This is to certify that the
dissertation entitled

EXPERIMENTAL AND COMPUTATIONAL INVESTIGATION OF EARLY EVENTS IN PROTEIN FOLDING

presented by

Vijay R. Singh

has been accepted towards fulfillment
of the requirements for the

Ph.D.	degree in	Physics and Astronomy Biochemistry and Molecular Biology

Major Professor's Signature

8/27/09

Date

EXPERIMENTAL AND COMPUTATIONAL INVESTIGATION OF EARLY EVENTS IN PROTEIN FOLDING

By

Vijay R. Singh

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Physics and Astronomy
Biochemistry and Molecular Biology

2009

ABSTRACT

EXPERIMENTAL AND COMPUTATIONAL INVESTIGATION OF EARLY EVENTS IN PROTEIN FOLDING

By

Vijay R. Singh

Knowledge of the early events in protein folding and its characterization is essential for complete comprehension of the protein folding problem: structure, pathways, and mechanisms. The chemical heterogeneity, topological entanglements, and dynamic nature of the unfolded state conformation pose a huge challenge to a accurate description. Recent years have witnessed a boost in the efforts to develop experimental techniques to characterize the unfolded state of protein under folding conditions. This will have immense implications not only in diagnostic and therapeutic strategies for tackling protein misfolding and aggregation based diseases but also in paving way for development of evolutionarily superior biological structures and design of effective drugs against resistant pathogens.

Using an experimental technique that involves monitoring the tryptophan triplet-triplet optical absorption to measure its lifetime against intramolecular Trp/Cys contact quenching, we investigated the unfolded state of two structurally similar but sequentially nonhomologous still well behaved B1 domains of proteins - L and G. Employing the Szabo, Schulten, and Schulten (SSS) theory with a wormlike chain polymer model, we observe that the loss of denaturant yields an unfolded state that is less diffusive and more compact than the fully denatured state. This reflects the complex internal dynamics of the proteins mediated by transient interactions through

the chain.

Polyglutamine amino acid motif is implicated in several neurodegenerative diseases, all of which exhibit aggregation *in vivo*. To understand the mechanism and correlation between the long glutamine stretches and consequent destabilizion of the proteins leading to amyloid fibrilation, we probed the structural properties of polyglutamine polypeptides using Trp/Cys contact quenching. Modeling the length dependence of contact formation rates with a wormlike chain with excluded volume, we find the polyglutamine peptides to be a unusually "stiff" polymer with a persistence length of $\sim$ 13.0 $\AA$. This propensity for extended conformation can explain both the decrease in stability of the host protein and the propensity to form amyloids once unfolded.

A detailed atomic level characterization and the intricate interplay between kinetic and thermodynamic controls at various phases of protein folding can be obtained using accurate molecular modeling and simulations. This complements the experimental techniques by providing insights at higher temporal and structural resolutions, generally inaccessible to experiments. We consolidated this technique with Trp/Cys contact quenching to characterize the unfolded states of protein L. Making quantitative comparisons between experimentally obtained denaturant-induced unfolded ensemble and computationally simulated temperature-induced unfolded ensemble, we observe a low intrachain diffusion rate that decreases with denaturant concentration. This low diffusion can limit the folding speed and its origin can be quantified by close inspection of the simulated trajectories.

ACKNOWLEDGMENT

It has been a five year long journey and as I bring my graduate life to a close, I realize how various people have helped and guided me to be the person I am today.

First and foremost I would like to thank my advisor Dr. Lisa Lapidus.....You have given me the freedom to explore and gently nudged me in the right direction when I was stranded. You have also helped me put my thoughts in newer and better perspective and listened to me patiently while I made my naive arguments and proposals.....I surely could not have asked for a better advisor, mentor and friend. I will always take pride in acknowledging you as my mentor through my graduate experience. Apart from science, I have also learned valuable time management and leadership skills from you that are helping me develop a more harmonious personality.

I also welcome the rewarding experience and guidance that I received from Dr. Wedemeyer.....You walked me through the areas of computational biophysics. Your thoughts and ideas infuse a lot a positive aggression in me to tackle the problems in science from multiple perspectives.

During the course of this endeavor I have had the opportunity to meet and grow with many people, colleagues and otherwise. Michaela executed many of the initial experiments in addition to performing protein expression and purification.....I picked up my first lessons in cell culture from you. A very special thanks to Terry.....But for your hard work, skill, and dedication, many of us in the lab would not be able to confront our research problems with such ease and planning.

Beyond friends and colleagues is family. My family has been a source of inspiration and perspiration for me....I would like to express my gratitude to you. My Uncle-

Aunt and their extended family (Sandhya and family, Vandana and family)....You stand behind me all through my life and teach me the value of education and responsibility. I am the man I am because of you. To Kavitha and Sanjay.....These words would not have been written but for you. My parents.....You had the foresight to send me to better schools, away from home, when I was yet to comprehend the magnitude of your decision.

And finally to my wife Kasturi Chatterjee, who is on the way to her own graduation, for always being there for me. For always having the faith and trust in my endeavors. And for understanding me even when I do not make any sense.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction to Proteins and The Protein Folding Problem

1.1 The Protein Folding Problem

Proteins are biological heteropolymers understood to be constituted from twenty different species of monomers called amino acids. They form some of the most versatile molecules in living beings, playing significant role in almost all the biological functions. About half the dry mass of the human body is proteins. Protein functions include operating as antibodies to defend against pathogens, transport and storage of ions and molecules (like oxygen by hemoglobin, iron by transferrin), form the building blocks of other biological structures (for instance, bones and hairs are constituted mainly by collagen), and enzymatic catalysis of most chemical reactions. For proteins to successfully perform their biological function, they are known to adopt a well

defined unique three-dimensional structure.

The architecture of the biologically active protein depends on its amino acid sequence and the biological function of the protein itself is strongly correlated to its structure. Therefore, any deviation from the native state conformation can lead to the so called "protein-misfolding diseases". A common feature among many such diseases, such as the Alzheimer's, Parkinson's, and Prion diseases such as bovine spongiform encephalopathy (BSE) and its human equivalent Creutzfeld-Jakob disease (CJD), is formation of toxic aggregation or amyloid fibrils *in vivo* [1–3]. Development of any therapeutic treatment against such diseases requires a knowledge of the dynamics, kinetics, mechanisms, and thermodynamics of the folding pathways and concomitant conformations [4, 5]. This constitutes **THE PROTEIN FOLDING PROBLEM** - accurately predicting the three-dimensional (most often) thermodynamically stable native state of a protein from the knowledge of its amino acid sequence under normal physiological conditions [6]. This involves not only predicting of the native state but also delineating the mechanism as to how the protein folds. A complete solution to the protein folding problem is possible only through a comprehensive understanding of the unfolded state ensemble under folding conditions [7, 8]. The characteristics of nonnative states has yet to be fully comprehended and its role in protein folding is now being actively investigated. Large amounts of research has been invested on studying the native states of proteins and hence we currently have plenty of information on the folded configurations. It is essential to understand the properties of the unfolded states of a protein under native conditions to predict the overall and complete behavior of the protein. This will have deep impacts in pro-

tein engineering and drug design to not only combat the current pathogens through immune resistant therapeutic strategies but also to design evolutionarily superior biological machinery.

1.2 Amino Acids and Proteins

In the cells, proteins and polypeptide chains are synthesized linearly in the ribosomes through a polycondensation reaction of amino acids. An α-amino acid is constituted of a carboxyl group, an amino group, a hydrogen atom and a distinctive side chain (R group) chemically bonded to the α-carbon atom. The structure of this R-group distinguishes the amino acids from one another by according them different physical characteristics - polar and non polar, aromatic, hydrophobic or hydrophilic, basic or acidic, etc. The α-carbon atom is asymmetric and all C_α atoms except for glycine have the same chirality of left-handedness. The 20 most common naturally occurring amino acids are:

- Non-Polar Hydrophobic Sidechain Residues: Alanine, Isoleucine, Leucine, Methionine, Phenylalanine, Proline, Tryptophan, Valine.
- Polar Neutral Residues: Asparagine, Cysteine, Glutamine, Glycine, Serine, Threonine, Tyrosine.
- Positively Charged (Basic) Residues: Arginine, Histidine, Lysine.
- Negatively Charged (Acidic) Residues: Aspartic acid, Glutamic acid.

A peptide bond between the carboxyl and amino groups of adjacent amino acid residues causes the polymerization of the resultant protein as shown in figure 1.1.

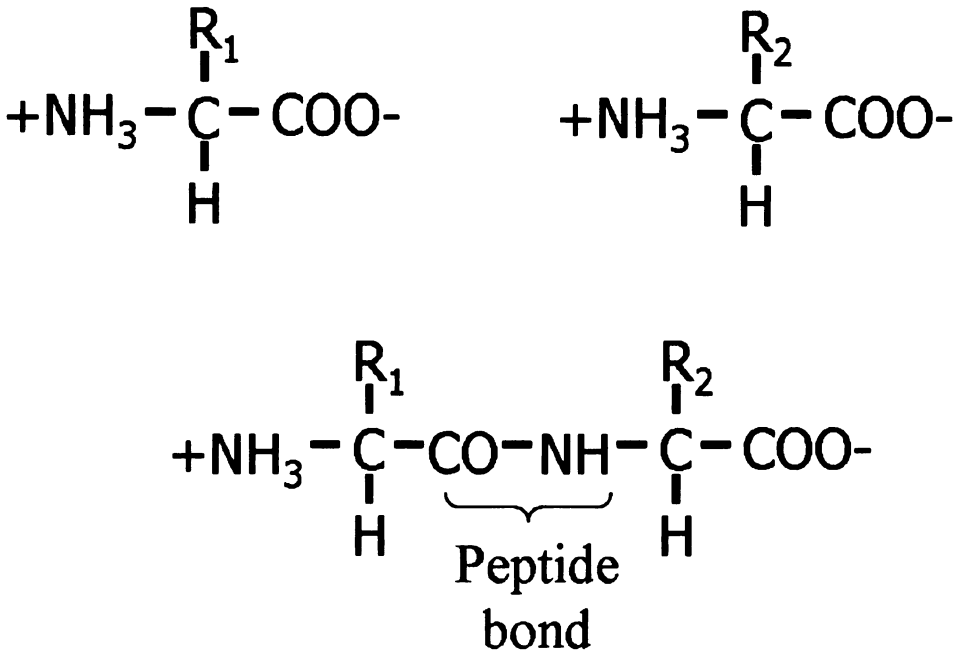


Figure 1.1: Amino Acids and Protein Polymerization.

Protein sequence is determined by gene sequence, the protein sequence governs the protein structure; and the protein function is determined by its structure. There are various physical interactions - electrostatics of charge and partial charges, geometry of hydrophobic collapse and stereochemistry of amino acid residues, and strength of various chemical bonds, etc., - at work right from its genesis in the ribosomes that define a protein's native structure [9]. Proteins with similar structure appear to have a similar physiological function. Therefore, the key to understanding protein function is in its geometrical architecture. Existence of regular local ordering in the structure of proteins was theoretically predicted by Pauling and Corey in the year 1951 [10, 11]. They predicted the existences of α-helices and β-sheets that are en-

ergetically more stable and have a well defined network of hydrogen bonds. This represents a loss in the degree of freedom when the protein adopts its native state. Following that, innumerable protein structures have been solved and a few repetitive folding motifs seem to be a common theme in all the protein architectures. They in fact seem to form the building block for the comprehensive structure. Based on the current database of solved protein structures, the strata in protein architecture is segregated into four levels of structural hierarchy (Figure 1.2):

1. Primary Structure: This is just the chemical composition of amino acid sequence linearly along the main chain backbone. The primary structure of every protein is unique. By convention, the protein primary structure is numbered from the amino terminal to the carboxyl terminal.
2. Secondary Structure: Local ordering in the linear sequences of the chemically bonded amino acids form the secondary structure. Governed and driven by the energy considerations, they are defined by backbone hydrogen bonding and bond rotations. Depending on the network of the hydrogen-bonds between the amide group and the carboxyl group, the secondary structure motif could either be a α-helix, a β-sheet or a turn. In α-helices the H-bonds are aligned parallel to the helical axis, while in the β-sheets the H-bonds lies perpendicular to axis of the strand.
3. Tertiary Structure: The thermodynamically and kinetically stable three dimensional structure formed by packing together of the secondary structure elements is referred to as the tertiary structure. This, in most cases, is the biologically active and functional protein self-assembled by the considerations of energy,

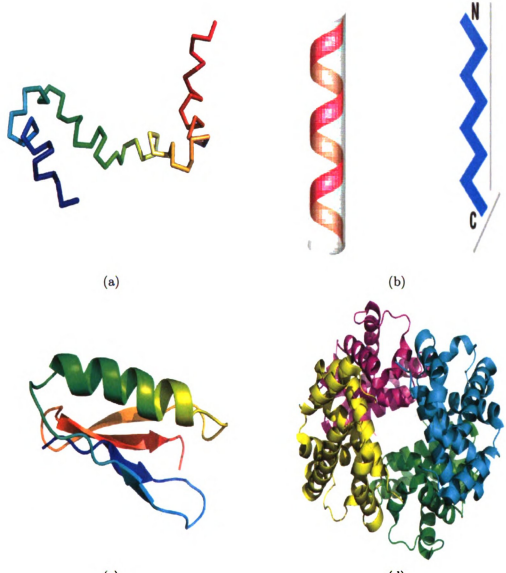


Figure 1.2: Structural Hierarchy in Protein Structure: (a) Primary structure being the linear chemical sequence, (b) Secondary structures α-helix and β-sheet, (c) Tertiary structure made up by synthesis of secondary structure elements, and (d) Quaternary structure. Figure (b) is adopted from Petsko and Ringe [12].

entropy, and hydrophobicity.

4. Quarternary Structure: An ensemble of smaller units of tertiary structures, usually stabilized by weak inter-residue interactions, to form an integrated structure of multiple domains where each domain is constituted by a single distinct polypeptide chain.

1.3 Forces and Interactions

The secondary structures in proteins and polypeptides are usually not stable in isolation and the tertiary and quaternary protein structures are only marginally more stable compared to the unfolded conformation ensemble. A delicate balance between the forces that drive the protein towards folding and those that oppose it lends the native state a stability that generally corresponds to a thermodynamically global energy minimum. Typically the free energy difference due to protein folding is in the range of -5 to -15 kcal/mol [13,14]. The current consensus for the dominating effects in protein folding are hydrophobic forces, network of hydrogen bonding, and entropic considerations [9]. Other factors involved in folding and stability are: disulfide salt-bridges, Van der Waals forces, electrostatic Coulomb interactions, solvent pH, and temperature.

The Hydrophobic Effect: Hydrophobic effect is defined as the free energy associated with transfer of hydrophobic surface from protein interior to water. Seven of the twenty amino acids are considered to have strongly hydrophobic side chains. Residues valine, isoleucine, leucine, methionine, phenylalanine, alanine, and proline

each have a non polar side chain and so cannot form any hydrogen-bond with water molecules. They in fact disrupt the hydrogen-bond network favorable to water molecules and force them in forming an ordered (decreased entropy) cage-like or ice-like structure around the non-polar group [15]. This entropic loss can be minimized by minimizing the surface area of the non-polar group accessible to the polar solvent. This results in the molecules with hydrophobic side chains coalescing together to exclude the water molecules and thus collapsing into a compact globular state. The degree of compaction and relative strength of the stability of native conformation depends on distribution of hydrophobic residues along the primary sequence of the protein chain. However, the hydrophobic effect alone cannot account for the observed regular geometric structures in a protein core and therefore other molecular forces and interactions must be taken into consideration to describe the unique native state.

Hydrogen-Bonding: This is a partial sharing of a hydrogen between two atoms where the hydrogen is covalently bonded to one of the electronegative atoms. The link is provided by electrostatics and strength dictated by the electronegativity and orientation of the bonding atoms. Thermodynamically each hydrogen-bond is hypothesized to contribute -1 to -5 kcal/mole, depending on its chemical environment, towards the protein native state stability. If the hydrogen-bonding between protein-solvent becomes dominant it will lead to unfolding of the protein chain fractions. Therefore, if the protein chain segments need to maintain the identities of their local secondary structure, it is necessary for them to be hydrophobic enough so that protein intramolecular hydrogen-bonding dominates. Those segments of the protein chain that have protein-solvent hydrogen-bond dominating will primarily have a dis-

ordered or coil structure [9]. Since the network of hydrogen bond is well defined in the secondary structure motifs of a protein, it seems plausible that the hydrophobic interactions first collapse the relevant protein chains segments; following this collapse and consequent geometric rearrangement of the hydrophobic side chains into an minimal "frustrating" configuration, the spatially short range hydrogen-bonds begin to take effect and cause the formation of secondary structures.

Configurational Entropy: As far as the protein alone is considered, entropic effects would be an opposing force to protein folding since the native state configurations are extremely compact and the effective volume occupied by a folded protein is much smaller than an unfolded protein conformation. The entropy considerations cannot, in general, be isolated from the hydrophobic effects. The nonpolar protein segments trigger the hydrophobic effect in the solvent resulting in high entropic barrier of the collapsed state to minimize the accessible surface area of nonpolar residues to the polar solvent.

A delicate balance between these two opposing factors - forces that drive protein towards folding and forces that inhibit protein folding - provides the marginal stability of the native state conformation.

1.4 Time Scales and Energy

Protein folding kinetics and dynamics is largely mapped onto two time scales: the events occurring from femtoseconds to milliseconds are referred to as the microscopic regime and the timescales beyond a millisecond are referred to as the macroscopic

regime. Events in protein folding reactions span about 15 orders of magnitude in time. This is a rather large temporal range considering the dimensions of the protein. The slowest folding proteins take a few minutes or sometimes even hours to fold while some smaller proteins (less than $\sim$ 100 amino acid residues) fold within 100 microseconds. For a comprehensive understanding of protein folding - pathways, kinetics, thermodynamics and mechanisms - it is essential to decipher the relation between the motions at various time scales. Understanding the interlink between the faster and slower processes will help better predict protein folding and native structure from sequence and consequently the functions. A typical distribution of protein folding events and associated time scales is shown in table 1.1. The transition from an apparently random configuration to the native structure involves events such as bond stretching, bond angle bending, formations of loops, secondary structure elements and their self-assembly, formation of disulfide bonds, folding of the protein itself, and breathing motion of the native state. To capture all the cardinal processes in the relaxation kinetics of the protein, it is essential to accurately capture and understand the time scales of the motions involved.

The evolving experimental techniques are pushing towards improved temporal resolution for quantifying and studying the protein folding events on a wide range of timescales. The bonded interactions generally make the highest contributions to the net energy of any molecule at the atomic level. An example of energy scales involved for various interaction in the aqueous environment of protein is shown in table 1.2. The electrostatic interaction is governed by the Coulomb's law ($U = kq_1q_2/\epsilon r$) and is dependent on the charge of the atoms under consideration. Fundamentally, both

Time Scale (s)	Event Description
10^{-15} - 10^{-12}	Bond stretching and angle bending.
10^{-12} - 10^{-9}	Surface side chain motion and loop motion; Local breathing and collective motion.
10^{-9} - 10^{-6}	Formation of secondary structure and loops; Helix-coil transition and global hydrophobic collapse.
10^{-6} - 10^{-3}	Formation of "intermediates"; Folding of smaller proteins.
$>10^{-3}$	Cis-trans prolyl-peptidyl isomerization; Protein folding.

Table 1.1: Typical time scales in protein folding events. Note the wide temporal range.

Van der Waals and hydrogen bond interactions are electrostatic in their origin. The Van der Waals interaction arises from transient asymmetrical charge distribution in atoms resulting in a dipole formation. This causes a mutual attraction between the atoms until they approach the minimum mutual distance (Van der Waals contact) where repulsive forces between the outer electron clouds begin to dominate.

Energy Scale(kcal/mole)	Description
20-150	Covalent bonds
1-5	Hydrogen bonds; Electrostatic interaction
1-2	Aromatic-Aromatic interaction
<1	Van der Waals attraction

Table 1.2: Typical energy scales measured in protein folding reactions.

1.5 Foundation for Protein Folding Studies

The central dogma of genetics enunciates the process of biosynthesis of proteins. It explains the transcription of DNA to RNA and translation of RNA into proteins.

Therefore, knowledge of the genes can reveal the sequence information of the protein. The primary sequence of proteins can also be determined by mass spectrometry. The very first protein structures to be solved experimentally (X-ray crystallography) were for myoglobin and hemoglobin in around 1958 [16,17]. By 1962 Anfinsen demonstrated that the amino acid sequence contains all the information necessary for the formation of the native state. Under right conditions, folding and unfolding of a protein can be reversibly achieved *in vitro* [18, 19]. These results had far reaching consequences in that it opened up the possibility of studying the protein in isolation - experimentally and computationally.

The Anfinsen Experiment: One of the fundamental observations [18] that promises success in the study of protein folding is provided by the Anfinsens experiment. It provided the very first evidence of the possibility that all the information needed for correct folding of the protein from any arbitrary unfolded state is latent in the amino acid sequence. The experiment demonstrated that Ribonuclease A can be fully denatured by reducing the disulfide bonds and dissolving it in 8M urea. Subsequently it can be reversibly renatured by diluting away the urea and oxidizing it to allow the disulfide formation. More than 90% of the native state activity was restored.

He therefore proposed the thermodynamic hypothesis of protein folding according to which the native conformation of a protein is thermodynamically the most stable state (corresponding to lowest free energy) and the protein adopts this structure spontaneously. This mean that all the information needed to form the three-dimensional structure of the protein was inherent in the sequence itself. This hypothesis was

eventually challenged as further research into the protein folding and mechanisms revealed the role of the other molecules *in vivo* that assist protein folding.

Levinthal's Paradox: Consider a polypeptide chain with one hundred amino acid residues. Assuming each amino acid to have only 3 possible conformations, the whole polypeptide chain will have 3^{100} possible conformations. If the transition time between each conformational state is 1 picosecond, it would take the protein 16×10^{27} years to find its native state. The Age of the universe is believed to be 15×10^{9} years. If the protein is to attain its correctly folded configuration by sequentially sampling all the possible conformations, it would require a time longer than the age of the universe to arrive at its correct native conformation. But most proteins have been observed to fold on a submillisecond timescale. Therefore, the folding pathway towards the native state cannot really be an unbiased random search but rather has to be some specific pathway that is biased towards the native structure which is kinetically the most accessible state [20].

1.6 Mechanism and Hypothesis for Protein Folding

Traditionally two main mechanisms for protein folding has been proposed - the Thermodynamic hypothesis and the Kinetic hypothesis [21]. The thermodynamic hypothesis proposes that the native state adopted by a protein is thermodynamically the most most stable conformation and corresponds to the global free energy minimum. The native state is completely determined by the interatomic interactions which are

in turn determined by the amino acid sequence of the protein chain and the environment it is in. The Kinetic hypothesis, on the other hand, suggests that the native state is that which is kinetically the most accessible conformation, and the biologically active protein need not necessarily correspond to the state of minimal free energy.

The thermodynamic hypothesis of protein folding gained support primarily from Anfinsen's experiment on protein denaturation and renaturation which established that the native state of the protein correspond to the global free energy minimum under the given constraints of the physiological conditions. The native state established is the same *in vivo* and *in vitro*. The support for the kinetic pathway mechanism sprouts mainly from the Levinthal's paradox. He argued that if the protein were to sample all of the conformation space, it would take astronomical time for the protein to reach its native state. Since the proteins are known to adopt their native structure in a fraction of that time, it must be that only a small fraction of this pathway is kinetically sampled. This kinetic sampling leads to a native state that is lower in energy to most other states, but is not necessarily the lowest energy state (the global energy minima) [22]. For example, the biologically active state of plasminogen activator inhibitor(PAI-1) is not the stable conformation state with minimal free energy but is a metastable state [23]. The large volume of the configuration space makes it impossible for the protein to always achieve the lowest energy state. However, given long enough sampling time, the protein will eventually find this state.

Although the debate about the validity of one over the other is not absolutely settled, a synergy between the two seems to be fast emerging that makes the kinetic

hypothesis a larger view that subsumes the thermodynamic hypothesis [21,22,24,25]. It has been proposed that an initial collapse of the protein is governed by kinetics, this is followed by the rate-limiting step of slower relaxations of the protein into the native state and is dictated by thermodynamics [26]. Further advances in experimental techniques have probed the early events in protein folding to a greater atomic and temporal resolution to reveal finer details of the protein folding process. It has led to the "New View" of protein folding wherein the pathway to native state is not single, unique, well defined, and constrained but there is a muliplicity of pathways to choose from that can lead to the native state conformation. The folding pathways can be represented by a free energy landscape as shown in figure 1.3 or equivalently by a entropic landscape in which the protein diffuses through sampling the kinetically accessible conformations and eventually adopt the appropriate native state. The energy landscape is often like a funnel in which the minimum energy conformation connotes the native state. In the context of the kinetic hypothesis, the misfolded states of any protein are essentially kinetically trapped conformations.

