ABSTRACT

Title of dissertation: Protein folding and amyloid formation in various environments

Edward P. O'Brien, Doctor of Philosophy, 2008

Dissertation directed by: Devarajan Thirumalai
Department of Chemistry and Biochemsitry
Bernard R. Brooks
National Institutes of Health

Understanding and predicting the effect of various environments that differ in terms of pH and the presence of cosolutes and macromolecules on protein properties is a formidable challenge. Yet this knowledge is crucial in understanding the effect of cellular environments on a protein. By combining thermodynamic theories of solution condition effects with statistical mechanics and computer simulations we develop a molecular perspective of protein folding and amyloid formation that was previously unobtainable. The resulting Molecular Transfer Model offers, in some instances, quantitatively accurate predictions of cosolute and pH effects on various protein properties. We show that protein denatured state properties can change significantly with osmolyte concentration, and that residual structure can persist at high denaturant concentrations. We study the single molecule mechanical unfolding of proteins at various pH values and varying osmolyte and denaturant concentrations. We find that the the effect of varying solution conditions on a protein under tension can be understood and qualitatively predicted based on knowledge of that

protein's behavior in the absence of force. We test the accuracy of FRET inferred denatured state properties and find that currently, only qualitative estimates of denatured state properties can be obtained with these experimental methods. We also explore the factors governing helix formation in peptides confined to carbon nanotubes. We find that the interplay of the peptide's sequence and dimensions, the nanotube's diameter, hydrophobicity and chemical heterogeneity, lead to a rich diversity of behavior in helix formation. We determine the structural and thermodynamic basis for the dock-lock mechanism of peptide deposition to a mature amyloid fibril. We find multiple basins of attraction on the free energy surface associated with structural transitions of the adding monomer. The models we introduce offer a better understanding of protein folding and amyloid formation in various environments and take us closer to understanding and predicting how the complex environment of the cell can effect protein properties.

Protein folding and amyloid formation in various environments

by

Edward P. O'Brien

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2008

Advisory Committee:
Professor Devarajan Thirumalai, Chair/Advisor
Dr. Bernard Brooks, Co-advisor
Professor Christopher Jarzynski
Professor John Weeks
Professor Dorothy Beckett

© Copyright by
Edward P. O'Brien
2008

Dedication

This dissertation is dedicated to my parents, Maryann and Edward 'Obie' O'Brien, and to my wife Stephanie.

Acknowledgments

There are many people I need to thank in bringing the past six years of my dissertation research to a successful completion. First to my parents, whose constant love, support and encouragement throughout my life has allowed me to flourish. To Stephanie, my wife, who was always there for me when I needed her and did not demand attention when my attention needed to be focused on research. To my sister's, Megan and Kara, whom, by virtue of their age and maturity, have always led the way in the various stages of life and been excellent role models on this journey.

I thank Dmitri Klimov (now a Professor at George Mason University), whose patience and guidance as a mentor the first two years of my research helped immensely. Changbong Hyeon, whose conversations I have almost always found to be enlightening. Greg Morrison, whose unflagging willingness to discuss research ideas and aid in solving complex mathematics was valued. Guy Ziv, whom I had many valuable conversations with ranging from modeling to the intricacies of single molecule FRET experiments. Subramanian Vaitheeswaran, Govardhan Reddy, Riina Tehver, David Pincus, Sam Cho and other Thirumalai group members who were always useful in bouncing research ideas off of and offering advice on manuscripts. George Stan, Lee Woodcock, Richard Pastor, and Herbert Geller, who were all kind enough to offer invaluable career advice, especially related to my post-doctoral research grant proposal. Ruxandra Dima, with whom I carried out a research project early in my graduate career. Tim Miller and Mary Cornio at the NIH, without whom

my graduate research and life would have been much more difficult. Tim Miller, the information technology guru in the Brooks' group, was always great at keeping the computer cluster running and offering help when computer crises struck. Mary, the administrative assistant at NIH, navigated me through the bureaucracy at NIH and often took on the onerous task of making sure all of my paper work was filled out and in order. I thank Prof. David Wayne Bolen (Univ. Texas Medical Branch) for many useful conversations over the past several years related to modeling of osmolyte and pH effects on proteins.

Prof. Yuko Okamoto (Nagoya University, Japan) was kind enough to host me in his lab for three months during the summer of 2007 where he taught me a variety of Replica exchange sampling methods. We had many enjoyable discussions on the history and politics of Japan. His generosity, and the kindness and humility of the Japanese people has left a lasting impression on me.

I thank Dr. Bernard Brooks, who has been an invaluable co-advisor on my doctoral research. Bernie was always willing to spend countless hours answering questions and giving me advice on computational methods. In addition, Bernie has been an avid supporter of my career development. He has supported me in going to numerous conferences and seminars. Bernie was very helpful in discussing options for my future career path, including post-doctoral positions and beyond.

I thank Prof. Dave Thirumalai, my advisor, who set high research standards for me and helped extinguish my youthful impetuousness. Dave has been an excellent role model in how to think about research problems and how to write good research papers.

Finally, I would like to acknowledge financial support from the National Institutes of Health and the National Science Foundation.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

FRET	Föster Resonance Energy Transfer
GRM	Generalized Rouse Model
NBA	Native Basin of Attraction
NSE	Native State Ensemble
DSE	Denatured State Ensemble
MREX	Multidimensional Replica Exchange
WHAM	Weighted Histogram Analysis Method
TTM	Tanford Transfer Model
MTM	Molecular Transfer Model
SSE	Secondary Structural Element
CG	Coarse grain
SASA	Solvent Accessible Surface Area
$C_\alpha - SCM$	C_α-Side Chain Model
CI2	Chymotrypsin Inhibitor 2
CspTm	Cold Shock Protein
TMAO	Trimethylamine-N Oxide
GdmCl	Guandinium Chloride
PMF	Potential of Mean Force
GSC	Gaussian Self-Consistency Test
R_g	Radius of Gyration
R_{ee}	End-to-End Distance
f	Constant Pulling Force

Chapter 1

Introduction

1.1 Protein Folding and Amyloid Formation

How proteins fold into an ordered structure known as the native state from the disordered structural ensemble of the denatured state has been a question of interest for over 55 years [2]. Over the past five decades, a significant amount of knowledge and understanding of the process of protein folding and the factors governing it has been gained [3]. A central finding that has emerged is that a combination of a protein's amino acid sequence and the environment encode a protein's native structure [4, 5]. In the 1970's and early 1980's, it was shown that the process of protein folding is energetically biased towards the native state and is stochastic [5, 6], which means that multiple folding pathways between the denatured and native states are possible [7, 6]. Assuming that the native state is the lowest free energy structure [5], it was suggested that native state topology determines folding pathways [8], meaning that conformational fluctuations in the denatured state towards the native state are energetically favored [8, 5] and therefore the free energy surface is funnel-like [9, 10, 11]. These findings, coupled with statistical mechanics [12, 13, 6], were incorporated into a perspective on protein folding referred to as the energy landscape picture [9, 10, 11].

A large focus of protein folding research has been on the behavior of proteins at

ideal, infinite dilution conditions where no other solutes or cosolutes are present [3]. While this focus has provided a wealth of invaluable insights, protein folding in vivo occurs under non-ideal conditions where other proteins and cosolutes are present at non-negligible concentrations [14, 15] and the pH can vary depending on a proteins location [16, 17, 18]. These non-ideal conditions can lead to protein misfolding and aggregation, which can be deleterious to a cell. If the protein clearance system in a cell fails, misfolded proteins can aggregate and form ordered structures referred to as amyloid. Amyloid is associated with over forty different human diseases [19]. Therefore, it is important to understand the structural and thermodynamic basis of amyloid formation, as well as the interplay of solution conditions.

For these reasons a number of questions remain unanswered, including (1) Can we predict the effect of various solution conditions (osmolyte, denaturant and pH effects) on protein properties? (2) How does the denatured state vary with solution conditions? (3) Does the presence of other macromolecules impact amyloid formation? In this dissertation, we develop a molecular perspective of protein folding and amyloid formation, that was previously unobtainable, by combining thermodynamic theories of solution condition effects with statistical mechanics and computer simulations. We discuss these macroscopic theories in the next section. Then in Section 1.3 we discuss the novel method we propose to combine them with a statistical physics perspective.

1.2 Theoretical Background

1.2.1 Modeling denaturant and osmolyte effects on proteins

In 1961 it was found that some cosolutes referred to as denaturants can *destabilize* the native state of a protein in direct proportion to that denaturants concentration [C] in solution [2, 20], i.e. $\Delta\Delta G_{ND} = m[C]$, where $m(\equiv d\Delta G_{ND}/d[C])$ is a constant of proportionality and is conventionally referred to as the 'm-value' [21].* In 1963, it was found that the free energy cost, denoted δg, of transferring model compounds that mimic individual amino-acids from water to denaturant solution conditions was *also* directly proportional to [C] [23]. This suggested that the mechanism by which denaturants act on individual amino acids is the *same* mechanism by which denaturants act on proteins, and therefore $\Delta\Delta G_{ND} \propto \delta g$. In addition, the measured δg values were found to be so small that, from a ligand binding perspective, the binding constant of denaturant molecules for the model compounds was on the order of 0.1 M [24, 25]. Such a weak binding affinity suggested that to a first approximation, the number of ligand binding sites on an amino acid (or protein) was proportional to that amino acid's solvent accessible surface area (α).

Charles Tanford utilized this information to develop a phenomenological model to estimate the free energy of transferring a protein in conformation state l, defined

*It was found later that another class of cosolutes, referred to as counteracting osmolytes, can *stabilize* a protein's native state in direct proportion to their concentration in solution [14, 22]. The Tanford transfer model, discussed later in this section with regards to denaturants, applies equally well to osmolytes.

by the coordinates of the atoms in the protein, from water to a solution containing denaturant at concentration [C]. This free energy cost is denoted $\Delta G_{tr}(l,[C])$ [26]. In the 'Tanford Transfer Model' (TTM)

$$\Delta G_{tr}(l,[C]) = \sum_{i=1}^{N_S} \delta g_i^{S'}(l,[C]) + \sum_{i=1}^{N_B} \delta g_i^{B'}(l,[C]) \quad (1.1)$$

$$= \sum_{i=1}^{N_S} \frac{\alpha_i^S(l)}{\alpha_{G-i-G}^S} \delta g_i^S([C]) + \sum_{i=1}^{N_B} \frac{\alpha_i^B(l)}{\alpha_{G-i-G}^B} \delta g^B([C]), \quad (1.2)$$

where the summations are over the N_S and N_B side chain (S) and backbone (B) groups of the protein, respectively. $\delta g_i^{P'}(l,[C])$, in Eq. 1.1, is the free energy cost of transferring group $P(= S$ or $B)$ of residue i in protein conformation l from water to a solution containing [C] M denaturant. $\delta g_i^P([C])$, in Eq. 1.2, is the free energy cost of transferring group P in a model compound of amino acid type i from water to a solution containing [C] M denaturant. $\alpha_i^P(l)$ is the solvent accessible surface area of group P of amino acid type i in protein conformation l. α_{G-i-G}^P is the solvent accessible surface area of group P of amino acid type i in the tripeptide $Gly-i-Gly$, which is the model compound used to experimentally measure δg_i^P.

Comparing Eqs. 1.1 and 1.2 it is clear that $\delta g_i^{P'}(l,[C]) = \frac{\alpha_i^P(l)}{\alpha_{G-i-G}^P} \delta g_i^P([C])$, i.e. the transfer free energy of amino acid i, when it is part of a protein, is equal to its solvent accessible surface area in the protein divided by its solvent accessible surface area when it is in the model compound, multiplied by that model compound's experimentally measured transfer free energy. Thus, when reside i is fully exposed to solvent $\alpha_i^P(l) = \alpha_{G-i-G}^P$ the ratio $\frac{\alpha_i^P(l)}{\alpha_{G-i-G}^P} = 1$ and $\delta g_i^{P'}(l,[C]) = \delta g_i^P([C])$. On the other hand, when residue i is completely buried in the protein, it is not in direct contact with denaturant molecules and $\alpha_i^P(l) = 0$ and $\delta g_i^{P'}(l,[C]) = 0$.

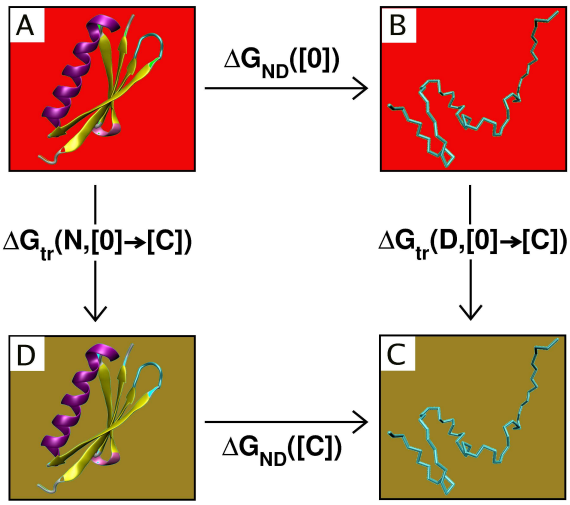


Figure 1.1: Using this thermodynamic cycle, it can be shown that $\Delta\Delta G_{ND} = \Delta G_{tr}(N,[C]) - \Delta G_{tr}(D,[C])$. In states labeled A and D, the protein is folded. In states labeled B and C, the protein is unfolded. The change in stability upon going from the native to the denatured state in the absence of cosolutes (i.e. going from state A to B) is labeled $\Delta G_{ND}([0])$ and in the presence of cosolutes (i.e. going from state D to C) is labeled $\Delta G_{ND}([C])$. The transfer of the native or denatured states from water to aqueous cosolute solution are labeled, respectively, $\Delta G_{tr}(N,[C])$ and $\Delta G_{tr}(D,[C])$.

