulatory, and cytoskeletal function [46, 122]. It adopts a meandering, tripled-stranded β-sheet fold and is named after the two highly conserved tryptophans which binds to proline-rich peptide motifs [165]. In their study, the conservation and coevolution analysis was performed on the multiple sequence alignment (MSA) of 120 members of the natural WW domain first. Then in order to test the necessity and sufficiency of those conversation and coevolution information for specifying protein fold, two libraries of artificial sequences were constructed using computational algorithms [227]: i) site-independent conservation (IC) sequences which only preserve the amino acide composition (conservation) at each single site but diminish the pairwise coevolution between sites. ii) coupled-conservation (CC) sequences which maintain both the pattern of conservation and pairwise coevolution information. Those artificial sequences, along with positive and negative control (natural sequences and

random sequences) were expressed later in Ecoli and their foldabiliies were examined in a series of experiments. The results showed that none of IC sequences can fold, even though the mean amino acid identities of both CC and IC sequences are similar to that of natural sequences. In contrast, 28 percent of CC sequences not only fold, but also show excellent structural agreement with natural WW domain [227]. This study suggested that the conservation and coevolution information might be necessary and sufficient to specify the fold of a protein family.

I aim to decipher the folding code encoded in natural sequences and reveal how such evolutionary information helps specify a protein fold. The understanding might provide an alternative way to design proteins or help improving current protein design approaches. For this purpose, I simulated a repertoire of artificial WW domain sequences with known foldability using a physics-based protein structure search method called ZAM, which samples conformational space effectively towards native-like conformations through zipping and assembly search mechanism. These artificial sequences were designed by Ranganathan group based on only statistical information encoded in MSA and no tertiary structure information. I focus on protein folding from evolutionary perspective and explored the sequence-structure relationship for WW domain protein (especially how mutations affect folding) through analysis of evolutionary information and the simulation data.

4.2 Methods

4.2.1 A Database of WW Domain Sequences

The dataset of WW domain sequences includes 28 natural (NT) sequences, 31 CC sequences and 30 IC sequences with known foldability. A breakdown of the sequences is shown in table 4.1. I excluded the insoluble and poorly expressed sequences as well as unfoldable NT sequences in original Ranganathans sequence sets, because the reasons for insolubility, poor expression and unfoldbality of native sequences are beyond the scope of the study here. The length of these sequences ranges from 33 to 37, and most sequences (61 out of 89) have 34 residues.

Table 4.1: The simulated WW domain sequences with known foldability

	NT	CC	IC	Total
Foldable	28	12	0	40
Unfoldable	0	19	30	49

4.2.2 Simulation Details

All the WW domain sequences in table 4.1 were simulated using ZAM. I performed independent simulations of full-coverage 8-mer fragments (the first step of ZAM) of these sequences, which explored the possible nucleation sites at the beginning of folding process. Acetyl and N-methylamine blocking group were used to cap the N and C termini of the fragments respectively. Each simulation was 5 ns in length with 2 fs time step and 15 replicas distributed exponentially over the range of 270 to 450 K and giving an average ≈50% acceptance ratio. Replica swaps were attempted every picosecond. I also grew

some sequences up to the full sequences to confirm their fold *in silico*. The representative conformations were clustered using K-mean algorithm from the last nanosecond of the trajectory at the lowest replica temperature. I used the AMBER ff96 forcefield with the GB implicit solvent model, which has been shown to predict the structures of small peptides with better accuracy than other combinations of AMBER forcefields with GB models.

4.2.3 Contact Probability Metric

Contact probability (CPROB) is the equilibrium probability of a contact, calculated as the fraction of sampled configurations with inter-residue distance less than 8.0 Å. It has been shown to be the best single predictor to determine whether a contact is native or not in the simulations of fragments [248]. Here only local contacts (with sequence separation less than 8) are considered and they are sampled extensively in the simulation of 8-mer fragments. As a contact of a given sequence may be included in many 8-mer fragments, the CPROB of the contact is computed by averaging over all the 8-mer fragments containing this contact. Therefore, a sequence with total N possible local contacts can be represented by a CPROB vector $\vec{x} = (x_1, x_2, \cdots x_n)$ ($1 \leq n \leq N$ and x_i is the average CPROB of local contact i)

The size of CPROB vector $\vec{x}$ varies arising from the sequences with different length. The location of contacts may also change in different sequences due to insertions or deletions. It is necessary to uniform the CPROB vector and the definition of contacts before making comparisons. To accommodate the sequences of different length, I ignored the contacts whose starting position is an insertion or deletion (positions colored in gray in figure 4.1) and only calculated CPROB for the contacts whose starting position is not a insertion

Figure 4.1: A sample contact (11, 16) in the multiple sequence alignment of representative WW domain sequences. The insertion or deletion are colored in gray. The positions which do not have insertion or deletion in all sequences are colored in bold black. I only studied the contacts that starts with residues in bold black. The red transparent shadows mark the starting position (red) and ending position (green) of the contact (11, 16). The starting positions are aligned but the ending position varies in different sequences in order to maintain the same sequence separation of 5. The contact (11, 16) is name after the starting residue 11 and ending residue 16, based on a 34-residue long sequences (the sequence in the first row).

or deletion in all sequences (positions colored in black in figure 4.1). There are 28 positions like this which gives 115 local contacts (with sequence separation from 4 to 7) in 8-mer fragments, i.e., a uniform size of CPROB vector with 115 elements or $N = 115$. Furthermore, since sequence separation (contact order) is a very important feature defining the topology of a contact, I chose to maintain the same sequence separation for a given contact, and thus the ending position of the contact may vary in different sequences due to insertion or deletion. Figure 4.1 shows the location of the contact (11, 16) in different sequences as an example.

4.2.4 Classification Model

For each sequence in the dataset, I know its foldability from earlier experiments and have a CPROB vector. Based on these data, I wish to train a probabilistic model to estimate the probability of a sequence being foldable versus unfoldable, given only the CPROB vector obtained from the 8-mer fragment

simulations. This is a binary classification problem, where I have an unknown outcome z which could be either foldable ($z = 1$) or unfoldable ($z = 0$) and I want to calculate $P(z = 1|\vec{x})$, the probability of the sequence being foldable given its CPROB vector (Pfold). Such problem could be solved using logistic regression model where the log odds (logit), a function of $P(z = 1|\vec{x})$, is assumed to be linearly related to $\vec{x}$:

$$\log \frac{P(z = 1|\vec{x})}{1 - P(z = 1|\vec{x})} = \alpha + \vec{\beta} \cdot \vec{x} \tag{4.1}$$

Solving for $P(z = 1|\vec{x})$ yields:

$$P(z = 1|\vec{x}) = \frac{\exp(\alpha + \vec{\beta} \cdot \vec{x})}{1 + \exp(\alpha + \vec{\beta} \cdot \vec{x})} \tag{4.2}$$

Those linear coefficients α and $\vec{\beta}$ are estimated using maximize likelihood estimation. The Wald statistics of β_i indicates the significance of the contact i.

Given a large selection of possible potential contacts, I followed forward stepwise regression approach to search a large space of possible models. Starting with no predictors in the model, this approach tests addition of each predictor, adds the predictor that improve the model most and repeat this process until none improves the model significantly. Despite the advantage of efficient model search, this approach is prone to over-fit the data. Thus I also used cross validation method to prevent over-fitting, where the training data used to construct each model is divided randomly into three groups so that independent models could be built for each group. 1/3 of the data is set aside for testing the model, and the rest is used to train the model.

4.2.5 *Strategy of Designing New Foldable Sequence*

Five local contacts are found to be sufficient for the construction of the classification model to differentiate foldable and unfoldable sequences. It indicates that the modification of CPROB of the five local contacts may lead to the change of foldability of a given sequence. Thus new foldable sequences can be designed by introducing mutations to the five local contacts of unfoldable sequences. Taking an unfolded sequence as a template, I tried to maximize the expected Pfold $P(z = 1|x_1, x_2, \cdots x_5)$ for the template by swapping its five local contacts (or ten residues) with those of a foldable natural sequence. To achieve this, I enumerated all possible combinations of swaps (i.e., swapping only one certain contact or two contacts, etc). The expected Pfold after swapping was calculated with equation (4.2), where the CPROBs of the swapped contacts were represented by those from the foldable natural sequence and unswapped ones were kept as originally in the unfolded sequence (figure 4.2). The hybrid sequence (a mixture of unfoldable template and amino acids from foldable sequence) corresponding to the maximum expected Pfold would be further examined in ZAM simulation and experiment.

4.3 Results

4.3.1 *Crucial Local Contacts Highly Impacts on Foldability*

Every 8-mer fragment of all sequences was simulated with ZAM, to explore the possible nucleation sites at the beginning of folding process. For each sequence, I computed the CPROB vector $\vec{x} = (x_1, x_2, \cdots x_n)$ $(1 \leq n \leq N$ and $N = 115)$ for all 115 possible local contacts . There are totally 89 CPROB vectors, arising from 40 foldable and 49 unfoldable sequences (table 4.1). To make

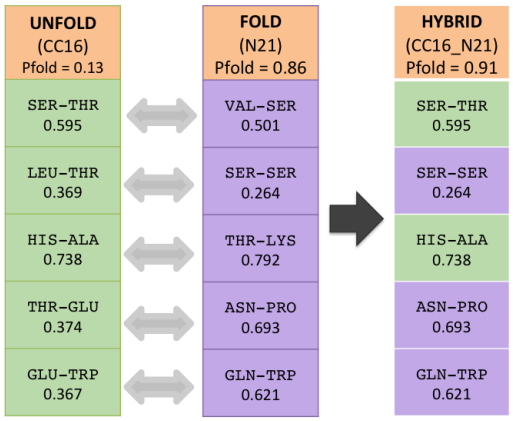


Figure 4.2: Generate the hybrid sequence CC16_N21 based on an unfoldable scaffold CC16 and a foldable sequence N21. Two contacts remain the same as in CC16 (green) and three contacts are chosen to swap (purple).

comparison, CPROB histograms of every contact were generated for foldable and unfoldable sequences. Figure 4.3 displays the CPROB histograms of contact (11,16). It shows that this contact is more favored by foldable sequences than unfoldable sequences. In fact, this contact is important for the stabilization of the N-terminal hairpin (figure 4.7). The maximum likelihood CPROB (MLCPROB) for foldable and unfoldable sequences can also be obtained from the CPROB histogram (figure 4.3). Here a normal Kernel Density Estimate is used to smooth the histogram which removes the dependence on the bin starting points and better reflect the underlying. The x coordinate of the peak is MLCPROB.

MLCPROB maps were constructed for foldable and unfoldable sequences separately based on 8-mer fragment simulations. On the map, each rectangle stands for a contact and the color represents the value of MLCPROB from 0 (blue) to 1 (red). The MLCPROB maps of 8-mer fragments show that there are strong local interactions in the turn segment of N-terminal hairpin for foldable sequences of WW domain (i.e., high local contact probabilities

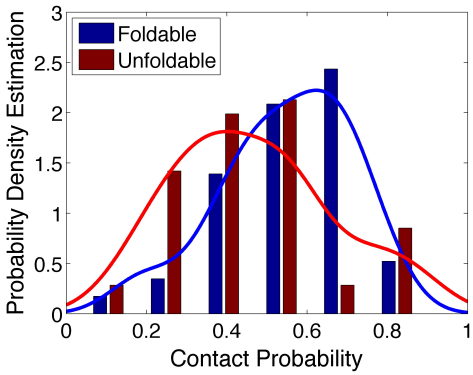


Figure 4.3: Contact probability (CPROB) histogram of contact (11,16). A normal Kernel Density Estimate is used to smooth samples (solid lines). The x coordinates of the peaks on solid lines are MLCPROB for foldable and unfoldable sequences respectively.

are observed around the region corresponding to the N-terminal hairpin) (figure 4.4(A)). On the contrary, on the map of unfoldable sequences, weak local interactions are observed in N-terminal hairpin (figure 4.4(B)). Furthermore, such difference becomes even more significant when I grew them to 16-mer fragments. Figure 4.5 shows the CPROB maps calculated form 16-mer fragment simulations, and it strongly shows strong local interactions in N-terminal hairpin and the growth of N-terminal hairpin for the foldable sequence. However, for the unfoldable sequence, CPROB near N-terminal hairpin are very low indicating weak interactions, and strong interactions are observed in another region. These strong non-native interactions create a frustration and weaken the formation of N-terminal hairpin which may cause it unfold.

4.3.2 Contact Probabilities of Local Interactions can Predict Foldability

Since foldable and unfoldable sequences show different behaviors about local contacts at the early stage of protein folding, I ask if there is a way to make possible prediction about whether a sequence is foldable or not, given

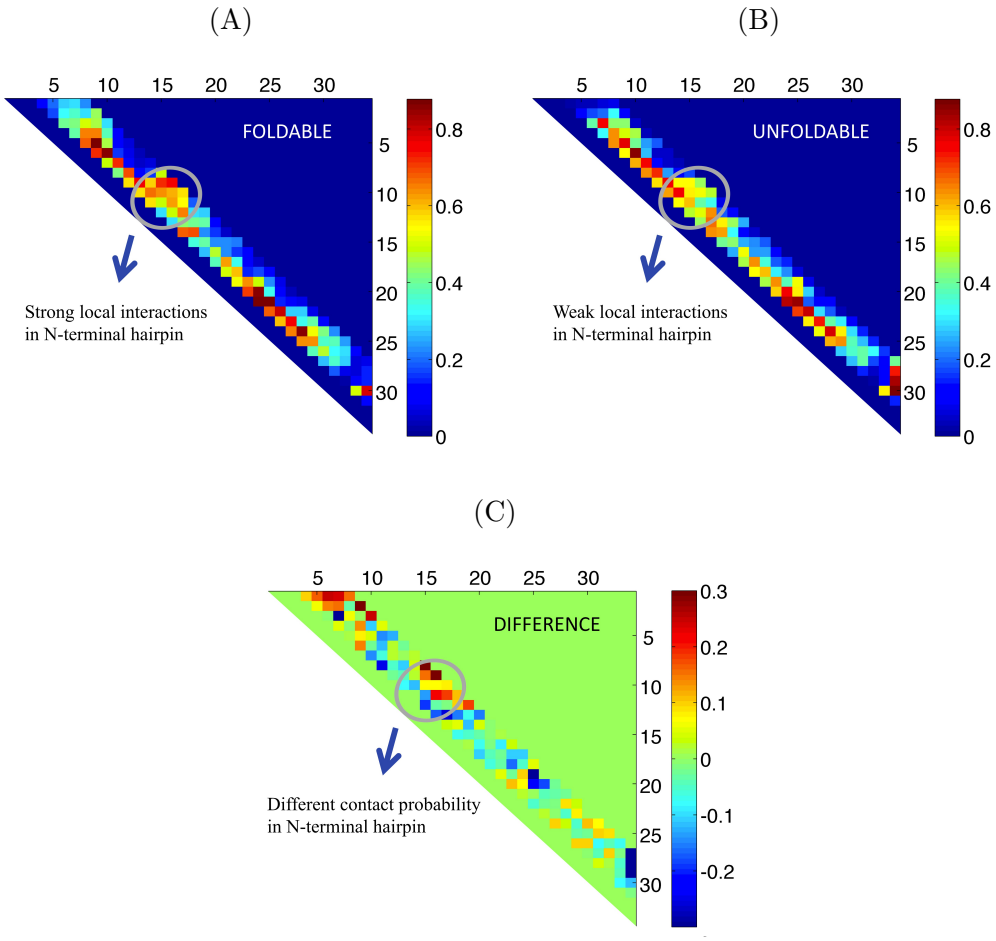


Figure 4.4: MLCPROB maps from 8-mer fragment simulations for (A) foldable sequences, (B) unfoldable sequences and (C) their difference (MLCPROB of foldable sequences subtracts that of unfoldable sequences).

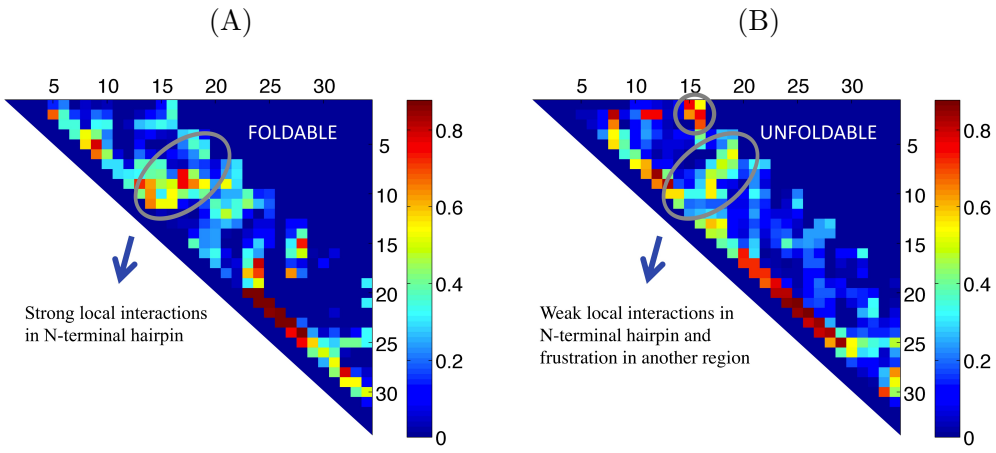


Figure 4.5: CPROB maps from 16-mer fragment simulations for (A) a representative foldable sequence N2 and (B) a unfoldable sequence IC1.

Table 4.2: List of five crucial local contacts in the classification model

Crucial Contacts	Coefficient	S.E.	Wald Statictics	p-value
(2, 7)	3.426	1.434	5.713	0.017
(4, 7)	-6.278	1.770	12.581	0.000
(10, 13)	-3.330	1.234	7.282	0.007
(11, 16)	6.685	1.995	11.233	0.001
(25, 28)	5.554	1.843	9.086	0.003

its CPROB vector. Furthermore, if is is true, are there certain local contacts much more critical for initiation of folding (nucleation sites) than others? To answer these questions, I built a classification model, where the probability of a sequence being foldable given its CPROB vector $P(z = 1|\vec{x})$ is expressed as a function of $\vec{x}$ and solved using maximum likelihood estimation (see Method).

From such classification analysis, I find that only five elements of the CPROB vector $\vec{(x)}$, i.e., five of 115 local contacts, are enough to differentiate the foldability of sequences in the dataset with high accuracy. These five crucial contacts are listed in table 4.2. Using CPROB of these five local

Table 4.3: The result of prediction using five crucial local contacts

	Predicted		
Oberved	*unfoldable*	*foldable*	Percent Correct
unfoldable	41	8	83.7
foldable	9	31	77.5

contacts computed from ZAM simulation, I achieve to predict the foldability of WW sequences with average true prediction rate 80.9% (table 4.3), when the sequences are classified to be foldable if the conditional probability of foldability (Pfold) $P(z = 1|x_1, x_2, \cdots x_5) > 0.5$ and unfoldable if $P(z = 1|x_1, x_2, \cdots x_5) \leq 0.5$. This model also shows excellent statistical significance compared with random models, with high true prediction rate (figure 4.6(A)) and low deviance (figure 4.6(B)). Mapping those contacts onto a crystallographic structure of WW domain, four of them locate in or around the N-terminal hairpin, which have been shown to form early and control the folding process in many experiments and simulations (figure 4.7).

Given the fact that the formation of N-terminal hairpin is a critical step of folding, as well as the importance of five local contacts found from statistical analysis on simulation data, I wonder whether the stabilization of those crucial local contact could assist the formation of N-terminal hairpin and hence avoid of misfolding. To test this idea, for a few unfolded sequences, I artificially constrained two crucial local contacts ((10, 13) and (11, 16)) at the N-terminal hairpin in the simulation. It turns out the artificial constraints of the two contacts indeed increases the probability of the formation of N-terminal hairpin and lead to the correct fold finally (figure 4.8(A)-(B)).