1.7 Unfolded States and Protein Folding

Quite often the protein folding problem is viewed as three different, but closely related, issues [6]. The first issue is that of the folding code through which interatomic forces acting on and by the amino acid residues of the polypeptide chain determine the three-dimensional native state by thermodynamic considerations. The second issue is the computational problem of predicting the native state using both

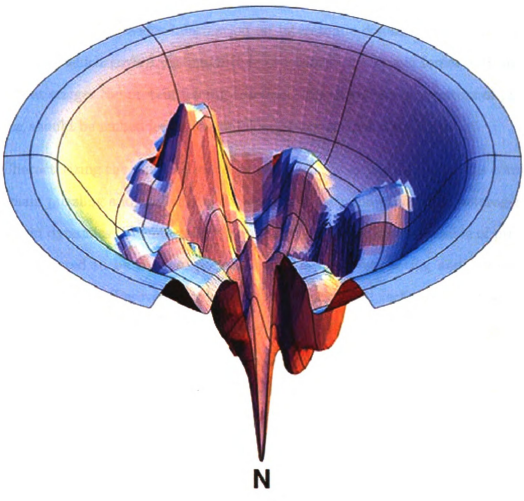


Figure 1.3: Depiction of a typical energy landscape for protein folding. This figure is adopted from Dill and Chan [27]

knowledge-based (homology modeling) and physics-based methods. The third issue is that of kinetics - the speed at which protein folds and factors that govern this rate. To obtain solutions to any of these problems it is essential to not only characterize the native state conformation but also investigate the details and characteristics of

the unfolded conformation ensembles. It is the distribution of the unfolded chain conformation that provides information on interatomic distances and their relative orientations which govern the magnitude of various interatomic forces and interactions that trigger the protein folding reaction. Therefore, the progenitors of protein folding should be embedded in the unfolded state conformation.

Characterizing the unfolded states of a protein under folding conditions continue to remain puzzling and challenging on account of their conformational heterogeneity and rapid dynamics. Traditionally the unfolded conformational ensemble of proteins were considered to be random polymers that are non-interacting and lacking any defined structure. However, development of experimental techniques and advances in NMR spectroscopy suggest the presence of residual structure in the unfolded states of many proteins [28–30]. Another structural element that is yet to be fully understood is the loops that connect other secondary structures. The stability of the native state conformations have been observed to be dependent on the loop lengths too. Loop formation is one of the most fundamental process occurring in protein folding and dynamics. It is only when nonlocal chain segments come in close spatial contact that the long range and short range interatomic interactions begin to guide the protein folding towards the energy minimum of the folding funnel. The rate of intramolecular contact formation in the protein chains also characterizes the kinetics of sampling the free energy landscape. It also sets the timescales for the loop closure and consequently for speed of protein folding. The characteristics of this intrachain loop diffusion forms a significant aspect of the unfolded state attributes.

A quantitative description of the unfolded state topology and the intramolecular

contact formation in the early stage of protein folding can be obtained through a statistical treatment of the conformations by investigating them through well described physics based polymer models. This will enable us to develop a picture of the unfolded state conformation distribution which is sequence dependent and will have all the information needed to achieve the native state conformation. It would be impossible to describe discrete conformations of the unfolded states owing to its dynamic heterogeneity and the degeneracy of free energy associated with multiple conformations. For an analytical treatment of the experimental data and to obtain a macroscopic description of probability distributions, often the freely-jointed chain does well to capture the characteristics of long chains. But for shorter chains and for chains in mild denaturing conditions, a wormlike chain with excluded volume interaction is employed that incorporates an intrinsic stiffness. The details of some of the polymer models are discussed in appendix A.

This probability distribution obtained from the polymer models can be used to model the experimental data and describe the conformational distribution of the unfolded protein. Time evolution of the structural reconfiguration and other folding events can be characterized through the diffusion coefficient and the intramolecular end-to-end distance probability. In this work I have used a wormlike chain model to incorporate the property of "stiffness" and excluded volume, but this model did not consider the hydrophobic effects and the residual structure in the unfolded fractions of the protein under folding conditions. The growing experimental evidence of nascent structures in early stages of protein folding suggests that we definitely need to invoke these effects in the polymer models for a more accurate description of the

characteristics of unfolded states. Many of these shortcomings can be transcended by using all-atom molecular dynamics simulations that incorporate all the chemical and structural details of amino acid.

It is now being acknowledged that protein misfolding and aggregation may have its seed in unfolded fractions or the partly structured folding intermediates. Hence, the earliest steps in the protein folding process, such as loop formation mentioned above, may hold the key to understanding the pathogenesis related to aggregation, misfolding and amyloid formation. Adding another step towards understanding of protein and peptide structures and their early stages in particular, I present in this thesis, a few experimental results for molecules with and without known propensities for aggregation and model them with a wormlike chain polymer model and all-atom molecular dynamics simulations.

Chapter 2

Transient Absorption Spectroscopy Applied to Polymer Dynamics

2.1 Introduction

To study and characterize the unfolded states of a protein, it is essential to understand one of the most fundamental processes - intrachain loop formation [31]. Since the slower processes are, in principle, a consequence of the faster processes, the driving mechanism for various phenomena can be better understood by estimating the underlying time scales involved. Such fast processes as loop formation can be measured using pump-probe laser spectroscopy to cover a large orders of magnitude in time (until at least a millisecond as depicted in figure 2.1) limited only by the relative lifetime of the metastable states.

Function of the pump is to induce a time evolving perturbation in the sample.

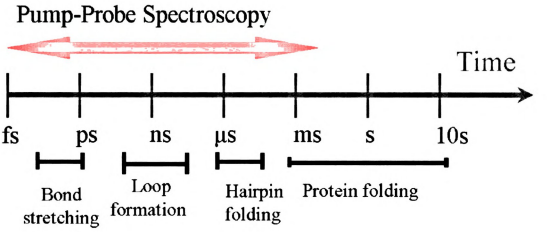


Figure 2.1: Time Span of Folding Events and Pump-Probe Spectroscopy.

This perturbation is studied in the presence of a quencher which systematically represses the perturbation by processes such as collision or an electron transfer among other possible pathways. The probe beam monitors the donor electronic states for any changes. It is therefore essential to ensure a choice of the probe wavelength such that it has an intraband resonance with the electronic excitations of the donor. To minimize possible disruptions of the sample, and also to ensure that the detector is not saturated, the probe is generally chosen to have energy lower than the pump beam. For more true and accurate measurements it is desirable to have a pulse repetition rate in the pump to be less than the ground state repopulation time of the donor.

Operating in the linear regime of the detector and within the damage threshold of the sample, one of most straightforward and direct measurements is that of the optical density (also called absorbance). Its relation to the probe intensity is given

by the Beer-Lambert law as

$$OD = \log\left(\frac{I_0}{I}\right) \tag{2.1}$$

where I_0 is the intensity of the reference beam and I is the intensity of the beam after it passes through the sample. The probe signal decay thus recorded can largely be fit to a sum of first order decaying exponentials.

2.2 Application of Pump-Probe Spectroscopy

For the study of loop formation dynamics (intramolecular contact formation) in proteins and peptides, using pump-probe laser spectroscopy in the UV-Vis spectral domain it is essential to have:

1. Two lasers: A probe beam and a short pulsed laser for pump.
2. A long lived target (donor) for perturbation/excitation.
3. A quencher (acceptor) to monitor the donor perturbation against.

The roles of donor and acceptor are very well adopted by two naturally occurring amino acids tryptophan and cysteine respectively [32]. Upon absorption of UV light at about 289 nm, the triplet energy states of tryptophan is populated via a non radiative intersystem crossing from the higher singlet energy states as shown in figure 2.2.

Quantum mechanically this transition between states of different spin multiplicities is forbidden. But the spin-orbit interaction is known to partially eliminate such

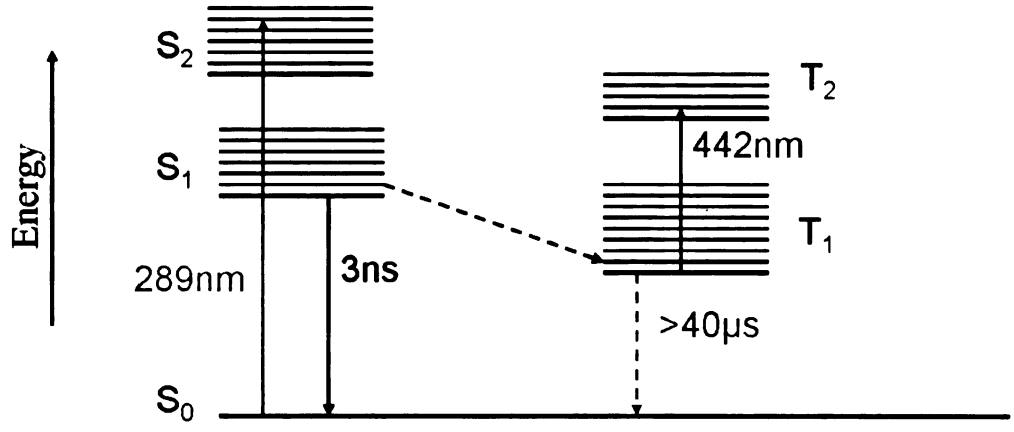


Figure 2.2: Electronic energy levels of tryptophan. The spin of an excited electron can be reversed achieving intersystem cross over.

forbiddeness and the transition probability is increased with overlapping of the corresponding vibrational states [33]. In the absence of any quencher this triplet state lives for more than $\sim$ 40 microseconds [34–36]. The lifetime of this triplet state is reduced in presence of cysteine, an efficient quencher. At physiological pH, cysteine is the most efficient quencher, with a rate of at least 400-fold faster than all other amino acids with the exception of tryptophan. The rate of quenching with cysteine is about $2.0 \times 10^8 M^{-1} s^{-1}$ [32], where M stands for mole.

The experiment yields an observed rate for varying temperatures and viscosities in a range of denaturant concentrations. To minimize the decay rate uncertainties, the signals are averaged for 128 pulsed laser shots. The general form for the observed triplet decay rate can be expressed as [31],

$$k_{obs} = k_0 + \sum_i k_i^{uni} + \sum_i k_i^{bi}[i] \tag{2.2}$$

where k_0 is the decay rate in absence of any quencher i. k_i^{uni} and k_i^{bi} are the unimolecular and bimolecular quenching rates respectively. $[i]$ is the quencher concentration. In almost every experiment performed, we typically employ a single quencher. The bimolecular rates do not make any significant contributions to the observed rate on account of the low sample concentrations ($\sim 30\mu M$) used in the experiments. Also, the tryptophan and cysteine are engineered to be close enough in sequence separation to make $k_i^{uni} \sim$ 10x faster than k_0. Consequently equation (2.2) simplifies to $k_{obs} = k_i^{uni}$, with effectively only one quencher i in the protein chain.

To study the rate of intramolecular contact formation in a loop formed by tryptophan at one end and cysteine at the other (figure 2.3), we engineer these residues into the protein or peptide if they do not already exist. The rate of end-to-end contact formation is then studied using optical (triplet to triplet [34, 36, 37]) absorption by measuring the lifetime of the excited triplet state of tryptophan. In presence of cysteine the quenching rates follow an exponential decaying distance dependence [38,39]:

$$q(r) = q_0 \exp\left[-\beta(r - a_0)\right] \tag{2.3}$$

where r is the intramolecular distance, a_0 is the distance of closest approach (defined to be $4\mathring{A}$), $q_0 = 4.2$ ns^{-1}, and $\beta = 40$nm^{-1}. The quenching is a very short range process.

In the process of diffusing towards and away from each other, the metastable triplet state of the donor is quenched by the close van der Waal contact with the acceptor. Using the kinetic model of figure 2.3, the observed lifetime of the donor

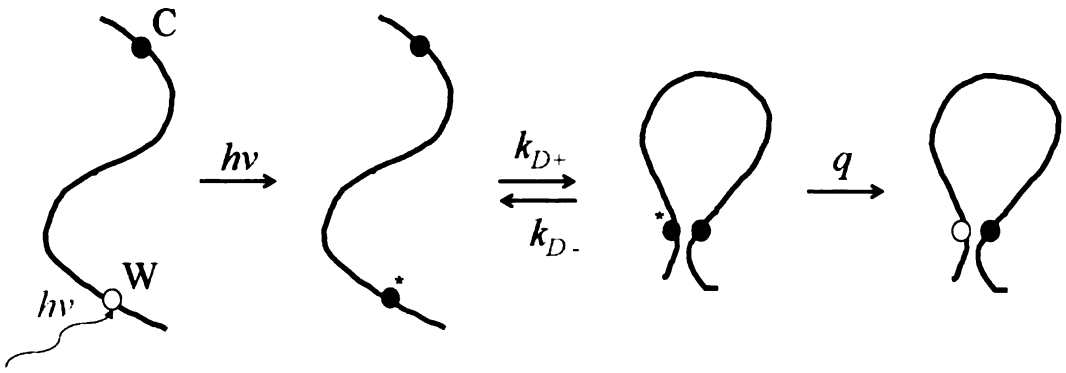


Figure 2.3: Schematic of loop formation and quenching. Tryptophan triplet states are selectively populated by UV pulse. The loop ends diffuse towards and away from each other at rates k_{D+} and k_{D-} respectively. Upon a van der Waals contact, cysteine quenches the tryptophan triplet states at a rate q.

can be approximated by treating it as a two step process -

1. The two ends diffuse towards each other at a rate k_{D+} to form the encounter complex (C*).
2. Cysteine either quenches the excited triplet state with rate q or they diffuse away from each other at a rate k_{D-}.

Representing it in the form of a chemical reaction to determine the rate equation,

$$R \underset{k_{D-}}{\overset{k_{D+}}{\rightleftharpoons}} C^* \xrightarrow{q} P \tag{2.4}$$

Where R represents the ensemble of protein molecules with various end-to-end distances. C^* is the ensemble of molecules forming the encounter complex and P depicts the conformational ensemble after quenching. Using a steady state approximation for the encounter complex formation,

$$\frac{d[C^*]}{dt} = 0 \tag{2.5}$$

that is, the encounter complex does not accumulate over time.

$$\frac{d[C^*]}{dt} = 0 = k_{D+}[R] - (k_{D-} + q)[C^*] \tag{2.6}$$

$$\frac{d[P]}{dt} \equiv k_{obs}[R] = q[C^*] = q\left(\frac{k_{D+}}{k_{D-} + q}\right)[R] \tag{2.7}$$

Hence, the observed rate is given as

$$k_{obs} = k_{D+}\frac{q}{k_{D-} + q} \equiv k_{D+}\phi \tag{2.8}$$

where k_{D+} is rate of diffusion of the loop ends towards each other, k_{D-} is the rate of diffusion away from each other, q is the quenching rate and ϕ is probability of quenching. For $q >> k_{D-}$, the observed rate reduces to the diffusion limited rate: $k_{obs} = k_{D+}$. And for $q << k_{D-}$, it gives the reaction limited rate:

$$k_{obs} = q\left(\frac{k_{D+}}{k_{D-}}\right) = qK_{eq} \equiv k_R, \tag{2.9}$$

where k_R is the reaction limited rate, and K_{eq} is the equilibrium constant for forming the encounter complex. In our experiments usually $q \sim k_{D-}$, and hence we need to isolate both the reaction-limited and diffusion-limited rates. The observed rate can be rearranged to be written as,

$$\frac{1}{k_{obs}} = \frac{1}{k_{D+}(\eta, T)} + \frac{1}{k_R(T)} \tag{2.10}$$

where we posit that the diffusion limited rate, k_{D+}, depends on both temperature (T) and viscosity(η) of the solvent and the reaction-limited rate, k_R, depends on temperature alone [38]. This makes it possible to extract these individual rate coefficients by performing the experiments at varying temperatures and viscosities. The technique was developed and first built at the Laboratory of Chemical Physics, National Institutes of Health.

2.3 The Instrumentation

For the measurement of transient absorption in a transmission mode, the instrument is designed to have a collinear geometry as shown in figure 2.4.

For the tryptophan excitation a 1-mJ, 8-ns UV pulse at 289 nm is employed. A 266nm fourth harmonic of Nd:YAG laser is Raman shifted using $1m$ long methane cell with a pressure of about 250 psi. It has been observed and documented that photodestruction of tryptophan increases at lower excitation wavelengths [40]. To minimize the photodamage, we shift the wavelength of the excitation pulse to 289

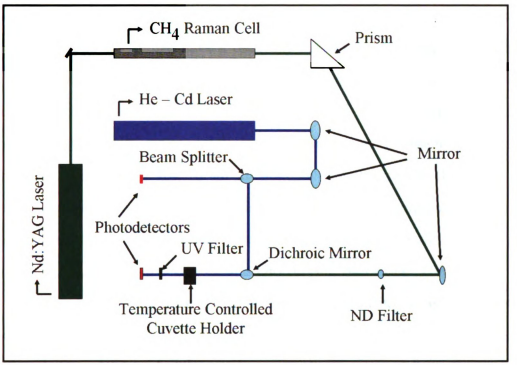


Figure 2.4: Transient Absorption Instrumentation

nm.

After the nanosecond UV excitation, the tryptophan triplet states are monitored using a continuous wave 441 nm (He-Cd laser). The output beam from the CW laser is split into two parts: a reference beam and a sample probe beam. Changes in the transmitted probe beam intensity is measured using a silicon photodiode (New-Focus), recorded with a digital oscilloscope (Tektronix TDS), and stored on a computer using GPIB interface. Any possible pump leakage and high frequency cable noise is subtracted from both parts of the probe beam by recording a background with the pump in absence of the probe beam. Since the triplet lifetime can range from nanosecond to few milliseconds, the absorption is recorded to cover a wide dy-

namic temporal range of 10 ns to 10 ms. This is accomplished by using two different oscilloscopes. The UV pulses have a tendency also to produce radicals which can have a lifetime of about a millisecond and absorb light near 450nm, and this gets convolved with the triplet lifetime measurement [32,36]. It is therefore necessary to record and observe for a long enough time that encompasses the essential and critical experimental events to be able to sift all possible contributions to the decay. The protein/peptide sample solution cuvette is placed in a Peltier temperature controlled sample holder(Quantum Northwest). The data is collected at five different temperatures of 0, 10, 20, 30 and 40 degree Celsius and varying viscosities. The viscosity is an important parameter in equation (2.10) and is varied by using measured quantities of sucrose in the solvent. More recently, we have also incorporated a digital signal amplifier (LeCroy DA1886A differential amplifier). The output from this amplifier of 100MHz bandwidth used in comparator mode with gain of 1x is coupled to a 350 MHz four channel preamplifier (SR445A, SRS Inc.) that can be used in a cascaded configuration with a gain of 5x at each stage. I predominantly used a total gain of 5x for most experiments. This allows to lower the sample concentration and also decrease the pump pulse power thereby reducing the production of free radicals and hydrated electrons.

The optical alignment of the instrument can be optimized using benzophenone, an organic compound, in acetonitrile. At a concentration of about $50\mu M$, it has a decay time of almost $200ns$. N-acetyl-L-tryptophanamide (NATA) is also regularly used for optical alignment of this instrument. NATA dissolved in deionized water and degassed with nitrous oxide (discussed below), has a measured lifetime of about

$40\mu s$.

2.4 The Sample Preparation

The crux of measuring the transient absorption using this pump-probe spectroscopy lies in the excited triplet state of tryptophan. It is hence necessary to eliminate other detrimental competing quenchers which can be an encumbrance to the success of the experiment in terms of its accuracy. Molecular oxygen, with its unique triplet ground state, is an efficient quencher of the tryptophan triplet state [41]. The succession of UV pulse excitations in the sample generates other detrimental photoproducts like radicals and hydrated electrons that interfere with triplet lifetime measurements. A proposed mechanism for tryptophan triplet decay is via an electron transfer to the sulfur in cysteine side chain [36, 42]. Presence of free radicals can also pose as efficient competers for the triplet quenching through electron transfer. The radicals also have an absorption band which is in resonance with triplet-triplet absorption. A spectral amplitude at $3\mu s$ is attributed to combination of radicals and hydrated electrons [32].

The buffer is therefore treated with nitrous oxide (N_2O) by bubbling it to remove any oxygen in the solvent. This also helps scavenge solvated electrons created by the UV pulse. A $10mm$ sealable cuvette serves well to hold the buffer for it. Protein and peptide samples are then diluted to about 20-30 μM in this buffer. The formation of weak covalent disulfide bonds between the sidechains (the Thiol groups -SH) of two cysteine residues by oxidation also hampers the quenching process. To keep

the cysteines reduced and avoid dimerization, measurements are made in presence of $1mM$ Tris(2-carboxyethyl)phosphine hydrochloride (TCEP). Typically dithiothreitol (DTT) is employed as a reducing agent during protein purification. But since DTT is a competing quencher for cysteine, it is preferable to use TCEP in place of DTT. The experiments are then performed with varying the solvent viscosity. Viscosity of the solution containing sucrose, denaturant(usually guanidinium hydochloride) and the buffer is measured in a temperature controlled cone-cup viscometer(LVDV-II+CP, Brookfield Engineering). The temperature is varied from 0 through 40 degree Celsius.

2.5 Data Acquisition and Analysis

Using the fundamental idea of the experiment shown in figure 2.3, the absorbance profile of the tryptophan triplet at 441nm is monitored and recorded after its excitation using the pulsed laser. A representative trace of this triplet kinetics in shown in figure 2.5. The absorption profile exhibits an exponential decay and the triplet lifetime undergoes a rapid decay with close Van der Waals contact between tryptophan and cysteine.

Using the relationship between the observed triplet lifetime and the microscopic rate coefficients (equation (2.10)), a plot of $1/k_{obs}$ against the solution viscosity (η) at each temperature, the intercept gives the reaction-limited rate, $1/k_R$, and the slope is proportional to the diffusion-limited rate k_{D+} as shown in figure 2.6. For fitting we assume the data can be well modeled by

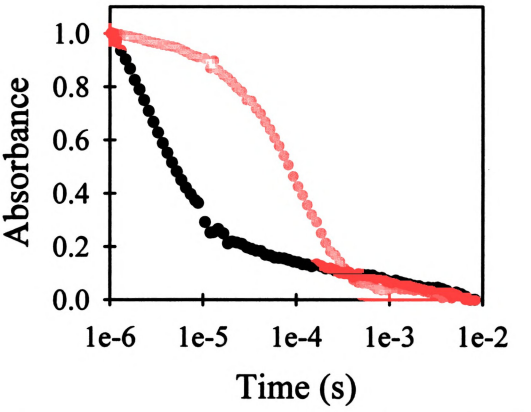


Figure 2.5: Representative Tryptophan Triplet Kinetics for Protein G (T51C). The red points correspond to the kinetics in 2M guanidinium hydrochloride (GdnHCl) and can be fit to a sum of two single exponentials. The faster rate is a contribution from the unfolded fractions of the protein and slower rate from the folded fraction. The black points correspond to the kinetics in 5M GdnHCl and is fit to a single exponential.

$$k_R = k_{R0} \exp\left(\frac{E_0(T - T_0)}{RTT_0}\right) \tag{2.11}$$

and,

$$k_{D+} = \frac{k_{D+0}T}{\eta T_0} \exp(\gamma(T - T_0)) \tag{2.12}$$

where, T is the temperature, η is the solution viscosity, and k_{R0}, E_0, k_{D+0}, and γ are fitting parameters. k_{R0} and k_{D+0} are the values of k_R and k_{D+} at the reference temperature T_0 of 293K.

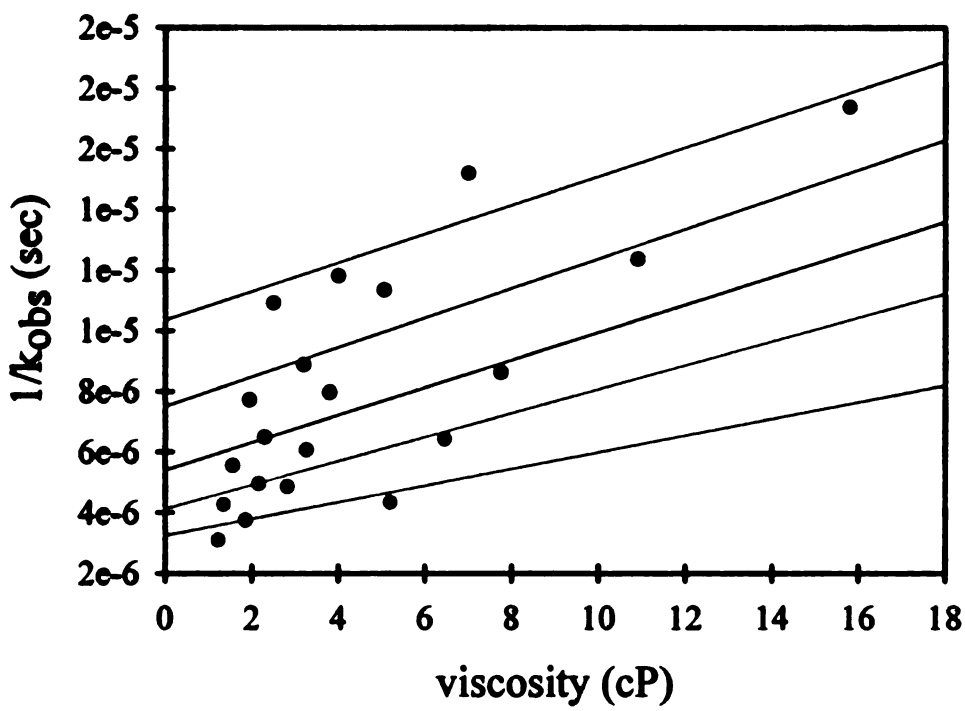


Figure 2.6: Observed Rate fit to equations (2.12) and (2.11). This representative plot is for protein G (E19C) at 0 C (red), 10 C (green), 20 C (blue), 30 C (pink), and 40 C (cyan).

