ABSTRACT

Title of dissertation: COMPUTER SIMULATIONS OF PROTEIN FOLDING

Aram Davtyan, Doctor of Philosophy, 2013

Dissertation directed by: Professor Garegin A. Papoian
Department of Chemistry and Institute
for Physical Science and Technology

Understanding how proteins fold and interact with each other is key to understanding virtually all biological processes. Recent advances in computer power and modeling techniques make it possible to study proteins and other microscopic systems on biologically relevant time and length scales, closing the gap between simulations and experiments. At the same time, the emergence of more accurate models, derived from more rigorous physical principles, allows us to address a number of fundamental questions. The present work relies on molecular dynamics (MD) simulations to investigate several important aspects of protein behavior. First, we introduce the associative memory, water mediated, structure and energy model (AWSEM) and demonstrate its structure prediction capabilities. AWSEM is a coarse-grained protein force field that consists of many physically motivated potentials and a bioinformatically based term, which accounts for many-body local effects by matching its short sequential fragments to the sequences of experimentally resolved structures. We show that the AWSEM force field can be used for

de novo structure prediction, as well as for kinetics and dynamics studies. Next, we use AWSEM to study protein-protein association. Our results indicate that the model not only can successfully predict the native dimeric interfaces but can also correctly reproduce the two and three state behavior of obligatory and nonobligatory dimers. We also find that both monomer geometry and specific non-bonded interactions play an important role in protein-protein association. Subsequently, we investigate protein folding under environmental fluctuations with a simple Gō-like model. More specifically, we study the effect of an oscillating cellular environment on protein folding dynamics through modulating the strength of inter-residue interactions. The results show that, when occurring at some specific timescales, both deterministic and random fluctuations significantly accelerate the folding.

COMPUTER SIMULATIONS OF PROTEIN FOLDING

by

Aram Davtyan

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2013

Advisory Committee:
Professor Garegin A. Papoian, Chair/Advisor
Professor John D. Weeks
Professor Christopher Jarzynski
Professor David Fushman
Professor Sergei Sukharev, Dean's Representative

Acknowledgments

I would like to express my gratitude to all the people who in one way or another assisted and advised me over the past 5 years. Without them the completion of this dissertation would not have been possible.

First and foremost, I am grateful to my advisor, Professor Garegin A. Papoian, for his guidance, continuous support and encouragement throughout my doctoral studies. He has always made himself available for discussions, which has helped me to avoid mistakes and overcome many challenges I faced. His enthusiasm and broad knowledge has inspired me. It has been a pleasure to work with and learn from such an extraordinary individual.

My co-authors have been incredibly valuable in this process. Professor Peter G. Wolynes, Weihua Zheng, and Nicholas Schafer have enabled me to think of new horizons and opportunities. I am looking forward to continuing our joint work in the future.

I thank my fellow lab-members: Ignacia Echeverria, Longhua Hu, Maria Minakova, Konstantin Popov, Davit Potoyan, David Winogradoff, and Pavel Zhuravlev, for many stimulating discussions, and for all the help and support they provided.

I would like to thank the members of my committee: Professors David Fushman, Christopher Jarzynski, Sergei Sukharev, and John D. Weeks, for their time and invaluable feedback.

Last but not the least, I would like to thank my family for believing in me

and helping me achieve my dreams. My parents, Marietta Manucharyan and Sergey Davtyan, have always shared my passion for science and research, without which I would have been in no position to complete this thesis.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

MD	Molecular Dynamics
MC	Monte Carlo
BD	Brownian Dynamics
GPU	graphics processing unit
CPU	central processing units
ASIC	Application-specific integrated circuit
BPTI	bovine pancreatic trypsin inhibitor
CG	coarse-grained
PDB	Protein Data Bank
AWSEM	Associative memory, Water mediated, Structure and Energy Model
AMH	Associated Memory Hamiltonian
AMW	Associated Memory, Water mediated
FRET	Förster resonance energy transfer
WHAM	Weighted Histogram Analysis Method
RMSD	Root-mean-square deviation
MMSI	Maximum mutual sequence identity
AFPT	Average first-passage time

Chapter 1: Introduction and Outline of Thesis

1.1 Introduction

Natural proteins have been specially selected by evolution to perform a vast variety of different functions inside living organisms. One of the most fascinating abilities of proteins is the ability to fold into well-defined three-dimensional structures on surprisingly short time scales. A typical protein consists of several dozen to several hundred, and sometimes even thousands, of consecutively connected amino acids. The amino acids themselves are made of amine and carboxylic acid groups, and a sidechain, which varies with type. Figure 1.1 shows the chemical structures of the 20 standard amino acid types, classified according to polarity and charge. The ability of a protein to fold, as well as its native three-dimensional structure, are largely determined by the properties of individual amino acids in its sequence [1, 2].

Since the discovery of the first three-dimensional structure of myoglobin in 1958 [3], until only a few decades ago, there were a number of controversies about how proteins fold. One of the open questions was how, given the enormous number of degrees of freedom, proteins fold on biologically relevant timescales of the order of seconds or less. This controversy is known as the Levinthal's paradox [4]. He estimated that even in a simplified imaginary experiment, where a conformation

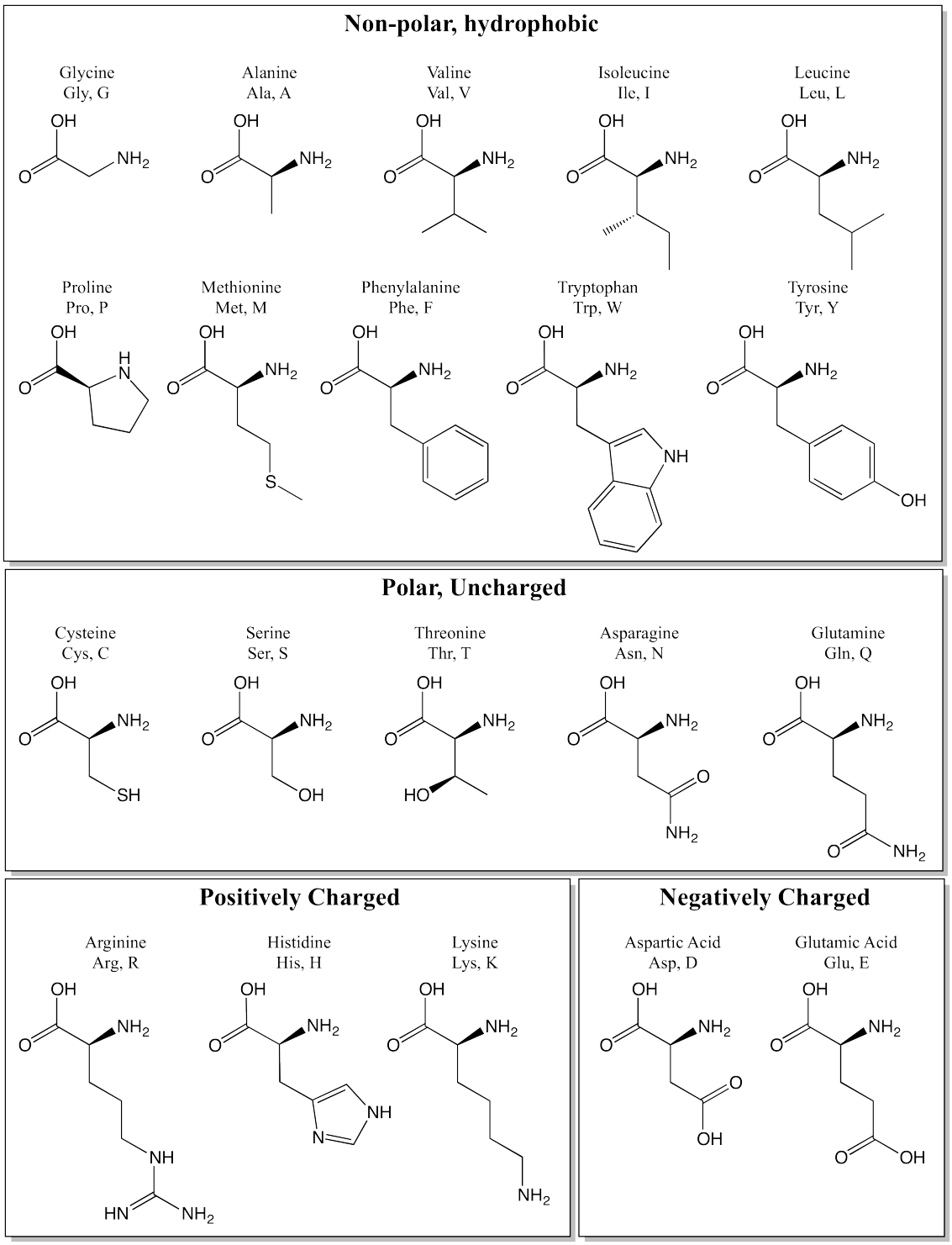


Figure 1.1: The chemical structure of the 20 standard amino acids.

of a chain is defined by two dihedral angles per residue with 3 possible values for each, it will require longer than the age of the universe to sample all the possible states, even if the sampling is done at an extremely fast rate. Later it became clear that natural proteins were carefully selected and designed by evolution in a very specific way, which allows them to find their native state without sampling the vast fraction of the total available phase space. The breakthrough in this so-called protein folding problem was made with the introduction of energy landscape theory [5–8]. According to this theory, the energy landscape of a natural protein is funneled towards its native state, making it energetically favorable and accessible. The former statement means that at room temperature the landscape does not have any high barriers that will prevent the protein from folding on biologically relevant timescales. This is a result of the principle of minimal frustration formulated in the energy landscape theory, which states that the native conformation has minimal unfavorable contacts between the amino acids in the sequence.

Besides helping to understand protein folding from a purely theoretical viewpoint, the energy landscape theory has a large practical significance. Given the high multidimensionality of the energy landscape of a protein, it is important to find a small set of generalized coordinates that can be used to describe the system. Choosing such a set of coordinates is highly non-trivial, as the answer is usually problem specific and not unique. In particular, as a result of the funneled energy landscape and the minimal frustration principle, the protein folding process can be often conveniently described by a single reaction coordinate. One of the most commonly used ones is the fraction of native contacts, denoted by Q [9]. The advantage of using Q

as the folding reaction coordinate is that it is correlated with average contact energy, which decreases with increasing Q. This allows one to schematically draw the energy landscape as a funnel, as shown in Figure 1.2. The vertical dimension indicates the energy relative to the native state, and the width of the funnel corresponds to the average conformational entropy. At the very top of the funnel the protein is in the coil state, which possesses high energy and high conformational entropy. This corresponds to Q values close to zero. As the protein moves down the folding funnel, it compacts, making a transition to a molten globule state [10], where it still has a significant amount of conformational entropy. Molten globule conformations may have most of the native-like secondary structure formed, but are still not as tightly packed as the native state. For molten globules the Q is typically lower than 0.5. At the bottom of the funnel there is a relatively small set of conformations highly similar to the native state (usually with $Q \geq 0.7$). These conformations represent the so-called functional part of the landscape [11–15].

The emergence of energy landscape theory triggered the computational exploration of protein folding problem. The key ideas behind it were directly used to construct models for structure prediction and folding dynamics studies. On the other hand, recent advances in computational techniques, on both a hardware and software level, have opened large horizons for studying diverse biological problems using methods like Molecular Dynamics (MD) [17, 18] and Monte Carlo (MC) [16] simulations. Throughout my work, I predominantly used MD simulations, thus I will only talk about those here.

Over the last four decades Molecular Dynamics simulations became a widely

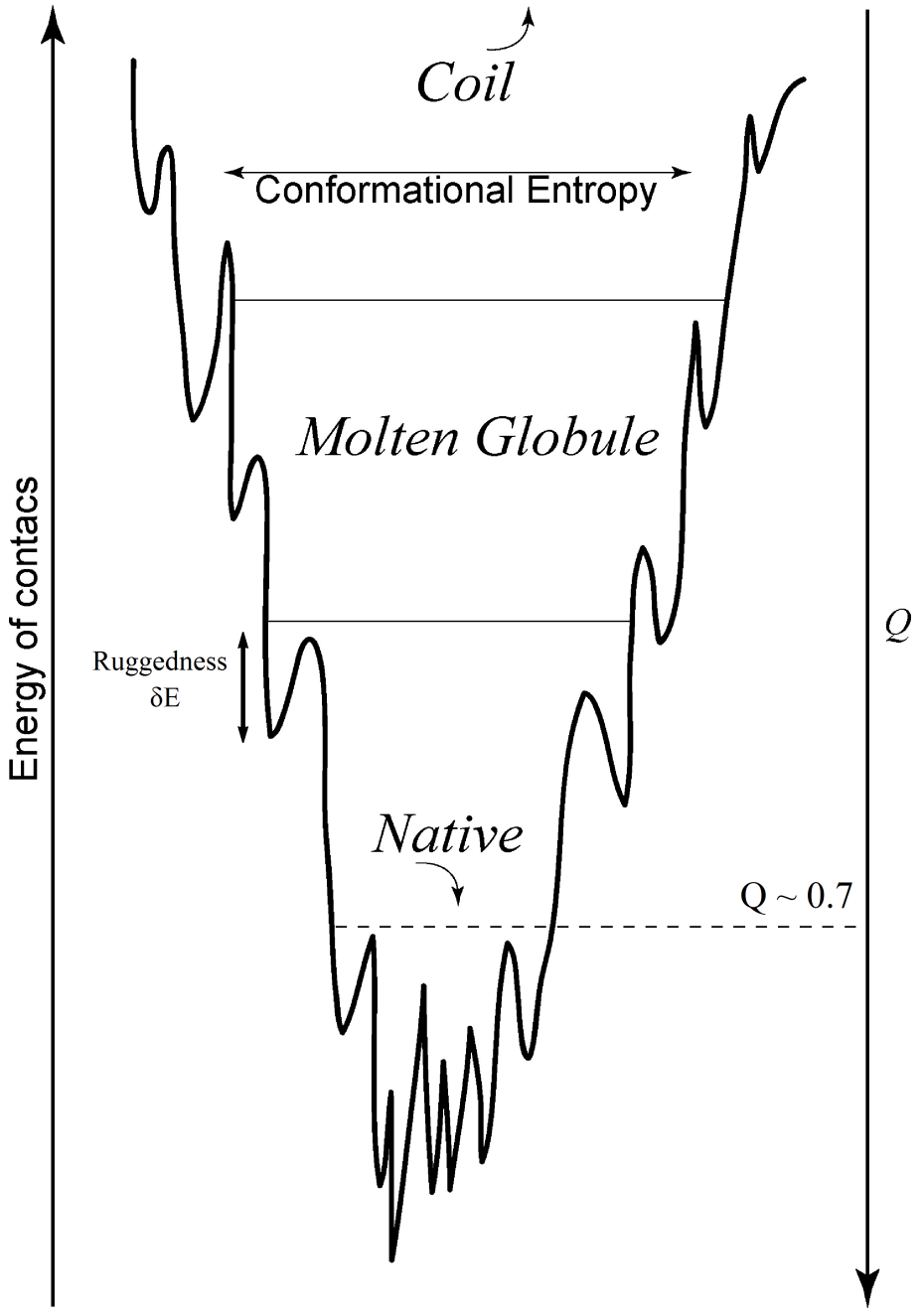


Figure 1.2: The energy landscape of a protein can be schematically drawn as a funnel with the native state located at its bottom. The energy of contacts increases in vertical direction, while the reaction coordinate Q is correspondingly decreases. The width of the funnel indicates the average conformational entropy of the state.

accepted and well established tool for studying various biological systems. The basic idea behind MD simulations is the determination of individual trajectories for each atom in the system by numerically solving Newton's equations of motion, given the interactions between the atoms. MD simulations allow one to model small-scale atomic movements and medium scale conformational changes, which usually cannot be directly observed experimentally. Further, they can be used to interpret experiments by calculating certain time-averaged macroscopic properties of the system, which can either be directly compared to the ones obtained from experiments or used to refine the measurements. The former approach is widely used in X-ray crystallography and NMR spectroscopy structure determination.

Historically the first all-atom Molecular Dynamics simulation of a small protein *bovine pancreatic trypsin inhibitor* (BPTI) 58 residues long was performed in 1977 [19]. The length of this simulation was less than 10 picoseconds. From that time on, both computers and numerical algorithms evolved drastically, allowing for longer and larger scale simulations, carried out with increased accuracy. Of particular interest are the two recent developments, both with very promising perspectives. The first of them is related to the design of special hardware, customized for MD simulations, and the second to the use of graphics processing unit (GPU) cards.

A typical all-atom simulation of a small protein, with only a few tens of residues, immersed in the solvent, involves a few tens of thousands of atoms. Each of those atoms are propagated in space in discrete time steps (usually on the order of femtosecond). On each timestep all interaction energies and forces on individual atoms need to be calculated. This requires an enormous number of calculations

of bonded (including angles and dihedrals) and non-bonded interactions, with non-bonded interactions taking the vast majority of the computational time. Thus, all-atom simulations are usually done in parallel, using several tens or hundred of general-purpose central processing units (CPU). Commonly, this makes it possible to achieve a performance of up to several tens of nanoseconds per day for a typical system discussed above. General-purpose CPUs, though, are not ideal for heavily parallelized and calculation rich MD simulations. A large portion of a CPU chipset is dedicated to memory management, rather than arithmetical calculations. A special design hardware, on the other hand, can enable better performance by several orders of magnitudes. The best example in this regard is the massively-parallel supercomputer called Anton, which was introduced by D. E. Shaw Research [20, 21]. With the focus on faster non-bonded calculations and data exchange between individual computing unites (called Application-specific integrated circuit; ASIC), Anton achieves rates up to several tens of microseconds per day for MD simulation containing tens of thousands of atoms. This new hardware, along with specially designed novel algorithms, enables all-atom MD simulations to be used for many new interesting biological problems, including conformational changes of a voltage-gated potassium channel between activated and deactivated states [22], and folding of small, fast-folding proteins [23].

Another revolutionary idea which gets more and more attention is the use of GPU cards for MD simulations. GPUs where initially designed to calculate and render two and three dimensional computer graphics. Over time, with increasing demand for better resolution computer interfaces, they evolved into multithreaded

computing devices. In modern GPUs the majority of the logic is dedicated to arithmetical calculations, with each of them containing hundreds of arithmetic units. This makes GPUs highly attractive for numerically intensive parallelizable scientific calculations, such as MD modeling. Even in the cases when calculations are only partially ported to GPU, as is done in many standard MD packages including NAMD [24], GROMACS [25], AMBER [26], and LAMMPS [27–29] (with usually only per particle force calculations done on GPU), performance up to twenty times of a single CPU's can be achieved for all-atom simulations. This is, however, not nearly the limit of what one can attain. In works of Zhmurov A. et al. [30, 31] the up to 200 times acceleration of MD simulations (compared to single CPU) made it possible to carry out implicit solvent force-extension simulations using experimental pulling speeds. The most challenging issue for GPU computing is the implementation of long-range electrostatic interactions. This requires Ewald summation, which unlike other calculations included in MD, is not well parallelizable because it involves Fast Fourier transformation. Nevertheless, several attempts to implement Ewald summation have been made, reporting the overall speed of all-GPU MD simulations equivalent to up to 40 CPUs [32, 33]. In conclusion, this is a promising direction, which even today offers the best performance to cost ratio, and is highly supported by the scientific community and hardware manufacturers. For instance, responding to the recent trend, NVIDIA released a series of GPUs called Tesla and a set of tools (Tesla Bio Workbench) specifically designed for scientific calculations.

Many important biological processes involve conformational changes in proteins and protein complexes of several hundred and even thousand amino acids,

which occur on millisecond or second timescales. One such example is the 393 amino acid long tumor suppressor p53 protein, which was originally discovered in 1979, and is known to play a crucial role throughout the cell cycle. p53 can prevent tumor formation by either activating DNA repair mechanisms, stopping cell cycle or, in desperate case, initiating self-destruction of the cell [34–37]. To stop a cell cycle, p53 undergoes a conformational change, transitioning into an active state, in which it can bind DNA. It is also known that p53 mutants are expressed by more than half of the human cancer cells [38]. In addition, larger proteins usually require much longer time to fold; reaching to seconds or even minutes. Unfortunately state of art computer technology today does not allow to access the length and time scales necessary to study those processes with all-atom simulations. Therefore, a number of attempts have been made to construct simplified, coarse-grained (CG) models for proteins. In CG models the number of degrees of freedom of the system are significantly reduced by means of grouping atoms and treating most or all of the solvent implicitly. The interactions between the resulting effective "atoms" (usually called beads) are then derived based on underlying key physical concepts, experimental data and sometimes even all-atom simulations. One of the most apparent advantages of using CG simulations is that they can often be directly compared with experimental results, because they allow to reach length scales and timescales accessible to today's single-molecule experiments.

Designing a coarse-grained model one need to make a compromise between simplicity, accuracy and transferability. In the recent years a large number of protein CG models have been introduced. Aimed to approach a wide variety of problems,

they highly differ in their level of coarse-graining and the degree of specificity; the amount of target specific information they employ. For instance, the models where each amino acid is represented with only one bead usually rely on one or more reference structures, which can either be the native state or any other desirable configuration. The structure of the chain is then biased towards those configurations using bonded and non-bonded potentials. One of the early entirely structure based models was the Gō model [39]. Despite its extreme simplicity the Gō model often correctly reproduces kinetics of protein folding. This surprising result is the consequence of the principle of minimal frustration, according to which the folding rate is strongly coupled with the native state topology [40–42]. The Gō-like approach was also recently used to construct an all-heavy atom model, which due to the steric constraints was able to reproduce some important sequence-dependent effects [43, 44].

The low resolution of the one bead representation does not allow to completely eliminate the need for reference structure. Nevertheless, many attempts have been made to introduce more sophisticated terms to the Gō model and several similar potentials, while retaining some or all biasing terms [45–51].

Inclusion of more beads per amino acid makes it possible to directly account for such important aspects of protein physics as backbone geometry, sequence specificity, hydrogen bonding, and solvation effects without seriously relying on use of reference structures. One key advantage of this is the full or partial transferability of the model from one amino acid sequence to another. Besides the number of beads used to describe backbone and sidechain conformations, those models can also be

distinguished by the methods which have been employed to obtain the interaction parameters. While the complexity of inter-protein interactions make it extremely hard to explicitly derive those parameters from purely physical perspective, the large amount of the available experimental data in form of RCSB Protein Data Bank (PDB) [52] about tens of thousands of proteins and their native folds, makes it very practical to do the parametrization using the occurrence of certain structural and sequential motives in proteins which were already resolved. The models parameterized this way are usually called knowledge-based models. The first attempt to construct a knowledge-based protein model dates back to 1970s [53]. In many cases this is just the first step in model building, followed by further optimization of the energy function either by using a training set of representative experimental structures [54,55] or by matching the model's behavior to all-atom simulations [56].

In the recent years a number of multi-bead protein CG models have been introduced [57–61]. Here, I will briefly discuss the particular model I used throughout my work, dubbed AWSEM.

The Associative memory, Water mediated, Structure and Energy Model (AWSEM) has been introduced and fully described in our recent work published in [62] (see also Chapter 2 and Appendix A). According to AWSEM, each amino acid is described with two beads per backbone and one bead per sidechain (except for Glycine), placed in the center of C_α, C_β, and the carboxylic oxygen (see Figure 1.3). It is primarily based on the Associated Memory Hamiltonian (AMH) developed over many years by Professor Peter Wolynes and his group at the University of Illinois at Urbana-Champaign and University of California in San Diego [63–69].

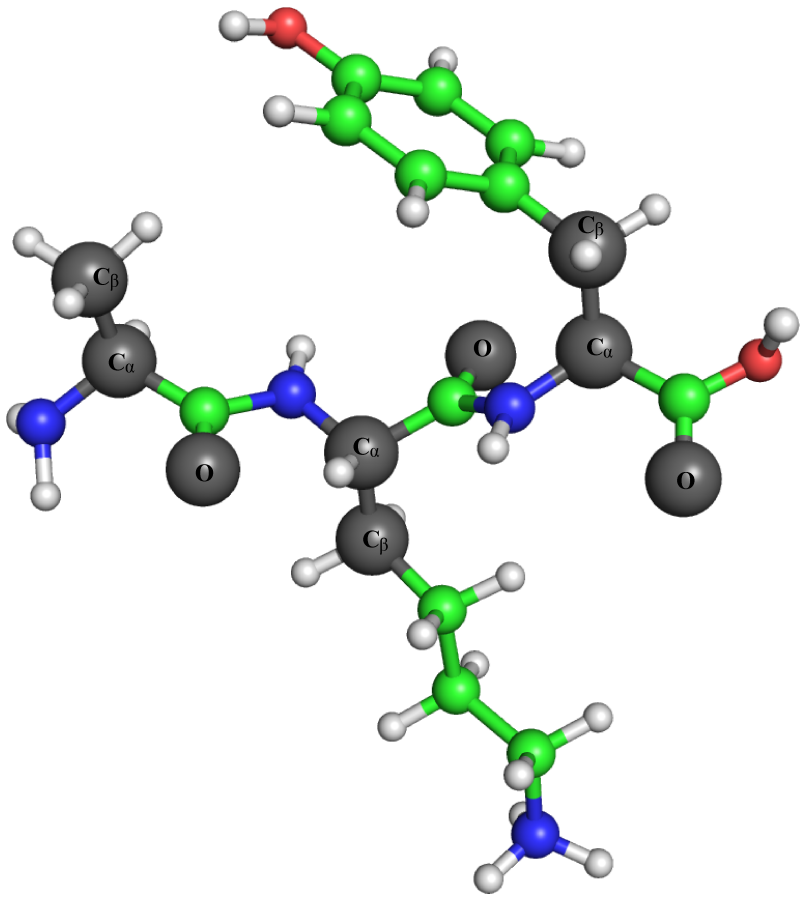


Figure 1.3: According to the Associative memory, Water mediated, Structure and Energy Model (AWSEM) each amino acid is described by three beads placed in the center of C_α, C_β and O. The figure shows the placement of those beads (grey spheres) on Alanine Lysine Tyrosine tri-peptide.

After addition of the contact and water-mediated interactions in 2004 by Papoian, Wolynes and co-workers [70,78,79], it was renamed AMW. AMH/AMW was derived and optimized based on the energy landscape theory and the principle of minimal frustration. Unlike the early versions of AMH, AWSEM does not heavily rely on the knowledge-based part of the Hamiltonian. The interactions are dominated by the physical components, with exception of only one knowledge-based term, which biases

the local structure within nine residue fragments, using a large number (usually up to 20) of conformations found in proteins containing analogous fragment sequences. This model was recently successfully used to study protein-protein association [71] and misfolding in multidomain proteins [72].

Current coarse-grained models are still not as accurate and predictive as all-atom force-fields. However, they make it possible to investigate complex bimolecular systems on several order of magnitudes larger time and length scales. This makes them an extremely valuable tool for interpreting mesoscopic and macroscopic experimental results, where the low to medium experimental resolution makes the use of all-atom simulations excessive.

1.2 Outline of Thesis

In Chapter 2 we present our results on protein structure prediction using AWSEM; a primarily physical protein coarse-grained model with knowledge-based local structure biasing term. As the discussed combination of the alpha-helical hydrogen bonding potential and the knowledge-based term has not been used before, we evaluated the structure prediction capability of AWSEM on 13 alpha-helical proteins. The three different levels of global homology between the target sequence and those proteins used for local structure biasing were tested, showing from medium to high quality predictions. Generally, AWSEM calculations produce structural predictions that are somewhat improved compared with the previous models. The inclusion of a small number of structures from homologous sequences improves struc-

ture prediction only marginally, but, when the fragment search is restricted to only homologous sequences, AWSEM can perform high resolution structure prediction and can be used for kinetics and dynamics studies.

In Chapter 3 we discuss the capability of AWSEM to predict protein-protein association. The non-bonded term of the potential was initially parameterized using a database of dimeric interfaces, but was later optimized to fold individual proteins. However, the ability of the model to predict binding interfaces was not previously tested. The present results demonstrate that the model indeed is able to predict the native interfaces of 8 homodimers and 4 heterodimers. We further discuss the importance of the monomer geometry and flexibility, and the significant role of non-native intermonomeric contacts in the association process for the homodimers. We find that even though the monomer geometry is important for correct association, it is a necessary, but not sufficient condition. Using a uniform inter-monomer contact potential, results in sampling of geometrically preferred misbound states, which are, however, energetically disfavored by AWSEM. Non-native contacts may play different roles in the association process, depending on the stability of the monomers. They can either stabilize a productive, on-pathway intermediates, catalyzing binding through a fly-casting mechanism in the case of unstable monomers, or create an off-pathway trap that obstruct binding. It has also been found that AWSEM correctly reproduces the two and three state behavior of obligatory and nonobligatory dimers.

In Chapter 4 we report on our work on protein folding in the presense of background noise. The one-bead Gō model was used to investigate the folding of two proteins in case of applied correlated random noise or harmonic fluctuations

of the strength of inter-residue interaction potential. The MD simulations show that existence of resonance in mean first-passage time of folding in both cases. When applied, correlated random noise or harmonic fluctuations result in increase transition rate for certain values of the characteristic parameter. In the case of the harmonic noise the resonance occurs when the fluctuation frequency approaches the folding rate of the unperturbed system. The former result was recently validated experimentally by Martin Gruebele and Max Platkov at University of Illinois at Urbana-Champaign using time-resolved Förster resonance energy transfer (FRET) technique. Despite the significantly smaller fluctuation amplitudes used, their results are consistent with the existence of the resonance.

The Chapter 5 is dedicated to further theoretical discussion of the phenomenon described in Chapter 4. In the first part we present the analytical kinetic theory of protein folding and unfolding under applied harmonic temperature wave (as it was done in the experiments), which does not explain the resonance phenomenon. Then, we discuses the results of the Brownian Dynamics (BD) simulations, which model the protein folding under the same conditions, and fully explain the results of the MD simulations. The comparison of the two approaches allows to interpret the experiments, and connect them to the results of the Molecular Dynamics simulations.

Chapter 2: AWSEM-MD: Protein Structure Prediction Using Coarse-grained Physical Potentials and Bioinformatically Based Local Structure Biasing

The chapter is based on the published work of the author:

A. Davtyan, N.P. Schafer, W. Zheng, C. Clementi, P.G. Wolynes, G.A. Papoian; J. Phys. Chem. B **116**(29), 8494–8503 (2012)

2.1 Introduction

Over the last decades what has been called "the Protein Folding Problem" [73] has evolved dramatically. Throughout this period both the practical and philosophical aspects of the problem have changed in the minds of scientists. Practical people want to find the structure of a protein from its sequence alone by whatever means necessary. Those of a more philosophical bent have been intrigued by the puzzle presented by a chain molecule organizing itself to a small family of structures in the face of incessant thermal buffeting, seemingly violating our notions of entropy. How this happens is probed in the laboratory through studies of folding kinetics, often by mutating various residues in the protein [74], to explore their contribution

to folding.

The hope has always been that conceptually understanding the physical process of folding will help in the practical task of structure prediction. Like the entwined histories of thermodynamics and the steam engine the interaction of the practical and theoretical sides of the folding problem has been mutually supportive. Interestingly, many of the key physical forces driving the folding process, in particular the hydrophobic interactions and the necessity for backbone hydrogen bonds, were predicted by Pauling and Kauzmann before crystal structure determination of proteins [75, 76]. It has turned out that many other more subtle interactions also contribute to precise sculpting of folding landscapes, with unique native basins that are kinetically accessible and these have been learnt in the process of improving structure prediction algorithms. Among these subtle forces, it has been shown that water-mediated interactions between hydrophilic residues are used as weak but specific forces that complement hydrophobic interactions and help guide early folding events [70, 77–79]. In addition, water-mediated interactions may allow larger proteins to partition into foldons, stabilizing intra-protein hydrophilic interfaces. We see there has been a decades long quest to identify key interactions stabilizing the native basins of globular proteins, which, in turn, has lead to subsequent improvements in the quality of structure prediction efforts.

The most powerful tool for practically predicting tertiary structure, however, remains homology: structure evolves more slowly than sequence so structures can be predicted if closely related molecules have already had their structures determined. This conservation of structure seems to be a consequence of the funneled nature of

real protein energy landscapes. Thus, while prediction by analogy does not explicitly use an understanding of the physical folding process, the funneled nature of the folding landscape is crucial. The funnel landscape ultimately is also responsible for the cooperativity of folding and is thus an essential feature of models of the folding process in the laboratory [80, 81]. Energy landscape theory allows the funneled nature of a landscape to be quantified [82–84]. Using this quantification, energy landscape theory has led to a way of learning the forms and parameters of energy functions for modeling folding kinetics and structure prediction by studying the database of existing structures. The main idea of the learning algorithm is that the folding landscape should be as strongly funneled as possible, while still remaining transferable from one sequence to another. Over the years this approach has led to a family of energy functions whose simulated dynamics mimics many observed features of laboratory folding and that also allow low resolution prediction of protein tertiary structure from sequence, even when no homology information is known ("*de novo* prediction"). In this paper, we report on a further development of this family of methods that uses local sequence similarity to encode structure short range in sequence while a coarse-grained water mediated interaction is used to determine tertiary structural themes.

The Associative memory, Water mediated, Structure and Energy Model (AWSEM) force field, presented in this work, is a direct successor in a series of protein structure prediction models [63–69] called early on the Associative Memory Hamiltonian (AMH) model because of its similarity to neural network models [85] and in later works was called the AMW model, to emphasize the addition of water-

mediated interactions [70, 78, 79]. The key idea behind AMH is to simultaneously sculpt deep folding funnels for multiple unrelated proteins, using the same set of parameters, which then produces a transferrable protein folding force field. The physical principle from landscape theory that drives the optimization learning algorithm in AMH/AMW is the maximization of the ratio of folding temperature[1] over glass transition temperature[2] for each training protein. While this key principle has remained steady over two decades, the underlying force field components have substantially grown in scope. Specifically, while the earlier versions of AMH had to rely almost entirely on the knowledge-based part of the Hamiltonian derived from global homology to memory proteins, the later iterations emphasized more and more the role of physical interactions, such as hydrogen bonds and water-mediated interactions which have a novel character going beyond Kauzmann's hydrophobicity. The AWSEM force field of the current work continues this tradition, and is actually dominated by the physical interactions. The only explicitly knowledge-based component of the AWSEM Hamiltonian is a term which biases local sequences that are of length nine residues or shorter, towards conformations found in proteins containing analogous fragment sequences. A related local fragment based approach has been successfully used by Baker and coworkers in a variety of works to assemble candidate conformations for protein structure prediction [86].

Even for this knowledge-based component based on peptide fragments, there exists a sound physical justification based on modern ideas of coarse-graining [87–

[1] At the folding temperature, the populations of the folded state and unfolded states are equal.
[2] Below the glass transition temperature, the dynamics of the protein chain is arrested.

89]. In the three-bead per residue structural model adopted in AWSEM, the vast number of original atomic degrees of freedom have been integrated out, both from the solvent and the protein. Hence, a priori, one expects this integration to result in a coarse-grained force field that contains a large number of complicated many-body terms, especially at the local in sequence level, where detailed interactions of specific neighboring sidechains may favor one local conformation over another. In terms of model building, the choice here is to either to determine explicitly what these many-body potentials are [90] and determine a huge number of associated parameters, or, alternatively, use similarity to local sequences in other proteins to infer the same many-body interactions using a knowledge-based approach. The latter is the strategy adopted in AWSEM and it seems to be a useful compromise that one needs to make for coarse-grained protein structure prediction in the foreseeable future.

The idea that a significant amount of the funneling of the folding landscape lies in the short range in sequence details is consistent with our knowledge of the thermodynamics of peptide fragments. Saven and Wolynes [91] showed that local structural signals which only weakly bias the helical state of peptides become much more effective when the protein chain has collapsed and, indeed, if they are not in conflict with tertiary structure should provide more than a third of the native structure seeking energy gap in the folding funnel. In this regard, local fragment energy terms are also appropriate as a realistic first step in describing laboratory kinetics faithfully.

While the efficacy of combining fragment energy terms with water mediated interactions has already been established [92], the specific combination of elements

in the current combination of physical potentials, such as the alpha-helical hydrogen bonding potential along with the locally determined fragment memory potential have not been studied before. In addition, as a significant technological improvement in its computer implementation, AWSEM has been written from ground up as new software in C++, leveraging the popular LAMMPS molecular dynamics package [27]. This flexible implementation, in turn, provides opportunities for applying AWSEM to modeling situations that were difficult to program because of the limitations of the previous FORTRAN codes for AMH and AMW. In particular, assembly of multi-protein complexes, interactions of proteins with coarse-grained models of DNA [93], mechanical pulling [94], and many other studies now become straightforward. The AWSEM MD package is available for download as an open source software (http://code.google.com/p/awsemmd/).

In this work, we have benchmarked the AWSEM code by predicting folding of 13 alpha-helical proteins which we have studied before with earlier versions of the AMW [78]. Not surprisingly, the quality of predictions depends on the fraction of global homologs that are similar to the particular target protein in the fragment memory database. To quantitatively explore this issue, we prepared three database versions that mimic practical situations one encounters in real life structural prediction: 1) homologs excluded, 2) homologs allowed, and 3) homolog-only. The homologs excluded version is tantamount to the situation one faces in predicting a new fold, a fold currently unrepresented in the structural database. For smaller proteins, such "novel" folds are becoming ever more rare. We found that for "homologs excluded" databases, the predictions from AWSEM were slightly improved

over previous AMW results, where for two proteins, 1R69 and 3ICB, impressive improvements are achieved. Especially for larger proteins, over 100 residues, inclusion of a few homologs can result in somewhat better predictions but for smaller proteins the effect is marginal. Allowing the inclusion of some homologs mimics the practical situation where one may be unaware there are, in fact, structural homologs available because they haven't been singled out by the alignment scheme. Finally, when the fragment memory database consists of only homologs, even distant ones, surprisingly high resolution predictions are made even for larger proteins. This homology only instantiation represents a common practical situation these days for smaller proteins where such distant homologs can often be recognized with sensitive alignment tools. Although specialized homology modeling algorithms, such as MODELLER [95–97], are already able to produce structures that are within 1 to 2 Å RMSD to the native structures vs. the 2 to 3 Å structures that are generated with AWSEM with "homologs only" fragment memories, the former very high quality results are based on a complete atomistic structural representation, while AWSEM is rather coarse-grained, with only three beads representing each residue. Because of its coarse-grained representation, AWSEM can be used to study the dynamics of real protein systems on experimentally relevant time scales using ordinary computer hardware. AWSEM provides an appealing alternative to purely structure based models, which are efficient and can be accurate but lack non-native interactions, and all atom simulations, which, while increasingly reliable, require specially designed computer hardware to access experimental time scales.