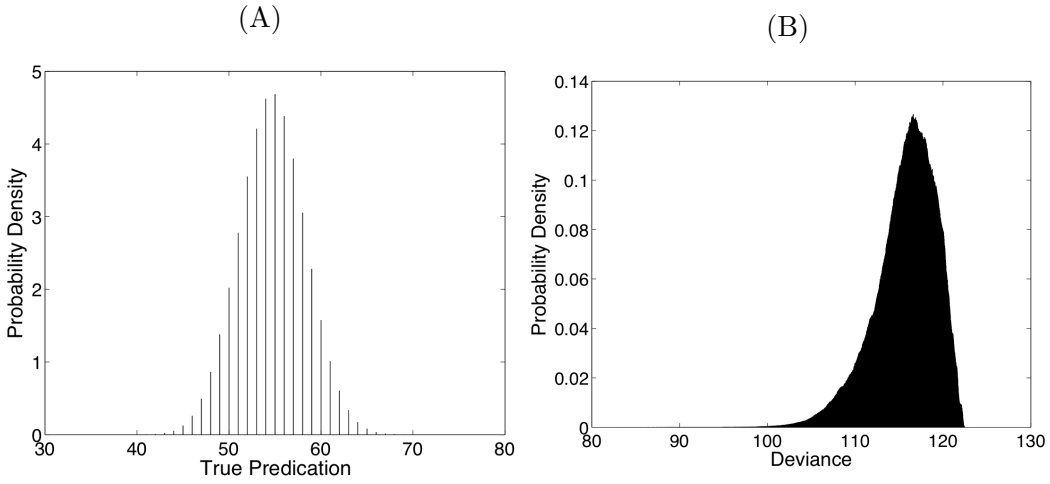


Figure 4.6: Histograms of (A) true prediction and (B) deviance (a measure of the lack of fit to the data) for all possible models using any five of 115 local contacts. The model with the five crucial local contacts in table 4.2 yields a true prediction 82 (out of 89 cases) and a deviance of 83.7, which is statistically significant.

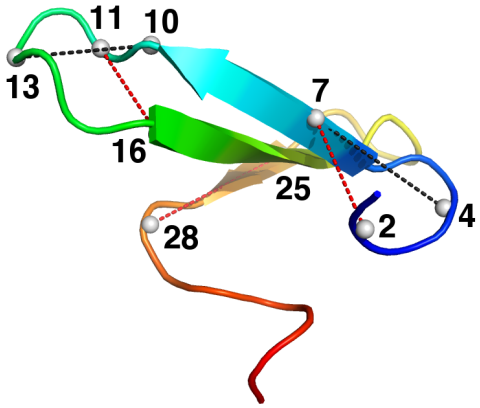


Figure 4.7: Location of five crucial local contacts (dash lines) on 3D crystallographic structure of a representative WW domain (PDB code: 1I5H). The contacts with positive and negative coefficients are colored in red and black respectively.

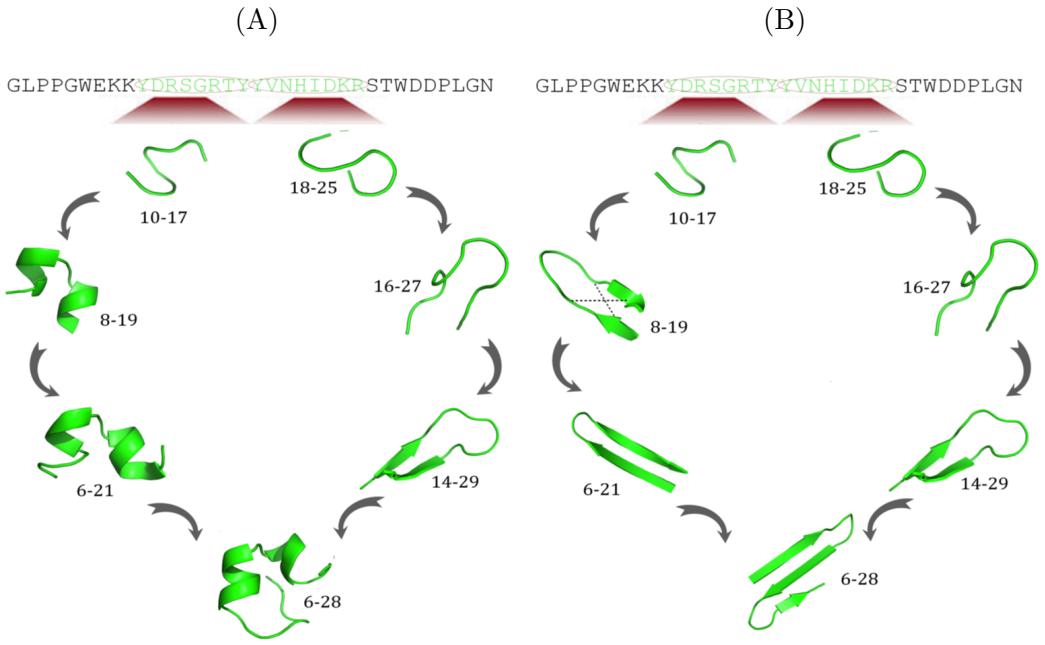


Figure 4.8: (A) The folding pathway of an unfolded WW sequence (CC36) using ZAM. CC36 turns out misfolded in the simulation. (B) Adding constraints to the crucial local contacts helps form the N-terminal hairpin correctly and make this unfolded sequence foldalbe.

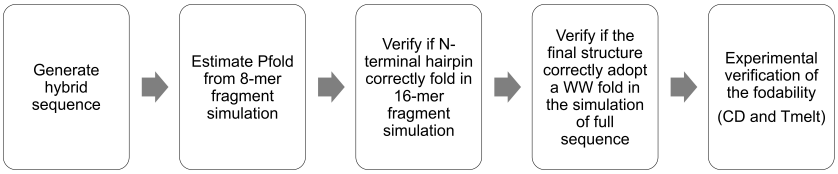


Figure 4.9: A flow chart of analysis for selecting fodlable sequences from hybrid sequences.

4.3.3 Design Foldable WW Domain Sequences

Inspired by the fact that enforcing crucial contacts as restraints in ZAM simulations makes those unfolded sequences foldable, I developed an approach to design new foldable sequences by introducing mutations to unfolded sequences at positions involved in those five crucial contacts. This approach is totally different from the common approach of computational protein design mentioned in Chapter 1, which usually searches for optimal amino acid side chains given a fixed backbone topology (fold) by optimizing the energy, or stability of the native state.

Figure 4.9 shows a flowchart of analysis for screening foldable sequences from the hybrid sequences. So far I have generated 227 hybrid sequences and simulated their 8-mer fragments by ZAM. I re-estimated their Pfold $P(z = 1|x_1, x_2, \cdots x_5)$ using the CPROBs of five crucial contacts using equation (4.2) and selected those with high Pfold values to grow to 16-mer fragments. At 16-mer fragment step, those failing to form N-terminal hairpin were filtered out and the rest were grown to the full sequences. At the end, 11 sequences with correctly folded WW structures (table 4.4) became the foldable sequence candidates and were subjective to experimental verification.

My collaborator, Dr. Ghirlanda's group, have been synthesizing these designed foldable candidates. A few preliminary experimental results have

Table 4.4: Designed foldable candicates and their unfolded templates

Unfolded Template		Foldable Candidate	
Name	Pfold	Name	Pfold
cc16	0.13	cc16_n21	0.72
cc46	0.16	cc46_n39	0.84
cc4	0.36	cc4_n6	0.87
cc46	0.16	cc46_n46	0.89
ic41	0.30	ic41_n6	0.94
ic23	0.07	ic23_n37	0.72
ic13	0.18	ic13_n40	0.91
ic4	0.29	ic4_n39	0.99
cc16	0.13	cc16_n39	0.87
cc46	0.16	cc46_n15	0.53
ic34	0.32	ic34_n15	0.91

been obtained for certain sequences. For example, a designed sequence called CC16_N21 exhibits a strong maximum at 227nm in its circular dichroism (CD) spectra, which is a distinctive quantitative feature of the correct WW fold propensity (figure 4.10(A)). The ellipticity at this wavelength decreased reversibly but not cooperatively when the temperature increases, indicating the sequence is not very stable at room temperature (figure 4.10(B)). However, it have already been a big improvement from the unfolded template CC16. Furthermore, it is also worthwhile to mention that not all natural WW domains are stable. It has been reported that only about 71% WW domains are correctly fold without ligand at 12 °C.

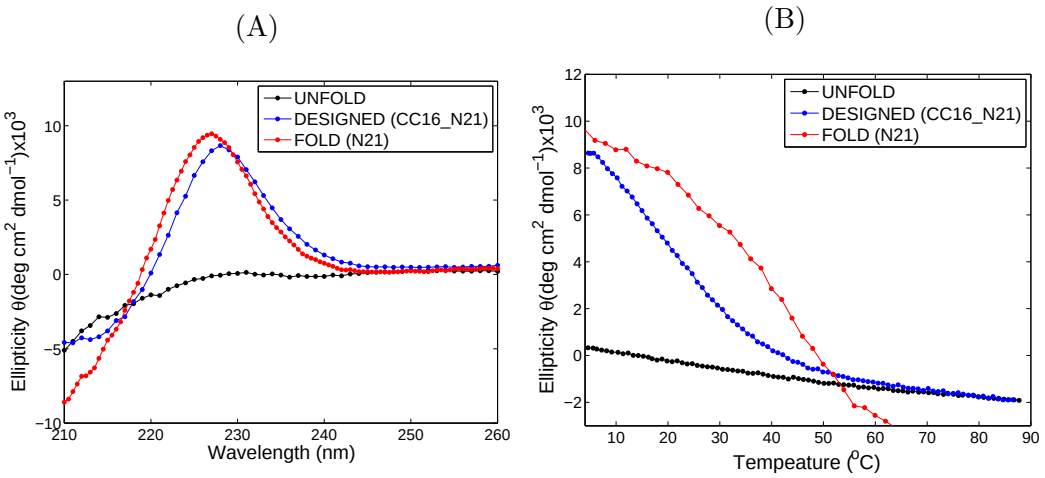


Figure 4.10: (A)CD spectra of CC16_N21 and controls. (B)Thermal denaturation profile of CC16_N21 and controls.

4.4 Conclusion

Earlier studies have shown that the evolutionary information is necessary and sufficient to specify a protein fold. In this chapter, through studying a repertoire of WW domain sequences using ZAM simulation, I tried to reveal the role of the evolutionary information playing on the folding. It turns out that the evolutionary information may affect the formation of local native contacts, which is crucial for the formation of N-terminal hairpin of WW domain. Based on the contact probability of five local contacts at the early stage of simulation (8-mer fragment simulation), I built a classification model which could predict the foldabilty of WW domain sequence with high accuracy. Enforcing the formation of certain local contacts in WW domain also help avoid misfold and lead to correct structure. Moreover, I proposed a novel approach to design foldable WW domain sequence by hybridizing an unfoldable template with a foldable sequence.

Chapter 5

MECHANISTIC INSIGHTS OF PROTEIN EVOLUTION

5.1 Introduction

Proteins are remarkable central machines of living cells which participate in a marvelous range of biological processes. They are not only efficient and robust, but also capable of evolving to acquire new functions and structures. In fact, the evolution of many modern proteins can be traced back to a limited set of common ancestors over millions to billions of years ago. Recently the topic of protein functional evolution attracts more and more attention, as the emergence of drug resistance in bacteria and the acquisition of capacity to degrade new chemicals for enzymes. An fundamental question about the evolution of protein is how a change in protein sequence determines the change of structure and function. In the past many research on protein evolution have focused on the functional alteration as the result of structural change induced by the sequence variation in a static viewpoint of a protein structure. Most attention has been paid to identifying mutations that disrupt or stabilize structures. However, proteins are not static in cellular environment. They have inherent conformational dynamics in the native state which is actually the primary determinant of a protein's function [178, 16]. Thus it is necessary to take into account the dynamics properties in the investigation of protein functional evolution. In fact, a single mutation which induces little structural modification can lead to a large change in conformational dynamics, even at quite distant residues due to structural allostery [130]. This suggests that the

switch of structural dynamics can be achieved by a few mutations without the change of overall structure, which may bring in new functions of the protein under the same structural fold [240, 239, 135]. The functional evolution through the modification of conformational dynamics may be a more common way in nature, since the evolutionary process starts from limited sequences and structure diversity [127].

From a phylogenetic point of view, one may find out a set of critical mutations for the change of protein function through horizontal and vertical approaches [111]. The horizontal approach compares the primary sequence and structure of a given modern protein with other modern proteins which are also leaves in the phylogenetic tree. These proteins belong to the same protein family and exist in the same era but may exhibit divergent functions. Such comparison can help identify mutations that may play important roles in protein function. Experimental site-directed mutagenesis study can also been carried out at those mutational sites to characterize their functional roles. However, this horizontal approach often fails to determine the exact set of critical mutations which may be necessary or sufficient to alter the function, as it ignores enormous evolutionary information encoded in the linkage from the modern protein from its ancestors. In fact, the protein function evolves from their ancestors in a vertical way through the history. It is important to incorporate such vertical evolutionary information in the determination of functional critical mutations. The major challenge in the vertical approach is that the ancestral proteins are no longer exist. Scientist has made significant effort to resurrect those ancestors. With the advances in phylogenetic analysis methods and the development of genomic databases, it becomes possible to reconstruct the sequences of proteins in the ancient using Bayesian and

Maximum Likelihood approaches. Given the reconstructed sequences, several proteins, including opsins [263, 264], GFP-like proteins [88, 243], steroid receptors[35, 186], β-lactamases [206] and others [94, 95, 128, 145, 237], have been synthesized in the laboratory. It provided a basis for the characterization of their structures, stabilities and biochemical functions.

Interestingly, it turns out many modern proteins remain significantly high structural similarity to their ancestors but exhibit divergent functions [206]. This suggests that the function divergence is achieved through the modification conformational dynamics. To test this hypothesis, I investigated the evolution of two protein systems, β-lactamases and GFP-like proteins. I carried out a comprehensive analysis on structures and conformational dynamics of related proteins through multiscale methods consisting of MD simulation at the atomistic level and PRS analysis at the residue level, in order to answer whether structure-encoded conformational dynamics can provide mechanistic insights about the evolution of protein function. The detailed results and discussions about the two protein systems are presented in the subsequent two sections respectively.

5.2 Case I: β-lactamase

5.2.1 Introduction

Antibiotic resistance is one of the most serious threat to public health. When the bacteria is exposed to the antibiotics that threaten its existence, it adopts genetic change rapidly under the powerful selection pressure and acquires new mutations that confer resistance to the drugs. Therefore, the bacteria becomes less and less susceptible to the currently available antibiotics,

whereas the development of new antibiotics becomes more and more difficult and expensive. It is urgent to understand the evolution of antibiotic resistance in order to continually combat bacterial infection [161, 156, 43, 194, 244]. Moreover, the rapid evolution of antibiotic resistance also provides us an ideal system with a lot of sequencing data which allows us to obtain the ancestral resistance gene, to understand the environment it used to inhabit, and learn how it evolved [25].

The central machinery delivering antibiotic resistance is β-lactamase. Since the introduction of penicillin in 1940s, β-lactam antibiotics have became the most popular antibiotic agents which account for about 65% antibiotics consumption across the world [82]. They kill a broad spectrum of bacteria with low toxicity to humans and livestock. The essential nucleus of β-lactam antibiotics is a four-atom ring known as a β-lactam ring. Resistance to β-lactam antibiotics is the result of production of the enzyme β-lactamases [161]. β-lactamases are capable of catalyzing the hydrolysis of β-lactam ring and thus deactivating the antibiotic activity. Because of the great clinical importance of β-lactam antibiotics, β-lactamases have been studied extensively in decades [203, 161, 170, 251, 194]. Hundreds of β-lactamases with different amino acid sequences have been discovered and many of them differ in phenotypes [170, 213, 41]. There are currently two major classification schemes for β-lactamases: Bush's functional classification [40, 41] and Ambler's molecular classification [5]. The former scheme is based on the substrates that the enzymes hydrolyzes and the inhibition of enzyme activity by clavulanic acid [40, 41]. The latter scheme is simpler and more popular, which classifies β-lactamases based on the primary sequences of the enzyme [5]. Presently, four classes have been identified (class A, B C and D). Classes A, C and D are three

classes of serine β-lactamases which utilize serine as a catalytic site. Class B β-lactamases are metallo-β-lactamases which require at least a bivalent metal ion like Z_n^{2+} for activity. Furthermore, the three classes of serine β-lactamases are homologous, i.e., descending from a common ancestor, because they share the similar fold but their sequences are sufficiently different [109].

β-lactamases are ancient enzymes, originating over two billion years (Gyr) ago, and some have been on plasmids for millions of years. In fact, many divergent and ancient reisistance genes have also been found in the antibiotic-free environment, such as Alaskan soil [4], sediments from the bottom of pacific ocean [242] and even 30,000-year-old Beringian permafrost sediments [60]. To understand the evolution of β-lactamases as well as antibiotic resistance, my collaborators (Dr. Jose Manuel Sanchez-Ruiz *et al*) have resurrected a series of ancestral Class A β-lactamases in the laboratory, including the last common ancestor of enterobacteria (ENCA), the last common ancestor of gamma-proteobacteria (GPBA), the last common ancestor of various Gram-negative bacteria (GNCA) and the last common ancestor of Gram-positive and Gram-negative bacteria (PNCA) [206]. Those ancestors used to exist on Earth about 1 Gyr (ENCA), 1.5 Gyr (GPBCA), 2 Gyr (GNCA) and 3 Gyr (PNCA) based on the estimates of divergence times. The protein sequences of those ancestors were derived through Bayesian Maximum Likelihood approach in a phylogenetic framework, targeting the Precambrian nodes in the evolution of Class A β-lactamases. The sequence identities of these ancestral proteins range from 53% to 79% in pairwise alignments with TEM-1 β-lactamases, one of their modern decedents. However, despite the significant variations in sequence, they adopt similar conformation to TEM-1 β-lactamase, especially at the active sites. More interestingly, those 2-3 Gyr-old β-lactamases are highly stable

with melting temperate (T_m) about 35 degrees higher than the modern one. Additionally, they can degrade a variety of antibiotics *in vitro* with similar levels of catalytic efficiency, while the modern β-lactamases have remarkable susceptibility bias to specific antibiotics (benzylpenicillin), suggesting that β-lactamases evolve from substrate-promiscuous generalists to specialists [206].