For apparent two-state folding proteins, which exist in either native or denatured ensembles, the thermodynamic cycle shown in Fig. 1.1 requires that $\Delta\Delta G_{ND} = m[C] = \Delta G_{tr}(N,[C]) - \Delta G_{tr}(D,[C])$ [26, 27]. Inserting Eq. 1.2 into this result, we find that

$$m[C] = \sum_{i=1}^{N_S} \frac{\Delta\alpha_i^S}{\alpha_{G-i-G}^S} \delta g_i^S([C]) + \sum_{i=1}^{N_B} \frac{\Delta\alpha_i^B}{\alpha_{G-i-G}^B} \delta g^B([C]), \tag{1.3}$$

where $\Delta\alpha_i^P = \alpha_i^P(N) - \alpha_i^P(D)$. Thus, by knowing a protein's amino acid sequence, its native structure, and using experimentally measured δg data, the Tanford Transfer Model can be used to predict $\Delta\Delta G_{ND}$, or equivalently the m-value, for the denaturant (or osmolyte) of interest [26, 27]. It was not until 2004, when Wayne Bolen and colleagues overcame several experimental hurdles [28, 29], that the TTM

was finally shown [27] to quantitatively predict $\Delta\Delta G_{ND}$ for a large number of proteins (Fig. 1.2). This experimental validation of the TTM is important because it gives insight into the forces governing denaturant and osmolyte effects on proteins. We realized that it also means that if the partition function of a protein in water ($Z([0])$) is known, then the partition function in [C] M solution can be computed as $Z([C]) = \sum_j e^{-\beta E(j,[0]) - \beta \Delta G_{tr}(j,[C])}$. Thus, the effect of *any* osmolyte at *any* [C] on a protein's thermodynamic properties can be predicted provided $Z([0])$ is known accurately. We will utilize this fact in Section 1.3.

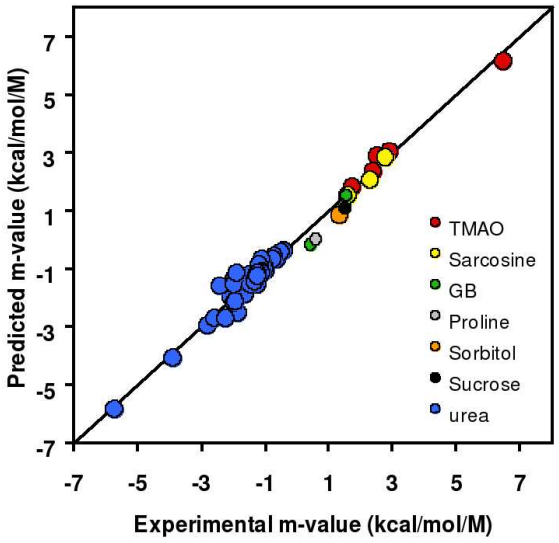


Figure 1.2: A comparison of predicted m-values (using Eq. 1.3) versus experimentally measured m-values is shown as circles for the different cosolutes listed in the legend. Note that denaturants have negative m-values while counteracting osmolytes have positive m-values. The solid line is used to illustrate a 1-to-1 correspondence between predicted and measured m-values. This figure was generously provided by D. Wayne Bolen.

1.2.2 Modeling pH effects on proteins

pH, the $\log_{10}$ of the proton concentration ($[H^+]$) in solution, was shown to have large effects on protein properties as early as 1951 [30, 31]. Aune and Tanford developed one of the most widely used theories to quantitatively account for pH effects on protein stability [32]. Using the well known Wyman Linkage result $\frac{dlog\{K_{NU}\}}{dpH} = \Delta Q$ [33], Tanford showed that they could fit an experimentally measured ΔG_{NU} vs. pH profile using the equation $\Delta\Delta G_{ND}(pH) = -2.3RT \int_{pH_1}^{pH_2} \Delta Q(pH) dpH$, where $\Delta Q(pH) = \langle Q_N(pH) \rangle - \langle Q_D(pH) \rangle$, which is the difference in the average number of protons (Q) bound to the native and denatured states, respectively. They also showed that by using the Henderson-Hasselbach equation, the $\Delta\Delta G_{ND}$ vs. pH profile could be *predicted* based solely on the knowledge of the titratible groups proton binding constants [34], referred to as pK_a values, via

$$\begin{aligned}
\Delta\Delta G_{ND}(pH_1 \rightarrow pH_2) &= \Delta G_{tr}(N, pH_1 \rightarrow pH_2) - \Delta G_{tr}(D, pH_1 \rightarrow pH_2) \quad (1.4)\\
&= \sum_{k=1}^{N_t} \delta g_k(N, pH_1 \rightarrow pH_2) - \sum_{k=1}^{N_t} \delta g_k(D, pH_1 \rightarrow pH_2) \quad (1.5)\\
&= -k_B T \sum_{k=1}^{N_t} ln \left[\frac{10^{pH_2} + 10^{pK_{k,N}}}{10^{pH_1} + 10^{pK_{k,N}}} \right]\\
&\quad + k_B T \sum_{k=1}^{N_t} ln \left[\frac{10^{pH_2} + 10^{pK_{k,D}}}{10^{pH_1} + 10^{pK_{k,D}}} \right] \quad (1.6)
\end{aligned}$$

where $\Delta\Delta G_{ND}(pH_1 \rightarrow pH_2)$ is the change in free energy of ΔG_{ND} upon a change in pH from a value of pH_1 to pH_2. $\Delta G_{tr}(l, pH_1 \rightarrow pH_2)$ is the free energy cost of transferring the l^{th} protein conformation from pH_1 to pH_2, where l, for a two state system, is limited to N or D. Comparing Eqs. 1.4 and 1.6, it can be seen that $\Delta G_{tr}(l, pH_1 \rightarrow pH_2) = -k_B T \sum_{k=1}^{N_t} ln \left[\frac{10^{pH_2} + 10^{pK_{k,l}}}{10^{pH_1} + 10^{pK_{k,l}}} \right]$ where the summation is over

the N_t titratible groups and $pK_{k,l}$ is the pK_a value of the k^{th} titratible group in the l^{th} protein conformation.

Eq. 1.6 was shown by Bashford and Karplus [35] to be the mean field result of integrating over all possible protonation states of a protein with N_t titratible groups in the native and denatured states. The success of Eq. 1.6 in modeling experimental ΔG_{ND} vs. pH data not only offers insight into the mechanism of pH denaturation[†], it also means that the free energy cost of transferring individual protein conformations from one solution pH to another can be estimated.

1.3 Computational Background

Simulating protein folding involves a system with a large number of degrees of freedom. As such, ergodically sampling the configurational space of the model's Hamiltonian is a formidable challenge and is often intractable with current computer resources. However, achieving ergodicity in simulations is required to justify the use of thermodynamics and equilibrium statistical mechanics in the analysis of molecular simulations, and for a valid comparison to experiments that are at equilibrium. There are a number of methods to enhance sampling in molecular simulations, including so-called coarse graining of the system [10], Multidimensional Replica Exchange [36], and post-simulation techniques such as the Weighted Histogram Analysis Method [37]. We describe these methods below and show how

[†]Eq. 1.6 implies that pH denatures proteins by the excess number of protons bound to the denatured state as compared to the native state and not necessarily by charge repulsion between like-charged groups.

they can be combined in a model we developed called the Molecular Transfer Model (MTM) that allows osmolyte, denaturant, and pH effects to be accurately modeled and connected to the underlying ensemble of protein conformations.

1.3.1 Coarse grained models

An approach used extensively in this dissertation is to remove 'non-essential' degrees of freedom from the system and thereby reduce the dimensionality of phase space and coarse grain (CG) the structural resolution of the model [10, 38]. Deciding what features of a model are essential depends on the questions you want to answer. For the purposes of this thesis, we use a utilitarian definition that non-essential features are those that can be removed such that the resulting model still exhibits the phenomenon of interest. For example, in our CG model of proteins, individual amino acids are represented as just one or two interaction sites instead of explicitly representing all of the atoms (Fig. 1.3). This coarse graining allows most of the features of protein folding to be retained, including properties that are experimentally measured. What is lost in terms of structural resolution in this CG is made up for by achieving ergodicity that allows us to apply the tools of thermodynamics and equilibrium statistical mechanics in our analysis.

1.3.2 Multidimensional Replica Exchange

To significantly enhance sampling during a simulation, Multidimensional Replica Exchange (MREX) can be used [36, 39]. In MREX, simulations (referred to as repli-

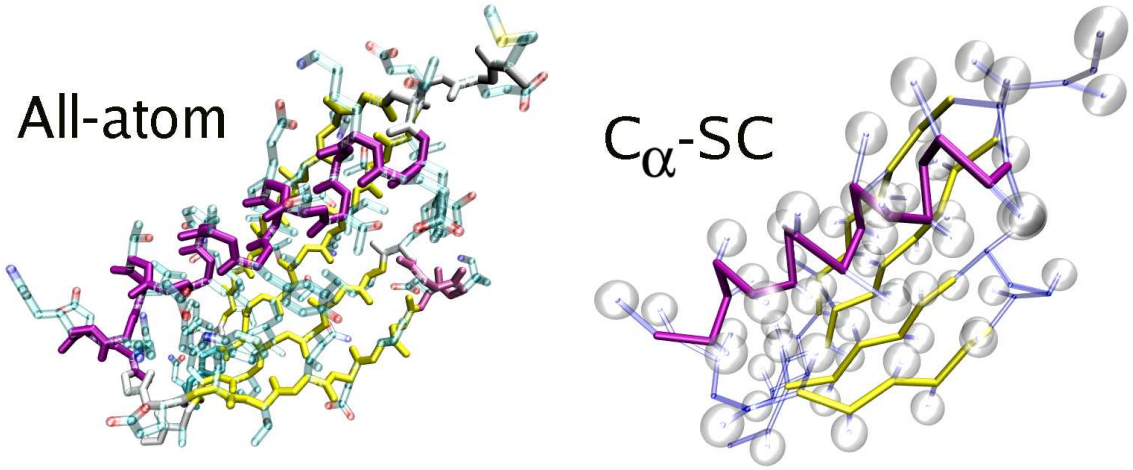


Figure 1.3: (Left) An all-atom model and (Right) C_α-side chain model of protein L. Achieving ergodicity in all-atom models of globular proteins is currently not possible. We use coarse-grained models to achieve an effective ergodicity.

cas) at different temperatures or evolving under different Hamiltonians are simulated simultaneously in their respective NVT ensemble. These replicas are periodically allowed to swap their system coordinates with other replicas at different temperatures or Hamiltonians while preserving detailed balance. The probability of swapping between replicas i and j ($P(i,j)$) in MREX uses the standard Metropolis criterion

$$P(i,j) = \begin{cases} 1, & for \quad \Delta \leq 1 \\ exp(-\Delta), & for \quad \Delta > 1. \end{cases} \tag{1.7}$$

For swapping between replicas i and j that are at different temperatures, but have the same Hamiltonian denoted k, $\Delta = (\beta_i - \beta_j)(E_k(j) - E_k(i))$ in Eq. 1.7, where $E_k(i)$ and $E_k(j)$ are the potential energies of the system coordinates of replicas i and$j$ using Hamiltonian k and $\beta_i = \frac{1}{k_B T_i}$[36]. For swapping between replicas i and j that have different Hamiltonians (labeled k and l respectively) but are at the same temperature denoted m, $\Delta = \beta_m(E_k(i) - E_l(i) + E_k(j) - E_l(j))$ in Eq. 1.7. Just

as in traditional Monte Carlo simulations, these acceptance criteria preserve the underlying thermodynamic ensemble. The swapping in MREX enhances sampling by allowing the replicas to perform a random walk in temperature and Hamiltonian space. Thus, free energy barriers that might inhibit sampling at low temperatures or in a given Hamiltonian can usually be overcome when the replica is swapped to higher temperatures or to a different Hamiltonian [36].

1.3.3 The Weighted Histogram Analysis Method

The partition function Z in statistical mechanics is the central quantity that connects the molecular configurations of a system with that system's experimentally measured thermodynamic properties. Z cannot be computed analytically for most protein model Hamiltonians. However, Z can be inferred based on the time series of potential energies from a simulation. We compute Z using the Weighted Histogram Analysis Method (WHAM) [40, 37]. In WHAM, time series data from simulations at various temperatures and Hamiltonians are used to obtain an optimal estimate of the density of states. WHAM does this by self-consistently solving for a free energy weighting term (F_m in Eq. 1.8 below) which minimizes the error associated with the density of states estimate. The resulting partition function[‡] is

$$Z(T_i) = \sum_{k=1}^{R} \sum_{t=1}^{n_k} \frac{e^{-\beta_i E_P(k,t)}}{\sum_{m=1}^{R} n_m e^{F_m - \beta_m E_P(k,t)}} \tag{1.8}$$

[‡]For the sake of brevity we present the WHAM partition function for simulations under one Hamiltonian and different temperatures, the reader is referred to Chapter 3 for the equation for multiple temperatures and multiple Hamiltonians.

where $Z(T_i)$ is the partition function at temperature T_i, the summations are over the R simulations at R different temperatures, and the n_k are data points of the time series from the k^{th} simulation. $E_P(k,t)$ is the potential energy of the t^{th} data point from the k^{th} simulation. In the denominator, the summation is over the R simulations at R different temperatures. n_m is the number of data points saved during the simulation at the m^{th} temperature. F_m is the free energy of the m^{th} simulation and is solved for self-consistently [37] and $\beta_m = \frac{1}{k_B T_m}$.