2.2 Methods

2.2.1 Model

According to AWSEM, the position and orientation of each amino acid residue is dictated by the positions of its C_α , C_β and O atoms (with the exception of glycine, which lacks a C_β atom). The positions of the other atoms in the backbone are calculated assuming an ideal peptide bond. A complete description of the structural model and the force field is given in the Appendix A. For the current study, we used only the alpha helical part of hydrogen bonding potential [70] and a variation of the associative memory term (herein denoted FM for "fragment memory"), which imposes a local bias using short, overlapping fragments of 9 residues or less. The total energy function is given in Equation 2.1,

$$V_{total} = V_{backbone} + V_{contact} + V_{burial} + V_{helical} + V_{FM} \tag{2.1}$$

$V_{backbone}$ is responsible for maintaining protein-like backbone geometries. The full form of the backbone potential is shown in Equation 2.2.

$$V_{backbone} = V_{con} + V_{chain} + V_{\chi} + V_{rama} + V_{excl} \tag{2.2}$$

V_{con} ensures the chain connectivity through number of harmonic bonds. The correct bond angles are achieved by the V_{chain} potential. V_{χ}, V_{rama}, and V_{excl} are responsible for chirality of the C_α atom, correct dihedral angle distribution, and inter-bead excluded volume interactions respectively.

$V_{contact}$, V_{burial}, and $V_{helical}$ are each based on a different aspect of protein

physics. $V_{contact}$ is an amino acid type dependent tertiary interaction term. It acts between pairs of residues which are 9 or more residues apart in sequence. In addition to being amino acid type dependent, the strength of the $V_{contact}$ potential also depends on distance separation and a local density. In the case of low local density, we say that the interactions are water-mediated and that they are protein-mediated in the opposite case. The burial term represents the preference of an amino acid of a specific type to be buried inside the protein or to be on the surface. Parameters for $V_{contact}$ and V_{burial} potentials were obtained by self-consistent optimization which maximizes the ratio of the folding temperature to the glass transition temperature for the model, $\frac{T_f}{T_g}$ [78].

$V_{helical}$ is an explicit hydrogen bonding term that acts between the carbonyl oxygen of residue i and the amide hydrogen of residue i+4. The strength of the interaction depends on the helical propensity of both residues participating in the interaction. This potential was recently introduced in the work of V. Oklejas *et al* [70].

V_{FM} is a purely bioinformatical term, and makes use of available experimental information from the RCSB PDB [52]. The form of V_{FM} is given in Equation 2.3

$$V_{FM} = -\lambda_{FM} \sum_{m} \sum_{ij} \exp\left[-\frac{(r_{ij} - r_{ij}^{m})^2}{2\sigma_{ij}^2}\right] \tag{2.3}$$

where the outer sum is over aligned memory fragments, and the inner sum is over all possible pairs of C_α and C_β atoms within the memory fragment that are separated by two or more residues. r_{ij} is the instantaneous distance between the atoms, r_{ij}^m is the corresponding distance in the memory fragment, λ_{FM} is a scaling factor

that can be used to change the strength of V_{FM} relative to other terms, and σ_{ij} is a sequence separation dependent width, which is given explicitly in the Supplementary Information.

2.2.2 Fragment Library

To generate the fragment memory libraries, we first used the online protein sequence culling server PISCES [98] to generate a database of sequences that has known structures in the PDB [52] with a resolution of 3 Å or better, and a specified maximum mutual sequence identity (MMSI). Two databases were generated for 80% and 95% MMSI. We then divided each target sequence into overlapping 9-residue segments and used PSI-BLAST [99] to find the 20 best matching fragments in the databases described above. We used PSI-BLAST's E-value to determine the quality of an alignment.

For each target sequence, we generated three different fragment libraries. For the first library, we excluded all related sequences from the search by setting an E-value cutoff in PSI-BLAST of 0.005. This typically leaves only those sequences with less than 20% sequence identity with the target sequence. We refer to this as the "homologs excluded" (HE) library. Predictions made with this library are similar to "free modeling" predictions, where no globally homologous sequences have experimentally resolved structures. For the second library, we will call "homologs allowed" (HA), we excluded a sequence from fragment search if and only if it had 95% or higher sequence identity with the target sequence. For the first two libraries,

we used PSI-BLAST to search the sequence database with 80% MMSI. For the third library, we used the sequence database with 95% MMSI and chose memory fragments only from sequences related to the target sequence, but again excluded sequences with 95% or higher sequence identity to the target sequence. We will refer to this library as the "homologs only" (HO) fragment library. As the number of related sequences in the database was typically small, we adjusted the strength of the V_{FM} term based on the average number of fragment memories found.

2.2.3 Targets

We looked at 13 alpha-helical proteins which were considered in an earlier work [78]. Some of them were used in past Critical Assessment of protein Structure Prediction (CASP) contests. The length of the target sequences ranged from 63 to 172 residues. Information about the target proteins is summarized in Table 2.1.

2.2.4 Simulation Protocol

All simulations were carried out using the LAMMPS molecular dynamics package [27], where we implemented the AWSEM force field. To evaluate the *de novo* structure prediction capability of our model, we first performed simulations with the "homologs excluded" fragment libraries for all target sequences. Next, to determine the effect of including fragments from globally homologous sequences, we performed a set of "homologs allowed" simulations on a subset of the proteins (see Figure 2.1). Finally, for seven of the target sequences, including the six largest, we performed

Table 2.1: **Target sequences information.**

Code	CASP Contest	Length	Homologs			
			Database with 80% MMSI		Database with 95% MMSI	
			Count	Best	Count	Best
1R69		63	1	52.38%	1	52.38%
1UZC	CASP5	69	1	40.00%	1	40.00%
1UTG		70	2	57.35%	2	57.35%
3ICB		75	15	78.67%	16	78.67%
1BG8	CASP3	76	0		0	
1N2Xb[1]	CASP5	101	1	51.92%	1	51.92%
256B		106	0		2	88.68%
4CPV		108	13	79.63%	19	79.63%
1CCR		111	14	64.08%	21	66.99%
1JWE	CASP3	114	4	48.21%	4	48.21%
2MHR		118	2	45.76%	2	45.76%
1MBA		146	20	31.03%	26	32.64%
2FHA		172	16	83.14%	21	94.77%

[1] b indicates domain

"homologs only" simulations, where the fragment memory search included only the homologs of the target sequence found in the database with 95% MMSI (see Figure 2.5 and Table 2.1). For each target sequence/fragment library combination, we ran 20 molecular dynamics annealing simulations starting from an extended conformation. We used the Nose-Hoover thermostat to cool the simulations over 4 million steps from above to below the folding transition temperature and recorded the coordinates every 1000 steps.

2.2.5 Analyses

To evaluate the predictive capability of our model, we calculated the structural similarity of all snapshots from the 20 trajectories of a given target sequence against the corresponding experimentally determined structure. As specific measures of similarity, both Q and RMSD were used, where Q is an order parameter which compares pairwise distances among residues between two structures, as elaborated below. It varies between 0 and 1, with higher values corresponding to higher similarity between the structures. The form of Q is given in Equation 2.4,

$$Q = \frac{2}{(N-2)(N-3)} \sum_{i<j-2} \exp\left[-\frac{(r_{ij} - r_{ij}^N)^2}{2\sigma_{ij}^2}\right], \tag{2.4}$$

where N is the total number of residues, r_{ij} is the instantaneous distance between C_α atoms of residues i and j, r_{ij}^N is the same distance in the experimentally determined structure and σ_{ij} is given as $\sigma_{ij} = (1 + |i - j|)^{0.15}$.

To demonstrate the prediction quality for each of our targets, we have plotted the best Q values from each of the 20 annealing runs, sorting them in descending order (see Figures 2.2, 2.3, 2.4, 2.9). These plots show how stable the predictions are, *i.e.*, what maximum Q values could be expected if fewer simulated annealing runs were performed.

We used the CE alignment server [100] to align the maximum Q structures with native structures for visual comparison; see Figures 2.5, 2.6, 2.7.

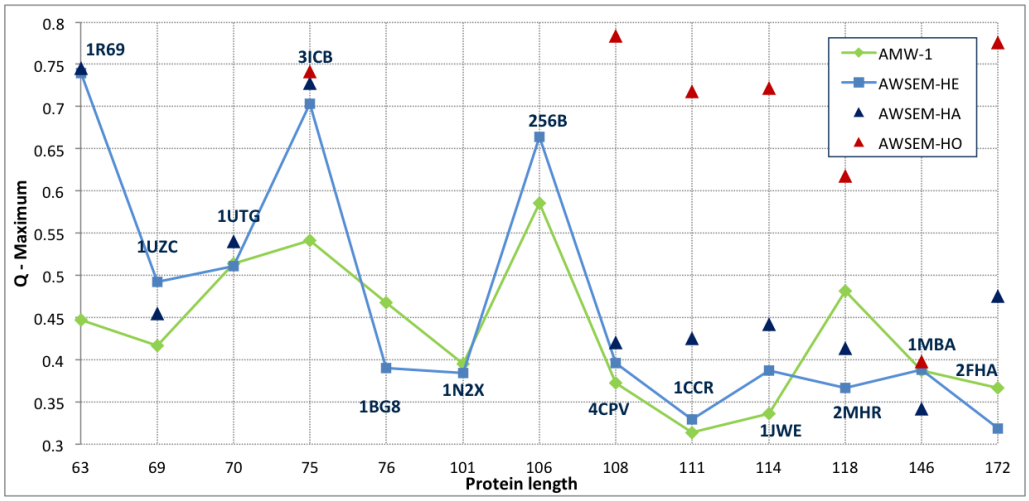


Figure 2.1: Maximum Q score versus sequence length for "homologs excluded" AWSEM (AWSEM-HE, light blue squares) and AMW-1 (green diamonds) models. Maximum Q scores for "homologs allowed" (AWSEM-HA, dark blue triangles) and "homologs only" (AWSEM-HO, red triangles) are also shown where available.

2.3 Results

We have summarized our structure prediction results in Figure 2.1, wherein we have plotted the maximum Q value achieved for a particular target sequence versus its sequence length. The 3 data sets are for the "homologs excluded" (light blue squares), "homologs allowed" (dark blue triangles) and "homologs only" (red triangles) fragment libraries. We have also plotted the AMW-1 results [78] (green diamonds) for comparison.

The results from both the "homologs excluded" and "homologs allowed" fragment libraries are overall slightly improved compared to the results of the AMW-1 model. The "homologs only" library, which we generated only for sequences with

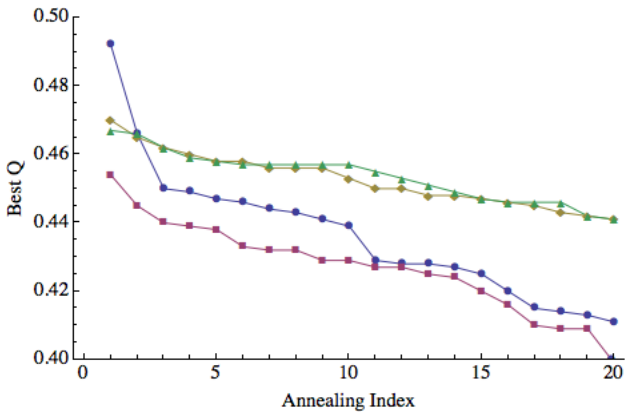


Figure 2.2: Prediction quality for 1UZC, including and excluding disordered region. For each of the 20 annealing simulations, the maximum Q values obtained are plotted in descending order. Blue circles correspond to "homologs excluded" predictions and red squares to "homologs allowed" predictions when the disordered region is included in the calculation of Q. Green triangles correspond to "homologs excluded" predictions and orange diamonds to "homologs allowed" predictions when the disordered region is excluded from the calculation of Q.

a sufficient number of homologs in our culled database, significantly outperformed the AMW-1, "homologs excluded" and "homologs allowed" models for all target sequences except 1MBA and 3ICB.

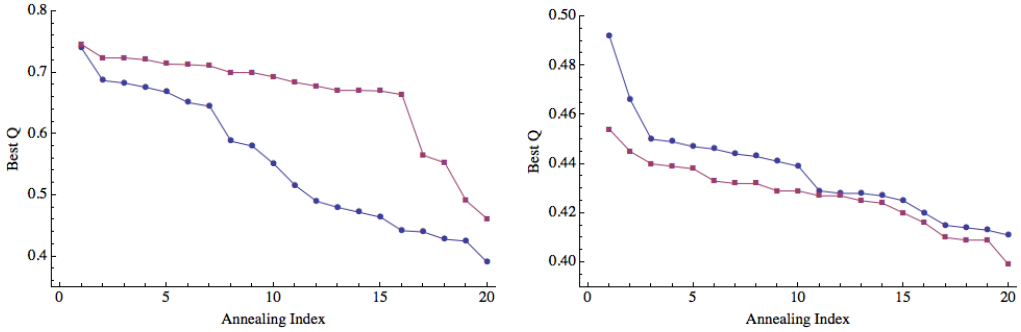
Maximum Q values for each of the 20 annealing runs (sorted in descending order) are shown in Figure 2.3 and Figure 2.4. These figures show that, in most cases, the predictions are stable, meaning that performing only 5 to 10 annealing runs would have yielded a similar maximum Q value. Two exceptions worth mentioning here are the "homologs only" prediction for 1JWE, and the "homologs excluded" prediction for 1UZC. For the former, the maximum Q value of 0.7 is the only point above $Q = 0.4$. For the latter, there is a more modest "jump" from $Q = 0.45$ to $Q = 0.47$ and 0.49. A close examination of the results for 1UZC indicated that a

disordered region on the N-terminal was likely responsible for the erratic results. Figure 2.2 shows the results when this 11 residue segment was excluded from the calculations of Q. Without this region, the prediction is better on average and significantly more stable.

Finally, we compared our "homologs only" results with the popular comparative structure prediction package MODELLER [95–97] using the same homologs that were used for the "homologs only" simulations. The results are summarized in Figure 2.8, where blue squares are the best RMSD values for "homologs only" AWSEM, and orange diamonds are the MODELLER results.

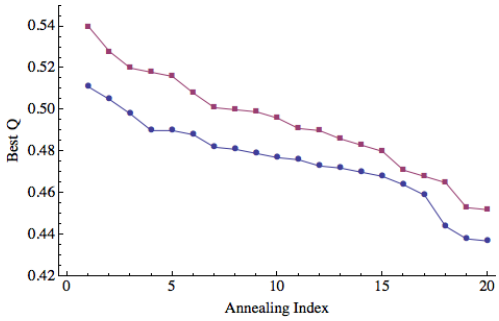
2.4 Discussion

As shown in Figure 2.1, the predictions made by AWSEM using the "homologs excluded" fragment library are in general improved compared to the AMW-1 results [78]. Before giving a more comprehensive comparison, we will briefly mention the key differences between AWSEM and AMW-1. The two models share the same backbone, direct contact, protein/water-mediated contact and burial potentials. However, AMW-1 used globally aligned protein sequences to specify associative memory interactions, whereas AWSEM uses short fragments to bias the local conformational search. In addition, AWSEM includes an explicit helical hydrogen bonding potential, and does not use a radius of gyration biasing term. The latter was shown to play an important role in correctly predicting the structure of large, non-spherical proteins [79].

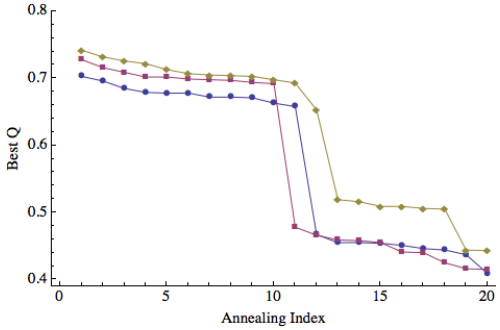


(a) 1R69

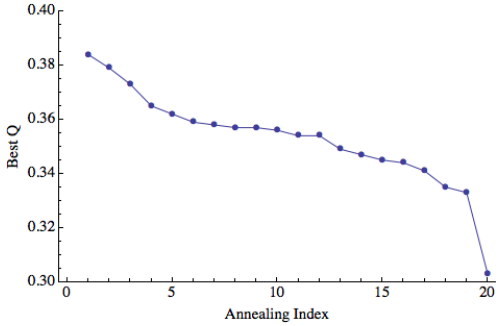
(b) 1UZC



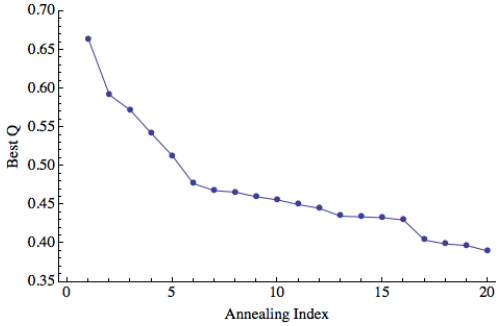
(c) 1UTG



(d) 3ICB

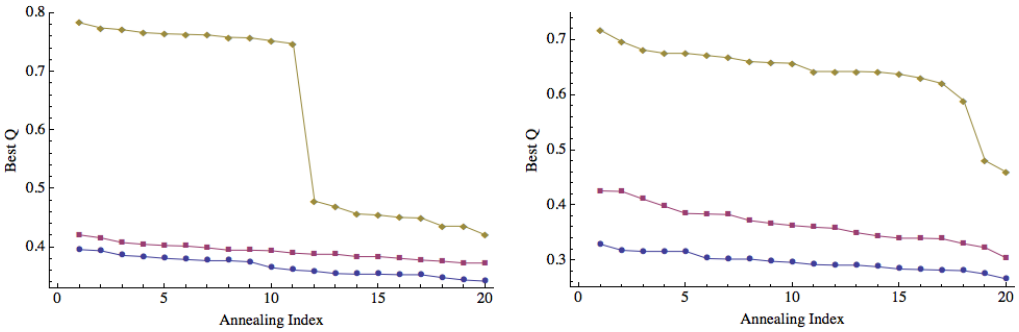


(e) 1N2Xb



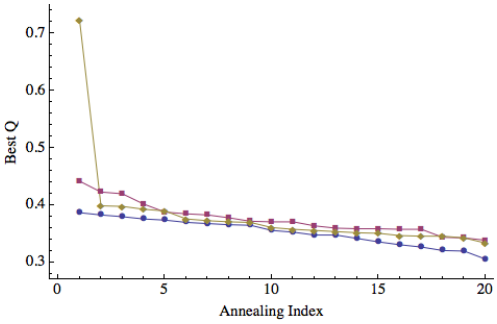
(f) 256B

Figure 2.3: Prediction quality for 1R69 (a), 1UZC (b), 1UTG (c), 3ICB (d), 1N2Xb (e), and 256B (f). Blue circles correspond to "homologs excluded" predictions, red squares to "homologs allowed" predictions and orange diamonds correspond to "homologs only" predictions.

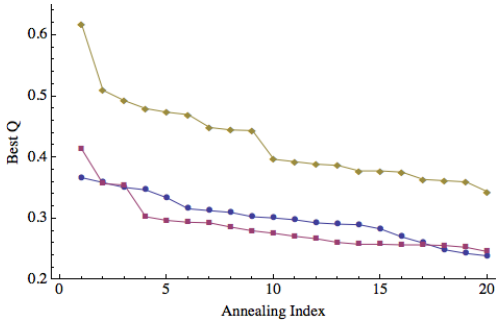


(a) 4CPV

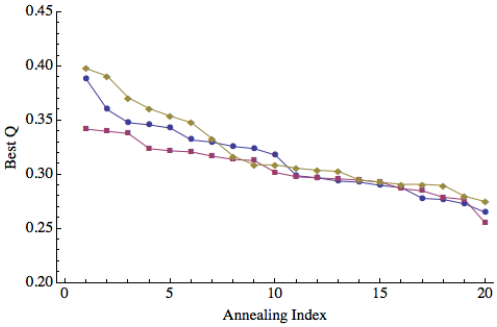
(b) 1CCR



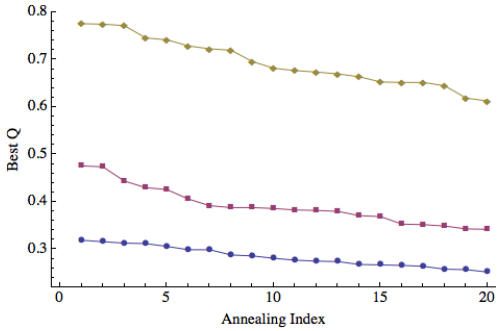
(c) 1JWE



(d) 2MHR



(e) 1MBA



(f) 2FHA

Figure 2.4: Prediction quality for 4CPV (a), 1CCR (b), 1JWE (c), 2MHR (d), 1MBA (e), and 2FHA (f). Blue circles correspond to "homologs excluded" predictions, red squares to "homologs allowed" predictions and orange diamonds correspond to "homologs only" predictions.

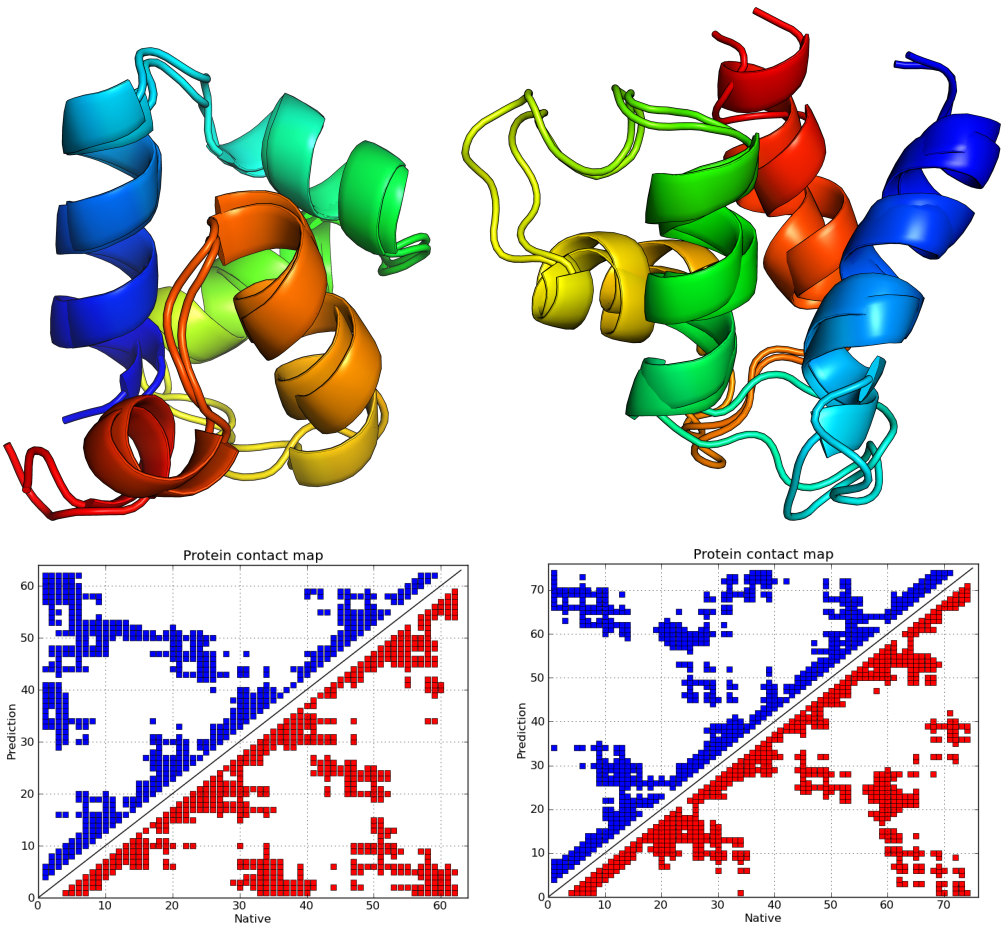


Figure 2.5: Structural alignments and comparative contact maps of the maximum Q score structures obtained from "homologs excluded" predictions for 1R69 (on the left, $Q = 0.74$, RMSD 1.6Å) and 3ICB (on the right, $Q = 0.703$, RMSD 2.4Å).

For 1R69 and 3ICB, maximum Q values of $\sim$ 0.75 and $\sim$ 0.7 are highly significant improvements of ~ 0.3 and ~ 0.15, respectively, compared to the AMW-1 predictions. Figure 2.5 shows an alignment of the predicted and native structures, and comparative contact maps for 1R69 and 3ICB, which indicate precise prediction of all secondary structure elements as well as good agreement of the global folds. AWSEM predictions of 1BG8, 2MHR and 2FHA were slightly worse than those of

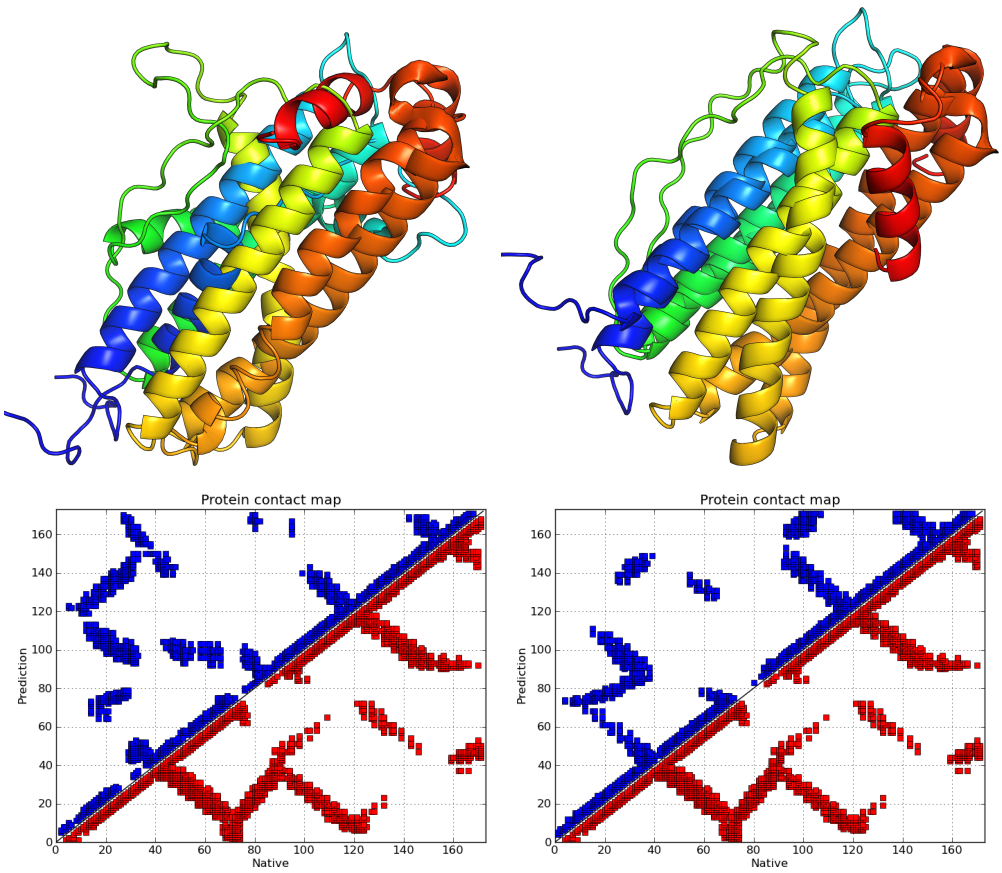


Figure 2.6: Structural alignments and comparative contact maps of the maximum Q score structures for 2FHA, with the "homologs excluded" prediction on the left ($Q = 0.319$, RMSD 12.383Å) and the "homologs allowed" prediction on the right ($Q = 0.476$, RMSD 8.781Å).

AMW-1.

The number of homologs available for each sequence varied from one to twenty (see Table 2.1). By performing predictions with the "homologs allowed" fragment library, we determined that the effect of including fragments from globally homologous sequences among other fragments from non-homologous sequences on the quality of prediction is small. In fact, the improvement was statistically significant for four proteins, of which only two had a change in the maximum Q value of 0.1 or more.

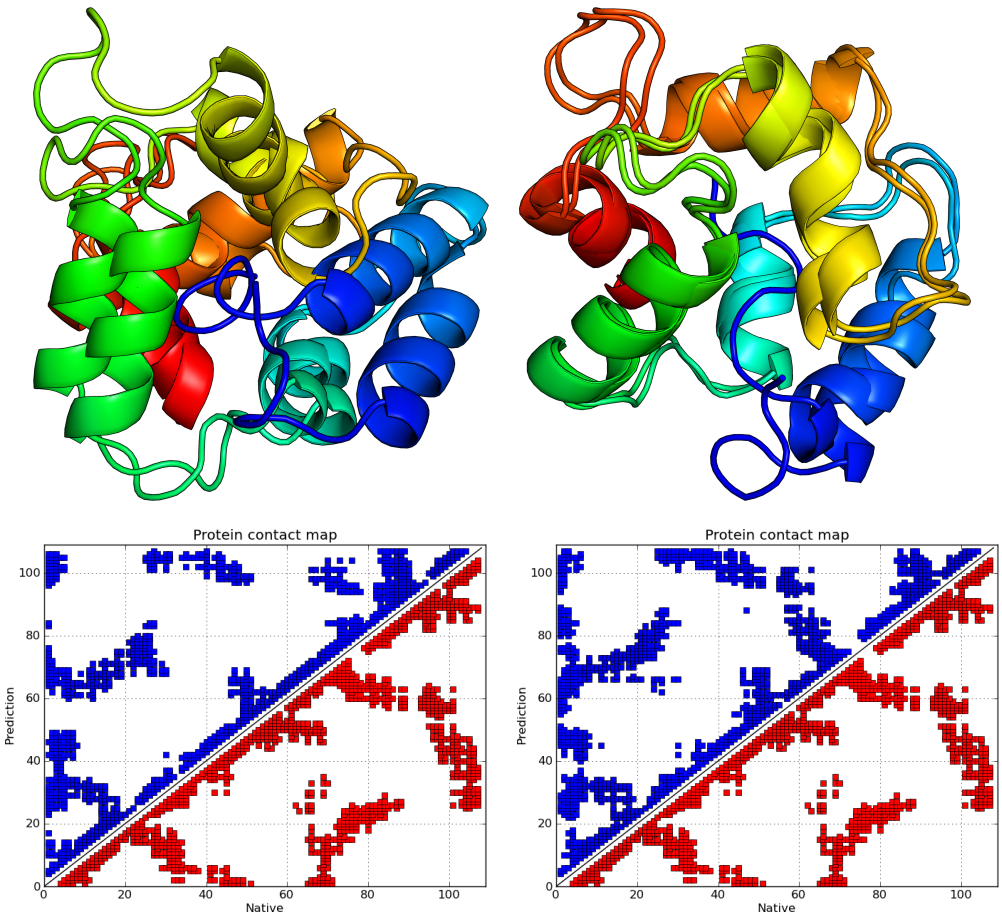


Figure 2.7: Structural alignments and comparative contact maps of the maximum Q score structures for 4CPV, with the "homologs excluded" prediction on the left ($Q = 0.396$, RMSD 5.8Å) and the "homologs only" prediction on the right ($Q = 0.784$, RMSD 1.3Å).

Specifically, the maximum Q values for 1CCR and 2FHA improved by 0.1 (from 0.33 to 0.43) and by 0.16 (from 0.319 to 0.474), respectively. The improvement for 2FHA can be seen in the structural alignment and contact maps in Figure 2.6. Unlike the "homologs excluded" prediction, wherein only 3 of the 5 helices are well formed, in the "homologs allowed" prediction all helices are formed and 4 of them, with the exception of the small C-terminal helix, have the correct mutual orienta-

tion and packing. This is particularly impressive given the size (172 residues) and non-symmetric shape of 2FHA.

For five of the seven targets predicted using the "homologs only" library, AWSEM achieved a maximum Q greater than 0.7. For 2MHR, a maximum $Q = 0.62$ and minimum RMSD of 3.44Å was obtained. For 1MBA, the maximum Q obtained was 0.4. To evaluate these results, we compared them with structure prediction results obtained using the MODELLER package. This package can do all-atom comparative modeling of proteins using experimentally determined structures, and their sequence alignments with the target sequence by satisfying spatial restraints. MODELLER was able to predict the structure of all larger proteins within 2Å RMSD resolution (Figure 2.8). Except for 1MBA, the difference in RMSD between the AWSEM prediction and the MODELLER prediction is between 1 and 2 Å . This implies that, despite being a coarse-grained model lacking explicit side chains, AWSEM can be used to make high resolution predictions for sequences that have homologs with experimentally determined structures.

There are several possible contributing factors to AWSEM's relatively poor prediction of 1MBA. Of all the target sequences, 1MBA has the homologs with the lowest sequence identity, with a maximum of 32.64%. As a result, even though there are 26 homologs in the database with 95% MMSI, the number of fragments assigned per position varied from 0 to 14 with an average value of 3. This inhomogeneity cannot be overcome simply by scaling the strength of the fragment memory term. In such cases it would be useful to introduce a smarter normalization and weighting scheme within the fragment memory potential based on the number of interactions

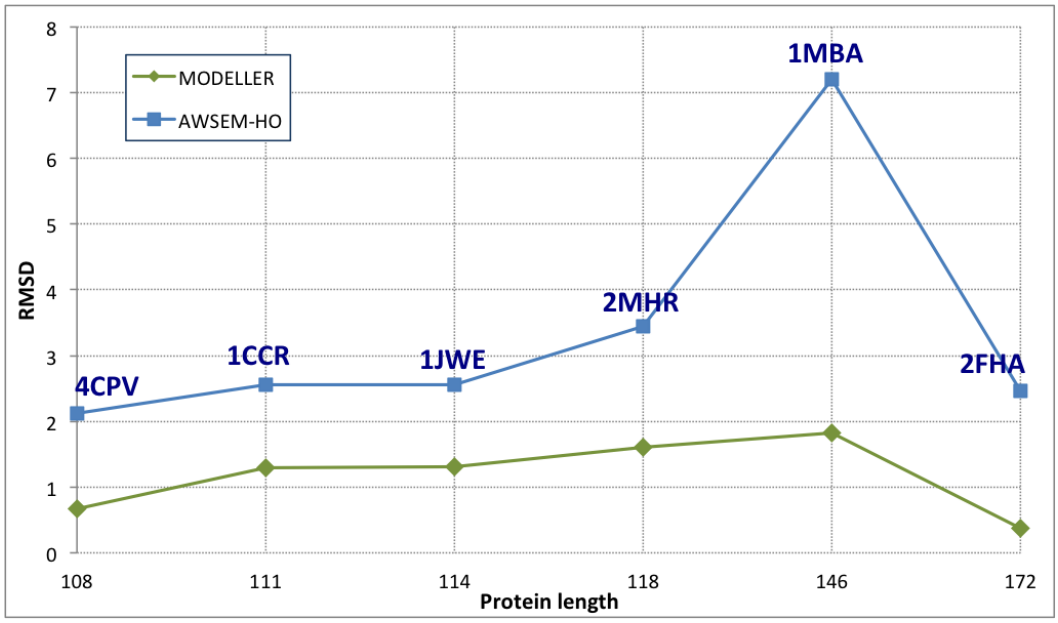


Figure 2.8: Comparison of MODELLER (green diamonds) and AWSEM (blue squares) prediction quality, showing RMSD in Å to the experimental structure versus sequence length in amino acids.

per residue, fragment length and alignment quality. The fragment memory potential could also potentially be improved by optimizing with respect to the fragment length and fragments per position. We did not test these possibilities here. Finally, unlike MODELLER, AWSEM lacks all-atom side chains, which may play an important role in 3 dimensional packing. This type of effect might accumulate and become particularly important for large proteins, such as 1MBA (146 residues). On the other hand, we should also bear in mind that MBA has a heme cofactor which is entirely omitted in our present simulations.

Another important factor to consider when analyzing the quality of prediction results is the presence of disordered and flexible terminal regions (or tails). Because these regions lack a static structure, "errors" in the prediction of these regions will

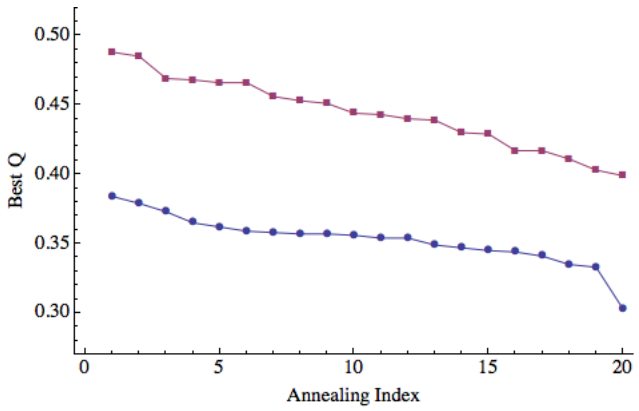


Figure 2.9: Prediction quality for 1N2X, including and excluding the first 22 residues, a disordered region. For each of the 20 annealing simulations, the maximum Q values obtained are plotted in descending order. Blue circles correspond to the maximum Q values when the disordered region is included and the red squares correspond to the maximum Q value when the disordered region is excluded from the calculation of Q.

have the effect of artificially lowering and broadening the distribution of Q values and RMSD scores we get. This broadening effect is apparent in Figure 2.2, where exclusion of the flexible tail from Q calculations of 1UZC collapses the "homologs excluded" and "homologs allowed" results, causing them to both be more similar to each other and making them individually more stable. Similarly, excluding the flexible tail (first 22 residues) from the Q calculations of 1N2X (see Figure 2.9) systematically increases the maximum we obtain Q in each simulation by 0.1.