2.6 Reaction and Diffusion Limited Rates

The reaction limited rate is the observed rate measured when diffusion is sufficiently fast and predominates over the quenching rate, $k_{D-} >> q$. The encounter complex is formed and broken multiple times before quenching of the donor occurs. Consequently an equilibrium end-to-end distance distribution is maintained at every instant and hence k_R provides direct information about the distance distribution. The population of molecules forming the encounter complex is in equilibrium with the ensemble of all chain conformations. The reaction limited rate depends only on the quenching rate and the equilibrium end-to-end distance distribution. The quenching rate being well described implies a relationship of direct proportionality between k_R and $P_{eq}(r)$.

$$k_{obs} = \frac{k_{D+}}{k_{D-}} q = K_{eq} q \equiv k_R = \int_a^{l_c} P_{eq}(r) q(r) dr \tag{2.13}$$

For a constant K_{eq} and a well defined q, k_R will be determined solely by $P_{eq}(r)$. This implies, on the quenching timescale, since the encounter complex and the end-to-end distance of rest of the protein ensemble will be in a dynamic equilibrium, the essential shape of $P_{eq}(r)$ will not change with time as illustrated in the figure 2.7. Only the peak of the distribution diminishes with time.

The diffusion limited rate, k_{D+}, is the rate of bringing the ends of the loop together. This is the measured rate for the case when $k_{D-} << q$. The result is formation of short range contact between the donor and acceptor resulting in the encounter complex. The diffusion limited rate depends in a complex way on both

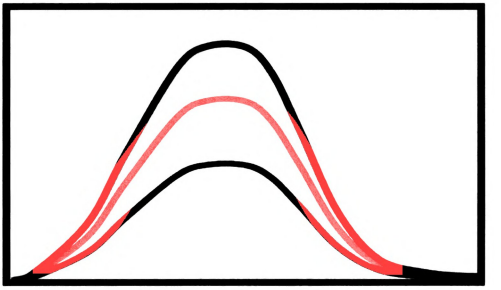


Figure 2.7: Illustration of Reaction Limited Rate: The abscissa shows the end-to-end loop distance. Ordinate gives the probability of unquenched molecules.

on the diffusional dynamics of chain ends and the end-to-end distance distribution. When the diffusion rate-limit predominates, more and more of the encounter complex is formed and the donor immediately quenched since the loop ends cannot diffuse away fast enough. With time, one would notice a asymmetrical rightwards shift in $P_{eq}(r)$ for smaller end-to-end distance regime as shown in figure 2.8 below.

2.7 SSS Theory and Polymer Chain Model

The theory of Szabo, Schulten, and Schulten (SSS) constitutes a description of the dynamics of end-to-end distance of a polymer. It treats polymer dynamics as motion

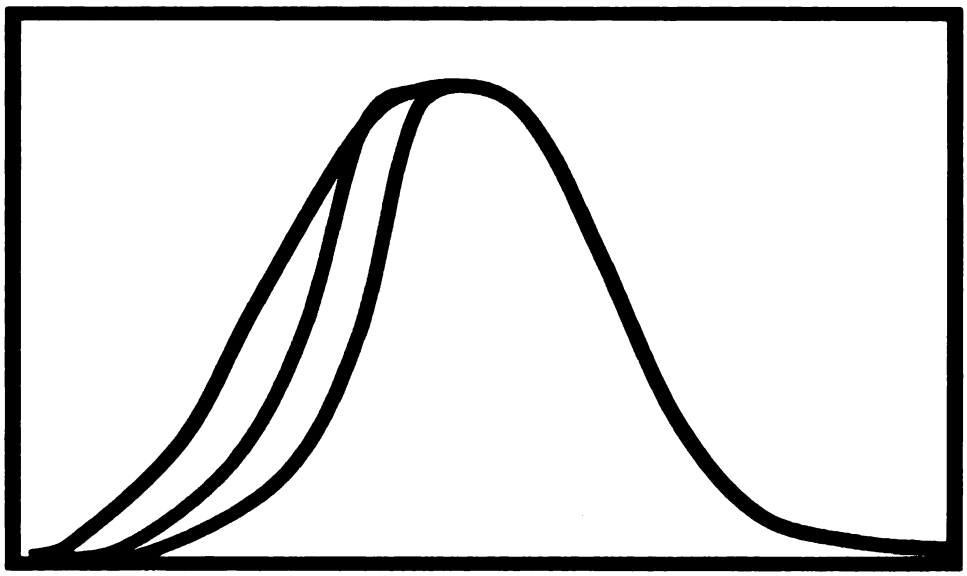


Figure 2.8: Illustration of Diffusion Limited Rate: The abscissa shows the end-to-end loop distance. Ordinate gives the probability of unquenched molecules.

on a one-dimensional potential of mean force corresponding to the equilibrium end-to-end distance distribution $P_{eq}(r)$ [43]. The rate of bringing two reactive ends of a polymer chain in close contact by diffusion under the influence of potential gives an estimate of the first passage time. The observed rate is calculated by solving a Smoluchowski-like diffusion equation for a diffusion process in a field defined by -

$$U(r) = -k_B T \ln P_{eq}(r) \tag{2.14}$$

where k_B is the Boltzmann constant and $P_{eq}(r)$ is the equilibrium end-to-end distance distribution of the loop ends. For a distance dependent quenching rate, the observed rate derived through SSS theory is found to be [38, 43],

$$\frac{1}{k_{obs}} = \frac{1}{k_R} + \frac{1}{k_R^2} \int_0^\infty \langle \delta q(t) \delta q(0) \rangle dt, \tag{2.15}$$

where $\delta q = q(r) - k_R$ and $k_R = \langle q \rangle = \int_a^{lc} P_{eq}(x) q(x) dx$ is the reaction-limited rate. The time integral of the sink-sink correlation function is analytically evaluated to be,

$$\langle \delta q(t) \delta q(0) \rangle = \frac{1}{t_{max} - t} \int_0^{t_{max}-t} \delta q(t_0) \delta q(t + t_0) dt_0, \tag{2.16}$$

where t_{max} is length of the trajectory. For diffusion on a 1D potential, the expression for the observed rate becomes (using equation (2.16)),

$$\frac{1}{k_{obs}} = \frac{1}{k_R} + \frac{1}{k_R^2} \int_a^{lc} (D(x) P_{eq}(x))^{-1} \left[\int_x^{lc} \delta q(y) P_{eq}(y) dy \right]^2 dx \tag{2.17}$$

where the reaction-limited rate is given by,

$$k_R = \int_a^{lc} P_{eq}(x) q(x) dx \tag{2.18}$$

and the second term in equation (2.17) gives the diffusion limited-rate,

$$\frac{1}{k_{D+}} = \frac{1}{k_R^2} \int_a^{l_c} (D(x)P_{eq}(x))^{-1} \left[\int_x^{l_c} \delta q(y) P_{eq}(y) dy \right]^2 dx \tag{2.19}$$

In order to be able to take advantage of the above analytical expression (equation (2.17)) for numerical evaluation, we will need to generate a distribution of the Trp-Cys distances to form $P_{eq}(r)$. Some of the common ways of generating the end-to-end distance distribution are -

a) Gaussian (random-walk) Chain approximations: The connected monomer subunits, having no orientational correlation, are approximated as a random walk with a Gaussian probability.

$$P_{eq}(r) = 4\pi r^2 \left(\frac{2\pi \langle r^2 \rangle}{3} \right)^{-3/2} \exp\left(\frac{-3r^2}{2\langle r^2 \rangle} \right) \tag{2.20}$$

The reaction-limited and diffusion-limited rate for small a is given by,

$$k_R = \frac{4\pi q a}{(2\pi \langle r^2 \rangle)^{3/2}} \exp\left(-\frac{3a^2}{2\langle r^2 \rangle} \right) \tag{2.21}$$

$$k_{D+} = \frac{4\pi D a}{(2\pi \langle r^2 \rangle / 3)^{3/2}} \tag{2.22}$$

b) WormLike Chain: This is the simplest model of a polymer chain that takes

into consideration the stiffness of the chain characterized by Kuhn length (κ) and/or the persistence length (l_p). The chains are made more realistic by invoking the excluded volume effects. Assume a sphere of diameter d_α at each end of the peptide bond. The volume excluded by the backbone and side chain is treated as a hard-sphere interaction. There is no one specific expression that gives the probability distribution. Typically the wormlike chains can be generated using a Monte Carlo algorithm using parameters of a persistence length l_p, and excluded volume diameter d_α [44]. A normalized histogram of Trp-Cys distances is used as $P(r)$.

c) Molecular Dynamics (MD): Using MD simulations to obtain the probability distribution of end-to-end distances tenders the most realistic chain statistics owing to the all atom details and realistic bond chemistry employed in these simulations. The conformational distributions are generated using the high performance computing cluster (HPCC) using AMBER molecular dynamics simulation package. Performing simulations with explicit-solvent is computationally more expensive. Since our goal here is mainly to generate the equilibrium distribution $P(r)$ and not be concerned with the precise kinetics we can perform the simulation using implicit-solvent models. Following the simulation of trajectories, a normalized histogram (0.1 Åbinning) of Trp-Cys distances was used as $P(r)$.

Chapter 3

Unfolded States of Proteins and Peptides: Experimental Investigation

3.1 Introduction

It is now being acknowledged that protein misfolding and aggregation may have its seed in unfolded fractions or the partly structured folding intermediate. Hence, the earliest steps in the protein folding process, such as loop formation, may hold the key to understanding the pathogenesis related to aggregation, misfolding and amyloid formation. Adding another step towards understanding of protein and peptide structure, I present in this chapter some experimental results for molecules with and without known propensities for aggregation.

Protein L and Protein G are two bacterial immunoglobulin-binding α/β proteins that have been widely studied and are known to be relatively aggregation free under physiological conditions. The thermodynamic and kinetic properties of their folding have been addressed and recorded in great detail [45, 46]. This makes them a good target for our experimentation. Both proteins L and G share the same native topology with a central $\alpha - helix$ resting on a four-stranded $\beta - sheet$ composed of N- and C-terminal $\beta - hairpins$. The Trp/Cys triplet quenching method has been used to investigate the unfolded states of two domains of proteins L and G under various concentrations of denaturant guanidinium hydrochloride (GdnHCl) [44]. Our experimental results suggest that under conditions that favor folding, the unfolded fractions of proteins L and G are compact and viscous [44, 47].

The polyglutamine amino acid stretches in numerous proteins are implicated in least nine neurodegenerative diseases like the Huntington and Spinobulbar muscular atrophy, all of which seem to exhibit aggregation *in vivo* [48]. Not much is known about the structure of monomeric polyglutamine peptide sequence. Even the mechanism of its aggregation is poorly understood. The length of glutamine repeat is strongly related to the onset of the disease. The greater the number of glutamines in the sequence, the earlier will be the disease onset. These proteins appear to have a pathological threshold of glutamine repeats and become cytotoxic with aggregation. It has been suggested that the polyglutamine expansion induces a significant conformational distortion in the host protein that initiates the formation of aggregation [49]. Generally any abnormal, misfolded or other extraneous protein in cells is subjected to degradation by proteasomes. The fact that many proteins with polyglu-

tamine sequences beyond the pathological threshold seem to evade this fate suggests that the misfolded structure may inhibit binding to the proteasome. This could be a consequence of very large structural change. To understand the mechanism and correlation between the long glutamine stretches and consequent destabilization of the proteins leading to amyloid fibrillation, it is essential to obtain information on structural properties of polyglutamine. This would provide insights into pathogenesis of polyglutamine related diseases and possible therapeutic development.

In first part of this chapter, I present results of experimental measurements that seek to describe the accessible conformational ensemble of monomeric unstructured polyglutamine. The technique of intramolecular Trp/Cys contact quenching was used to measure intramolecular diffusion and rate of end-to-end contact formation. Modeling the length dependence of contact formation rates with a wormlike chain, we find the persistence length of polyglutamine peptides to be $\sim 13.0\ \mathring{A}$ [50]. This is a rather stiff polymer in comparison to that reported for other peptide sequences and denatured proteins. This polymer "stiffness" possibly renders the polyglutamine sequence an extended (α-helical) conformation that hinders the host protein in adopting its intrinsic native state. Consequently the destable misfolded state forms a nucleus for aggregation and amyloid formation.

3.2 Results for Polyglutamine

Five different lengths of polypeptides of polyglutamine sequences were experimentally studied. They were synthesized using solid-phase synthesis with the sequence,

$KKCQ_nWKK$, with n = 4, 7, 10, 13 and 16. Lengths 4-13 were made by Syn-BioSci (Livermore, CA) and n = 16 was a kind gift from Regina Murphy of the University of Wisconsin. The polyglutamine peptide sequences by themselves are insoluble in most solvents. However, to make them soluble they are often synthesized by introducing lysines at the N- and C- terminus. Following the solubilization and disaggregation method of Chen and Wetzel [51], the synthesized polyglutamine samples were treated with volatile solvents of Trifluoroacetic acid (TFA) and Hexafluoroisopropanol (HFIP). As predicted, the solubility of polyglutamine increased in aqueous buffer and rate of aggregation drastically reduced. In spite of this, the peptide sequence is known to eventually aggregate and form amyloid fibrils at physiological pH. The sample is therefore stored at, and experiments performed in aqueous buffer at pH 3.0.

Experiments were then performed following the protocol discussed earlier (chapter 2). Since these peptides are easily prone to aggregation, all samples were used within 30 minutes after thawing so aggregation is unlikely. However, if aggregation did occur, the kinetics of the tryptophan triplet would exhibit two decays: a fast decay corresponding to the monomeric species that can undergo contact quenching ($\sim 3\mu s$), and a slower one for the aggregated species that cannot undergo intramolecular contact quenching ($\sim 40\mu s$). This longer decay was not observed in the experiment, confirming absence of any aggregation.

Figure 3.1 shows the plot of $1/k_{obs}$ versus viscosity (η) for the various polyglutamine length sequences. Each subplot displays the data set of k_{obs} at five temperatures and four concentrations of sucrose. The results of the fits of observated rates

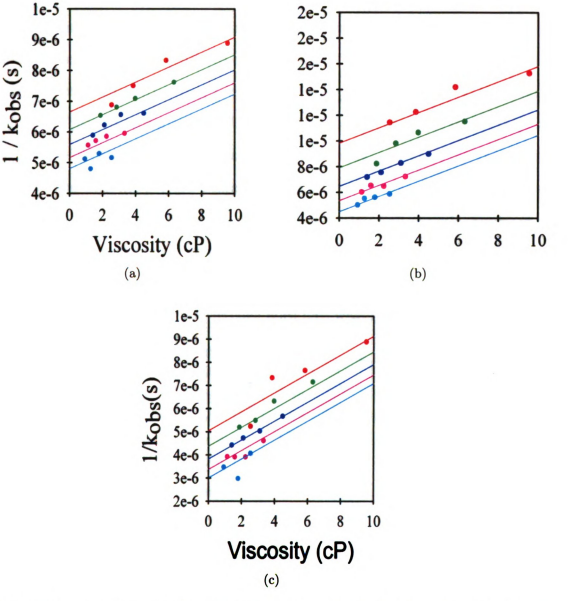


Figure 3.1: Viscosity versus observed tryptophan triplet lifetimes at various temperatures for $KKCQ_nWKK$ where (a) n = 4, (b) n = 10, and (c) n = 7. The lines are the fits to equations (2.11) and (2.12) with the fit parameters given in Table 3.1.

to Equation (2.11) and (2.12) for different length peptides are shown in Table 3.1. However, for fitting purposes the diffusion-limited rate, Equation (2.12), was modified to discard the temperature dependence since empirically, in this work, we do not observe a temperature dependence on the slopes. Consequently, the k_{D+} equation takes the form: $k_{D+}(\eta) = k_{D+0}/\eta$. The fitted k_R and k_{D+} at T = 293 K and η = 1 cP are plotted in Figure 3.2. An interesting feature emerging from this plot is the trend of reaction-limited rate, k_R (the blue points in figure 3.2). It is seen to be increasing with the polyglutamine sequence length, and there is an apparent turn over in this rate at $n \sim 13$. Is this a general property of all peptide sequences approaching this length scale? Or, do the chemical and physical properties of glutamine and possibly other closely associated amino acid residues have some implication in this trend? We will seek to answer such questions with further analysis.

n	$k_{R0}(s^{-1})$	$k_{D+0}(s^{-1})$	E_0
4	1.79×10^5	4.11×10^6	1.38
7	2.61×10^5	2.45×10^6	2.19
10	2.42×10^5	8.04×10^5	4.81
13	2.77×10^5	1.36×10^6	1.23
16	3.13×10^5	2.31×10^5	0

Table 3.1: Polyglutamine Fit Parameters from equations (2.11) and (2.12). Q_{16} was constrained to have $E_0 = 0$ for good convergence of the fit.

Figure 3.3 shows a comparison with previous measurements on polypeptide sequence cys-(ala-gly-gln)$_n$-trp with n = 1 to 9. We notice that the polyglutamine rates are much lower than that for unstructured AGQ peptides. Moreover, the reaction-limited rates in both sequences have an opposing trend. The k_R in AGQ sequences are monotonically decreasing with the sequence length [52]. This opposing

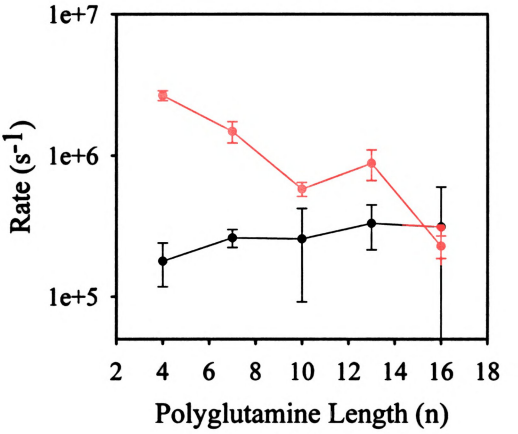


Figure 3.2: Polyglutamine: Reaction-Limited (Blue) and Diffusion-Limited Rates (Red).

trend between the two different polypeptides rates is limited to the reaction-limited rate alone. The nature of diffusion-limited rate is very similar in both, although the absolute values of k_{D+} are lower for the polyglutamine peptides. Since AGQ polypeptides have only 33% of the glutamine content compared to polyglutamine, the former would be expected to have higher a flexibility. This suggests that the turn over of k_R observed in polyglutamine sequence is a special property associated with

this sequence and possibly with closely related (physical and chemical) sequences. Huang and Nau have shown that glutamine is more rigid than most amino acid residues except for His, Arg, Lys, Val, Ile and Pro [53].

3.3 Wormlike Chain Modeling of Polyglutamine

For a more quantitative analysis of the observed rate we use the Szabo, Schulten, and Schulten (SSS) theory which gives the reaction-limited and diffusion-limited rates as equations (2.18) and (2.19) respectively [38, 43]. As is evident from figure 3.3 the length dependence of observed rates cannot be explained with a simple freely jointed chain model. The peptides are therefore modeled as wormlike chains with excluded volume. The one-dimensional end-to-end distance distribution $P_{eq}(r)$, needed for analytical evaluation of k_R and k_{D+} is obtained from the wormlike chains [50].

Figure 3.4 shows the sum of squares of difference between measured and predicted k_R for all five lengths for various values of l_p and d_α. The curve of least-squares is not uniform in all directions and the axis of shallow descent is not along either parameter axis. Therefore, it is difficult to assign an uncorrelated error to either of these parameters. Nevertheless, changing either of these parameters by 10% results in increase of sum of squares of at least a factor of 5 so we chose 0.4 $\AA$ as the error for d_α and 1.3 $\AA$ as the error for l_p.

The AGQ peptides show a reaction-limited rate monotonically decreasing with sequence length, hallmark of a flexible polymer. Contrastingly, for the shortest lengths of polyglutamine peptides, the reaction-limited rate (as also the observed

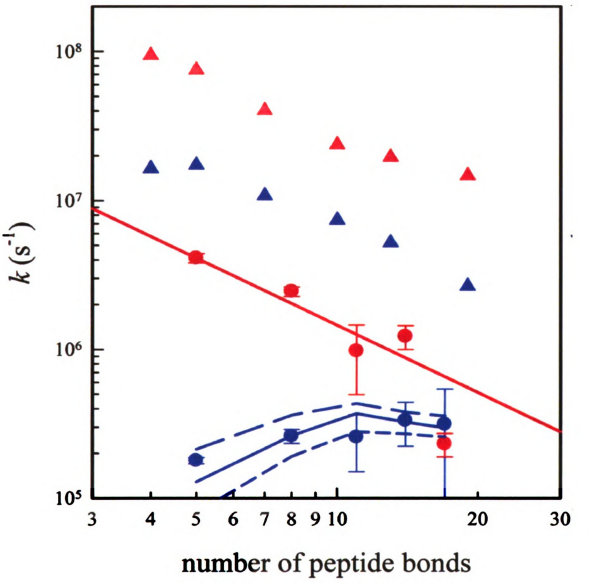


Figure 3.3: Number of peptide bonds between the Trp and Cys (n + 1) versus reaction-limited (blue points) and diffusion-limited (red points) rates for polyglutamine peptides (circles). The triangles are rates for Ala-Gly-Gln peptides in water at pH 7 reported in [52]. The red line is power law fit to $k \sim (n+1)^{-3/2}$. The blue line plots the rates predicted by equation (2.18) using the wormlike chain parameters $l_p = 13$ Å and $d_\alpha = 4$ Å (solid); $l_p = 12$ Å and $d_\alpha = 4$ Å (long dash); and $l_p = 14$ Å and $d_\alpha = 4$ Å (short dash).

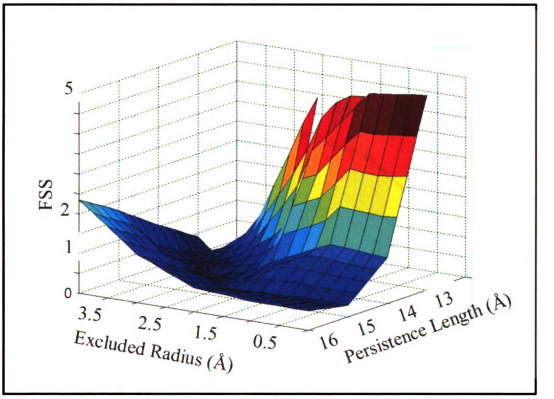


Figure 3.4: Sum of squared difference between the measured k_R and the k_R calculated from Equation (2.18) for all 5 peptides. $P(r)$ was calculated from ensembles of wormlike chains for various values of l_p and d_α.

rate shown in figure 3.5) increases with length rather than decrease. This indicates that the polyglutamine peptides behave more like the flexible peptides (in so far as the trend in the rates), after a threshold length of about 13 residues. Modeling these peptides as wormlike chains with excluded volume suggest that an apparent persistence length for this sequence is ~ 13.0 Å whereas for AGQ, the persistence length was found to be ~ 5.5 Å [52].