1.3.4 The Molecular Transfer Model

As noted in Section 1.2, if we know $Z(A)$, which is the partition function at solution condition A[§], and we know ΔG_{tr} for each protein conformation (microstate) upon transfer from A to solution condition B, then $Z(B)$ is also known. This means that by achieving effectively ergodic simulations (using CG and MREX) at one solution condition, $Z(A)$ can be computed using WHAM (Eq. 1.8) and $Z(B)$ predicted using the ΔG_{tr} models described in Section 1.2. This approach, which we refer to as the Molecular Transfer Model (MTM), expresses the partition function as

$$Z(B) \;=\; \sum_{k=1}^{R}\sum_{t=1}^{n_k} \frac{e^{-\beta_B E_P(k,t) - \beta_B \Delta G_{tr}(k,t,B)}}{\sum_{m=1}^{R} n_m e^{F_m - \beta_m E_P(k,t)}}, \tag{1.9}$$

where all terms in Eq. 1.9 are the same as Eq. 1.8 except for the term $\Delta G_{tr}(k,t,B)$. $\Delta G_{tr}(k,t,B) = \Delta G_{tr}(k,t,[0] \rightarrow [C_B]) + \Delta G_{tr}(k,t,pH_A \rightarrow pH_B)$, where $\Delta G_{tr}(k,t,0M \rightarrow$

[§]where A is uniquely defined by specifying the temperature (T_A), pH (pH_A), osmolyte type, and osmolyte concentration ($[C_A]$), of solution.

$[C_B]$) is the free energy cost of transferring the k^{th} protein conformation from the t^{th} simulation from 0 M cosolute solution to [C_B] M cosolute solution of cosolute type B. $\Delta G_{tr}(k,t,pH_A \rightarrow pH_B)$ is the free energy cost of transferring that same conformation from the pH at which the simulation was carried out (denoted pH_A) to a solution at pH value pH_B. We use Eqs. 1.2 and 1.6 to model $\Delta G_{tr}(k,t,[0] \rightarrow [C_B])$ and $\Delta G_{tr}(k,t,pH_A \rightarrow pH_B)$, respectively.

We emphasize that $\Delta G_{tr}(k,t,B)$ is *not* a single body term. It incorporates multibody effects that explicitly depend on the configuration of amino acid groups within a given protein conformation and it also depends on the solvent averaged enthalpic and entropic interactions between A and W, A and O, W and O, and W and W, where A, W and O correspond to amino acid, water, and osmolyte molecules respectively. The configuration of amino acid groups in a given protein conformation is accounted for in $\Delta G_{tr}(k,t,[0] \rightarrow [C_B])$ by the surface area term $\alpha(l)$. While δg_k^P contains the solvent averaged interactions. For these reasons, the $\Delta G_{tr}(k,t,B)$ term depends sensitively upon the conformation of the protein and the aqueous osmolyte solution conditions.

As we show in Chapters 2, 3, and 4, this is a powerful, accurate approach for modeling and predicting the effects of osmolytes and pH on proteins. It is powerful because after the simulations are completed, any thermodynamic property under any other set of osmolyte or pH conditions can be predicted in a matter of minutes. It is accurate because we show that we can achieve, in several instances, quantitative agreement between predicted and experimental data. Thus, this approach offers a molecular level perspective on solution condition effects on proteins. The accuracy

of Eq. 1.9 is limited by the accuracy of the protein model Hamiltonian (i.e. the force field), the accuracy of the ΔG_{tr} models, and the extent of simulation sampling in solution condition A.

1.4 Overview of Chapters

This thesis presents theoretical studies of protein folding and amyloid formation in various environments ranging from various osmolyte and pH solution conditions to protein folding under tension and inside carbon nanotubes.

In Chapter 2, we introduce the Molecular Transfer Model for modeling osmolyte and denaturant effects on proteins. We validate the MTM against experimental data from two proteins: protein L and a cold shock protein. We find excellent agreement between the MTM predicted FRET efficiency $\langle E \rangle$ and the experimentally measured $\langle E \rangle$. We examine how denatured state properties change with osmolyte and denaturant concentration. We find that R_g of the denatured state can vary by several angstroms depending on the type of cosolute and its concentration. Residual structure in the denatured state of protein L is found even at high denaturant concentrations. Fitting to an analytic polymer model, we show that the denatured state of these two proteins behave as excluded volume polymers at high denaturant concentrations.

In Chapter 3, we introduce the MTM for pH effects on proteins. We validate this model of pH denaturation against experimental data for protein G and Chymotrypsin Inhibitor 2. Excellent agreement is found between the MTM predicted

ΔG_{ND} vs. pH profile and the experimentally measured profile. We then study the effect of pH and osmolytes on proteins under constant force, as in Atomic Force Microscopy and Laser Optical Tweezer experiments that are being carried out. We find that urea facilitates mechanical unfolding, while TMAO counteracts it. We also find that pH effects are protein dependent. $f_{1/2}$, the midpoint unfolding force, is found to be linear with temperature and urea concentration and non-linear with pH. The transition state location exhibits classic Hammond-Leffler behavior. Surprisingly, the m-value is found to change dramatically with the applied force f. The central conclusion of this chapter is that the effect of varying solution conditions on a protein under tension can be understood and qualitatively predicted based on the knowledge of that protein's behavior in the absence of force.

Single molecule experiments using FRET are being used to infer properties of the denatured state of proteins, including R_g, l_p, R, and $P(R)$. While it is often assumed that the procedure for inferring these properties from FRET data yield quantitatively accurate results, there is *no* independent experimental means to determine their accuracy. In Chapter 4, we test the accuracy of FRET inferred protein properties using both the MTM and a polymer model for which all properties are independently known. By applying the same analysis procedure that experimentalists use to FRET data generated from the MTM, the accuracy of the resulting FRET inferred properties can be tested. We find that while R is accurately inferred (less than 10% relative error under all solution conditions), R_g and l_p are not (with errors of up to 25%). The inferred $P(R)$ distribution, while qualitatively correct, is quantitatively inaccurate. These findings are important because they suggest that

single molecule FRET data on unfolded proteins, as currently analyzed, give only a qualitative measure of denatured state properties.

From kinetic experiments, it is known that the process of monomer addition to a fully formed amyloid fibril is complex. A thermodynamic characterization of this process is not experimentally possible due to issues of reversibility and signal to noise ratios. In Chapter 5, we use molecular simulations of a peptide from the Aβ protein to understand the thermodynamic and structural basis for the 'dock-lock' mechanism. We find that the reversible association of the monomer to a fibril surface has multiple basins of attraction and undergoes multiple structural transitions as it adds to the surface. The free energy barrier separating the docked and locked phases arises from the loss of internal monomer interactions. The impact of urea, TMAO and molecular crowders on the critical concentration is examined.

The behavior of proteins under confinement is relevant to a number of in vivo situations, including protein transport through protein membrane channels and the synthesis of nascent peptides in the ribosome exit tunnel. In Chapter 6, we explore the effect of a range of parameters on helix formation of peptides confined to carbon nanotubes including protein sequence, nanotube diameter, hydrophobic strength and the chemical heterogeneity of the nanotube. We find a rich diversity of behavior in helix formation as a function of these parameters. Narrow, weakly hydrophobic nanotubes stabilize the helix for all sequences. Increasing the hydrophobic strength of the nanotube causes amphiphilic sequences to form helices and a polyalanine to lose helical content. Decreasing the size of the hydrophobic patch lining the nanotube enhances helix formation of the polyalanine when the hydrophobic strength

is strong. The relevance of these findings to in vivo situations is discussed.

Chapter 2

Effects of denaturants and osmolytes on proteins are accurately predicted using the Molecular Transfer Model

2.1 Introduction

To function proteins fold [3], while misfolding is linked to a number of conformational diseases [19, 41], thus making it important to determine the factors that control their stabilities [3] and the assembly mechanisms [42, 43, 44]. A molecular understanding of protein folding requires quantitative estimates of the energetic changes [45, 29] in the folding reaction and characterization of the populated structures along the folding pathway. A large number of studies have dissected the interactions that contribute to the stability of proteins [3, 45, 29, 46, 47, 48, 49, 50, 48, 51, 32]. In contrast, only relatively recently has there been a concerted effort to determine the structures of the denatured state ensemble (DSE) [52] whose experimental resolution is difficult due to fluctuations in the unfolded structures. In particular, it is difficult to determine the properties of the DSE under conditions in which the native state is stable because the population of the unfolded structures is low [53]. Single molecule FRET experiments have begun to investigate the variations in the global properties of the DSE under native conditions [54, 55, 56]. Despite these intense efforts, structural characterization of the DSE, and its link to

global thermodynamic properties and the folding process is lacking.

Denaturants, such as urea and guanadinium chloride (GdmCl), destabilize proteins. In contrast, osmolytes that protect cells against environmental stresses such as high temperature, dessication, and pressure can stabilize proteins [14]. Thus, a complete understanding of the stability of proteins and a description of the structures in the diverse DSE requires experimental and theoretical studies that provide a quantitative description of the effects of both osmolytes and denaturants.

From a theoretical perspective, significant advances in our understanding of how proteins fold have come from molecular simulations using coarse-grained (CG) off-lattice models [57, 10, 38, 58, 59, 60]. However, the CG models only probe the folding of proteins by changing temperature, making it difficult to compare directly with many experiments that use denaturants. In principle, all-atom simulations of proteins in aqueous denaturant solutions can be used to calculate the conformational properties of proteins. However, the difficulty in adequately sampling the protein conformational space makes most of these simulations inherently non-ergodic [61]. Here, we overcome these problems by combining Tanford's transfer model (TM) [23, 26] together with simulations using an off-lattice side chain representation of polypeptide chains [59] to predict the dependence of the size of the protein, fraction of molecules in the native state, and FRET efficiencies as a function of the concentration ([C]) of denaturants and osmolytes. We introduce a novel method that combines molecular simulations of a protein of interest at [C]=0, and the experimental transfer free energies [27, 62] (see Methods) to predict the thermodynamic averages at [C] $\neq$ 0. In the process, we have greatly expanded the power and scope

of CG off-lattice models [10, 58, 60] in predicting the outcomes of experiments. Applications of the resulting Molecular Transfer Model (MTM) to protein L and Cold shock protein (CspTm) show that calculated changes in the fraction of folded conformations, and the average FRET efficiency as a function of [GdmCl] are in excellent agreement with experiments [63, 64, 56]. The stability in the presence of glycine betaine, proline, sucrose, sarcosine, sorbitol and TMAO for the two proteins increases linearly as [C] increases. Our results also give plausible explanations for the inability of scattering methods to directly infer protein collapse at low [C]. The heat capacity changes in proteins in denaturants and osmolytes are interpreted in terms of changes in the folding landscape.

2.2 Results

2.2.1 MTM accurately captures denaturant-induced unfolding of protein L and CspTm

To establish the efficacy of the MTM, we calculate a number of quantities that can be directly compared with data from ensemble and single molecule experiments [54, 55, 63, 64, 56]. As with most molecular force fields, the absolute interaction energies in the C_α-SCM at [C]=0 are not accurate. We set the temperature ($T = T_S$) so that the calculated free energy of stability of the native state $\Delta G_{NU}(T_S)$, with respect to the unfolded structures, and the measured $\Delta G_{NU}(T_E)$ at $T = T_E$ coincide. In the absence of denaturants, T_S = 328 K and T_E = 295 K and $\Delta G_{NU}(T_S) = \Delta G_{NU}(T_E) = -4.6\ kcal/mol$ [65] for protein L (Fig. 2.1A). For CspTm (Fig. 2.1A)

T_S = 326 K and T_E = 298 K and $\Delta G_{NU}(T_S) = \Delta G_{NU}(T_E) = -6.3\ kcal/mol$ [66]. By adjusting T_S appropriately, we find that the dependence of the calculated fraction of molecules in the native basin of attraction (NBA), f_{NBA}, as a function of [C] for GdmCl is in excellent agreement with experiments (Fig. 2.1B). The values for C_m, the midpoint concentration at which $f_{NBA} = 0.5$, for both proteins also reproduce the measured values accurately (Table 2.1).