2.5 Conclusions

Steady progress has been made in the last two decades in addressing the practical aspect of the "Protein Folding Problem", namely predicting the three-dimensional structures of proteins from their sequence. While early efforts were

almost exclusively based on knowledge-based potentials, more recent work uses a mix of physical and bioinformatic approaches. The rapid advances in designing and building specialized computer hardware already allow the use of all-atom explicit solvent simulations to successfully predict structures of some small proteins [21]. Nevertheless, given that the average human protein is over 400 residues, and many important and poorly understood biological processes involve complex multi-protein or nucleic acid assemblies, it will be rather difficult to apply atomistic simulations to routinely address these large length- and time-scales processes for some time. Hence, there remains a significant need for the development of coarse-grained, yet preferably accurate protein force fields. Most prior kinetic and mechanistic studies using coarse-grained protein force fields relied on native structure based approaches, which assign favorable interactions to native contacts, giving in concrete terms a folding funnel. While such approaches are physically meaningful, rooted in the energy landscape theory of protein folding, they can underestimate or often completely ignore the role of non-native interactions, cannot be used for proteins without solved structure, and also cannot be directly applied without modification to partially or fully disordered proteins. The above discussion underlines a need for development of a coarse-grained protein force field which is substantially based on known physical interactions, is amenable to Molecular Dynamics simulations and can be used for both *de novo* protein structure prediction as well as for probing protein folding and dynamics.

AWSEM, which is a successor to the AMH and AMW approaches to protein structure prediction, represents one such force field. It combines a large number

of physical interactions, from backbone terms to direct- and water-mediated interactions and hydrogen bonding, with structural biases that are local in sequences, based on the alignments of fragments of nine-residues or less of the target protein to the local segments found in a protein database. The force field was implemented from the ground up in C++, leveraging the LAMMPS molecular dynamics package. It can be used not only for protein structure prediction, but also, for example, to study protein folding kinetics, functional dynamics of the native state, and binding and folding processes. In ongoing works, our research groups plan to explore the extensions of AWSEM to simulate disordered proteins and interactions of proteins with membranes and DNA.

In this work, we have shown that the best structures produced by AWSEM in the "blind prediction mode", where we ensured that no global homologs were included in the local fragments database, were either comparable in quality or improved over the prior AMW efforts in blind prediction. We have also analyzed the consistency of prediction runs. We find that when poorly-defined loops or tails are excluded from the structural comparisons, then there is considerable consistency between different runs for almost all proteins. For some proteins, such as 1R69 and 3ICB, impressive predictions were achieved, with 1.6 Å and 2.4 Å RMSDs to the corresponding native structures. For larger proteins, over 100 residues, the consistency of predictions has somewhat improved compared to AMW. For these larger proteins, AWSEM obtains maximal Q values in the range of 0.35 to 0.4. This is often indicative of many native-like structural elements and even a roughly correct overall fold in some cases, but with a number of packing defects among the secondary

structural elements. How to take *de novo* structure prediction of large proteins to the high-resolution levels that are achievable for many smaller proteins is a challenging question, no doubt requiring further efforts in force field development and parameter optimization.

If the goal is not blind protein structure prediction, but instead investigation of protein folding kinetics and protein function, it may be advisable to bias the fragment library with homologs of the target protein, even distant ones. While exploring this possibility, we have shown in this work that even large proteins (on the order of 200 residues) fold to structures that are similar to the corresponding native structures within 1-3 Å RMSD. Hence, by appropriately tuning the fragment library, one may use AWSEM-based coarse-grained modeling of proteins either for *de novo* structure prediction, or in cases where the structures of distant homologs are known, kinetics and dynamics can be the main aims of the study. As an alternative to using experimentally determined structures for memories, snapshots of highly populated states sampled in atomistic simulations can be used as fragment memories for subsequent AWSEM coarse-grained simulations of the same protein [101].

Since AWSEM is an open-source package, many groups may choose to contribute to its further development and applications to new areas of research. The comprehensive description of the AWSEM force field, along with all force field parameters, are elaborated in the supplemental Information of [62] (also Appendix A), allowing the possibility of reimplementing AWSEM in alternative programming environments.

Chapter 3: Predictive Energy Landscapes for Protein-Protein Association

The chapter is based on the published work of the author:

W. Zheng, N. Schafer, A. Davtyan, G.A. Papoian, P.G. Wolynes; Proc. Natl. Acad. Sci. USA **109**(47), 19244–19429 (2012)

3.1 Introduction

Protein-protein interfaces encode information that is key to a molecular understanding of biological functions. The folding of proteins is well understood in the framework of energy landscape theory and its principle of minimal frustration. Are binding landscapes also funneled? Mechanistic consequences of funneled binding landscapes have been investigated using structure based models [47, 102–105]. The agreement of these mechanisms with observation suggests that binding landscapes are generally funneled, explaining why topology is indeed a major factor in determining binding mechanisms [47]. A statistical analysis of a large database of protein complexes revealed that for many of the complexes the binding energy gap is indeed larger than expected knowing the variance of the binding energy [106], the hallmark feature of a funneled landscape [80]. Further testing this idea, Papoian et al. dis-

covered that for other complexes, in order to have a funneled landscape for binding, unanticipated water-mediated interactions were required. They developed a water-mediated potential encoding these interactions [77]. This transferable potential was later optimized to create funneled folding landscapes that successfully predict the structure of monomeric proteins [62, 78]. Therefore there is considerable support for the idea that, like folding landscapes, protein-protein recognition landscapes are funneled.

In this paper, we test whether the AWSEM potential can predict binding interfaces, the problem which motivated its original invention. Unlike rigid docking programs [107–109], our approach uses molecular dynamics with simulated annealing to search for structures energetically favored by the AWSEM potential. While many docking protocols entail multiple stages [110] to accomplish interface prediction, including rigid body search to locate regions of interest [107,109,111] and refinement of docked structures and selecting the best models [108,112,113], simulated annealing of the AWSEM potential proves directly able to predict the binding interface of the dimers we have tested. The molecular dynamics implementation allows one also to compute free energy profiles in order to predict mechanisms. Using this predictive transferable potential model we now revisit the role of topology in determining binding mechanisms and explore the additional role played by non-native contacts in coupled folding and binding reactions.

3.2 Binding Interface Prediction

We used AWSEM to predict the binding interfaces of 8 homodimers and 4 heterodimers. The homodimers were previously studied with pure structure based models [47]. The only structural information used by AWSEM was local backbone information of the monomers from the PDB structure of the dimeric complex - no information about dimeric contacts was included. The tertiary contacts within the monomers are also not used as input. The input of native monomeric information guides only local-in-sequence structure formation. Both the tertiary contacts within the monomers and between the two monomers are determined by the same transferrable tertiary contact potential, which is described briefly in the Methods section and in full detail in the supplementary information of the paper of Davtyan et al. [62]. The starting states of all simulations consisted of two completely unfolded and unbound monomers, and molecular dynamics with simulated annealing was performed to search for the bound state.

As shown in Figures 3.1 and 3.2, the binding interfaces for the 12 dimers are generally very well predicted. One can argue that the interfaces of homodimers might be easier to predict because their binding interactions are usually stronger due to symmetry [114, 115]. Homodimers are in general observed in a symmetric binding geometry, where strong contacts on the interface are doubled. We therefore also tested 4 heterodimers with relatively weak interfaces, and AWSEM was able to predict the interfaces to similar accuracy as the homodimers discussed herein. The heterodimers that we tested have mostly hydrophilic interfaces, which are rel-

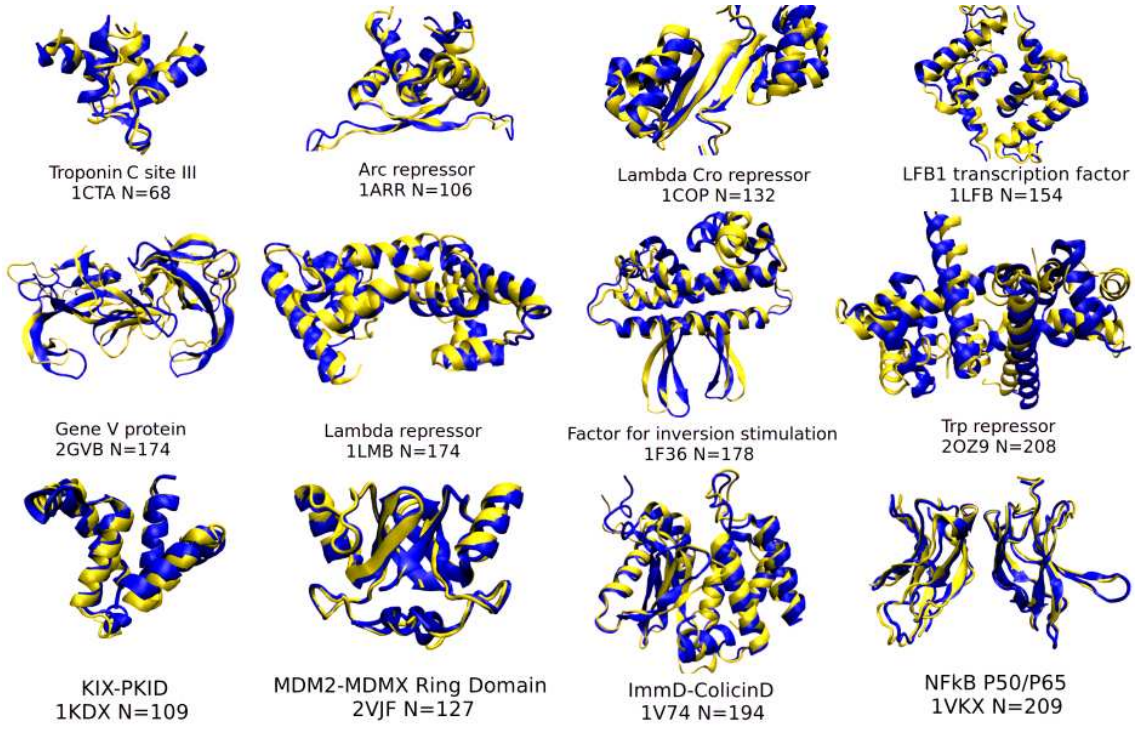


Figure 3.1: Snapshots of best predicted structures (yellow) using AWSEM, compared with the PDB structure (blue). The name of the proteins, their PDB ID and the number of residues are shown in the figure. The first 8 dimers are homodimers, the last 4 are heterodimers.

atively weak when compared to hydrophobic interfaces, and are therefore harder to predict. The water-mediated interactions (see *Methods* section) play a major role in predicting hydrophilic interfaces. When it is turned off, the prediction quality of dimers with hydrophilic interfaces get significantly worse, as shown in Figures B.3 and B.4. Having successfully passed the prediction test for both homodimers and heterodimes, AWSEM was applied to study the mechanism of homodimer binding in greater detail in the following part of the paper.

In Figure 3.2, the seemingly worst prediction in our test set is for the homeodomain of LFB1 (PDB ID: 1LFB). Its intermonomeric contacts in the PDB struc-

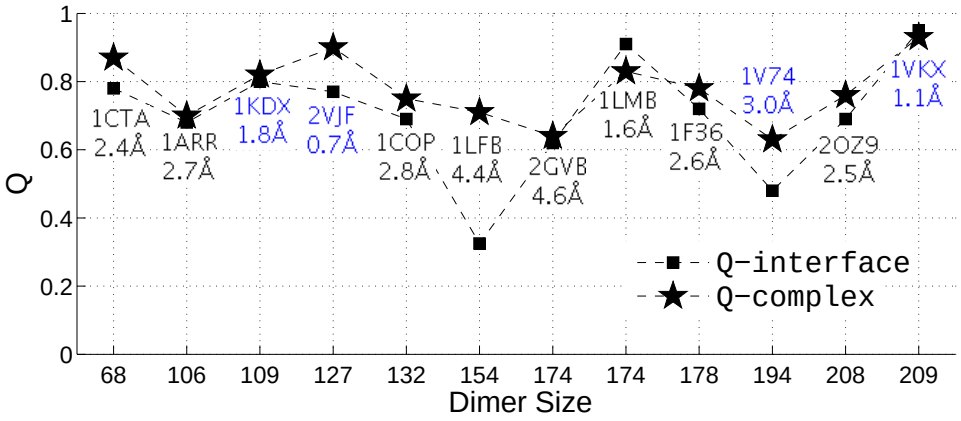


Figure 3.2: The accuracy of the AWSEM predictions is measured by Q and RMSD of the C_α atoms of the complex. PDB codes for homodimers and heterodimes are in black and blue, respectively. $Q_{complex}$ is shown as a star symbol and $Q_{interface}$ a square symbol. Note that for 1LFB, there is a relatively large difference between the two Q values. This can be explained by its very small ratio of the number of interfacial contacts to the number of total contacts, as shown in Figure B.1. Dimer size refers to the number of residues in the dimer complex. 20 or 40 independent annealing runs were performed for each dimer, starting from two monomers completely unfolded and separated. The final structure obtained at the end of the annealing runs with the best Q is selected for each dimer.

ture are weaker than for all other proteins in the test set. Significant native or non-native contacts only form at temperatures well below the binding temperatures[1] of the other proteins. The PDB structure with which we compared our prediction is in fact only a model, not a directly determined crystal structure, proposed by Ceska et. al. [116] and involves a simple 2-fold rotation of the crystallographically determined monomer structure. It has been suggested that the homeodomain might be dimeric when bound to DNA [117,118]. But we have been unable to find a crystal structure of the homeodomain of LFB1 in dimeric form in the presence or absence of DNA.

[1] At the binding temperature, the populations of the bound state and dissociated state are equal.

Inspired by the principle of minimal frustration, AWSEM was optimized by maximizing the ratio of the folding temperature[2] to the glass transition temperature[3], similar to the Z-score optimization algorithm. But the parameters found by optimization were developed using a training set containing only monomers. The success of the model in actually predicting binding structures buttresses the idea that the same energy landscape principles are applicable to binding processes as to monomeric folding. For Arc repressor (1ARR) and Lambda repressor (1LMB), Figure 3.4 shows the total energy of the predicted complex at the end of each annealing simulation as a function of $Q_{interface}$, the fraction of native contacts formed on the interface. Low energy structures are seen to correspond to near native states and there appear to be few competing (low energy but low $Q_{interface}$) traps.

3.3 Experimental and Theoretical Descriptions of Protein Dimers

Homodimers are often categorized as being either obligatory or nonobligatory dimers, meaning that the monomers must associate to complete folding (obligatory) or are stably folded in isolation at physiological temperature (nonobligatory). This distinction can be made in the laboratory by performing equilibrium denaturation experiments. In these experiments, obligatory dimers show only two states - one with both monomers unfolded (or partially folded) and the other with the native dimer structure - and are therefore sometimes referred to as two-state dimers. Nonobligatory dimers have three populated states under physiological condition -

[2]At the folding temperature, the populations of the folded state and unfolded states are equal.
[3]Below the glass transition temperature, the dynamics of the protein chain is arrested.

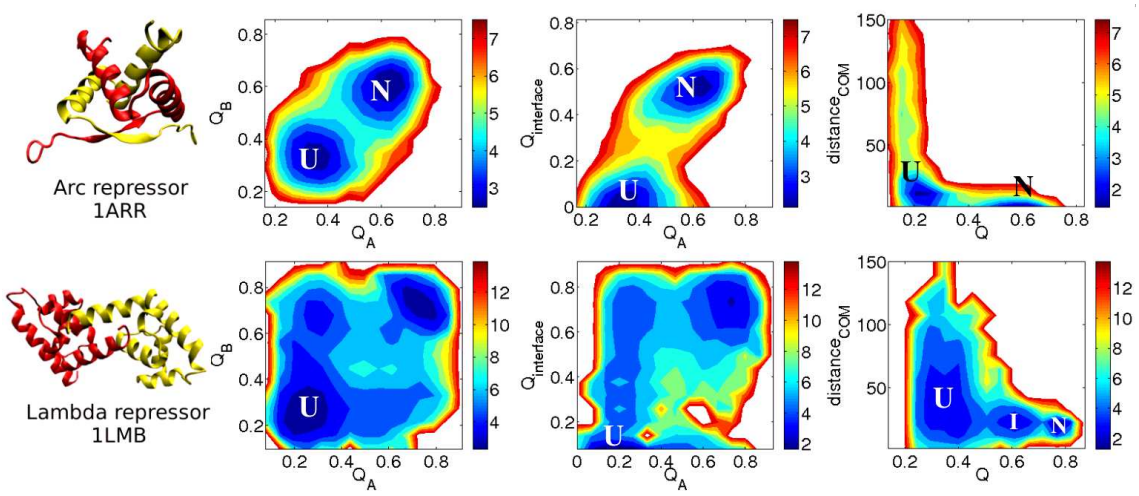


Figure 3.3: Free energy surfaces of folding and binding of obligatory (two-state) and nonobligatory (three-state) dimers obtained using AWSEM. Free energy surfaces are plotted as a function of the fraction of native contacts within the individual subunit Q_A (Q_B), $Q_{interface}$, Q of the complex and the distance between the centers of mass of the two subunits $distance_{COM}$. State U refers to the unfolded and unbound state, state N is the native bound state. The intermediate state I is observed only in the free energy plot of the non-obligatory dimer 1LMB. For 1LMB, the marcobasins that contain a mixture of different states are left unlabeled. The simulations reproduce the binding mechanism inferred from experimental and previous theoretical modeling results [47]. The free energy surfaces are calculated at the temperature where the heat capacity has a peak as a function of temperature.

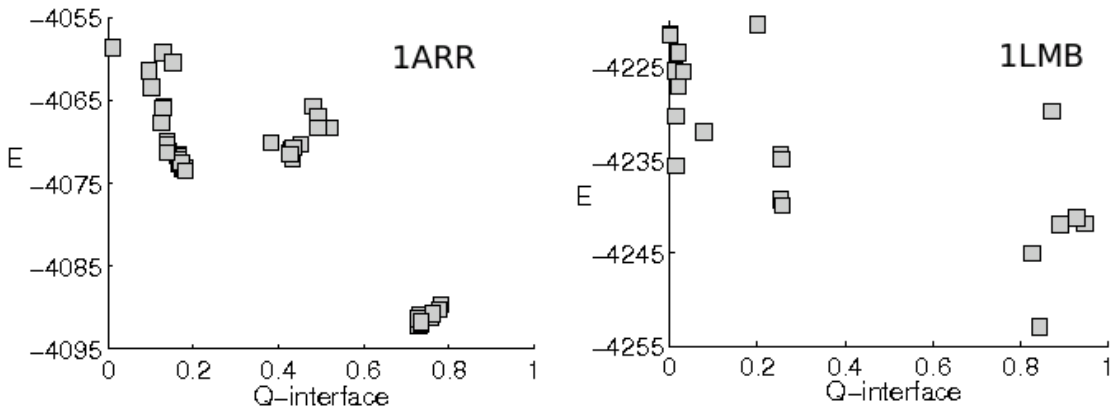


Figure 3.4: The total energies of the final complexes at the end of annealing simulations are plotted against $Q_{interface}$. Near native bound structures have lower energy than non-specific bound structures.

one with unfolded monomers, another with folded but unbound monomers, and yet another with the folded monomers bound together.

The binding-folding mechanism has been found to correlate with several global characterizations of the native dimer structure: interface hydrophobicity and the ratio of the number of interfacial contacts to the number of intramonomeric contacts being most important. A dimer with a highly hydrophobic interface and a large ratio of interfacial to monomeric contacts is typically two-state. Dimerization in these cases is, in some ways, reminiscent of monomeric protein folding in so far as the dimer as a whole can be thought of as a single domain folding cooperatively with the interface playing the part of the hydrophobic core. This type of folding mechanism is sometimes referred to as involving "induced fit" [119,120], meaning that the presence of the binding partner is needed to induce the monomer to adopt its folded structure. Nonobligatory dimers typically have more hydrophilic interfaces and smaller ratios

of interfacial contacts to monomeric contacts. These dimers associate via a lock and key type mechanism [121] wherein complementary interfacial geometry and favorable contact energies drive association.

Knowing the size and shape of the interface has often proved sufficient to determine whether a homodimer will associate via a two-state or three-state mechanism [47]; Using a structure-based model with uniform contact energies for only native interfacial contacts and native monomeric contacts, Levy et al. were able to accurately reconstruct the experimentally determined binding mechanisms for 11 homodimers. As shown in Figure 3.3, the current model also correctly reproduces the observed pattern of two-state and three-state behaviors for these examples. Two stable states are observed for two-state dimer Arc repressor, the unfolded, unbound state U and the native bound state N. There is no stable intermediate state, indicating a folding-upon-binding mechanism. For the three-state dimer Lambda repressor, on the other hand, there is an additional intermediate state I which consists of an ensemble of a variety of encounter complexes. These complexes have only one monomer folded and partially bound or are complexes in which both monomers have folded but remain unbound. The free energy surfaces calculated using AWSEM are consistent with the experimental observations and previous theoretical modeling results. Unlike the previously employed structure-based model, however, the current model which can predict dimer interfaces can also shed light on the role of non-native interactions in the association process.

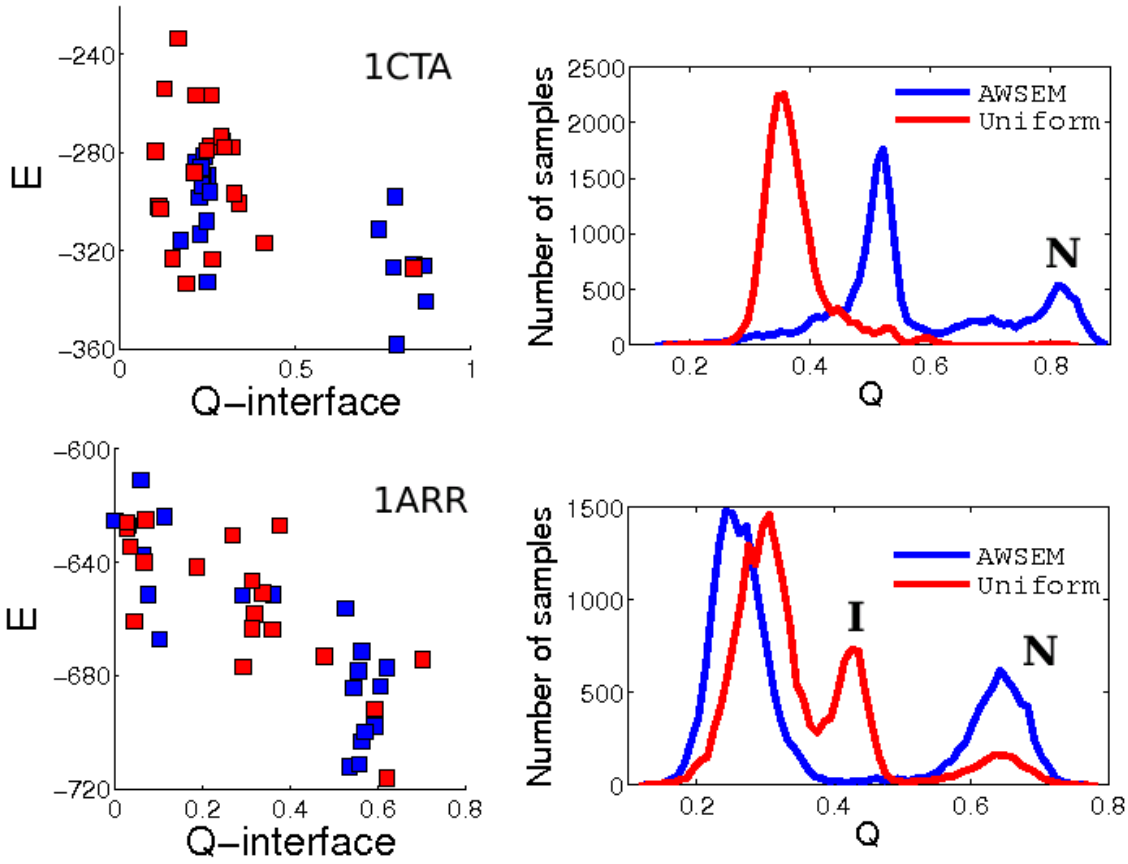


Figure 3.5: The AWSEM predictions vs. the predictions using a non-optimized energy function with uniform intermonomer contact strength for Troponin C site III (1CTA) and Arc repressor (1ARR). In the plots on the left, the energies of the final configurations from all simulations are plotted as a function of $Q_{interface}$. Blue and red colors are for the AWSEM predictions and the uniform predictions, respectively. In the plots on the right, the distribution of the number of samples collected from 20 simulations is plotted along Q. For 1CTA, the native bound state N is energetically less favored for uniform contact energy function than for AWSEM. For 1ARR, on the other hand, the native bound state is the lowest energy state for both energy functions. However, uniform intermonomeric contacts create an intermediate state I which drastically reduces the binding efficiency.

3.4 Role of Monomer Geometry in Interface Determination

One might argue that the successful predictions of dimer interfaces could be due to geometrical factors related to the limited number of ways that two dimers of prescribed geometry can associate. In this case of course the monomer is flexible so doesn't have a fixed tertiary structure *a priori*. Nevertheless, to investigate the possible role of monomer geometry by itself, we changed the strength of any contacts between monomers to have a residue-independent, uniform value while retaining the transferable potential within the monomers. The strength of the intermonomeric interaction is rescaled so that the stability of the native bound state was the same as with the AWSEM potential. New prediction simulations using the uniform intermonomer contact strength with the same annealing schedule were performed. The simulations using the uniform intermonomer contact energy were significantly worse than the AWSEM predictions that used the optimized potential. Interestingly, the effects of changing the intermonomer contact strength to a uniform value are different for different dimers. For example, as shown in Figure 3.5, for Troponin C (1CTA), with uniform contact energies the native bound state is no longer an energetically favored state. Instead there are numerous misbound states with more favorable binding energies than the native. For Troponin C the native bound state is not the state with the maximal number of intermonomeric contacts. On the other hand, for arc repressor (1ARR), the native bound state does still remain the lowest energy state when the contacts have uniform weight. Nevertheless, uniform intermonomeric contacts create an intermediate state I in this system which drastically

reduces the binding efficiency.

The structure of the arc repressor monomer in the bound dimer consists of two helices and a beta strand, and the resulting dimer interface forces the two monomers to significantly intertwine. The large size of the interface allows the uniform interaction energy described above to still favor the correct bound structure, albeit with a rougher landscape as indicated by the presence of misbound structures encountered during annealing. Energetic heterogeneity is not the only contributor to high binding efficiency. In instances where the native binding interface geometry forces the monomers to interweave, the flexibility of the local structure of the monomers also modulates the binding efficiency, as shown in Figure B.2. When the strength of the energetic term encoding the local in sequence structure bias is decreased, the percentage of successful binding simulations at first increases but finally decreases when the local bias becomes too weak. This is consistent with the suggestion that flexibility allows proteins to adjust to achieve optimal fit upon binding in order to perform specific biological functions [122]. Binding is a dynamic process on a funneled landscape; Geometry of the monomers alone does not completely explain the binding process.

3.5 Role of Nonnative Contacts in Dimer Formation and the Fly-casting Mechanism

The water-mediated potential in AWSEM is a transferable potential that can be used to model the intermonomer tertiary interactions. This part of the model

allows us to study the role of non-native intermonomeric contacts in dimer formation. In order to discuss the role of non-native interactions, it is informative to single out a special class of non-native contacts called swapped contacts. The name comes from a type of intermonomer contact pair that is observed in domain-swapped dimers [123, 124]. Swapped contacts are defined as non-native intermonomeric contacts formed between the ith residue in monomer A and jth residue in monomer B that correspond to i and j being a native contact pair within the monomer. Note that sometimes there are pairs of residue indices (i, j) corresponding to a native monomer contact pair that are also native interfacial contact pairs. These contact pairs are excluded from the computation of the number of swapped contacts since they are considered to be native contacts. The swapped contacts are of special importance in dimer association because they are on average stronger than other random contacts that have no analog in the native monomer structure. According to the principle of minimal frustration, the native contacts within a stable monomeric protein are on average stronger than other random contacts so, likewise, swapped contacts are more stable than random ones.

Non-native interactions play different roles for obligatory and non-obligatory dimers as seen in Figure 3.6. An example of an obligatory dimer, Arc repressor, is shown on the top of Figure 3.6. States stabilized by non-native interactions correspond to on-pathway intermediates that catalyze the association process through a fly-casting mechanism [125]; The individual monomers, which are both in extended conformations before the association, have significantly larger capture radii than those of the folded monomers. The large capture radius increases the rate of

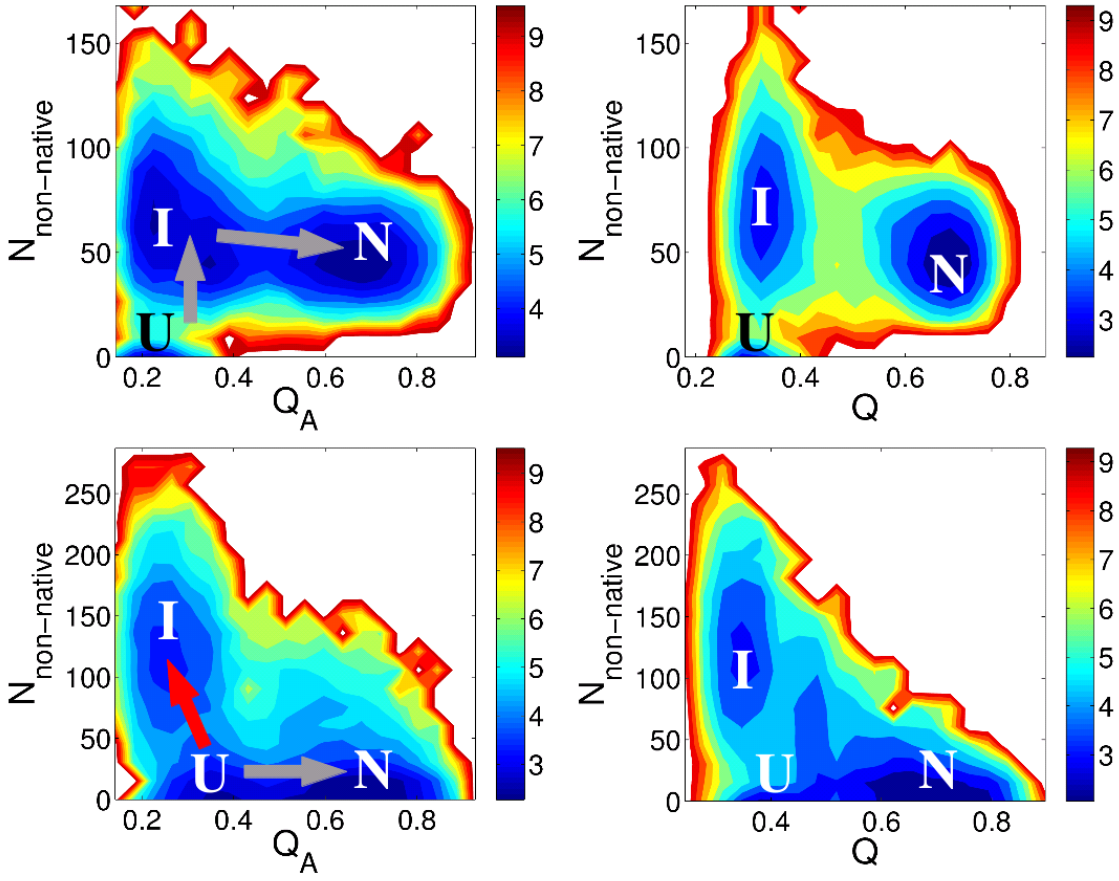


Figure 3.6: For Arc repressor (top) and Lambda repressor (bottom), free energy surfaces at the folding temperature are plotted as a function of the number of non-native intermonomeric contacts $N_{non-native}$, Q_A and Q of the complex. I, U and N stand for intermediate, unbound and native bound states, respectively. Non-native interactions have different consequences for obligatory and non-obligatory dimers; In the case of obligatory dimers, as shown on the top of the figure, states stabilized by non-native interactions correspond to on-pathway (indicated as gray arrow) intermediates that can catalyze the association process through fly-casting mechanism. In the case of non-obligatory dimers, these states appear to be off-pathway (indicated as red arrow) and can thereby impede binding by acting as a trap.

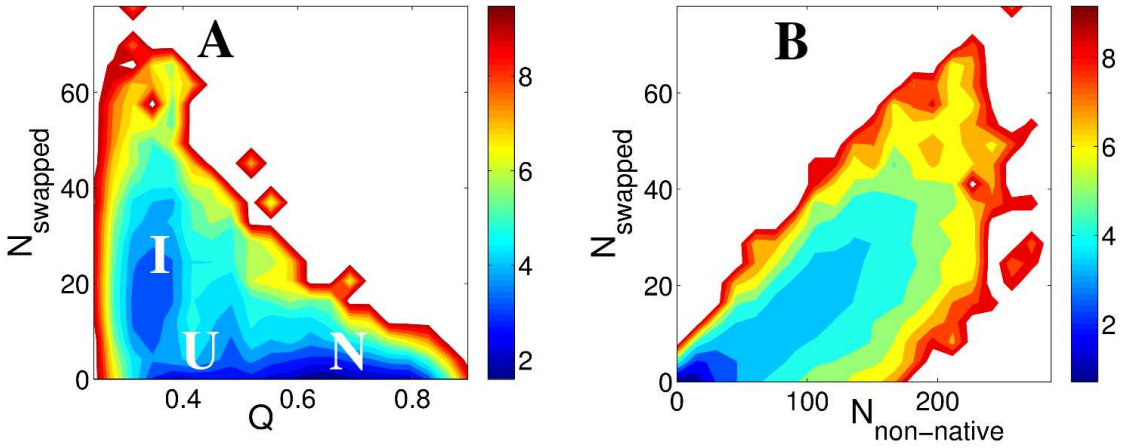


Figure 3.7: Free energy surfaces as a function of $N_{swapped}$ for Lambda repressor (1LMB). *A*) Similar as in Figure 3.6, we observe an off-pathway intermediate state, stabilized by the swapped contacts. *B*) There is a linear increase of the number of swapped contacts when the number of non-native contacts increases. The ratio of the number of swapped contacts to the number of non-native contacts is about $10 \sim 20\%$. These suggest that the intermediate state stabilized by non-native contact pairs contains a significant number of swapped contacts.

binding. In the case of non-obligatory dimers, however, the states with non-native contacts appear to be off-pathway and impede binding by acting as kinetic traps. We investigated further these off-pathway intermediates for the case of Lambda repressor (1LMB). In Figure 3.7, free energy surfaces of Lambda repressor are plotted as functions of the number of swapped contacts $N_{swapped}$, Q and $N_{non-native}$. As in Figure 3.6, we observe an off-pathway intermediate state stabilized by swapped contacts in the left plot of Figure 3.7. The plot on the right shows a linear increase in the number of swapped contacts when the number of non-native contacts increases. Figure 3.7 suggests that the intermediate state I consists of a significant number of swapped contact pairs. These intermediate states stabilized by swapped contacts

are kinetic traps in the binding of non-obligatory dimers. If both of the monomers of a non-obligatory dimer are significantly unfolded when they encounter each other, they may fall into the trap state I as shown in Figures 3.6 and 3.7.

To summarize the roles of the different types of contacts during binding, we plot their average contact strength and their contributions to the total binding energy against Q in Figure 3.8. As the complex approaches the native state, as shown in Figure 3.8A, the major contributor to the binding energy switches from being non-native contacts to native contacts, as expected. This change of contribution is steep around $Q = 0.5$, near the transition state region. At low Q region, total energy of swapped contacts is about 20% of the total binding energy. Consistent with the principle of minimal frustration, swapped contacts are on average stronger than other non-native contacts throughout the whole binding process, as shown in Figure 3.8B. At low Q, where the two monomers are first coming into contact, the average strength of the swapped contacts is even larger than the strength of the native contacts alone, suggesting their important role in stabilizing non-specific bound structures at the start of the binding process. As binding progresses, the strengths of both swapped and non-native contacts decrease, while the strength of the native contacts is, interestingly, more or less constant. These observations of the ubiquity of domain swapping are consistent with the experimental observation by Oliveberg of the universality of transient aggregation at high protein concentration [126].

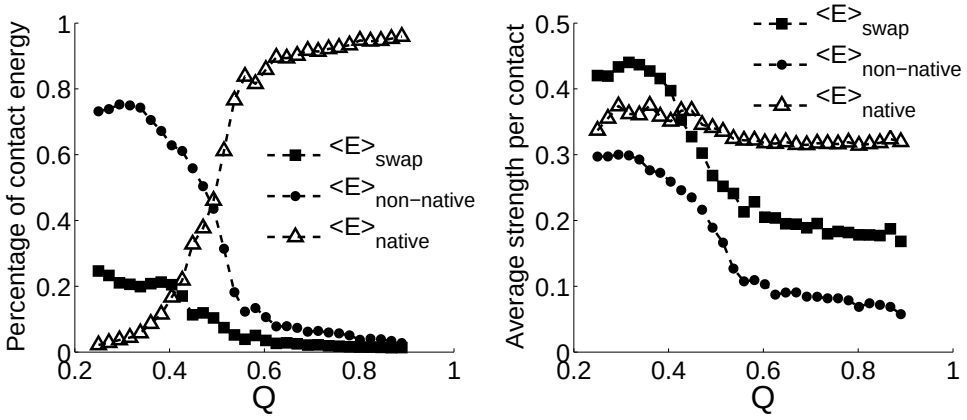


Figure 3.8: For all three different types of contacts, their contributions to the total binding energy and their average strength are plotted against Q of the Lambda repressor complex. Native contacts, swapped contacts and non-native contacts excluding swapped contacts are in triangle, square and circle symbols, respectively. A) As the complex approaches the native state, the major contributor to the binding energy switches from non-native contacts to native contacts. This change of contribution is steep around $Q = 0.5$, near the transition state region. Total energy of swapped contacts is about 20% of the total binding energy at the low Q region. B) Swapped contacts are on average stronger than other non-native contacts throughout the whole Q region. As Q increases, the average strengths of both swapped and non-native contacts decrease, while the strength of the native contacts is more or less constant. At low Q region where the two monomers are in initial encountering, the average strength of the swapped contacts is even larger than the native contacts.