In summary, despited the fact that those 2-3 Gyr-old β-lactamases have distinct thermal profile and enhanced substrate-promiscuity compared with the modern decedents, they have very similar structures. This prompts the question whether structure-encoded conformational dynamics can provide mechanistic insights about the evolution of β-lactamases. From the required necessary flexibility of a ligand-binding site to the conformational transitions of allosteric proteins, proteins must fluctuate to function. Indeed, previous structural dynamics analysis on ancestral steroid receptors [99] has shown that inherent structural dynamics is crucial to give a more complete understanding of protein evolution. Thus I ask here whether a comprehensive conformational dynamics analysis can elucidate how β-lactamases evolve to function as specialists from their generalists ancestors. To this aim, the three ancestral β-lactamase (PNCA, GNCA and ENCA) and a modern decedent (TEM-1) are simulated using reservoir replica-exchange molecular dynamics (r-REMD), an efficient simulation technique incorporating geometric simulation with REMD algorithm [208]. The simulations provide us the dynamic information of those β-lactamases. The analysis of residue fluctuations indicates that the ancient lactamases are more flexible than TEM-1 lactamase. Moreover, to explore how their conformational dynamics alters from the unbound conformation upon approaching ligand (i.e., β-lactam), I performed PRS analysis on each β-lactamase. PRS relies on sequentially applying an

externally random force (i.e., random Brownian kick) on a single residue and record the response of other residues [13, 14, 96]. After PRS, a metric called Dynamic Flexibility Index (dfi) is introduced to measure the resilience of individual site to perturbations [176]. These perturbations indeed mimic nature, since protein is exposed to many random forces as a first-order approximation in a crowded cell while interacting with other proteins or ligand such as β-lactam as in the case of β-lactamase. Overall, the conformational dynamics of individual β-lacamteses shows changes in the conformational dynamics are in agreement with the functional divergence: while the dfi distributions of PNCA and GNCA lactamase are similar to each other and distinctively separated from the functionally divergent TEM-1 lactamase, ENCA lactamase (the most substrate-specific ancient β-lactamase) shows more similar dfi distribution to TEM-1 lactamase. Moreover, TEM-1 lactamase has a more rigid catalytic pocket, suggesting that the shape of the pocket has also evolved towards a specific target, as the catalysis becomes benzylpenicillin-specific. I also analyzed the statistical pattern of their dynamics profiles using Single Value Decomposition (SVD), which enhance the signal-to-noise ratio of the data by expressing them as a linear combination of a few dominant principal components. On the basis of their pairwise distances in the subspace of principal components, a cladogram is constructed to illustrate the evolutionary relationship of these β-lactamase in terms of dynamics. Furthermore, through SVD, I identified the residue sites which are critical for the dynamic divergence among β-lactamse. Changes of the dynamics at those critical sites may lead to the change of the substrate-specificity of the protein. In summary, the findings suggests that the change in structural dynamics best explains the

evolution of catalytic function in β-lactamases and the analysis of the detailed conformational dynamics could help us understand the underlying mechanism.

5.2.2 Methods

Structure Refinement and Simulation

The refinement and equilibrium sampling of ancestral β-lactamases is accomplished with reservoir REMD (r-REMD) [208]. REMD samples the system by molecular dynamics at different temperatures (replicas) and allow the system to attempt exchange between replicas [232]. By doing so, systems at high temperature might overcome potential energy barriers and explore a large volume of configuration space. To further improve sampling efficiency, a structure reservior is prepared and coupled with REMD. The system at the highest temperate replica is also allowed to exchange the configuration with the reservior structure periodically. To prepare the structural configuration in the reservoir, I generated the unfolding pathway of the protein under external pulling forces using a highly efficient geometric based sampling technique called FRODAN [257], which decomposes a protein into a set of small rigid units and models the interactions as harmonic constraints. Then I clustered the unfolding trajectory and use a large ensemble of partially unfolded configurations as reservoir structures.

In detail, I first ran restrained simulation for 1.5 ns with 40 replicas from 270K to 450K in the AMBER96 force field [189] with generalized born implicit solvent model [184]. The residue-residue pairs are constrained if their C_α atoms are within 8.0 Å cutoff distance among 90% of reservoir structures. The residue-residue constraints are applied at the C_α atoms of the residue

and the force constant is 0.5 kcal $\cdot$ mol^{-1} $\cdot$ Å^{-2}. After the restrained run, an unrestrained r-REMD with identical parameters continues for at least 5 ns. A convergence analysis is performed by the evaluating the correlation between the covariance matrices of different windows using MD trajectory. The simulation is considered to be converged when the correlation is above 0.80. If the analysis indicates that the simulation has converged, then no further simulation is performed. If it is not, an additional 2 ns of simulation is run and the convergence is re-examined. Such process is repeated until it is converged.

Perturbation Scanning Response and Dynamic Flexiblity Index (dfi)

The canonical PRS model is based on Elastic Network Model (ENM), where the protein is viewed as an elastic network (section 2.6.2 on page 42). A disadvantage of the ENM-based PRS model is that the coarse-grained network makes it insensitive to changes arising from the biochemical specificity of amino acids. Therefore, in order to compare the ancestral β-lactamases with similar backbone structures, I replaced the ENM basis of PRS with all-atom r-REMD simulations, where the inverse of Hessian matrix was substituted by the covariance matrix $[G]$ derived from the MD trajectory equation (2.56). MD simulations take into account long-range interactions as well as the biochemical specificity of amino acids. Thus incorporating MD allows PRS to provide more insights about specific residues beyond the scope of the canonical PRS.

The metric by which PRS quantifies the flexibility of a residue upon the perturbation of other residues is called Dynamic Flexibility Index (dfi). To compute dfi, a unit external force is applied on a single residue. The response vector of positional displacement $\mathbf{\Delta R}$ is computed by equation 2.56. To en-

sure the isotropicity of perturbation, the perturbations are attempted in ten directions and then the average of response vectors is computed. The pertrubations are repeated for each residue site and one can obtain the perturbation matrix which records the displacement for each residue upon the perturbation of the other residues like

$$[\mathbf{A}]_{N\times N} = \begin{bmatrix} |\Delta R^1|_1 & |\Delta R^2|_1 & \cdots & |\Delta R^N|_1 \\ |\Delta R^1|_2 & |\Delta R^2|_2 & \cdots & |\Delta R^N|_2 \\ \vdots & \vdots & \ddots & \vdots \\ |\Delta R^1|_{N-1} & |\Delta R^2|_{N-1} & \cdots & |\Delta R^N|_{N-1} \\ |\Delta R^1|_N & |\Delta R^2|_N & \cdots & |\Delta R^N|_N \end{bmatrix} \tag{5.1}$$

where $|\Delta R^j|_i = \sqrt{\langle \Delta R^2 \rangle}$ denotes the magnitude of the displacement by residue i in response to the perturbation at residue j. Each row presents the average displacement of a specific residue from its equilibrium position upon perturbation of the remaining residues one at a time, while each column represents the response profile of each of the remaining residues upon perturbation of one specific residue. dfi is defined as the total displacement of residue i induced by perturbation of the remainder of the residues in the protein, i.e., the sum of elements in row i of the matrix above, normalized by the total displacement of all residues in the protein:

$$dfi(i) = \frac{\sum_{j=1}^{N} |\Delta R^j|_i}{\sum_{i=1}^{N}\sum_{j=1}^{N} |\Delta R^j|_i} \tag{5.2}$$

Singular Value Decomposition for Clustering and Identifying Functionally Important Dynamics

SVD analysis is used here to classify β-lactamases by examining their dynamics profiles (i.e., dfi values) at different residue sites, following the standard

procedure present in section 2.6.2 (page 42). The subjects of the study is the four β-lactamases and the attributes are the dfi values. To accomodate β-lactamases with varied length, I focus on the 262 residue sites where each β-lactamases has a residue pesent (i.e., not a gap) in multiple sequence alignment. Therefore in this application of SVD to β-lactamases, each column of $\mathbf{X}$, conventionally denoted as $\mathbf{x}_j$, is a 262-dimensional vector describing the dynamics profile at those residue sites of a given β-lactamases i ($1 \leq j \leq 4$). I move the origin to mean of the data by subtracting the mean of row i from each element x_{ij}. The resulting $\mathbf{X}$ matrix eliminates the generic characteristics of particular residue sites and emphasizes more clearly the differences among dfi patterns of those β-lactamases. As I wish to understand the relationship of the β-lactamases, the signal of interest in this case is the dynamics profile $\mathbf{x}_j$ of β-lactamases i. The dynamic profiles are transformed to the left-singular subspace through SVD, where the modern and ancestral β-lactamases are represented emphasizing their differences. The pairwise distance of β-lactamases in the subspace reveals their dynamics similarities and differences. Moreover, the residue sites with high weight which have significant contribution to the top left-singular vectors may account for the major dynamic differences among those β-lactamases. The mutation occurred at those sites may have a large impact on the protein dynamics.

5.2.3 Results

Minor Structural Change in Evolution

The sequences of ENCA, GNCA and PNCA lactamases differ from a modern decedent β-lactamase, where the sequence identity is 79.2%, 53.6% and

52.9% with respect to TEM-1. Despite of the extensive sequence differences, they all share the canonical β-lactamase fold with all-atom RMSDs of 0.53 Å, 0.76 Å and 0.86 Å with respect to the TEM-1 lactamase (figure 5.1(A)). The structure of ENCA lactamase is more similar to the TEM-1 lactamase than those of GNCA and PNCA lactamases as in sequence. Closer inspection of RMSD at individual residue sites reveals minor movement with RMSD$\leq$ 2Å in the $\alpha + \beta$ domains of the GNCA and PNCA lactamases corresponding to N-terminal helix and solvent-exposed loops (figure 5.1(C)). Moreover, no substantial difference are found in the α-domain and all active sites occupying canonical space (figure 5.1(B)). Therefore, the structural analysis is not sensitive enough to address the cause of the functional divergence, i.e., how the β-lactamases evolve from substrate-promiscuous generalists to specialists.

Structural Dynamics Related with Functional Divergence

Here I turn to investigate the role of structural dynamics on functional divergence observed among the β-lactamase. The unbound conformations of the three ancestral β-lactamase (PNCA, GNCA and ENCA) and a modern decedent (TEM-1) are simulated using reservoir replica-exchange molecular dynamics (r-REMD). r-REMD incorporates the conformations generated by the geometric simulation algorithm FRODA as reservoir structures, which in turn increases the efficiency of conformation sampling [99]. I first analyzed the root mean square fluctuation (RMSF) of residues for each β-lactamase. RMSF is a measure of the positional deviation of a residue over time from its time-averaged position. While the structural analysis does not show any differences, the RMSF profiles slightly do. Indeed, the ancient lactamases (PNCA, GNCA and ENCA) fluctuate a little more than TEM-1 lactamase. More in-

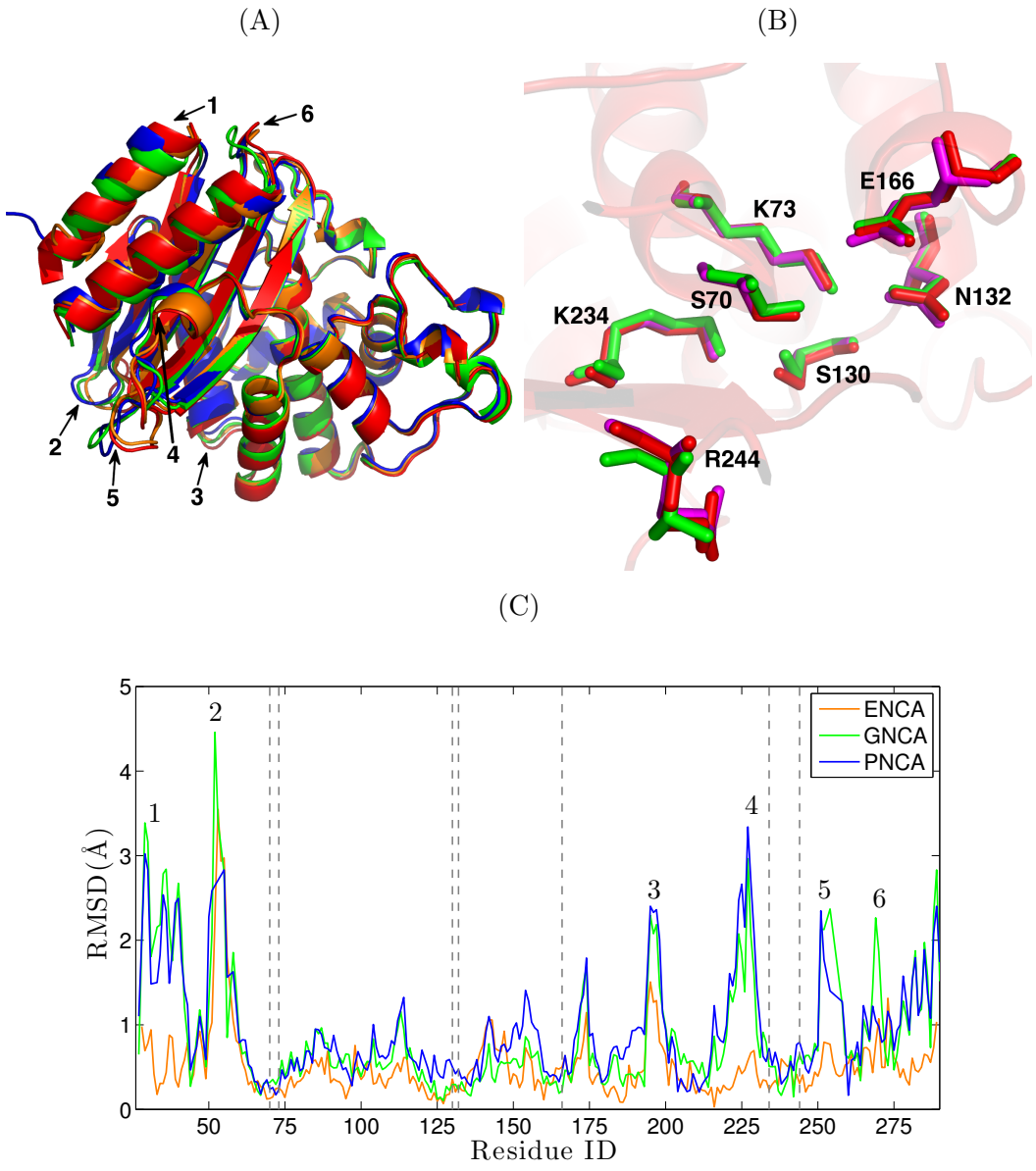


Figure 5.1: Structural characterization of laboratory resurrections of Precambrian β-lactamases. (A) Structural comparison of the TEM-1 β-lactamases (PDB: 1BTL; red), the last common ancestor of enterobacteria (ENCA; PDB: 3ZDJ; orange), the last common ancestor of various Gram-negative bacteria (GNCA; PDB: 4B88; green) and the last common ancestor of Gram-positive and Gram-negative bacteria (PNCA; blue). (B) Close examination of the structural differences at active sites. (C) RMSD of individual residue site along the sequence. The vertical dash lines mark the location of active sites. Minor structural differences are seen in the N-terminal helix and solvent-exposed loops (labeled 1 to 6).

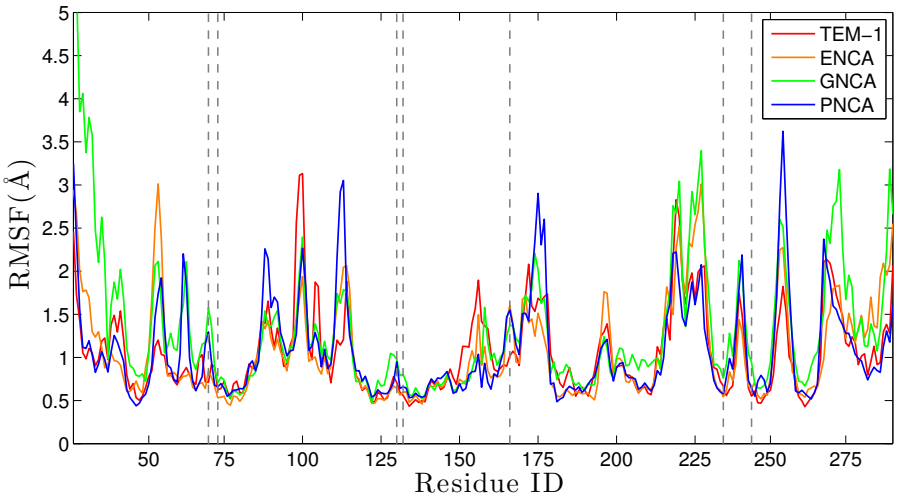


Figure 5.2: The root mean square fluctutaion of C_α atoms in TEM-1 (red), ENCA (orange; 1 Gyr), GNCA (green; 2 Gyr) and PNCA (blue; 3 Gyr). The vertical dash lines mark the location of active sites.

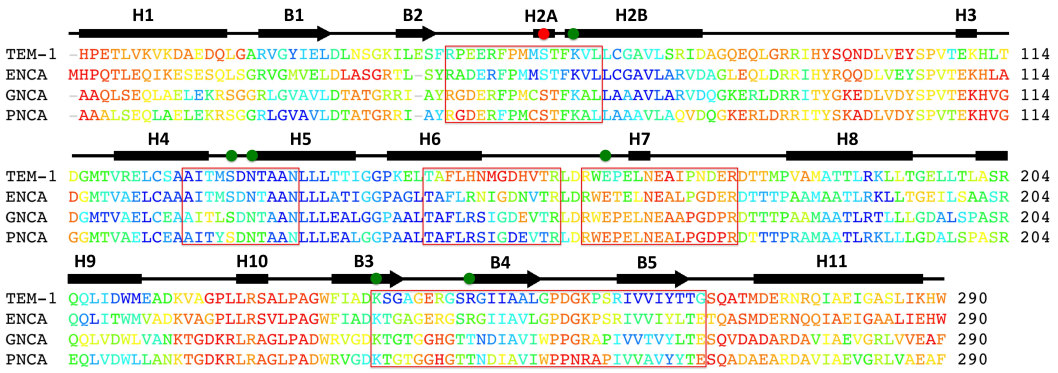
terestingly, at the primary active site S70, PNCA and GNCA lactamase show significantly higher flucatuation (i.e., more flexible) than ENCA and TEM-1 lactamase (figure 5.2).

As RMSF profiles present the structural dynamics of β-lactamase extracted from their equilibrated unbound conformational dynamics, I am more interested in capturing the dynamics profiles of each position as they deviate from unbound equilibrium such as their response to an approaching ligand as it exerts forces on the protein. For this purpose, I applied PRS approach. With PRS, I introduced perturbations by applying a random external unit force on the selected single residues, and then analyze the residue response fluctuation profile of the rest of the chain using linear response theory. It has been shown in the past that PRS and its derivative are powerful tools to i) capture conformational changes upon binding [13, 14]; ii) reveal allosteric pathways and identify critical residues that mediate long-range communica-

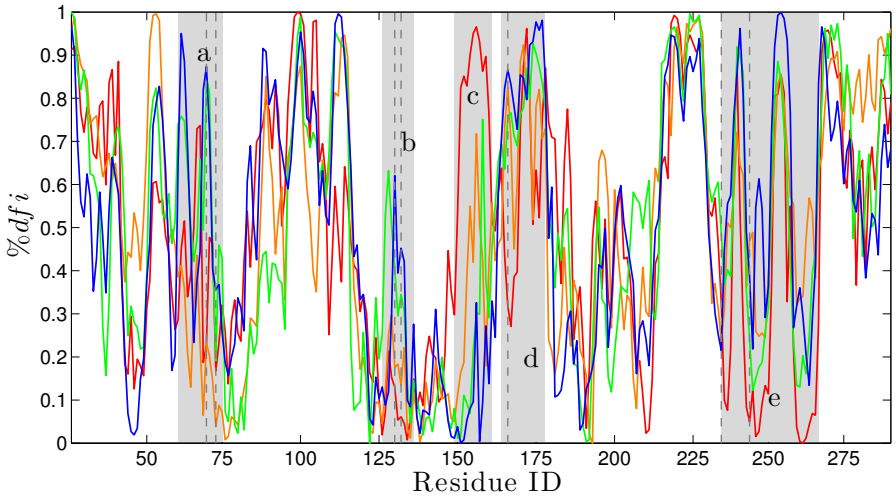
tion [96]; iii) generate an ensemble of configuration rapidly for flexible docking that improves binding affinity score [33]; iv) distinguish disease-associated and putatively neutral population variations in human proteome [176]. PRS mimics the natural process of binding as a first-order approximation by exerting a Brownian kick on a single residue in the unbound conformation without presence of the ligand and then computes the fluctuation response of the rest of the residues (both direction and magnitude) through linear response theory [121]. To ensure the isotropicity of perturbation, the Brownian kick is applied at ten different directions for individual site one at a time. The magnitude of displacement by residue i in respond to the perturbation at residue j is given by the mean square fluctuation $|\Delta R^i|_j$. Then the perturbation is repeated at all other residues and and dynamics flexibility index dfi is normalized average mean square fluctuation of a site upon perturbations of others as shown in equation 5.2.