2.2.2 Measured and predicted FRET efficiencies are in good agreement:

In an attempt to characterize the nature of unfolded states of proteins under folding conditions (low denaturant concentrations) several groups have used single-molecule FRET spectroscopy [54, 63, 56, 64]. By attaching fluorescent dyes at two points (typically, but not always [56], located at the termini of the protein) the average FRET efficiency $\langle E \rangle$ as a function of [GdmCl] has been measured for protein L and CspTm. We calculated $\langle E \rangle$ as a function of [GdmCl] for protein L (Fig. 2.1C) and CspTm (Fig. 2.1D). The discrepancies between different experiments not withstanding [54, 56, 64, 63], the simulated and the measured $\langle E \rangle$ for protein L and CspTm, for the subpopulation of unfolded states, are in excellent agreement (Figs. 2.1C and 2.1D) with each other. The average FRET efficiency, that weights the subpopulations of folded and unfolded states, reflects the cooperativity observed in f_{NBA} (Fig. 2.1B). The values of $\langle E \rangle$ for the structures in the NBA are roughly constant as [GdmCl] changes (Figs. 2.1C and 2.1D). Even though the simulated

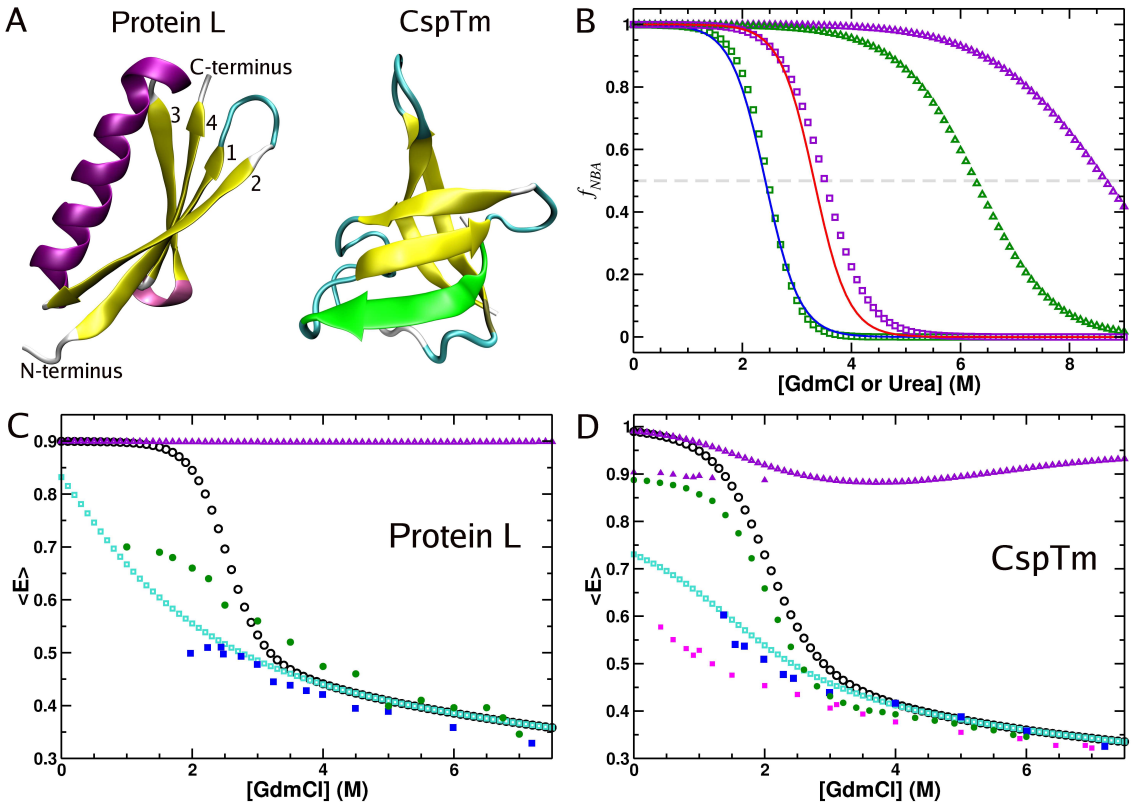


Figure 2.1: Native structures and comparison of calculated and experimental results. (A) The numbers in protein L label the strands starting from the N-terminus. The N-terminal β-strand in CspTm is colored green. (B) The fraction of molecules in the NBA (f_{NBA}) as a function of GdmCl (green squares) and urea (green triangles) for protein L. Results for CspTm in GdmCl and urea are shown in violet squares and violet triangles, respectively. Blue line is the result of $f_{NBA}([C])$ for protein L [65]. Results in red lined is for CspTm [66]. Dashed line shows $f_{NBA} = 0.5$. (C) The dependence of $\langle E \rangle$ for protein L (open circles) versus GdmCl concentration. Open triangles show $\langle E \rangle$ for the native state and the squares are for the DSE. The experimental values for the average $\langle E \rangle$ and $\langle E \rangle$ of the DSE are shown as green circles [63] and blue squares [64], respectively. (D) Results for CspTm using the same notation as in (C). The filled blue squares are experimental results from [64]. Filled green circles, violet triangles, and magenta squares correspond to experimental measurements of $\langle E \rangle$, the NBA $\langle E \rangle$, and the DSE $\langle E \rangle$, respectively [56]. To account for the destabilization of CspTm due to the attachment of dyes we set $T_S = 341$ K, which gives C_m in agreement with experiment [54]. In (C) and (D) we use $R_o = 55$ Å (see Eq 3 in Appendix A). Changes in R_o with [C] cause small corrections to $\langle E \rangle$.

value of $\langle E \rangle$(=0.9) for protein L at zero [GdmCl] agrees with the calculated FRET efficiency using Protein Data Bank (PDB) coordinates (PDB ID 1HZ6), it is larger than the measured value, which is in the range of 0.7-0.8. The discrepancy could also arise because the present simulations do not explicitly include the dyes with flexible linkers which can have a large effect [64]. Despite the difference, at [C]=0, our simulations accurately reproduce the experimental measurements.

2.2.3 Changes in R_g depend on the nature of cosolvents:

The R_g distribution ($P(R_g)$) for protein L in urea, at the folding (or melting) temperature $T_F = 356K$, shows the expected behavior (Fig. 2.2). At 0 M, there is a sharp peak in $P(R_g)$ at $\overline{R_g^N}$ (the value in the native state) $\sim$ 12 Å, whereas a relatively broad ensemble of conformations, with larger R_g values ($>$ 12 Å), is populated at 6 M urea (Fig. 2.2A). The distribution $P(R_g)$ at 6 M urea compares favorably with recent all atom simulations (see Fig. 10 in [64]). In 6 M TMAO the peak height at $R_g \sim$ 12 Å increases which reflects its stabilizing influence. The average $\overline{R_g}$ for protein L expands continuously as urea concentration increases from 0 to 6 M (Fig. 2.2B). Decomposition of the ensemble of structures into the DSE subpopulation shows that $\overline{R_g^{DSE}}$ expands from 21.6 Å at 0 M urea to 24 Å at 6 M whereas $\overline{R_g^N}$ is independent of urea concentration (Fig. 2.2B). At physiological concentrations ($\sim$ 1 M) the change in $\overline{R_g}$ induced by TMAO is small (Fig. 2.2B). Just as for urea, the value of $\overline{R_g^N}$ remains constant at all TMAO concentrations (Fig. 2.2B).

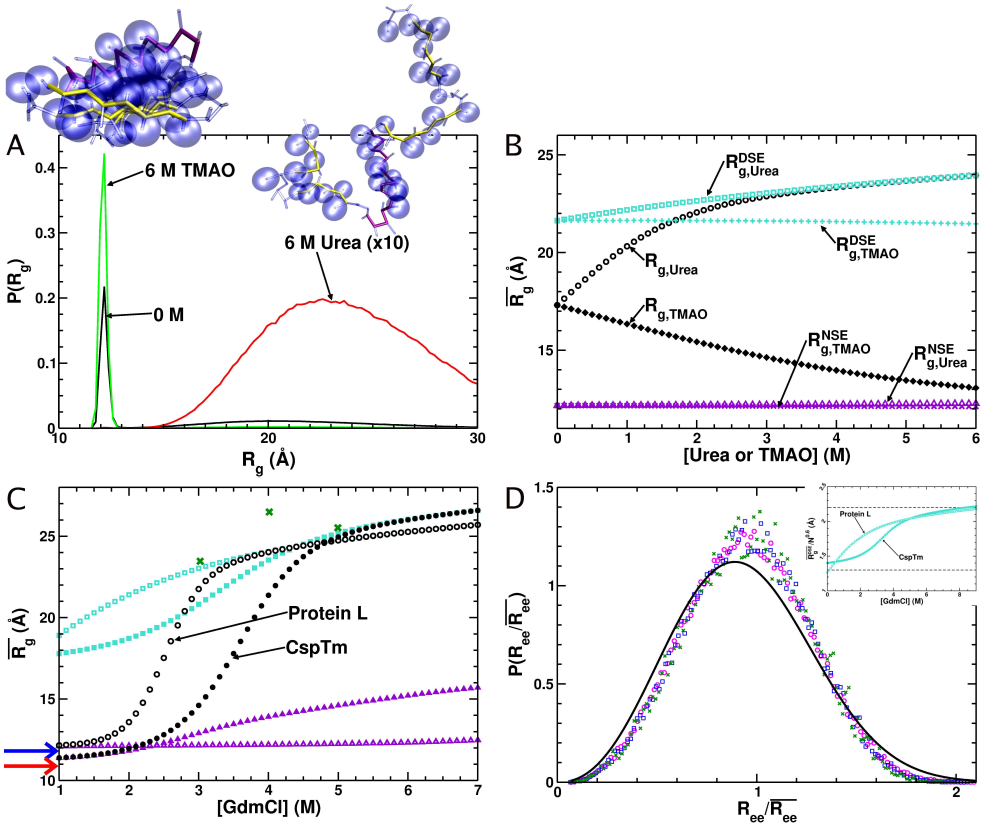


Figure 2.2: $P(R_g)$ and $\overline{R_g}$: (A) The distribution $P(R_g)$ for protein L at 0 M and 6 M TMAO and urea at $T_F = 356K$, the melting temperature at 0 M. For $P(R_g)$ at 0 M, the area under native and denatured ensembles are equal. The $P(R_g)$ at 6 M urea is multiplied by ten. The structure on the left corresponds to the C_α-SCM representation of the native state and the one to the right is an example of a conformation in the DSE. For clarity, only hydrophobic side chains are displayed as blue spheres (B) The average $\overline{R_g}$ of protein L as a function of urea (open black circles) and TMAO (black diamonds) concentration at T_F. The values of $\overline{R_g}$ of the NBA is in violet squares for urea and plus signs for TMAO. The results for $\overline{R_g^{DSE}}$ in urea and TMAO are in turquoise triangles and x-symbols respectively. Using Flory theory $\overline{R_g}$ at $[C] = 0$ is $0.5(\overline{R_g^N} + \overline{R_g^D}) \sim 17.7$ Å, which agrees with the simulations. (C) The $\overline{R_g}$ for protein L (open black circles) and CspTm (filled black circles) as a function of GdmCl concentration at a temperature of 328 K (protein L) and 326 K (CspTm). $\overline{R_g}$ for the NBA is in triangles and DSE in squares for protein L (open symbols) and CspTm (filled symbols). Blue and red arrows show $\overline{R_g}$ computed from crystal structures for protein L and CspTm, respectively, The green X's are $\overline{R_g}$ from SAXS experiments [67] for protein L with His tag ($N = 79$). (D) The DSE distribution $P(R_{ee}^{DSE}/\overline{R_{ee}^{DSE}})$ for protein L in 5, 7, and 9 M GdmCl at 328 K. The solid black line is the theoretical universal curve for a self avoiding polymer chain. Inset shows the effective Kuhn length $a_D([C], T) = \overline{R_g^{DSE}}/N^{0.6}$ versus GdmCl concentration. The dashed lines show the range of experimentally measured Kuhn lengths [68].

There are substantial changes in the size of protein L and CspTm in aqueous GdmCl solution (Fig. 2.2C). (1) For both proteins, the precipitous change in $\overline{R_g}$ occurs at [C] $\sim C_m$ which suggests that global unfolding is accompanied by expansion of the proteins (compare Figs. 2.1C, 2.1D, and 2.2C). Unfolding in GdmCl is considerably more cooperative than in urea (data not shown). (2) In contrast to protein L, whose $\overline{R_g^N}$ is nearly independent of the concentration of GdmCl (Fig. 2.2C), $\overline{R_g^N}$ for CspTm increases marginally when [C] exceeds $\sim$ 2.5 M. Moderate denaturant-induced increase in $\overline{R_g^N}$ at high concentrations of GdmCL indicates that packing is somewhat compromised in CspTm, arising from enhanced fluctuations in the N-terminal β-strand (Fig. 2.1A and see below). (3) The values of $\overline{R_g^{DSE}}$ for both proteins increase nearly continuously as [C] increases. In CspTm, there may be an inflection point at [C]$\sim$ 2.5 M which coincides with the onset of a modest increase in $\overline{R_g^N}$. (4) At high [C], $\overline{R_g^{DSE}} \sim$ 25.5 Å for protein L and $\overline{R_g^{DSE}} \sim$ 26.5 Å for CspTm (Fig. 2.2C). These values are in near quantitative agreement with the analysis of FRET efficiency using a highly simplified Gaussian model for the end-to-end distribution function [63, 64, 56].

2.2.4 Dissecting denaturant-induced loss of secondary and tertiary structures:

The native structure of protein L has a β-sheet comprised of two β-hairpins formed by strands 1 and 2, and 3 and 4 that interact with a central helix (Fig. 2.1A). The loss in the β-strand contacts in GdmCl and urea mirror the overall unfolding

of the protein (compare Fig. 2.3A and Fig. 2.1B). Chain expansion, and the loss of secondary and tertiary contacts occur at nearly similar concentrations (see Figs. 2.1B, Fig. 2.2C, and Fig. 2.3A). For protein L, at high denaturant concentrations there is near complete loss of β-strand content, while residual helical content persists (Fig. 2.3A).