3.6 Conclusions

The intent of this study was to investigate the extent to which protein-protein association is funneled by the same forces that determine the landscapes of monomeric proteins. We see that the association is well described by a funneled model but that there are residual effects of energetic frustration which allow non-native interactions to play a role. The picture which emerges from the study is that folding

and binding are dynamic processes that are often coupled and that both take place via diffusion on rugged but nevertheless largely funneled energy landscapes. Interactions that successfully predict the structure of monomeric proteins also prove sufficient to predict native dimeric interfaces. Monomer geometry alone does not lead to the successful prediction of binding modes: both energetic heterogeneity and flexibility of the monomers are important. Non-native interactions can stabilize on-pathway or off-pathway conformations depending on the stability of the monomers, and swapped contacts in particular are stronger than other, non-specific, non-native contacts, in accordance with the principle of minimal frustration. Swapped contacts play an important role in stabilizing non-specific bound structures at the start of the association process. Other non-native interactions, on the other hand, also sometimes play a role but in general dimeric proteins have evolved so as to eliminate traps on the combined folding and binding landscape.

3.7 Methods

AWSEM was described in detail in the supplementary material of a recent paper [62]. The tertiary contact potential $V_{contact}$ consists of two terms, the direct contact V_{direct} and the mediated contact V_{water}. In simulations with multiple chains, the associative memory potential V_{AM}, acts only locally in sequence within each monomer. On the other hand, $V_{contact}$, the burial potential and the beta hydrogen bonding terms act both within and among the monomeric chains. When calculating the local density of residues, which is used by the helix and burial potentials as well

as $V_{contact}$, all chains are included. V_{AM} is determined by a single memory, which is the structure of the monomer in the experimentally determined dimer structure. As mentioned previously, this interaction includes only those pairs of residues that have a sequence separation of less than or equal to 9. No information about contacts within the monomers or on the dimer interface is included.

The predictions were performed using molecular dynamics with simulated annealing. The annealing simulations were initialized by completely unfolding the individual monomers and separating them. The temperature is then lowered to below the empirically determined binding temperature and a weak bias is applied between the centers of mass of the two monomers to ensure that contact is made during the course of the simulation. The free energy surfaces were calculated using the Weighted Histogram Analysis Method (WHAM) [127] on the data collected from constant temperature simulations with umbrella sampling along Q.

Chapter 4: The Role of Micro-environmental Fluctuations in Protein Folding

4.1 Introduction

Protein folding experiments *in vitro* have taught us a great deal about various folding mechanisms. In particular, the early puzzle about protein folding timescales being unreasonably short, based on a naive expectation of very long search times on a random landscape, have been resolved using the modern language of energy landscapes, invoking, among other physical ideas, the key idea of the principle of minimal frustration [5, 128]. The latter postulates that the energetic correlations among various conformations of a globular protein sculpt a funnel with the native basin residing deep, at the bottom, which allows many unfolded conformations to glide down during folding. Additionally, residual frustration renders transient traps along the folding pathway. However, the ruggedness of energy landscapes of evolved protein sequences should not be too large, such that folding dynamics is not arrested in these traps. Interestingly, one source of frustration is very difficult to do away using sequence change via evolution, namely the topological frustration of chain segments not being able to cross each other. For proteins with large contact

order, which often have complicated folds and function, topological misfolding may represent a major source of slowdown towards the native state.

Although major advances have been made in the last three decades in understanding protein folding *in vitro*, only recently folding *in vivo* has attracted considerable attention. In particular, there are two effects that have been throughly investigated: 1) the role of excluded volume effects, induced by crowding inside a cell, which may affect not only folding rates but also result in misfolding and aggregation [129]; 2) the role of chaperons that help to prevent misfolding and aggregation through various means. Even proteins which are capable of autonomously folding into their native structure *in vitro* very often need chaperon assistance *in vivo* [130–132]. A number of mechanisms were suggested for the physical basis of chaperon action, including iterative annealing [133] and confined space mechanisms [134, 135].

It has recently been discussed that in a dense environment folding may be affected not only through excluded volume interactions, but also through formation of transient, yet chemically specific, contacts with the crowding agent. In the context of a biological cell, this becomes even more complicated, since there are thousands of crowding agents present, and the composition of molecules evolves in time and space. Hence, from the viewpoint of any specific protein which undergoes folding, the chemical environment around it is fluctuating on either one or, more likely, multiple timescales. In this work, we address the question of whether transient fluctuations of chemical micro-environment of a protein may significantly affect its folding and unfolding dynamics.

A large extent of interactions between protein residues are solvent mediated, such as the hydrophobic effect or electrostatic and water-mediated interactions between hydrophilic residues. Hence, when the chemical composition of the solvent changes, one expects that the strength of attraction between the protein residues will vary as well. Since the chemical composition varies spatially within a biological cell, and the composition in one spatial location fluctuates temporally, the strength of inter-residue interactions of any particular protein in the cell will fluctuate in sync. In analogy to various barrier crossing problems studied in other fields of condensed matter physics, one expects that these fluctuations under certain conditions may induce a stochastic resonance, accelerating both folding and unfolding kinetics. Indeed, this is what we have discovered in our numerical simulations of protein folding using a simple coarse-grained model. We found that both deterministic and random fluctuations of the strength of inter-residue interaction potential, when occurring at some specific timescale, significantly accelerate the folding dynamics. Hence, we suggest a new mechanism which may accelerate protein folding in the dense cellular environment, assisted by transient fluctuations of the chemical micro-environment for any protein that undergoes folding and unfolding. This mechanism may be helpful for larger proteins that are difficult to fold, particularly due to having significant topological frustration, providing an assistance pathway complementing the action of chaperons.

4.2 Results and Discussion

In this work, we investigated the effect of correlated random noise and harmonic fluctuations on protein folding dynamics. Using Molecular Dynamics (MD) simulations we showed that both harmonic and random noise increase the transition rate if applied at a certain characteristic frequency or correlation time, respectively.

In our computational studies, we looked at two structures; a 53 residue 1SRL and a 415 residue PGK (see Figure 4.3). We modeled the environmental noise by inducing fluctuations of native binding interactions (see Methods section). The results summarized in the Figure 4.1 illustrate the dependence of average first-passage time of folding from the harmonic period θ and correlation parameter τ for different noise amplitudes (corresponding to the root-mean-square deviation values $\frac{\sqrt{\langle \delta\epsilon^2 \rangle}}{\epsilon}$). The horizontal solid line in each of the graphs corresponds to the average folding time in the absence of any noise.

Figure 4.1 clearly shows that, for both proteins, and for both harmonic fluctuations and correlated random noise, there exists a range of values of the characteristic parameter (τ or θ) for which the average first-passage times are smaller compared to the unperturbed system. Furthermore, for each value of the harmonic or correlated random noise amplitude, there exists an optimal τ or θ value for which the effect is maximal. For the harmonic noise the optimal fluctuation frequency ($\nu = 1/\theta$) is in close order to the transition rate of the unperturbed system. This is in agreement with earlier theoretical calculations for resonance of nonlinear systems of differential equations, which applies to a wide class of problems including equations of

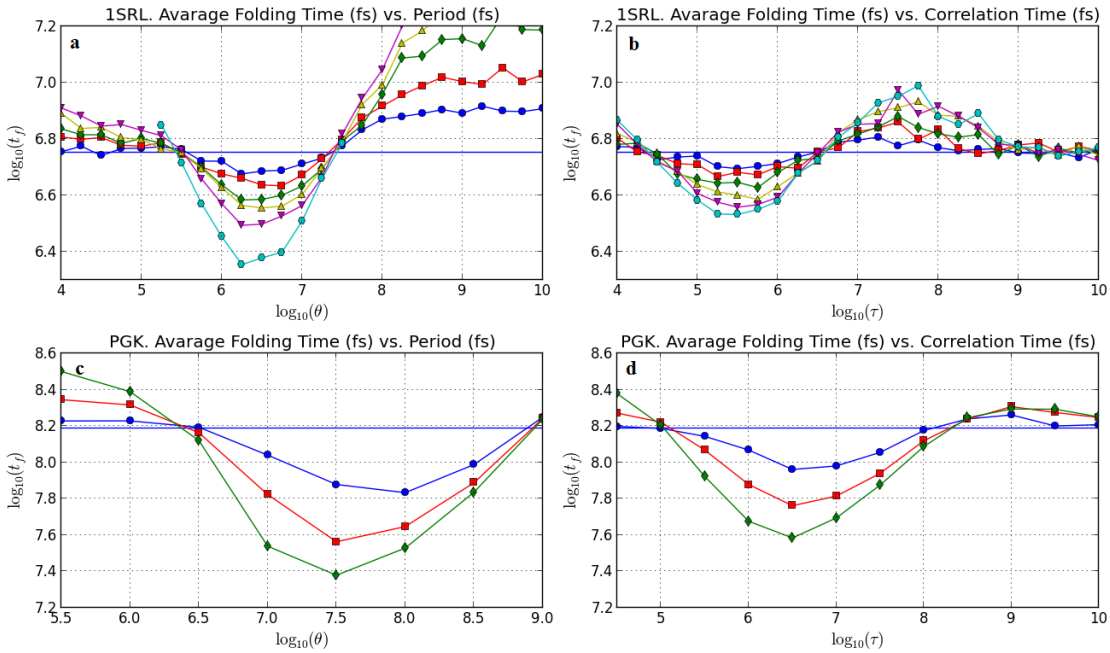


Figure 4.1: a) 1SRL. Harmonic fluctuations. Plot of average first-passage time vs θ period for root-mean-square deviations $\frac{\sqrt{\langle \delta\epsilon^2 \rangle}}{\epsilon}$=0.03, 0.04, 0.05, 0.06, 0.07, 0.1; blue circles, red squares, green diamonds, yellow triangles, magenta down triangles, cyan hexagons correspondingly. b) 1SRL. Correlated random noise. Plot of average first-passage time vs τ correlation time for root-mean-square deviations $\frac{\sqrt{\langle \delta\epsilon^2 \rangle}}{\epsilon}$=0.04, 0.05, 0.06, 0.08, 0.09, 0.1; blue circles, red squares, green diamonds, yellow triangles, magenta down triangles, cyan hexagons correspondingly. c) PGK. Harmonic fluctuations. Plot of average first-passage time vs θ period for root-mean-square deviations $\frac{\sqrt{\langle \delta\epsilon^2 \rangle}}{\epsilon}$=0.03, 0.05, 0.07; blue circles, red squares, green diamonds correspondingly. d) PGK. Correlated random noise. Plot of average first-passage time vs τ correlation time for root-mean-square deviations $\frac{\sqrt{\langle \delta\epsilon^2 \rangle}}{\epsilon}$=0.03, 0.05, 0.07; blue circles, red squares, green diamonds correspondingly. The horizontal line in each plot corresponds to the average folding time in the absence of noise. Folding time, τ and θ are in units of femtoseconds.

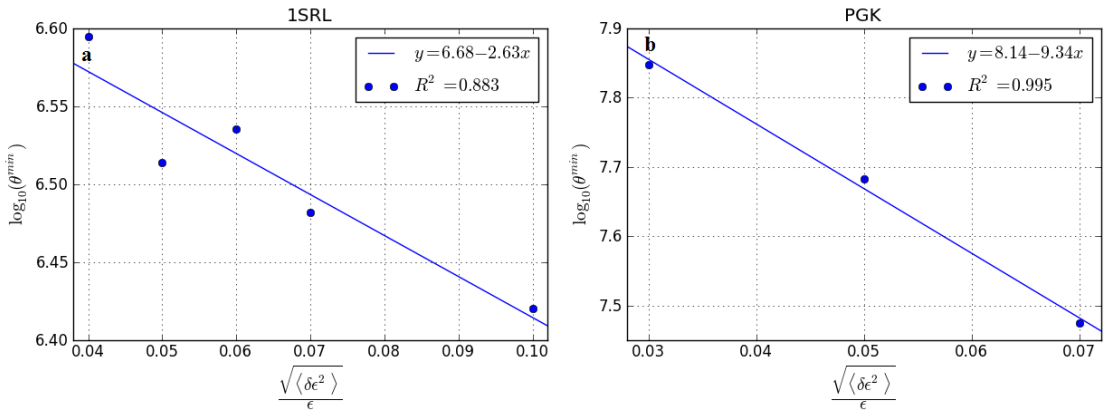


Figure 4.2: a) 1SRL. Plot of θ^{min} vs. noise amplitude. $R^2 = 0.883$. b) PGK. Plot of θ^{min} vs. noise amplitude. $R^2 = 0.995$.

motion [136, 137]. We also showed that, for the harmonic fluctuations, the optimal θ value decreases with increasing noise amplitude (see Figure 4.2). On the contrary, in the case of correlated random noise, the optimal τ value does not seem to depend on noise amplitude.

In other words, our computational results show the existence of a resonance when the driving frequency approaches the folding rate of the unperturbed system. For PGK, this finding was recently validated experimentally using time-resolved Förster resonance energy transfer (FRET) technique. In the experiments done by Martin Gruebele and Max Platkov at University of Illinois at Urbana-Champaign, the temperature of the aqueous solution of PGK was harmonically modulated around average temperature $T_0 = 37.0\,^{\circ}\mathrm{C}$ with amplitudes $\delta T = \pm 1 \pm 2\,^{\circ}\mathrm{C}$ [138, 139].

$$T(t) = T_0 + \delta T \sin[2\pi\nu t] \tag{4.1}$$

The melting temperature and the folding/unfolding relaxation rate of the PGK

construct used in the experiments are 39.0 °C and 0.5 Hz, respectively [140]. The applied temperature wave will induce folding and unfolding, affecting the donor and acceptor intensities. At low frequencies, they will follow the temperature modulation, but they will lag behind when the frequency of modulation approaches the reaction rate [141, 142]. In principle, measuring the phase difference between donor and acceptor intensities ($\phi = \phi_A - \phi_D$) should make it possible to detect the change in reaction rate. However, the small temperature signal used in the experiments only make it possible to examine the $\Delta\phi$ difference between two amplitudes.

According to our estimates (see Section C.2), ± 1 and ± 2 °C correspond to $\frac{\sqrt{\langle \delta\epsilon^2 \rangle}}{\epsilon} = \mp 0.0033$ and ∓ 0.0066, respectively. This is approximately 10 times weaker then the signal we applied in our simulations. (± 2 °C is an experimental limit). Using interpolation technique (see Section C.3) we estimated

$$\ln \left[\frac{t_f^{min}\left(\sqrt{\langle \delta\epsilon^2 \rangle}/\epsilon = 0.0066\right)}{t_f^{min}\left(\sqrt{\langle \delta\epsilon^2 \rangle}/\epsilon = 0.0033\right)} \right] = -0.085 \tag{4.2}$$

This is in agreement with the estimate made from the experimentally measured $\Delta\phi$ vs. frequency dependence.

It is important to note that the experimentally observed folding times and the times obtained from simulation differ in order significantly. This is in result of using a Gō-like model (see Methods) for the computational studies. Such models allow us to explore the folding dynamics of rather big proteins, PGK for example, through the use of reduced representation and by smoothing the energy landscape [42]. This results in a significant reduction of physical folding time for a protein.

Our computational studies (see Figure 4.1) also show that, for both harmonic

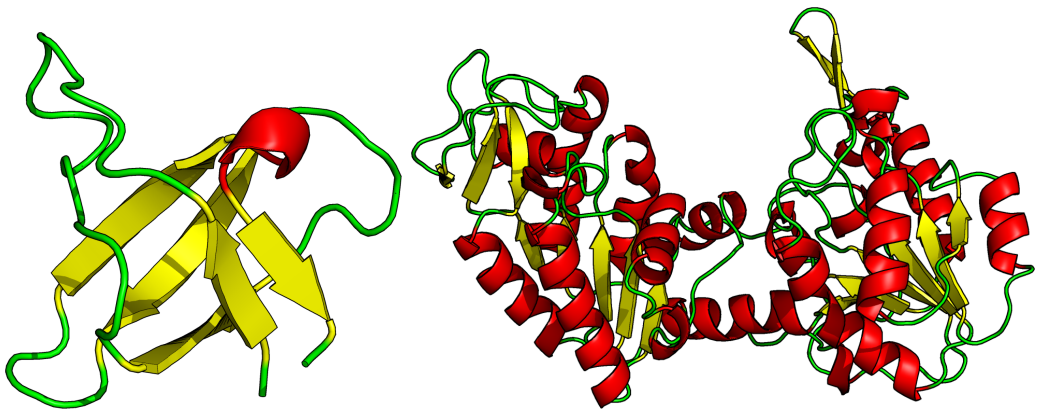


Figure 4.3: Left: SH3 domain of tyrosine kinases (PDB ID: 1SRL). 56 residues. Right: Phosphoglycerate kinase (PGK). 415 residues.

and correlated random noise, the average first-passage time increases relative to the unperturbed system before and after the system experience resonance. This is due to the exponential nature of the partition function; fluctuations of an equal strength do not cancel each other if they have opposite signs. In the case of harmonic noise, the average first-passage time first increases rapidly when the period becomes greater than the folding time of the unperturbed system. It further levels out to a certain constant value. In the latest case, when the period becomes much larger than the folding time, the applied wave acts like a constant driving force instead. Because the phase of the harmonic wave is randomly chosen at the beginning of the simulation, and because of the non-linearity of the system, this results in the average first-passage time being above its value in the absence of noise. In the case of correlated random noise, we always started from the desired average value of the ϵ. Thus, for large correlation times, the average first-passage time coincides with its value for the unperturbed system.

A further theoretical discussion of the phenomenon described here, in the context of the experimental measurements, is presented in Chapter 5.

4.3 Conclusions

Using a simple computational model, we showed the existence of resonance in the protein folding reaction. Our simulations directly indicate that both harmonic fluctuations and correlated random noise increase the transition rate for certain values of the characteristic parameters. In the case of harmonic noise, we found that resonance occurs when the fluctuation frequency approaches the folding rate of the unperturbed system. This former result was validated experimentally, where resonance was also observed near the natural folding rate. Another interesting feature we saw from the simulations was the increase of resonance frequency with the amplitude of the harmonic wave, while, in the case of the correlated random noise, the resonance correlation time did not depend on the amplitude.

4.4 Methods

4.4.1 Gō-like Model

For this study we used a C_α only Gō-like model developed by Onuchic and coworkers [42]. According to this model, the energy of a specific conformation of a

protein is,

$$
\begin{aligned}
V_{Go-model} &= \sum_{bonds} K_r(r-r_0)^2 + \sum_{angles} K_\theta(\theta-\theta_0)^2 \\
&+ \sum_{dihedrals} \sum_{n=1,3} K_\phi^{(n)}(1+cos[n(\phi-\phi_0)]) \\
&+ \sum_{i<j-3} \epsilon_{i,j} \left[5\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - 6\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{10} \right] + \sum \epsilon'_{i,j} \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12}
\end{aligned}
\tag{4.3}
$$

where r and r_0 are the distances between two subsequent C_α atoms in current and native states. Similarly θ and θ_0 are angles between three subsequent C_α atoms and ϕ and ϕ_0 are dihedral angles between four subsequent C_α atoms. The last two terms describe non-local native interactions and short range repulsion between non-native pairs. $\epsilon_{i,j} = \epsilon > 0$ and $\epsilon'_{i,j} = 0$ if i and j are in contact in native state, and $\epsilon_{i,j} = 0$ and $\epsilon'_{i,j} = \epsilon_2 > 0$ in the opposite case. σ_{ij} is taken to be equal to the distance between C_{α_i} and C_{α_j} for native pairs and $4\mathring{A}$ for non-native ones. K parameters are taken to be $K_r = 100\epsilon$, $K_\theta = 20\epsilon$, $K_\phi = \epsilon$ and $K_\phi = 0.5\epsilon$. We choose ϵ and ϵ_2, equal to the same ϵ_0, so that the folding temperature[1] falls between 300 and 350K.

4.4.2 Simulations

We performed molecular dynamics simulations of the SH3 domain of tyrosine kinases (PDB ID: 1SRL) and phosphoglycerate kinase (PGK) protein. We used Gō-like model described above, but with modified ϵ for the native contact term. This way, we aim to model the environmental noise present in realistic biological environment. We modeled ϵ in two ways: 1) as a random process with some characteristic correlation time τ, using the Langevin equation, 2) as a sinusoidal wave

[1] At the folding temperature, the populations of the folded state and unfolded states are equal.

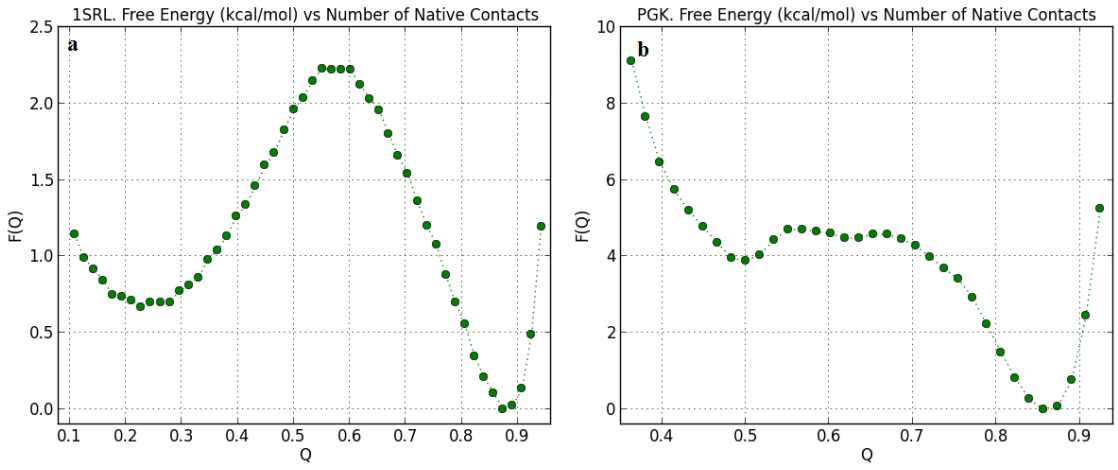


Figure 4.4: a) Free energy of 1SRL vs. the ratio of native contact Q. b) Free energy of PGK vs. the ratio of native contact Q.

with period θ. In both cases, the average value of ϵ was equal to the original value of ϵ_0. We computed the average first-passage time of folding for different values of standard deviation of ϵ, and the characteristic correlation time τ in the 1st case and period θ in the 2nd case.

In the case of correlated random noise, $\epsilon_{eff} = \epsilon + \delta\epsilon$, where $\delta\epsilon$ is a solution of the equation

$$\frac{d\delta\epsilon}{dt} = -\xi\delta\epsilon + \omega^2\eta(t) \tag{4.4}$$

where $\eta(t)$ is Gaussian white noise. ξ and ω determine properties of the random process, namely τ correlation time and $\sqrt{\langle\delta\epsilon^2\rangle}$ standard deviation (see Section C.1 for more details). In the case of sinusoidal wave, $\epsilon_{eff} = \epsilon + \frac{\delta\epsilon}{\sqrt{0.5}}\sin(\frac{2\pi t}{\theta} + \phi_0)$ where $\delta\epsilon$ now is a numerical parameter. ϕ_0 is the initial phase of the wave, which was uniformly randomized for all simulations with harmonic noise.

For the 1SRL protein, we chose $\epsilon_0 = 0.56 kcal/mol$. This resulted in the free energy barrier of approximately $1.5 kcal/mol$ or $2.5 k_B T$. For PGK, we chose a slightly higher $\epsilon_0 = 0.58 kcal/mol$. In this case, the free energy barrier was approximately $0.9 kcal/mol$ or $1.5 k_B T$ (see Figure 4.4).

For each data point we ran 1000 MD simulations and found the dependence of the average first-passage time of folding from noise amplitude and the characteristic parameters. We considered a protein folded if the fraction of native contacts (Q) formed was above 0.8.

Chapter 5: Further Discussion of Fluctuation-induced Resonance Phenomenon in Protein Folding

5.1 Introduction

Here we further discuss the phenomenon described in Chapter 4. First, we present a simple kinetic theory of folding under applied temperature wave. We find the number fraction of folded (and unfolded, respectively) population vs. time, and its phase shift relative to the driving wave. As expected, at slow modulation we see that the reaction closely follows the temperature fluctuations and lags behind them at higher frequencies. Knowing the folded population allows us to estimate the phase shift between donor and acceptor fluorescence intensities in FRET experiments. Even though the kinetic theory does not explain the emergence of the resonance in the folding reaction, it helps to understand the general effect of the temperature fluctuations on experimental measurements. Next, we approach the same problem using Brownian Dynamics (BD) simulations of a particle moving on a two-state free energy landscape. Varying the frequency and the amplitude of the applied temperature wave, we get very similar results to the ones obtained with our MD simulations (see Chapter 4), including the decrease of average first-passage time of

folding and resonance near the frequency value corresponding to the transition rate of the unperturbed system. Then again, we calculated the phase shifts between the folded population and the temperature wave, and between donor and accepter intensities. The comparison of BD results to the ones from the kinetic theory allows us to understand the effect of resonance on those phase shifts and to interpret the experimental results.

5.2 Kinetic Theory. Analytical Solutions

5.2.1 Finding folded and unfolded fractions

Let's assume we have a homogenous protein solution, and that the protein itself is a two-state system (see Figure 5.1) with folding temperature[1] $T_m = 312K$. We will denote the number fractions of folded and unfolded proteins by F and U. Then, we can describe the evolution of F with the following kinetic equation:

$$\frac{dF}{dt} = k_+ U - k_- F \tag{5.1}$$

where k_+ and k_- are the folding and unfolding rates respectively. They can be written as

$$k_+ = k_0 e^{-\frac{\Delta G_{UF}}{k_B T}}, k_- = k_0 e^{-\frac{\Delta G_{FU}}{k_B T}} \tag{5.2}$$

We are interested in the case when there is a harmonic temperature wave applied to the system:

$$T(t) = T_0 + \delta T \sin(2\pi\nu t) \tag{5.3}$$

[1] At the folding temperature, the populations of the folded state and unfolded states are equal.

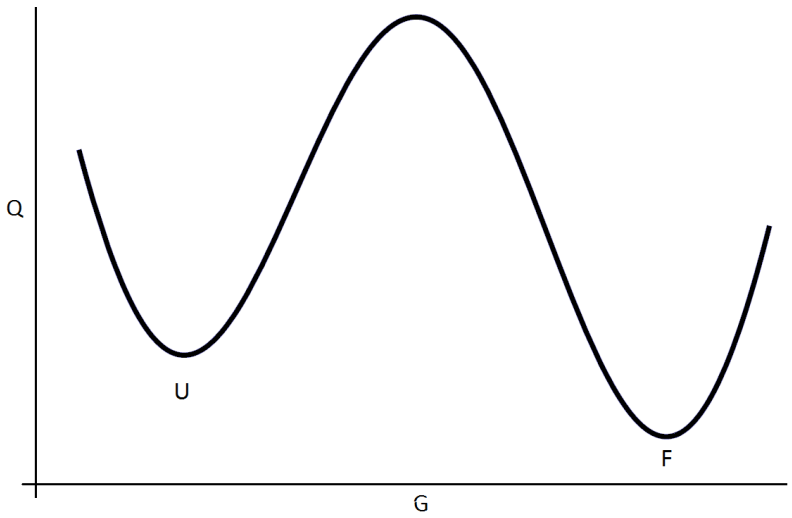


Figure 5.1: Free energy of a two-state protein.

where $T_0 = 310K$ and δT is either ± 1 or $\pm 2K$. As a result of the applied temperature wave, ΔG_{UF} and ΔG_{FU}, and thus k_+ and k_-, will depend on time. In the linear approximation

$$\begin{aligned} \Delta G_f &= \Delta G_{UF} - \Delta G_{FU} = \Delta G_f^{(1)}(T(t) - T_m) \\ \Delta G_{UF} &= \Delta G^{(0)} + \Delta G_{UF}^{(1)}(T(t) - T_m) \\ \Delta G_{FU} &= \Delta G^{(0)} + \Delta G_{FU}^{(1)}(T(t) - T_m) \end{aligned} \tag{5.4}$$

The exact analytical solution of Equation 5.1 cannot be found. Thus, we will first discuss the analytical solution of the equation, which we found through a number of approximations. As we will show next, this analytical solution is a good first order approximation, which explains the phase shift vs. frequency dependence well but fails to explain the dependence of the phase shift on the temperature wave amplitude.

Using Equation 5.4 and taking into account that U + F = 1, we can rewrite Equa-

tion 5.1 in the following way:

$$\frac{dF}{dt} = k_0^{'}\left[(1-F)e^{-\frac{\Delta G_{UF}^{(1)}(T(t)-T_m)}{k_B T(t)}} - Fe^{-\frac{\Delta G_{FU}^{(1)}(T(t)-T_m)}{k_B T(t)}}\right] \tag{5.5}$$

where $k_0^{'} = k_0 e^{-\frac{\Delta G^{(0)}}{k_B T}}$ is the slow changing prefactor, which we will assume to be constant. By taking into account that $\frac{\delta T}{T_0} \ll 1$, the expressions in the exponents can be simplified in the following manner

$$\frac{\Delta G_{UF}^{(1)}(T(t)-T_m)}{k_B T(t)} = \frac{\Delta G_{UF}^{(1)}}{k_B}\left[1-\frac{T_m}{T_0+\delta T\sin(2\pi\nu t)}\right] \approx \frac{\Delta G_{UF}^{(1)}}{k_B}\left[1-\frac{T_m}{T_0}\right] + \frac{\Delta G_{UF}^{(1)}}{k_B}\frac{T_m\delta T}{T_0^2}\sin(2\pi\nu t).$$

Accordingly, we will get

$$\frac{dF}{dt} = k_0^{'}\left[(1-F)\mu_1 e^{-\frac{\Delta G_{UF}^{(1)}}{k_B}\frac{T_m\delta T}{T_0^2}\sin(2\pi\nu t)} - F\mu_2 e^{-\frac{\Delta G_{FU}^{(1)}}{k_B}\frac{T_m\delta T}{T_0^2}\sin(2\pi\nu t)}\right] \tag{5.6}$$

where μ_1 and μ_2 are time independent constants given by the equations

$$\mu_1 = e^{-\frac{\Delta G_{UF}^{(1)}}{k_B}\left[1-\frac{T_m}{T_0}\right]},\ \mu_2 = e^{-\frac{\Delta G_{FU}^{(1)}}{k_B}\left[1-\frac{T_m}{T_0}\right]}. \tag{5.7}$$

After Taylor expanding Equation 5.6 (assuming $\frac{\Delta G^{(1)}}{k_B}\frac{T_m\delta T}{T_0^2} \ll 1$) we will get the equation of the following form

$$\frac{dF}{dt} = a_0 + a_1\sin(2\pi\nu t) + a_2 F + a_3 F\sin(2\pi\nu t) \tag{5.8}$$

where

$$\begin{aligned} a0 &= k_0^{'}\mu_1, \\ a1 &= -k_0^{'}\mu_1\Delta G_{UF}^{(1)}\frac{T_m\delta T}{k_B T_0^2}, \\ a2 &= -k_0^{'}(\mu_1+\mu_2), \\ a3 &= k_0^{'}\left(\mu_1 G_{UF}^{(1)} + \mu_2 G_{FU}^{(1)}\right)\frac{T_m\delta T}{k_B T_0^2}. \end{aligned} \tag{5.9}$$

The solution of the Equation 5.8 can be written as

$$F(t) = C_0 e^{a_2 t - \frac{a_3\cos(2\pi\nu t)}{2\pi\nu}} + e^{a_2 t - \frac{a_3\cos(2\pi\nu t)}{2\pi\nu}}\int_0^t e^{-a_2\tau + \frac{a_3\cos(2\pi\nu\tau)}{2\pi\nu}}[a_0 + a_1\sin(2\pi\nu\tau)]\mathrm{d}\tau \tag{5.10}$$

The first term of this expression has a multiplier $e^{a_2 t}$, where $a_2 < 0$ (see Equation 5.9), and will vanish with $t \to \infty$. We are only interested in the steady solution, thus we can throw the first term away. By doing Taylor expansion under the integral in respect of $\frac{a_3 \cos(2\pi\nu\tau)}{2\pi\nu} \ll 1$ we will get

$$F(t) = e^{a_2 t - \frac{a_3 \cos(2\pi\nu t)}{2\pi\nu}} \int_1^t e^{-a_2\tau} \left(1 + \frac{a_3 \cos(2\pi\nu\tau)}{2\pi\nu}\right) [a_0 + a_1 \sin(2\pi\nu\tau)] \mathrm{d}\tau. \quad (5.11)$$

After integrating, we will find the following steady expression for $F(t)$,

$$F(t) = e^{-\frac{a_3 \cos(2\pi\nu t)}{2\pi\nu}} \left[-\frac{a_0}{a_2} + \frac{a_0 a_3 - a_1 a_2}{a_2^2 + 4\pi^2\nu^2} \sin(2\pi\nu t) - \frac{a_0 a_2 a_3 + 4\pi^2\nu^2 a_1}{2\pi\nu(a_2^2 + 4\pi^2\nu^2)} \cos(2\pi\nu t) \right.$$
$$\left. - \frac{a_1 a_2 a_3}{4\pi\nu(a_2^2 + 16\pi^2\nu^2)} \sin(4\pi\nu t) - \frac{a_1 a_3}{a_2^2 + 16\pi^2\nu^2} \cos(4\pi\nu t) \right]. \quad (5.12)$$

Expanding further we will finally get

$$F(t) = \left(1 - \frac{a_3 \cos(2\pi\nu t)}{2\pi\nu}\right) \left[-\frac{a_0}{a_2} + \frac{a_0 a_3 - a_1 a_2}{a_2^2 + 4\pi^2\nu^2} \sin(2\pi\nu t) - \frac{a_0 a_2 a_3 + 4\pi^2\nu^2 a_1}{2\pi\nu(a_2^2 + 4\pi^2\nu^2)} \cos(2\pi\nu t) \right.$$
$$\left. - \frac{a_1 a_2 a_3}{4\pi\nu(a_2^2 + 16\pi^2\nu^2)} \sin(4\pi\nu t) - \frac{a_1 a_3}{a_2^2 + 16\pi^2\nu^2} \cos(4\pi\nu t) \right]. \quad (5.13)$$

After opening the brackets and throwing away the terms with smaller multipliers, we will find

$$F(t) = F_{eq} + A \left[\sin(2\pi\nu t) + B \cos(2\pi\nu t)\right], \quad (5.14)$$

where

$$F_{eq} = -\frac{a_0}{a_2} = \frac{k_+^0}{k_{obs}^0} = \frac{1}{1 + e^{\frac{\Delta G_f^{(1)}(T_0 - T_m)}{k_B T_0}}}, \quad (5.15)$$

$$A = \frac{a_0 a_3 - a_1 a_2}{a_2^2 + 4\pi^2\nu^2} = -\frac{k_+^0 k_-^0}{(k_{obs}^0)^2 + 4\pi^2\nu^2} \frac{\Delta G_f^{(1)} T_m \delta T}{k_B T_0^2}, \quad (5.16)$$

$$B = \frac{2\pi\nu}{a_2} = -\frac{2\pi\nu}{k_{obs}^0}. \quad (5.17)$$

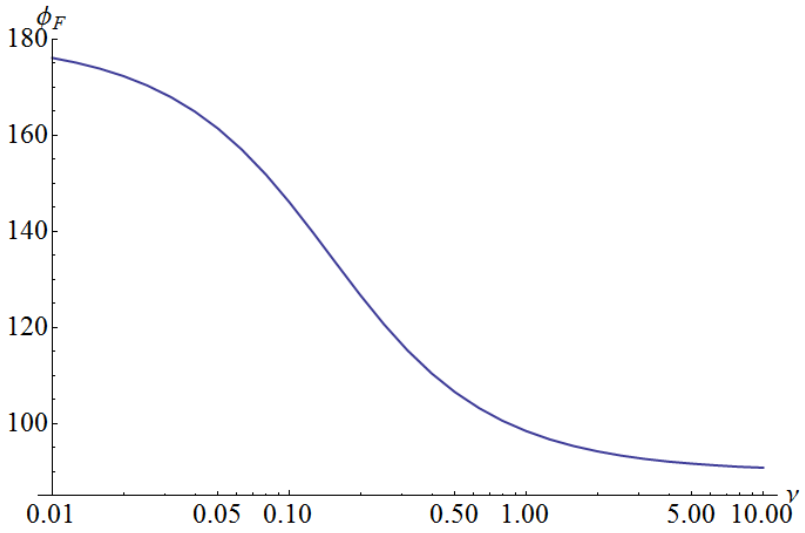


Figure 5.2: Kinetic theory, analytical solution. Phase shift of the folded population relative to the driving wave vs. frequency.