As such, the dfi value provides a measure of the resilience of individual sites to perturbation by events such as residue substitution. A high dfi value is associated with a high degree of structural flexibility in response to perturbation elsewhere, whereas a low dfi value indicates that this site may transfer the perturbation energy to its surroundings. Low dfi values tend to involve hinge regions that may control motions like joints in skeleton. Therefore, the dfi could evaluate the contribution of each site to the functionally important dynamics. To eliminate the effect of the global flexibility of difference proteins, I here compute the rank of the dfi profile and label it as $\%dfi$. Figure 5.3(A)-(C) compares the $\%dfi$ profile in each β-lactamase. Visual observation identifies five regions (a to e) showing significant flexibility discrepancy among the β-lactamases (figure 5.3(B)): i) Region a (residues 61-75) consists of part of helix

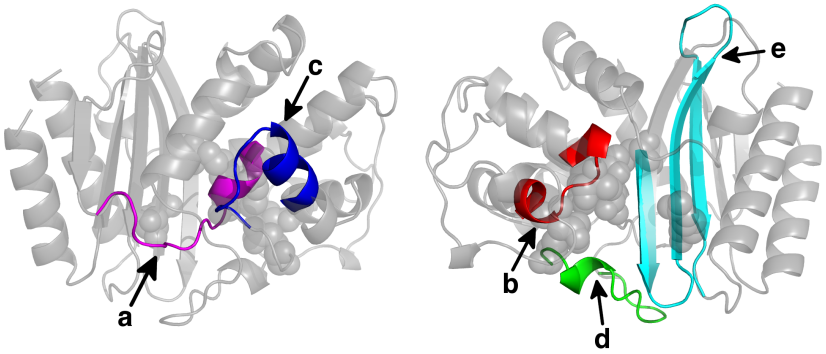
(A)


H1
B1
B2
H2A
H2B
H3
TEM-1 -HPETLVKVKDAEDQLGARVGYIELDLNSGKILESFRPEERFPMMSTFKVLLCGAVLSRIDAGQEQLGRRIHYSQNDLVEYSPVTEKHLT 114
ENCA MHPQTLEQIKESESQLSGRVGMVELDLASGRTL-SYRADERFPMMSTFKVLLCGAVLARVDAGLEQLDRRIHYRQQDLVEYSPVTEKHLA 114
GNCA -AAQLSEQLAELEKRSGGRLGVAVLDTATGRRI-AYRGDERFPMCSTFKALLAAAVLARVDQGKERLDRRITYGKEDLVDYSPVTEKHVG 114
PNCA -AAALSEQLAELEKRSGGRLGVAVLDTATGRRI-AYRGDERFPMCSTFKALLAAAVLAQVDQGKERLDRRITYSKADLVDYSPVTEKHVG 114
H4
H5
H6
H7
H8
TEM-1 DGMTVRELCSAAITMSDNTAANLLLTTIGGPKELTAFLHNMGDHVTRLDRWEPELNEAIPNDERDTTMPVAMATTLRKLLTGELLTLASR 204
ENCA DGMTVAELCAAAITMSDNTAANLLLATIGGPAGLTAFLRNIGDNVTRLDRWETELNEALPGDERDTTTPAAMAATLRKLLTGEILSAASR 204
GNCA DGMTVAELCEAAITLSDNTAANLLLEALGGPAALTAFLRSIGDEVTRLDRWEPELNEAAPGDPRDTTTPAAMAATLRTLLLGDALSPASR 204
PNCA GGMTVAELCEAAITYSDNTAANLLLEALGGPAALTAFLRSIGDEVTRLDRWEPELNEALPGDPRDTTTPRAMAATLRKLLLGDALSPASR 204
H9
H10
B3
B4
B5
H11
TEM-1 QQLIDWMEADKVAGPLLRSALPAGWFIADKSGAGERGSRGIIAALGPDGKPSRIVVIYTTGSQATMDERNRQIAEIGASLIKHW 290
ENCA QQLITWMVADKVAGPLLRSVLPAGWFIADKTGAGERGSRGIIAVLGPDGKPSRIVVIYLTETQASMDERNQQIAEIGAALIEHW 290
GNCA QQLVDWLVANKTGDKRLRAGLPADWRVGDKTGTGGHGTTNDIAVIWPPGRAPIVVTVYLTESQVDADARDAVIAEVGRLVVEAF 290
PNCA EQLVDWLLANKTGDKRLRAGLPADWRVGDKTGTGGHGTTNDIAVIWPPNRAPIVVAVYYTESQADAEARDAVIAEVGRLVAEAF 290

(B)


%dfi
Residue ID
a
b
c
d
e
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
50
75
100
125
150
175
200
225
250
275

(C)

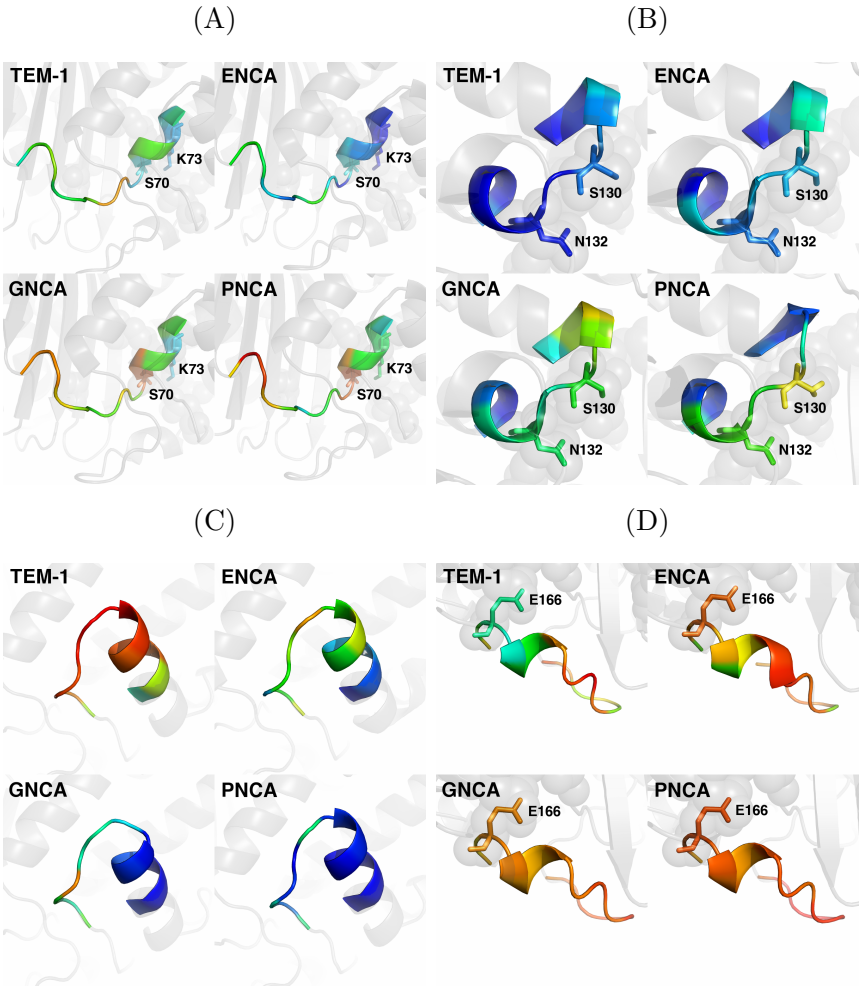

a
b
c
d
e

H2 and a loop region between strand B2 and helix H2. ii) Region b (residues 126-136) includes a loop region between H4 and H5. iii) Region c (residues 149-161) consists of part of helix H6 and a loop region between helix H6 and Ω-loop. iv) Region d (residues 164-178) is mostly the Ω loop (residues 163-178). v) Region e (residues 234-267) spans from strands B3 to B5. The dynamics and structual details of those regions in close investigation are provided in figure 5.4. The overall flexibility increases from TEM-1 to PNCA lactamase in those regions except the region c, where the trend becomes exactly opposite. I also notice that the Ω loop, which is important for substrate recognition and catalysis, become more flexibile from TEM-1 to PNCA lactamase. Furthermore, it is worthwhile to compare the dynamics at the active sites participating in catalysis. The catalytic mechanism of class A β-lactamase involves the acylation of the active site S70, followed by deacylation. During this process, a general base is expected to activate the primary catalytic site S70 by accepting the proton from it [149, 231]. Although whether the identification of the

---

Figure 5.3 *(preceding page)*: The dynamcis profile (flexibility) of residues in TEM-1 (modern), ENCA (1 Gyr), GNCA (2 Gyr) and PNCA (3Gyr). (A) The $\%dfi$ index is mapped onto the multiple sequence alignment of the four β-lactamases. Residues are colored with a spectrum of red to blue, where rigid residues are denoted by blue/green and flexible regions are denoted with red/orange. The primary active site S70 are marked by red dot and other active sites are marked by green dots. Five regions where the β-lactamases show high discrepancy by visual obervation are marked in red box (region a: residues 61-75; region b: residues 126-136; region c: residues 149-161; region d: residues 164-178; region e: residues 234-267). (C) The $\%dfi$ distribution in the four β-lactamases: TEM-1 (red), ENCA (orange), GNCA (green) and PNCA (blue). The vertical dash lines mark the location of active sites. The five regions with high discrepancy are marked in grey shadow. (B) Mapping the five regions (a to e) with significant flexibility difference among β-lactamase to the structure. The active sites are displayed in spheres. The dynamics and structual details of those regions are shown in figure 5.4

general base is K73 or E66 remains controversial, it is well known that the two sites are critical in this proton transfer event [45, 59, 11]. In addition, serveral residues, such as S130, N132, K234, R244 (K244 in PNCA/GNCA), are also identified as important active sites for catalysis [149, 65, 11]. According to the $\%dfi$ profile, TEM-1 and ENCA lactamases get more flexible than PNCA/GNCA at those active sites as well. In summary, I observe that the PNCA and GNCA lactamases show higher flexibility than ENCA and TEM-1 lactamases in four regions (regions a-b and d-e). The four regions span the active sites and nearby sites, indicating that compared to specialists ENCA and TEM-1 lactamases, the generalists ancestors, PNCA and GNCA lactamases, have flexible pockets which accommodate necessary biochemistry to inhibit various antibiotics (figure 5.4(F)). This finding also suggests the increase of catalytic specificity in the modern β-lactamase evolove through the decrease of flexibility in the catalytic pocket as observed earlier for the evolution of stress hormone receptor [99]. PNCA and GNCA lactamases have more flexibility around the active sites and thus higher catalytic promiscuity, while TEM-1 and ENCA lactamases are less flexible and thus more substrate-specific. Interestingly, the last region shows an exactly opposite trend, where the flexibility of TEM-1 is higher and there is trend of decreasing flexibility from TEM-1 to PNCA. This trend can be explained due to fact that the loss in the flexibility of catalytic pocket of specialist is compensated by the gain of the flexibility of the region c.


(A)
(B)
TEM-1
ENCA
GNCA
PNCA
K73
S70
S130
N132
(C)
(D)
E166
(E)
(F)
K234
R244
T244
a
b
d
e

Clustering Proteins Based on Dynamics Profile and Identify Potentially Function Altering Mutations

Despite the conserved structures of these four β-lactamases, the experimental characterization has showns that their antibiotic catalytic patterns are different, where PNCA and GNCA shows promiscuous trend, responding to different antibiotics, and TEM-1 and ENCA lactamases are more specific. In order to understand whether the underlying structural dynamics of these four proteins can indicate the functional differences, I performed SVD and cluster the statistical pattern of dfi profiles of the four β-lactamases. To accommodate β-lactamases with varied length, I selected 262 residue sites where each β-lactamase has a residue present (i.e., not a gap) in multiple sequence alignment. According to the pairwise distances in the left subspace of SVD (figure 5.5(A)), a cladogram for clusters is constructed in figure 5.5(B). In-

---

Figure 5.4 *(preceding page)*: Close investigation of five regions with significant flexibility difference among β-lactamase. These regions are colored with a spectrum of red to blue, where rigid regions are denoted by blue/green and flexible regions are denoted with red/orange. The active sites are displayed in sticks. (A). In Region a (residues 61-75), TEM-1 and ENCA lactamases are less flexible than GNCA and PNCA lactamases, especially at the active sites S70 and K73. (B). In Region b (residues 126-136) , TEM-1 and ENCA lactamases are also more rigid than GNCA and PNCA lactamases, especially at the active sites S130 and N132. (C). The trend of dynamics profiles in Region c (residues 149-161) is opposite to the other four regions, where the flexbility decreases from TEM-1 to PNCA lactamase. (D). Region d (residues 164 to 178) spans the Ω-loop. It becomes more flexible from TEM-1 to PNCA lactamase. The active site E166 is more rigid in TEM-1 lactamase then the others. (E). Region e (residues 234-269) gets more flexible from TEM-1 to PNCA lactamase and TEM-1 lactamase is significantly rigid than the other three proteins in this region. (F). The catalytic pocket, surrounded by regions a-b and d-e, exhibits overall increased flexibility from the specialists (ENCA and TEM1 lactamases) to the ancestral generalists (PNCA and GNCA lactamases).

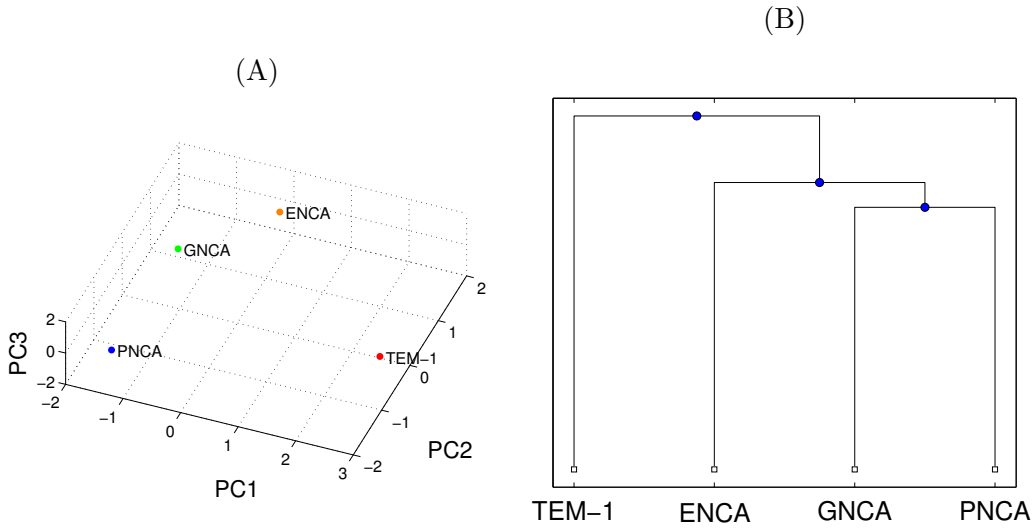


Figure 5.5: Clustering of β-lactamase based on the dynamics profiles. (A) Distribution of β-lactamase in the subspace formed by top three left-singular vectors (principal compont or PC). (B) Cladogram of SVD distances for β-lactamases determined from their dfi data at 262 residue sites.

teretingly, the β-lactamase are separated into two major branches, with one branch consisting only TEM-1 lactamase (the most modern one), and the remaining three lactamases appearing in the other branch. The larger branch found in this analysis is divided into two sub-branches separating ENCA from PNCA and GNCA lactamases. This result shows that PNCA and GNCA lactamases are very similar to each other but further separated from TEM-1 lactamase based on their dynamic characterization.

SVD analysis also enables us to identify residue sites which are critical for the structural dynamics divergence. The weight of each residue site, given by its contribution in the top principal components, measures how important the site is to distinguish the dynamic difference of β-lactamases (see Method). Figure 5.6 displays the weights of all residue sites and marks statistically critical sites with large weights. Since those sites are important for the protein

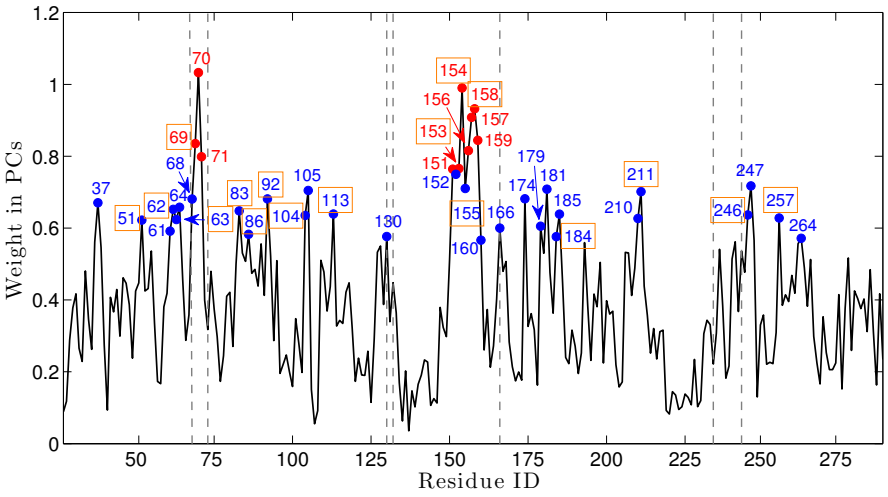


Figure 5.6: Weights of residue sites based on their contribution in the top principal components. The sites whose weights deviate more than twice of standard deviation and one standard deviation from the mean are labeled in red and blue. The sites where the residue types are not consistent (mutational sites) in the four β-lactamases are marked in box.

dynamics and suggestign that change of the dynamics leads to functional divergence, it is likely that mutations at those sites would lead to the change of the substrate-specificity.

5.3 Case II: GFP-like protein

5.3.1 Introduction

Most natural green fluorescent protein like (GFP-like) proteins exhibit bring fluorescence within distinct color classes such as cyan, green and red when exposed to ordinary day light [204, 205]. A large number of GFP-like protein family members have been identified in reef-building (stony) corals (class *Anthozoa*, order *Scleractinia*), where three clades have been classified as clades B, C and D based on a phylogenetic analysis of the protein family tree [3]. Clade D, in which extant cyan, green, and red-fluorescent proteins are well-

represented, provides a large sequence space for further analysis and inference. However, in this clade, the red fluorescent proteins only consist of green-to-red (G/R) photoconvertible FPs (Kaede-type) [249], where the red-fluorescent proteins of the DsRed-type appear absent [243]. The G/R photoconvertible proteins can undergo a light dependent color conversion process, in which the green form (emission at $\sim$ 518 nm) is irreversibly modified to a red-fluorescing form (emission at $\sim$ 582 nm) upon exposure to UV or violet light [8].

A series of ancestral GFP-like proteins in Clade D have been reconstructed to study the evolution of photoconvertible competency [137]. One of them called Least Evolved Ancestor (LEA), which includes a total of 12 residue modifications (11 residue substitutions and one deletion) on a reconstructed ancestral green-fluorescent protein, ALL-GFP, demonstrates a photoconversion efficiency equal to that of extant G/R photoconvertible FPs [88, 137]. The set of modifications present in LEA encompasses six residues with internal side chains that cluster near the chromophore [A60(63)V, Q62(65)H, T69(72)A, S105(110)N, Y116(121)N and V157(165)I] [88], two residues, T104(109)R and R194(204)C, with their side chain located within the antiparallel subunit interface of the tetrameric assembly, one residue with exterior side chains that do not partake in subunit-subunit interactions [M154(162)T], and three residues that are located along the C-terminal tail [R216(227)H, delY217(228), M218/219(229)G] [137]. Here the residues are labeled as the actual residue IDs in the PDB files of resurrected proteins, while the corresponding residue IDs in the conventional GFP numbering system are given in parentheses.

Here I try to understand the mechanism of the G/R color evolution of GFP-like protein, i.e., why and how those residue modifications endow the LEA with photoconvertible competency. LEA shows high structure similarity

to other ancestral green-fluorescent protein. The all atom RMSD between LEA and ALL-Q62H, an ancestral green-fluorescent protein with Q62(65)H mutation but also green phenotype, is only 0.54 Å. The G/R color evolution in GFP-like proteins is likely achieved through the change in conformational dynamics caused by those residue modifications. Therefore, I performed PRS analysis in together with MD simulations on LEA and ALL-Q62H, one with G/R phenotype and the other with green phenotype, in order to obtain the global protein breathing motions and the local structural flexibility around the chromophore. I find that LEA exhibits increased dynamics of chromophore and surrounding regions, agrees with its photoconvertible competency. For convenience, only the numbering system in the PDB files of the resurrected proteins is used below.