Comparison of the plots (Fig. 2.3A) of the tertiary contacts involving the secondary structural elements (SSEs) and the total number of contacts in protein L as a function of urea concentration shows that most of the curves overlap. These results (Fig. 2.3B) show that the loss of secondary and tertiary interactions occurs cooperatively. The fluctuations of the various SSEs $\sigma^2_{Q_i} = \langle Q_i^2 \rangle - \langle Q_i \rangle^2$, as a function of urea concentration (Fig. 2.3C) show that the strands 1 and 4, that join the two β-hairpins together to form the full β-sheet, have the most cooperative transition (Fig. 2.3C). These strands, which are far apart in sequence space, form the longest-range contacts in the NBA. Similarly, contacts involving the two hairpins S12 and S34 also unfold cooperatively. Thus, SSEs that form long range contacts in the NBA unfold most cooperatively.

2.2.5 Heat capacity of proteins are greatly altered by osmolytes:

The temperature dependence of the heat capacity (C_V) for protein L and CspTm shows that as urea concentration increases from 0 M to 8 M the curves shift to the left (Figs. 2.4A and 2.4B). In contrast, in the presence of the osmolyte TMAO the curves move to the right (Figs. 2.4A and 2.4B). For proteins that

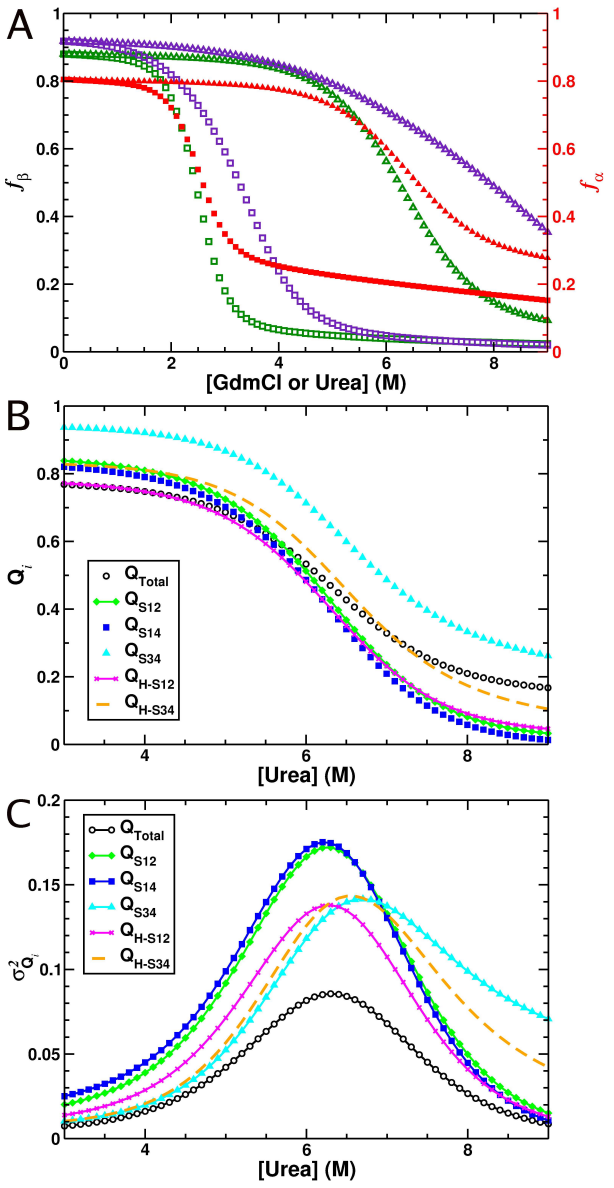


Figure 2.3: Changes in the secondary structural elements of protein L as a function of urea and GdmCl concentration at 328 K. (A) The dependence of β-sheet (green and violet symbols) and helix (red symbols) content of protein L and CspTm on the concentration of GdmCl and urea using the same notation as Fig. 2.1B. (B) The dependence of the fraction of native contacts in urea for protein L. The fraction of native contacts for the entire protein is denoted Q_T, between strands 1 and 2 as Q_{S12}, between strands 1 and 4 as Q_{S14}, between strands 3 and 4 as Q_{S34}, between strands 1,2 and the helix as Q_{H-S12}, and between strands 3,4 and the helix as Q_{H-S34}. (C) Variance in the fraction of native contacts versus urea concentration.

fold in an apparent two-state manner the peak in C_V can be identified with the folding temperature, T_F. The decrease in the folding temperature $\Delta T_F([C]) \equiv T_F([C]) - T_F(0)$ as the concentration of urea increases from 0 to 8 M can be as large as $35°C$. As the concentration of TMAO increases from 0 to 8 M, $\Delta T_F([C])$ increases by as much as $12°C$ for protein L and $\sim 20°C$ for CspTm. These results (Figs. 2.4A and 2.4B) indicate that there are large variations in thermal stability of CspTm and protein L as the concentrations of urea and TMAO are increased.

In contrast to the behavior of C_V for protein L (Fig. 2.4A), the peak heights and the widths change significantly for CspTm in urea and TMAO (Fig. 2.4B). For CspTm the maximum in C_V goes from 6.5 $kcal\ °C^{-1}\ M^{-1}$ at 0 M to ∼9.0 $kcal\ °C^{-1}\ M^{-1}$ in 8 M TMAO and ∼5.0 $kcal\ °C^{-1}\ M^{-1}$ in 8 M urea. The maximum in C_V for protein L on the other hand changes by only ∼0.2 $kcal\ °C^{-1}\ M^{-1}$ under these same solution conditions (Fig. 2.4A).

2.2.6 Protein stability changes linearly as denaturant and osmolyte concentrations increase:

Denaturants: Although the changes in native state stability $\Delta G_{NU}([C])$ as a function of [C] for protein L (Fig. 2.4C) and CspTm (Fig. 2.4D) at $T \sim 328$ K, shows evidence for non-linearity in some of the curves, the free energy change can be approximately fit using $\Delta G_{NU}([C]) = \Delta G_{NU}(0) - m[C]$ [69, 22]. The m-values show that GdmCl is significantly more efficient in denaturing protein L and CspTm than urea (Table 2.1). As a result, the denaturation midpoint C_m for protein L,

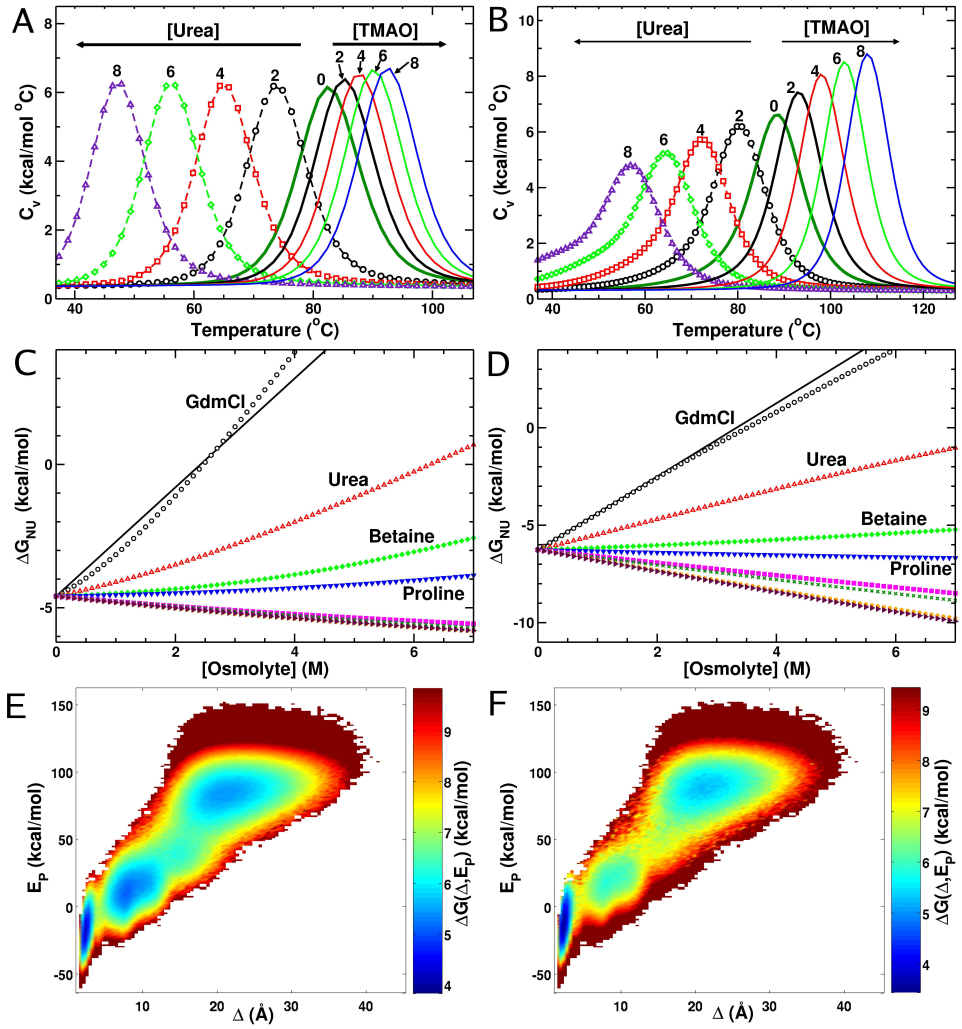


Figure 2.4: Thermodynamic properties of protein L and CspTm in denaturant and osmolyte solutions. (A) Heat capacity of protein-L versus temperature as a function of urea and TMAO concentration. Numbers above the maxima of each trace give the osmolyte concentration in Molar units. Curves to the left of the 0 M plot correspond to increasing urea concentrations, while those to the right represent increasing TMAO concentrations. (B) Results for CspTm using the same notation as in (A). (C) The stability of the native state ensemble of protein L as a function of concentration of various osmolytes at 328 K. The data corresponding to GdmCl, urea, betaine and proline are labeled. The variation of ΔG_{NU} in aqueous sorbitol, sucrose, sarcosine and TMAO solutions are similar, and are unlabeled. The solid black line is the experimental result for GdmCl denaturation [65]. (D) Same as (C) except the results are for CspTm and experimental results are taken from [66]. (E) The free energy surface of CspTm as a function of the root mean square deviation relative to the crystal structure (Δ) and the potential energy (E_P) at 0 M and T_F (361 K). (F) The same as (E) except at 8 M TMAO and 381 K.

obtained by using $\Delta G_{NU}([C_m]) = 0$, is 2.4 M in aqueous GdmCl and is 6.3 M in aqueous urea.

The calculated (2.4 $kcal\ mol^{-1}\ M^{-1}$ for protein L and 1.7 $kcal\ mol^{-1}\ M^{-1}$ for CspTm) and measured (1.9 $kcal\ mol^{-1}\ M^{-1}$) GdmCl m-values for protein L and CspTm (Table 2.1) are in excellent agreement. The predicted m-value for betaine is relatively small ($m \simeq 0.2\ kcal\ mol^{-1}\ M^{-1}$) which implies that betaine only marginally affects the stability of CspTm and protein L (Table 2.1 and Figs. 2.4C and 2.4D). Therefore, the efficiency of denaturation follows the trend $GdmCl > Urea > betaine$. The predictions in aqueous urea and betaine await future experiments.

Osmolytes: The stability changes for osmolytes (proline, sorbitol, sucrose, TMAO, and sarcosine) for protein L (Fig. 2.4C) and CspTm (Fig. 2.4D) at $T \simeq 328$ K vary linearly over a broad range of concentrations. The extracted m-values for all these osmolytes vary only moderately for protein L (m = -(0.1 to 0.2) $kcal\ mol^{-1}\ M^{-1}$) and for CspTm (m =-(0.5 to 0.3) $kcal\ mol^{-1}\ M^{-1}$, see Table 2.1). The nearly constant m-values for the osmolytes is consistent with experiments that have found that m-values for TMAO and sarcosine are roughly the same for barstar [70]. As a result of the small m-values the osmolytes increase the stability of the small proteins only modestly ($\sim 1\ kcal/mol$).

Table 2.1: Calculated thermodynamic parameters for protein L and CspTm

	protein L		CspTm	
Osmolyte	m-value [a]	$\Delta G_{NU}[0]$ [b]	m-value	$\Delta G_{NU}[0]$
GdmCl	2.4 [c]	-6.0[d]	1.7 [e]	-5.8 [f]
Urea	0.9	-5.7	0.7	-6.1
Betaine	0.2	-4.8	0.2	-6.3
Proline	-0.1	-4.7	0.1	-6.3
Sorbitol	-0.1	-4.7	-0.3	-6.3
Sucrose	-0.2	-4.7	-0.4	-6.3
TMAO	-0.2	-4.7	-0.5	-6.3
Sarcosine	-0.2	-4.7	-0.5	-6.3

[a] units in $kcal\ M^{-1}\ mol^{-1}$.

[b] Native state stability, in $kcal\ mol^{-1}$ units, at 0 M using the linear extrapolation method [21].

[c] Experimental value is 1.9 $kcal\ mol^{-1}\ M^{-1}$ [65].

[d] Two state fit to experimental data gives $\Delta G_{NU}[0] = -4.6\ kcal\ mol^{-1}$ to $6.0\ kcal\ mol^{-1}$ [65, 71].

[e] Experimental value is $1.9 \pm 0.08\ kcal\ mol^{-1}\ M^{-1}$ [66].