For the polyglutamine sequences, although k_R increases with length and begins to turn over at the longest length, k_{D+} decreases monotonically with length and can

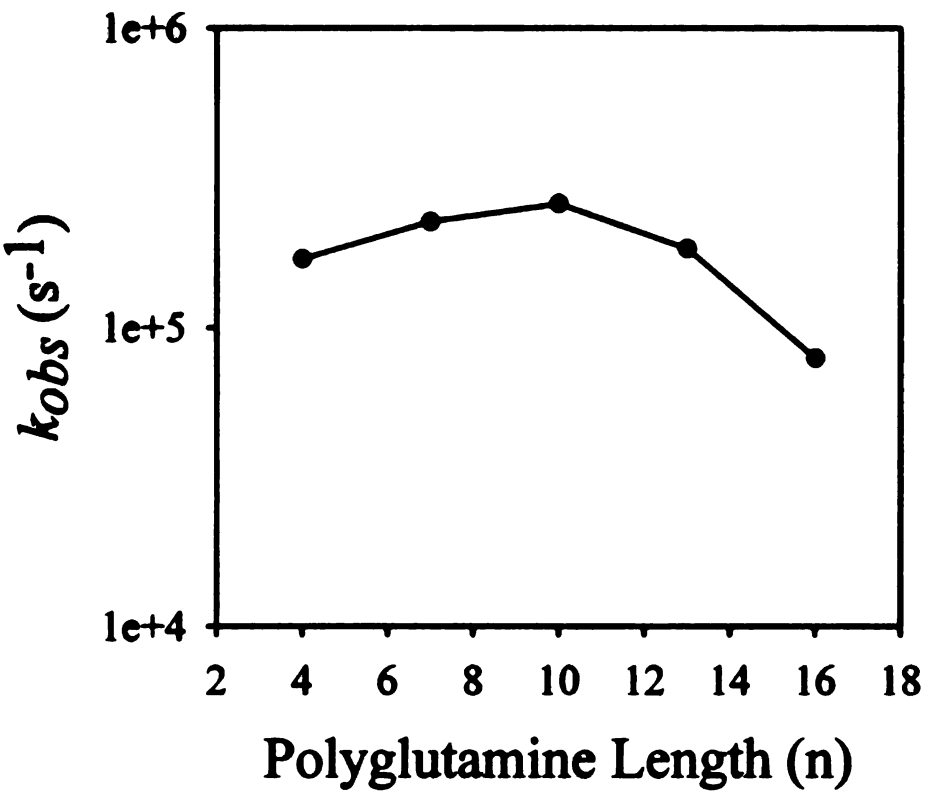


Figure 3.5: Polyglutamine observed kinetics at 20 degree C.

be reasonably fit with a power law dependence $k_{D+} \sim n^{-3/2}$, as would be expected for a flexible peptide [52]. Looking at the mathematical relation between k_R and k_{D+} (equations (2.21) and (2.22)), one would expect, for a fixed (loop independent) diffusion coefficient, a similar length dependence of k_R and k_{D+}. Given that the only free parameters in diffusion-limited rate are D and $P(r)$, it must be that the diffusion coefficient cannot be regarded as loop length independent and must be allowed to change to reflect the trend in k_{D+}. This appears to be a property peculiar to polyglutamine sequences and not observed in AGQ polypeptides. To evaluate the

diffusion coefficient D, we use the experimentally determined value of k_{D+} and the $P(r)$ calculated for each peptide with $l_p = 13.0$ $\AA$ and $d_\alpha = 4.0$ $\AA$, in the diffusion-limited rate equation (2.19). For each peptide length, the diffusion coefficients are shown in Table 3.2. Evidently, there is a substantial decrease in the dynamics of increasing loop lengths as manifested through D.

n	$D \times 10^{-7}(cm^2 s^{-1})$
4	16.6
7	10.5
10	4.3
13	6.6
16	1.5

Table 3.2: Diffusion Coefficients from Equation (2.19) using wormlike chain models with $l_p = 13.0$ $\AA$, $d_\alpha = 4.0$ $\AA$.

There have been a number of computational and structural studies on polyglutamine peptides of multiple stretches. The crystal structure of a Q_{10} inserted in CI2 shows domain swapping within a dimer with the glutamine stretch extended between the two domains, but the structure of the stretch itself could not be determined, indicating it had much conformational flexibility [54, 55]. Various plausible structures have been proposed for monomeric polyglutamine peptides. Initial studies by Chen et.al., [56], suggest peptides with stretches containing 5 to 44 consecutive glutamines to be unstructured. According Perutz et al., [57] shorter polyglutamine repeats have a random coil conformation, whereas longer repeats tend to form β-strand structures. Other simulation studies have suggested a highly compact random coil structure. Using a coarse grain model, Khare et. al., [58] and Marchut et. al., [59] have independently hinted that all polyglutamine chains have a tendency to fold

into a beta-helical structure. Crick and Pappu [60,61], using molecular dynamics at room temperature, show that polyglutamine peptides exhibit existence of partially collapsed states. Armen et. al., [62] have used explicit solvent all-atom molecular dynamics to conclude that polyglutamines of various lengths fold predominately into an alpha-extended chain conformation. Other simulation of this system suggests an extended random coil, rather than α-helix, β-sheet, or PPII structure, best fits the measured thermodynamics with a Flory characteristic ratio of $\sim$3.2 [63]. Thus there has been disparate views on the structure of polyglutamine. Apart from the possibility that a high conformational flexibility and large scale fluctuations makes it hard to experimentally get a handle on the structural and dynamics information, the disparities could arise for other reasons - not incorporating any water interaction with the polypeptide; the force-fields employed in the simulations may not be accurate enough; the water model used may not accurately produce solvent-mediated hydrogen bonded interactions.

Presence of positively charged lysine residues flanking the glutamines may be a reason the contact formation rates are 10-100 times lower for this sequence compared to the same length of AGQ peptides. The lysines are incorporated to achieve better solvent solubility of the polypeptide. A mutual electrostatic repulsion between these charged residues could prevent the formation of close intramolecular contact in polyglutamine. But control experiments on the peptides with up to 170 mM NaCl showed a similar contact formation rate with almost no repression. Therefore, long-range Coulomb interactions do not appear to affect the peptide dynamics. A short-range interactions due to lysine being next to both cysteine and tryptophan could play a

role in slowing of the contact formation rates. But this should affect all lengths of polyglutamine in a homogeneous way. However, the observed length dependence of k_R, which largely determined the ~ 13.0 $\AA$ persistence length, cannot be explained by this short-range interaction.

For the polyglutamine peptides, a single set of wormlike chain parameters (l_p and d_α) were found to simultaneously fit the reaction-limited rates at all lengths. However, the diffusion coefficient varied by a factor of 10 over the length range. In contrast, the diffusion coefficients for the AGQ peptides, found to be typically about $1.5 \times 10^{-6} cm^2 s^{-1}$, varied by only 15%. This suggests that the homogeneous (chemically insensitive) wormlike chain is not a perfect model for the dynamics of polyglutamine. Also, all but the shortest lengths of polyglutamine has a significantly lower D than the AGQ peptides, perhaps because fluctuating residual structure influences the conformational population. One alternative model is to add random β-strand structure to a wormlike chain. To simulate fluctuating β structure within the wormlike chain, individual amino acids (10 links) were randomly assigned a polar angle of 0°. While the inclusion of β-strand orientation on randomly selected residues does decrease k_R in proportion to the relative population of β structure, it cannot produce the observed turnover in length dependence if the underlying wormlike chain has a short (4 $\AA$) persistence length. Another alternative would be to add a β-helical structure. Although not attempted, it could reproduce the turn around in the observed rates, depending on the helical pitch and relative positions of tryptophan and cysteine in the structure. Nonetheless any description of polyglutamine peptides must have a high intrinsic stiffness.

Polyglutamine appears to be much more stiffer than most random peptides or unfolded proteins and is therefore likely significantly stiffer than the rest of the host protein *in vivo*. Small stretches of glutamine may not significantly alter the native conformation, but long stretches could put significant mechanical stress on the folded protein, making it more prone to aggregation from an unfolded state. Wetzel and others have suggested that the aggregation of polyglutamine follows a nucleation-propagation model in which the monomeric nucleus is in rapid and reversible pre-equilibrium with normal monomeric protein [64].

However, recent results by Klein et. al., show that monomeric polyglutamine peptides both above and below the pathogenic threshold show no difference in structure [65]. This suggests that there is no structural transition in different lengths of polyglutamine sequence part of the host protein. The aggregation nucleus may simply be a highly extended conformation which is usually very improbable in a very flexible chain. Thus a pathogenic protein may be one that is destabilized by an intrinsically stiff sequence and then prone to aggregation through bimolecular contact of extended conformations. Huang and Nau have showed that most amino acids were more flexible than glutamine. using a similar contact quenching technique, they measured $k_{D+} \sim 7 \times 10^6 s^{-1}$ for a Q_6 peptide [53]. This is in agreement with our results if one accounts for the fact that the previous peptide had no tails beyond the probe and quencher which decreases the contact rate by about a factor of 3 [52].

In summary, we find the polyglutamine peptide to be much more "rigid" than most peptides. Characterized by the persistence length, the polyglutamine sequences ($l_p \sim 13$ $\AA$) are about 2-3 peptide bonds more stiffer. In the context of amyloid

formation and pathogenesis, this stiffness could lead to an extended conformation between domain ends. Shorter lengths may not alter the native state conformation, but longer glutamine stretches can prevent certain intramolecular interactions and bond formations and leaving the protein deprived of necessary native contacts. This leads to a misfolded and destable protein conformation with propensities for aggregation and consequent cytotoxicity. Could this be the general algorithm of protein aggregation based diseases: higher intrinsic stiffness of at least a part of the protein, formation of extended conformation depriving the native contacts, resulting in misfolded, unstable conformation and aggregation, reduction in proteosome binding, ending up cytotoxic and neurodegenarative? Even if this were to be true, a detailed knowledge of the mechanisms involved would necessitate probing the various phenomena at the atomic scale. Knowing the underlying principles, will then provide insights for development of effective therapeutic agents.

3.4 Proteins L and G

We investigated the nature of the unfolded states of structurally similar but sequentially nonhomologous B1 domains of proteins L and G using end-to-end contact formation measurements. The B1 domain of Protein L is 8 kDa, 63 residues long. To express this protein, the plasmids were transformed into BL21(DE3) *Escherichia coli* cells. On account of its N-terminal hexa-His tag, the protein L was purified by Ni-affinity chromatography. The purified proteins were verified by N-terminal sequencing and MALDI-TOF mass spectrometry. Protein G B1 domain is 6.8 kDa, 56

residues long. After expression of the protein in BL21(DE3) *Escherichia coli* cells, it was purified by anion exchange chromatography [44].

Scalley et.al., used a variety of experimental techniques to conclude that folding kinetics of protein L can be characterized by a single exponential. Thus, a simple two-state model adequately describes the thermodynamics and kinetics. However they also suggest the possibility of a partially collapsed unfolded chain at lower dentaurant concentrations [66]. Measurement of fluorescence quenching by iodide suggests a submillisecond conformational change: a partial chain collapse. Unlike protein L, Park et.al., demonstrated, using continuous-flow fluorescence measurements, that protein G exhibits presence of an intermediate and is not a simple two-state folder [67, 68]. In the backdrop of these observations, we find the unfolded fractions of proteins L and G to be compact and viscous.

3.5 Structure and Stability of Proteins L and G

Both domains (B1) of proteins L and G have a single tryptophan residue near the beginning of the C-terminal hairpin, Trp43 of the wild-type protein G and Trp47 of the well-studied Y47W mutant of the protein L [69]. Both domains have been mutated to have a cysteine in one of two positions (figures 3.6 and 3.7):

1. Near the end of the N-terminal hairpin (K23C and E19C in proteins L and G, respectively), forming a 24-residue loop with the tryptophan; or
2. In the final strand of the C-terminal hairpin (T57C and T51C in protein L and G, respectively), forming a 10- and 8-residue loop.

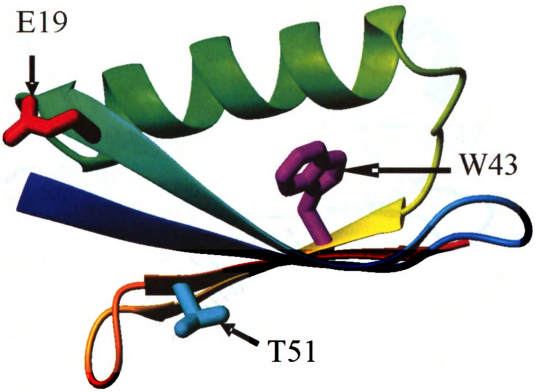


Figure 3.6: Structure and locations of mutations for the Protein G B1 Domain.

The model for contact formation measurement between tryptophan and cysteine is depicted in figure 3.8. The two pertinent amino acid residues are always placed far apart relative to each other in the native state. However, in terms of their sequence separation, they are less than 50 residues apart. Following the optical excitation of the tryptophan triplet states, the decay to ground state is either by a natural process or through contact quenching by cysteine. A Trp/Cys contact formation will be observed only in the denatured or unfolded fractions of the protein. The observed triplet decay is rapid enough that the protein conformations cannot interconvert

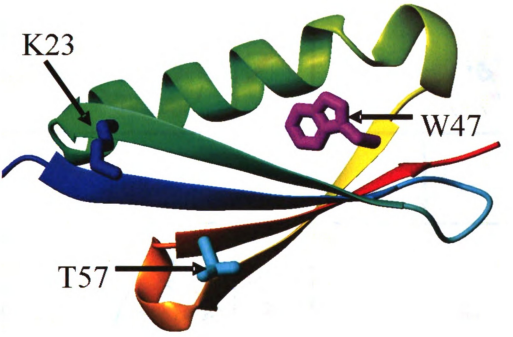


Figure 3.7: Structure and locations of mutations for the Protein L B1 Domain.

between any disordered state and the native folded conformation in the time span of the experiment.

For proteins L and G in intermediate concentrations of denaturant (GdnHCl), two kinetic phases are observed for the tryptophan triplet decay (figure 3.9).

The fast rate is attributed to intramolecular contact with cysteine in the unfolded state and the slow rate to natural decay of tryptophan triplet in the folded or partially folded state. The observed contact formation rates provides stability information from fast phase amplitudes, and structural and kinetics information from life-time

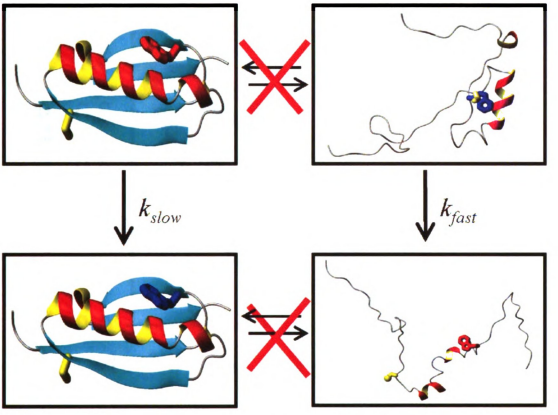


Figure 3.8: Model for contact formation in proteins and peptides. A Trp/Cyc contact quenching is possible only in the denatured ensemble of the protein chain. Under the influence of chemical denaturants the protein is forced to reconfigure and adopt a "disordered" state.

measurements. The two phases observed at intermediate denaturant concentrations are extracted by fitting the observed rate to a sum of two first order decays. The fast phase amplitude provides estimate on the relative population of unfolded fraction as shown in figure 3.10. This plot indicates that the equilibrium stability of both proteins, regardless of mutant, is essentially the same.

These equilibrium unfolding curves are generally in agreement with that reported by Park et al. for protein G [67,68] but are significantly different from those reported

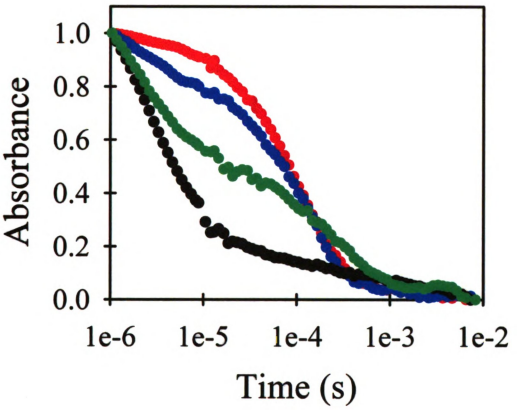


Figure 3.9: Normalized absorbance of the tryptophan triplet state of the T51C mutant as a function of time for various concentrations of GdnHCl. The absorbance decay in solution with 2 M (red), 2.5 M (blue), 3 M (green), and 4 M (dark) GdnHCl.

by Scalley et al. for protein L [66], both of which measured equilibrium fluorescence. The kinetic and equilibrium data for protein L was found by Scalley et al. and Yi et al. to be well described by a two-state folding model [66, 70]. However, folding equilibria determined by a variety of spectroscopic methods yields a wide range of thermodynamic parameters for protein L (Table 3.3), which may mean that a simple two-state transition model is not appropriate for protein L.

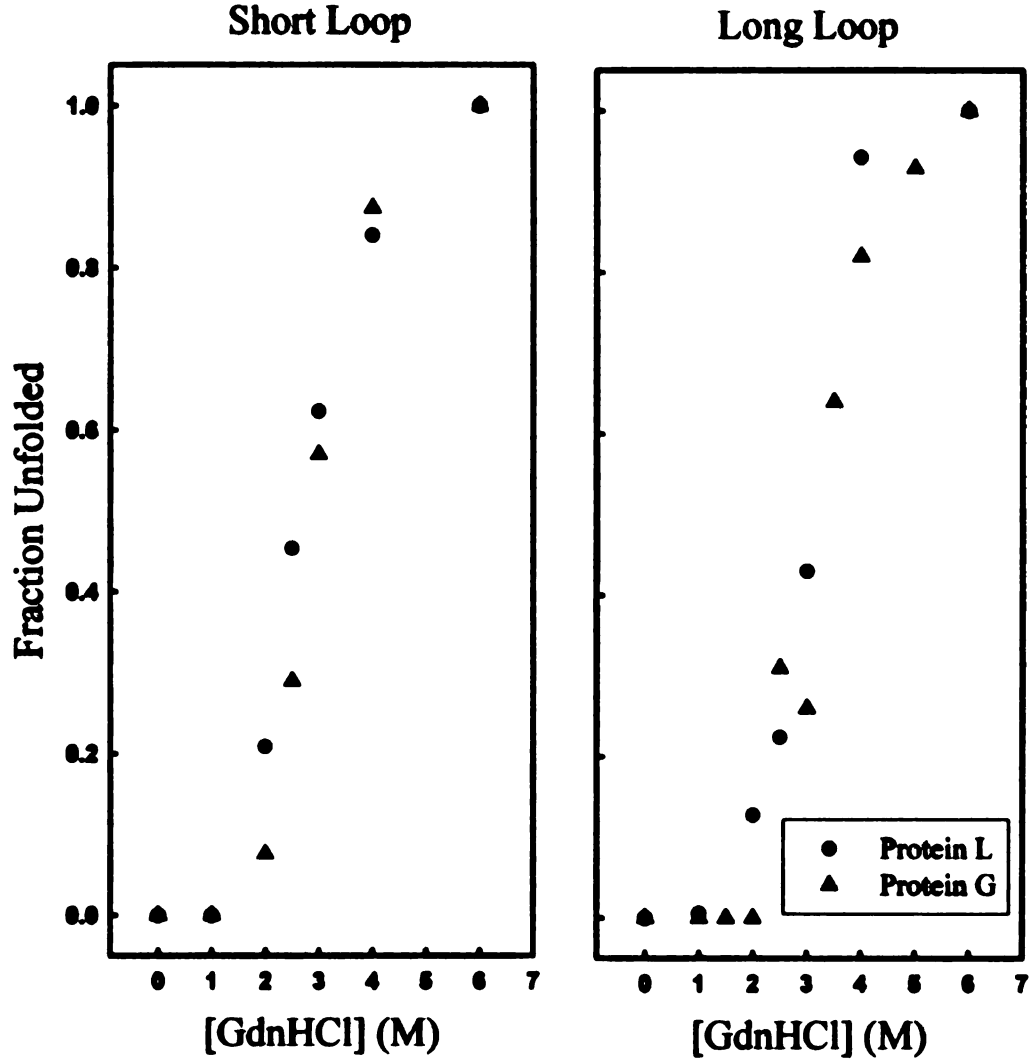


Figure 3.10: Fraction of unfolded molecules as measured by the relative fast phase amplitudes. The equilibrium stability of both proteins and all mutants appear to be essentially the same. The red points correspond to the protein L data and dark points to protein G.

3.6 Contact Formation Kinetics

Figure 3.11 shows the fast and slow rates of both proteins grouped by loop length.

For protein G, the slow rate is similar at all concentrations of denaturant signify-

type	method	$\Delta G(kcal/mol)$	$m(kcal/mol)$
WT	Fluor, $\lambda_{ex} = 280, \lambda_{em} = 320$[a]	4.6	1.85
WT	Fluor, $\lambda_{ex} = 280, \lambda_{em} = 334$	3.86 ± 1.01	1.73 ± 0.44
WT	Fluor, $\lambda_{ex} = 297, \lambda_{em} = 334$	1.26 ± 0.66	0.69 ± 0.26
WT	Fluor, max wavelength[b]	5.15 ± 1.23	1.78 ± 0.43
WT	CD, $\lambda_{ex} = 220$	2.24 ± 0.42	0.88 ± 0.16
T57C	Fluor, $\lambda_{ex} = 297, \lambda_{em} = 334$	1.18 ± 0.95	0.82 ± 0.41
T57C	Fluor, max wavelength	5.13 ± 1.01	2.12 ± 0.41
T57C	CD, $\lambda_{ex} = 220$	1.75 ± 0.33	1.02 ± 0.14
T57C	intramolecular contact[c]	2.61 ± 0.33	0.96 ± 0.12

Table 3.3: Thermodynamic Parameters for protein L Wildtype (Trp only) and T57C Mutant. [a]All measurements were made at room temperature in 0.1 M potassium phosphate buffer (pH 7). The difference between exciting fluorescence at 280 and 297 nm is that the 297 nm excites only tryptophan while 280 nm also excites tyrosine emission. [b]The peak fluorescence intensity wavelength, $\lambda_{ex} = 297nm$. [c]Data from Figure 3.10 (triangles).

ing that the tryptophan remains hydrophobically buried until the protein unfolds completely. However, for protein L, the slow rate decreases from ~22000 s^{-1} to ~5000 s^{-1} in an apparently cooperative transition. This is observed in both the protein L mutants and in the control protein, Trp-only protein L (black circles in figure 3.11). The higher protein L rate is approximately equal to that measured for N-Acetyltryptophan amide in water whereas the lower rate is consistent with rates for tryptophan triplet measured in the hydrophobic core of a protein [44]. Neither rate is fast enough to reflect an intramolecular contact.

These observations suggests that proteins L and G each have a structurally different "intermediates". In architecture of protein G, formation of the second β-hairpin is the first step in its folding event [68]. Hairpin 2 forms a stable intermediate that effectively buries the tryptophan at position 43. Formation of hairpin 1 is the rate-

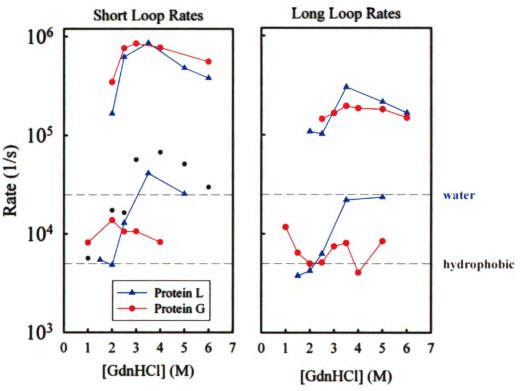


Figure 3.11: Fast and slow triplet decay rates measured at various concentrations of GdnHCl. Graph on the left correspond to the short loop mutants (T57C and T51C), and graph on the right is for the long loop mutants (K23C and E19C). The dark points are the measured tryptophan triplet rates in the control protein L Trp-only mutant, which has no cysteine.

limiting step in this well characterized sequence of folding events. Since the W43 on hairpin 2 remains hydrophobically buried until the complete unfolding of the protein, triplet contact quenching is insensitive to the rate limiting step and does not capture the kinetic intermediate. In contrast, hairpin 1 forms first in protein L. This results in a compact, partially folded intermediate and leaves the tryptophan at position 47 on hairpin 2 of protein L solvent exposed. The rate limiting step here, is formation of hairpin 2 which hydrophobically buries the tryptophan eventually. This kinetic

intermediate in protein L is therefore captured as a decrease in the slow rate from $\sim$22000 s^{-1} to $\sim$5000 s^{-1}. Could this be an indicator of the partially collapsed unfolded chain, hinted by fluorescence-quenching experiments with sodium iodide by Scalley et.al. [66]? A definitive answer to questions such as this can be obtained by closer inspection of the equilibrium data and well resolved kinetic experiments. A detailed analysis of thermodynamics and kinetics of unfolding/refolding of protein L will expose any stabilized intermediates that may have significant accumulation during folding.

Further evidence for deviation of protein L folding kinetics from a simple two-state model is obtained from microfluidic ultrarapid mixer experiments performed in this lab. Improved mixing techniques have enabled us to unravel the fast folding events and phases that previously remained obscured in the long (>1 ms) dead time of the stopped-flow instrument. Kinetic measurements using FRET and tryptophan fluorescence on a 2-4 μs mixing time scale discerns, in addition to a slow phase, a faster phase of about 50 μs [47]. Presence of two phases in the folding kinetic of protein L suggests that a simple two-state model may not offer a complete description of the folding characteristics. The dynamics of the unfolded state may be marked by multiple energy barriers and hence a multidimensional energy landscape is necessary to describe the folding of protein L. Figure 3.12 shows a conceptual representation of such a folding energy landscape. Between the unfolded and the native state basins, there exists a large energy barrier. The fast phase of 50 μs, corresponds to energy roughness fluctuations of the order of 1 kcal/mol and also corresponds to the medium phase triplet rates at moderate [GdnHCl]. From figure 3.12, it appears

that denatured or unfolded states close to points such as "a" will tend to exhibit a cooperative folding transition while those near the region marked "unfolded" could show a more complex behavior deviating from two-state model.