5.3.2 Methods

Parameterization and Structure Simulation

Molecular dynamics simulations were performed on the All-Q62H and LEA protein structures using the NAMD package [192] with the AMBER FF99SB force field [118] and a Generalized Born implicit solvent model [184]. The initial configurations were obtained from the X-ray structures of the tetrameric assemblies of All-Q62H and LEA (PDB entries 4DXM and 4DXN). After energy minimization for 100,000 steps using conjugate gradient and line search algorithms, the tetrameric structure of these fluorescent proteins were subjected to 220 ns Langevin dynamics simulations at 300 K while keeping the volume constant. The Langevin equations of motion are integrated using the Verlet form of the BBK integrator equation (2.24) [38]. Using the SHAKE

algorithm, constraints on bond stretching were applied to all bonds involving hydrogen atoms. Therefore, a larger time step of 2 fs was employed in the integration. The non-bonded van der Waals interactions were truncated at 16 Å with a switching function that gradually reduces the van der Waals potential to 0. Non-bonded electrostatic interactions were evaluated through a multiple time stepping scheme by calculating the energy every 10 steps for short-range and every 20 steps for long-range electrostatic interactions (cutoff distance 16 Å).

In All-Q62H and LEA, the chromophore is formed from His62, Tyr63 and Gly64. Geometry optimization of the chromophore in its ground state was performed through a quantum chemical calculation employing a restricted Hartree-Fock method and the 6-31G(d) basis set using GAMESS [217]. The atomic charges of the chromophore were calculated by fitting the molecular electrostatic potential of the molecular surface using the RED package [77]. All force field parameters of the imidazolinone ring and briding atoms, such as bond length, angles, dihedrals and related force constants, were obtained from the work of Xu *et al* [261]. For the chromophore's phenolic group, I used the parameters for a tyrosine phenolate, since the crystal structures represent the anionic form of the chromophore. The parameters for His61 were taken from the AMBER force field [118].

Allostric Response Ratio

The PRS approach can also be used to measure the effects of a collective perturbation of a subset of residues. The allosteric response ratio (χ) is defined as the ratio of the displacement of residue i upon sequential perturbation of a group of selected residues, divided by its total displacement upon sequential

perturbation of all other residues. This expression provides a quantitative measure of the sensitivity of residue i to perturbation of a particular subset of residues formulated as

$$\chi_i = \frac{\sum_{j=k_1}^{k_M} |\Delta R^j|_i}{M} \Bigg/ \frac{\sum_{j=1}^{N} |\Delta R^j|_i}{N-1} \tag{5.3}$$

where k_1, k_2,$\cdots$, k_M are the indices of M selected residues. The higher the value of χ , the more significant is the response of residue i due to perturbation of a specific subset of other residues, and vice versa.

5.3.3 Results

The Photoconversion-competent Chromophore Exhibiting Increased Flexibility

Here I turn to investigate the role of structural dynamics of GFPs on their fluorescent function. MD methods were used to simulate protein motion for a 220 ns time period, while keeping the chromophores in their ground electronic states. As both LEA and ALL-Q62H are equilibrated at about 100 ns, I took the last 100 ns trajectory to perform dynamics analysis. Using a 5 ns sliding window, the average RMSF was calculated for the four chromophores in each tetramer (figure 5.7). The positional deviations from their time-averaged positions are significantly higher in LEA than in ALL-Q62H, suggesting that the LEA chromophores exhibit increased dynamic motion. This result is independent of window size or averaging of individual chromophores in the tetramer. Therefore, the MD results suggest that non-planar chromophore conformations are more accessible to LEA than to ALL-Q62H.

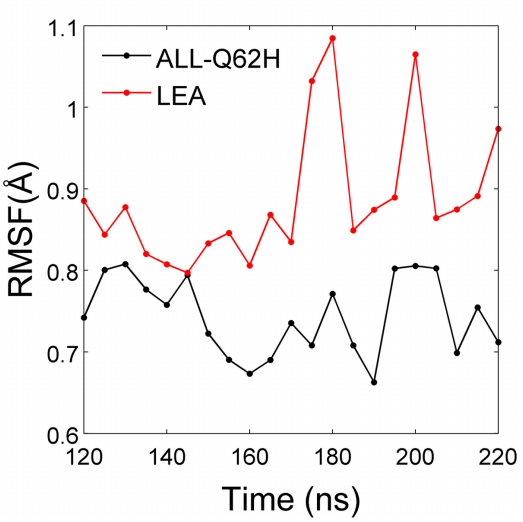


Figure 5.7: The average RMSF of the chromophore in ALL-Q62H and LEA calculated in a sliding window of 5 ns.

Structural Dynamics Related to Functional Divergence

To identify any collective protein motions that correlate with chromophore dynamics, I employed the same PRS method as used in the study of β-latamases. As mentioned earlier, sites with high dfi are structurally flexible sites and prone to the perturbation of other residues, whereas sites with low dfi may absorb and transfer the perturbation throughout the protein in a cascade fashion, which are usually involved with hinge parts of the protein that control the domain motion. The difference in $\%dfi$ between ALL-Q62H and LEA ($\Delta(\%dfi)$; the $\%dfi$ of LEA subtractes the $\%dfi$ of ALL-Q62H) allows for the identification of sites with substantially increased dynamics (red, positive $\Delta(\%dfi)$) or decreased dynamics (blue, negative $\Delta(\%dfi)$) from ALL-Q62H to LEA (figure 5.8). In this way, 16 red and 13 blue sites were identified as the top 15% sites showing significant flexibility change), with the red sites clustered near the chromophore's phenolic end, and the blue sites clustered in loop regions located diagonally across the β-barrel (figure 5.9).

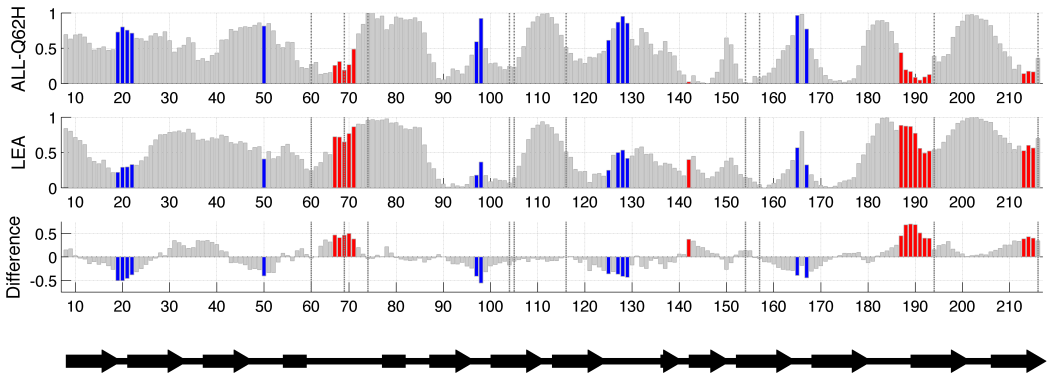


Figure 5.8: The dynamics profile ($\%dfi$) of ALL-Q62H and LEA. The residue sites where LEA gets more flexible than ALL-Q62H are marked as red, and those where LEA are less flexible are marked as blue. The difference of $\%dfi$ is calculated by subtracting the $\%dfi$ of LEA from that of All-Q62H at the same site. The vertical dash lines mark the mutation sites. Sites colored red are: 67-71, 142, 187-193 and 213-215. Sites colored blue are: 19-22, 50, 97-98, 125, 127-129, 165 and 167.

The dfi analysis provided strikingly different dynamic features for ALL-Q62H and LEA. Some regions with relatively high dfi values in ALL-Q62H were found to correspond to regions with relatively low dfi values in LEA, and vice versa (figure 5.10). These results suggests that global protein breathing motions may involve opening of the β-barrel near one end, while the other end serves as a hinge region similar to the joints of a skeleton. Surprisingly, the almost perfect switch of red and blue regions between ALL-Q62H and LEA suggests that the hinge region has moved diagonally across the β-barrel in response to the set of 12 residue replacements promoting photoconversion.

Regions with increased dynamics comprise the chromophore and attached peptide, as well as segments of β-strands contacted by these groups. The 16 top red sites includes residues 67-71, 142, 187-193 and 213-215. Based on the $\%dfi$ value, the chromophore exhibits increased flexibility in LEA, consistent with the RMSF analysis above. Locally, only two residues in direct con-

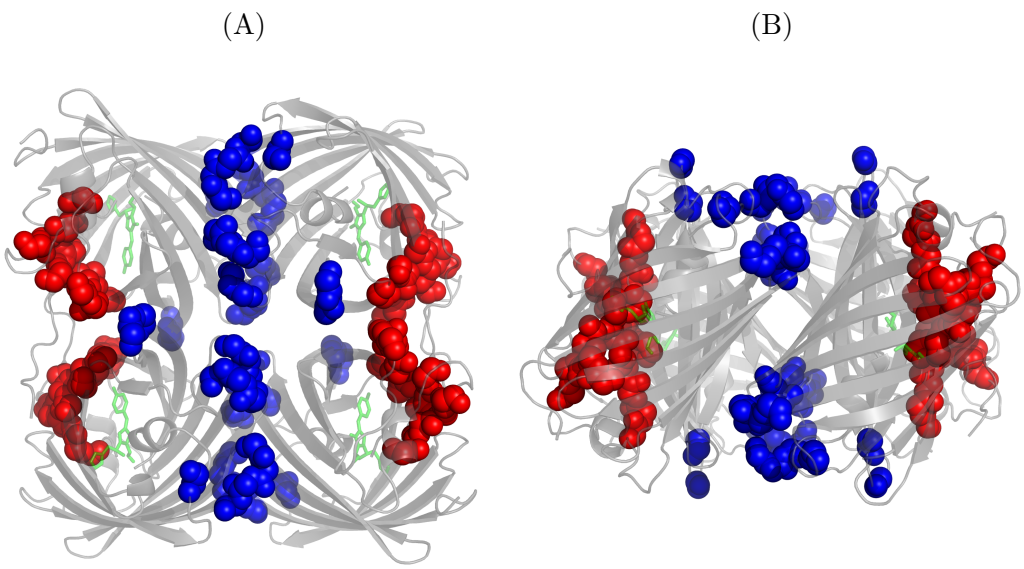


Figure 5.9: The sites where ALL-Q62H and LEA show significant flexibility difference are mapped on the 3D structure of ALL-Q62H. (A) Top view. (B) Side view.

tact with the chromophore exhibit substantially increased dynamics, and are therefore flagged as red, Ser142 (strand #7) and His193 (strand #10). These observations provide support for the proposed mechanism of light-activated remodeling of the active site. The primary effect of chromophore twisting would entail the transient disruption of H-bonding with Ser142 and π-stacking with His193, a process that would be greatly facilitated by more flexible interactions with these residues.

Based on the $\%dfi$ analysis, several regions of the protein exhibit substantially increased mobility that may be transmitted to the chromophore binding pocket. An internal peptide classified as red, comprising residues 67-71, forms an irregular structure that extends from the 310-helical turn (65 to 68) attached to the chromophore to the β-barrel cap near the N- and C-termini. The primary source of increased fluctuations appears to be the packing defect introduced by the T69A substitution. Main-chain H-bonding between T69A

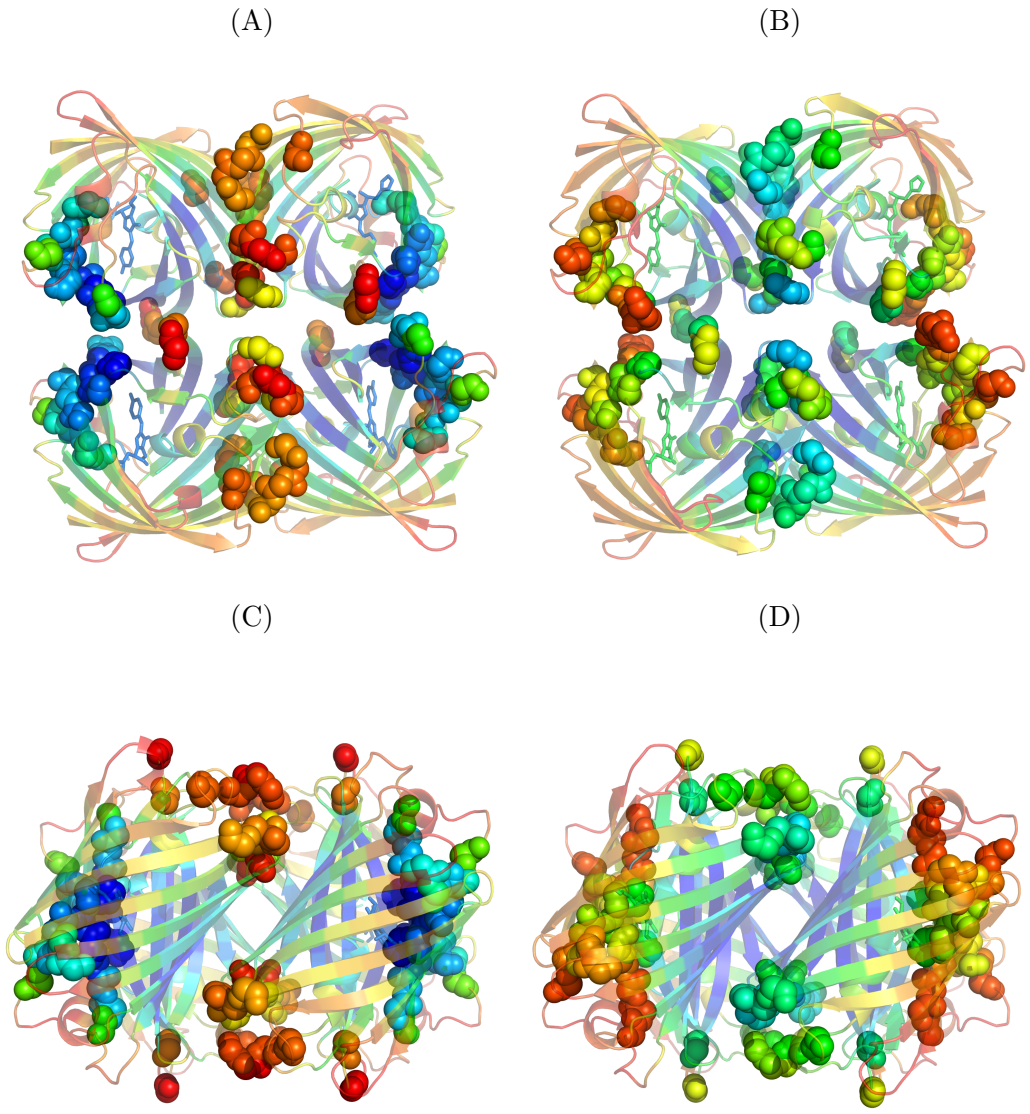


Figure 5.10: Comparison of the dynamics profile of ALL-Q62H and LEA. Residues are colored with a spectrum of red to blue according to their $\%dfi$ values, where rigid regions are denoted by blue/green and fexible regions are denoted with red/orange. Sites where ALL-Q62H and LEA show significant flexibility difference are shown in spheres. (A) Top view of ALL-Q62H; (B) Top view of LEA; (C) Side view of ALL-Q62H; (D) Side view of LEA. A and B, C and D are in the same orientation.

and Ala213 (strand #11) may propagate dynamic motions to the red β-strand segments 213-215 (strand #11) and 187-193 (adjacent strand #10) bearing the π-stacked His193. β-sheet interactions may propagate thermal motions to the red residue Ser142 (strand #7 adjacent to strand #10), which is H-bonded to the chromophore. As the C-terminal tail is largely disordered in LEA, residues 220-225 demonstrate the most pronounced increase in dynamics (large negative $\%dfi$), providing a rationale for the increased dynamics of residues 187-193 (strand #10) and 213-215 (strand #11). In addition, the main chain of Ser142 makes inter-chain contact with the aromatic group of Phe190 (strand #10), which in turn makes intra-chain contact with the aromatic group of Tyr188. In this way, increased motions may be transmitted across the core of the A-B interface as well.

Regions with decreased dynamics map to loop regions located diagonally across the β-barrel. Residues that exhibit substantially elevated $\%dfi$ values, and are therefore rigidified in LEA compared to ALL-Q62H (blue residues), map onto five loop regions forming the β-barrel cap opposite to that bearing the N- and C-termini. The 13 top-scoring blue residues consists of 19-22, 50, 97-98, 125, 127-129, 165 and 167. Four of these make molecular contact across the A-D interface. The blue residue Lys22 (beginning of strand #2) forms an ionic bond with neighboring chain Glu117 (strand #6) and vice versa, such that the two ionic bonds fortified by reduced fluctuation flank one side of the A-D interface, likely stabilizing the anti-parallel dimer. Additional stabilization of this interface may be provided by a hydrogen bond between blue Asn19 and neighboring-chain Lys178, and van der Waals contacts involving blue Pro127 and Asn128 (loop connecting strands #6 and #7) and adjacent chain Asp150 (loop connecting strands #7 and #8). Regardless, the majority

of rigidified residues cluster onto loop regions that delineate a section of the β-barrel cap diagonally across the C-terminus and near the A-D interface. I propose that this region serves as a hinge that is electrostatically anchored to the neighboring chain by Lys22.

Mutational sites that correlate with photoconversion-competency tend to exhibit average $\%dfi$ values. For the most part, the mutational sites themselves are found to carry average $\%dfi$ values. Most of the substituted residues promoting photoconversion (10 out of 11) do not exhibit modified dynamics, but instead, transmit the perturbation to sites elsewhere in the protein. The only exception is the T69A substitution, which is part of the internal 67-71 red peptide with substantially increased dynamic features, as discussed above. On a different note, two of the amino acid replacements flank the segments of β-strands #10 and #11 that were tagged as red, R194C and R216H. Both of these substitutions disrupt inter-molecular interactions in the A-B interface, thereby contributing to the observed disorder in the LEA C-tail, which in turn appears to contribute to the rise in β-strand dynamics.

Structural Dynamics Related with Allosteric Regulation

Any change in protein dynamics, whether short- or long-range, must originate from the set of 12 modifications (11 mutations and 1 deletion) introduced into ALL-Q62H to generate LEA. However, with the exception of T104R, the locations of the mutational sites are found to be rather remote from the blue regions in the LEA tetramer. Therefore, rigidification of the blue sites must result from allosteric effects propagated through the protein matrix over longer distances. To better understand long-range effects, the allosteric response fluctuation profile was calculated upon perturbation of the set of mutational

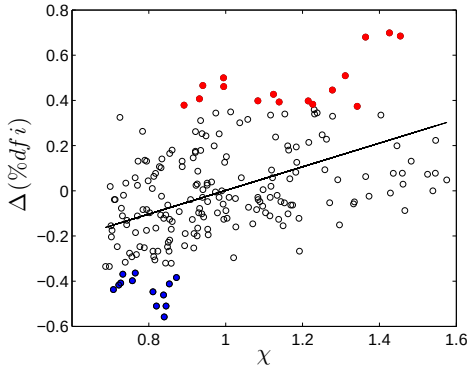


Figure 5.11: The relationship of allosteric response ratio (χ) and the change of $\%dfi$ ($\Delta(\%dfi)$. The blue residues (decreased dynamics) are well separated from the red residues (increased dynamics), based on their χ values.

sites only [96]. For each residue i, the allosteric response ratio is defined as the average displacement of residue i upon perturbation of mutational sites only, divided by the average displacement of residue i upon perturbation of all residues. Thus, the allosteric response ratio can be used to identify residues that are particularly sensitive to the perturbation of mutational sites. Accordingly, an elevated χ value ($\chi > 1$) indicates increased flexibility specifically in response to the mutations. A plot of χ values versus $\Delta(\%dfi)$ as shown in figure 5.11 indicates a positive correlation, consistent with the notion that red sites ($\Delta(\%dfi) > 0$ specifically respond to the 12 modifications with increased dynamic fluctuations. Interestingly, the positions that are rigidified due to mutations in LEA ($\Delta(\%dfi) < 0$) tend to have an allosteric response ratio of less than 1. Therefore, blue positions seem to have a particularly strong ability to absorb perturbations at the mutational sites, and do not fluctuate as much as the remaining positions. In summary, the positive correlation between the allosteric response ratio and the $\Delta(\%dfi)$ values suggests that the change in dynamics is due to allosteric regulation.