[f] The experimental value is $\Delta G_{NU}[0] = -6.3 \pm 0.3\ kcal\ mol^{-1}$ [66].

2.3 Discussion

2.3.1 Flory theory, simulations, and experiments for R_g and the end-to-end distance distribution $P(R_{ee})$:

The $\overline{R_g}$ values of proteins scales as $\overline{R_g^D} = a_D([C], T)N^{\nu}$ (where ν, the Flory exponent, is $\nu \simeq 0.59$) [68]. The Kuhn length $a_D([C], T)$, reflects the quality of the solvent, which depends on [C], T, and the protein sequence, is found to be a constant $a_D \sim 2$ Å [68] (however, see Fig. A.1 in Appendix A). Analysis of the folded structures of proteins shows that $\overline{R_g^N} = a_N N^{1/3}$ with $a_N \sim 3$ Å [72]. For protein L (N=64) and CspTm (N=66) we expect that $\overline{R_g^N} \sim 12$ Å and 12.1 Å, respectively. Direct calculation of $\overline{R_g^N}$ using coordinates from the structures of protein L and CspTm give 12 Å and 11 Å respectively.

If $a_D([C],T) \sim a_D = 2$ Å is a constant then Flory theory predicts that $\overline{R_g^D} \sim$ 23.3 Å for protein L ($N = 64$), which is in excellent agreement with the simulation results (Fig. 2.2B). Small angle X-ray scattering (SAXS) measurements of protein L with a histidine tag, resulting in $N = 79$ [68], show that $\overline{R_g} = 26 \pm 1.5$ Å and 25 ± 1.5 Å at 4 M and 5 M GdmCl, respectively. From Flory theory we expect $\overline{R_g^D} \sim 27.8$ Å. The agreement between theory, simulations and SAXS data show that, as far as $\overline{R_g}$ is concerned, protein L behaves as a random coil at high GdmCl concentrations.

In apparent contrast to SAXS measurements [73], our simulations and analysis of FRET data show that protein L [54, 55, 63, 64] and CspTm [56] collapse at low [C]. The differences could arise for the following reasons. (1) At [C] < [C_m] almost all of the scattering intensity arises from the folded state, just as at [C] > [C_m] the scattering is dominated by the conformations in the DSE. Thus, it is unlikely that SAXS measurements can resolve the small contributions of $\overline{R_g^{DSE}}$ at low values of [C].

(2) At a fixed T, the "non-universal" Kuhn length $a_D([C],T)$ should be [C]-dependent. The Kuhn length $a_D([C],T) \rightarrow a_D$ only when $x = [C]/C_m >> 1$ so that inter-residue attractive interactions are negligible, and hence the conformational characteristics of proteins are determined solely by excluded volume interactions. To ascertain the variations of the Kuhn length as [C] changes we computed $a_D([C],T) = \overline{R_g^{DSE}}/N^{\nu}$, which increase from about 1.3 Å to about 2.2 Å (see inset in Fig. 2.2D). Recent, SAXS experiments (see Fig. 3B in [74]) also show that $\overline{R_g}$ for the 159-residue *E. Coli* Dihydrofolate Reductase continues to increase as urea concentration increases in the range 4.5M to 8M which can be rationalized in terms of a [C]-

dependent Kuhn length.

(3) There are a large changes in the distribution $P(R_g^{DSE})$ as [C] changes (Fig. A.2 in Appendix A). If proteins are random coils at high [C] then $P(R_{ee}^{DSE})$, for sufficiently large $y = R_{ee}^{DSE}/\overline{R_{ee}^{DSE}}$ should be given by the universal curve $P(y) = c_1 y^{2+\theta} exp(-c_2 y^{1/(1-\nu)})$ [75], where $\theta = (\gamma - 1)/2 \sim 1/3$, $c_1 = 3.7$ and $c_2 = 1.2$ (see Appendix A). The simulation results show that, to an excellent approximation, this is indeed the case for $P(R_{ee}^{DSE}/\overline{R_{ee}^{DSE}})$ (Fig. 2.2D) for $y > 1.5$ and [C]$>$ 5 M GdmCl (see also Fig. A.2 in Appendix A). Thus only at high [C], when the residual intrapeptide attraction is negligible, the random-coil nature of proteins emerges, while at low [C] there are substantial deviations from the self-avoiding $P(y)$ (Fig. A.3 in Appendix A).

The incorrect assumption that $a_D([C], T)$ is a constant (or equivalently that $\overline{R_g^{DSE}}$ is [C] independent) when analyzing experimental results (see Fig. A.4 in Appendix A for further discussion), and the limited data at [C] beyond the transition region [73] make it difficult to infer protein collapse using SAXS measurements. In addition, it has been suggested [64] that inter-protein interactions could also have affected the SAXs measurements. At the very least, the protein L measurements have to be extended beyond 5M GdmCl to decipher the changes in $\overline{R_g^{DSE}}$.

2.3.2 Structural interpretation of the heat capacity curves:

The origin of the contrasting behaviors in C_V between protein L and CspTm in urea and TMAO (Figs. 2.4A and 2.4B) is reflected in the free energy surfaces

(FESs) at T_F. The two-dimensional FES, expressed in terms of the potential energy (E_P) and the root-mean-square deviation (Δ) from the native state, of protein L has two distinct basins at all osmolyte concentrations (data not shown). On the other hand, CspTm displays three distinct basins at 0 M (Fig. 2.4E). The basin centered at $\Delta \sim$3 Å corresponds to conformations that closely resemble the crystal structure. The basin, at $\Delta \sim$9 Å, corresponds to conformations in which the N-terminal strand (Fig. 2.1A) is disordered while the rest of the barrel is intact. The basin centered at $\Delta \sim$22 Å consists of mostly random coil conformations that have little β-sheet content. At 8 M TMAO the basin of attraction centered at $\Delta \sim$ 9 Å at 0 M is significantly destabilized (Fig. 2.4F) resulting in a sharper transition in C_V (Fig. 2.4B). In contrast, urea expands the area of the denatured basin in the FES (data not shown), which in turn leads to a reduction in the height of C_V and an increase in the width of the transition.

2.4 Conclusions

By using converged simulations in the absence of denaturants and osmolytes, together with the measured transfer free energies, the MTM accurately predicts the dependence of any thermodynamic property at arbitrary denaturant or osmolyte concentration. The striking agreement between the computed and measured GdmCl-induced changes in the FRET efficiencies for protein L and CspTm attests to the success of the MTM. The structures of the denatured states, as measured by the residual secondary and tertiary structure content, can be greatly perturbed

by adjusting the osmolyte concentration. As a consequence, the folding trajectories may change significantly depending on the initial conditions. Predictions for urea-induced changes in the DSE and the profound differences between the heat capacity changes in urea and TMAO between protein L and CspTm are amenable to experimental tests. More generally, the MTM provides a structural interpretation of the cooperative thermal melting of proteins in osmolytes. In addition, we have made a number of testable predictions for the changes in equilibrium properties of these small single domain proteins in osmolytes. The present theory sets the stage for using the MTM not only in the context of the $C_\alpha - SCM$, but also in conjunction with all-atom Go models for which exhaustive sampling can be carried out.

2.5 Methods

2.5.1 C_α-Side chain model (C_α-SCM) for proteins:

We use the coarse-grained C_α-side chain model (for details see Appendix A) in which each residue in the polypeptide chain is represented using two interaction sites, one that is centered on the α-carbon atom and another that is located at the center-of-mass of the side chain [59].

2.5.2 The Molecular Transfer Model:

The energy of a protein conformation at non-zero [C] is taken to be a sum of the potential energy E_P of the protein (see Appendix A) and the transfer free energy $\Delta G_{tr}([C])$ based on TM. According to the TM, the free energy of transferring

a protein to osmolyte solution is equal to the sum of the transfer free energies (TFEs) of the individual groups (side chain and backbone moieties) that are solvent exposed. The free energy cost of transferring the i^{th} protein conformation from water to aqueous osmolyte solution at concentration [C] is written as

$$\Delta G_{tr}(i,[C]) = \sum_{k=1}^{N_{SC}} \delta g_{tr,k}^{SC}([C]) n_k \alpha_{i,k}^{SC} / \alpha_{k,Gly-k-Gly}^{SC} + \sum_{k=1}^{N_{BB}} \delta g_{tr}^{BB}([C]) n_k \alpha_{i,k}^{BB} / \alpha_{k,Gly-k-Gly}^{BB} \tag{2.1}$$

where the sums are over the different amino acid types in the protein, n_k is the number of amino acid residues of type k, $\delta g_{tr,k}^{SC}$ and $\delta g_{tr,k}^{BB}$ are the transfer free energies of the side chain and backbone group of amino acid type k, respectively [26, 62]. For denaturants $\delta g_{tr} < 0$, i.e. thermodynamically favorable, for the peptide backbone and many types of amino acid side chains [23, 46, 28]. The transfer of some of these substituents to an osmolyte solution results in $\delta g_{tr} > 0$ [28]. The solvent accessible surface areas of the side chain and backbone group of amino acid type k are $\alpha_{i,k}^{SC}$ and $\alpha_{i,k}^{BB}$, respectively, and $\alpha_{k,Gly-k-Gly}^{SC}$ is the solvent accessible surface area of the side chain and backbone in the tripeptide $Gly-k-Gly$.

To combine experimentally measured $\delta g_{tr,k}$'s with simulations at [C]=0 we introduce the primary equation of MTM, that has the form of the Weighted Histogram Analysis Method [40, 37, 76], namely,

$$\langle A[C_i,T] \rangle \;\; = \;\; Z([C_i],T)^{-1} \sum_{k=1}^{R} \sum_{t=1}^{n_k} \frac{A_{k,t} e^{-\beta(E_P(k,t,[0]) + \Delta G_{tr}(k,t,[C_i]))}}{\sum_{m=1}^{R} n_m e^{f_m - \beta_m E_m(k,t,[0])}}, \tag{2.2}$$

where $Z([C_i],T)$ is the partition function. Thus, if $Z([0],T)$ is computed and the transfer free energy of each protein conformation is known then any thermodynamic property, at arbitrary $[C_i]$, can be predicted. In Eq. 3.2, R is the number of in-

dependent simulated trajectories, n_k is the number of conformations from the the k^{th} simulation, $A_{k,t}$ is the value of property A for the t^{th} conformation, $\beta = 1/k_B T$, where k_B is Boltzmann's constant and T is the temperature. The potential energy of the t^{th} conformation from the k^{th} simulation in the presence of osmolyte i at concentration C_i is $E(k,t,[C_i]) = E_P(k,t,[0]) + \Delta G_{tr}(k,t,[C_i])$, where $E_P(k,t,[0])$ is the corresponding value at 0 M. The free energy cost of transferring the t^{th} conformation in the k^{th} simulation from 0 M to $[C_i]$ M is $\Delta G_{tr}(k,t,[C_i])$. In the denominator of Eq. 3.2, n_m and f_m are, respectively, the number of conformations and the free energy in the m^{th} simulation.

The values $\alpha_{k,Gly-k-Gly}$ for the side chain and backbone groups (Eq. 3.4) are listed in Table II in Appendix A. For the osmolytes considered here (urea, glycine betaine, proline, sucrose, sarcosine, sorbitol, and TMAO) we use the TFEs given in [28], and for aqueous GdmCl we use the transfer free energies listed in [46]. We extrapolate to osmolyte concentrations that were not experimentally measured by fitting the TFE data to a straight line [77] (see Appendix A for details).

Chapter 3

pH and osmolyte effects on single molecule mechanical unfolding of proteins

3.1 Introduction:

Single molecule constant force (smCF) experiments using Atomic Force Microscopy [78, 79] are capable of characterizing the thermodynamics and kinetics of protein folding under external tension [80]. In many smCF experiments, a constant force (f) is applied to the N and C termini of a protein. Using this technique, information on the characteristics of the protein folding energy landscape (e.g. native state stability, roughness, transition state barrier height, and location) can be obtained [80]. smCF also provides insight into in vivo situations in which external tension is applied to proteins, such as the stretching of proteins by the chaperone GroEL [81, 82, 83, 84], unfoldases such as ClpX [85], and translocons [86], which transport proteins across membranes. From in vitro ensemble experiments it is known that pH and osmolytes can have profound effects on the thermodynamics of protein folding [20, 87, 21, 88, 89, 90, 91]. Yet, surprisingly little attention has been given to exploring these solution condition effects in smCF experiments [92, 93].

To our knowledge, only two studies, both of which used the non-equilibrium constant pulling velocity technique, have investigated the effect of pH and an os-

molyte (guanidinium chloride, denoted GdmCl) on the mechanical unfolding of proteins. The first study on protein G in aqueous GdmCl found that the critical force of folding/unfolding was linear with the change in GdmCl concentration and that the position of the transition state was unchanged [93]. The other study, which focused on ubiquitin, found that the critical force changed only when the pH was acidic and well below the protein's isoelectric point [92]. While these studies start to shed light on the interplay of osmolytes and pH on the response of proteins to mechanical forces, a variety of questions remain open, including (1) How does urea, trimethylamine-N-oxide (TMAO) and pH effect the force-temperature phase diagram? (2) Does the midpoint unfolding force ($f_{1/2}$) change linearly with solution conditions? (3) Does the transition state location move with changing solution conditions? (4) Does the relative mechanical stability of 2° and 3° structural elements (SE's) change with solution conditions? (5) Does the m-value ($\equiv \Delta\Delta G_{ND}/\Delta[C]$), associated with urea denaturation, change under tension?