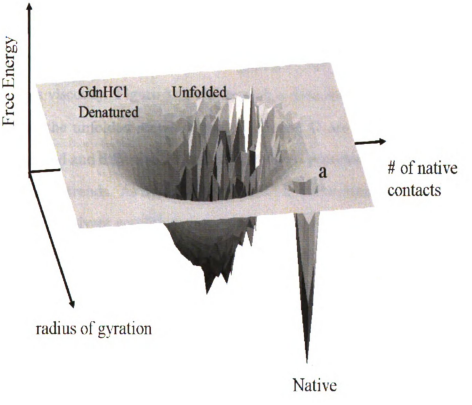


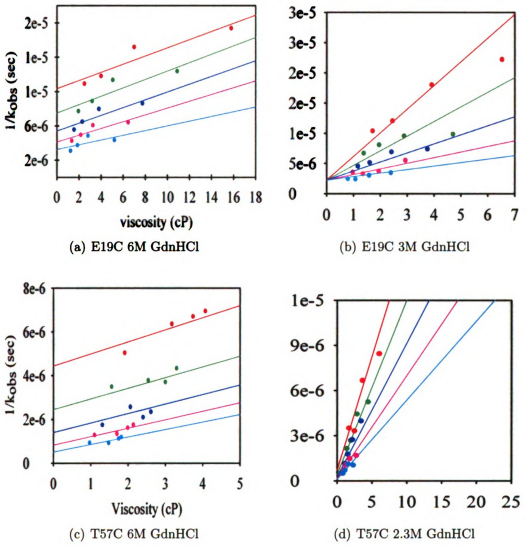
Figure 3.12: Energy landscape representation under the final folding conditions. The unfolded basin reflects the roughness that slows down the relaxations and intramolecular diffusion.

3.7 Dynamics of Proteins L and G

The contact formation (fast) rate of the unfolded state for both proteins increases slightly with decreasing denaturant concentration but then decreases significantly below 3 M GdnHCl (fast rates in Figure 3.11). This contrasts sharply with similar measurements of the cold shock protein of *Thermotoga maritime* in which the fast rate increased monotonically as denaturant decreased to at least 1 M GdnHCl [71]. From the measurements made at varying viscosities and temperatures, the reaction-limited rate, $k_R(T)$, and diffusion-limited rate, $k_{D+}(\eta, T)$ are extracted through a plot of $1/k_{obs}$ versus viscosity. Figure 3.13 shows such a data set.

The dynamics of the unfolded states of proteins L and G are remarkably similar. The reaction-limited and diffusion-limited rates for both proteins and all mutants appear to have opposing trends. As depicted in figure 3.14, as the denaturant concentration is decreased and solvent conditions become favorable for folding, the reaction limited rate increases and the diffusion-limited rate decreases. At 6 M GdnHCl, the diffusion-limited rate is 2-5 times faster than the reaction-limited rates for all mutants. Comparing it to the AGQ peptides under the same conditions, the measured rates are found to be very similar. To understand this behavior of the reaction-limited and diffusion-limited rates of proteins L and G quantitatively, we employ Szabo, Schulten, and Schulten (SSS) theory and model the proteins as wormlike chain with excluded volume. k_R and k_{D+}, according to SSS theory are given by equations (2.18) and (2.19) respectively.

To give a qualitative description of the observed rates, we can model the unfolded protein as a Gaussian chain. The reaction-limited and diffusion-limited rate is in-



(a) E19C 6M GdnHCl

(b) E19C 3M GdnHCl

(c) T57C 6M GdnHCl

(d) T57C 2.3M GdnHCl

Figure 3.13: Temperature and viscosity dependence of observed quenching rates of protein L T57C at 6 M GdnHCl. Lines are fits to equations (2.11), and (2.12). The y-intercept gives the reaction-limited rate $1/k_R$ and slope gives $1/\eta k_{D+}$. For the lowest [GdnHCl] in which the unfolded state was observed, the temperature dependence of k_R disappeared. Errors of these measurements are typically less than 10%.

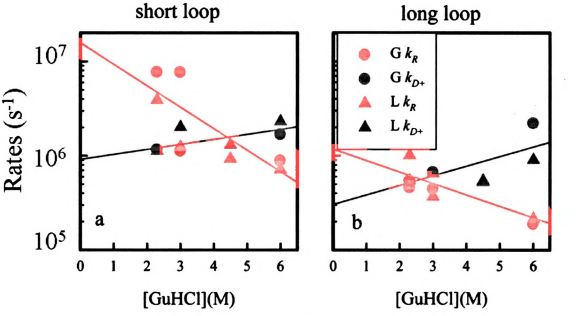


Figure 3.14: Reaction-limited and diffusion-limited rates for the short loop (left panel) and long loop (right panel) in proteins L and G determined by equations (2.11), and (2.12). Based on the sum of squares of the fit, the error of these rates is typically less than 10%.

versely proportional to the chain volume (equations (2.21) and (2.22)). The k_R being larger at the lower denaturant concentrations implies that the chain volume is lower in conditions that favor folding. The ratio of the measured reaction-limited rates at two denaturant concentrations gives an approximate change in the volume of the unfolded chain. For the long loops, E19C and K23C, the ratio of $k_R(2.3M)/k_R(6M)$ is 2.4 and 4.7 respectively. This means that the chain volume decreases by a factor of 2 (4) for the loop E19C (K23C) at 2.3 M GdnHCl as compared to that at 6M GdnHCl. Since we are measuring the unfolded fractions, this denotes a compact unfolded state under folding conditions. The diffusion-limited rate k_{D+}, is observed to be decreasing with denaturant concentration. For the mathematical consistency,

an observed decrease in k_{D+} and a concomitant decrease in average chain volume must correspond to a significant decrease in the diffusion coefficient D. Taking the ratio $(k_R/k_{D+})_{2.3M} \times (k_{D+}/k_R)_{6M} \equiv D_{6M}/D_{2.3M}$ we find that the effective diffusion coefficient decreases by 8-10-fold for the long loops and 15-30-fold for the short loops. This suggests that the unfolded fractions of the protein have lower rate of intramolecular diffusion, i.e., the chain is more viscous under folding conditions.

To further investigate the polymer dynamics of the unfolded state under folding conditions, we modeled the protein as a wormlike chain with excluded volume and use Szabo, Schulten, and Schulten (SSS) theory to estimate the effective persistence length and intramolecular diffusion constant at various concentrations of GdnHCl. We assume that the persistence length, l_p, is an intrinsic property of the chain and does not change with solvent conditions, but that the excluded volume diameter, d_α, and diffusion constant, D, depend on solvent and intramolecular interactions and therefore depend on the concentration of denaturant. For each loop length (8, 10, and 24 residues), 2×10^6 wormlike chains were generated for each persistence and contour length. The simulated tails were always set to 5 residues since Buscaglia et al. found that reaction limited rates were depressed by about a factor of 1.7 for each tail due to excluded volume and that the effect was very insensitive to the length of the tail [52]. A normalized histogram (0.1 $\AA$ binning) of Trp-Cys distances was used as $P(r)$ in equations (2.18) and (2.19). The $P(r)$ in conjunction with equation (2.18) was used to calculate the reaction-limited rate for a variety of persistence lengths and excluded volume diameters and compared to the measured rates at 6 M GdnHCl. We observed that the calculated rates were much more sensitive to d_α than to l_p.

Therefore the persistence length was kept fixed at 4 $\AA$ as was found for unstructured AGQ peptides [52], and the excluded volume radius was allowed to vary. As shown in Table 3.4, for each mutant at 6 M GdnHCl, the best fits of d_α are very close to 4 $\AA$, indicating that the polymer properties of the fully denatured proteins are very similar to the completely unstructured peptides from [52] in 6 M GdnHCl. Chains were then generated for smaller d_α to match the measured reaction-limited rates at 2.3 M GdnHCl (see Table 3.4). Column 6 of Table 3.4 gives the root mean squared Trp-Cys distance calculated using these probability distributions. Between 6 and 2.3 M GdnHCl, the average Trp-Cys distance decreases by 10 percent at lower denaturant concentration.

mutant	n	$l_p(\AA)$	[GdnHCl(M)	$d_\alpha(\AA)$	$\langle r^2 \rangle^{1/2}(\AA)$	calc $k_R(s^{-1})$	meas $k_R(s^{-1})$
T51C	8	4.0	6.0	4.05	16.8	8.0×10^5	8.8×10^5
			2.3	2.0	15.1	8.0×10^6	7.7×10^6
T57C	10	4.0	6.0	3.9	19.2	7.4×10^5	7.2×10^5
			2.3	2.3	17.5	4.3×10^6	3.9×10^6
E19C	24	4.0	6.0	3.9	32.4	1.8×10^5	1.9×10^5
			2.3	3.3	30.7	4.0×10^5	4.6×10^5
K23C	24	4.0	6.0	3.8	32.1	2.0×10^5	2.1×10^5
			2.3	2.3	28.6	1.1×10^6	1.0×10^5

Table 3.4: Parameters Used in Wormlike Chain Simulations and the Resulting k_R Calculated Using Equation (2.18). The persistence length l_p was kept fixed at 4 $\AA$.

Using equation (2.19), the effective diffusion coefficient, D, was calculated from the wormlike chain probability distributions given in Table 3.4 and the measured diffusion-limited rates. These values are given in Table 3.5. They vary quite substantially by mutant and, except for K23C, are significantly lower than that reported for unstructured AGQ polypeptides in 6 M GdnHCl. We conclude that the diffusion coefficient is a local property of the loop sequence and support this claim by two observations. First, the sequences measured by Buscaglia et al. are 33 % glycine, so intramolecular diffusion should be faster than the sequences in this study. Second, the diffusion coefficients for the T51C and T57C loops, which form $\beta - strands$, are much lower than for the K23C and E19C loops, which form mostly $\alpha - helix$; this difference likely reflects the propensity for extended structure in the $\beta - strand$ sequences. From table 3.5, the calculated effective diffusion coefficients in 2.3 M GdnHCl for all mutants decrease by about a factor of 6 relative to 6 M GdnHCl. This uniformity of scaling for each mutant suggests that this trend is a global property of the chain. Thus, the loss of denaturant in a real protein yields an unfolded state that is less diffusive than the fully denatured state and reflects transient interactions throughout the entire chain. A 6-fold decrease in D represents a very significant change in the internal dynamics of the proteins. This decrease in D is completely different than the trend observed by Buscaglia et al. on unstructured AGQ peptides in which the coefficient increases by about a factor of two between 6 and 0 M GdnHCl [52]. We attribute this qualitatively different behavior to the fact that the peptides measured previously were completely unstructured and contained no hydrophobic residues.

protein	mutant	[GdnHCl(M)	meas $k_{D+}(s^{-1})(\eta = 1cP)$	$D \times 10^6(cm^2s^{-1})$	$D_{6M}/D_{2.3M}$
G	T51C	6	1.7×10^6	0.17	5.7
		2.3	1.2×10^6	0.03	
L	T57C	6	2.3×10^6	0.33	6.6
		2.3	1.1×10^6	0.05	
G	E19C	6	2.2×10^6	1.5	6.8
		2.3	5.4×10^5	0.22	
L	K23C	6	9.1×10^5	0.57	5.2
		2.3	5.4×10^5	0.11	

Table 3.5: Diffusion Coefficients for the loops in proteins L and G calculated from Equation (2.19) using the Wormlike Chain Probability Distributions given in Table 3.4.

The intramolecular diffusion rate of unstructured peptides has been used as a measure of the protein folding "speed limit" [72]. Low intramolecular diffusivity of the unfolded state ultimately limits the folding process. The D values measured in this work under folding conditions are up to 20 times lower than measured for unstructured AGQ peptides. Our extrapolation of k_{D+} to 0 M GdnHCl in figure 3.14 indicates that this internal friction time is on the order of microseconds, in good agreement with measurements by Hagen et al. of fast folding protein rates as a function of viscosity and temperature [73]. This low diffusivity could arise through formation of backbone-backbone interaction (hydrogen bonds). A decrease in the diffusion coefficient has been observed by other research groups too. Using single molecule fluorescence correlation spectroscopy, a two-fold decrease in D from 6 M to 2.3 M GdnHCl was reported for the unfolded state of the cold shock protein [74]. Using bulk FRET on Gly-Ser unstructured peptides, the average end-to-end distance and the diffusion coefficient seemed to be lower in water as compared to high levels of denaturant [75]. Except for the Ala-Gly-Gln measurements of Buscaglia et al., it does appear that presence of hydrophobes have an ubiquitous effect of lowering the diffusion coefficient and also the average end-to-end distance in conditions that favor folding. This degree of compaction and the extent of diffusion coefficient lowering varies from sequence to sequence and will be expected to also depend on solution conditions such as temperature, nature of denaturant chemical used, solvent pH conditions, solution viscosity and molecular crowding.

In conclusion, we have investigated the contact formation rates in amyloidogenic peptides - polyglutamine - as well as proteins that are not very prone to aggrega-

tion (proteins L and G). While the proteins that do not easily aggregate have an unfolded state fraction that is relatively more collapsed and less diffusive compared to the typical denatured states, the same is not necessarily true for the aggregation prone peptides. A further detailed atomic level characterization can be obtained using molecular modeling and simulations. The unfolded fractions of protein can be observed at higher temporal and structural resolution, generally inaccessible to experiments. This could lay a strong foundation to directly relate experimental data and simulation results.

Chapter 4

Protein L Simulations and Experiment

4.1 Introduction

Molecular dynamics (MD) is a computational technique for modeling time evolution of a molecular system. The central governing equation of motion is provided by Newtonian mechanics and statistical mechanics. Each atom in the system experiences a net force due to the potential created by all other atoms. The direction of motion is thus determined and atomic motions simulated as a function of time by integrating the Newton's equations of motion. Trajectories of all the atoms are saved as the molecular system progresses in time.

In the previous chapter we observed the significance of characterizing the unfolded states of a protein to understand its folding landscape. A number of experimental and computational tools have helped better our understanding of protein folding. The various experimental techniques provide a wealth of information (structural, dynamic, kinetic and thermodynamic) on a macroscopic length and time scales.

Molecular dynamics simulations complement the experiments by providing the information at atomic scales too. In spite of numerous advances in techniques and computational throughput, there is no settled doctrine on the pathways and mechanisms of protein folding or the exact nature of the folding landscape. Lack of better connectivity between experimental results and computational modeling is one reason there is still no comprehensive understanding of the problem of protein folding and misfolding. In this chapter we present a comparison of experimental results with simulation data to characterize the structure and dynamics of unfolded states of protein.

Experimentally, we observe the unfolded states in proteins L and G to be more viscous and compact under folding conditions [44]. Using a wormlike chain model with these results, the diffusion coefficients in both proteins appear to decrease by a factor of about 6 at 2.3 M GdnHCl as compared to that at 6 M GdnHCl. Also, the average end-to-end distance decreases by almost 10% for both the proteins at lower denaturant (GdnHCl) concentration exhibiting the dynamic coil-to-globule transition. Identification of the primary contributors to this attenuated dynamics is necessary to break the protein folding code. For a better understanding of the molecular basis and origin of the observed nature of unfolded state, we compute the sampling of the configuration space using molecular dynamics simulations and compare it to experimental results. In the following sections we present a study of the B1 domain of protein L using molecular simulations and Trp/Cys contact quenching experiments to characterize its folding dynamics and unfolded states. The simulations reveal a coil-to-globule transition with decreasing temperature that closely resembles the

results found with decreasing GdnHCl concentration.

4.2 Methods: Simulations and Experiment

A typical MD simulation using AMBER molecular dynamics package involves the following:

1. Obtain or create the starting structure (typically a PDB structure file).
2. From the standardized PDB file, create the topology and parameters files.
3. Perform an initial energy minimization to remove steric hindrances and find the local energy minimum.
4. Perform an optional equilibration to overcome low energy barriers and escape the local minima.
5. Do the final production run for the protein to sample the thermally and kinetically accessible configuration space.

The starting extended structure for protein L can be constructed using the molecular visualization program Pymol. Generally, the starting structures do not have hydrogens atoms in them. However, AMBER offers that functionality through the program *tleap*. All our initial simulations are done with AMBER9 Molecular Dynamics package. For the protein representation, we have adopted the AMBER ff99 force field [76] in all the runs. The solvation effect was modeled by choosing the standard pairwise Born continuum solvent (igb=1) [77]. The MD simulations were carried out at 9 different temperatures: 250, 300, 350, 400, 500, 600, 700, 800, and 1000K. For all simulations, I started with the fully extended initial conformation of

the 63 residue protein L mutant T57C. This was then energy minimized with an initial minimization of 500 steps. This energy minimized state formed the starting structure for all our further production runs. Initial velocities are assigned randomly from a Maxwell-Boltzman distribution. To maintain the temperature of our system, we used the Langevin thermostat with a collision frequency of 1 ps^{-1}. The time step of our simulation was set to 1fs. Coordinates of the entire system is saved every picosecond. The Trp-Cys end-to-end distance was obtained as the distance between the sulfur atom of cysteine side chain and the center of mass of tryptophan. The production run was for 40 ns at all the temperatures except 300K. We performed 70 independent 10 ns simulation runs at 300K. This makes for a total accumulated time of 0.94 μs.

For temperatures below 500K, the simulation trajectories do not appear to converge to a stationary state (discussed below) even in about 40 ns. Therefore, to achieve a more accurate description further runs were performed in collaboration with Folding@Home research group at Stanford. Longer trajectories were obtained using graphical processing units (GPU) accelerated GROMACS. For the protein representation, they have used the AMBER ff96 and ff03 forcefields with the generalized Born/surface area (GBSA) implicit solvent model of Onufriev, Bashford and Case (igb=5) [78]. Simulations were performed at temperatures of 300K, 330K, 370K and 450K. The simulations produced multiple independent trajectories of 10 microseconds starting from native, extended and random coil conformations.

For a direct comparison between experiments and simulations, we performed further Trp/Cys contact quenching experiments with a destabilized protein L mu-

tant: F22A, K23C. A substitution of the hydrophobic core residue of phenylalanine with alanine considerably destabilizes the protein. This makes for a better model for studying the nature of unfolded protein segments using both molecular dynamics and experiments since various conformations from native to denatured states can now be accessed and studied using minimal inclusion of non-physical entities (chemical denaturants). This destabilization results in unfolded state ensembles for denaturant concentrations even below 0.5M GdnHCl. Figure 4.1 shows the normalized observed tryptophan triplet kinetics in 6M denaturant (green) and 1M GdnHcl (red) at 20 degree C. The observed rates appear to be very similar at this viscosity and denaturant concentrations and can both be fit to first order decays. This shows that the mutation F22A does in fact destabilize the protein because we would have normally seen, at 1M GdnHCl, a slow rate corresponding to natural tryptophan decay; we now observe a fast decay signifying an intramolecular contact. Figure 4.2 shows a plot of $1/k_{obs}$ against viscosity for protein L F22A, K23C in 1M GdnHCl. As discussed earlier, the slope in figure 4.2 gives $1/\eta k_{D+}$ and y-intercept gives $1/k_R$.

4.3 Results and Discussion

By comparing the experimentally obtained parameters for protein L with that of the computed ones from MD trajectories, we can quantify and benchmark the accuracy of the simulations and also characterize the nature of unfolded proteins at atomic level. We also intend to check the effectiveness of the MD simulations in sampling of the configuration space and the degree to which the sampling is realistic. We observe

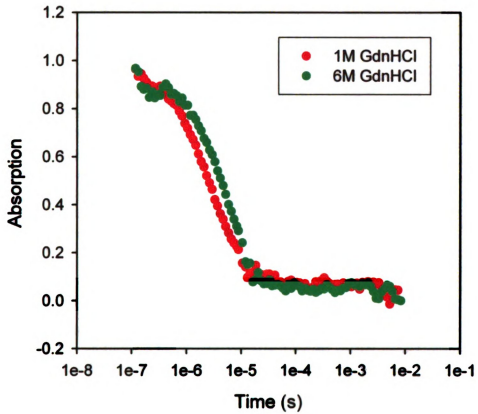


Figure 4.1: Protein L F22A, K23C, W47 Observed Tryptophan Triplet Kinetics: A fast decay even at 1M GdnHcl (red points) caused by significant protein destabilization as a consequence of deleting a hydrophobic core residue.

that at higher temperatures, the end-to-end distance distribution from simulation trajectories can be fit well to a Gaussian model indicating that the protein behaves like an ideal freely-jointed chain at higher temperatures.

Figure 4.3 shows a plot of the root mean square deviation (RMSD) of the backbone C_α atoms at 400K with respect to various conformational states with the MD performed on high performance computing cluster (HPCC) using AMBER9. There

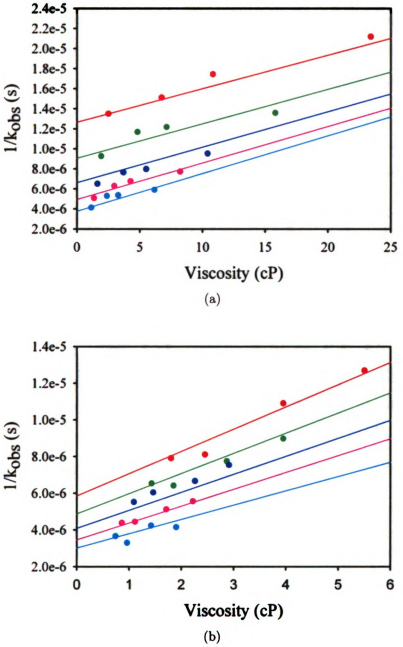


Figure 4.2: Temperature and viscosity dependence of quenching rate for protein L F22A, K23C, W47 mutant at 0 C (red), 10 C (green), 20 C (blue), 30 C (pink), and 40 C (cyan). (a) Protein L F22A K23C in 6M GdnHCl, (b) Protein L F22A K23C in 1.5M GdnHCl

is no indication of relaxation through multiple local minima as would be indicated by distinct jumps representing spatial deviation between the structures in time with respect to the reference structure [79]. This is yet another indication of limited and narrow conformation sampling for simulations below 500K, the protein remains stuck in local energy minima even till about 40 ns.

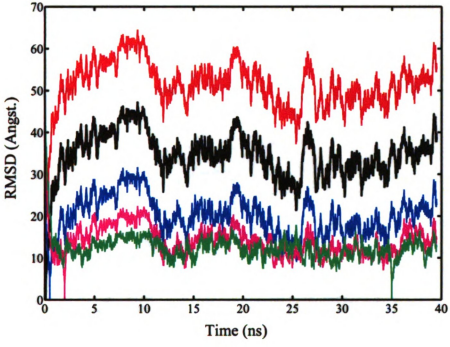


Figure 4.3: Time evolution of backbone (C_α) RMSD with respect to various conformations for protein L simulations at 400K. The deviation is obtained with respect to conformations at: starting extended structure(red), 0.1ns (black), 0.5ns (blue), 2ns (pink), and 35ns (green).

Figure 4.4 shows the probability distribution of W47-T57C distances for temperatures from 400K to 1000K. The distribution gets broader at higher temperatures. Figure 4.5 shows the autocorrelation function at 800K for various parameter mea-

surements all of which seem to exhibit exponential relaxation kinetics. The estimated distribution of relaxation time constants was found to vary from 7 ps for the dihedral angle $\phi(6)$ to about 0.30 ns for the radius of gyration. The global chain relaxation will be a convolution of different local relaxations. At very large simulation times, all the different relaxation rates should collapse into a single global rate constant. Figure 4.6 shows the relaxation time as estimated from the autocorrelation of radius of gyration measurements at various temperatures.

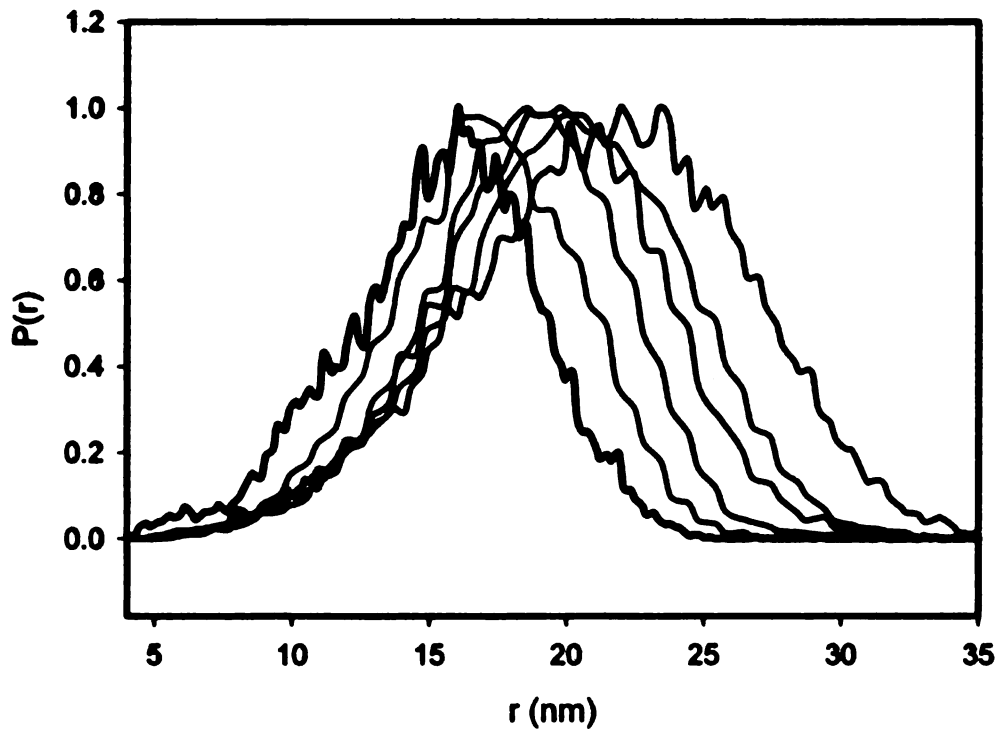


Figure 4.4: Protein L T57C probability distribution of W47-T57C distances from 400K (extreme left, black) to 1000K (extreme right, dark red).