Here $k_+^0 = k_0^{'}\ \mu_1$, $k_-^0 = k_0^{'}\ \mu_2$ and $k_{obs}^0 = k_0^{'}(\mu_1 + \mu_2)$ are the folding, unfolding and observed reaction rates at the temperature T_0, correspondingly. Equation 5.14 can be rewritten as

$$F(t) = F_{eq} + A^{'} \sin(2\pi\nu t + \phi_F) \tag{5.18}$$

where

$$A^{'} = -A\sqrt{1+B^2} = \frac{k_+^0 k_-^0}{k_{obs}^0\sqrt{(k_{obs}^0)^2 + 4\pi^2\nu^2}} \frac{\Delta G_f^{(1)} T_m \delta T}{k_B T_0^2} \tag{5.19}$$

$$\phi_F = \pi - \arctan\left[2\pi\frac{\nu}{k_{obs}^0}\right]. \tag{5.20}$$

In Equation 5.18, ϕ_F is the phase shift of the folded population relative to the driving temperature wave. As expected, it depends on the $\frac{\nu}{k_{obs}^0}$ ratio. Equation 5.20 shows, that at zero frequency of the driving wave, the phase shift is equal to 180 degrees (folded population decreases with increasing temperature). As the frequency

increases, the phase shift asymptotically decreases to 90 degrees. This is illustrated in Figure 5.2. It is also important to note, that according to Equation 5.20, ϕ_F does not depend on δT.

5.2.2 FRET Donor and Acceptor Intensities

In the FRET experiment, the donor and acceptor fluorescence intensities can be written down as

$$\begin{aligned} D(T) &= \gamma_D I_{exc} D_0(T)(1 - E(T)) \\ A(T) &= \gamma_A I_{exc} D_0(T) A_0(T) E(T) \end{aligned} \tag{5.21}$$

where γ_D and γ_D are constants that depend on collection efficiency and fluorescence filters, I_{exc} is the excitation laser intensity, D_0 and A_0 are the donor and acceptor quantum yields, and E is the FRET efficiency [142]. In the range of interest, the dependence of D_0 and A_0 from temperature is strictly linear. The FRET efficiency $E = E[\rho(R, T)]$ is given by the following formula

$$E[\rho] = \int_0^\infty \mathrm{d}R \frac{1}{1 + \left(\frac{R}{R_0}\right)^6} \rho(R) \tag{5.22}$$

where R is the donor-acceptor distance, R_0 is the Förster distance, and $\rho(R, T)$ is the protein population at the given temperature. For a two-state protein we can write $E(T)$ as

$$E(T) = E_F F(T) + E_U U(T) = (E_F - E_U) F(T) + E_U. \tag{5.23}$$

If we represent $D_0(T)$ and $A_0(T)$ as

$$\begin{aligned} D_0(T) &= p_1 + p_2 T \\ A_0(T) &= q_1 + q_2 T \end{aligned} \tag{5.24}$$

and plug in the expression 5.14 for $F(t)$, we will get the following expressions for $D(T)$ and $A(T)$:

$$\begin{aligned} D(T) &= \gamma_D I_{exc}[D_{eq} + D_1 \sin(2\pi\nu t) + D_2 \cos(2\pi\nu t)] \\ A(T) &= \gamma_A I_{exc}[A_{eq} + A_1 \sin(2\pi\nu t) + A_2 \cos(2\pi\nu t)] \end{aligned} \tag{5.25}$$

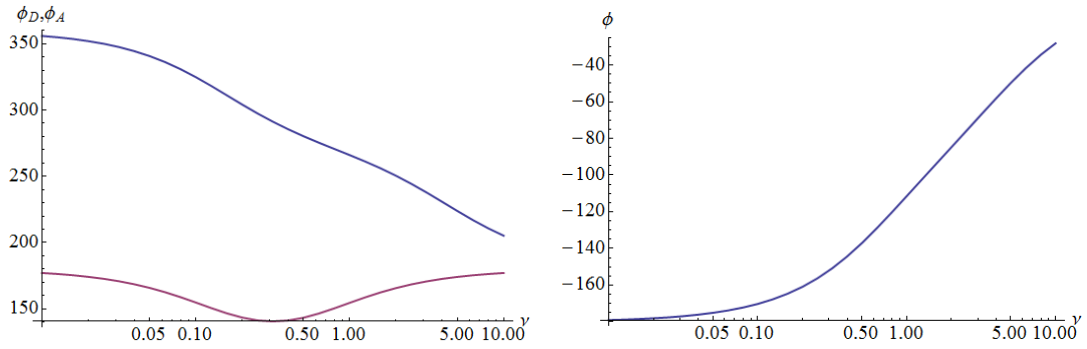
where

$$\begin{aligned} D_{eq} &= (1 - E_U - (E_F - E_U)F_{eq})(p_1 + p_2 T_0) \\ D_1 &= \delta T(1 - E_U - (E_F - E_U)F_{eq})p_2 - A(E_F - E_U)(p_1 + p_2 T_0) \\ D_2 &= -AB(E_F + E_U)(p_1 + p_2 T_0) \\ A_{eq} &= (E_U + (E_F - E_U)F_{eq})(p_1 + p_2 T_0)(q_1 + q_2 T_0) \\ A_1 &= \delta T(E_U + (E_F - E_U)F_{eq})((p_1 + p_2 T_0)q_2 + (q_1 + q_2 T_0)p_2) \\ &\quad -A(E_F - E_U)(p_1 + p_2 T_0)(q_1 + q_2 T_0) \\ A_2 &= AB(E_F + E_U)(p_1 + p_2 T_0)(q_1 + q_2 T_0). \end{aligned} \tag{5.26}$$

Here we threw away all square terms (by p_2, q_2 and A) using the fact that p_2 and q_2 are orders of magnitude smaller then p_1 and p_2. The phase shifts of D and A will obey the following expressions.

$$\begin{aligned} \phi_D &= -arctan\left[\frac{A^* B(E_F - E_U)(p_1 + p_2 T_0)}{(1 - E_U - (E_F - E_U)F_{eq})p_2 - A^*(E_F - E_U)(p_1 + p_2 T_0)}\right] \\ \phi_A &= arctan\left[\frac{A^* B(E_F + E_U)(p_1 + p_2 T_0)(q_1 + q_2 T_0)}{(E_U + (E_F - E_U)F_{eq})((p_1 + p_2 T_0)q_2 + (q_1 + q_2 T_0)p_2) + A^*(E_F - E_U)(p_1 + p_2 T_0)(q_1 + q_2 T_0)}\right]. \end{aligned} \tag{5.27}$$

Here $A^* = A/\delta T$ and is independent of δT. Thus ϕ_D and ϕ_A are also independent of δT. In reality ϕ_F, ϕ_D and ϕ_A weakly depend on δT. For ϕ_F, for example, between



(a) (b)

Figure 5.3: Kinetic theory, analytical solution. a) Donor (blue) and accepter (red) phase shifts vs. frequency. b) Phase shift between donor and accepter intensities vs. frequency.

δT equals 1 and 2 the difference is less then 0.5%. We will show this below by numerically solving the Equation 5.1.

5.3 Kinetic Theory. Numerical Solutions

5.3.1 Finding folded and unfolded fractions

Here we will discuss the numerical solution of Equation 5.1, which reveals the dependence of the phase shift on the temperature wave amplitude. We used the following assumptions $\Delta G^{(1)}_{UF} = \lambda \Delta G^{(1)}_f$, $\Delta G^{(1)}_{FU} = (\lambda - 1)\Delta G^{(1)}_f$, $\Delta G^{(1)}_f = 1.5 kJ/mol/K$ and $\lambda = 0.3$ [143]. The folding and unfolding rates at $312K$ are equal to 0.5 Hz. Thus, $k'_0 = k_0 e^{-\frac{\Delta G^{(0)}}{k_B T}} = 0.5$ Hz.

We numerically solved Equation 5.1 for a range of driving frequencies $\nu = 0.01 - 4Hz$ and for the amplitudes $\delta T = \pm 1, \pm 2K$. Figure 5.4 illustrates one such

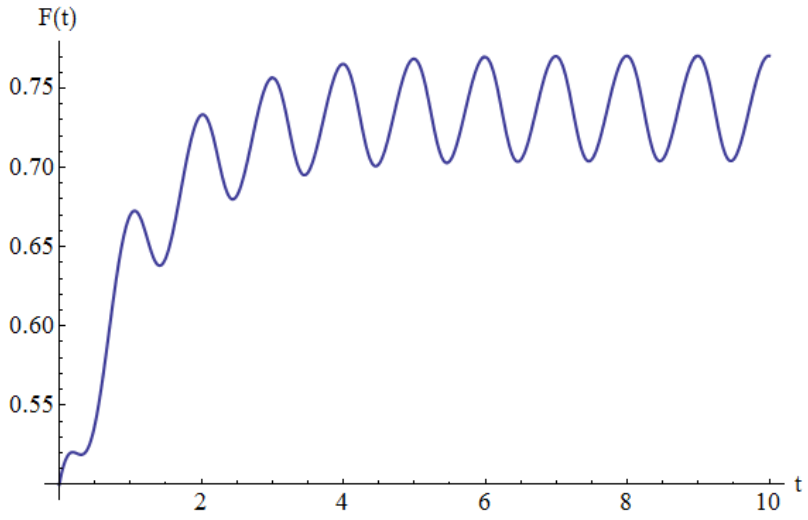


Figure 5.4: Numerical solution for folding population $F(t)$ for $\nu = 1Hz$, $\delta T = 2K$ and $F(0) = 0.5$.

solution. As expected, starting from its initial value at $t = 0$, $F(t)$ soon equilibrates to a harmonic function, which has the same frequency as the driving wave but is phase-shifted. At low frequencies, this phase shift will be equal to π, and it will asymptotically decrease to $\frac{\pi}{2}$ with increasing frequency. The unfolded population $U(t) = 1 - F(t)$ will fluctuate in-phase with the driving wave at low frequencies and approach $\frac{\pi}{2}$ with increasing ν. To find the phase shift of $F(t)$ at a given frequency, we fitted the equilibrated part of the $F(t)$ solution with a sinusoidal function. The dependence of the phase shift of $F(t)$ from ν for amplitudes $\delta T = \pm 1, \pm 2K$, and the difference between them are shown in Figure 5.5. As it can be seen, the phase shift difference is very small (less than 1 degree), but is not monotonic, with a maximum at about $\nu = 0.16$Hz.

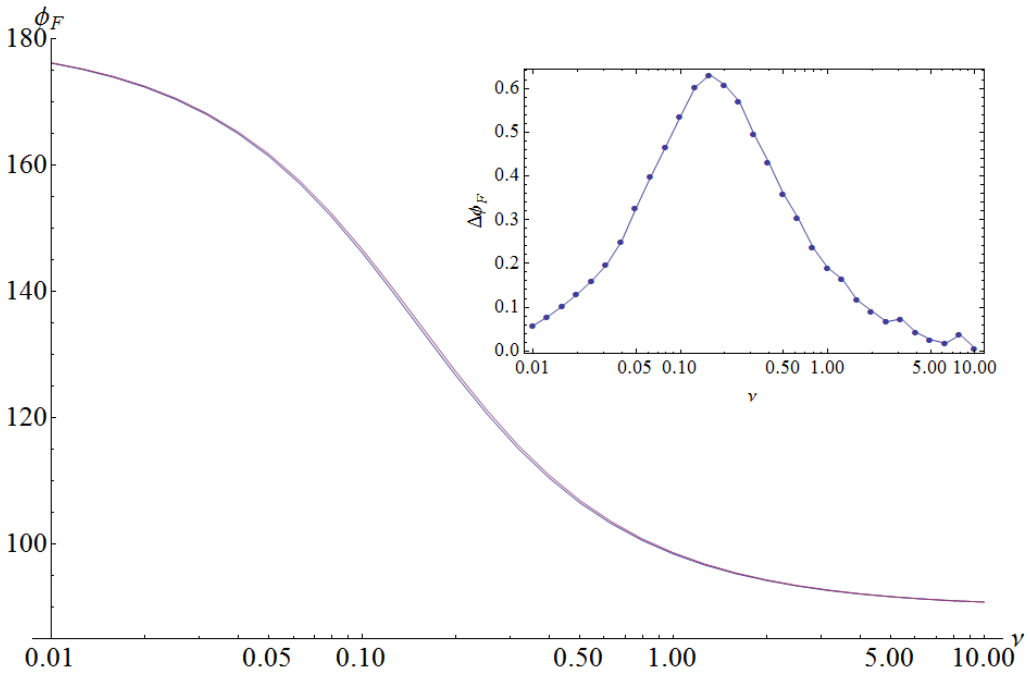


Figure 5.5: Kinetic theory, numerical solution. Phase-shift of folded population $F(t)$ vs. frequency ν (Hz) for $\delta T = \pm 1, \pm 2K$ (blue and red curves respectively). The insert shows the difference between the curves.

5.3.2 FRET Donor and Acceptor Intensities

Similar to the case of the analytical solution described in Section 5.2, we used the obtained numerical solutions of Equation 5.1 to find the phase shifts of donor and accepter intensities relative to the driving frequency. For that, we plugged the found $F(t)$ into Equation 5.23, and then into Equation 5.21. We used the following numerical parameters:

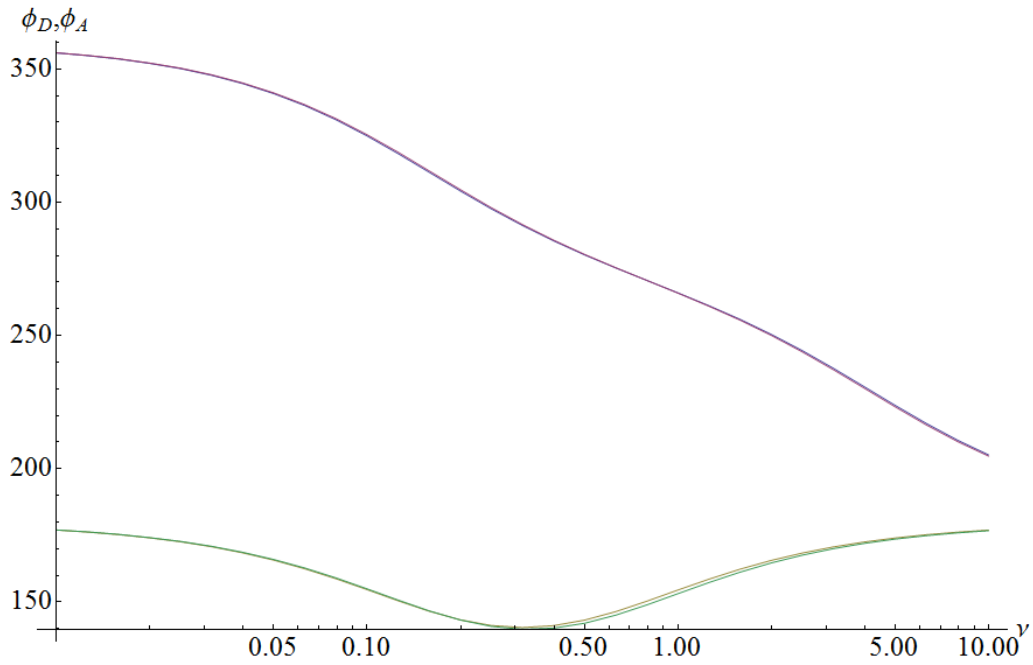


Figure 5.6: Kinetic theory, numerical solution. Donor (curves in the top) and accepter (curves in the bottom) phase shifts relative to driving wave vs. frequency.

p_1	5.72	p_2	-0.016
q_1	4.422	q_2	-0.0116
E_F	1.0	E_U	0.1

The dependence of donor and acceptor phase shifts vs. driving frequency are shown in Figure 5.6. The phase shifts between donor and accepter ($\phi = \phi_A - \phi_D$) for two different amplitudes are shown in Figure 5.7. The insert in Figure 5.7 shows the difference of the donor/accepter phase shifts between two driving amplitudes $\delta T = \pm 1, \pm 2K$. Same as in the case of $\Delta\phi_F$, $\Delta\phi$ has a maximum, but around $0.5Hz$.

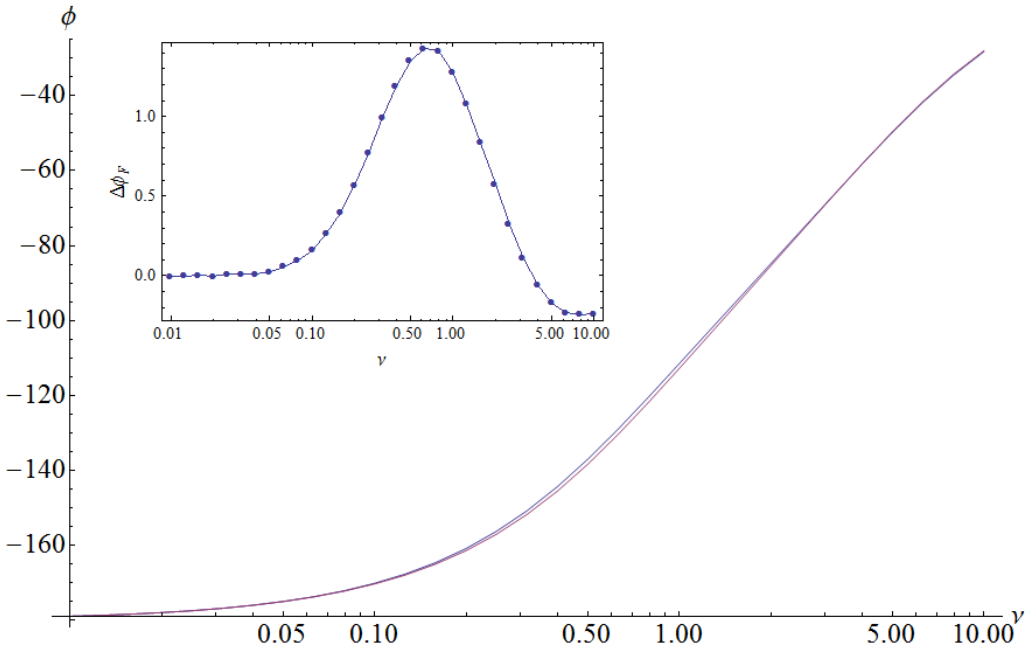


Figure 5.7: Kinetic theory, numerical solution. Phase shift between donor and accepter intensities vs. frequency ($\delta T = \pm 1, \pm 2K$ blue and red respectively). The insert shows the difference between the curves.

5.4 Brownian Dynamics

The kinetic transition theory described above does not explain the resonance phenomenon seen in our MD simulations. The average first-passage time (AFPT) calculations show a frequency independent acceleration compared to the unperturbed case when no wave is applied. Thus, to explain the phenomena we saw previously and to connect them to the experiments, we carried out a serious of Brownian Dynamics (BD) simulations of particles diffusing on a two-state free energy landscape.

5.4.1 Brownian Dynamics Simulations

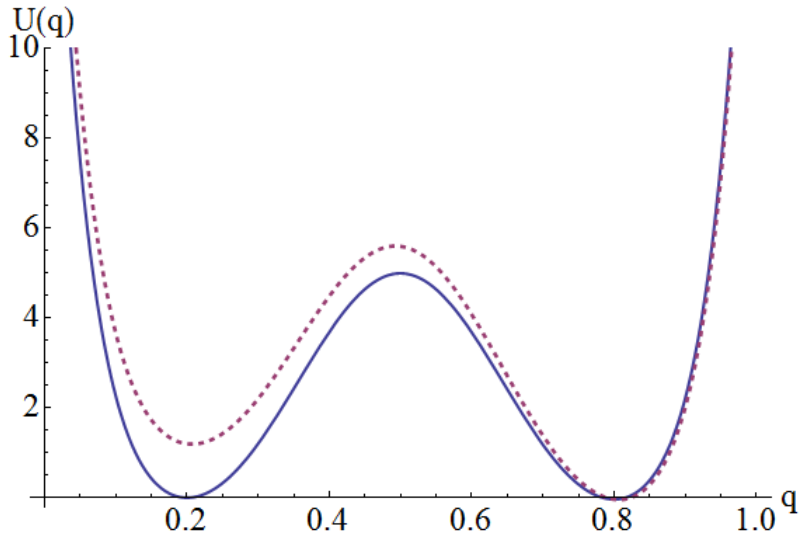


Figure 5.8: Energy landscape for Brownian particle moving along coordinate q at T=312K (solid) and T=311K (dashed). The unites of U(q) are kJ/mol. Right and left wells correspond to folded and unfolded states respectively.

In order to simplify the problem, we assumed that our system moves over a one dimensional energy landscape along a reaction coordinate q (see Figure 5.8). At the folding temperature of $T = 312K$, the energy landscape consists of two wells of equal depth, divided by a barrier of height $5kJ/mol$. The right well corresponds to the folded state of the system, and the left well to the unfolded state. As illustrated in Figure 5.8, the landscape shifts towards folded state with decreasing temperature. In high dumping regime, this system can be described using the Brownian equation of motion. By solving this equation we arrive to Equation 5.28,

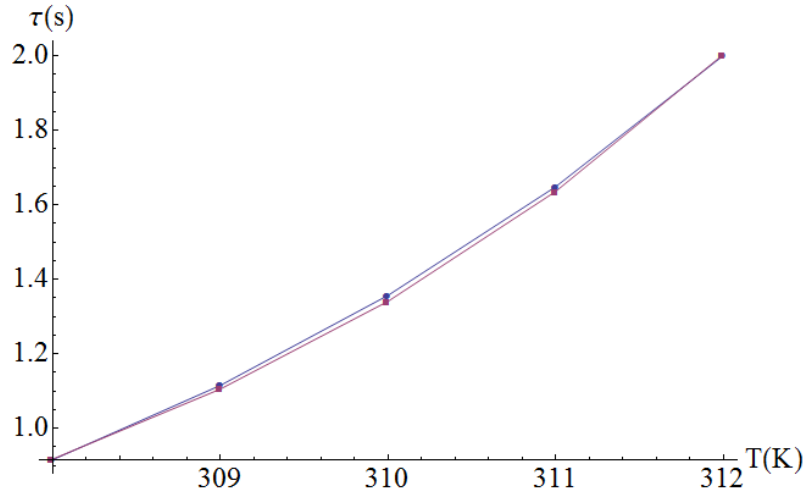


Figure 5.9: The figure shows the plot of Average First-Passage Time (AFPT) τ vs. temperature T. The blue circles correspond to the kinetic theory and the red squares to Brownian Dynamics results.

$$q(t+\Delta t) = q(t) + \frac{1}{\gamma}U'[(q(t), T(t)]\Delta t + \sqrt{\frac{2k_B T(t)\Delta t}{\gamma}}\eta(t) \tag{5.28}$$

where γ is the friction coefficient, k_B is the Boltzmann constant, $T(t)$ is the current temperature, $\eta(t)$ is a random number from the normal distribution, and $U' = \partial U/\partial q$. We chose the time step Δt equal to $0.001s$, which is more then 2 orders of magnitude smaller then the fastest characteristic timescale in our system. The γ friction coefficient and the rate of the change of the landscape were chosen so that the AFPT at $T = 312K$ equals to the experimentally known value of $2s$, and the AFPT vs. temperature dependence reproduces the one obtained using the kinetic theory (see Figure 5.9). The corresponding value of the friction coefficient used was $\gamma = 13.5 \times 10^3\ kg\ s^{-1}\ mol^{-1}$.

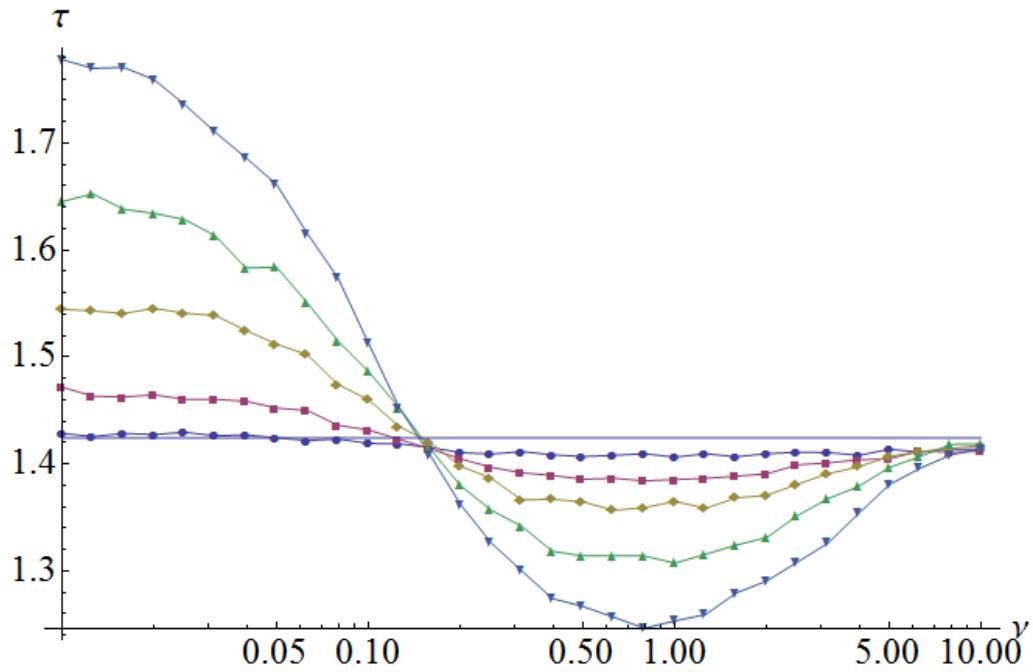


Figure 5.10: Brownian Dynamics results. Average First-Passage Time (AFPT) τ vs. fluctuation frequency ν for wave amplitudes 1K (blue circles), 2K (red squares), 3K (orange diamonds), 4K (green triangle) and 5K (blue inverted triangle). The horizontal line corresponds to the value of AFPT of unperturbed system.

5.4.2 Average First-Passage Time

To calculate AFPT, we run BD simulations for a system of one "particle" starting from a randomized unfolded state. The first time the system crossed the barrier, we recorded it and terminated the simulation. For each data point, we averaged over $100,000$ runs. Figure 5.9 shows the AFPT values for a range of constant temperatures from $308K$ to $312K$.

We calculated AFPT for the case of an applied harmonic temperature wave with average temperature $T_0 = 310K$ and amplitudes $\delta T = 1 - 5K$ (see Equa-

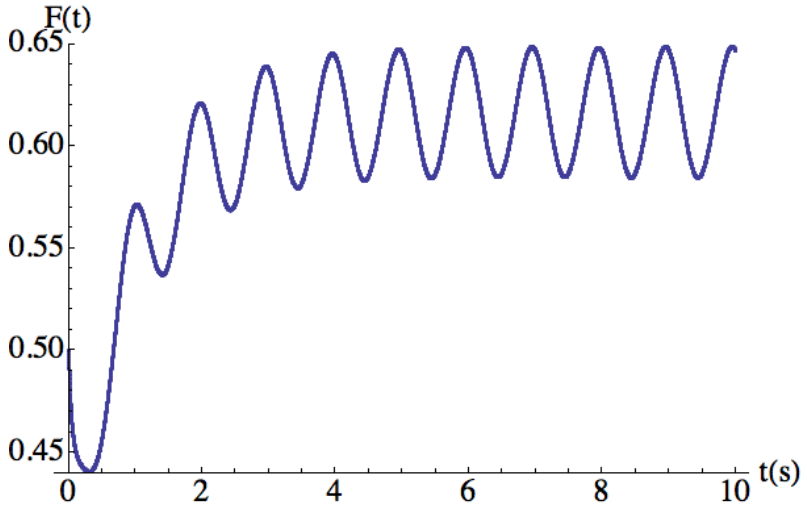


Figure 5.11: Brownian Dynamics results. Folded fraction population $F(t)$ vs. time for $\nu = 1Hz$, $\delta T = 2K$ and $F(0) = 0.5$.

tion 5.29) for the range of fluctuation frequencies from $0.1Hz$ to $10Hz$. For each run, the value of the initial phase ψ_0 was randomly chosen. Each data point in this case was averaged over $1,000,000$ trajectories. The results illustrated in Figure 5.10 are similar to the ones obtained from MD protein folding simulations. In both cases, we clearly see a decrease in AFPT around frequency values $\nu \approx k_f$. Also, in both cases the AFPT increases for low frequency fluctuation. In the former case, the applied temperature wave acts like a slowly increasing or decreasing driving force, depending on the initial phase ψ_0. The positive and negative contributions do not balance each other, resulting in the increase of AFPT relative to the unperturbed system.

$$T(t) = T_0 + \delta T \sin(2\pi\nu t + \psi_0) \tag{5.29}$$

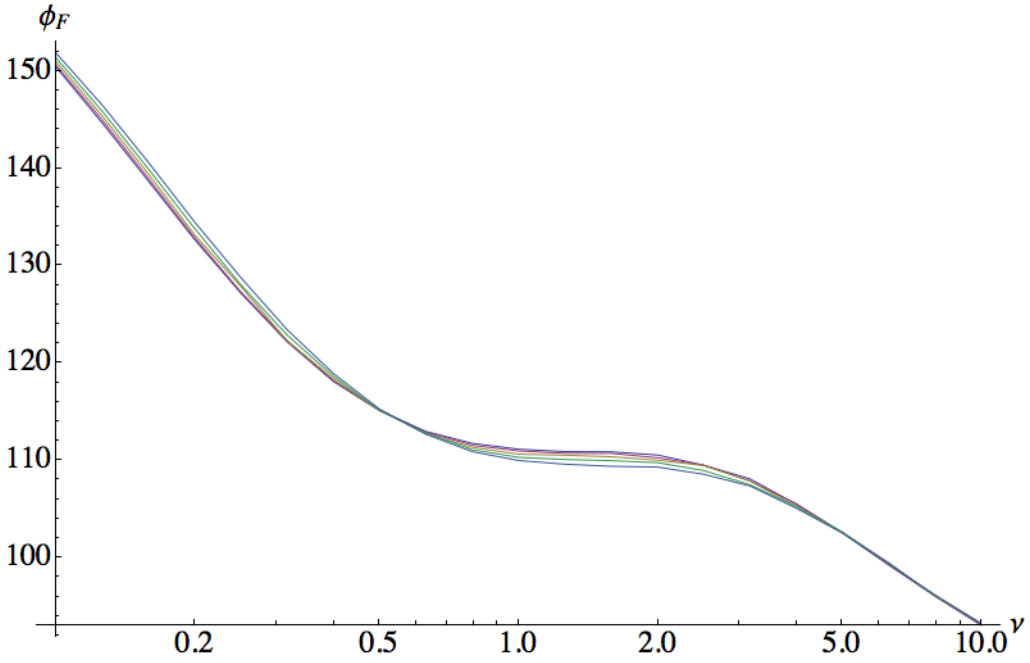


Figure 5.12: Brownian Dynamics results. Phase shift of folded population $F(t)$ vs frequency ν (Hz) for $\delta T = \pm 1K, \pm 2, \pm 3K, \pm 4K, \pm 5K$.

5.4.3 Calculations of Folded Fractions and Phase Shifts

To compare Brownian Dynamics simulations with the kinetic theory described previously, we computed the evolution of folded fractions over time for a range of values of driving frequencies and temperature wave amplitudes. For that reason, we ran BD simulations for a system containing $1,000,000$ particles for each case. Figure 5.11 illustrates folded fraction $F(t)$ for driving frequency $\nu = 1Hz$ and amplitude $\delta T = 2K$. In all the cases, we used initial phase $\psi_0 = 0$.

We fitted the equilibrated portion of $F(t)$ curves with a harmonic function and determined the phase shift of $F(t)$ relative to the driving temperature wave $T(t)$. Figure 5.12 shows the phase shift ϕ_F for a range of amplitudes.

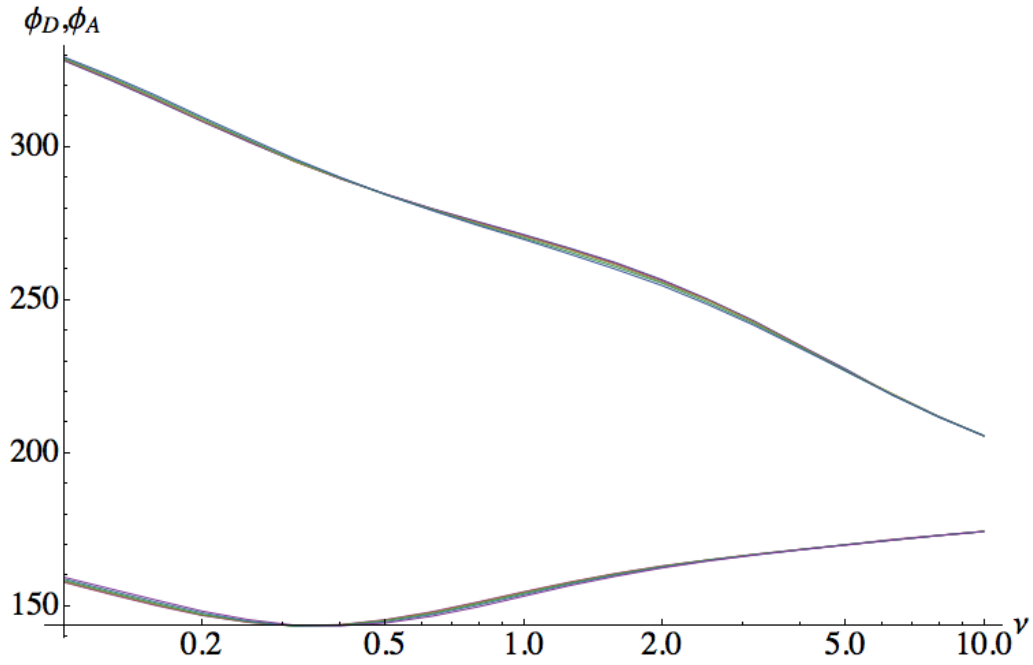


Figure 5.13: Brownian Dynamics results. Donor (curves in the top) and acceptor (curves in the bottom) phase shifts relative to temperature wave vs. frequency ν (Hz) for $\delta T = \pm 1K, \pm 2, \pm 3K, \pm 4K, \pm 5K$.

5.4.4 Donor and Acceptor Intensities and Phase Shifts

Using calculated $F(t)$ Brownian Dynamics trajectories and Equations 5.21, 5.23 and 5.24, we calculated corresponding FRET donor and accepter fluorescence intensities. For E_F, E_U, p_1, p_2, q_1 and q_2 we used the same values from Section 5.3.2. Here again, we estimated phase shifts relative to temperature wave from the equilibrated portions of $D(t)$ and $A(t)$ curves. Figures 5.13 and 5.14 show ϕ_A and ϕ_D phase shifts, and $\phi = \phi_A - \phi_D$ phase-shift difference vs. frequency over a range of amplitudes $1 - 5K$. The shape of these curves are very similar to ones obtained from the kinetic calculations.

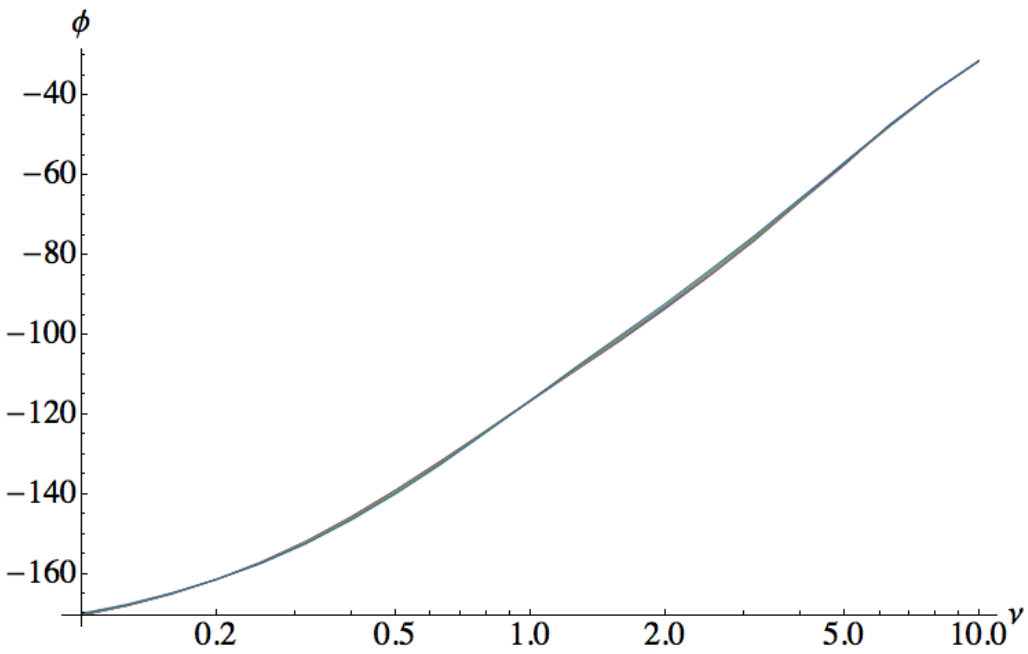


Figure 5.14: Brownian Dynamics results. Phase shift between donor and acceptor intensities vs. frequency ν (Hz) for $\delta T = \pm 1K, \pm 2, \pm 3K, \pm 4K, \pm 5K$.