5.4 Conclusion

In this chapter, I investigated the relationship between conformational dynamics and function divergence of two protein systems: β-lactamases and GFP-like proteins. I found that change in structural dynamics best explains how β-lactamases evolve from substrate-promiscuous generalists to specialists. The modern decedent TEM-1 lactamase has a rigid catalytic pocket, while the ancestral β-lactamases, expecially PNCA and GNCA lactamase, have very flexible binding pockets. It suggests that as the catalysis becomes benzylpenicillin-specific, the catalytic pocket has also been shaped toward the specific target in evolution. Clustering analysis based on dynamics profile indicates PNCA and GNCA lactamases are much more similar to each other and distinctively separated from TEM-1 lactamase. Sites of significant dynamic importance are identified and mutations at those site may be able to alter the substrate-specificity of the protein. The change in conformational dynamics has also been observed in GFP-like proteins, in agreement with their G/R color evolution. Compared with the photocoversion-incompetent protein, the photocoversion-competent protein exhibits increased flexibility at its chromophore and attached peptide, as well as segments of β-strands contacted by these groups, but shows decreased flexibility at loop regions located diagonally across the β-barrel. Such change of conformational dynamics may facilitate G/R photoconversion in the photocoversion-competent protein.

Chapter 6

CONCLUSION

This thesis provides insights about the following puzzles related to the protein folding problem: i) what is the folding mechanism, especially its relationship with protein topology; ii) how the evolutionary information encoded in the sequence of a protein determines the structure; and iii) how a protein acquires new function in the evolution.

The recent view of of protein folding assumes a funnel-shaped free energy landscape for the protein to pass from the vast unfolded ensemble down to the native state. As the funnel shape is largely determined by the entropy of the protein, the topology of the native structure must be a determinant in folding mechanism and folding kinetics. Indeed, the folding rates of proteins correlate with the average degree to which native contacts are "local" within the chain sequence: fast-folders usually have mostly local structures. In Chapter 3, I dissected the native topology further by focusing on non-local and local contacts using lower and upper bounds of allowable sequence separation in computing the average contact order. I analyzed non-local and local contacts of 82 two-state proteins whose experimental folding rates span over six orders of magnitude. I observed that both the number of non-local contacts and the average sequence separation of non-local contacts (non-local CO) are both negatively correlated with the folding rate, showing that the non-local contacts dominate the barrier-crossing process. Surprisingly, the local contact orders of the proteins also correlate with the folding rates. However, this correlation

shows a strong positive trend indicating the role of a diffusive search in the denatured basin.

Based on the topology of the native structure, I predicted the folding rate distribution, weighted by protein copy number, for E. coli and Yeast proteome. E. coli and Yeast proteomes yield very similar distributions with average folding time of 100 milliseconds and 170 milliseconds, respectively. While the fastest time scale of the distribution is near the speed limit of 1 microsecond (typical of barrier less folders), it is postulated that the lower speed limit is determined by protein degradation time scale. A diffusion-drift population model in the sequence space - with these two speed limits as boundary conditions - well captures the predicted folding time distribution, and quantitatively demonstrates the lower speed to be eight seconds, close to protein half life. Direct comparison between the predicted folding time and experimentally measured half life shows 97% of the proteome have a folding time faster than their corresponding degradation time, further supporting that proteome folding kinetics is limited by protein half life.

The native structure of a protein is dictated by it sequence. Earlier experiments suggested that the evolutionary information (conservation of amino acids and coevolution between amino acids) encoded in protein sequences is necessary and sufficient to specify the fold of a protein family. However, there was no computational work to quantify the effect of such evolutionary information on the folding process. In Chapter 4, I simulated a repertoire of native and artificial WW domain sequences using a physics-based protein structure search method called Zipping and Assembly method (ZAM), which samples conformational space effectively towards native-like conformations through zipping and assembly search mechanism. I explored the sequence-structure relationship

for WW domains and found that the coevolution information has a remarkable influence on local contacts of N-terminal β-turn of WW domains . This turn would not form correctly in the absence of such information. Moreover, through the maximum likelihood approach, I identified five local contacts that play a critical role in folding. Using the contact probability of those five local contacts at the early stage of folding, I built a classification model. This enables me to predict the foldability of a WW sequence with 81% accuracy. Based on this classification model, I re-designed the unfoldable WW domain sequences and make them foldable by introducing a few mutations that lead to stabilization of these critical contacts.

Rather than a single structure, protein exists as an ensemble of structures in the native equilibrium with conformationally dynamics. In Chapter 5, I postulated the divergence of new functions can take place within existing fold by modification of conformational dynamics. The postulation was demonstrated in both β-lactamases and GFP-like proteins, whose conformational dynamics were investigated through molecular dynamics (MD) simulations in conjunction with Perturbation Response Scanning (PRS). For β-lactamases, the modern TEM-1 lactamase shows a comparatively rigid active-site region, likely reflecting adaptation for efficient degradation of a specific substrate (penicillin), while enhanced active-site flexibility in the resurrected proteins likely allows for the binding and subsequent degradation of antibiotic molecules of different size and shape. Clustering of the conformational dynamics on the basis of Singular Value Decomposition (SVD) is in agreement with the functional divergence, as the ancient β-lactamases cluster together, separated from their modern descendant. For GFP-liked proteins, the trend of change in conformational dynamics is also consistent with their G/R color

evolution. Compared with the photocoversion-incompetent protein, the chromophore of photocoversion-competent protein exhibits increased flexibility. Its chromophore-attached peptide, as well as segments of β-strands contacted by these groups also show increased dynamics, while loop regions located diagonally across the β-barrel indicate decreased dynamics. Such modification in conformational dynamics may be a consequence of allosteric regulation due to mutations.

REFERENCES

[1] "http://predictioncenter.org", (2014).

[2] "http://scop.mrc-lmb.cam.ac.uk/scop/count.html", (2014).

[3] Alieva, N. O., K. A. Konzen, S. F. Field, E. A. Meleshkevitch, M. E. Hunt, V. Beltran-Ramirez, D. J. Miller, J. Wiedenmann, A. Salih and M. V. Matz, "Diversity and evolution of coral fluorescent proteins", PLoS ONE **3**, 7 (2008).

[4] Allen, H. K., L. a. Moe, J. Rodbumrer, A. Gaarder and J. Handelsman, "Functional metagenomics reveals diverse beta-lactamases in a remote Alaskan soil.", The ISME journal **3**, 2, 243–51, (2009).

[5] Ambler, R. P., "The structure of beta-lactamases", Philosophical Transactions of the Royal Society B: Biological Sciences **289**, 1036, 321–331, (1980).

[6] Andersen, H. C., "Molecular dynamics simulations at constant pressure and/or temperature", The Journal of Chemical Physics **72**, 4, 2384–2393, (1980).

[7] Andersen, H. C., "Rattle: A velocity version of the shake algorithm for molecular dynamics calculations", Journal of Computational Physics **52**, 1, 24–34, (1983).

[8] Ando, R., H. Hama, M. Yamamoto-Hino, H. Mizuno and A. Miyawaki, "An optical marker based on the UV-induced green-to-red photoconversion of a fluorescent protein.", Proceedings of the National Academy of Sciences of the United States of America **99**, 20, 12651–12656 (2002).

[9] Anfinsen, C. B., "Principles that govern the folding of protein chains.", Science (New York, N.Y.) **181**, 96, 223–230 (1973).

[10] ANFINSEN, C. B., E. HABER, M. SELA and F. H. WHITE, "The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain.", Proceedings of the National Academy of Sciences of the United States of America **47**, 1309–1314 (1961).

[11] Atanasov, B. P., D. Mustafi and M. W. Makinen, "Protonation of the beta-lactam nitrogen is the trigger event in the catalytic action of class A beta-lactamases.", Proceedings of the National Academy of Sciences of the United States of America **97**, 7, 3160–5, (2000).

[12] Atilgan, a. R., S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin and I. Bahar, "Anisotropy of fluctuation dynamics of proteins with an elastic network model.", Biophysical journal **80**, 1, 505–15, (2001).

[13] Atilgan, C. and A. R. Atilgan, "Perturbation-response scanning reveals ligand entry-exit mechanisms of ferric binding protein.", PLoS computational biology **5**, 10, e1000544, (2009).

[14] Atilgan, C., Z. N. Gerek, S. B. Ozkan and a. R. Atilgan, "Manipulation of conformational change in proteins by single-residue perturbations.", Biophysical journal **99**, 3, 933–43, (2010).

[15] Bagler, G. and S. Sinha, "Assortative mixing in Protein Contact Networks and protein folding kinetics.", Bioinformatics (Oxford, England) **23**, 14, 1760–7, (2007).

[16] Bahar, I., T. R. Lezon, L.-W. Yang and E. Eyal, "Global dynamics of proteins: bridging between structure and function.", Annual review of biophysics **39**, 23–42 (2010).

[17] Bahar, I. and a. J. Rader, "Coarse-grained normal mode analysis in structural biology.", Current opinion in structural biology **15**, 5, 586–92, (2005).

[18] Bakan, A., L. M. Meireles and I. Bahar, "ProDy: protein dynamics inferred from theory and experiments.", Bioinformatics (Oxford, England) **27**, 11, 1575–7, (2011).

[19] Baker, D., "A surprising simplicity to protein folding.", Nature **405**, 6782, 39–42, (2000).

[20] Baker, D. and D. a. Agard, "Kinetics versus thermodynamics in protein folding.", Biochemistry **33**, 24, 7505–9, (1994).

[21] Baker, N., M. Holst and F. Wang, "Adaptive multilevel finite element solution of the Poisson-Boltzmann equation II. Refinement at solvent-accessible surfaces in biomolecular systems", Journal of Computational Chemistry **21**, 15, 1343–1352, (2000).

[22] Baker, N. A., D. Sept, S. Joseph, M. J. Holst and J. A. McCammon, "Electrostatics of nanosystems: application to microtubules and the ribosome.", Proceedings of the National Academy of Sciences of the United States of America **98**, 18, 10037–10041 (2001).

[23] Baldwin, R. L. and G. D. Rose, "Is protein folding hierarchic? I. Local structure and peptide folding", (1999).

[24] Baldwin, R. L. and G. D. Rose, "Is protein folding hierarchic? II. Folding intermediates and transition states.", Trends in biochemical sciences **24**, 2, 77–83, (1999).

[25] Barlow, M. and B. G. Hall, "Predicting evolutionary potential: in vitro evolution accurately reproduces natural evolution of the tem beta-lactamase.", Genetics **160**, 3, 823–32, (2002).

[26] Batey, S. and J. Clarke, "Apparent cooperativity in the folding of multidomain proteins depends on the relative rates of folding of the constituent domains.", Proceedings of the National Academy of Sciences of the United States of America **103**, 48, 18113–18118 (2006).

[27] Beberg, A., D. Ensign, G. Jayachandran, S. Khaliq and V. Pande, "Folding@home: Lessons from eight years of volunteer distributed computing", 2009 IEEE International Symposium on Parallel & Distributed Processing (2009).

[28] Beeman, D., "Some multistep methods for use in molecular dynamics calculations", Journal of Computational Physics **20**, 2, 130–139, (1976).

[29] Belle, A., A. Tanay, L. Bitincka, R. Shamir and E. K. O'Shea, "Quantification of protein half-lives in the budding yeast proteome.", Proceedings of the National Academy of Sciences of the United States of America **103**, 35, 13004–13009 (2006).

[30] Berendsen, H. J. C., J. P. M. Postma, W. F. van Gunsteren, A. DiNola and J. R. Haak, "Molecular dynamics with coupling to an external bath", The Journal of Chemical Physics **81**, 8, 3684, (1984).

[31] Berry, M., S. Dumais and G. O'Brien, "Using linear algebra for intelligent information retrieval", SIAM review **37**, 4, 573–595, (1995).

[32] Bieri, O., J. Wirz, B. Hellrung, M. Schutkowski, M. Drewello and T. Kiefhaber, "The speed limit for protein folding measured by triplet-triplet energy transfer.", Proceedings of the National Academy of Sciences of the United States of America **96**, 17, 9597–601, (1999).

[33] Bolia, A., Z. N. Gerek, O. Keskin, S. Banu Ozkan and K. K. Dev, "The binding affinities of proteins interacting with the PDZ domain of PICK1.", Proteins **80**, 5, 1393–408, (2012).

[34] Bowler, B. E., "Thermodynamics of protein denatured states.", Molecular bioSystems **3**, 2, 88–99, (2007).

[35] Bridgham, J. T., E. A. Ortlund and J. W. Thornton, "An epistatic ratchet constrains the direction of glucocorticoid receptor evolution.", Nature **461**, 7263, 515–519 (2009).

[36] Brinker, A., G. Pfeifer, M. J. Kerner, D. J. Naylor, F. U. Hartl and M. Hayer-Hartl, "Dual function of protein confinement in chaperonin-assisted protein folding", Cell **107**, 2, 223–233 (2001).

[37] Brooks, B. R., C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L.

Woodcock, X. Wu, W. Yang, D. M. York and M. Karplus, "CHARMM: the biomolecular simulation program.", Journal of computational chemistry **30**, 10, 1545–1614 (2009).

[38] Brünger, A., C. B. III and M. Karplus, "Stochastic boundary conditions for molecular dynamics simulations of ST2 water", Chemical physics letters **105**, 5, (1984).

[39] Bryngelson, J. D. and P. G. Wolynes, "Spin glasses and the statistical mechanics of protein folding.", Proceedings of the National Academy of Sciences of the United States of America **84**, 21, 7524–7528 (1987).

[40] Bush, K., G. Jacoby and A. Medeiros, "A functional classification scheme for beta-lactamases and its correlation with molecular structure.", Antimicrobial agents and ... **39**, 6, (1995).

[41] Bush, K. and G. a. Jacoby, "Updated functional classification of beta-lactamases.", Antimicrobial agents and chemotherapy **54**, 3, 969–76, (2010).

[42] Camarero, J. a., D. Fushman, S. Sato, I. Giriat, D. Cowburn, D. P. Raleigh and T. W. Muir, "Rescuing a destabilized protein fold through backbone cyclization.", Journal of molecular biology **308**, 5, 1045–62, (2001).

[43] Cantón, R. and T. M. Coque, "The CTX-M beta-lactamase pandemic.", Current opinion in microbiology **9**, 5, 466–75, (2006).

[44] Chelliah, V., L. Chen, T. L. Blundell and S. C. Lovell, "Distinguishing structural and functional restraints in evolution in order to identify interaction sites", Journal of Molecular Biology **342**, 5, 1487–1504 (2004).

[45] Chen, C. C., T. J. Smith, G. Kapadia, S. Wäsch, L. E. Zawadzke, a. Coulson and O. Herzberg, "Structure and kinetics of the beta-lactamase mutants S70A and K73H from Staphylococcus aureus PC1.", Biochemistry **35**, 38, 12251–8, (1996).

[46] Chen, H. I. and M. Sudol, "The WW domain of Yes-associated protein binds a proline-rich ligand that differs from the consensus established for Src homology 3-binding modules.", Proceedings of the National Academy of Sciences of the United States of America **92**, 17, 7819–23, (1995).

[47] Chen, J. and C. L. Brooks, "Implicit modeling of nonpolar solvation for simulating protein folding and conformational transitions.", Physical chemistry chemical physics : PCCP **10**, 4, 471–81, (2008).

[48] Cheng, X., G. Cui, V. Hornak and C. Simmerling, "Modified replica exchange simulation methods for local structure refinement.", The journal of physical chemistry. B **109**, 16, 8220–30, (2005).

[49] Chennubhotla, C., A. J. Rader, L.-W. Yang and I. Bahar, "Elastic network models for understanding biomolecular machinery: from enzymes to supramolecular assemblies.", Physical biology **2**, 4, S173–S180 (2005).

[50] Cheung, M. S., L. L. Chavez and J. N. Onuchic, "The energy landscape for protein folding and possible connections to function", (2004).

[51] Cho, J.-H., S. Sato and D. P. Raleigh, "Thermodynamics and kinetics of non-native interactions in protein folding: a single point mutant significantly stabilizes the N-terminal domain of L9 by modulating non-native interactions in the denatured state.", Journal of molecular biology **338**, 4, 827–37, (2004).

[52] Choy, W. Y. and J. D. Forman-Kay, "Calculation of ensembles of structures representing the unfolded state of an SH3 domain.", Journal of molecular biology **308**, 5, 1011–32, (2001).

[53] Ciryam, P., R. I. Morimoto, M. Vendruscolo, C. M. Dobson and E. P. O'Brien, "In vivo translation rates can substantially delay the cotranslational folding of the Escherichia coli cytosolic proteome.", Proceedings of the National Academy of Sciences of the United States of America **110**, 2, E132–40, (2013).

[54] Coyle, J. E., F. L. Texter, A. E. Ashcroft, D. Masselos, C. V. Robinson and S. E. Radford, "GroEL accelerates the refolding of hen lysozyme without changing its folding mechanism.", Nature structural biology **6**, 7, 683–690 (1999).

[55] Creighton, T. E., *Proteins: Structures and Molecular Properties* (W. H. Freeman, 1993), .

[56] Csermely, P., R. Palotai and R. Nussinov, "Induced fit, conformational selection and independent dynamic segments: An extended view of binding events", Trends in Biochemical Sciences **35**, 10, 539–546 (2010).

[57] Cui, Q. and I. Bahar, *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems*, Chapman & Hall/CRC Mathematical & Computational Biology (Taylor & Francis, 2005), .

[58] Dahiyat, B. I., C. A. Sarisky and S. L. Mayo, "De novo protein design: towards fully automated sequence selection.", Journal of molecular biology **273**, 4, 789–796 (1997).

[59] Damblon, C., X. Raquet, L. Y. Lian, J. Lamotte-Brasseur, E. Fonze, P. Charlier, G. C. Roberts and J. M. Frère, "The catalytic mechanism of beta-lactamases: NMR titration of an active-site lysine residue of the TEM-1 enzyme.", Proceedings of the National Academy of Sciences of the United States of America **93**, 5, 1747–52, (1996).

[60] D'Costa, V. M., C. E. King, L. Kalan, M. Morar, W. W. L. Sung, C. Schwarz, D. Froese, G. Zazula, F. Calmels, R. Debruyne, G. B. Golding, H. N. Poinar and G. D. Wright, "Antibiotic resistance is ancient.", Nature **477**, 7365, 457–61, (2011).

[61] de Graff, A. M. R., G. Shannon, D. W. Farrell, P. M. Williams and M. F. Thorpe, "Protein unfolding under force: crack propagation in a network.", Biophysical journal **101**, 3, 736–44, (2011).

[62] De Sancho, D., U. Doshi and V. Muñoz, "Protein folding rates and stability: how much is there beyond size?", Journal of the American Chemical Society **131**, 6, 2074–5, (2009).

[63] Deerwester, S., S. Dumais and T. Landauer, "Indexing by latent semantic analysis", JASIS (1990).

[64] Dekker, J. P., A. Fodor, R. W. Aldrich and G. Yellen, "A perturbation-based method for calculating explicit likelihood of evolutionary covariance in multiple sequence alignments.", Bioinformatics (Oxford, England) **20**, 10, 1565–1572 (2004).