The central hypothesis of this study is that the change in native state stability due to a change in solution conditions (denoted $\Delta\xi$), when $f \neq 0$, is equal to the change in stability when no force is present. That is,

$$\Delta\Delta G_{ND}(f \neq 0, \Delta\xi) \approx \Delta\Delta G_{ND}(f = 0, \Delta\xi), \tag{3.1}$$

where $\Delta\Delta G_{ND}(f, \Delta\xi) = \Delta G_{ND}(f, \xi_2) - \Delta G_{ND}(f, \xi_1)$ and $\Delta\xi$ corresponds to either a change in temperature, pH, or osmolyte concentration. This hypothesis, while formally inexact for anything other than a system with two microstates (as opposed to thermodynamic states), predicts that changes in solution conditions that

destabilize the native state when $f = 0$ will also destabilize the native state when $f \neq 0$ and vice versa. This hypothesis makes testable predictions for questions one through four. Regarding question five, we note that the m-value is conventionally interpreted to be proportional to the difference in solvent accessible surface area (SASA) between the native state ensemble (NSE) and denatured state ensemble (DSE) [32, 27]. Since the force applied to the termini of the protein tends to lead to extended DSE structures [94] with greater SASA, we predict that the m-value will increase with increasing f.

In this study, we address the questions and test the predictions discussed above by utilizing and further developing the Molecular Transfer Model (MTM) [95]. The MTM predicts how thermodynamic properties of a protein change with changing osmolyte and pH conditions. It does this by computing the partition function under the solution condition of interest by combining experimentally measured, or theoretically computed, transfer free energies of individual amino acids with molecular simulations. The MTM is a post-simulation technique and rapidly predicts a protein's properties under a wide range of solution conditions. Previously, we validated the MTM against experimental results of osmolyte effects on proteins [95]. In this study, we extend the capabilities of the MTM to be able to model pH effects. We validate this approach against experimentally measured ΔG_{ND} vs. pH profiles. We study Chymotrypsin inhibitor 2 (CI2) [69, 96] and protein G [97, 98] under constant force using coarse-grained simulations that allow equilibrium simulations to be carried out.

We find that many of the effects of varying solution conditions on protein G

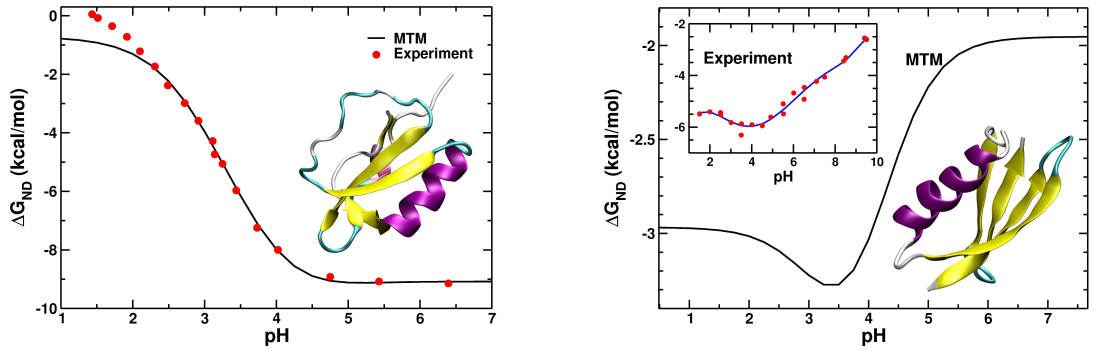
and CI2, when $f \neq 0$, can be qualitatively predicted based on knowledge of the behavior of $\Delta G_{ND}(f = 0)$ vs. ξ. The stabilizing effect of TMAO was found to counteract mechanical unfolding while urea facilitated it. $f_{1/2}$ is found to be linear over a range of temperature and urea values and non-linear as a function of pH. The transition state location (x_{TS}) changes significantly as a function of solution conditions and exhibits classic Hammond-Leffler behavior, with x_{TS} shifting towards unfolded values in TMAO and towards folded values in urea. The m-value is found to increase significantly with f due to changes in the solvent accessible surface area of the DSE.

Our results are relevant to cellular machinery such as chaperone's, translocons, and proteosomes, which may mechanically unfold proteins [81, 82, 83, 84, 85, 86]. The results suggest that low concentrations of urea (or other incompatible osmolytes) make it easier for these machines to do there jobs, whereas the presence of counteracting osmolytes (such as TMAO) make it harder to force unfold proteins in the cell.

3.2 Results and Discussion

3.2.1 MTM accurately models pH denaturation:

To validate the MTM model of pH denaturation, which is used extensively in this study, we plot the experimentally measured and MTM predicted ΔG_{ND} vs. pH profile. The excellent agreement between experimental results for CI2 [99] and the MTM prediction (Fig. 3.1A) indicates that the MTM accurately models pH



(a) (b)

Figure 3.1: The NSE stability, relative to the DSE ($\Delta G_{ND} = -k_B T ln(P_N/P_D)$, where P_N and P_D are the probabilities of being in the NSE and DSE, respectively), of (a) CI2 and (b) protein G versus pH. Structural insets display the native state of CI2 and protein G in a secondary structure representation based on crystal structures with PDB accession codes of 2CI2 and 1GB1 respectively. The inset in (b) is experimental data (red circles) for a triple mutant protein G (T2Q, N8D, N37D). The blue line is a 5th order polynomial fit to the data and is used to guide the eye. Experimental data for wild-type protein G is unavailable. For the CI2 data in (a), T_S and T_E are 302 K and 298 K, respectively. For the protein G data in (b), T_S and T_E are 317 K and 298 K, respectively.

effects on the thermodynamics of folding and unfolding. For wild-type protein G no experimental ΔG_{ND} vs. pH data is available. However, $\Delta G_{ND}(pH)$ data does exist for a triple mutant (T2Q, N8D, N37D) of protein G [98]. While these mutations can be expected to alter the native state stability, we expect that the response of ΔG_{ND} to pH for the wild-type to be qualitatively similar. Fig. 3.1B shows that indeed the overall shape of the MTM predicted ΔG_{ND} vs. pH profile is similar to the experimental data from the mutant protein. These data give us confidence that the MTM is accurate.

3.2.2 Protein properties at $f = 0$ and predictions for $f \neq 0$:

It is relevant to discuss CI2 and protein G's properties when $f = 0$ since the central hypothesis of this study is that the effect of solution conditions on proteins under tension can be predicted by knowledge of solution condition effects on proteins experiencing no external tension.

For both CI2 and protein G, we find that $\Delta G_{ND}(f = 0)$ is approximately linear as a function of temperature, urea, and TMAO concentration (data not shown), while $\Delta G_{ND}(f = 0)$ as a function of pH is non-linear (Fig. 3.1). It is interesting to note that CI2 and protein G show opposite ΔG_{ND} vs. pH behavior. For CI2, decreasing pH monotonically destabilizes its native state (Fig. 3.1A), while decreasing pH non-monotonically stabilizes protein G, having a maximum stability at a pH value of 3.4 (Fig. 3.1B). Based on this, we predict that under external tension, $f_{1/2}$ values will be a linear function of temperature, urea and TMAO concentration, and non-linearly related to changes in pH. The stability of CI2's SEs change with solution conditions, however, the relative ordering of their midpoints of denaturation or renaturation (such as $T_{1/2}$, $pH_{1/2}$, $Urea_{1/2}$, $TMAO_{1/2}$) are largely unchanged by the means of denaturation (see Table 3.1), although some differences exist. Based on this, we predict that the relative ordering of the SE's $f_{1/2}$ values will also be unchanged when $\Delta\xi$. Assuming the folding reaction is analogous to a classical chemical reaction, we predict the proteins will exhibit Hammond-Leffler behavior - with the transition state location shifting towards the native state when the urea concentration or temperature increase, and shifting towards denatured val-

Table 3.1: CI2's denaturation and renaturation midpoints by temperature, pH, urea, and TMAO at $f = 0$.

Structural element	$T_{1/2}$ (K)[a]	$pH_{1/2}$[b]	$Urea_{1/2}$ (M)[c]	$TMAO_{1/2}$ (M)[d]
S_{12}	340.9	Fld[e]	5.0	1.4
S_{23}	341.7	4.6	5.1	1.2
S_{24}	337.2	3.8	3.4	1.9
H_1	343.8	3.8	6.6	1.0
$H_1 - all$	340.3	Fld	4.7	1.5
$H_1 - S_3$	340.1	Fld	4.6	1.5

[a] pH 3.5, 0 M cosolute
[b] 302 K, 0 M cosolute
[c] 325 K, pH 3.5
[d] 350 K, pH 3.5
[e] This structural element, and all of those labeled 'Fld', were folded under these solution conditions.

ues when TMAO is added to solution. Such behavior has been previously observed in the mechanical unfolding of RNA hairpins at various temperatures [80].

3.2.3 Urea facilitates mechanical unfolding, TMAO counteracts it, pH effects are protein dependent:

The native state stability of CI2 and protein G, as a function of f and T, are shown in Fig. 3.2 and Fig. 3.3, respectively. For both CI2 and protein G we find that increasing TMAO concentration counteracts force unfolding (for CI2 compare Figs. 3.2D to 3.2A, and for protein G compare Figs. 3.3D to 3.3A), while increasing urea concentration facilitates force denaturation (for CI2 compare Figs. 3.2C to 3.2A, and for protein G compare Figs. 3.3C to 3.3A). These results are in line with the well-known denaturing effect of urea and stabilizing effect of TMAO, and their effect on native stability at $f = 0$. While naively one might expect that $\Delta G(f \neq 0, [C]) = \Delta G(f = 0, [0]) + m[C]$, we show below that this is an

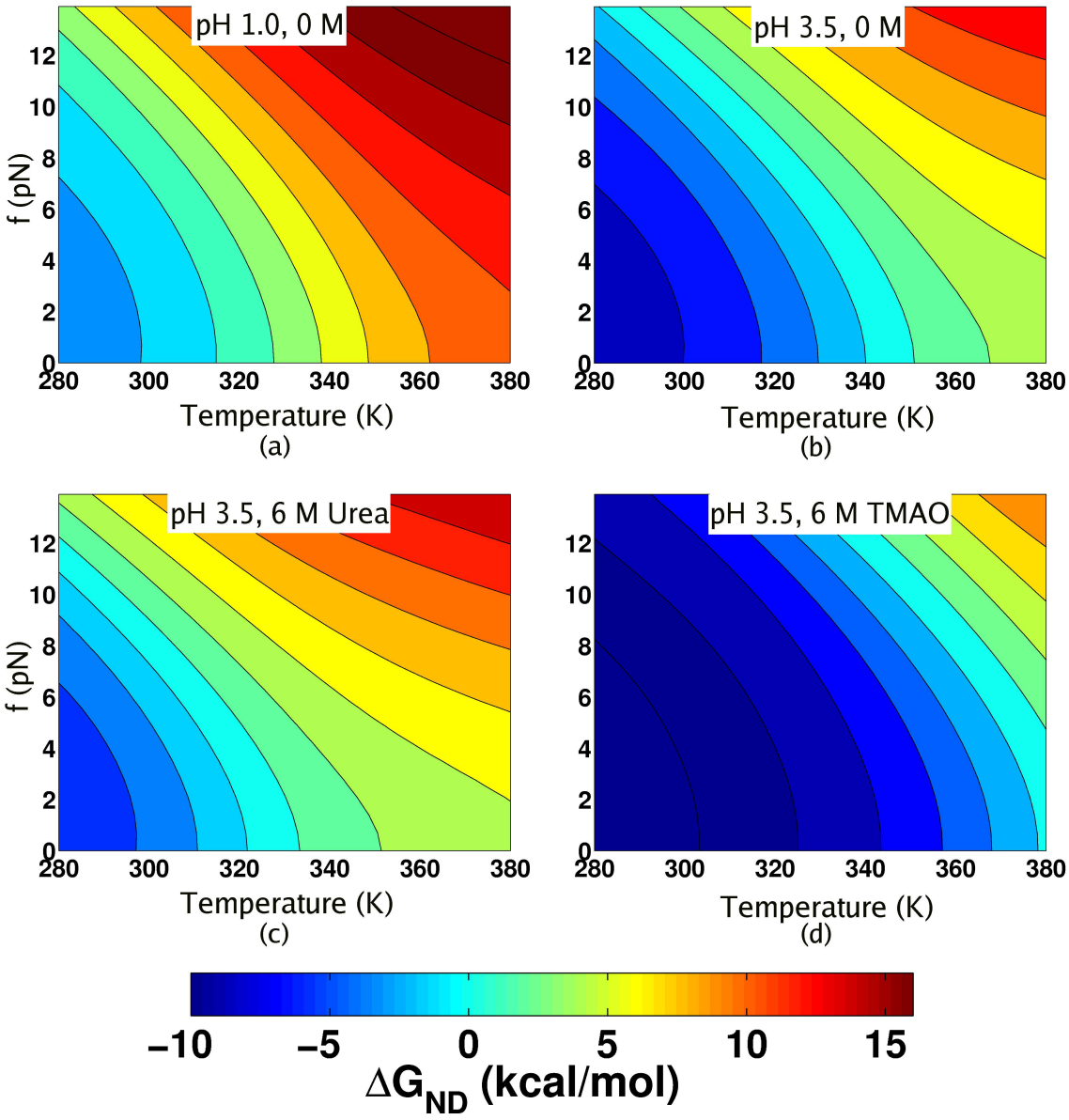


Figure 3.2: The native state stability of CI2 as a function of force and temperature under various solution conditions. Unless otherwise stated, the solution conditions in all panels correspond to 0 M cosolutes and a pH of 3.5. In (a) pH=1.0, (c) 6 M urea is added to solution, and in (d) 6 M TMAO is added to solution.

approximation. The m-value is not constant. It is in fact a function of f, and its dependence cannot be easily anticipated from its value measured at $f = 0$.

pH effects on the $f - T$ phase diagram differ dramatically between CI2 and protein G. For CI2, increasing pH counteracts force unfolding by stabilizing the native state (compare Figs. 3.2A and 3.2B), while for protein G, increasing pH destabilizes the NSE (when pH > 3.4) allowing smaller mechanical forces to unfold the protein (compare Figs. 3.3A and 3.3B). These divergent pH effects on mechanical unfolding are particularly clear when the native state stability is plotted as a function of f and pH (Fig. 3.4).