5.4.5 Comparison to the Kinetic Theory

The kinetic theory described in Sections 5.2 and 5.3 predicts a weak dependence on the driving wave amplitude for the phase shift ϕ_F between the folded population and driving temperature wave, and the phase shift ϕ between donor and acceptor fluorescence intensities. This is illustrated in Figure 5.15 where a) and b) show how ϕ_F and ϕ depend on frequency for a range of amplitudes from $1K$ to $5K$. c) and d) show the phase shift differences between $2K$, $3K$, $4K$, $5K$ and $1K$ amplitudes for ϕ_F and ϕ, plotted over driving frequency. In general, the ϕ_F and ϕ obtained from BD simulations show somewhat similar dependence for both frequency and amplitude. For lower frequencies in particular, the agreement is very

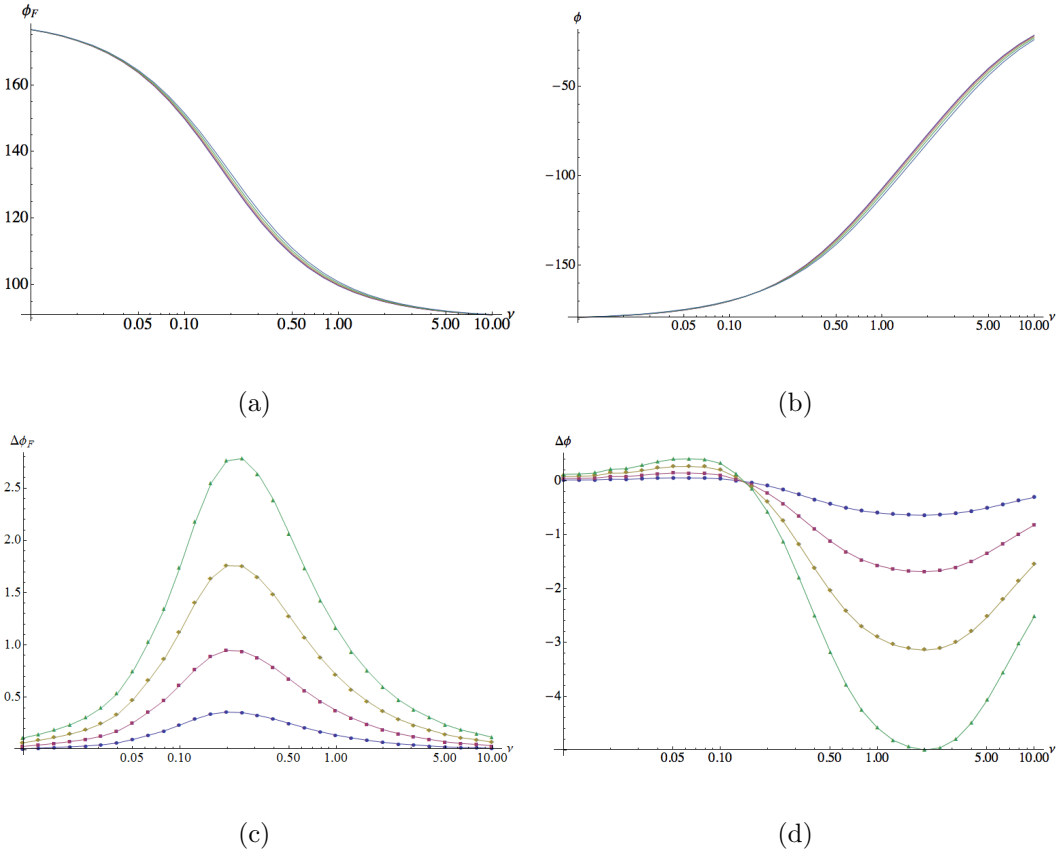


Figure 5.15: Kinetic theory, summary of numerical results. a) and b) phase shift ϕ_F between folded population and driving temperature wave, and phase shift ϕ between donor and acceptor fluorescence intensities respectively vs. frequency ν(Hz) for temperature wave amplitudes $\delta T = 1K, 2K, 3K, 4K, 5K$ c) and d) the differences between temperature wave amplitudes $2K$, $3K$, $4K$, $5K$ and $1K$ (blue circles. red squares, orange diamonds, and green triangles correspondingly) for ϕ_F and ϕ respectively vs. frequency ν(Hz).

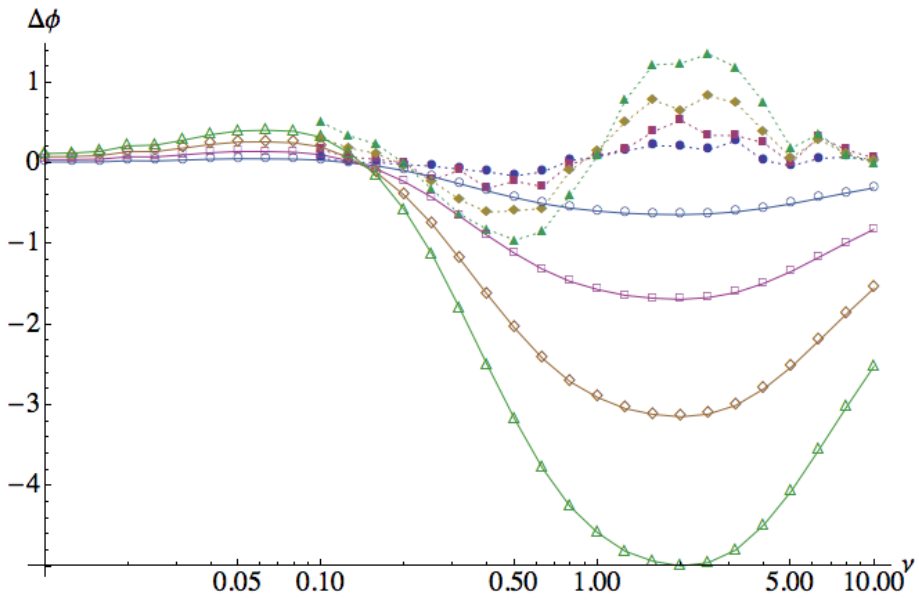


Figure 5.16: Difference of donor/acceptor phase shifts between temperature wave amplitudes $2K$, $3K$, $4K$, $5K$ and $1K$ (circles, squares, diamonds and triangles correspondingly). The solid curves correspond to kinetic theory and the dotted curves to BD simulations.

good, while for higher frequencies, and in particular for $\nu \sim k_f$, we see a relatively large deviation due to resonance, which, contrary to the kinetic theory, is observed in BD simulations (see Figure 5.9). For example, when we compare the $\phi_F(\nu)$ curves (see Figures 5.11 and 5.15(a)), we see a plateau at around $\nu = 1$Hz, which suggests that the folding reaction accelerates near that frequency range. This is supported by the fact that the average first-passage time calculations also show a minimum around that frequency value.

The experiments measure the difference of donor/acceptor phase shifts between different amplitudes. Thus, next we compared $\Delta\phi$ curves obtained from kinetic theory and BD simulations. As can be seen in Figure 5.15(d), $\Delta\phi$ calculated with the kinetic theory has a well pronounced minimum in the region of higher

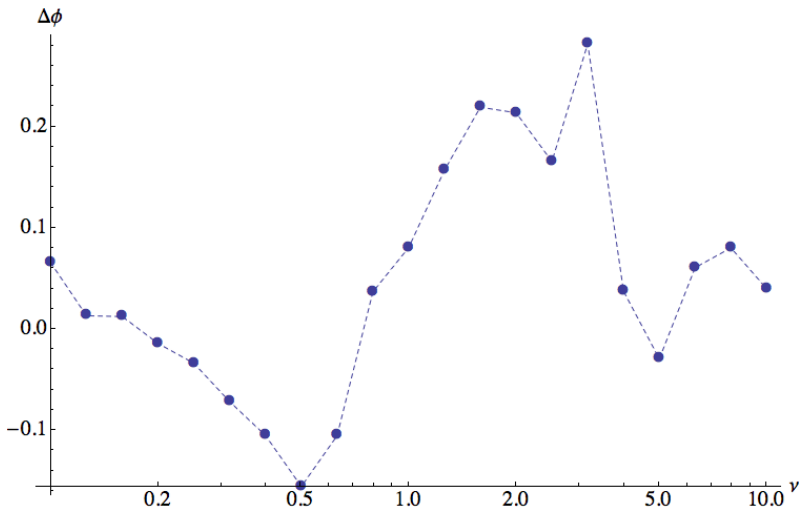


Figure 5.17: Difference of donor/acceptor phase shift between temperature wave amplitudes $2K$ and $1K$ calculated from Brownian Dynamics simulations.

frequency for all temperature wave amplitudes. This is not due to resonance, since the kinetic theory does not explain it. Figure 5.16 compares $\Delta\phi$ curves obtained from the kinetic theory to the ones from BD simulations. At lower frequencies, both sets of curves follow the same trend, but, for higher frequencies, BD curves invert, each showing a minimum at $\nu \sim 0.5Hz$ and then a maximum at $\nu \sim 3.0Hz$. This is consistent with the experimental results where the phase shift difference between amplitudes $2K$ and $1K$ also shows a minimum and a maximum of a similar magnitude near the same frequency values. The corresponding curve obtained from BD simulations is shown again in Figure 5.17.

5.5 Conclusion

Here we showed that Brownian Dynamics (BD) simulations of a particle moving on one dimensional energy landscape fully explain the results we obtained from Molecular Dynamics (MD) simulations of protein folding. They both show the existence of resonance under an applied harmonic wave when the fluctuation frequency approaches the folding rate of the unperturbed system. We also discussed the kinetic transition theory, which, as expected, does not explain the resonance phenomenon, but allows us to connect the BD simulations, and thus our MD results, to the experiments.

Chapter A: Supporting Information for Chapter 2

A.1 Introduction

Below we present the details of a coarse-grained model for protein simulations dubbed the Associative memory, Water mediated, Structure and Energy Model (AWSEM). This model has been continually developed over approximately two decades and successfully applied to many problems in protein physics [54, 55, 63–69, 77–79, 83, 84, 92, 101, 106, 144–159].

In this text and in our calculations in general we use kcal/mol for units of energy, Angstroms for length and radians for angles.

A.2 Description of the Coarse-grained Protein Chain

According to AWSEM, the position and orientation of each amino acid residue is dictated by the positions of its C_α, C_β and O atoms (with the exception of glycine, which lacks a C_β atom). The positions of the other atoms in the backbone are

calculated assuming an ideal geometry (Equation A.1).

$$\begin{aligned}
\mathbf{r}_{N_i} &= 0.48318\mathbf{r}_{C_{\alpha_{i-1}}} + 0.70328\mathbf{r}_{C_{\alpha_i}} - 0.18643\mathbf{r}_{O_{i-1}} \\
\mathbf{r}_{C'_i} &= 0.44365\mathbf{r}_{C_{\alpha_i}} + 0.23520\mathbf{r}_{C_{\alpha_{i+1}}} + 0.32115\mathbf{r}_{O_i} \\
\mathbf{r}_{H_i} &= 0.84100\mathbf{r}_{C_{\alpha_{i-1}}} + 0.89296\mathbf{r}_{C_{\alpha_i}} - 0.73389\mathbf{r}_{O_{i-1}}
\end{aligned} \tag{A.1}$$

The third line in Equation A.1 gives the position of the hydrogen atom that is attached to the backbone nitrogen. Side chains and solvent are not explicitly present in the model; instead, the effects of side chains and solvent are aliased onto various interactions described in the next section.

A.3 The AWSEM Hamiltonian

The solvent averaged free energy function of the protein chain is given in Equation A.2.

$$V_{total} = V_{backbone} + V_{contact} + V_{burial} + V_{HB} + V_{AM} + V_{DSB} \tag{A.2}$$

The backbone term, $V_{backbone}$, is responsible for restricting the chain to "protein-like" conformations. It consists of several parts, which are shown in Equation A.3.

$$V_{backbone} = V_{con} + V_{chain} + V_{\chi} + V_{rama} + V_{excl} \tag{A.3}$$

The connectivity of the protein chain is maintained by V_{con}, which is a sum of harmonic potentials. Its explicit form is given in Equation A.4, and a schematic of several amino acids is shown in Figure A.1.

$$\begin{aligned}
V_{con} &= \lambda_{con} \textstyle\sum_{i=1}^{N} [(\mathbf{r}_{C\alpha_i O_i} - \mathbf{r}^0_{C\alpha_i O_i})^2 + (\mathbf{r}_{C\alpha_i C_{\beta_i}} - \mathbf{r}^0_{C\alpha_i C_{\beta_i}})^2] \\
&+ \lambda_{con} \textstyle\sum_{i=1}^{N-1} [(\mathbf{r}_{C\alpha_i C\alpha_{i+1}} - \mathbf{r}^0_{C\alpha_i C\alpha_{i+1}})^2 + (\mathbf{r}_{O_i C\alpha_{i+1}} - \mathbf{r}^0_{O_i C\alpha_{i+1}})^2]
\end{aligned} \tag{A.4}$$

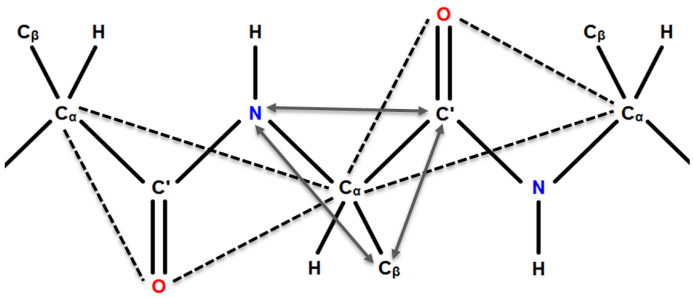


Figure A.1: The connectivity of the chain is maintained by a combination of harmonic potentials. The distances constrained by V_{con} are shown as dashed lines and the distances constrained by V_{chain} are shown as double headed arrows.

The values of λ_{con} and the equilibrium distances are given in Table A.1. In Equation A.4 and elsewhere, unless otherwise noted, i and j are residue indices and N is the total number of residues in the chain.

The correct bond angles around the C_α atom are also achieved using harmonic potentials, as shown in Equation A.5. The values of the parameters in Equation A.5 are given in Table A.1.

$$V_{chain} = \lambda_{chain}\left[\sum_{i=2}^{N}(\mathbf{r}_{N_iC_{\beta_i}} - \mathbf{r}^0_{N_iC_{\beta_i}})^2 + \sum_{i=1}^{N-1}(\mathbf{r}_{C'_iC_{\beta_i}} - \mathbf{r}^0_{C'_iC_{\beta_i}})^2 + \sum_{i=2}^{N-1}(\mathbf{r}_{N_iC'_i} - \mathbf{r}^0_{N_iC'_i})^2\right] \tag{A.5}$$

The chirality term, V_χ, given in Equation A.6, ensures the correct orientation of the C_β atom relative to the plane formed by the C', C_α and N atoms. A value of $\chi_0 = -0.83\text{Å}^3$, corresponding to an L-amino acid, is used for all residues except Glycine, which is excluded from this potential because it lacks a C_β atom. The

Table A.1: **Protein backbone potential parameters**

Parameter	Value	Unites
λ_{con}	10.0	kcal/Å^2 mol
λ_{chain}	5.0	kcal/Å^2 mol
λ_{χ}	20.0	kcal/Å^6 mol
λ_{rama}	2.0	kcal/mol
λ_{excl}	20.0	kcal/Å^2 mol
$\mathbf{r}^0_{C\alpha_i C\alpha_{i+1}}$	3.80	Å
$\mathbf{r}^0_{C\alpha_i CO_i}$	2.43	Å
$\mathbf{r}^0_{CO_i C\alpha_{i+1}}$	2.82	Å
$\mathbf{r}^0_{C\alpha_i C\beta_i}$	1.54	Å
$\mathbf{r}^0_{N_i C_{\beta_i}}$	2.46	Å
$\mathbf{r}^0_{C'_i C_{\beta_i}}$	2.70	Å
$\mathbf{r}^0_{N_i C'_i}$	2.46	Å
χ_0	-0.83	Å^3

values of the parameters in Equation A.6 are given in Table A.1.

$$V_{\chi} = \lambda_{\chi} \sum_{i=2}^{N-1} (\chi_i - \chi_0)^2$$
$$\chi_i = \left(\mathbf{r}_{C'_i C_{\alpha_i}} \times \mathbf{r}_{C_{\alpha_i} N_i}\right) \cdot \mathbf{r}_{C_{\alpha_i} C_{\beta_i}} \tag{A.6}$$

To reproduce the experimental distribution of backbone dihedral angles, we use a Ramachandran potential, V_{rama}, shown in Equation A.7. The resulting potential is plotted in Figure A.2. The value of λ_{rama} used is given in Table A.1. All other parameters are given in Table A.2, where ϕ_0 and ψ_0 are given in radians and W, σ, ω_ϕ and ω_ψ are unitless weights.

$$V_{rama} = -\lambda_{rama} \sum_{i=2}^{N-1} \sum_{j} W_j e^{-\sigma_j(\omega_{\phi_j}(\cos(\phi_i - \phi_{0j})-1)^2 + \omega_{\psi_j}(\cos(\psi_i - \psi_{0j})-1)^2)} \tag{A.7}$$

The first, last and glycine residues are not included in this potential. ϕ_i is the dihedral angle between the C'_{i-1}, N_i, C_{α_i} and C'_i atoms, and ψ_i is the dihedral angle

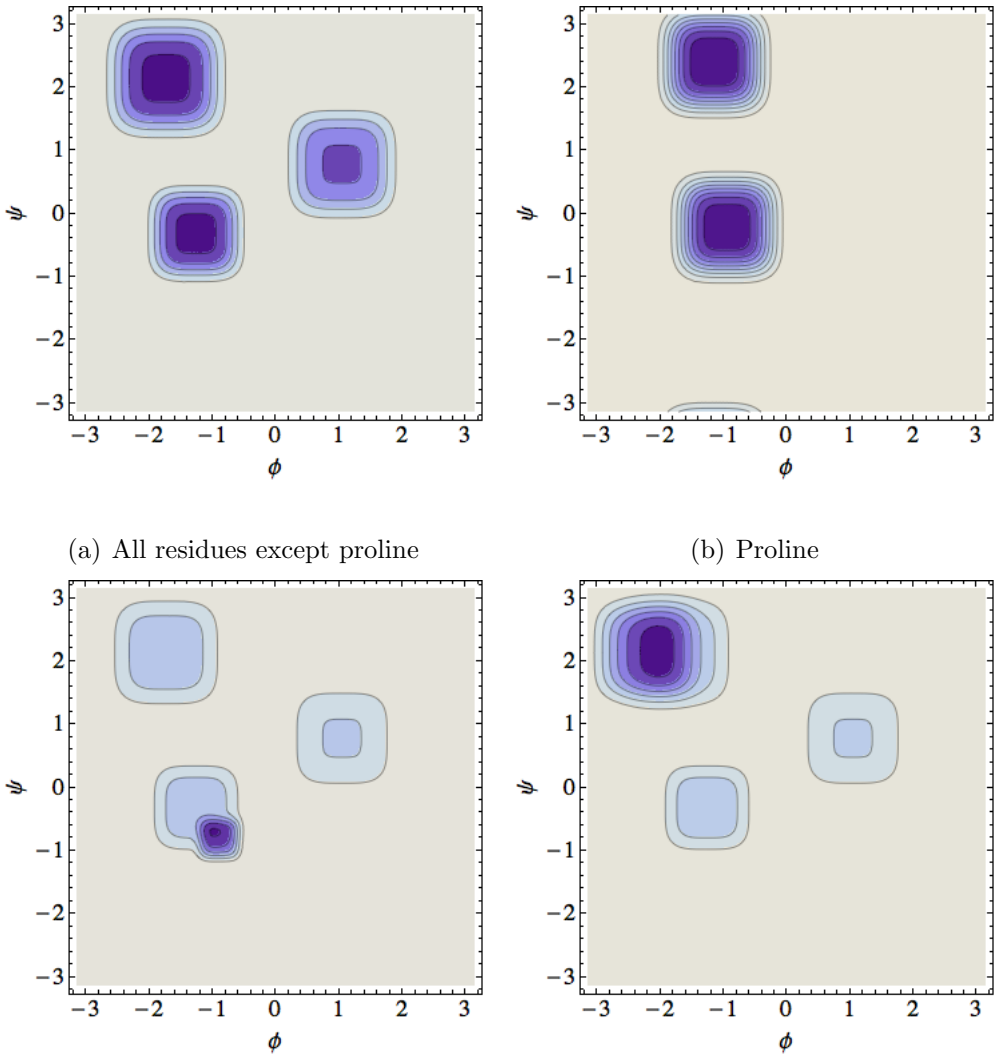
between the N_i, C_{α_i}, C'_i and N_{i+1} atoms.

Table A.2: **Ramachandran potential parameters**

	General Case			Alpha Helix	Beta Sheet	Proline	
W	1.3149	1.32016	1.0264	2.0	2.0	2.17	2.15
σ	15.398	49.0521	49.0954	419.0	15.398	105.52	109.09
ω_ϕ	0.15	0.25	0.65	1.0	1.0	1.0	1.0
ϕ_0	-1.74	-1.265	1.041	-0.895	-2.25	-1.153	-0.95
ω_ψ	0.65	0.45	0.25	1.0	1.0	0.15	0.15
ψ_0	2.138	-0.318	0.78	-0.82	2.16	2.4	-0.218

The first three columns of Table A.2 represent the set of the parameters for the general case of non-proline residues. These three columns correspond to right handed helix, left handed helix and β regions of the Ramachandran plot (see Figure A.2). Parameters from the next two columns can be used to bias the secondary structure towards right handed alpha helix or beta sheet based on a secondary structure prediction server (*e.g.,* JPRED [160]). The final two columns of Table A.2 refer to proline residues, which are known to have different allowed regions for the dihedral angles. The index j in Equation A.6 in this case is not a residue index; instead, it runs over each column of parameters that is appropriate for residue i.

V_{excl} is the excluded volume interaction that provides a repulsion between atoms at short distances, preventing them from overlapping. It has the form given in Equation A.8 where $r_{ex}^C = 3.5$Å for sequence separation less then 5 and 4.5Å otherwise, whereas $r_{ex}^O = 3.5$Å for any sequence separation. The subscript C refers to both C_α and C_β atoms. In Equation A.8, i and j are atom indices, which run over all pairs of C or O atoms that are not directly connected by V_{con}. $\Theta(x)$ is the



(a) All residues except proline

(b) Proline

(c) Biased toward alpha helix

(d) Biased toward beta sheet

Figure A.2: Ramachandran potential, V_{rama}. Secondary structure biasing is achieved by adding additional wells to the Ramachandran potential. The colors in (a), (b), (c) and (d) are not normalized to the same scale.

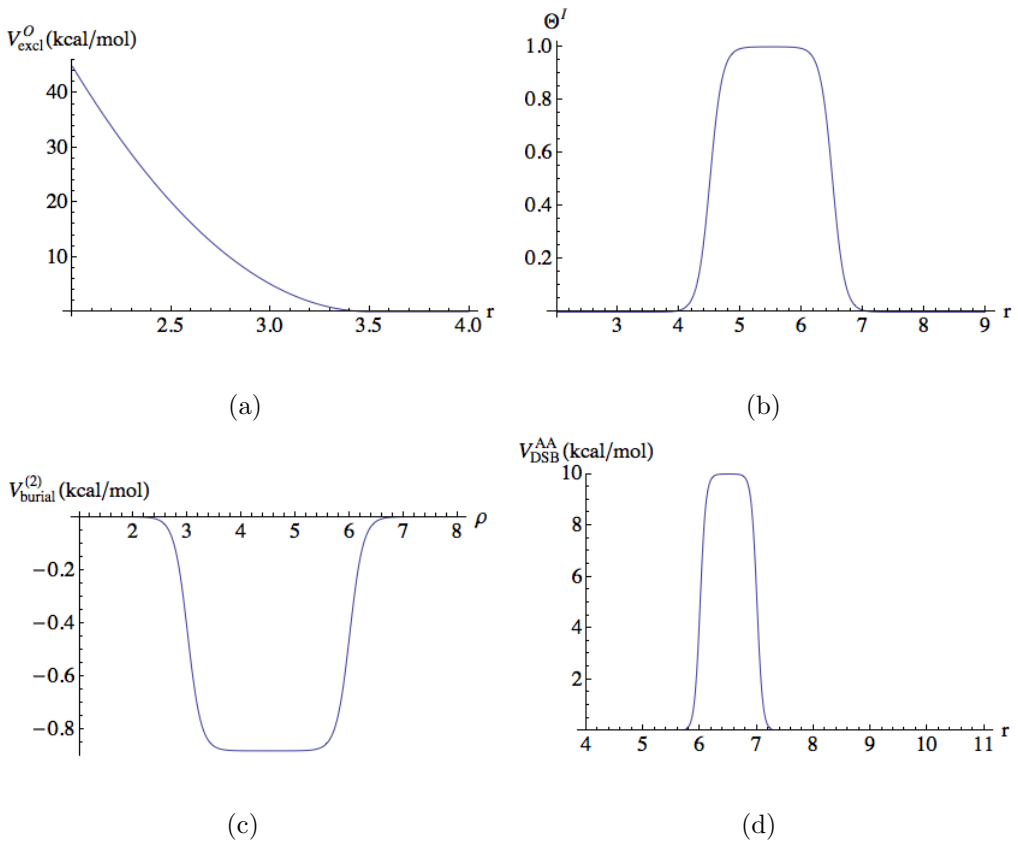


Figure A.3: (a) Plot of excluded volume potential vs. distance between two oxygens. (b) Plot of Θ function defined in Equation A.9 for direct contact well vs. distance between two residues. (c) Plot of burial potential function for $\mu = 2$ vs. local density value. (d) Plot of desolvation barrier between two alanines vs. distance between them.

Heaviside step function. The form of V_{excl} for single pair of oxygens is plotted in Figure A.3(a).

$$\begin{aligned} V_{excl} = \quad & \lambda_{excl} \sum_{ij} \Theta(r_{C_i,C_j} - r^{C}_{ex})(r_{C_i,C_j} - r^{C}_{ex})^2 \\ & + \lambda_{excl} \sum_{ij} \Theta(r_{O_i,O_j} - r^{O}_{ex})(r_{O_i,O_j} - r^{O}_{ex})^2 \end{aligned} \tag{A.8}$$

When describing $V_{contact}$, it is useful to define two C_β-C_β distance ranges (replaced by C_α atom in the case of glycine), hereafter identified by the superscripts I and II. The first distance range, the "direct contact well", goes from $r^{I}_{min} = 4.5$Å

to $r^I_{max} = 6.5$Å. The second distance range, the "water or protein mediated well", goes from $r^{II}_{min} = 6.5$Å to $r^{II}_{max} = 9.5$Å. If the C_β atoms of two residues i and j are separated by a distance between r^μ_{min} and r^μ_{max}, then the function Θ^μ_{ij}, given in Equation A.9, will be equal to 1; otherwise, it will be 0. It switches smoothly from 1 to 0 near the extremes of the distance ranges (see Figure A.3(b)).

$$\Theta^\mu_{ij} = \frac{1}{4}\left(1 + \tanh\left[\eta\left(r_{ij} - r^\mu_{min}\right)\right]\right)\left(1 + \tanh\left[\eta\left(r^\mu_{max} - r_{ij}\right)\right]\right) \tag{A.9}$$

By summing Θ^μ_{ij} over j, you can obtain the number of residues in the μ-well of residue i. The local density, ρ_i, of residue i is defined as $\rho_i = \sum_{j=1}^{N} \Theta^I_{ij}$, which is equal to the number of residues in its "direct contact well".

$V_{contact}$ is a contact interaction term between residues far apart in sequence [78]. It consists of V_{direct} and V_{water}. V_{direct} is a pairwise additive potential with the form given in Equation A.10

$$V_{direct} = -\lambda_{direct} \sum_{j-i>9}^{N} \gamma_{ij}(a_i, a_j)\Theta^I_{ij} \tag{A.10}$$

where r_{ij} is the C_β-C_β distance between residues i and j, and $\gamma(a_i, a_j)$ is a residue type specific constant. The γ parameters were optimized to maximize the ratio of the folding temperature[1] to the glass transition temperature[2] of the model, $\frac{T_f}{T_g}$ [78]. In Equation A.10 and elsewhere, a_i refers to the residue type of residue i.

V_{water} is a many-body interaction term that switches between water-mediated and protein-mediated interaction weights depending on the local density around the

[1] At the folding temperature, the populations of the folded state and unfolded states are equal.
[2] Below the glass transition temperature, the dynamics of the protein chain is arrested.

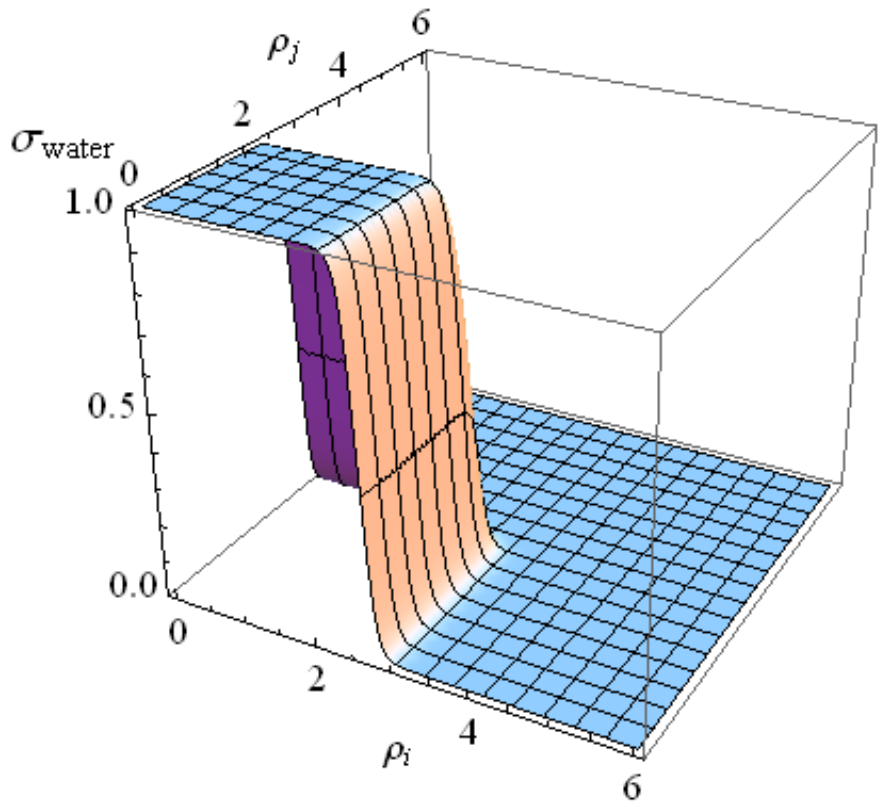


Figure A.4: Plot of σ_{ij}^{wat} in Equation A.12; adapted with permission from [78].

interacting residues. The explicit form is given in Equation A.11

$$V_{water} = -\lambda_{water} \sum_{j-i>9}^{N} \Theta_{ij}^{II} \left(\sigma_{ij}^{wat} \gamma_{ij}^{wat}(a_i, a_j) + \sigma_{ij}^{prot} \gamma_{ij}^{prot}(a_i, a_j) \right) \tag{A.11}$$

where σ_{ij}^{wat} and σ_{ij}^{prot} are the switching functions defined in Equation A.12.

$$\begin{aligned} \sigma_{ij}^{wat} &= \tfrac{1}{4} \left(1 - \tanh\left[\eta_\sigma \left(\rho_i - \rho_0\right)\right]\right) \left(1 - \tanh\left[\eta_\sigma \left(\rho_j - \rho_0\right)\right]\right) \\ \sigma_{ij}^{prot} &= 1 - \sigma_{ij}^{wat} \end{aligned} \tag{A.12}$$

σ_{ij}^{prot} and σ_{ij}^{wat} switch smoothly from 0 to 1 and 1 to 0, respectively, as either of the local densities, ρ_i or ρ_j, exceeds a threshold $\rho_0 = 2.6$. A plot of σ_{ij}^{wat} is given in Figure A.4.

Table A.3: **Potential parameters**

Parameter	Value	Unites	Parameter	Value	Unites
V_{direct}					
λ_{direct}	1.0	kcal/mol	η	5.0	$Å^{-1}$
V_{water}					
λ_{water}	1.0	kcal/mol	η	5.0	$Å^{-1}$
ρ_0	2.6		η_σ	7.0	
$V_{helical}$					
$\lambda_{helical}$	1.5	kcal/mol	ρ_0	3.0	
γ_{prot}	2.0		$\langle r^{ON} \rangle$	2.98	Å
γ_{wat}	-1.0		$\langle r^{OH} \rangle$	2.06	Å
η	7.0	$Å^{-1}$	σ_{ON}	0.68	Å
η_σ	7.0		σ_{OH}	0.76	Å
V_β					
$\langle r^{ON} \rangle$	2.98	Å	η^I	1.0	$Å^{-1}$
$\langle r^{OH} \rangle$	2.06	Å	η^{II}	0.5	$Å^{-1}$
σ_{ON}	0.68	Å	r_c^{HB}	12.0	Å
σ_{OH}	0.76	Å			
V_{P-AP}					
γ_{APH}	1.0	kcal/mol	η	7.0	$Å^{-1}$
γ_{AP}	0.4	kcal/mol	r_0	8.0	Å
γ_P	0.4	kcal/mol			
V_{burial}					
λ_{burial}	1.0	kcal/mol	η	4.0	
V_{AM}					
λ_{AM}	1.0	kcal/mol			
V_{DSB}					
λ_{DSB}	10.0	kcal/mol	r_{min}^0	6.0	Å
κ_{DSB}	10.0	$Å^{-1}$	r_{max}^0	7.0	Å

The burial term, V_{burial}, given in Equation A.13, is a many body interaction which is based on a particular residue type's propensity to be in a low ($\mu = 1$, $\rho^1_{min} = 0.0$, $\rho^1_{max} = 3.0$), medium ($\mu = 2$, $\rho^2_{min} = 3.0$, $\rho^2_{max} = 6.0$), or high ($\mu = 3$, $\rho^3_{min} = 6.0$, $\rho^3_{max} = 9.0$) density environment [78]. These propensities are given by the $\gamma_{burial}(a_i, \rho_i)$ coefficients in Table A.4.

$$V_{burial} = -\frac{1}{2}\lambda_{burial} \sum_{i=1}^{N} \sum_{\mu=1}^{3} \gamma_{burial}(a_i, \rho_i) \left(\tanh\left[\eta\left(\rho_i - \rho^{\mu}_{min}\right)\right] + \tanh\left[\eta\left(\rho^{\mu}_{max} - \rho_i\right)\right]\right) \tag{A.13}$$

Table A.4: **Burial potential, V_{burial}, coefficients $\gamma_{burial}(a_i, \rho_i)$**

a_i	ρ_i		
	0.0-3.0	3.0-6.0	6.0-9.0
Ala	0.84	0.88	0.57
Arg	0.94	0.83	0.13
Asn	0.96	0.79	0.25
Asp	0.98	0.75	0.20
Cys	0.67	0.94	0.66
Gln	0.96	0.79	0.24
Glu	0.97	0.78	0.16
Gly	0.94	0.81	0.34
His	0.92	0.85	0.13
Ile	0.78	0.92	0.55
Leu	0.78	0.94	0.46
Lys	0.98	0.75	0.00
Met	0.82	0.92	0.46
Phe	0.81	0.94	0.33
Pro	0.97	0.76	0.25
Ser	0.94	0.79	0.38
Thr	0.92	0.82	0.40
Trp	0.85	0.91	0.34
Tyr	0.83	0.92	0.34
Val	0.77	0.93	0.55

The hydrogen bonding potential, V_{HB}, given in Equation A.14, is a sum of

three terms.

$$V_{HB} = V_{\beta} + V_{P-AP} + V_{helical} \tag{A.14}$$

The first two terms of Equation A.14 are β hydrogen bonding terms. The V_{β} potential has the form given in Equations A.15, A.16, A.17, and A.18 where r_{ij}^{ON} is the distance from the carbonyl oxygen of residue i to the nitrogen of residue j, and r_{ij}^{OH} is the distance from carbonyl oxygen of residue i to the backbone amide hydrogen of residue j. $\langle r^{ON} \rangle$ and $\langle r^{OH} \rangle$ are the corresponding equilibrium bond lengths, and σ_{NO} and σ_{HO} are their variances.

$$V_{\beta}^{ij} = -[\Lambda_1(|j-i|)\theta_{i,j} + \Lambda_2(a_i, a_j, |j-i|)\theta_{i,j}\theta_{j,i} + \Lambda_3(a_i, a_j, |j-i|)\theta_{i,j}\theta_{j,i+2}]\nu_i^I \nu_j^{II} \tag{A.15}$$

$$\theta_{i,j} = \exp\left[-\frac{(r_{ij}^{ON} - \langle r^{ON} \rangle)^2}{2\sigma_{NO}^2} - \frac{(r_{ij}^{OH} - \langle r^{OH} \rangle)^2}{2\sigma_{HO}^2}\right] \tag{A.16}$$

$$\nu_i^{\mu} = \frac{1}{2}\left(1 + \tanh\left[\eta^{\mu}\left(r_{i-2,i+2}^{C\alpha} - r_c^{HB}\right)\right]\right) \tag{A.17}$$

$$\begin{aligned}
\Lambda_1(|j-i|) &= \lambda_1(|j-i|) \\
\Lambda_2(a_i, a_j, |j-i|) &= \lambda_2(|j-i|) - 0.5\alpha_1(|j-i|)lnP_{HB}(a_i, a_j) \\
&\quad -0.25\alpha_2(|j-i|)[lnP_{NHB}(a_{i+1}, a_{j-1}) + lnP_{NHB}(a_{i-1}, a_{j+1})] \\
&\quad -\alpha_3(|j-i|)[lnP_{anti}(a_i) + lnP_{anti}(a_j)] \\
\Lambda_3(a_i, a_j, |j-i|) &= \lambda_3(|j-i|) - \alpha_4(|j-i|)lnP_{parHB}(a_{i+1}, a_j) \\
&\quad -\alpha_5(|j-i|)lnP_{par}(a_{i+1}) + \alpha_4(|j-i|)lnP_{par}(a_j)
\end{aligned} \tag{A.18}$$

The first term in the Equation A.15 describes simple pairwise additive hydrogen bonding interactions. The second term gives additional cooperative stabilization to anti-parallel β conformations and the third term gives additional cooperative

stabilization to parallel β conformations. All of the Λ_k coefficients depend on the sequence separation of residues i and j, and the coefficients Λ_2 and Λ_3 are also amino acid type (a_i and a_j) dependent. The constants $\langle r^{ON} \rangle$, $\langle r^{OH} \rangle$, σ_{NO}, σ_{HO} (see Table A.3) and probabilities, P, for amino acids to be hydrogen bonded (HB) or not hydrogen bonded (NHB) were extracted from a database of well-resolved protein structures [69]. The parameters λ and α of Equation A.18 were optimized to maximize the T_f/T_g ratio [78]. Their values for different sequence separation classes are given in Table A.5. For $|j-i| < 18$, $\lambda_3 = 0$ because parallel hydrogen bonds rarely form between residues which are less than 18 amino acids apart. The ν_i and ν_j terms ensure that β hydrogen bonding does not occur between residues that are in the middle of a five residue segment that is shorter than $r_c^{HB} = 12.0$Å, as β hydrogen bonding networks tend not to form between chain segments that are not at least somewhat extended.