[65] Delairesq, M., R. Labiay and J.-p. Samamall, "Site-directed Mutagenesis at the Active Site of Escherichia coli TEM-1 β-Lactamase", The journal of biological chemistry **26**, 29, 20600–20606 (1992).

[66] Dill, K. and S. Ozkan, "The protein folding problem", Annual review of ... pp. 289–316, (2008).

[67] Dill, K. A., "Theory for the folding and stability of globular proteins.", Biochemistry **24**, 6, 1501–1509 (1985).

[68] Dill, K. a., "Dominant forces in protein folding.", Biochemistry **29**, 31, 7133–55, (1990).

[69] Dill, K. A., "The meaning of hydrophobicity.", Science (New York, N.Y.) **250**, 4978, 297–298 (1990).

[70] Dill, K. A., "Polymer principles and protein folding.", Protein science : a publication of the Protein Society **8**, 6, 1166–1180 (1999).

[71] Dill, K. A., S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas and H. S. Chan, "Principles of protein folding–a perspective from simple exact models.", Protein science : a publication of the Protein Society **4**, 4, 561–602 (1995).

[72] Dill, K. A. and H. S. Chan, "From Levinthal to pathways to funnels.", Nature structural biology **4**, 1, 10–19 (1997).

[73] Dill, K. a., K. M. Fiebig and H. S. Chan, "Cooperativity in protein-folding kinetics.", Proceedings of the National Academy of Sciences of the United States of America **90**, 5, 1942–6, (1993).

[74] Dill, K. a. and D. Shortle, "Denatured states of proteins.", Annual review of biochemistry **60**, 795–825, (1991).

[75] Drew, K., P. Winters, G. L. Butterfoss, V. Berstis, K. Uplinger, J. Armstrong, M. Riffle, E. Schweighofer, B. Bovermann, D. R. Goodlett, T. N. Davis, D. Shasha, L. Malmstrom and R. Bonneau, "The Proteome Folding Project: Proteome-scale prediction of structure and function", (2011).

[76] Duan, Y. and P. A. Kollman, "Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution.", Science (New York, N.Y.) **282**, 5389, 740–744 (1998).

[77] Dupradeau, F.-Y., A. Pigache, T. Zaffran, C. Savineau, R. Lelong, N. Grivel, D. Lelong, W. Rosanski and P. Cieplak, "The R.E.D. tools: advances in RESP and ESP charge derivation and force field library building.", Physical chemistry chemical physics : PCCP **12**, 28, 7821–39, (2010).

[78] Durbin, R., *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* (Cambridge University Press, 1998), .

[79] Earl, D. J. and M. W. Deem, "Parallel tempering: theory, applications, and new perspectives.", Physical chemistry chemical physics : PCCP **7**, 23, 3910–3916 (2005).

[80] Eaton, W., V. Muñoz and S. Hagen, "Fast Kinetics and Mechanisms in Protein Folding 1", Annual review of ... (2000).

[81] Eaton, W., V. Munoz and P. Thompson, "Kinetics and Dynamics of Loops, alpha-Helices, beta-Hairpins, and Fast-Folding Proteins", Accounts of Chemical Research **31**, 11, 745–753, (1998).

[82] Elander, R. P., "Industrial production of beta-lactam antibiotics.", Applied microbiology and biotechnology **61**, 5-6, 385–92, (2003).

[83] Ensign, D. L. and V. S. Pande, "The Fip35 WW domain folds with structural and mechanistic heterogeneity in molecular dynamics simulations", Biophysical Journal **96**, 8 (2009).

[84] Farrell, D. W., M. Lei and M. F. Thorpe, "Comparison of pathways from the geometric targeting method and targeted molecular dynamics in nitrogen regulatory protein C.", Physical biology **8**, 2, 026017, (2011).

[85] Farrell, D. W., M. Lei and M. F. Thorpe, "Comparison of pathways from the geometric targeting method and targeted molecular dynamics in nitrogen regulatory protein C.", Physical biology **8**, 2, 026017, (2011).

[86] Farrell, D. W., K. Speranskiy and M. F. Thorpe, "Generating stereochemically acceptable protein pathways.", Proteins **78**, 14, 2908–21, (2010).

[87] Fersht, a. R., "Transition-state structure as a unifying basis in protein-folding mechanisms: contact order, chain topology, stability, and the extended nucleus mechanism.", Proceedings of the National Academy of Sciences of the United States of America **97**, 4, 1525–9, (2000).

[88] Field, S. F. and M. V. Matz, "Retracing evolution of red fluorescence in GFP-like proteins from Faviina corals.", Molecular biology and evolution **27**, 2, 225–233 (2010).

[89] Flanagan, J. M., M. Kataoka, D. Shortle and D. M. Engelman, "Truncated staphylococcal nuclease is compact but disordered.", Proceedings of the National Academy of Sciences of the United States of America **89**, 2, 748–52, (1992).

[90] Fogolari, F., P. Zuccato, G. Esposito and P. Viglino, "Biomolecular electrostatics with the linearized Poisson-Boltzmann equation.", Biophysical journal **76**, 1 Pt 1, 1–16 (1999).

[91] Frauenfelder, H., S. G. Sligar and P. G. Wolynes, "The energy landscapes and motions of proteins.", Science (New York, N.Y.) **254**, 5038, 1598–1603 (1991).

[92] Gallicchio, E. and R. M. Levy, "AGBNP: an analytic implicit solvent model suitable for molecular dynamics simulations and high-resolution modeling.", Journal of computational chemistry **25**, 4, 479–499 (2004).

[93] Garcia, A. E., H. Herce and D. Paschek, "Chapter 5 Simulations of Temperature and Pressure Unfolding of Peptides and Proteins with Replica Exchange Molecular Dynamics", vol. 2 of *Annual Reports in Computational Chemistry*, pp. 83–95 (Elsevier, 2006), .

[94] Gaucher, E. A., S. Govindarajan and O. K. Ganesh, "Palaeotemperature trend for Precambrian life inferred from resurrected proteins.", Nature **451**, 7179, 704–707 (2008).

[95] Gaucher, E. A., J. M. Thomson, M. F. Burgan and S. A. Benner, "Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins.", Nature **425**, 6955, 285–288 (2003).

[96] Gerek, Z. N. and S. B. Ozkan, "Change in allosteric network affects binding affinities of PDZ domains: analysis through perturbation response scanning.", PLoS computational biology **7**, 10, e1002154, (2011).

[97] Ghosh, K., S. B. Ozkan and K. a. Dill, "The ultimate speed limit to protein folding is conformational searching.", Journal of the American Chemical Society **129**, 39, 11920–7, (2007).

[98] Glembo, T. J., D. W. Farrell, Z. N. Gerek, M. F. Thorpe and S. B. Ozkan, "Collective dynamics differentiates functional divergence in protein evolution", PLoS Computational Biology **8**, 3 (2012).

[99] Glembo, T. J., D. W. Farrell, Z. N. Gerek, M. F. Thorpe and S. B. Ozkan, "Collective dynamics differentiates functional divergence in protein evolution.", PLoS computational biology **8**, 3, e1002428, (2012).

[100] Glembo, T. J. and S. B. Ozkan, "Union of geometric constraint-based simulations with molecular dynamics for protein structure prediction.", Biophysical journal **98**, 6, 1046–54, (2010).

[101] Gloor, G. B., L. C. Martin, L. M. Wahl and S. D. Dunn, "Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions.", Biochemistry **44**, 19, 7156–7165 (2005).

[102] Göbel, U., C. Sander, R. Schneider and A. Valencia, "Correlated mutations and residue contacts in proteins.", Proteins **18**, 4, 309–317 (1994).

[103] Gromiha, M. M., "Multiple contact network is a key determinant to protein folding rates.", Journal of chemical information and modeling **49**, 4, 1130–5, (2009).

[104] Gromiha, M. M. and S. Selvaraj, "Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction.", Journal of molecular biology **310**, 1, 27–32, (2001).

[105] Gromiha, M. M., a. M. Thangakani and S. Selvaraj, "FOLD-RATE: prediction of protein folding rates from amino acid sequence.", Nucleic acids research **34**, Web Server issue, W70–4, (2006).

[106] Guo, M., Y. Xu and M. Gruebele, "Temperature dependence of protein folding kinetics in living cells.", Proceedings of the National Academy of Sciences of the United States of America **109**, 44, 17863–7, (2012).

[107] Hagen, S. J., J. Hofrichter, a. Szabo and W. a. Eaton, "Diffusion-limited contact formation in unfolded cytochrome c: estimating the maximum rate of protein folding.", Proceedings of the National Academy of Sciences of the United States of America **93**, 21, 11615–7, (1996).

[108] Halabi, N., O. Rivoire, S. Leibler and R. Ranganathan, "Protein Sectors: Evolutionary Units of Three-Dimensional Structure", Cell **138**, 4, 774–786 (2009).

[109] Hall, B. G. and M. Barlow, "Structure-based phylogenies of the serine beta-lactamases.", Journal of molecular evolution **57**, 3, 255–60, (2003).

[110] Harihar, B. and S. Selvaraj, "Refinement of the long-range order parameter in predicting folding rates of two-state proteins.", Biopolymers **91**, 11, 928–35, (2009).

[111] Harms, M. J. and J. W. Thornton, "Analyzing protein structure and function using ancestral gene reconstruction", (2010).

[112] Hawkins, G. D., C. J. Cramer and D. G. Truhlar, "Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium", Journal of Physical Chemistry **100**, 51, 19824–19839, (1996).

[113] Hills, R. D. and C. L. Brooks, "Insights from coarse-grained g models for protein folding and dynamics.", International journal of molecular sciences **10**, 3, 889–905, (2009).

[114] Hockney, R. W., "The Potential Calculations and Some Applications", Methods in Computational Physics **9**, 136–211, (1970).

[115] Holst, M. J. and F. Saied, "Numerical solution of the nonlinear Poisson–Boltzmann equation: Developing more robust and efficient methods", Journal of Computational Chemistry **16**, 3, 337–364, (1995).

[116] Hoover, W., "Canonical dynamics: Equilibrium phase-space distributions", (1985).

[117] Hopf, T. A., L. J. Colwell, R. Sheridan, B. Rost, C. Sander and D. S. Marks, "Three-dimensional structures of membrane proteins from genomic sequencing", Cell **149**, 7, 1607–1621 (2012).

[118] Hornak, V., R. Abel and A. Okur, "Comparison of multiple Amber force fields and development of improved protein backbone parameters", Proteins: Structure, ... **725**, May, 712–725, (2006).

[119] Huang, J., J. Cheng and H. Chen, "Secondary structure length as a determinant of folding rate of proteins with twoâand threeâstate kinetics", PROTEINS: Structure, Function, ... **17**, September 2006, 12–17, (2007).

[120] Hyeon, C. and D. Thirumalai, "Chain length determines the folding rates of RNA.", Biophysical journal **102**, 3, L11–3, (2012).

[121] Ikeguchi, M., J. Ueno, M. Sato and A. Kidera, "Protein Structural Change Upon Ligand Binding: Linear Response Theory", Physical Review Letters **94**, 7, 078102, (2005).

[122] Ilsley, J. L., M. Sudol and S. J. Winder, "The WW domain: linking cell signalling to the membrane cytoskeleton.", Cellular signalling **14**, 3, 183–9, (2002).

[123] Itzhaki, L. S., 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 **254**, 2, 260–88, (1995).

[124] Ivankov, D. N. and A. V. Finkelstein, "Prediction of protein folding rates from the amino acid sequence-predicted secondary structure.", Proceedings of the National Academy of Sciences of the United States of America **101**, 24, 8942–4, (2004).

[125] Jackson, C. J., J.-L. Foo, N. Tokuriki, L. Afriat, P. D. Carr, H.-K. Kim, G. Schenk, D. S. Tawfik and D. L. Ollis, "Conformational sampling, catalysis, and evolution of the bacterial phosphotriesterase.", Proceedings of the National Academy of Sciences of the United States of America **106**, 51, 21631–21636 (2009).

[126] Jacobs, D. J., a. J. Rader, L. a. Kuhn and M. F. Thorpe, "Protein flexibility predictions using graph theory.", Proteins **44**, 2, 150–65, (2001).

[127] James, L. C. and D. S. Tawfik, "Conformational diversity and protein evolution - A 60-year-old hypothesis revisited", (2003).

[128] Kaiser, S. M., H. S. Malik and M. Emerman, "Restriction of an extinct retrovirus by the human TRIM5alpha antiviral protein.", Science (New York, N.Y.) **316**, 5832, 1756–1758 (2007).

[129] Kaminski, G. A., R. A. Friesner, J. Tirado-Rives and W. L. Jorgensen, "Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides", Journal of physical chemistry B **105**, 6474–6487 (2001).

[130] Kar, G., O. Keskin, A. Gursoy and R. Nussinov, "Allostery and population shift in drug discovery", (2010).

[131] Karplus, M. and D. L. Weaver, "Protein-folding dynamics.", Nature **260**, 5550, 404–406 (1976).

[132] Karplus, M. and D. L. Weaver, "Diffusion-collision Model for Protein Folding", Biopolymers **18**, 1421–1437 (1979).

[133] Karplus, M. and D. L. Weaver, "Protein folding dynamics: the diffusion-collision model and experimental data.", Protein science : a publication of the Protein Society **3**, 4, 650–668 (1994).

[134] Keskin, O., I. Bahar, R. L. Jernigan, J. a. Beutler, R. H. Shoemaker, E. a. Sausville and D. G. Covell, "Characterization of anticancer agents by their growth inhibitory activity and relationships to mechanism of action and structure.", Anti-cancer drug design **15**, 2, 79–98, (2000).

[135] Khersonsky, O. and D. S. Tawfik, "Enzyme promiscuity: a mechanistic and evolutionary perspective.", Annual review of biochemistry **79**, 471–505 (2010).

[136] Kieseritzky, G. and E.-W. Knapp, "Optimizing pKa computation in proteins with pH adapted conformations.", Proteins **71**, 3, 1335–1348 (2008).

[137] Kim, H., T. J. Grunkemeyer, C. Modi, L. Chen, R. Fromme, M. V. Matz and R. M. Wachter, "Acid-Base Catalysis and Crystal Structures of a Least Evolved Ancestral GFP-like Protein Undergoing Green-to-Red Photoconversion.", Biochemistry **52**, 45, 8048–59, (2013).

[138] Kim, P. S. and R. L. Baldwin, "Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding.", Annual review of biochemistry **51**, 459–489 (1982).

[139] Kirkpatrick, S., C. D. Gelatt and M. P. Vecchi, "Optimization by simulated annealing.", Science (New York, N.Y.) **220**, 4598, 671–680 (1983).

[140] Kluger, Y., R. Basri, J. T. Chang and M. Gerstein, "Spectral biclustering of microarray data: coclustering genes and conditions.", Genome research **13**, 4, 703–16, (2003).

[141] Knight, A. M., P. H. Culviner, N. Kurt-Yilmaz, T. Zou, S. B. Ozkan and S. Cavagnero, "Electrostatic effect of the ribosomal surface on nascent polypeptide dynamics.", ACS chemical biology **8**, 6, 1195–204, (2013).

[142] Kofke, D. a., "On the acceptance probability of replica-exchange Monte Carlo trials", The Journal of Chemical Physics **117**, 15, 6911, (2002).

[143] KraemerPecore, C., "A de novo redesign of the WW domain", Protein ... **2**, 2194–2205, (2003).

[144] Krantz, B. a., A. K. Srivastava, S. Nauli, D. Baker, R. T. Sauer and T. R. Sosnick, "Understanding protein hydrogen bond formation with kinetic H/D amide isotope effects.", Nature structural biology **9**, 6, 458–63, (2002).

[145] Kuang, D., Y. Yao, D. Maclean, M. Wang, D. R. Hampson and B. S. W. Chang, "Ancestral reconstruction of the ligand-binding pocket of Family C G protein-coupled receptors.", Proceedings of the National Academy of Sciences of the United States of America **103**, 38, 14050–14055 (2006).

[146] Kubelka, J., J. Hofrichter and W. A. Eaton, "The protein folding 'speed limit'", (2004).

[147] Kuhlman, B., G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard and D. Baker, "Design of a novel globular protein fold with atomic-level accuracy.", Science (New York, N.Y.) **302**, 5649, 1364–8, (2003).

[148] Kumar, S., J. M. Rosenberg, D. Bouzida, R. H. Swendsen and P. A. Kollman, "THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method", Journal of Computational Chemistry **13**, 1011–1021, (1992).

[149] Lamotte-Brasseur, J., G. Dive, O. Dideberg, P. Charlier, J. M. Frère and J. M. Ghuysen, "Mechanism of acyl transfer by the class A serine beta-lactamase of Streptomyces albus G.", The Biochemical journal **279 (Pt 1**, 213–21, (1991).

[150] Larson, S. M., A. A. Di Nardo and A. R. Davidson, "Analysis of covariation in an SH3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions.", Journal of molecular biology **303**, 3, 433–446 (2000).

[151] Leach, A. R., *Molecular Modelling: Principles and Applications* (2001).

[152] Lee, M. S. and M. a. Olson, "Comparison of volume and surface area nonpolar solvation free energy terms for implicit solvent simulations.", The Journal of chemical physics **139**, 4, 044119, (2013).

[153] Leopold, P. E., M. Montal and J. N. Onuchic, "Protein folding funnels: a kinetic approach to the sequence-structure relationship.", Proceedings of the National Academy of Sciences of the United States of America **89**, 18, 8721–8725 (1992).

[154] Levinthal, C., "Are there pathways for protein folding?", Journal de Chimie Physique et de Physico-Chimie Biologique **65**, 44–45, (1968).

[155] Levinthal, C., "How to fold graciously", in "Mössbaun Spectroscopy in Biological Systems Proceedings", vol. 24, pp. 22–24 (1969), .

[156] Levy, S. B. and B. Marshall, "Antibacterial resistance worldwide: causes, challenges and responses.", Nature medicine **10**, 12 Suppl, S122–9, (2004).

[157] Liang, J., "Predicting protein folding rates from geometric contact and amino acid sequence", Protein science , 312, 1256–1263 (2008).

[158] Lindorff-Larsen, K., 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 **126**, 10, 3291–9, (2004).

[159] Lindorff-Larsen, K., S. Piana, R. O. Dror and D. E. Shaw, "How fast-folding proteins fold.", Science (New York, N.Y.) **334**, 6055, 517–20, (2011).

[160] Liu, Y., L. M. Gierasch and I. Bahar, "Role of Hsp70 ATPase domain intrinsic dynamics and sequence evolution in enabling its functional interactions with NEFs.", PLoS computational biology **6**, 9, (2010).

[161] Livermore, D. M., "beta-Lactamases in laboratory and clinical resistance.", Clinical microbiology reviews **8**, 4, 557–84, (1995).

[162] Lockless, S. W. and R. Ranganathan, "Evolutionarily conserved pathways of energetic connectivity in protein families.", Science (New York, N.Y.) **286**, 5438, 295–299 (1999).

[163] Lovell, S. C. and D. L. Robertson, "An integrated view of molecular coevolution in protein-protein interactions.", Molecular biology and evolution **27**, 11, 2567–2575 (2010).

[164] Lwin, T. Z. and R. Luo, "Overcoming entropic barrier with coupled sampling at dual resolutions.", The Journal of chemical physics **123**, 19, 194904, (2005).

[165] Macias, M., M. Hyvönen, E. Baraldi and J. Schultz, "Structure of the WW domain of a kinase-associated protein complexed with a proline-rich peptide", (1996).

[166] Makarov, D. and K. Plaxco, "The topomer search model: A simple, quantitative theory of twoâstate protein folding kinetics", Protein science pp. 17–26, (2009).

[167] Marks, D. S., L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina and C. Sander, "Protein 3D structure computed from evolutionary sequence variation", PLoS ONE **6**, 12 (2011).