For temperatures below 500K, we obtained narrower sampling distributions in-

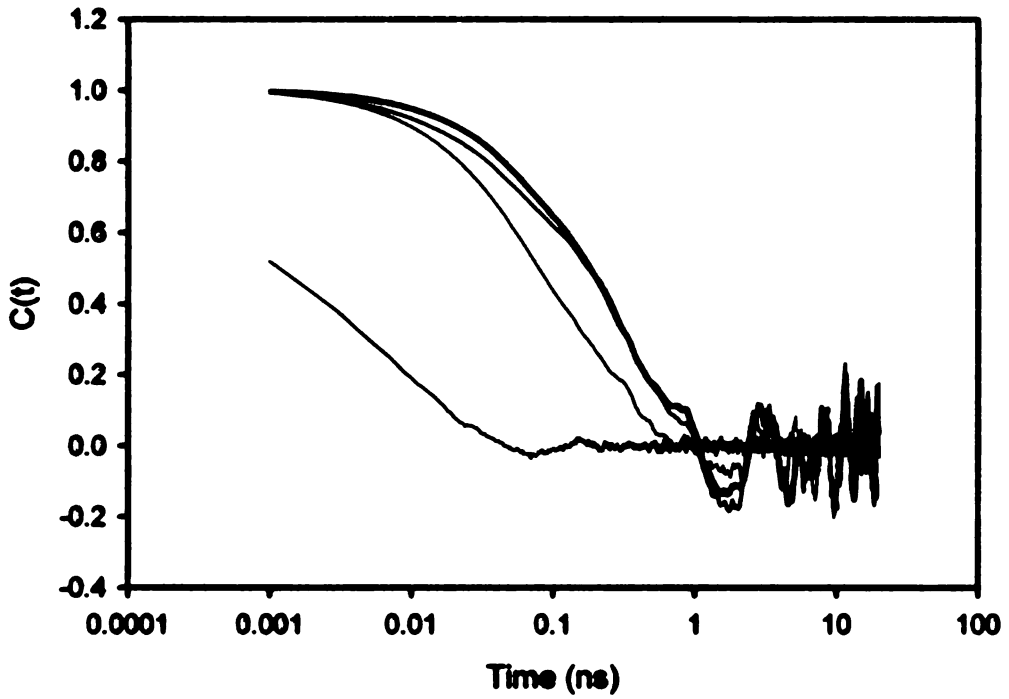


Figure 4.5: Protein L autocorrelation function for various parameters measures at 800K: dihedral angle $\phi(6)$ (green), RMSD from configuration at 35 ns (pink), W47-T57C distance (red), and radius of gyration (black).

dicative of a range of energy barriers on many different scales. The ensemble of unfolded states appear to be stuck in the rugged topography of folding energy landscape. Figure 4.7 shows the probability distribution of W47-T57C end-to-end distance of protein L T57C mutant at 300 K for various independent random coil starting configurations.

For 300K simulation runs, each of the starting configuration was randomly picked from the runs at 600, 700, and 800K. Each 10 ns simulation results in a different end-to-end average distance as shown in figure 4.7. Clearly, the protein is not sampling

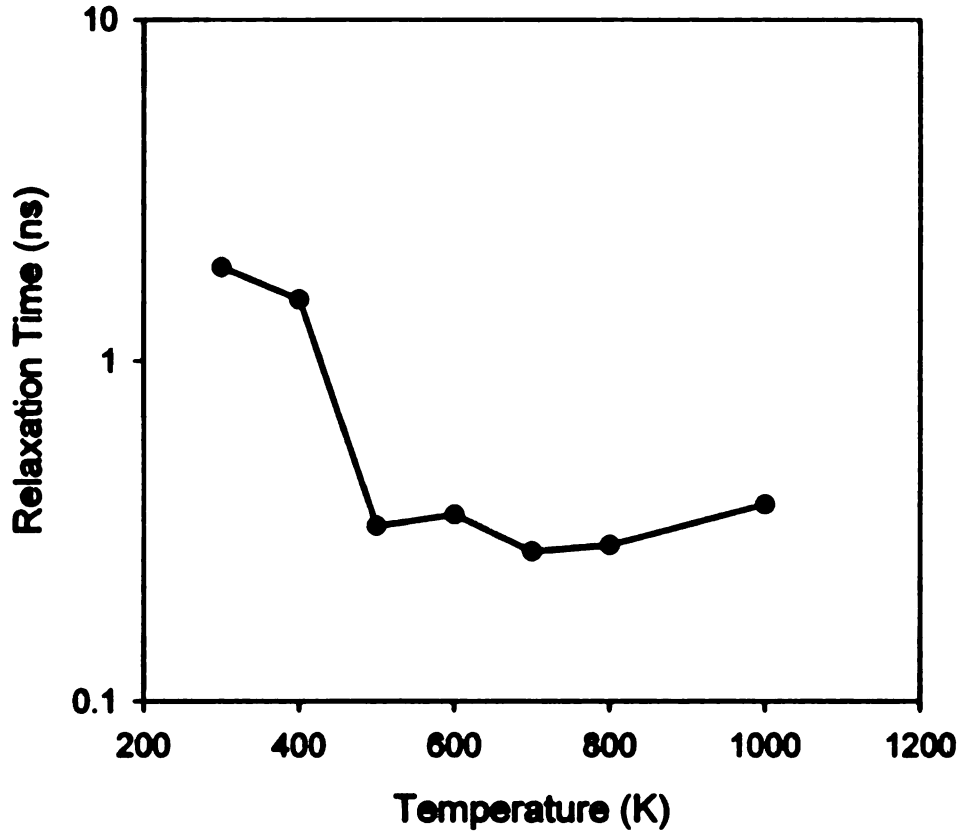


Figure 4.6: Protein L characteristic relaxation time constant estimated at various temperatures from single exponential fits to the autocorrelation function of radius of gyration estimates.

a broad configurational space and is rather immobile or trapped in one of the many local energy minima. Since the probability distributions of the intramolecular W47-T57C distances are not broad enough for the states to interconvert and converge, we conclude that the 10ns simulations at 300K does not produces an ergodic sampling. The net probability distribution over the 70 independent runs at 300K could not be be fit to a Gaussian curve suggesting that the protein does not behave like a freely

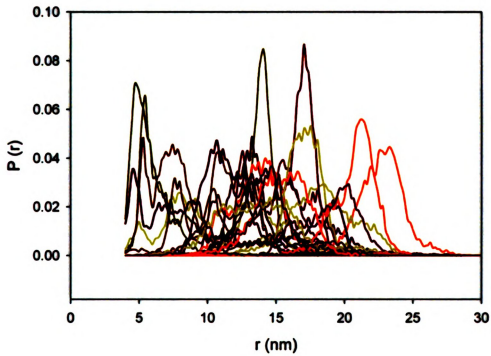


Figure 4.7: Simulated probability distributions of W47-T57C distances over time, for various starting configurations at 300K. Each 10 ns trajectory produces a completely independent and nonergodic distribution of distances.

jointed chain at 300 K.

Further evidence for a complex behavior of the chain at 300K emerges from the Trp-Cys distance autocorrelation plot shown in figure 4.8. An ergodic behavior would be reflected in this autocorrelation function rapidly converging to zero. On the time scale of our simulations, this end-to-end distance autocorrelation does not converge. Absence of any non-converging regime suggests that there is a very strong correlation between the Trp-Cys distances over the range of the trajectories averaged over 70 runs of 10 ns each. That is, each ensemble of states is bound and constrained to

diffuse in its own local potential well and cannot cross over the energy barriers within our simulation time.

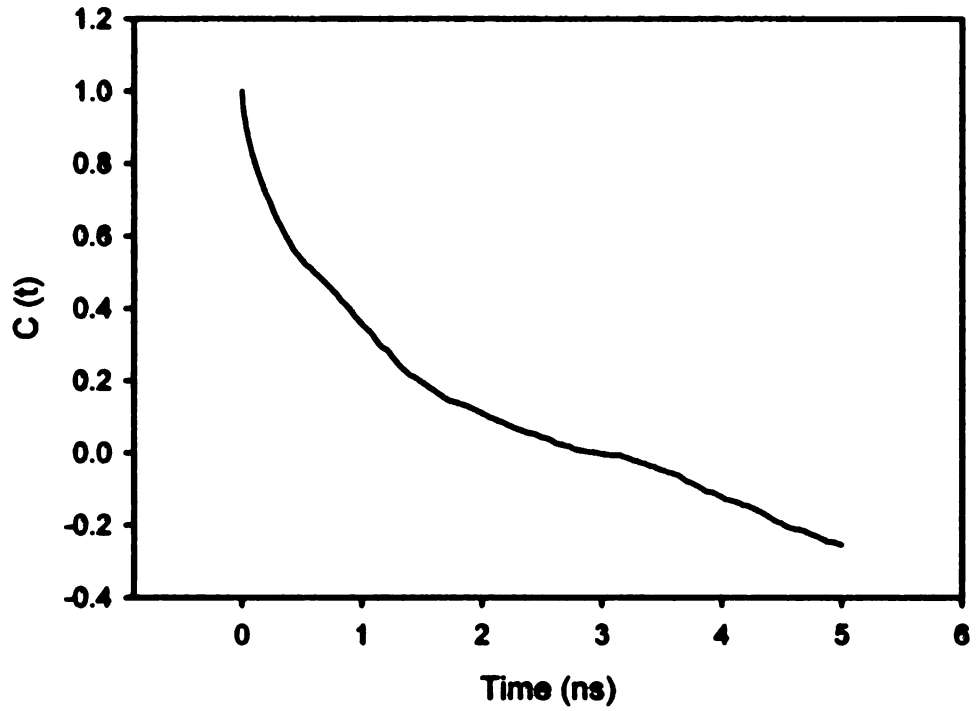


Figure 4.8: Autocorrelation function of W47-T57C distances in protein L T57C MD simulations at 300K. This is an average over 70 independent 10 ns simulation runs.

The relaxation to the equilibrium state is established much faster at higher temperatures. This suggests that the protein molecules are more diffusive at higher temperatures. At lower temperatures the average kinetic energy per molecule is small enough that intramolecular non bonded interactions begin to dominate. These dominating interactions (H-bonding, disulfide, aromatic-aromatic, charge-charge, van der Waals) define the kinetic and thermodynamical properties of the protein and conse-

quently structure of native states. The multiple energy barriers created by various interactions attenuate the protein chain dynamics. The energy required to cross over these barriers and escape the local minima are again obtained from Brownian motion and through these very interactions. Therefore, to accurately capture the protein dynamics through the rough energy landscape, one needs large enough simulation time to first relax the protein into equilibrium and then sample this equilibrium phase to obtain ergodic behavior/sampling. This cannot be easily accomplished on CPU's, but a GPU offers a better performance of up to about 300 ns/day as against about 9 ns/day achieved on the high performance computing cluster.

Taking advantage of the GPU accelerated GROMACS molecular dynamics simulation via the Folding@Home distributed computing platform, tens of thousands of all-atom MD trajectories spanning $\sim 10\mu s$ was generated. Simulations are performed from starting configurations of native, extended and coil states. The plots of W47-T57 distance $P(r)$ over time suggests that the simulation starting from random-coil and extended conformation produces ensembles that begin to converge at about 100 ns and are completely converged by about $1\mu s$ (figure 4.9). Convergence of the distributions can be quantified by computing the relative entropies of extended conformation probability distribution $P_{ext}(r)$ and native state probability distribution $P_{nat}(r)$ with respect to the reference distribution of coil conformation $P_{coil}(r)$:

$$S = \int dr P(r) \log \left[\frac{P(r)}{P_{coil}(r)} \right], \tag{4.1}$$

here $P_{coil}(r)$ forms the reference distribution. According to the relative entropy

metric, the unfolded states are converged on the 100ns-1μs time scale (figure 4.10). The smaller the value of relative entropy, the more similar are the distributions. An identical distribution will have zero relative entropy. The relative entropy metric decays roughly exponentially to small values only after about 100 ns; this explains why shorter 10 ns trajectories at 300K (Figure 4.7) fail to produce converged ensembles.

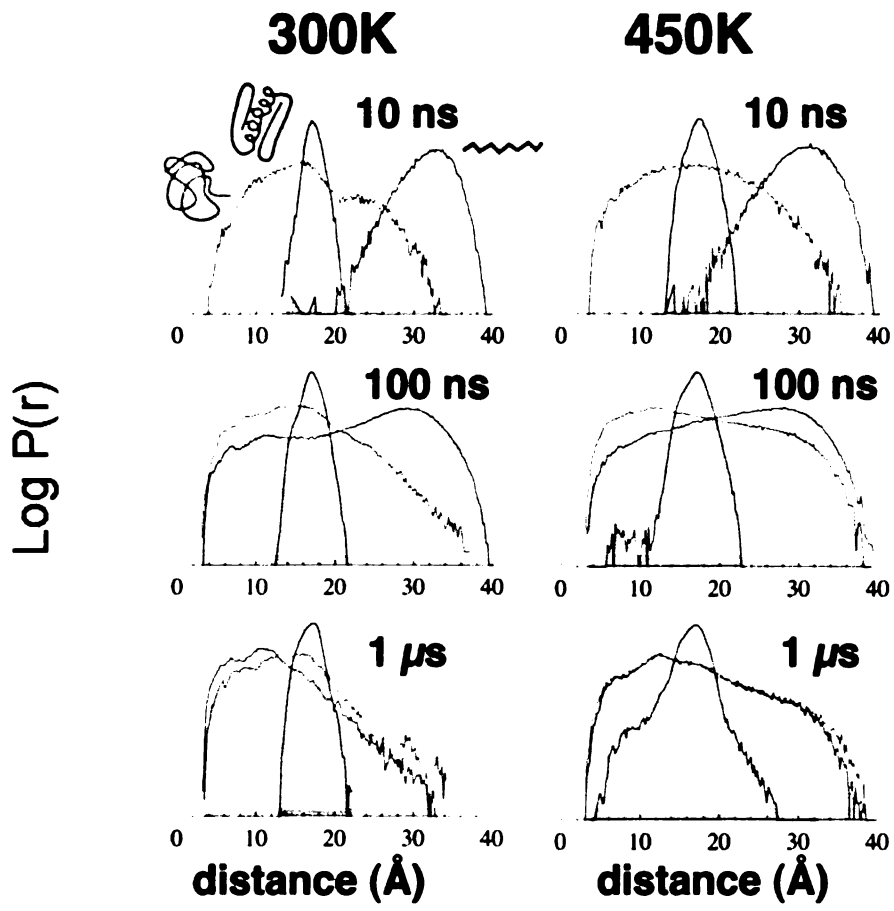


Figure 4.9: Simulated distributions of W47-T57C distances of protein L over time for various starting states at 300K and 450K. The unfolded state trajectories appear to converge over a μs time scale.

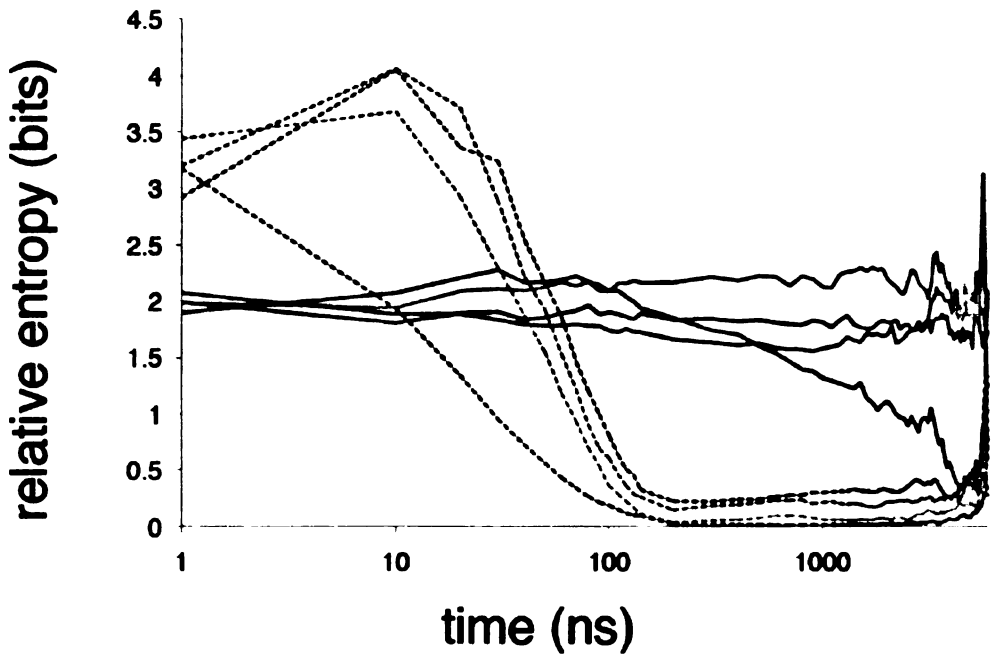


Figure 4.10: Relative entropy: A measure of convergence. The relative entropy of $P_{ext}(r)$ (dashed line) and P_{native}(r) (solid line) with respect to $P_{coil}(r)$ over time, at various temperatures. Smaller values of relative entropy reflect more similar distributions. For protein L T57C, the unfolded states converge on ~100ns-1μs time scale.

To draw a comparison between experiments and MD simulations, we have to compare denaturant (GdnHCl) concentration and temperature. While temperatures induces conformational changes by adding kinetic energy to the molecules and thus higher temperatures favor entropy of unfolding. Chemical denaturants, on the other hand, are understood to act by competing with protein-solvent interactions and intramolecular protein-protein interactions by binding along the protein surface thus

making the protein more soluble. It disrupts the hydrophobic core and induces disorder in the protein configuration. Monte Carlo simulation studies by Choi et al. [80] suggest that denaturant-induced unfolding exhibits a wider conformational sampling than temperature-induced unfolding. Although the destabilizing mechanism for temperature and GdnHCl are different, we would expect a qualitatively similar behavior of the unfolded states of protein under similar conditions. If the conformational sampling is broad enough and truly ergodic, the relaxation kinetics, thermodynamics and structural information would become independent of the type of denaturant (chemical, temperature, pH, or pressure). Consequently, nature and characteristics of the unfolded state will be context invariant. This is where the computing power of GPU's significantly overwhelms CPU's to easily achieve an ergodic sampling of the configuration space.

In our experimental measurements and results on proteins L and G, we observe an unfolded state that is less diffusive and more compact under folding conditions as compared to the denaturing condition (6M GdnHCl). Figure 4.11 shows the reaction-limited and diffusion-limited rates for the destable protein L mutant F22A, K23C. The trend observed in the destable mutant is mirrored in the other protein L mutants studied (T57C and K23C). Below 2 M GdnHCl, the reaction-limited rate increases rapidly with decreasing denaturant and the diffusion-limited rate decreases. At higher denaturant concentration (> 2M GdnHCl), the reaction limited rate decreases with increasing denaturant and diffusion-limited rate increases. In contrast, the rates of protein L K23C indicate that it is more compact than the destabilized protein L mutant (F22A,K23C). This suggests that deletion of the hydrophobic core

causes significant expansion and increased diffusion in unfolded protein states compared to the original protein sequences at higher denaturant concentrations.

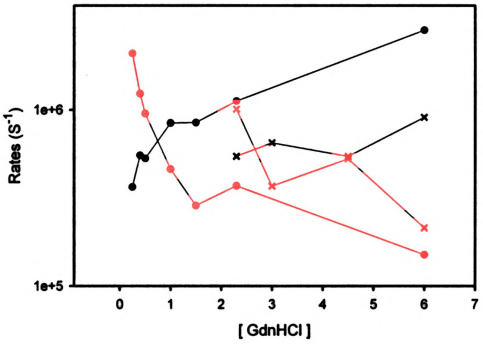


Figure 4.11: Reaction-Limited and Diffusion-Limited Rates for Protein L: The data points marked in red correspond to reaction-limited rates and the blue data points are the diffusion-limited rates. The "x" is for protein L K23C mutant and the circles represent the destabilized protein L F22A, K23C mutant. The rates of K23CC mutant are bounded by the F22A rates.

4.4 Comparison: Simulation and Experiment

A platform for comparing the simulated ensembles with experimental results in the context of intramolecular contact quenching experiments is provided by the SSS

theory. Probability distributions $P(r)$, generated from MD simulations at different temperatures was used to evaluate the reaction-limited rates using equation (2.18). These rates are compared to experimental rates as shown in figures 4.12 and 4.13. There is a very good agreement between the k_R measured as a function of GdnHCl and that computed from simulations using equation (2.18) in the higher denaturing regime (high temperature and high GdnHCl). The comparison reveals an intriguing relationship between temperature and chemical denaturant from the perspective of protein dynamics. For the estimation of k_R, simulated temperatures of 300K corresponds to 0M GdnHCl, 370K corresponds to 2.3M and 450K to ~3M. Simulations at lower temperatures corresponding to lower denaturant concentrations predict a very high k_R increasing by an order of magnitude indicating a compact globular state at lower denaturing conditions. The measured k_R values for F22A show an expanded ensemble at equivalent denaturant concentrations. However, simulations of the F22A mutant predict k_R values with little difference from wild-type k_R values except at lower simulation temperatures (300K and 330K).

Yet another calibration of the simulated ensemble with experimental measurements of protein L is obtained using two similar polymer theory approaches [81,82]. Based on this analysis we conclude that the simulated unfolded ensembles at low temperatures correspond to unfolded ensembles near zero concentration of denaturant, whereas the 450K ensembles correspond to [GdnHCl] $\sim 3.2M \pm 1M$ (for details see Appendix B).

Figures 4.14 and 4.15 show a comparison of the intramolecular diffusion coefficients D as obtained from experiments and simulations. The blue points in the

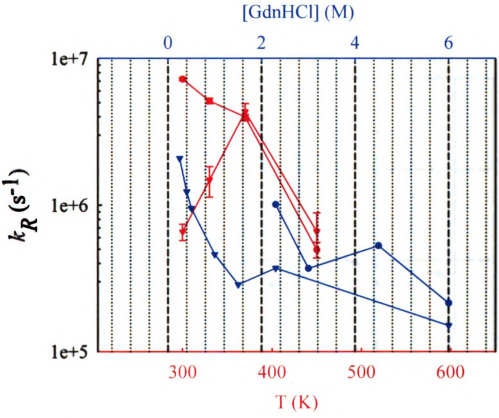


Figure 4.12: Reaction-Limited rate for protein L W47-K23 loop from experiment and MD simulation. Blue points are measured using Trp-Cys contact quenching in GdnHCl. Red points are calculated using equation (2.18) and the MD simulated $P(r)$. The "x" are for F22A mutant of protein L K23C.

figure show the diffusion coefficient calculated using worm like chain modeling of the experimental data, as discussed in chapter 4. The red points correspond to the diffusion coefficients calculated using equation (2.19) from the SSS theory, experimental values of k_{D+}, and $P(r)$ obtained from MD simulations. Experimentally deduced k_{D+} at 2.3M GdnHCl was used with simulated $P(r)$ at 370K to obtain D and the measured k_{D+} at 3M GdnHCl is used with $P(r)$ simulated at 450K. The

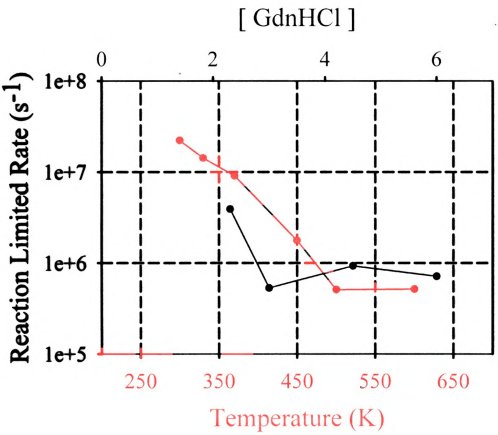


Figure 4.13: Reaction-Limited rate for protein L W47-T57 loop from experiment and MD simulation. Blue points are measured using Trp-Cys contact quenching in GdnHCl. Red points are calculated using equation (2.18) and the MD simulated $P(r)$.

green points in the plot represent the diffusion coefficients calculated directly from the simulated trajectory data by fitting the mean square displacements of Trp-Cys distances over time in 50-ns windows. At greater denaturing conditions (high T of 450K and excess GdnHCl of 3M) the agreement for D is quite good for all three methods. But at lower temperatures and [GdnHCl], the three methods seem to be exhibit differing absolute values of D. However the trend of decreasing diffusion coef-

ficient is captured fairly well. Since there is good agreement for k_R at low simulated temperatures and [GdnHCl], we conclude that the differing diffusion coefficients from mean square displacement is an artifact arising from GBSA solvation model, possibly due to lack of hydrodynamic interactions.