Table A.5: **Hydrogen bonding potential λ and α coefficients, in** $kcal/mol$

sequence separation	λ_1	λ_2	λ_3	α_1	α_2	α_3	α_4	α_5
$4 \leq \|j-i\| < 18$	1.37	3.89	0.0	1.30	1.32	1.22	0.0	0.0
$18 \leq \|j-i\| < 45$	1.36	3.50	3.47	1.30	1.32	1.22	0.33	1.01
$\|j-i\| \geq 45$	1.17	3.52	3.62	1.30	1.32	1.22	0.33	1.01

V_β will stabilize an already formed β hydrogen bonding network, but small deviations from an ideal β-sheet geometry will be significantly higher in energy. However, during secondary structure formation it is necessary to search through many possible conformations. The "liquid-crystal potential", V_{P-AP}, enables a pro-

tein chain to adopt approximate parallel or antiparallel β-sheet conformations before the hydrogen bonds are fully formed. The strength of this potential is chosen so that structures can easily fall apart and reassemble. The general form of this potential is given in Equation A.19.

$$\begin{aligned} V_{P-AP} = \quad & -\gamma_{APH} \sum_{i=1}^{N-13} \sum_{j=i+13}^{min(i+16,N)} \nu_{i,j}\nu_{i+4,j-4} \\ & -\gamma_{AP} \sum_{i=1}^{N-17} \sum_{j=i+17}^{N} \nu_{i,j}\nu_{i+4,j-4} - \gamma_P \sum_{i=1}^{N-13} \sum_{j=i+9}^{N-4} \nu_{i,j}\nu_{i+4,j+4} \end{aligned} \tag{A.19}$$

V_{P-AP} favors contacts between residues i and j if residues $i+4$ and $j+4$ (parallel, P) or $i+4$ and $j-4$ (antiparallel, AP) are already in contact. Formation of β-hairpins (APH) is separate from the general antiparallel case to allow for the possibility of assigning it a different weight. Two residues are considered to be in contact with each other if the distance between their C_α atoms is less than r_0. Thus, $\nu_{i,j}$ is defined as the smooth switching function $\nu_{i,j} = \frac{1}{2}\left(1 + \tanh\left[\eta\left(r_0 - r_{C\alpha_i,C\alpha_j}\right)\right]\right)$, where $\eta = 7.0\text{Å}^{-1}$ and $r_0 = 8.0$Å. γ_{AP} and γ_P usually take the value of 0.4 kcal/mol. Only in the case when secondary structure prediction information is available and both residues i and j are predicted to be in a β-strand do we use a value of $\gamma_{AP} = \gamma_P = 0.6$ kcal/mol instead.

The $V_{helical}$ term, given in Equation A.20, is responsible for the formation of alpha helices [161].

$$\begin{aligned} V_{helical} = \quad & -\lambda_{helical} \sum_{i=1}^{N-4} (f(a_i) + f(a_{i+4}))(\gamma_{prot}\sigma_{i,i+4}^{prot} + \gamma_{wat}\sigma_{i,i+4}^{wat}) \times \\ & \exp\left[-\frac{\left(r_{i,i+4}^{ON} - \langle r^{ON}\rangle\right)^2}{2\sigma_{ON}^2} - \frac{\left(r_{i,i+4}^{OH} - \langle r^{OH}\rangle\right)^2}{2\sigma_{OH}^2}\right] \end{aligned} \tag{A.20}$$

In Equation A.20, $f(a_i)$ (see Table A.6) is the probability of finding residue i in a helix. All residue types have positive values between 0 and 1 except for proline,

as it lacks a backbone amide hydrogen and therefore can only be a hydrogen bond acceptor, but never a donor. To reflect this we use $f(a_{i+4}) = -3.0$ if the $i+4$ residue is a proline. σ_{ij}^{prot} and σ_{ij}^{wat} are the same as in Equation A.12. γ_{prot} is the strength of the interaction when both residues are buried. When residues are exposed to water, they are allowed to form hydrogen bonds with surrounding water molecules and forming hydrogen bonds with each other is not as favorable. Thus γ_{wat} is negative, as shown in Table A.3.

Table A.6: $f(a_i)$ **values**

a_i	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE
$f(a_i)$	0.77	0.68	0.07	0.15	0.23	0.33	0.27	0.0	0.06	0.23
a_i	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
$f(a_i)$	0.62	0.65	0.50	0.41	0.4/-3.0	0.35	0.11	0.45	0.17	0.14

V_{AM} is the associative memory potential. When combined with known protein structures and an algorithm for aligning a target sequence to those structures, it can be used to limit the local (secondary structure) conformational search. Each portion of a known structure which is aligned to a particular set of residues in the target sequence is known as a "memory". In this paper, we have used a "fragment memory" approach wherein the memories are short (9 residues or less) and the fragments are chosen using BLAST [99]. The maximum sequence separation of interacting residues is determined either by the length of the memory or a maximum cutoff, whichever is shorter. The form of the V_{AM} potential is given in Equation A.21 where i and j go over all C_α and C_β atoms up to a maximum sequence separation, which in this

case includes the entire fragment. In Equation A.21, i and j are not residue indices, but atom indices. ω_m is the weight of the memory, $\gamma_{AM}(a_i, a_j)$ is a residue type dependent interaction strength, and r_{ij}^m is the distance between the i and j atoms in the memory structure. In the simplest case, as was used in this paper, both ω_m and $\gamma_{AM}(a_i, a_j)$ are 1.0 for all memories and all residue types. λ_{AM} is an overall scaling factor for the associative memory term, which can be used to adjust the weight of the term relative to others in the Hamiltonian, and $\sigma_{IJ} = |I - J|^{0.15}$ is a sequence separation dependent width.

$$V_{AM} = -\lambda_{AM} \sum_m \omega_m \sum_{ij} \gamma_{ij} \exp\left[-\frac{(r_{ij} - r_{ij}^m)^2}{2\sigma_{IJ}^2}\right] \tag{A.21}$$

V_{DSB} is a desolvation barrier potential. When pairs of residues are separated by a distance that is less than the width of a water, but they are not in direct contact, there is an energetic barrier that comes from the formation of a vacuum [162]. The form of the potential is given in Equation A.22 where r_{ij} is the $C_\beta - C_\beta$ distance, except when a glycine is involved, in which case the C_α coordinates for the glycine are used.

$$\begin{aligned} V_{DSB} &= \lambda_{DSB} \sum_{j-i>9}^{N} \frac{1}{2} \left(\tanh\left[\kappa_{DSB}(r_{ij} - r_{min}^{DSB}(a_i, a_j))\right] + \right. \\ &\quad \left. \tanh\left[\kappa_{DSB}(r_{max}^{DSB}(a_i, a_j) - r_{ij})\right]\right) \\ r_{min}^{DSB}(a_i, a_j) &= r_{min}^0 + r_{shift}(a_i) + r_{shift}(a_j) \\ r_{max}^{DSB}(a_i, a_j) &= r_{max}^0 + r_{shift}(a_i) + r_{shift}(a_j) \end{aligned} \tag{A.22}$$

Typical values for the parameters in Equation A.22 are given in Table A.3. A sample plot of V_{DSB} interaction potential between two alanines is shown in Fig-

ure A.3(d). As indicated, the minimum and maximum distances at which the desolvation barrier is activated, r_{min}^{DSB} and r_{max}^{DSB}, are residue type dependent. The details of r_{shift} are given in Table A.7.

Table A.7: $r_{shift}(a_i)$ **values, in** Å

a_i	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE
$r_{shift}(a_i)$	0.00	2.04	0.57	0.57	0.36	1.11	1.17	−1.52	0.87	0.67
a_i	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
$r_{shift}(a_i)$	0.79	1.47	1.03	1.00	−0.10	0.26	0.37	1.21	1.15	0.39

A.4 Simulation Protocol

We performed all molecular dynamics simulations using the Nose-Hoover thermostat as implemented in the open source simulation package LAMMPS. We recently extended LAMMPS by implementing all of the AWSEM potentials described in this supplement and adding a special atom style (called peptide), which is suitable for heteropolymeric systems such as proteins. All of our extensions to LAMMPS, as well as all analysis tools used for the current study, are available under the GNU General Public License at http://code.google.com/p/awsemmd/.

We started all structure prediction simulations from an extended conformation at a temperature well above the folding temperature. The simulations ran for 4×10^6 steps under non-periodic boundary conditions to a temperature well below the folding temperature. We used a timestep of 3 femtoseconds and saved the coor-

dinates of the system every 1000 steps. For each saved snapshot, we calculated Q and RMSD values relative to an experimentally determined structure.

Table A.8: **Direct contact potential, V_{direct}, and Water potential, V_{water}, coefficients $\gamma^{dir}(a_i, a_j)$, $\gamma^{prot}(a_i, a_j)$, $\gamma^{wat}(a_i, a_j)$**

a_i	a_j	γ^{dir}	γ^{prot}	γ^{wat}	a_i	a_j	γ^{dir}	γ^{prot}	γ^{wat}
ALA	ALA	0.72	0.09	0.02	ALA	ARG	-0.27	0.04	-0.00
ALA	ASN	-0.26	0.01	-0.00	ALA	ASP	-0.40	0.00	-0.07
ALA	CYS	0.62	0.27	0.29	ALA	GLN	-0.24	-0.02	-0.12
ALA	GLU	-0.35	0.02	-0.09	ALA	GLY	-0.11	0.05	-0.04
ALA	HIS	-0.13	0.03	-0.16	ALA	ILE	1.00	0.12	0.21
ALA	LEU	1.00	0.10	0.26	ALA	LYS	-0.45	0.02	0.08
ALA	MET	0.51	0.16	0.06	ALA	PHE	0.57	0.31	0.31
ALA	PRO	-0.53	-0.00	-0.00	ALA	SER	-0.21	-0.00	0.04
ALA	THR	0.08	0.05	0.03	ALA	TRP	0.40	0.09	-0.08
ALA	TYR	0.11	0.19	0.14	ALA	VAL	0.92	0.33	0.25
ARG	ARG	-0.64	-0.05	0.62	ARG	ASN	-0.28	-0.05	0.64
ARG	ASP	0.41	0.02	1.00	ARG	CYS	-0.40	0.43	0.46
ARG	GLN	-0.21	-0.04	0.43	ARG	GLU	-0.03	-0.03	0.97
ARG	GLY	-0.33	-0.01	0.32	ARG	HIS	-0.53	-0.06	0.32
ARG	ILE	-0.14	-0.04	0.07	ARG	LEU	-0.25	-0.07	-0.04
ARG	LYS	-0.96	-0.08	0.47	ARG	MET	-0.02	-0.16	0.14
ARG	PHE	-0.18	-0.13	-0.11	ARG	PRO	-0.82	0.01	0.43
ARG	SER	-0.33	0.01	0.32	ARG	THR	-0.23	-0.01	0.35
ARG	TRP	-0.30	-0.20	-0.05	ARG	TYR	0.14	0.15	-0.47
ARG	VAL	-0.17	0.01	0.11	ASN	ASN	0.16	-0.03	0.58
ASN	ASP	0.02	-0.01	0.28	ASN	CYS	-0.09	0.16	0.17
ASN	GLN	-0.19	-0.02	0.39	ASN	GLU	-0.56	-0.03	0.27
ASN	GLY	-0.14	0.01	0.10	ASN	HIS	-0.07	0.00	0.13
ASN	ILE	-0.72	-0.22	0.24	ASN	LEU	-0.58	-0.13	0.19
ASN	LYS	-0.45	-0.05	0.44	ASN	MET	-0.60	-0.10	-0.10
ASN	PHE	-0.52	-0.11	0.10	ASN	PRO	-0.69	-0.01	0.57
ASN	SER	-0.02	0.00	0.31	ASN	THR	-0.31	-0.02	0.30
ASN	TRP	-0.37	0.08	-0.30	ASN	TYR	-0.27	0.14	-0.45
ASN	VAL	-0.59	-0.11	-0.00	ASP	ASP	-0.57	0.00	0.23
ASP	CYS	-0.37	-0.24	0.52	ASP	GLN	-0.39	-0.03	0.31
ASP	GLU	-0.85	-0.04	0.20	ASP	GLY	-0.30	-0.02	0.25
ASP	HIS	-0.08	0.01	0.61	ASP	ILE	-0.72	-0.18	0.27
ASP	LEU	-0.78	-0.20	0.24	ASP	LYS	0.11	-0.03	0.84
ASP	MET	-0.58	-0.18	-0.02	ASP	PHE	-0.76	-0.19	0.00
ASP	PRO	-0.82	-0.02	0.48	ASP	SER	-0.03	-0.00	0.09
ASP	THR	-0.22	-0.01	0.18	ASP	TRP	-0.74	-0.13	-0.14
ASP	TYR	-0.78	0.05	-0.43	ASP	VAL	-0.74	-0.15	0.18
CYS	CYS	0.98	0.39	0.64	CYS	GLN	-0.43	0.16	0.66
CYS	GLU	-0.36	0.15	-0.15	CYS	GLY	0.43	0.39	-0.08
CYS	HIS	0.69	0.03	-0.04	CYS	ILE	0.70	0.33	0.91
CYS	LEU	0.98	0.31	0.25	CYS	LYS	-0.58	-0.01	0.30
CYS	MET	0.30	0.73	-0.52	CYS	PHE	0.85	0.88	0.77
CYS	PRO	0.09	0.39	0.02	CYS	SER	0.47	0.52	0.15
CYS	THR	-0.18	0.34	-0.11	CYS	TRP	0.10	0.58	1.00
CYS	TYR	0.87	0.52	0.42	CYS	VAL	0.95	0.62	0.00

Table A.8 – continue

a_i	a_j	γ^{dir}	γ^{prot}	γ^{wat}	a_i	a_j	γ^{dir}	γ^{prot}	γ^{wat}
GLN	GLN	-0.29	0.03	0.32	GLN	GLU	-0.49	-0.04	0.59
GLN	GLY	-0.37	0.01	0.11	GLN	HIS	-0.72	0.04	0.57
GLN	ILE	-0.43	-0.09	0.11	GLN	LEU	-0.29	-0.13	0.02
GLN	LYS	-0.49	-0.07	0.44	GLN	MET	-0.33	-0.13	-0.07
GLN	PHE	-0.35	0.04	-0.08	GLN	PRO	-0.60	0.01	0.46
GLN	SER	-0.34	-0.02	0.33	GLN	THR	-0.03	-0.03	0.37
GLN	TRP	-0.56	-0.06	-0.27	GLN	TYR	-0.21	-0.10	-0.69
GLN	VAL	-0.28	0.09	-0.02	GLU	GLU	-0.86	-0.04	0.38
GLU	GLY	-0.55	-0.01	0.09	GLU	HIS	-0.50	-0.05	0.40
GLU	ILE	-0.49	-0.11	0.22	GLU	LEU	-0.56	-0.26	0.13
GLU	LYS	0.13	-0.03	1.00	GLU	MET	-0.77	-0.23	0.22
GLU	PHE	-0.75	-0.16	-0.07	GLU	PRO	-0.78	-0.02	0.48
GLU	SER	-0.31	-0.01	0.18	GLU	THR	0.05	-0.01	0.14
GLU	TRP	-0.46	0.00	-0.29	GLU	TYR	-0.32	-0.04	-0.47
GLU	VAL	-0.38	-0.12	0.14	GLY	GLY	0.37	0.09	-0.08
GLY	HIS	-0.42	-0.03	0.29	GLY	ILE	0.04	-0.05	0.17
GLY	LEU	-0.22	-0.05	0.17	GLY	LYS	-0.48	-0.03	0.27
GLY	MET	0.13	0.21	0.05	GLY	PHE	-0.05	-0.08	0.32
GLY	PRO	-0.42	0.06	0.37	GLY	SER	0.02	0.03	0.14
GLY	THR	-0.14	0.02	0.18	GLY	TRP	0.04	-0.03	0.13
GLY	TYR	0.15	0.08	0.00	GLY	VAL	-0.11	0.05	0.20
HIS	HIS	-0.16	0.11	0.76	HIS	ILE	-0.30	-0.00	0.37
HIS	LEU	0.08	-0.00	-0.00	HIS	LYS	-0.55	-0.10	0.63
HIS	MET	0.20	0.09	-0.12	HIS	PHE	0.37	0.39	-0.11
HIS	PRO	-0.60	0.03	0.53	HIS	SER	-0.03	0.05	0.13
HIS	THR	-0.09	0.02	0.41	HIS	TRP	-0.01	0.48	-0.29
HIS	TYR	0.26	0.35	-0.28	HIS	VAL	0.16	0.03	0.03
ILE	ILE	0.98	1.00	1.00	ILE	LEU	0.98	1.00	0.38
ILE	LYS	-0.71	-0.09	0.20	ILE	MET	0.74	0.72	0.74
ILE	PHE	0.88	0.93	0.35	ILE	PRO	-0.43	-0.19	0.27
ILE	SER	-0.43	0.02	0.31	ILE	THR	-0.02	0.11	0.24
ILE	TRP	0.82	0.34	0.37	ILE	TYR	0.90	0.23	0.37
ILE	VAL	0.98	0.69	0.77	LEU	LEU	0.98	1.00	0.37
LEU	LYS	-0.66	-0.07	0.07	LEU	MET	0.85	0.74	0.27
LEU	PHE	0.79	0.70	0.25	LEU	PRO	-0.54	-0.15	0.11
LEU	SER	-0.34	-0.13	0.29	LEU	THR	0.01	0.13	0.26
LEU	TRP	0.98	0.38	0.76	LEU	TYR	0.69	0.35	0.32
LEU	VAL	0.98	0.64	0.43	LYS	LYS	-0.97	-0.06	0.42
LYS	MET	-0.70	-0.14	0.06	LYS	PHE	-0.66	-0.17	-0.26
LYS	PRO	-1.00	-0.03	0.55	LYS	SER	-0.62	-0.03	0.33
LYS	THR	-0.55	-0.03	0.47	LYS	TRP	-0.40	-0.21	-0.62
LYS	TYR	-0.18	-0.29	-0.58	LYS	VAL	-0.62	-0.13	0.03
MET	MET	0.52	0.32	-1.00	MET	PHE	0.69	0.72	0.30
MET	PRO	-0.50	0.01	0.13	MET	SER	-0.33	0.07	-0.03
MET	THR	-0.09	0.06	0.22	MET	TRP	0.12	0.50	-0.85
MET	TYR	0.64	0.27	-0.14	MET	VAL	0.63	0.40	0.62
PHE	PHE	0.98	1.00	0.52	PHE	PRO	-0.22	-0.21	0.26
PHE	SER	-0.27	0.01	0.13	PHE	THR	-0.16	0.12	0.16

Table A.8 – continue

a_i	a_j	γ^{dir}	γ^{prot}	γ^{wat}	a_i	a_j	γ^{dir}	γ^{prot}	γ^{wat}
PHE	TRP	0.67	0.66	0.54	PHE	TYR	0.62	0.27	-0.11
PHE	VAL	0.78	0.83	0.20	PRO	PRO	-0.51	-0.01	0.33
PRO	SER	-0.56	-0.00	0.52	PRO	THR	-0.47	-0.01	0.07
PRO	TRP	0.01	0.47	-0.56	PRO	TYR	0.06	-0.07	-0.34
PRO	VAL	-0.33	-0.10	0.21	SER	SER	-0.10	0.02	0.23
SER	THR	-0.10	-0.01	0.19	SER	TRP	-0.32	0.11	0.05
SER	TYR	-0.30	-0.03	-0.09	SER	VAL	-0.25	-0.00	0.10
THR	THR	0.16	-0.01	0.37	THR	TRP	-0.44	0.13	-0.13
THR	TYR	-0.22	-0.06	-0.37	THR	VAL	0.18	-0.10	0.19
TRP	TRP	0.07	0.43	-1.00	TRP	TYR	0.21	0.15	-0.95
TRP	VAL	0.52	0.44	1.00	TYR	TYR	0.55	0.21	-0.45
TYR	VAL	0.62	0.59	0.38	VAL	VAL	0.98	0.73	0.87

Chapter B: Supporting Information for Chapter 3

The phase diagram below shows the correlation between the binding mechanism and the structural properties of the dimers studied in this paper. A two-state dimer tends to have a hydrophobic interface and a large ratio of interfacial to monomeric contacts. On the contrary, a three-state dimer usually has a hydrophilic interface and a small ratio of interfacial to monomeric contacts.

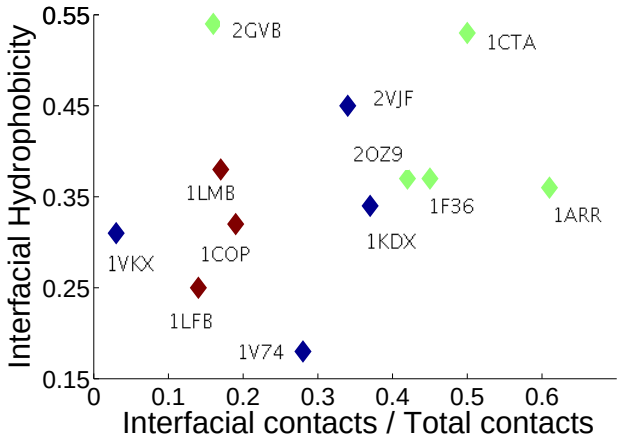


Figure B.1: A phase diagram that correlates the folding and binding mechanism of the dimers with their structural properties: interfacial hydrophobicity and the ratio of the number of interfacial contacts to the number of total contacts. Green and red colors are for two-state and three-state homodimers, respectively. The 4 heterodimers studied are plotted in blue. Typically, a two-state dimer has a highly hydrophobic interface and/or a large ratio of the number of interfacial contacts to the number of the total contacts. Three-state dimers usually have more hydrophilic interfaces and smaller ratios of interfacial to total contacts.

For Arc repressor with an intertwined dimer structure, the flexibility of the monomer modulates the binding efficiency.

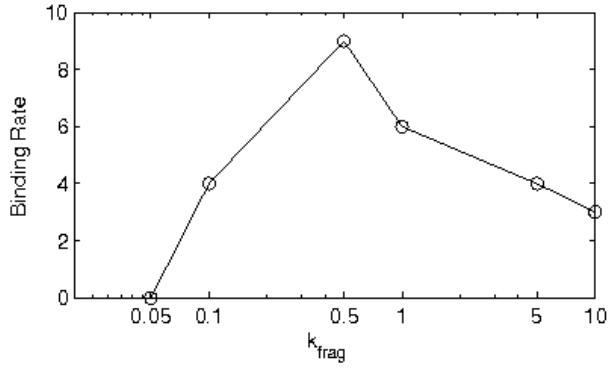


Figure B.2: The effect of the flexibility of the monomer structure on the successful rate of binding for Arc repressor (1ARR). k_{frag} is the interaction strength for the short range biasing potential acting on the monomers. Binding rate is calculated as the percentage of successful binding simulations for 40 independent runs. When k_{frag} is decreased, the binding rate first increases but finally decreases when the local bias is too weak.

The water-mediated interactions are very important in predicting dimer interfaces, especially for dimers with hydrophilic interfaces. The absence of the water-mediated interactions from the Hamiltonian greatly reduces the prediction quality, except for dimers with highly hydrophobic interfaces or having beta strands at the interface.

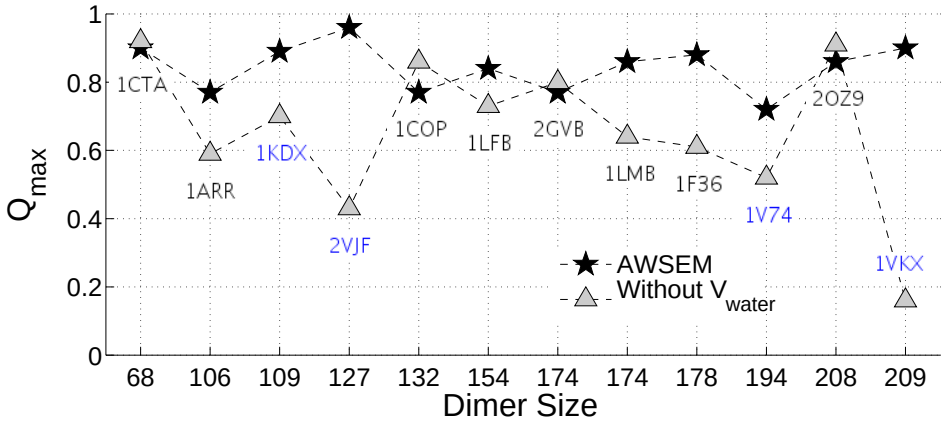


Figure B.3: The quality of interface prediction without the presence of mediated interactions V_{water} (see *Methods* section) is compared with the prediction quality of the standard AWSEM prediction. $Qmax$ is the Q of the best predicted structure in annealing simulations. The majority of the dimers, 8 out of 12, have worse predicted complex structures when the mediated interactions are turned off. The other 4 dimers, which are all homodimers and have comparable prediction quality, either have highly hydrophobic interfaces (2GVB, 1CTA and 2OZ9) or have beta strands at the interface to stabilize the dimer complex (1COP). The PDB id of heterodimers are in blue color.

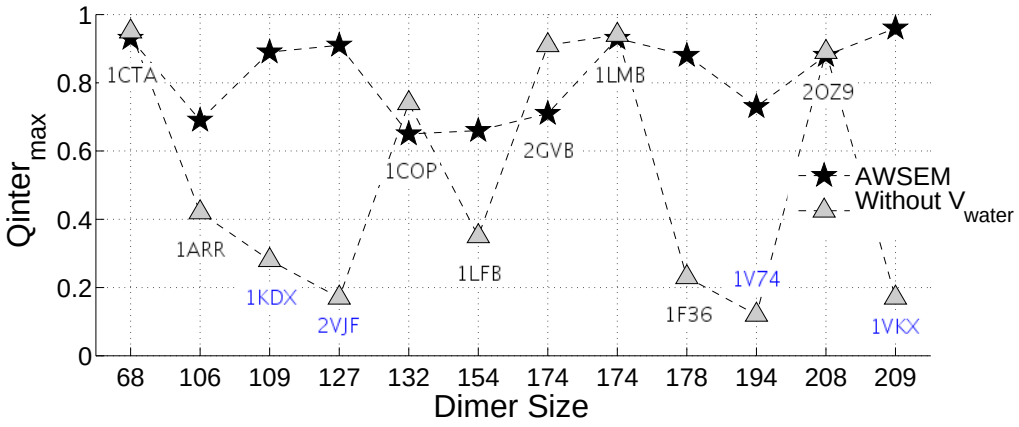


Figure B.4: Similar as Figure B.3, but the y axis is $Qinter_{max}$, the best predicted interface in annealing simulations, instead of Q_{max}. Note that the monomer structures are not necessarily well formed when Q of the interface is high.

Chapter C: Supporting Information for Chapter 4

C.1 The Theory of Correlated Random Noise

For correlated random noise $\epsilon_{eff} = \epsilon + \delta\epsilon$, where $\delta\epsilon$ is a solution of the following stochastic equation:

$$\frac{d\delta\epsilon}{dt} = -\xi\delta\epsilon + \omega^2\eta(t). \tag{C.1}$$

Here $\eta(t)$ is Gaussian white noise, and ξ and ω determine properties of the random process.

We can write the solution of the Equation C.1 in the following recursive form:

$$\begin{aligned} \delta\epsilon(t+dt) &= \delta\epsilon(t) - 0.5\xi(\delta\epsilon(t) + \delta\epsilon^*(t))dt + \omega\sqrt{2dt}\eta(t) \\ \delta\epsilon^*(t) &= \delta\epsilon(t) - \xi\delta\epsilon(t) + \omega\ \sqrt{2dt}\eta(t), \end{aligned} \tag{C.2}$$

or in more simplified manner

$$\delta\epsilon(t+dt) = \delta\epsilon(t) + (1 - 0.5\xi\delta\epsilon(t)dt)(\omega\sqrt{2dt}\eta(t) - \xi dt). \tag{C.3}$$

The autocorrelation function of the solution C.3 equals to

$$G(t') \equiv \langle\delta\epsilon(t)\delta\epsilon(t+t')\rangle = \frac{\omega^2}{\xi}\exp(-\xi t'). \tag{C.4}$$

Thus the correlation time and the mean-square deviation will be

$$\tau_{corr} = \frac{1}{\xi}, \qquad \langle \delta\epsilon^2 \rangle = \frac{\omega^2}{\xi}. \tag{C.5}$$

To find the spectrum of the noise we can use the following formula:

$$S(\Omega) = \frac{1}{\pi} \int_0^\infty \cos(\Omega\, t') G(t') dt'. \tag{C.6}$$

After substituting $G(t')$ we get

$$S(\Omega) = \frac{1}{\pi} \frac{\omega^2}{\xi} \int_0^\infty \cos(\Omega\, t') e^{-\xi t'} dt' = \frac{1}{\pi} \frac{\omega^2}{\xi^2 + \Omega^2}. \tag{C.7}$$

C.2 Connection Between Fluctuations in Potential Strength and Temperature

To connect fluctuations in contact potential strength and temperature, we calculated the free energy profiles for 1SRL vs. the ratio of native contacts Q for a variety of ϵ and T values (see a) and b) of Figure C.1). Then, we calculated the transition energy $\Delta G = F(Q_{folded}) - F(Q_{unfolded})$ for each case. For the constant temperature of 300K we found:

$\Delta G_\epsilon(0.54)$ = 0.70

$\Delta G_\epsilon(0.55)$ = 0.00

$\Delta G_\epsilon(0.56)$ = -0.74

$\Delta G_\epsilon(0.57)$ = -1.46

$\Delta G_\epsilon(0.58)$ = -2.14

For the constant ϵ=0.56kcal/mol, we correspondingly found:

$$\begin{aligned}
\Delta G_T(290) &= \text{-2.64} \\
\Delta G_T(295) &= \text{-1.66} \\
\Delta G_T(300) &= \text{-0.74} \\
\Delta G_T(305) &= \text{-0.22} \\
\Delta G_T(310) &= \text{ 1.28}
\end{aligned}$$

For both data sets above we did linear fits obtaining $\Delta G(\epsilon)$ and $\Delta G(T)$ relationships (See c) and d) of Figure C.1).

$$\begin{aligned}
\Delta G(\epsilon) &= 39.26 - 71.40\epsilon \\
\Delta G(T) &= -59.01 + 0.194T
\end{aligned} \tag{C.8}$$

Now, if we take the derivative of the first equation by ϵ and the second equation by T, we can write:

$$\left(\frac{\partial \Delta G}{\partial \epsilon}\right) d\epsilon = \left(\frac{\partial \Delta G}{\partial T}\right) dT. \tag{C.9}$$

Using expressions C.8 and C.9 we get $\frac{dT}{d\epsilon} = -367.42$. Thus, we can estimate the deviation in temperature ΔT corresponding to the deviation in potential strength $\Delta\epsilon$ by the following expression:

$$\Delta T = -367.42\ \Delta\epsilon. \tag{C.10}$$

In our simulations for PGK we applied the harmonic fluctuations with standard deviations 0.03ϵ, 0.05ϵ and 0.07ϵ. To get the amplitudes we need to multiply the standard deviations by $\sqrt{2.0}$. Then, we can get the corresponding temperature amplitudes.

$$\Delta T = -367.42 \times 0.58 \times 0.07 \times \sqrt{2.0} = -21.10K$$

$$\Delta T = -367.42 \times 0.58 \times 0.05 \times \sqrt{2.0} = -15.07K$$

$$\Delta T = -367.42 \times 0.58 \times 0.03 \times \sqrt{2.0} = -9.04K$$

Likewise, we can get that the temperature fluctuation amplitudes 1 and 2 °C, used in the experiments, correspond to the standard deviation of potential strengths of 0.0033ϵ and 0.0066ϵ, respectively.

C.3 Experimental vs. MD Results. Quantitative Comparison

To compare our computational results to the results obtained from the experiments, we need to estimate the value of $\ln\left[\frac{t_f^{min}(\sqrt{\langle\delta\epsilon^2\rangle}/\epsilon=0.0066)}{t_f^{min}(\sqrt{\langle\delta\epsilon^2\rangle}/\epsilon=0.0033)}\right]$. From the computational studies we know the $t_f(\theta)$ dependence for $\frac{\sqrt{\langle\delta\epsilon^2\rangle}}{\epsilon}$=0.03, 0.05, 0.07. In Figure C.2 we fitted $log_{10}(t_f)$ vs. $log_{10}(\theta)$ data with polynomial functions and found the minimums for each curve (denoted by a star). The obtained $\log_{10}(t_f^{min})$ values and the linear fit of $\log_{10}(t_f^{min})$ vs. $\frac{\sqrt{\langle\delta\epsilon^2\rangle}}{\epsilon}$ dependence is plotted in Figure C.3. Using it, we can find:

$$\ln\left[\frac{t_f^{min}(\sqrt{\langle\delta\epsilon^2\rangle}/\epsilon=0.0066)}{t_f^{min}(\sqrt{\langle\delta\epsilon^2\rangle}/\epsilon=0.0033)}\right] = 0.085. \tag{C.11}$$

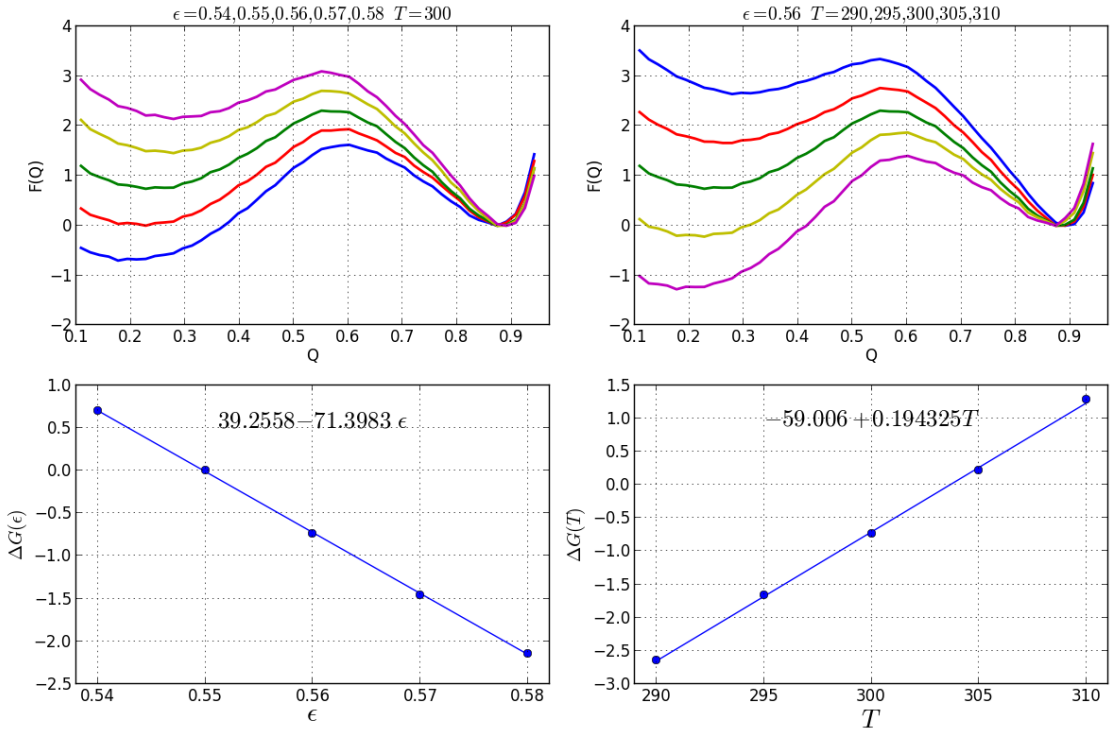


Figure C.1: a) Free energy plot for different ϵ values and fixed temperature b) Free energy plot for fixed ϵ value and different temperatures c) $\Delta G(\epsilon)$ plot. T=300K. d) $\Delta G(T)$ plot. ϵ=0.56kcal/mol

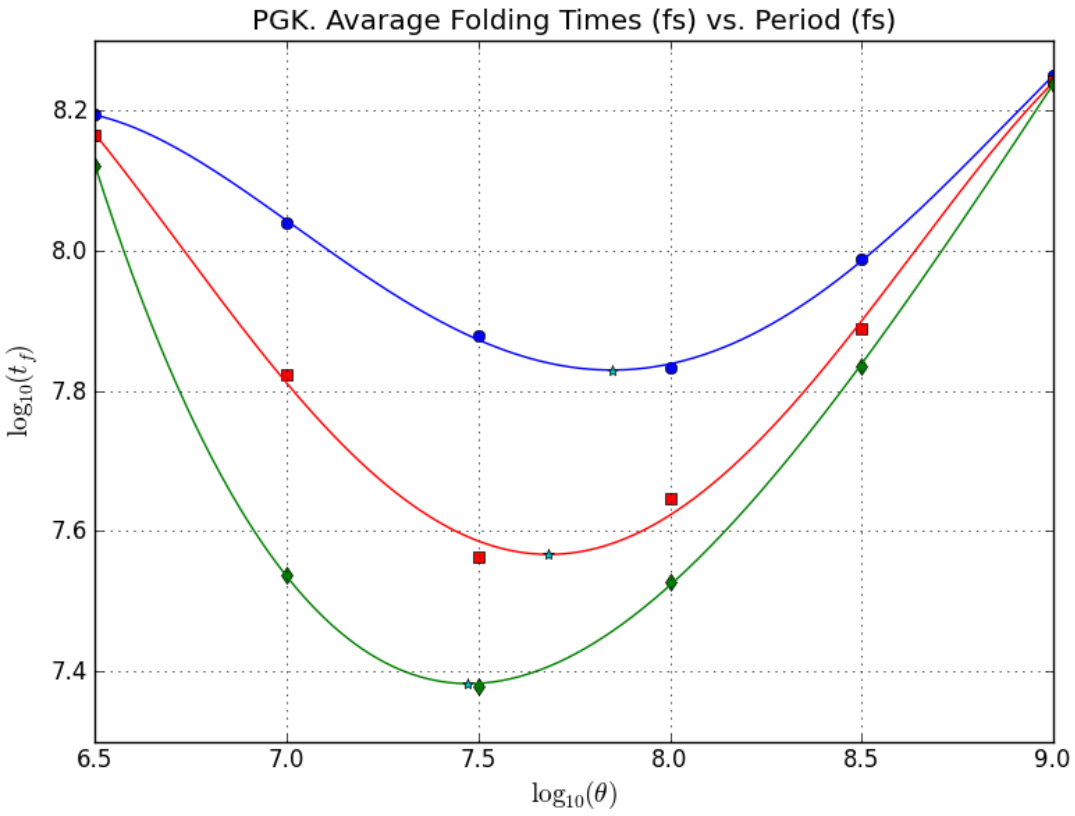


Figure C.2: The plot shows the polynomial fits of average first-passage time vs. period dependencies for PGK for $\frac{\sqrt{\langle \delta\epsilon^2 \rangle}}{\epsilon}$=0.03, 0.05, 0.07. The minimums are shown by stars.