[168] Marks, D. S., T. a. Hopf and C. Sander, "Protein structure prediction from sequence variation.", Nature biotechnology **30**, 11, 1072–80, (2012).

[169] Mashaghi, A., G. Kramer, P. Bechtluft, B. Zachmann-Brand, A. J. M. Driessen, B. Bukau and S. J. Tans, "Reshaping of the conformational search of a protein by the chaperone trigger factor.", Nature **500**, 7460, 98–101, (2013).

[170] Medeiros, A., "Evolution and dissemination of β-lactamases accelerated by generations of β-lactam antibiotics", Clinical Infectious Diseases **24**, Suppl 1, (1997).

[171] Merlo, C., K. a. Dill and T. R. Weikl, "Phi values in protein-folding kinetics have energetic and structural components.", Proceedings of the National Academy of Sciences of the United States of America **102**, 29, 10171–5, (2005).

[172] Micheletti, C., "Prediction of folding rates and transition-state placement from native-state geometry.", Proteins **51**, 1, 74–84, (2003).

[173] Miyazawa, S. and R. Jernigan, "Protein stability for single substitution mutants and the extent of local compactness in the denatured state", Protein engineering (1994).

[174] Morishita, T., "Fluctuation formulas in molecular-dynamics simulations with the weak coupling heat bath", J. Chem. Phys. **113**, 8, 2976–2982 (2000).

[175] Naganathan, A. N. and V. Muñoz, "Insights into protein folding mechanisms from large scale analysis of mutational effects.", Proceedings of the National Academy of Sciences of the United States of America **107**, 19, 8611–8616 (2010).

[176] Nevin Gerek, Z., S. Kumar and S. Banu Ozkan, "Structural dynamics flexibility informs function and evolution at a proteome scale.", Evolutionary applications **6**, 3, 423–33, (2013).

[177] Nosé, S., "A molecular dynamics method for simulations in the canonical ensemble", (1984).

[178] Okazaki, K.-I. and S. Takada, "Dynamic energy landscape view of coupled binding and protein conformational change: induced-fit versus population-shift mechanisms.", Proceedings of the National Academy of Sciences of the United States of America **105**, 32, 11182–11187 (2008).

[179] Okur, A., D. R. Roe, G. Cui, V. Hornak and C. Simmerling, "Improving Convergence of Replica-Exchange Simulations through Coupling to a High-Temperature Structure Reservoir", Journal of Chemical Theory and Computation **3**, 2, 557–568, (2007).

[180] Okur, A., L. Wickstrom, M. Layten, R. Geney, K. Song, V. Hornak and C. Simmerling, "Improved Efficiency of Replica Exchange Simulations through Use of a Hybrid Explicit/Implicit Solvation Model", Journal of Chemical Theory and Computation **2**, 2, 420–433, (2006).

[181] Onuchic, J. N., Z. Luthey-Schulten and P. G. Wolynes, "Theory of protein folding: the energy landscape perspective.", Annual review of physical chemistry **48**, 545–600 (1997).

[182] Onuchic, J. N. and P. G. Wolynes, "Theory of protein folding.", Current opinion in structural biology **14**, 1, 70–5, (2004).

[183] Onufriev, A., D. Bashford and D. A. Case, "Modification of the Generalized Born Model Suitable for Macromolecules", The Journal of Physical Chemistry B **104**, 15, 3712–3720, (2000).

[184] Onufriev, A., D. Bashford and D. a. Case, "Exploring protein native states and large-scale conformational changes with a modified generalized born model.", Proteins **55**, 2, 383–94, (2004).

[185] Onufriev, A., D. a. Case and D. Bashford, "Effective Born radii in the generalized Born approximation: the importance of being perfect.", Journal of computational chemistry **23**, 14, 1297–304, (2002).

[186] Ortlund, E. A., J. T. Bridgham, M. R. Redinbo and J. W. Thornton, "Crystal structure of an ancient protein: evolution by conformational epistasis.", Science (New York, N.Y.) **317**, 5844, 1544–1548 (2007).

[187] Ozkan, S., I. Bahar and K. Dill, "Transition states and the meaning of Phi-values in protein folding kinetics", Nature Structural & Molecular Biology , iii, 16–18, (2001).

[188] Ozkan, S. B., G. A. Wu, J. D. Chodera and K. a. Dill, "Protein folding by zipping and assembly.", Proceedings of the National Academy of Sciences of the United States of America **104**, 29, 11987–92, (2007).

[189] Pearlman, D. a., D. a. Case, J. W. Caldwell, W. S. Ross, T. E. Cheatham, S. DeBolt, D. Ferguson, G. Seibel and P. Kollman, "AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules", Computer Physics Communications **91**, 1-3, 1–41, (1995).

[190] Pei, J. and N. V. Grishin, "AL2CO: calculation of positional conservation in a protein sequence alignment.", Bioinformatics (Oxford, England) **17**, 8, 700–712 (2001).

[191] Petsko, G. A. and D. Ringe, *Protein Structure and Function*, Primers in biology (New Science Press, 2004), .

[192] Phillips, J. C., R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé and K. Schulten, "Scalable molecular dynamics with NAMD.", Journal of computational chemistry **26**, 16, 1781–802, (2005).

[193] Pitera, J. W. and W. Swope, "Understanding folding and design: replica-exchange simulations of "Trp-cage" miniproteins.", Proceedings of the National Academy of Sciences of the United States of America **100**, 13, 7587–7592 (2003).

[194] Pitout, J. and K. Laupland, "Extended-spectrum β-lactamase-producing Enterobacteriaceae: an emerging public-health concern", The Lancet infectious diseases **8**, March, 159–166, (2008).

[195] Plaxco, K. W., K. T. Simons and D. Baker, "Contact order, transition state placement and the refolding rates of single domain proteins.", Journal of molecular biology **277**, 4, 985–94, (1998).

[196] Plaxco, K. W., K. T. Simons, I. Ruczinski and D. Baker, "Topology , Stability , Sequence and Length : Defining the Determinants of Two-State Protein Folding Kinetics", Biochemistry **39**, 37 (2000).

[197] Pronk, S., S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess and E. Lindahl, "GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit.", Bioinformatics (Oxford, England) **29**, 7, 845–54, (2013).

[198] Ptitsyn, O., "Protein folding: Hypotheses and experiments", (1987).

[199] Ptitsyn, O. B. and A. A. Rashin, "A model of myoglobin self-organization.", Biophysical chemistry **3**, 1, 1–20 (1975).

[200] Ratcliff, K., J. Corn and S. Marqusee, "Structure, stability, and folding of ribonuclease H1 from the moderately thermophilic Chlorobium tepidum: comparison with thermophilic and mesophilic homologues.", Biochemistry **48**, 25, 5890–8, (2009).

[201] Ratcliff, K. and S. Marqusee, "Identification of residual structure in the unfolded state of ribonuclease H1 from the moderately thermophilic Chlorobium tepidum: comparison with thermophilic and mesophilic homologues.", Biochemistry **49**, 25, 5167–75, (2010).

[202] Rathore, N., M. Chopra and J. J. de Pablo, "Optimal allocation of replicas in parallel tempering simulations.", The Journal of chemical physics **122**, 2, 024111 (2005).

[203] Reading, C. and M. Cole, "Clavulanic acid: a beta-lactamase-inhibiting beta-lactam from Streptomyces clavuligerus", Antimicrobial Agents and Chemotherapy (1977).

[204] Remington, S., "Negotiating the speed bumps to fluorescence", Nature biotechnology **97403**, 28–29, (2002).

[205] Remington, S. J., "Fluorescent proteins: maturation, photochemistry and photophysics", (2006).

[206] Risso, V. a., J. a. Gavira, D. F. Mejia-Carmona, E. a. Gaucher and J. M. Sanchez-Ruiz, "Hyperstability and substrate promiscuity in laboratory resurrections of Precambrian β-lactamases.", Journal of the American Chemical Society **135**, 8, 2899–902, (2013).

[207] Robic, S., M. Guzman-Casado, J. M. Sanchez-Ruiz and S. Marqusee, "Role of residual structure in the unfolded state of a thermophilic protein.", Proceedings of the National Academy of Sciences of the United States of America **100**, 20, 11345–9, (2003).

[208] Roitberg, A. E., A. Okur and C. Simmerling, "Coupling of replica exchange simulations to a non-Boltzmann structure reservoir.", The journal of physical chemistry. B **111**, 10, 2415–8, (2007).

[209] Roux, B. and R. MacKinnon, "The cavity and pore helices in the KcsA K+ channel: electrostatic stabilization of monovalent cations.", Science (New York, N.Y.) **285**, 5424, 100–102 (1999).

[210] Rustad, M. and K. Ghosh, "Why and how does native topology dictate the folding speed of a protein?", The Journal of chemical physics **137**, 20, 205104, (2012).

[211] Ryckaert, J.-P., G. Ciccotti and H. J. C. Berendsen, "Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes", Journal of Computational Physics **23**, 3, 327–341, (1977).

[212] Salomon-Ferrer, R., D. a. Case and R. C. Walker, "An overview of the Amber biomolecular simulation package", Wiley Interdisciplinary Reviews: Computational Molecular Science **3**, 2, 198–210, (2013).

[213] Salverda, M. L. M., J. A. G. M. De Visser and M. Barlow, "Natural evolution of TEM-1 β-lactamase: experimental reconstruction and clinical relevance.", FEMS microbiology reviews **34**, 6, 1015–36, (2010).

[214] Scalley-Kim, M. 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.", Journal of molecular biology **338**, 3, 573–83, (2004).

[215] Schaefer, M. and M. Karplus, "A Comprehensive Analytical Treatment of Continuum Electrostatics", J. Phys. Chem. **100**, 5, 1578–1599 (1996).

[216] Schlick, T., "Molecular dynamics-based approaches for enhanced sampling of long-time, large-scale conformational changes in biomolecules.", F1000 biology reports **1**, July, 51, (2009).

[217] Schmidt, M. and K. Baldridge, "General atomic and molecular electronic structure system", Journal of ... **14**, 11, 1347–1363, (1993).

[218] Schug, A., T. Herges and W. Wenzel, "All-atom folding of the three-helix HIV accessory protein with an adaptive parallel tempering method.", Proteins **57**, 4, 792–798 (2004).

[219] Sept, D., N. A. Baker and J. A. McCammon, "The physical basis of microtubule structure and stability.", Protein science : a publication of the Protein Society **12**, 10, 2257–2261 (2003).

[220] Serohijos, A. W. R., S. Y. R. Lee and E. I. Shakhnovich, "Highly abundant proteins favor more stable 3D structures in yeast.", (2013).

[221] Serohijos, A. W. R., Z. Rimas and E. I. Shakhnovich, "Protein Biophysics Explains Why Highly Abundant Proteins Evolve Slowly", Cell Reports **2**, 2, 249–256 (2012).

[222] Shell, M. S., S. B. Ozkan, V. Voelz, G. A. Wu and K. a. Dill, "Blind test of physics-based prediction of protein structures.", Biophysical journal **96**, 3, 917–24, (2009).

[223] Shell, M. S., R. Ritterson and K. A. Dill, "A test on peptide stability of AMBER force fields with implicit solvation.", The journal of physical chemistry. B **112**, 22, 6878–6886 (2008).

[224] Shortle, D., "The denatured state (the other half of the folding equation) and its role in protein stability.", The FASEB journal pp. 27–34, (1996).

[225] Shortle, D. and M. S. Ackerman, "Persistence of native-like topology in a denatured protein in 8 M urea.", Science (New York, N.Y.) **293**, 5529, 487–9, (2001).

[226] Shortle, D., H. S. Chan and K. a. Dill, "Modeling the effects of mutations on the denatured states of proteins.", Protein science : a publication of the Protein Society **1**, 2, 201–15, (1992).

[227] Socolich, M., S. W. Lockless, W. P. Russ, H. Lee, K. H. Gardner and R. Ranganathan, "Evolutionary information for specifying a protein fold.", Nature **437**, 7058, 512–518 (2005).

[228] Sohl, J. L., S. S. Jaswal and D. A. Agard, "Unfolded conformations of alpha-lytic protease are more stable than its native state.", Nature **395**, 6704, 817–819 (1998).

[229] Srinivasan, J., M. W. Trevathan, P. Beroza and D. A. Case, "Application of a pairwise generalized Born model to proteins and nucleic acids. Inclusion of salt effects", Theoretical Chemistry Accounts **101**, 6, 426–434 (1999).

[230] Still, W., A. Tempczyk, R. Hawley and T. Hendrickson, "Semianalytical treatment of solvation for molecular mechanics and dynamics", Journal of the American Chemical Society **112**, 6127–6129, (1990).

[231] Strynadka NC, Adachi H, Jensen SE, Johns K, Sielecki A, Betzel C, Sutoh K, J. M., "Molecular structure of the acyl-enzyme intermediate in β-lactam hydrolysis at 1.7 Åresolution", Nature **359**, 6397, 700–705 (1992).

[232] Sugita, Y., A. Kitao and Y. Okamoto, "Multidimensional replica-exchange method for free-energy calculations", The Journal of Chemical Physics (2000).

[233] Sugita, Y. and Y. Okamoto, "Replica-exchange molecular dynamics method for protein folding", (1999).

[234] Swope, W. C., H. C. Andersen, P. H. Berens and K. R. Wilson, "A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters", The Journal of Chemical Physics **76**, 1, 637–649, (1982).

[235] Thirumalai, D., "From Minimal Models to Real Proteins: Time Scales for Protein Folding Kinetics", (1995).

[236] Thompson, P. a., W. a. Eaton and J. Hofrichter, "Laser temperature jump study of the helix<==>coil kinetics of an alanine peptide interpreted with a 'kinetic zipper' model.", Biochemistry **36**, 30, 9200–10, (1997).

[237] Thomson, J. M., E. A. Gaucher, M. F. Burgan, D. W. De Kee, T. Li, J. P. Aris and S. A. Benner, "Resurrecting ancestral alcohol dehydrogenases from yeast.", Nature genetics **37**, 6, 630–635 (2005).

[238] Tobias, D. J. and I. I. I. Brooks Charles L., "Molecular dynamics with internal coordinate constraints.", Journal of Chemical Physics **89**, 5115–5127 (1988).

[239] Tokuriki, N., C. J. Oldfield, V. N. Uversky, I. N. Berezovsky and D. S. Tawfik, "Do viral proteins possess unique biophysical features?", Trends in Biochemical Sciences **34**, 2, 53–59 (2009).

[240] Tokuriki, N. and D. S. Tawfik, "Protein dynamism and evolvability.", Science (New York, N.Y.) **324**, 5924, 203–207 (2009).

[241] Torrie, G. and J. Valleau, "Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling", Journal of Computational Physics **23**, 2, 187–199, (1977).

[242] Toth, M., C. Smith and H. Frase, "An antibiotic-resistance enzyme from a deep-sea bacterium", Journal of the American Chemical Society **132**, 2, 816–823, (2010).

[243] Ugalde, J. A., B. S. W. Chang and M. V. Matz, "Evolution of coral pigments recreated.", Science (New York, N.Y.) **305**, 5689, 1433 (2004).

[244] United States Centers for Disease Control and Prevention, "Antibiotic Resistance Threats in the United States 2013", (2013).

[245] Van de Velde, E. F., *Concurrent Scientific Computing* (Springer-Verlag, 1994).

[246] Verlet, L., "Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules", Physical Review **159**, 1, 98–103 (1967).

[247] Voelz, V. A., G. R. Bowman, K. Beauchamp and V. S. Pande, "Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1-39).", Journal of the American Chemical Society **132**, 5, 1526–1528 (2010).

[248] Voelz, V. a., M. S. Shell and K. a. Dill, "Predicting peptide structures in native proteins from physical simulations of fragments.", PLoS computational biology **5**, 2, e1000281, (2009).

[249] Wachter, R. M., J. L. Watkins and H. Kim, "Mechanistic diversity of red fluorescence acquisition by GFP-like proteins.", Biochemistry **49**, 35, 7417–7427 (2010).

[250] Wagoner, J. A. and N. A. Baker, "Assessing implicit models for nonpolar mean solvation forces: the importance of dispersion and volume terms.", Proceedings of the National Academy of Sciences of the United States of America **103**, 22, 8331–8336 (2006).

[251] Walsh, T. and M. Toleman, "Metallo-β-lactamases: the quiet before the storm?", Clinical microbiology ... **18**, 2, (2005).

[252] Wang, M., M. Weiss, M. Simonovic, G. Haertinger, S. P. Schrimpf, M. O. Hengartner and C. von Mering, "PaxDb, a Database of Protein Abundance Averages Across All Three Domains of Life", (2012).

[253] Wang, T., Y. Zhu and F. Gai, "Folding of a three-helix bundle at the folding speed limit", J. Phys. Chem. **108**, 12, 3694–3697, (2004).

[254] Wang, Z., J. Mottonen and E. J. Goldsmith, "Kinetically controlled folding of the serpin plasminogen activator inhibitor 1.", Biochemistry **35**, 51, 16443–16448 (1996).

[255] Weikl, T., M. Palassini and K. Dill, "Cooperativity in twoâstate protein folding kinetics", Protein science pp. 822–829, (2004).

[256] Weikl, T. R., "Loop-closure events during protein folding: rationalizing the shape of Phi-value distributions.", Proteins **60**, 4, 701–11, (2005).

[257] Wells, S., S. Menor, B. Hespenheide and M. F. Thorpe, "Constrained geometric simulation of diffusive motion in proteins.", Physical biology **2**, 4, S127–36, (2005).

[258] West, M., C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. a. Olson, J. R. Marks and J. R. Nevins, "Predicting the clinical status of human breast cancer by using gene expression profiles.", Proceedings of the National Academy of Sciences of the United States of America **98**, 20, 11462–7, (2001).

[259] Wong, K. B., J. Clarke, C. J. Bond, J. L. Neira, S. M. Freund, a. R. Fersht and V. Daggett, "Towards a complete description of the structural and dynamic properties of the denatured state of barnase and the role of residual structure in folding.", Journal of molecular biology **296**, 5, 1257–82, (2000).

[260] Woodcock, L., "Isothermal molecular dynamics calculations for liquid salts", (1971).

[261] Xu, Y., R. Gnanasekaran and D. M. Leitner, "The dielectric response to photoexcitation of GFP: A molecular dynamics study", Chemical Physics Letters **564**, 78–82, (2013).

[262] Yilmaz, L. S. and A. R. Atilgan, "Identifying the adaptive mechanism in globular proteins: Fluctuations in densely packed regions manipulate flexible parts", The Journal of Chemical Physics **113**, 10, 4454, (2000).

[263] Yokoyama, S., T. Tada, H. Zhang and L. Britt, "Elucidation of phenotypic adaptations: Molecular analyses of dim-light vision proteins in vertebrates.", Proceedings of the National Academy of Sciences of the United States of America **105**, 36, 13480–13485 (2008).

[264] Yokoyama, S., H. Yang and W. T. Starmer, "Molecular basis of spectral tuning in the red- and green-sensitive (M/LWS) pigments in vertebrates.", Genetics **179**, 4, 2037–2043 (2008).

[265] Zeldovich, K. B., P. Chen and E. I. Shakhnovich, "Protein stability imposes limits on organism complexity and speed of molecular evolution.", Proceedings of the National Academy of Sciences of the United States of America **104**, 41, 16152–16157 (2007).

[266] Zhou, H. and Y. Zhou, "Folding rate prediction using total contact distance.", Biophysical journal **82**, 1 Pt 1, 458–63, (2002).

[267] Zhou, H. X., "Boundary element solution of macromolecular electrostatics: interaction energy between two proteins.", Biophysical journal **65**, 2, 955–963 (1993).

[268] Zou, T. and S. B. Ozkan, "Local and non-local native topologies reveal the underlying folding landscape of proteins", Physical Biology **8**, 6, 066011 (2011).

[269] Zou, T., S. B. Ozkan, N. Williams and K. Ghosh, "Proteome folding kinetics is constrained by protein half life", Biophysical journal , submitted (2014).