These techniques (MD and Contact Quenching) of measurements and characterization through two different methods (temperature and chemical denaturant) and a comparison between them on at least two different scales provides an estimate on the realistic nature of the computational sampling methods. We find the probability distribution of Trp-Cys distances as sampled experimentally by denaturant and in simulations by temperature to exhibit similar conformational ensembles of the unfolded states in elevated denaturing regime (high T and [GdnHCl]) as measured by k_R. Equation (2.18) suggests that k_R is directly related to the probability distribution but is scaled by the distance dependent quenching rate. Since k_R is sensitive only to the lower end of $P(r)$, the comparison of the conformational ensembles with the yardstick of k_R will only be relevant to the tail of the distributions. However, the intramolecular diffusion coefficients incorporate more global dynamics and therefore compares the unfolded states on a larger length scale. In elevated denaturing regimes (450K, 3M GdnHCl), the diffusion coefficient estimated by MD simulations and contact quenching experiments seem to agree quite well. But in lower denaturing conditions of temperatures and chemical denaturants, the values of D obtained are quite different. This could be because in mild denaturing conditions of temperature and [GdnHCl], the unfolded fractions are observed to be less diffusive, more viscous and compact as compared to the fully denaturing conditions. Consequently

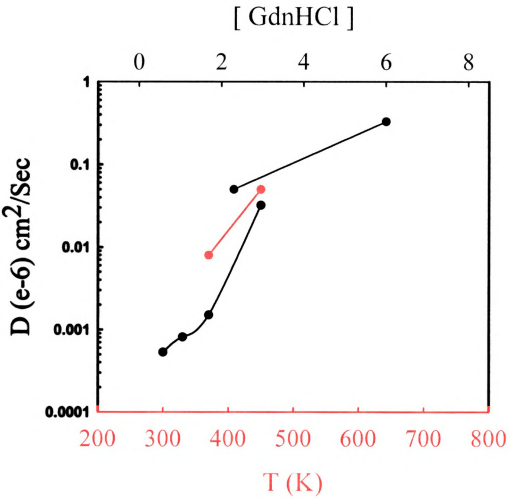


Figure 4.14: Comparing experimental and computationally deduced diffusion coefficient (D) for W47-T57C loop of B1 domain of protein L. The green points are calculated from the simulated mean squared displacement over time, blue points are calculated using worm-like chain model with equation (2.19) and experimentally obtained k_{D+}, the red points are calculated using equation (2.19), the MD simulated $P(r)$, and the measured k_{D+} at the equivalent [GdnHCl].

the average kinetic energy per protein molecules is smaller under folding conditions. Therefore, the amino acids have a long enough time to interact and influence each

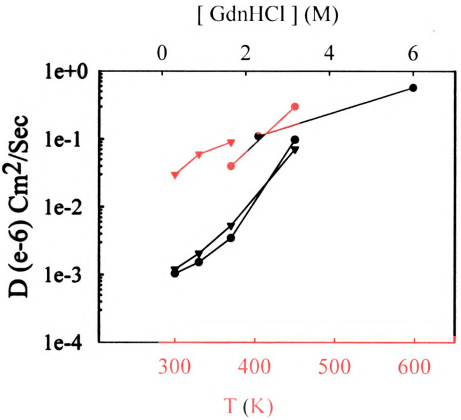


Figure 4.15: Comparing experimental and computationally deduced diffusion coefficient (D) for the K23C and the F22A mutants of B1 domain of protein L. The triangles are for the F22A mutant of protein L K23C. The green points are calculated from the simulated mean squared displacement over time, blue points are calculated using worm-like chain model with equation (2.19) and experimentally obtained k_{D+}, the red points are calculated using equation (2.19), the MD simulated $P(r)$, and the measured k_{D+} at the equivalent [GdnHCl].

other's motion causing roughness in the potential of the free energy landscape in which the proteins diffuse. It appears that for proteins to adopt the right native state, it is essential for it to have this low diffusivity to instigate the folding process. This may be the only way local interactions can take dominant effect and define

the kinetic and thermodynamical properties to lead the process of protein folding. But in elevated denaturing regimes, the kinetic energy per protein molecule is large enough that they can easily overcome and negate most of the local and nonlocal interactions (H-bonding, disulfide, aromatic-aromatic, charge-charge, Van der Waals) thus rendering the otherwise rugged landscape itself rather flat and planar.

To obtain a better understanding of protein folding, it is necessary to mark its characteristics at length and time scales (femtoseconds to seconds; angstroms to microns) that encompasses all of the relevant protein folding events. This can be accomplished only through a good connect between experimental results and molecular simulations. Since we observe a very slow equilibration of unfolded states on the timescales of a μs, this translates to, in the current scenario of computing power and throughput, adopting and further developing of parallel processing with increased number of graphical processing units (GPU).

Chapter 5

Summary

Anfinsen's experiment proved that all the information needed for the protein to adopt its native three-dimensional structure is embedded in its amino acid sequence. Although there are exceptions to this rule, but it holds true for most small fast folding single domains. Every protein synthesized *in vivo* has a function - either as an antibody or a catalytic enzyme or some other specific physiological function. The biological function of the protein therefore, is strongly dependent on its native conformation. To confront and successfully solve the protein folding problem of predicting the three-dimensional structure from its amino acid sequence, we need appropriate models for the initial unfolded state ensemble of the protein that is suffused with all the information necessary to spontaneously seek the viable native state through kinetic and thermodynamic considerations. Recent evidences from NMR experiments [30, 83] and results presented in this thesis have challenged the earlier perspectives of random coil nature of the unfolded state ensembles of a pro-

tein under folding conditions and have emphasized presence of residual structure in unfolded states. The dynamic and heterogeneous nature of the unfolded state conformations has proven to be an intriguing hindrance to its characterization.

The growing interest in the nature of a protein's unfolded state has resulted in rapidly developing methodologies to study them. Using experimental and computational techniques to study the early events of protein folding I have presented in this thesis -

1. Characteristics of polyglutamine polypeptides that are prone to aggregation. A possible pathway for this aggregation is proposed.
2. Early events in the folding process of well behaved proteins L and G that are not aggregation prone under normal physiological conditions.
3. An attempt to bring together the results from experiments, molecular simulation and polymer modeling to offer deeper and far reaching insights into the mechanisms of protein folding.

Aggregation Mechanism in Polypeptides: Using the experimental technique of transient intramolecular Trp/Cys contact quenching on polyglutamine polypeptides and modeling the length dependence of contact formation rates with a worm-like chain, we determined the persistence length of polyglutamine polypeptides to be $\sim 13.0\mathring{A}$. This corresponds to ~3.4 amino acid residues. Considering the observations reported for other peptide sequences and denatured proteins to be typically about $\sim 4\mathring{A}$, the polyglutamine sequences appear to be much more stiffer. This polymer "stiffness" possibly suggests the polyglutamine sequence prefers an extended conformation that hinders the host protein in adopting its intrinsic native state.

Consequently the destabilized misfolded state forms a nucleus for aggregation and amyloid formation.

A Primordial Hydrophobic Collapse: The technique of Trp/Cys contact quenching to measure intramolecular diffusion and rate of end-to-end contact formation, revealed for B1 domains of proteins L and G, a compact and viscous unfolded state under conditions that favor folding. This suggests that the primordial event after protein synthesis is the hydrophobic collapse of the relevant hydrophobic side chain amino acid residues. This entropy driven collapse is responsible for close intramolecular contacts between various chain segments triggering the interplay of thermodynamic, electrostatic forces and mechanical steric effects that induces ruggedness of the free energy landscape. For proteins L and G, the diffusion coefficient dropped by a factor of almost 6 at 2.3M GdnHCl compared to that at 6M GdnHCl, and the average end-to-end distance dropped by $\sim$10%. Similar compaction in the unfolded states of the protein has been reported by other research groups. Both proteins L and G share the same native topology with a central α-helix resting on a four-stranded β-sheet composed of N- and C-terminal β-hairpins, however, they follow different folding pathways. This is suggestive of at least two characteristics of proteins en route the folding - firstly of existence of a hydrophobic collapse as the precursor to the actual folding event, and secondly sequence dependent pathways for protein folding inspite of similar native topology.

Synthesis of Experiments, Molecular Simulation and Polymer Modeling: It is currently impossible to identify all the conformational variations in a protein as it progresses spontaneously towards its native basin on the energy land-

scape. The experimental techniques and protocols lack the fine temporal and spatial resolution. The simulations are deficient in their description of force fields, potentials and solvation. The polymer theory lacks the necessary chemical heterogeneity and other salient information on local structural ordering based on electrostatics and bond formations. A more complete understanding of the folding mechanism and pathways requires a union of all these techniques. Experimentally deciphered characteristics of early events of protein folding can be mapped onto the polymer models and the biochemical and physical characteristic ascribed to the respective local groups in molecular simulations so that they sample the required configurational space and adopt the accurate biologically functional native state conformation.

For a comprehensive understanding of the protein folding problem through a harmonious synthesis of the various fragments of folding characteristics, it is necessary to achieve a close collaboration between experimental techniques, computational simulations, and theoretical modeling. The computational simulations can be benchmarked by experimental results for providing the folding principles at a molecular level. Details of the molecular mechanism can be delved into only through computational techniques. For the simulations to be realistic and accurate, it should have a good overlap with predictions of experimental results and be well substantiated. The folded native conformation of a protein being unique, distinct and relatively rigid, its characterization is well accomplished through the recorded atomic coordinates, the local secondary structure elements defined by the network of hydrogen-bonds and well characterized dihedral angles. However, the unfolded state ensembles do not submit themselves for such a representation owing to their conformational hetero-

geneity. Therefore, the only way to characterize the distribution of unfolded fractions of the protein under folding conditions is to use physics based polymer models which use a statistical mechanics based depiction.

As a preliminary step the distribution of unfolded state conformations was sought by tuning the end-to-end distance probability distributions to the experimentally obtained reaction-limited rates and the diffusion coefficient deciphered through worm-like chain modeling. The simulations reveal a coil-to-globule transition with decreasing temperature that closely resembles the results found with decreasing GdnHCl concentration. The degree of such a chain collapse would depend on the nature of the hydrophobic pattern distribution along the chain length. As the protein is formed in the ribosome, the hydrophobic residues begin sequestering in an entropically driven process. Any polymer model that seeks to accurately describe the unfolded state fraction must incorporate this sequence dependent hydrophobic collapse which is responsible for bringing together in close contact various segments of the chain for short and long range forces to then start taking effect. Depending on the local short range and long range interactions, the protein adopts one of the multiple routes towards the native conformation basin.

In these molecular dynamics simulations, we observe a very slow equilibration of unfolded states on the timescales of a microsecond. To make more meaningful comparison between the experimental results and computational data, it would be essential to obtain a broad sampling of the configuration space that is essentially ergodic. The quest for the fundamental principles of protein folding will find its consummation in the synthesis of the experimental techniques, the molecular sim-

ulations and polymer modeling to define the free energy landscape on which the protein molecules diffuse to adopt their native conformation.

Appendix A

Polymer Models for Describing Unfolded State Conformations

A simple two-state kinetic model may not completely describe the folding characteristic of some proteins. Presence of compact non-native states under folding conditions are suggestive of transient "intermediate" states seemingly diffusing on a rugged energy landscape. The structural and dynamic characteristics of such states cannot easily be captured unless we adopt physics based polymer models. Many polymer models have developed that seek to describe various characteristics of the unfolded state ensembles in proteins and polypeptides. Here, I present a few polymer models that I have used in this work.

A.1 Freely-Jointed Chain

The freely-jointed chain model [84] of polymers is among the simplest model for polymer conformations. This model treats the polymer chain as being constituted of N rigid rods (monomers) of fixed length l as shown in figure A.1. The adjacent monomers have no orientational correlation and the pivot point is absolutely flexible. Therefore, there is no excluded volume in this model and different chain segments can occupy the same volume in space. The total contour length of the chain is $L = N \times l$.

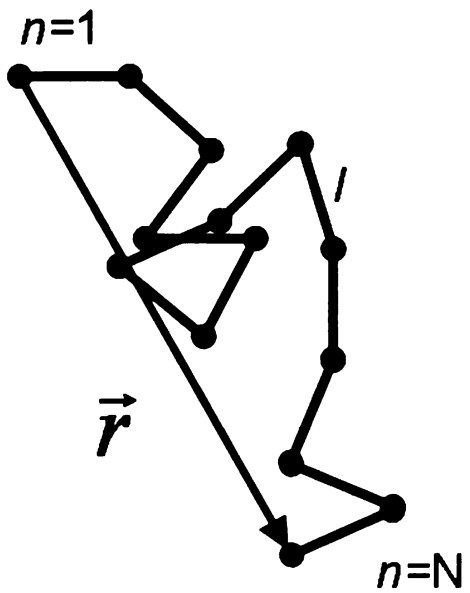


Figure A.1: Freely-Jointed Chain: Each monomer of length l has no orientational correlation with other monomers. There is no excluded volume interaction and two chain segments can occupy the same volume.

For an unbiased chain conformation, the ensemble average of end-to-end position vector is zero since the links are uncorrelated. The average of end-to-end distance squared, known as the mean squared displacement is

$$< r^2 >= \left\langle | \sum_{n=1}^{N} l_n |^2 \right\rangle = Nl^2 \tag{A.1}$$

therefore, the approximate size of the freely-jointed chain (as given by the root mean square of end-to-end distance, $\sqrt{N}l$) grows as square root of the number of monomers. For large N, size of the polymer would be much smaller than its contour length. For a large N, the probability density $P(r)$ such that $P(r)dr$ gives probability of finding the polymer chains with the end-to-end distance between r and $r + dr$ is given by a Gaussian distribution

$$P(r) = P(x)P(y)P(z) = \left(\frac{2\pi < r^2 >}{3} \right)^{-3/2} \exp\left(-\frac{3r^2}{2 < r^2 >} \right) \tag{A.2}$$

where the factor $(2\pi < r^2 > /3)^{-3/2}$ is obtained from the normalization condition $\int P(r)d^3r = 1$. For a freely-jointed chain of N segments each of length l, the radius if gyration is given by

$$R_g = \left(\frac{N}{6} \right)^{1/2} \times l \tag{A.3}$$

$$R_g^2 = \frac{Nl^2}{6} \equiv \frac{< r^2 >}{6} \tag{A.4}$$

and hence the ratio between radius of gyration and end-to-end distance is a constant

$$\frac{< r^2 >}{R_g^2} = \frac{Nl^2}{Nl^2/6} = 6 \tag{A.5}$$

The freely-jointed chain does not take into account any excluded volume. But real polymers do have an excluded volume interaction and also a stiffness. Chain stiffness is parametrized through either a Kuhn length or a persistence length. This model assumes that chemical bonds are free to rotate and posses a uniform distribution of bond angles. The uniform distribution of bond angles necessitates that the average chain vector over all conformations is zero. Flory used this relation to define the characteristic ratio (C_n) of a polymer

$$C_n = \frac{< r^2 >}{Nl^2} \tag{A.6}$$

By definition C_n is unity for freely jointed chains, for other models, such as the worm-like chain (WLC), which do not assume that the bond angle is free to rotate, C_n exceeds unity.

A.2 Gaussian Chain

In this model the length of the monomers (bond vectors) is no longer a constant, and the ideal polymer is treated as an effective freely jointed chains of Kuhn segments [85, 86]. Each link vector has a probability distribution given by

$$P(r) = \left(\frac{2\pi l^2}{3}\right)^{-3/2} \exp\left(-\frac{3r^2}{2l^2}\right) \tag{A.7}$$

where $< r^2 >= l^2$

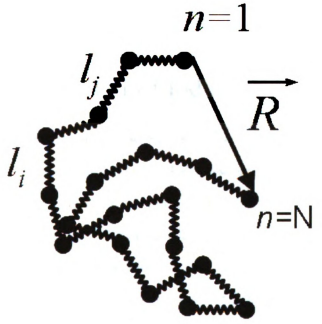


Figure A.2: Gaussian Chain Model: The length of each bond vector l_i is not a constant but has a Gaussian probability distribution.

The probability density of end-to-end distance distribution is given by

$$P(R) = \left(\frac{2\pi < R^2 >}{3}\right)^{-3/2} \exp\left(-\frac{3R^2}{2 < R^2 >}\right) \tag{A.8}$$

The Gaussian chain model gives a better description of the chain statistics for long chains. An important property in the chain characterization is the excluded volume, but the Gaussian chain ignores this. Local excluded volume interactions will play a large role at smaller chain lengths by constraining the backbone dihedral angles. Although the probability density may not be very accurately depicted, most denatured proteins and polypeptide chains do exhibit a Gaussian statistics for large chains lengths.

A.3 Worm-Like Chain Model

Also referred to as the Kratky-Porod worm-like chain model, it is the simplest model that incorporates stiffness to describe the behavior of semi-flexible polymers. The model describes the whole spectrum of chains with different degrees of chain stiffness from rigid rods to random coils. Proteins and polypeptides are typically inextensible unlike the Gaussian chain considered above. For short peptide chain lengths, the polymer exhibits memory effect in the chain direction. Persistence length is the distance over which the polymer begins to loose memory of the direction. Mathematically, it is expressed by the correlation between the tangent vectors at different points along the chain

$$\left\langle \hat{t}_1(s) \cdot \hat{t}_2(s') \right\rangle = \exp\left(-\frac{|s - s'|}{l_p} \right) \tag{A.9}$$

where $\hat{t}_1(s)$ and $\hat{t}_2(s')$ are the unit tangent vectors at points r_1 and r_2 respectively

separated by a distance $|s - s'|$ along the chain segment, and l_p is the persistence length. For $|s - s'| << l_p|$, $\hat{t}_1(s)$ and $\hat{t}_2(s')$ will be approximately collinear and the chain segment will appear like a stiff rod. For $|s - s'| >> l_p|$, the unit vectors would be fully independent (no memory) of each other and the chain segment would appear absolutely flexible. The mean square distance between two points for a chain of length l_c is given

$$\left\langle r^2 \right\rangle = 2l_p l_c - 2l_p^2[1 - \exp(-l_c/l_p)] \tag{A.10}$$

There is no one single expression that describes the $P(r)$, but to obtain the probability distribution of end-to-end distances, wormlike chain model is constructed with a Monte Carlo algorithm using the method of Hagerman and Zimm. Chains with a persistence length, lp, are randomly generated by sequentially adding random links to the C-terminus. Each link in the chain is 0.38 $\AA$ long, and is related to the previous link by two spherical angles, θ and ϕ, where the polar angle θ is a Gaussian distributed random number with a variance $4l_c/l_p$ around zero, and the azimuthal angle ϕ is a random number between 0 and 2π . To account for excluded volume effects, the pairwise distance between the last link added and every tenth link in the chain is computed. If a clash of less than the excluded volume diameter of d_α is detected, the chain is truncated 3 persistence lengths before the clash and the chain is regenerated. Millions of chains can be thus simulated and the end-to-end distance for each chain is calculated to create a normalized histogram with $0.1\AA$ binning. This normalized histogram is used as the probability distribution of distances $P(r)$.

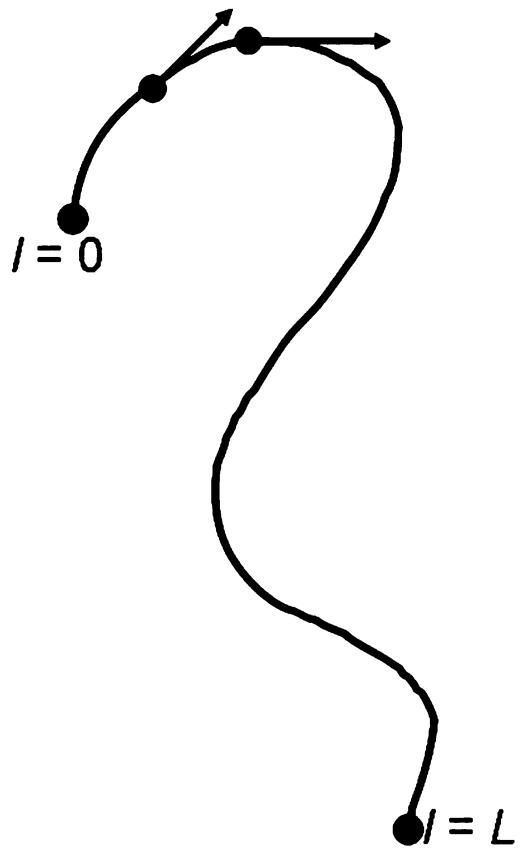


Figure A.3: Worm-Like Chain Polymer Model: This model incorporates an intrinsic polymer stiffness characterized by persistence length l_p. the excluded volume interactions are invoked by checking for spatial clash of the chain segments. The tangent vector at each point along the chain can be used to estimate the persistence length.

Appendix B

Calibrating Simulated Unfolded Ensembles with Experiment: Polymer Theory Approach

Protein stability is quantified by studying the denaturation-renaturation curve under different conditions of the denaturant - chemical (GdnHCl or urea), temperature or pH, etc. The molecular simulations most often employ temperature as the parameter of choice while many experimental techniques rely on a chemical denaturant. To be able to make a comparison between experiments and simulations and place them on a common platform, it is therefore necessary to establish a link between the unfolded state ensembles obtained with these two processes of denaturation.

A polymer-theory approach is used to examine the coil-globule transition in both experimental results and simulation data. The estimated radius of gyration (R_g)

is a good indicator of free volume occupied by the polymer chain (chain expansion and compaction). In the coil-globule reconfiguration process of the chain, some of the monomer residues undergo a shift in its physiochemical environment from the hydrophobic core to being solvent exposed. The free energy associated with this shift is referred to as the transfer free energy. A key features of the polymer theory is that this mean-field transfer free energy is independent of the molecular mechanism involved in protein destabilization. Therefore, the free energy for transfer of a monomer into solvent can be computed based on estimated R_g and compared across different [GdnHCl] and temperatures in order to map the unfolded state ensembles with respect to the two denaturing parameters.

To estimates the free energy of transfer (ε), two different polymer-models are used for fitting R_g vs temperature and a comparison is made with the experimentally obtained conformation. We refer to the first polymer model as the "Dill" model and the second as "Ziv" model:

B.1 Dill Polymer-Model

This model was proposed by Ken A. Dill [81, 87]. In this model, the free energy difference ΔF_{Dill} with respect to the maximally compact polymer state is modeled as

$$\frac{\Delta F_{Dill}}{nk_BT} = -\frac{7}{2}\left(\rho_0^{2/3} - (\rho_0/\rho)^{2/3}\right) + \frac{2}{n}\ln\rho + \left(\frac{1-\rho}{\rho}\right)\ln\rho + \frac{\varepsilon}{2k_BT}(1-\rho) \tag{B.1}$$

where n is the chain length, k_B is the Boltzmann constant, T is temperature, ρ is volume density of the monomers and is defined such that $\rho = 1$ in the maximally compact state (usually assumed to be the native state), ρ_0 is the volume density at the coil-globule transition, and ε reflects a mean-field transfer energy across all residues. The value of ρ_0 is fixed to be

$$\rho_0 = \left[\frac{19}{27n}\right]^{1/2} \tag{B.2}$$

B.2 Ziv Polymer-Model

The second of the models is the modified-Sanchez model of Ziv and Haran [82] which we refer to as the "Ziv" model. This model starts with the empirical ideal-chain Flory-Fisk distribution [88] of the radius of gyration $R_g =< R_g^2 >^{1/2}$ and weights it by a Boltzmann factor capturing the transfer free energy of the chain:

$$P(R_g) \sim R_g^6 \exp\left(-\frac{7R_g^2}{2R_0^2}\right) \exp\left(-\frac{ng(\rho,\varepsilon)}{k_BT}\right) \tag{B.3}$$

where

$$g(\rho, \varepsilon) = -\frac{1}{2}\rho\varepsilon + k_B T \left[\frac{1-\rho}{\rho}\right] \ln(1-\rho) \tag{B.4}$$

Here, R_0 is the radius of gyration of the ideal chain and is fixed by the relation

$$\rho_0 = \frac{R_{native}^3}{R_0^3} \tag{B.5}$$

B.3 Fitting Polymer Theory Models to Simulated Data

Since the equilibration of the unfolded states occurs on the timescale of about a microsecond, the radius of gyration is calculated using the simulated ensemble sampled after 1 microsecond. The ensemble corresponding to the starting structure as the extended conformation is used for estimation of R_g for the simulations at temperatures T = 300K, 330K, 370K and 400K. The calculated mean radius of gyration for ensembles started from the extended state at various simulated temperatures is shown in table B.1. The radius of gyration of the maximally compact state R_{native}, was set to the lowest value of mean R_g observed in our simulations (11.5$\AA$). The free energy for the Dill and Ziv models is minimized by numerical evaluation of the

volume density (ρ). A least-squares fit to the R_g vs T data is then performed to obtain the best estimate of transfer free energy ε.