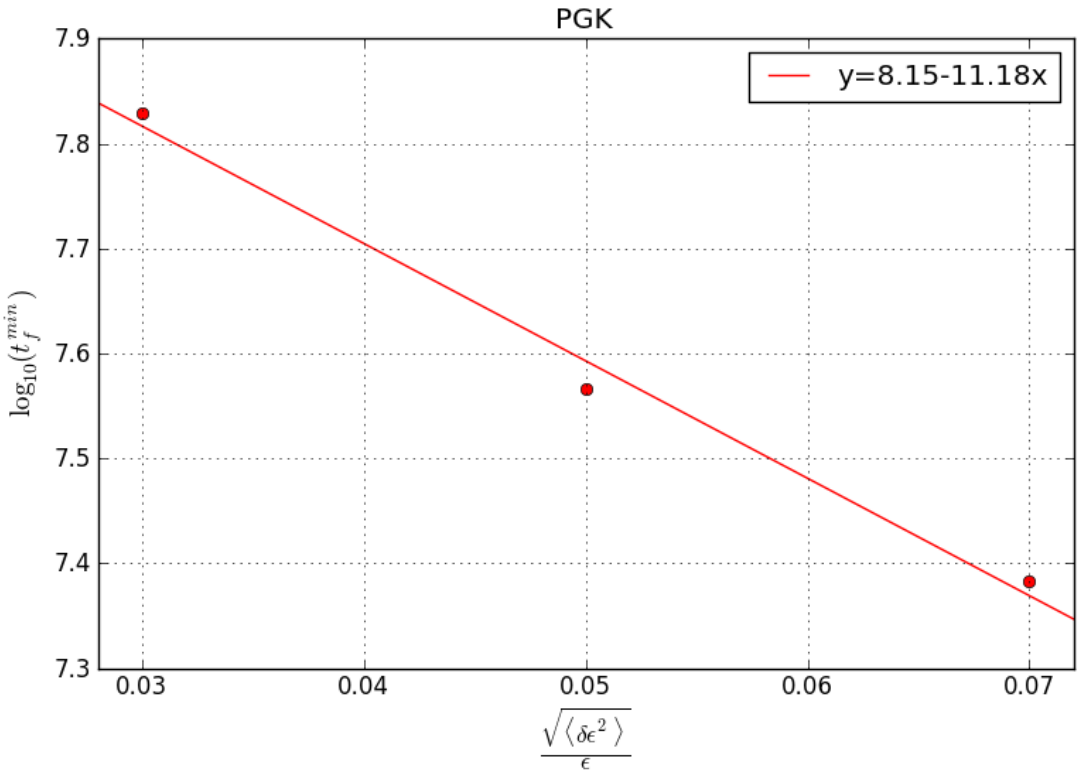


Figure C.3: Plot of $\log_{10}(t_f^{min})$ vs. $\frac{\sqrt{\langle \delta\epsilon^2 \rangle}}{\epsilon}$ and the linear fit.

Bibliography

[1] C.B. Anfinsen, "The formation and stabilization of protein structure", Biochem. J. **128**(4), 737–749 (1972).

[2] C.B. Anfinsen, "Principles that Govern the Folding of Protein Chains", Science **181**(4096), 223–230 (1973).

[3] J. Kendrew, G. Bodo, H. Dintzis, R. Parrish, H. Wyckoff, D. Phillips, "A three-dimensional model of the myoglobin molecule obtained by X-ray analysis", Nature **181**(4610), 662–666 (1958).

[4] C. Levinthal, "Are there pathways for protein folding?", J. Chim. Phys. **65**, 44–45 (1968).

[5] J.D. Bryngelson, P.G. Wolynes, "Spin glasses and the statistical mechanics of protein folding", Proc. Natl. Acad. Sci. USA **84**(21), 7524–7528 (1987).

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

[7] J.N. Onuchic, P.G. Wolynes, Z. Luthey-Schulten, N.D. Socci, "Toward an outline of the topography of a realistic protein-folding funnel", Proc. Natl. Acad. Sci. USA **92**(8), 3626–3630 (1995).

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

[9] S.S. Plotkin, J.N. Onuchic, "Understanding protein folding with energy landscape theory. Part II: Quantitative aspects", Q. Rev. Biophys. **35**(3), 205–286 (2002).

[10] M. Ohgushi, A. Wada, "'Molten-globule state': a compact form of globular proteins with mobile side-chains", FEBS Lett. **164**(1), 21–24 (1983).

[11] P.I. Zhuravlev, G.A. Papoian, "Functional versus folding landscapes: the same yet different", Curr. Opin. Struct. Biol. **20**(1), 16–22 (2010).

[12] P.I. Zhuravlev, C.K. Materese, G.A. Papoian, "Deconstructing the Native State: Energy Landscapes, Function, and Dynamics of Globular Proteins", J. Phys. Chem. B **113**(26), 8800–8812 (2009).

[13] P.W. Fenimore, H. Frauenfelder, B.H. Mcmahon, R.D. Young, "Bulk-solvent and hydration-shell uctuations, similar to alpha- and beta-uctuations in glasses, control protein motions and functions", Proc. Natl. Acad. Sci. USA **101**(40), 14408–14413 (2004).

[14] H. Frauenfelder, B. McMahon, "Dynamics and function of proteins: The search for general concepts", Proc. Natl. Acad. Sci. USA **95**(9), 4795–4797 (1998).

[15] J.D. Honeycutt, D. Thirumalai, "The nature of folded states of globular proteins", Biopolymers **32**(6), 695–709 (1992).

[16] B.F. Manly, *Randomization, bootstrap and Monte Carlo methods in biology* (Chapman & Hall/CRC, Laramie, WY, 2006).

[17] B.J. Alder, T.E. Wainwright, "Studies in Molecular Dynamics. I. General Method", J. Chem. Phys. **31**(2), 459 (1959).

[18] A. Rahman, "Correlations in the Motion of Atoms in Liquid Argon", Phys. Rev. **136**(2A), A405–A411 (1964).

[19] J.A. McCammon, B.R. Gelin, M. Karplus, "Dynamics of folded proteins", Nature **267**, 585–590 (1977).

[20] D.E. Shaw, R.O. Dror, J.K. Salmon, J.P. Grossman, K.M. Mackenzie, J.A. Bank, C. Young, M.M. Deneroff, B. Batson, K.J. Bowers, E. Chow, M.P. Eastwood, D.J. Ierardi, J.L. Klepeis, J.S. Kuskin, R.H. Larson, K. Lindorff-Larsen, P. Maragakis, M.A. Moraes, S. Piana, Y. Shan, B. Towles, *Millisecond-scale molecular dynamics simulations on Anton* (Proceedings of the Conference on High Performance Computing, Networking, Storage and Analysis (SC09), ACMPress, New York, 2009).

[21] D.E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R.O. Dror, M.P. Eastwood, J.A. Bank, J.M. Jumper, J.K. Salmon, Y. Shan, W. Wriggers,

"Atomic-level characterization of the structural dynamics of proteins", Science **330**(6002), 341–346 (2010).

[22] M.O. Jensen, V. Jogini, D.W. Borhani, A.E. Leffler, R.O. Dror, D.E. Shaw, "Mechanism of Voltage Gating in Potassium Channels", Science **336**(6078), 229–233 (2012).

[23] K. Lindorff-Larsen, S. Piana, R.O. Dror, D.E. Shaw, "How Fast-Folding Proteins Fold", Science **334**(6055), 517–520 (2011).

[24] J.C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R.D. Skeel, L. Kale, K. Schulten, "Scalable molecular dynamics with NAMD", J. Comput. Chem. **26**, 1781–1802 (2005).

[25] B. Hess, C. Kutzner, D. van der Spoel, E. Lindahl, "GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation", J. Chem. Theory Comput. **4**(3), 435–447 (2008).

[26] J. Wang, W. Wang, P.A. Kollman, D.A. Case, "Automatic atom type and bond type perception in molecular mechanical calculations", J. Mol. Graph. Model. **25**(2), 247–260 (2006).

[27] S. Plimpton, "Fast Parallel Algorithms for Short-Range Molecular Dynamics", J. Comput. Phys. **117**(1), 1–19 (1995).

[28] W.M. Brown, P. Wang, S.J. Plimpton, A.N. Tharrington, "Implementing Molecular Dynamics on Hybrid High Performance Computers – Short Range Forces", Comput. Phys. Commun. **182**(4), 898–911 (2011).

[29] W.M. Brown, A. Kohlmeyer, S.J. Plimpton, A.N. Tharrington, "Implementing Molecular Dynamics on Hybrid High Performance Computers - Particle-Particle Particle-Mesh", Comput. Phys. Commun. **183**(3), 449–459 (2012).

[30] A. Zhmurov, A.E.X. Brown, R.I. Litvinov, R.I. Dima, J.W. Weisel, V. Barsegov, "Mechanism of Fibrin(ogen) Forced Unfolding", Structure **19**(11), 1615–1624 (2011).

[31] A. Zhmurov, R.I. Dima, Y. Kholodov, V. Barsegov, "Sop-GPU: accelerating biomolecular simulations in the centisecond timescale using graphics processors", Proteins **78**(14), 2984–2999 (2010).

[32] D.N. LeBard, B.G. Levine, P. Mertmann, S.A. Barr, A. Jusufi, S. Sanders, M.L. Klein, A.Z. Panagiotopoulos, "Self-assembly of coarse-grained ionic sur-

factants accelerated by graphics processing units", Soft Matter **8**, 2385–2397 (2012).

[33] M.J. Harvey, G. De Fabritiis, "An Implementation of the Smooth Particle Mesh Ewald Method on GPU Hardware", J. Chem. Theory Comput. **5**(9), 2371–2377 (2009).

[34] C.A. Finlay, P.W. Hinds, A.J. Levine, "The p53 proto-oncogene can act as a suppressor of transformation", Cell **57**, 1083–1093 (1989).

[35] G.M. Wulf, Y.C. Liou, A. Ryo, S.W. Lee, K.P. Lu, "Role of Pin1 in the regulation of p53 stability and p21 transactivation, and cell cycle checkpoints in response to DNA damage", J. Biol. Chem. **277**(50), 47976–47979 (2002).

[36] S. Xie, H. Wu, Q. Wang, J.P. Cogswell, I. Husain, C. Conn, P. Stambrook, M. Jhanwar-Uniyal, W. Dai, "Plk3 functionally links DNA damage to cell cycle arrest and apoptosis at least in part via the p53 pathway", J. Biol. Chem. **276**(46), 43305–43312 (2001).

[37] E.S. Han, F.L. Muller, V.I. Pérez, W. Qi, H. Liang, L. Xi, C. Fu, E. Doyle, M. Hickey, J. Cornell, C.J. Epstein, L.J. Roberts, H. Van Remmen, A. Richardson, "The in vivo Gene Expression Signature of Oxidative Stress", Physiol. Genomics. **34**(1), 112–26 (2008).

[38] M. Hollstein, D. Sidransky, B. Vogelstein, C.C. Harris, "p53 mutations in human cancers", Science **253**(5015), 49–53 (1991).

[39] Y. Ueeda, H. Taketomi, N. Gō, "Studies on protein folding, unfolding and fluctuations by computer simulation. A three-dimensional lattice model of lysozyme", Biopolymers **17**(6), 1531–1548 (1978).

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

[41] N. Koga, S. Takada, "Roles of native topology and chain-length scaling in protein folding: a simulation study with a Gō-like model", J. Mol. Biol. **313**(1), 171–180 (2001).

[42] C. Clementi, H. Nymeyer, J.N. Onuchic, "Topological and energetic factors: what determines the structural details of the transition state ensemble and enroute intermediates for protein folding? An investigation for small globular proteins", J. Mol. Biol. **298**(5), 937–953 (2000).

[43] C. Clementi, A.E. Garcia, J.N. Onuchic, "Interplay among tertiary contacts, secondary structure formation and side chain packing in the protein folding mechanism: all-atom representation study of protein L", J. Mol. Biol. **326**(3), 933–954 (2003).

[44] I.A. Hubner, M. Oliveberg, E.I. Shakhnovich, "Simulation experiment and evolution: understanding nucleation in protein S6 folding", Proc. Natl. Acad. Sci. USA **101**(22), 8354–8359 (2004).

[45] M.S. Cheung, A.E. Garcia, J. Onuchic, "Protein folding mediated by solvation: water expulsion and formation of the hydrophobic core occur after structural collapse", Proc. Natl. Acad. Sci. USA **99**(2), 685–690 (2000).

[46] H. Kaya, H.S. Chan, "Solvation effects and driving forces for protein thermodynamics and kinetic cooperativity: how adequate is native-centric topological modeling?", J. Mol. Biol. **326**(3), 911–931 (2003).

[47] Y. Levy, P.G. Wolynes, J.N. Onuchic, "Protein topology determines binding mechanism", Proc. Natl. Acad. Sci. USA **101**(2), 511–516 (2004).

[48] F. Takagi, N. Koga, S. Takada, "How protein thermodynamics and folding mechanisms are altered by the chaperoning cage: molecular simulations", Proc. Natl. Acad. Sci. USA **100**(20), 11367–11372 (2003).

[49] S. Brown, N.J. Fawzi, T. Head-Gordon, "Coarse grained sequences for protein folding and design", Proc. Natl. Acad. Sci. USA **100**(19), 10712–10717 (2003).

[50] M. Friedel, J.E. Shea, "Self-assembly of peptides into a beta-barrel motif", J. Chem. Phys. **120**(12), 5809–5823 (2004).

[51] V. Tozzini, J. Trylska, C.E. Chang, A. McCammon, "Flap opening dynamics in HIV-1 protease explored with a coarse-grained model", J. Struct. Biol. **157**(3), 606–615 (2006).

[52] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, "The Protein Data Bank", Nucleic Acids Res. **28**(1), 235–242 (2000).

[53] M. Levitt, "A simplified representation of protein conformations for rapid simulation of protein folding", J. Mol. Biol. **104**(1), 59–107 (1976).

[54] K. Koretke, Z. Luthey-Schulten, P. Wolynes, "Self-consistently optimized statistical mechanical energy functions for sequence structure alignment", Protein Sci. **5**(6), 1043–1059 (1996).

[55] K.K. Koretke, Z. Luthey-Schulten, P.G. Wolynes, "Self-consistently optimized energy functions for protein structure prediction by molecular dynamics", Proc. Natl. Acad. Sci. USA **95**(6), 2932–2937 (1998).

[56] M.R. Betancourt, S.J. Omovie, "Pairwise energies for polypeptide coarse-grained models derived from atomic force fields", J. Chem. Phys. **130**(19), 195103 (2009).

[57] A.V. Smith, C.K. Hall, "α-helix formation: discontinuous molecular dynamics on an intermediate-resolution protein model", Proteins **44**(3), 344–360 (2001).

[58] A. Colubri, "Prediction of protein structure by simulating coarsegrained folding pathways: a preliminary report", J. Biomol. Struct. Dyn. **21**(5), 625–637 (2004).

[59] S. Takada, Z. Luthey-Schulten, P.G. Wolynes, "Folding dynamics with non-additive forces: a simulation study of a designed helical protein and random heteropolymer", J. Chem. Phys. **110**(23), 11616–11628 (1999).

[60] F. Forcellino, P. Derremaux, "Computer simulations aimed at structure prediction of supersecondary motifs in proteins", Proteins **45**(2), 159–166 (2001).

[61] M. Roca, B. Messer, D. Hilvert, A. Warshel, "On the relationship between folding and chemical landscapes in enzyme catalysis", Proc. Natl. Acad. Sci. USA **105**(37), 13877–13882 (2008).

[62] A. Davtyan, N.P. Schafer, W. Zheng, C. Clementi, P.G. Wolynes, G.A. Papoian, "AWSEM-MD: Protein Structure Prediction Using Coarse-Grained Physical Potentials and Bioinformatically Based Local Structure Biasing", J. Phys. Chem. B **116**(29), 8494–8503 (2012).

[63] M.S. Friedrichs, P.G. Wolynes, "Toward protein tertiary structure recognition by means of associative memory hamiltonians", Science **246**(4928), 371–373 (1989).

[64] M. Sasai, P. Wolynes, "Molecular theory of associative memory hamiltonian models of protein folding", Phys. Rev. Lett. **65**(21), 2740–2743 (1990).

[65] M.S. Friedrichs, P.G. Wolynes, "Molecular dynamics of associative memory hamiltonians for protein tertiary structure recognition", Tetrahedron Comput. Methodol. **3**(3-4), 175–190 (1990).

[66] M.S. Friedrichs, R.A. Goldstein, P.G. Wolynes, "Generalized protein tertiary structure recognition using associative memory hamiltonians* 1", J. Mol. Biol. **222**(4), 1013–1034 (1991).

[67] M. Sasai, P. Wolynes, "Unified theory of collapse, folding, and glass transitions in associative-memory Hamiltonian models of proteins", Phys. Rev. A **46**(12), 7979–7997 (1992).

[68] C. Hardin, M.P. Eastwood, Z. Luthey-Schulten, P.G. Wolynes, "Associative memory Hamiltonians for structure prediction without homology: alpha-helical proteins", Proc. Natl. Acad. Sci. USA **97**(26), 14235–14240 (2000).

[69] C. Hardin, M.P. Eastwood, M.C. Prentiss, Z. Luthey-Schulten, P.G. Wolynes, "Associative memory Hamiltonians for structure prediction without homology: α/β proteins", Proc. Natl. Acad. Sci. USA **100**(4), 1679–1684 (2003).

[70] V. Okeljas, C. Zong, G.A. Papoian, P.G. Wolynes, "Protein structure prediction: Do hydrogen bonding and water-mediated interactions suffice?", Methods **52**(1), 84–90 (2010).

[71] W. Zheng, N. Schafer, A. Davtyan, G. A. Papoian, P.G. Wolynes, "Predictive Energy Landscapes for Protein-Protein Association", Proc. Natl. Acad. Sci. USA **109**(47), 19244–19429 (2012).

[72] W. Zheng, N. Schafer, P.G. Wolynes, "Frustration in the energy landscapes of multidomain protein misfolding", Proc. Natl. Acad. Sci. USA **110**(5), 1680–1685 (2013).

[73] R.F. Service, "Problem solved* (*sort of)", Science **321**(5890), 784–786 (2008).

[74] A. Fersht, *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding* (W.H. Freeman and Co., New York, 1999).

[75] L. Pauling, R.B. Corey, H.R. Branson, "The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain", Proc. Natl. Acad. Sci. USA **37**(4), 205–211 (1951).

[76] W. Kauzmann, "Some factors in the interpretation of protein denaturation", Adv. Protein Chem. **14**, 1–63 (1959).

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

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

[79] C. Zong, G.A. Papoian, J. Ulander, P.G. Wolynes, "Role of topology, nonadditivity, and water-mediated interactions in predicting the structures of α/β proteins", J. Am. Chem. Soc. **128**(15), 5168–5176 (2006).

[80] J.D. Bryngelson, J.N. Onuchic, N.D. Socci, P.G. Wolynes, "Funnels, pathways, and the energy landscape of protein folding: A synthesis", Proteins **21**(3), 167–195 (1995).

[81] P.G. Wolynes, "Energy landscapes and solved protein–folding problems", Philos. Trans. R. Soc. Lond. A **363**(1827), 453–467 (2005).

[82] R.A. Goldstein, Z.A. Luthey-Schulten, P.G. Wolynes, "Optimal protein-folding codes from spin-glass theory", Proc. Natl. Acad. Sci. USA **89**(11), 4918–4922 (1992).

[83] R.A. Goldstein, Z.A. Luthey-Schulten, P.G. Wolynes, "Protein tertiary structure recognition using optimized Hamiltonians with local interactions", Proc. Natl. Acad. Sci. USA **89**(19), 9029–9033 (1992).

[84] M.P. Eastwood, C. Hardin, Z.A. Luthey-Schulten, P.G. Wolynes, "Evaluating protein structure-prediction schemes using energy landscape theory", IBM J. Res. Dev. **45**(3.4), 475–497 (2001).

[85] J.J. Hopfield, "Neurons with graded response have collective computational properties like those of two-state neurons", Proc. Natl. Acad. Sci. USA **81**(10), 3088–3092 (1984).

[86] B. Kuhlman, G. Dantas, G.C. Ireton, G. Varani, B.L. Stoddard, D. Baker, "Design of a novel globular protein fold with atomic-level accuracy", Science **302**(5649), 1364–1368 (2003).

[87] W.G. Noid, J. Chu, G.S. Ayton, V. Krishna, S. Izvekov, G.A. Voth, A. Das, H.C. Andersen, "The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models", J. Chem. Phys. **128**(24), 244114 (2008).

[88] A. Savelyev, G.A. Papoian, "Molecular Renormalization Group Coarse-Graining of Polymer Chains: Application to Double-Stranded DNA", Biophys. J. **96**(10), 4044–4052 (2009).

[89] A. Savelyev, G.A. Papoian, "Molecular Renormalization Group Coarse-Graining of Electrolyte Solutions: Application to Aqueous NaCl and KCl", J. Phys. Chem. B **113**(22), 7785–7793 (2009).

[90] G.G. Maisuradze, P. Senet, C. Czaplewski, A. Liwo, H.A. Scheraga, "Investigation of protein folding by coarse-grained molecular dynamics with the UNRES force field", J Phys Chem A **114**(13), 4471–4485 (2010).

[91] J.G. Saven, P.G. Wolynes, "Local conformational signals and the statistical thermodynamics of collapsed helical proteins", J. Mol. Biol. **257**(1), 199–216 (1996).

[92] J.A. Hegler, J. Lätzer, A. Shehu, C. Clementi, P.G. Wolynes, "Restriction versus guidance in protein structure prediction", Proc. Natl. Acad. Sci. USA **106**(36), 15302–15307 (2009).

[93] A. Savelyev, G.A. Papoian, "Chemically accurate coarse graining of double-stranded DNA", Proc. Natl. Acad. Sci. USA **107**(47), 20340–20345 (2010).

[94] C. Hyeon, G. Morrison, D.L. Pincus, D. Thirumalai, "Refolding dynamics of stretched biopolymers upon force quench", Proc. Natl. Acad. Sci. USA **106**(48), 20288–20293 (2009).

[95] M.A. Marti-Renom, A. Stuart, A. Fiser, R. Sánchez, F. Melo, A. Sali, "Comparative protein structure modeling of genes and genomes", Annu. Rev. Biophys. Biomol. Struct. **29**, 291–325 (2000).

[96] A. Sali, T.L. Blundell, "Comparative protein modelling by satisfaction of spatial restraints", J. Mol. Biol. **234**(3), 779–815 (1993).

[97] A. Fiser, R.K. Do, A. Sali, "Modeling of loops in protein structures", Protein Sci. **9**(9), 1753–1773 (2000).

[98] G. Wang, R.L. Dunbrack, "PISCES: a protein sequence culling server", Bioinformatics **19**(12), 1589–1591 (2003).

[99] S.L. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, A. Zhang, W. Miller, D.J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. **25**(17), 3389–3402 (1997).

[100] I.N. Shindyalov, P.E. Bourne, "Protein structure alignment by incremental combinatorial extension (CE) of the optimal path", Protein Eng. **11**(9), 739–747 (1998).

[101] K. Kwac, P.G. Wolynes, "Protein Structure Prediction Using an Associated Memory Hamiltonian and All-Atom Molecular Dynamics Simulations", Bull. Korean Chem. Soc. **29**(11), 2172–2182 (2008).

[102] Y. Levy, G.A. Papoian, J.N. Onuchic, P.G. Wolynes, "Energy Landscape Analysis of Protein Dimers", Isr. J. Chem. **44**(1-3), 281–297 (2004).

[103] Y. Levy, S.S. Cho, J.N. Onuchic, P.G. Wolynes, "A Survey of Flexible Protein Binding Mechanisms and their Transition States Using Native Topology Based Energy Landscapes", J. Mol. Biol. **346**(4), 1121–1145 (2005).

[104] Q. Lu, H.P. Lu, J. Wang, "Exploring the Mechanism of Flexible Biomolecular Recognition with Single Molecule Dynamics", Phys. Rev. Lett. **98**(12), 128105 (2007).

[105] J. Wang, Y. Wang, X. Chu, S.J. Hagen, W. Han, E. Wang, "Multi-Scaled Explorations of Binding-Induced Folding of Intrinsically Disordered Protein Inhibitor IA3 to its Target Enzyme", PLoS Comput. Biol. **7**(4), e1001118 (2011).

[106] G.A. Papoian, P.G. Wolynes, "The physics and bioinformatics of binding and folding–An energy landscape perspective", Biopolymers **68**(3), 333–349 (2003).

[107] S.J. de Vries, A.D.J. van Dijk, M. Krzeminski, M. van Dijk, A. Thureau, V. Hsu, T. Wassenaar, A.M.J.J. Bonvin, "HADDOCK versus HADDOCK: New features and performance of HADDOCK2.0 on the CAPRI targets", Proteins **69**(4), 726–733 (2007).

[108] B. Pierce, Z. Weng, "A combination of rescoring and refinement significantly improves protein docking performance", Proteins **72**(1), 270–279 (2008).

[109] C. Wang, P. Bradley, D. Baker, "Protein-Protein Docking with Backbone Flexibility", J. Mol. Biol. **373**(2), 503–519 (2007).

[110] S. Vajda, D. Kozakov, "Convergence and combination of methods in protein-protein docking", Curr. Opin. Struct. Biol. **19**(2), 164–170 (2009).

[111] J. Fernandez-Recio, M. Totrov, R. Abagyan, "ICM-DISCO docking by global energy optimization with fully flexible side-chains", Proteins **52**(1), 113–117 (2003).

[112] D. Kozakov, O. Schueler-Furman, S. Vajda, "Discrimination of near-native structures in protein-protein docking by testing the stability of local minima", Proteins **72**(3), 993–1004 (2008).

[113] E. Mashiach, D. Schneidman-Duhovny, N. Andrusier, R. Nussinov, H.J. Wolfson, "FireDock: a web server for fast interaction refinement in molecular docking", Nucleic. Acids. Res. **36**(Web Server), W229–W232 (2008).

[114] P.G. Wolynes, "Symmetry and the energy landscapes of biomolecules", Proc. Natl. Acad. Sci. USA **93**(25), 14249–14255 (1996).

[115] I. Andre, C.E.M. Strauss, D.B. Kaplan, P. Bradley, D. Baker, "Emergence of symmetry in homooligomeric biological assemblies", Proc. Natl. Acad. Sci. USA **105**(42), 16148–16152 (2008).

[116] T.A. Ceska, M. Lamers, P. Monaci, A. Nicosia, R. Cortese, D. Suck, "The X-ray structure of an atypical homeodomain present in the rat liver transcription factor LFB1/HNF1 and implications for DNA binding", EMBO J. **12**(5), 1805 (1993).

[117] M. Frain, G. Swart, P. Monaci, A. Nicosia, S. Stämpfli, R. Frank, R. Cortese, et al., "The liver-specific transcription factor LF-B1 contains a highly diverged homeobox DNA binding domain", Cell **59**(1), 145 (1989).

[118] L. Tomei, R. Cortese, R. De Francesco, "A POU-A related region dictates DNA binding specificity of LFB1/HNF1 by orienting the two XL-homeodomains in the dimer", EMBO J. **11**(11), 4119 (1992).

[119] D.E. Koshland Jr, "Application of a theory of enzyme specificity to protein synthesis", Proc. Natl. Acad. Sci. USA **44**(2), 98 (1958).

[120] P. Csermely, R. Palotai, R. Nussinov, "Induced fit, conformational selection and independent dynamic segments: an extended view of binding events", Trends Biochem. Sci. **35**(10), 539–546 (2010).

[121] E. Fischer, "Einfluss der Configuration auf die Wirkung der Enzyme", Ber. Dtsch. Chem. Ges. **27**(3), 2985–2993 (1894).

[122] J. Wang, L. Xu, E. Wang, "Optimal Specificity and Function for Flexible Biomolecular Recognition", Biophys. J. **92**(12), L109–L111 (2007).

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

[124] M.J. Bennett, M.R. Sawaya, D. Eisenberg, "Deposition Diseases and 3D Domain Swapping", Structure **14**(5), 811–824 (2006).

[125] B.A. Shoemaker, J.J. Portman, P.G. Wolynes, "Speeding molecular recognition by using the folding funnel: The fly-casting mechanism", Proc. Natl. Acad. Sci. USA **97**(16), 8868–8873 (2000).

[126] M. Oliveberg, "Alternative Explanations for Multistate Kinetics in Protein Folding: Transient Aggregation and Changing Transition-State Ensembles", Acc. Chem. Res. **31**(11), 765–772 (1998).

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

[128] P.G. Wolynes, J.N. Onuchic, D. Thirumalai, "Navigating the Folding Routes", Science **267**(5204), 1619–1620 (1995).

[129] C.M. Dobson, "Protein folding and misfolding", Nature **426**(6968), 884–890 (2003).

[130] M.J. Todd, G.H. Lorimer, D. Thirumalai, "Chaperonin-facilitated protein folding: optimization of rate and yield by an iterative annealing mechanism", Proc. Natl. Acad. Sci. USA **93**(9), 4030–4035 (1996).

[131] D. Thirumalai, G.H. Lorimer, "Chaperonin-mediated protein folding", Annu. Rev. Biophys. **30**(1), 245–269 (2001).

[132] G.H. Lorimer, "A quantitative assessment of the role of the chaperonin proteins in protein folding in vivo", FASEB J. **10**(1), 5–9 (1996).

[133] G. Stan, B.R. Brooks, D. Thirumalai, "Probing the "annealing" mechanism of GroEL minichaperone using molecular dynamics simulations", J. Mol. Biol. **350**(4), 817–829 (2005).

[134] F. Takagi, N. Koga, S. Takada, "How protein thermodynamics and folding mechanisms are altered by the chaperonin cage: molecular simulations", Proc. Natl. Acad. Sci. USA **100**(20), 11367–11372 (2003).

[135] H.X. Zhou, "Protein folding and binding in confined spaces and in crowded solutions", J. Mol. Recognit. **17**(5), 368–375 (2004).

[136] V. Gintautas, A.W. Hübler, "Resonant forcing of nonlinear systems of differential equations", Chaos **18**(3), 3118 (2008).

[137] V. Gintautas, G. Foster, A.W. Hübler, "Resonant forcing of chaotic dynamics", J. Stat. Phys. **130**(3), 617–629 (2008).

[138] H. Gelman, M. Platkov, M. Gruebele, "Rapid Perturbation of Free-Energy Landscapes: From In Vitro to In Vivo", Chemistry **18**(21), 6420–6427 (2012).

[139] I. Schoen, H. Krammer, D. Braun, "Hybridization kinetics is different inside cells", Proc. Natl. Acad. Sci. USA **106**(51), 21649–21654 (2009).

[140] A. Dhar, S. Ebbinghaus, Z. Shen, T. Mishra, M. Gruebele, "The diffusion coefficient for PGK folding in eukaryotic cells", Biophys J. **99**(9), L69–L71 (2010).

[141] J.R. Lakowicz, *Principles of fluorescence spectroscopy* (Plenum Press, New York, 1983).

[142] A. Dhar, K. Girdhar, D. Singh, H. Gelman, S. Ebbinghaus, M. Gruebele, "Protein Stability and Folding Kinetics in the Nucleus and Endoplasmic Reticulum of Eucaryotic Cells", Biophys J. **101**(2), 421–430 (2011).

[143] J. Ervin, M. Gruebele, "Quantifying Protein Folding Transition States with Φ", J. Biol. Phys. **22**(2), 115–128 (2002).

[144] D.U. Ferreiro, J.A. Hegler, E.A. Komives, P.G. Wolynes, "Localizing frustration in native proteins and protein assemblies", Proc. Natl. Acad. Sci. USA **104**(50), 19819–19824 (2007).

[145] M.P. Eastwood, P.G. Wolynes, "Role of explicitly cooperative interactions in protein folding funnels: A simulation study", J. Chem. Phys. **114**, 4702–4716 (2001).

[146] J. Latzer, T. Shen, P.G. Wolynes, "Conformational switching upon phosphorylation: a predictive framework based on energy landscape principles", Biochemistry **47**(7), 2110–2122 (2008).

[147] P. Weinkam, E.V. Pletneva, H.B. Gray, J.R. Winkler, P.G. Wolynes, "Electrostatic effects on funneled landscapes and structural diversity in denatured protein ensembles", Proc. Natl. Acad. Sci. USA **106**(6), 1796–1801 (2009).

[148] C. Hardin, Z. Luthey-Schulten, P.G. Wolynes, "Backbone dynamics, fast folding, and secondary structure formation in helical proteins and peptides", Proteins **34**(3), 281–294 (1999).

[149] J. Latzer, M.P. Eastwood, P.G. Wolynes, "Simulation studies of the fidelity of biomolecular structure ensemble recreation", J. Chem. Phys. **125**(21), 214905 (2006).

[150] J.A. Hegler, P. Weinkam, P.G. Wolynes, "The spectrum of biomolecular states and motions", HFSP J. **2**(6), 307–313 (2008).

[151] D.U. Ferreiro, J.A. Hegler, E.A. Komives, P.G. Wolynes, "On the role of frustration in the energy landscapes of allosteric proteins", Proc. Natl. Acad. Sci. USA **108**(9), 3499–3503 (2011).

[152] P. Weinkam, C. Zong, P.G. Wolynes, "A funneled energy landscape for cytochrome c directly predicts the sequential folding route inferred from hydrogen exchange experiments", Proc. Natl. Acad. Sci. USA **102**(35), 12401–12406 (2005).

[153] L. Sutto, J. Latzer, J.A. Hegler, D.U. Ferreiro, P.G. Wolynes, "Consequences of localized frustration for the folding mechanism of the IM7 protein", Proc. Natl. Acad. Sci. USA **104**(50), 19825–19830 (2007).

[154] P. Weinkam, F.E. Romesberg, P.G. Wolynes, "Chemical Frustration in the Protein Folding Landscape: Grand Canonical Ensemble Simulations of Cytochrome c", Biochemistry **48**(11), 2394–2402 (2009).

[155] P. Weinkam, J. Zimmermann, F.E. Romesberg, P.G. Wolynes, "The Folding Energy Landscape and Free Energy Excitations of Cytochrome c", Acc. Chem. Res. **43**(5), 652–660 (2010).

[156] M.P. Eastwood, C. Hardin, Z. Luthey-Schulten, P.G. Wolynes, "Statistical mechanical refinement of protein structure prediction schemes: Cumulant expansion approach", J. Chem. Phys. **117**, 4602–4615 (2002).

[157] M.P. Eastwood, C. Hardin, Z. Luthey-Schulten, P.G. Wolynes, "Statistical mechanical refinement of protein structure prediction schemes. II. Mayer cluster expansion approach", J. Chem. Phys. **118**, 8500–8512 (2003).

[158] M.C. Prentiss, C. Hardin, M.P. Eastwood, C. Zong, P.G. Wolynes, "Protein structure prediction: the next generation", J. Chem. Theory Comput. **2**(3), 705–716 (2006).

[159] V. Oklejas, C. Zong, G.A. Papoian, P.G. Wolynes, "Protein structure prediction: Do hydrogen bonding and water-mediated interactions suffice?", Methods **52**(1), 84–90 (2010).

[160] A.J. Cuff, E.M. Clamp, S.A. Siddiqui, M. Finlay, G.J. Barton, "JPred: a consensus secondary structure prediction server", Bioinformatics **14**(10), 892–893 (1998).

[161] V. Oklejas, C. Zong, G.A. Papoian, P.G. Wolynes, "Protein Structure Prediction: Do Hydrogen Bonding and Water-Mediated Interactions Suffice?", Methods **52**(1), 84–90 (2010).

[162] G. Hummer, S. Garde, A.E. Garcia, M.E. Paulaitis, L.R. Pratt, "The pressure dependence of hydrophobic interactions is consistent with the observed pressure denaturation of proteins", Proc. Natl. Acad. Sci. USA **95**(4), 1552–1555 (1998).