Temperature (K)	Average R_g (after 1 μs) ($\AA$)
300	12.42
330	12.32
370	12.63
450	21.99

Table B.1: Simulated temperature and mean radius of gyration for ensembles started from the extended state.

As is evident in figure B.1, the trend in R_g vs T cannot be accurately represented by fitting it with constant values for ε.

Therefore, a temperature dependent linear model of the transfer free energy, $\varepsilon(T) = \varepsilon_0 - (T - T_0)\Delta S$, is used for produce a more satisfactory and accurate fit to the data. For both the Dill and Ziv models, values of ε_0 and ΔS are obtained by least-squares fitting. Such a fit is shown in figure B.2. The fitting parameters obtained for Dill models is found to be $\varepsilon_0 \sim$ 4.0kcal/mol and $\Delta S \sim$ 20.4kcal/mol and that for Ziv mdoel is found to $\varepsilon_0 \sim$ 3.7kcal/mol and $\Delta S \sim$ 20.8kcal/mol. There is a very good agreement between both the Dill and Ziv models for transfer energy estimation.

B.4 Comparing Simulated and Experimental Unfolded Ensembles

From histograms of FRET efficiencies published in two different single-molecule FRET studies of protein L by Sherman et al. [89] and Merchant et al. [90], Ziv

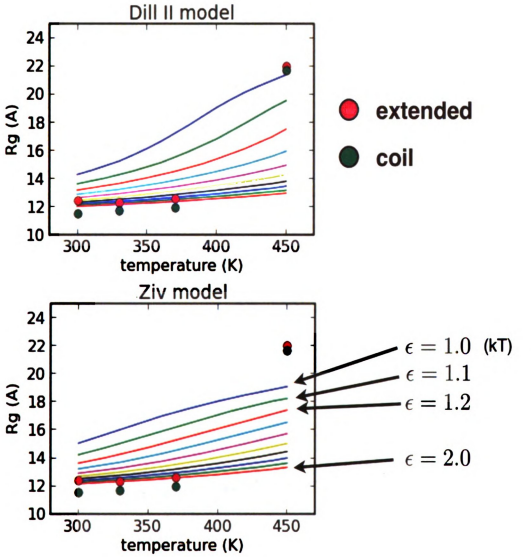


Figure B.1: The trend in R_g vs T cannot be accurately represented by fitting it with constant values for ε. Various constant values of $\varepsilon(T) = \varepsilon_0$ yield poor polymer theory fits to the data.

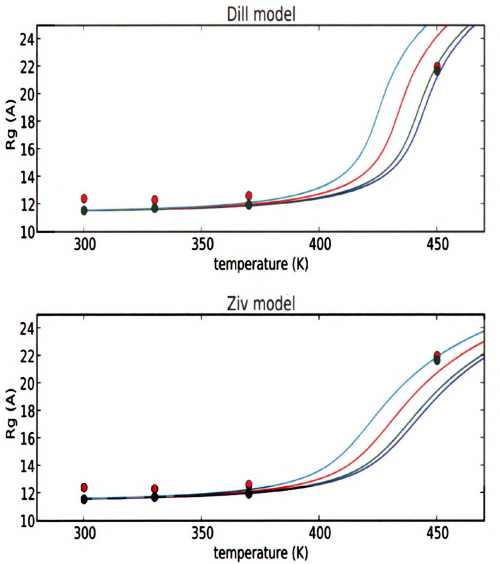


Figure B.2: A temperature dependent linear model of the transfer free energy, $\varepsilon(T) = \varepsilon_0 - (T - T_0)\Delta S$, yields a robust fit to the average R_g versus temperature.

and Haran [82] computed free energies of transfer ε([GdnHCl]) and expansion factors α^2([GdnHCl]).

The plot is shown in figure B.3. The expansion factor is equivalent to $\alpha^2 = R_g^2/R_0^2$, where R_g is the radius of gyration of the ensemble, and R_0 is the radius of gyration for the ideal chain at the coil-globule transition. Using $R_{native} = 11.5\mathring{A}$, $R_g(450K) = 21.99$ and $n = 64$, we compute the expansion factor for our 450K ensemble to be $\alpha^2 \sim 0.81$. As seen in figure B.3 the value of [GdnHCl] with an equivalent expansion factor is $\sim$ 2.45M for the Merchant et al. data, and $\sim$ 3.8M for the Sherman et al. data. Thus for purposes of comparing the simulated and experimental unfolded ensembles, we conclude that the 450K ensemble is comparable to an experimental GdnHCl concentration of $\sim 3.2M \pm 1M$.

The simulated unfolded states at low temperatures (300K, 330K and 370K) exhibit a high degree of compaction, becoming even more compact than the native ensembles beyond 1 μs. This observation, combined with the large free energies of transfer predicted by the polymer theory, and small values of the expansion factors for the observed R_g, indicate that the low-temperature simulations can best be compared to experimental conditions in the absence of denaturant.

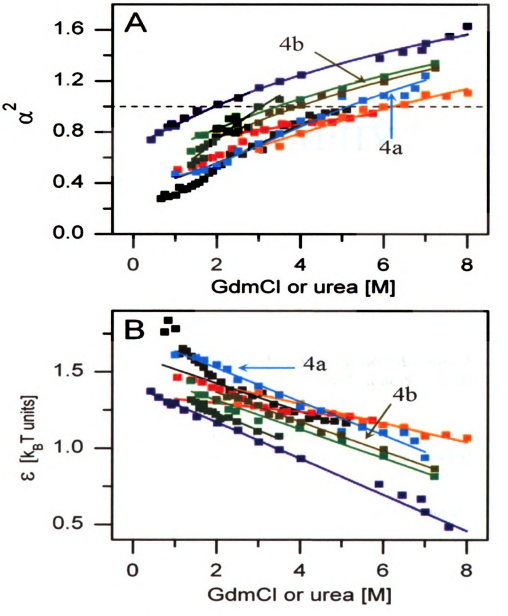


Figure B.3: Coil-Globule transition in denatured proteins obtained from smFRET experiments by analysis using Sanchez polymer theory. The plot in light blue (4a) corresponds to the Sherman et al. data and that in brown (4b) is for the Merchant et al. data for protein L. (A) The expansion factor α^2 as a function of denaturant concentration. (B) the mean-field interaction energy ε as a function of denaturant concentration. This plot is adopted from [82].

BIBLIOGRAPHY

[1] C. M. Dobson. Protein-misfolding diseases: Getting out of shape. *Nature*, 418:729–730, 2002.

[2] J. W. Kelly. Towards and understanding of amyloidgenesis. *Nat. Struct. Biol.*, 9:323, 2002.

[3] E. H. Koo, P. T. Lansbury Jr., and J. W. Kelly. Amyloid diseases: Abnormal protein aggregation in neurodegeneration. *Proc. Natl. Acad. Sci.*, 96:9989–9990, 1999.

[4] P. Hammarstrom, R. L. Wiseman, E. T. Powers, and J. W. Kelly. Prevention of transthryretin amyloid disease by changing protein misfolding energetics. *Science*, 299:713–716, 2003.

[5] J. C. Sacchettini and J. W. Kelly. Therapeutic strategies for human amyloid diseases. *Nature Rev. Drug Discov.*, 1:267–275, 200).

[6] K. A. Dill, S. B. Ozkan, T. R. Weik, J. D. Chodera, and V. A. Voelz. The protein folding problem: when will it be solved? *Current Opinion in Structural Biology*, 17:342–346, 2007.

[7] P. A. Ellison and S. Cavagnero. Role of unfolded state heterogeneity and en-route ruggedness in protein folding kinetics. *Protein Sci.*, 15 (96):564–582, 2006.

[8] P. Hammarstrom and U. Carlsson. Is the unfolded state the rosetta stone of the protein folding problem? *Biochemical and Biophysical Research Communications*, 276 (2):393–398, 2000.

[9] K. A. Dill. Dominant forces in protein folding. *Biochemistry*, 29 (31):7133–7155, 1990.

[10] L. Pauling and R. B. Corey. The pleated sheet, a new layer configuration of polypeptide chains. *Proc. Natl. Acad. Sci.*, 37:251–256, 1951.

[11] L. Pauling and R. B. Corey. The structure of feather rachis keratin. *Proc. Natl. Acad. Sci.*, 37:256–261, 1951.

[12] G. A. Petsko and D. Ringe. *Protein Structure and Function.* WileyBlackwell, third edition, 2004.

[13] A. Alm and D. Baker. Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proc. Natl. Acad. Sci.*, 96 (20):11305–11310, 1999.

[14] E. E. Lattman and G. D. Rose. Protein folding – what's the question? *Proc. Natl. Acad. Sci.*, 90:439–441, 1993.

[15] N. T. Southall, K. A. Dill, and A. D. J. Haymet. A view of the hydrophobic effect. *J. Phys. Chem. B*, 106 (3):521–533, 2002.

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

[17] H. Muirhead and M. Perutz. Structure of hemoglobin. a three-dimensional fourier synthesis of reduced human hemoglobin at 5.5 angstrom resolution. *Nature*, 199 (4894):633–638, 1963.

[18] C.B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181 (96):223–230, 1973.

[19] C.B. Anfinsen and E. Haber. Studies on the reduction and re-formation of protein disulfide bonds. *J. Biol. Chem.*, 236:1361–1363, 1961.

[20] L. Cyrus. Are there pathways for protein folding? *Journal de Chimie Physique et de Physico-Chimie Biologique*, 65:44–45, 1968.

[21] K. Baker and D. A. Agard. Kinetics versus thermodynamics in protein folding. *Biochemistry*, 33 (24):7505–7509, 1994.

[22] S. Govindarajan and R. A. Goldstein. On the thermodynamic hypothesis of protein folding. *Proc. Natl. Acad. Sci.*, 95 (10):5545–5549, 1998.

[23] M. B. Berkenpas, D. A. Lawrence, and D. Ginsburg. Molecular evolution of plasminogen activator inhibitor-1 functional stability. *EMBO J.*, 14 (13):2969–2977, 1995.

[24] T. Lazaridis and M. Karplus. Thermodynamics of protein folding: a microscopic view. *Biophys. Chem.*, 100:367–395, 2003.

[25] K. A. Dill and H. S. Chan. From levinthal to pathways to funnels. *Nat. Struct. Biol.*, 4:10–19, 1997.

[26] L. Cruzeiro-Hansson and P. A. S. Silva. Protein folding : thermodynamic versus kinetic control. *Journal of Biological Physics*, 27:S6–S8, 2001.

[27] K. A. Dill and H. S. Chan. Protein folding in the landscape perspective: Chevron plots and non-arrhenius kinetics. *Proteins.*, 30 (1):2–33, 1998.

[28] T. Mittag and J. D. Forman-Kay. Atomic-level characterization of disordered protein ensembles. *Critical Reviews In Biochemistry And Molecular Biology.*, 17 (1):3–14, 2007.

[29] McCarney E. R., Kohn J. E., and Plaxco K. W. Is there or isn't there? the case for (and against) residual structure in chemically denatured proteins. *Current Opinion In Structural Biology.*, 40 (4):181–189, 2004.

[30] Meier S., Blackledge M., and Grzesiek S. Conformational distributions of unfolded polypeptides from novel nmr techniques. *The Journal of chemical physics.*, 128 (5):052204–189, 2008.

[31] W. A. Eaton, V. Munoz, S. J. Hagen, G. S. Jas, L. J. Lapidus, E. R. Henry, and J. Hofrichter. Fast kinetics and mechanisms in protein folding. *Annu Rev Biophys Biomol Struct.*, 29:327–359, 2000.

[32] L.J. Lapidus, W. A. Eaton, and J. Hofrichter. Measuring the rate of intramolecular contact formation in polypeptides. *Proc. Natl. Acad. Sci.*, 97:7220–7225, 2000.

[33] D. R. Roberts and E. H. White. Energy transfer in chemiluminescence. iii. intramolecular triplet-singlet transfer in derivatives of 2,3-dihydrophthalazine-1,4-dione. *J. Am. Chem. Soc.*, 92 (16):4861–4867, 1970.

[34] K. Sudhakar, C. M. Phillips, C. S. Owen, and J. M. Vanderkooi. Dynamics of parvalbumin studied by fluorescence emission and triplet absorption spectroscopy of tryptophan. *Biochemistry.*, 34 (4):1355–1363, 1995.

[35] G. B. Strambini and M. Gonnelli. Tryptophan phosphorescence in fluid solution. *J. Am. Chem. Soc.*, 117:7646–7651, 1995.

[36] D. V. Bent and E. Hayon. Excited state chemistry of aromatic amino acids and related peptides: Iii. tryptophan. *J. Am. Chem. Soc.*, 97:2612–2619, 1995.

[37] W. A. Volkert, R. R. Kuntz, C. A. Ghiron, R. F. Evans, R. Santus, and M. Bazin. Flash photolysis of tryptophan and n-acetyl-l-tryptophanamide: the effect of bromide on transient yields. *Photochem. Photobiol.*, 26:3–9, 1977.

[38] L. J. Lapidus, P. J. Steinbach, W. A. Eaton, A. Szabo, and J. Hofrichter. Effects of chain stiffness on the dynamics of loop formation in polypeptides. appendix: Testing a 1-dimensional diffusion model for peptide dynamics. *J. Phys. Chem.B.*, 106:11628–11640, 2002.

[39] L. J. Lapidus, W. A. Eaton, and J. Hofrichter. Dynamics of intramolecular contact formation in polypeptides: distance dependence of quenching rates in a room-temperature glass. *Phys. Rev. Lett.*, 87:258101–1–258101–4, 2001.

[40] E. Amouyal, A. Bernas, and D. Grand. On the photoionization energy threshold of tryptophan in aqueous solutions. *Photochem. Photobiol.*, 29:1071–1077, 1979.

[41] D. B. Calhoun, W. S. Englander, W. W. Wright, and J. M. Vanderkooi. Quenching of room temperature protein phosphorescence by added small molecules. *Biochemistry.*, 27:8466–8474, 1988.

[42] M. Gonnelli and G. B. Strambini. Phosphorescence lifetime of tryptophan in proteins. *Biochemistry.*, 34:13847–13857, 1995.

[43] A. Szabo, K. Schulten, and Z. Schulten. 1st passage time approach to diffusion controlled reactions. *J. Chem. Phys.*, 72:4350–4357, 1980.

[44] V. R. Singh, M. Kopka, Y. Chen, W.J. Wedemeyer, and L. J. Lapidus. Dynamic similarity of the unfolded states of proteins l and g. *Biochemistry.*, 46 (35):10046–10054, 2007.

[45] S. Kmiecika and A. Kolinski. Folding pathway of the b1 domain of protein g explored by multiscale modeling. *Biophysical Journal.*, 94 (3):726–736, 2008.

[46] T Cellmera, R. Doumaa, A. Huebnera, J. Prausnitza, and H. Blanch. Kinetic studies of protein l aggregation and disaggregation. *Biophysical Chemistry.*, 125 (2-3):350–359, 2007.

[47] S. A. Waldauer, O. Bakajin, T. Ball, Y. Chen, S. J. DeCamp, M. Kopka, M. Jager, V. R. Singh, W. J. Wedemeyer, S. Weiss, S. Yao, and L. J. Lapidus. Ruggedness in the folding landscape of protein l. *HFSP J.*, 2 (6):388–395, 2008.

[48] H.Y. Zoghbi and H.T. Orr. Glutamine repeats and neurodegeneration. *Annu. Rev. Neurosci.*, 23:217–247, 2000.

[49] E. Scherzinger. Huntingtin-encoded polyglutamine expansions form amyloid-like protein aggregates in vitro and in vivo. *Cell.*, 90:549–558, 1997.

[50] V. R. Singh and L. J. Lapidus. The intrinsic stiffness of polyglutamine peptides. *J. Phys. Chem. B.*, 112 (42):13172–13176, 2008.

[51] S. Chen and R. Wetzel. Solubilization and disaggregation of polyglutamine peptides. *Protein Sci.*, 10 (4):887–891, 2001.

[52] M. Buscaglia, L. J. Lapidus, W. A. Eaton, and J. Hofrichter. Effects of denaturants on the dynamics of loop formation in polypeptides. *Biophys. J.*, 91 (1):276–288, 2006.

[53] F. Huang and W. M. Nau. A conformational flexibility scale for amino acids in peptides. *Angew. Chem., Int. Ed.*, 42 (20):2269–2272, 2003.

[54] S. Barton, R. Jacak, S. D. Khare, F. Ding, and N. V. Dokholyan. The length dependence of the polyq-mediated protein aggregation. *J. Biol. Chem.*, 282 (35):25487–25492, 2007.

[55] Y. W. Chen, K. Stott, and M. F. Perutz. Crystal structure of a dimeric chymotrypsin inhibitor 2 mutant containing an inserted glutamine repeat. *Proc. Natl. Acad. Sci.*, 96 (4):1257–1261, 1999.

[56] S. Chen, V. Berthelier, W. Yang, and R. Wetzel. Polyglutamine aggregation behavior in vitro supports a recruitment mechanism of cytotoxicity. *J. Mol. Biol.*, 311:173–182, 2001.

[57] M.F. Perutz. Glutamine repeats and neurodegenerative diseases: molecular aspects. *Trends Biochem. Sci.*, 24:58–63, 1999.

[58] S.D. Khare, F. Ding, K. N. Gwanmesia, and N. V. Dokholyan. Molecular origin of polyglutamine-mediated protein aggregation in neurodegenerative diseases. *PLoS Computational Biology.*, 1:230–235, 2005.

[59] A. J. Marchut and C. K. Hall. Effects of chain length on the aggregation of model polyglutamine peptides: molecular dynamics simulations. *Proteins.*, 66 (1):96–109, 2007.

[60] S. L. Crick, M. Jayaraman, C. Frieden, R. Wetzel, and R. V. Pappu. Fluorescence correlation spectroscopy shows that monomeric polyglutamine molecules form collapsed structures in aqueous solutions. *Proc. Natl. Acad. Sci.*, 103 (45):16764–16769, 2006.

[61] X. Wang, A. Vitalis, M. A. Wyczalkowski, and R.V. Pappu. Characterizing the conformational ensemble of monomeric polyglutamine. *Proteins: Struct. Funct. Bioinf.*, 63 (2):297–311, 2006.

[62] R. S. Armen, B. M. Bernard, R. Day, D. Alonso, and V. Daggett. Characterization of a possible amyloidogenic precursor in glutamine-repeat neurodegenerative diseases. *Proc. Natl. Acad. Sci.*, 102 (38):13433–13438, 2005.

[63] J. M. Finke, M. S. Cheung, and J. N. Onuchic. A structural model of polyglutamine determined from a host-guest method combining experiments and landscape theory. *Biophys. J.*, 87 (3):1900–1918, 2004.

[64] S. Chen, F. A. Ferrone, and R. Wetzel. Huntington's disease age-of-onset linked to polyglutamine aggregation nucleation. *Proc. Natl. Acad. Sci.*, 99 (18):11884–11889, 2002.

[65] F. A. Klein, A. Pastore, L. Masino, G. Zeder-Lutz, H. Nierengarten, M. Oulad-Abdeighani, D. Altschuh, J. L. Mandel, and Y. Trottier. Pathogenic and non-pathogenic polyglutamine tracts have similar structural properties: towards a length-dependent toxicity gradient. *J Mol Biol.*, 371 (1):235–244, 2007.

[66] M. L. Scalley, Q. Yi, H. Gu, A. McCormack, J. R. Yates III, and D. Baker. Kinetics of folding of the igg binding domain of peptostreptoccocal protein l. *Biochemistry.*, 36 (11):3373–3382, 1997.

[67] S. H. Park, K. T. O'Neil, and H. Roder. An early intermediate in the folding reaction of the b1 domain of protein g contains a native-like core. *Biochemistry.*, 36 (47):14277–14283, 1997.

[68] S. H. Park, M. C. Shastry, and H. Roder. Folding dynamics of the b1 domain of protein g explored by ultrarapid mixing. *Nat. Struct. Biol.*, 6 (10):943–947, 1999.

[69] M. L. Scalley, S. Nauli, S. T. Gladwin, and D. Baker. Structural transitions in the protein l denatured state ensemble. *Biochemistry.*, 38 (48):15927–15935, 1999.

[70] Q. Yi, M. L. Scalley, K. T. Simons, S. T. Gladwin, and D. Baker. Characterization of the free energy spectrum of peptostreptococcal protein l. *Fold. Des.*, 2 (5):271–280, 1997.

[71] M. Buscaglia, B. Schuler, L. J. Lapidus, W. A. Eaton, and J. Hofrichter. Kinetics of intramolecular contact formation in a denatured protein. *J. Mol. Biol.*, 332 (1):9–12, 2003.

[72] S. J. Hagen, J. Hofrichter, A. Szabo, and W. A. Eaton. Diffusion-limited contact formation in unfolded cytochrome c: Estimating the maximum rate of protein folding. *Proc. Natl. Acad. Sci.*, 93 (21):11615–11617, 1996.

[73] S. J. Hagen, L. L. Qiu, and S. A. Pabit. Diffusional limits to the speed of protein folding: fact or friction? *J. Phys.: Condens. Matter.*, 17:S1503–S1514, 2005.

[74] D. Nettels, I. V. Gopich, A. Hoffmann, and B. Schuler. Ultrafast dynamics of protein collapse from single-molecule photon statistics. *Proc. Natl. Acad. Sci.*, 104 (8):2655–2660, 2007.

[75] A. Moglich, K. Joder, and T. Kiefhaber. End-to-end distance distributions and intrachain diffusion constants in unfolded polypeptide chains indicate intramolecular hydrogen bond formation. *Proc. Natl. Acad. Sci.*, 103:12394–12399, 2006.

[76] J. Wang, P. Cieplak, and P. A. Kollman. How well does a resp (restrained electrostatic potential) model do in calculating the conformational energies of organic and biological molecules? *J. Comput. Chem.*, 21:1049–1074, 2000.

[77] V. Tsui and D. A. Case. Theory and applications of the generalized born solvation model in macromolecular simulations. *Biopolymers.*, 56:275–291, 2001.

[78] A. Onufriev, D. Bashford, and D. Case. Exploring native states and large-scale conformational changes with a modified generalized born model. *Proteins.*, 55:383–394, 2004.

[79] L. Stellaa and S. Melchionnab. Equilibration and sampling in molecular dynamics simulations of biomolecules. *J. Chem. Phy.*, 109 (23):10115–10117, 1998.

[80] H. S. Choi, J. Huh, and W. H. Jo. Comparison between denaturant- and temperature-induced unfolding pathways of protein: A lattice monte carlo simulation. *Biomacromolecules.*, 5 (6):2289–2296, 2004.

[81] K. A. Dill. Theory for the folding and stability of globular proteins. *Biochemistry.*, 24:1501–1509, 1985.

[82] G. Ziv and G. Haran. Protein folding, protein collapse, and tanfords transfer model: lessons from single-molecule fret. *J. Am. Chem. Soc.*, 131:2942–2947, 2009.

[83] H. J. Dyson and P. E. Wright. Unfolded proteins and protein folding studied by nmr. *Chem. Rev.*, 104 (8):3607–3622, 2004.

[84] A. Y. Grosberg and Khokhlov A. R. *Statistical Physics of Macromolecules.* AIP Press, 1994.

[85] C. Tanford. *Physical Chemistry of Macromolecules.* Wiley, New York, 1961.

[86] H. Zhou. A gaussian-chain model for treating residual chargecharge interactions in the unfolded state of proteins. *Proc. Natl. Acad. Sci.*, 99 (6):3569–3574, 2002.

[87] K. A. Dill, D. O. V. Alonso, and K. Hutchinson. Thermal stabilities of globular proteins. *Biochemistry.*, 28:5439–5449, 1989.

[88] P. J. Flory and S. Fisk. Effect of volume exclusion on the dimensions of polymer chains. *J. Chem. Phy.*, 44:2243–2248, 1966.

[89] E. Sherman and G. Haran. Coilglobule transition in the denatured state of a small protein. *Proc. Natl. Acad. Sci.*, 103 (31):11539–11543, 2006.

[90] K. A. Merchant, R. B. Best, J. M. Louis, I. V. Gopich, and W. A. Eaton. Characterizing the unfolded states of proteins using single-molecule fret spectroscopy and molecular simulations. *Proc. Natl. Acad. Sci.*, 104 (5):1528–1533, 2007.