3.2.4 $f_{1/2}$ is a linear function of temperature and urea concentration and is non-linear with pH:

For both CI2 and protein G, the midpoint unfolding force ($f_{1/2}$), behaves linearly over a wide temperature range (280 K to 320 K) and urea concentration range (Figs. 3.5A and 3.5B). Increasing temperature or urea leads to smaller $f_{1/2}$ values for both proteins. $f_{1/2}$ as a function of pH is non-linear (Figs. 3.5A and 3.5B). At high (pH > 5) and low (pH < 2) pH values, $f_{1/2}$ is largely unchanged. At intermediate pH values ($2 <$ pH < 5) $f_{1/2}$ increases for CI2 (Fig. 3.5A) - correlating with its increased native stability (Fig. 3.1A). For protein G, at these intermediate pH values, $f_{1/2}$ is non-monotonic - with the maximum $f_{1/2}$ value occurring at a pH value of 3.4, the same pH at which the maximum native state stability occurs when $f = 0$ (Fig. 3.1B).

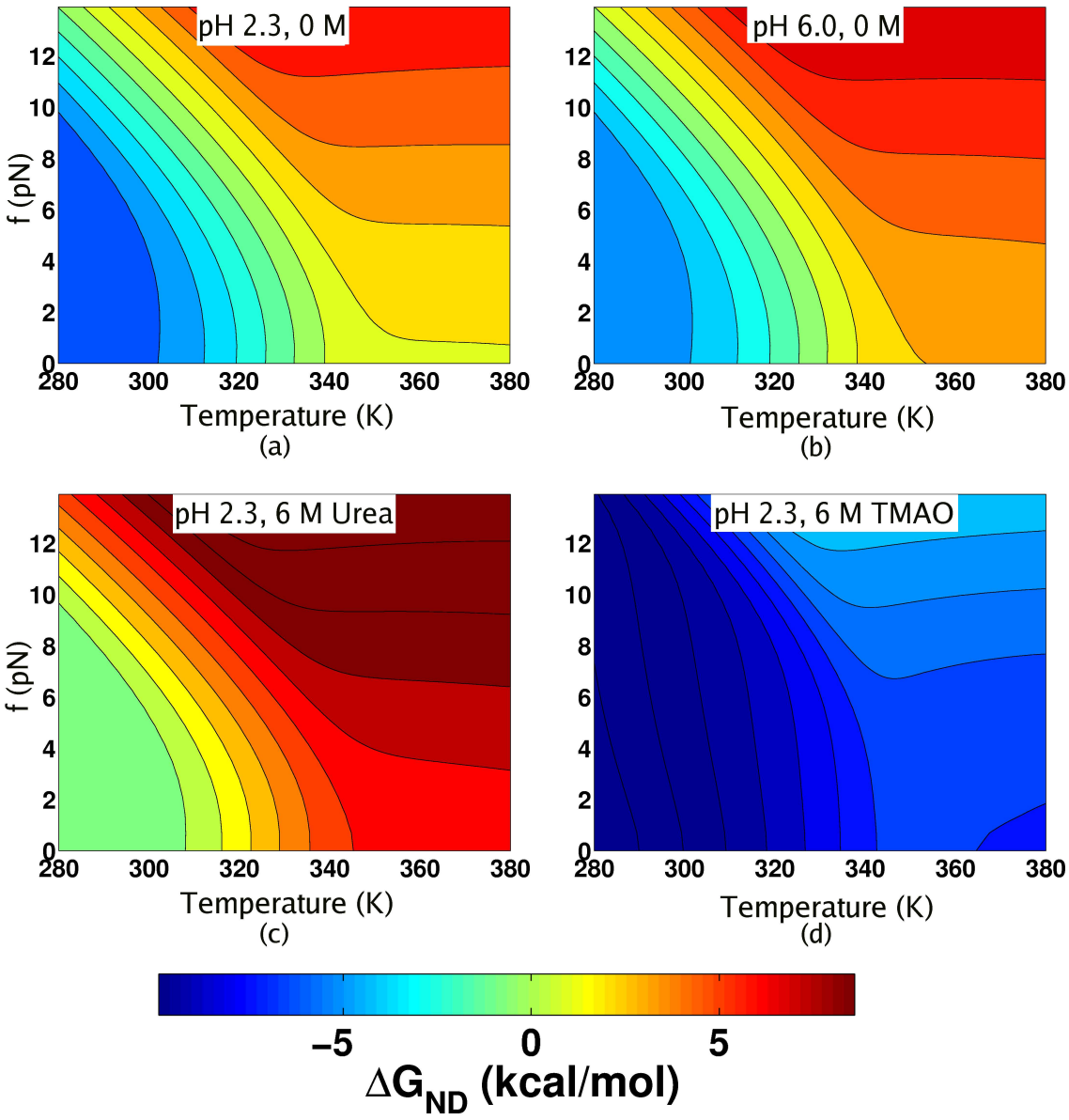
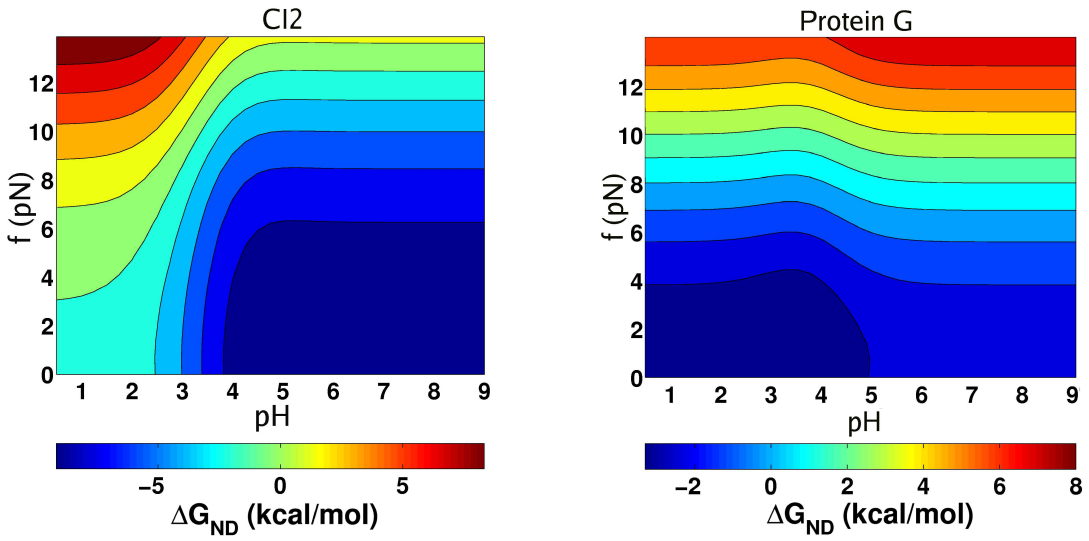


Figure 3.3: The native state stability of protein G as a function of force and temperature under various solution conditions. Unless otherwise stated, the solution conditions in all panels correspond to 0 M cosolutes and a pH of 2.3. In (b) pH=6, (c) 6 M urea is added to solution, and in (c) 6 M TMAO is added to solution.



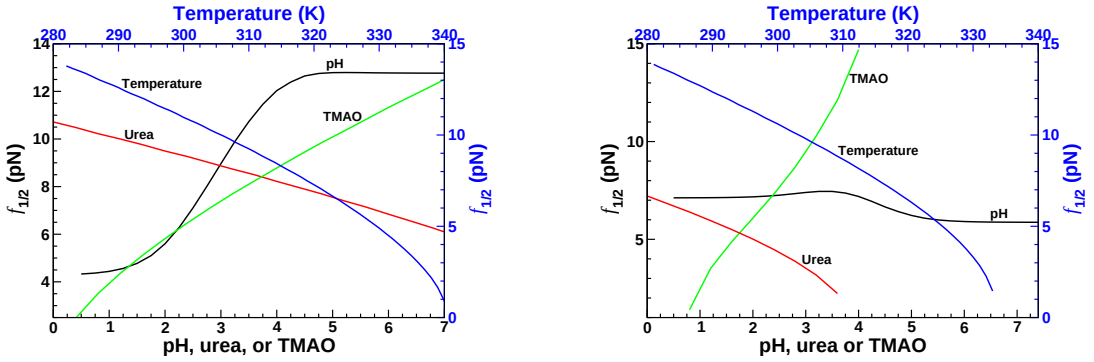
(a) (b)

Figure 3.4: The native state stability of (a) CI2 and (b) protein G as a function of force and pH at 0 M cosolutes and T_S is 302 K and 317 K for CI2 and protein G, respectively.

3.2.5 The change in transition state location exhibits Hammond-Leffler behavior:

According to the Hammond-Leffler postulate [100], the transition state (TS) should resemble the least stable species in the reaction. For proteins under tension, this implies that the location of the TS, denoted x_{TS} and defined as the x-value at which $F(x)$ is a maximum, should shift towards native state x-values when solution conditions destabilize the native state. Below we examine how changes in temperature, urea, TMAO and pH shift x_{TS} while protein G and CI2 are under an applied external tension.

Temperature: For both proteins, x_{TS} shifts towards the native state with increasing temperature (Figs. 3.6A and 3.6B). For example, for CI2 (protein G) at 0



(a) (b)

Figure 3.5: The midpoint unfolding force ($f_{1/2}$) versus pH, urea, or TMAO (lower abscissa) and temperature (upper abscissa) for (a) CI2 and (b) protein G. Unless otherwise stated, the solution conditions for CI2 are 302 K, pH 3.5 and 0 M cosolutes, and for protein G the conditions are 317 K, pH 2.3, and 0 M cosolute.

M cosolute and pH 3.5 (2.3), increasing the temperature from 290 K to 330 K shifts x_{TS} from 70 Å (30 Å) to 27 Å (27.5 Å). CI2 exhibits a much larger change in x_{TS} than protein G. This is due in part to a two-step force unfolding mechanism that CI2 undergoes at temperatures below 305 K. This two stage mechanical unfolding is indicated by the $F(x)$ profile at 285 K (Fig. 3.6C), which exhibits two plateaus for $x > 20$ Å. The fraction of native contacts for the various SE's (Q_{SE}, Fig. 3.7) indicate that at these low temperatures the transition from the native basin to the first plateau (located between 25 and 55 Å) corresponds to the unfolding of β-strands 2-3, 2-4 and loss of tertiary interactions between helix 1 and β-strand 3. The transition to the second plateau (located at $x > 70$ Å) corresponds to the unfolding of the rest of the SE's in the protein. At temperatures higher than 305 K the force unfolding of CI2 is a one-step 'all-or-none' transition as indicated by the two-basin $F(x)$ profile (Fig. 3.6C) and the decrease in dispersity of Q_{SE} vs. x (Fig. 3.7).

The mechanical unfolding of protein G on the other hand is an all-or-none process at all temperatures (Fig. 3.6B), leading to smaller shifts in x_{TS}. Thus, x_{TS} clearly exhibits Hammond-Leffler behavior as a function of temperature. The magnitude of these shifts are dependent in part on the presence of metastable states along the mechanical unfolding pathway, which can lead to non-continuous changes in x_{TS} as observed for CI2.

pH, Urea and TMAO: For CI2, acidic pH's shift x_{TS} towards the native state by up to 43 Å while increases in urea concentration can shift x_{TS} by a similar magnitude (Fig. 3.6A). For protein G, urea shifts x_{TS} by up to 8 Å towards the native state at concentrations above 2 M. pH has no effect on x_{TS} for protein G. TMAO is found to have a large effect on x_{TS} for both protein G and CI2, shifting x_{TS} towards the denatured state by up to 15 Å and 42 Å respectively. These results are in accord with the Hammond-Leffler postulate.

3.2.6 The rank ordering of $f_{1/2}$ for various structural elements is largely unchanging with solution conditions:

We characterize the relative mechanical stability of various SE's by computing their $f_{1/2}$ values at various solution conditions. (Table 3.2 and 3.3). For protein G we find that while the magnitude of $f_{1/2}$ values change with solution conditions (Table 3.2), the relative ordering of $f_{1/2}$ between various SE's does not. The only exception to this finding is the $f_{1/2}$ for protein G's helix (denoted $f_{1/2}^{H}$). At pH 7 $f_{1/2}^{H}$ is smaller in magnitude than all other SE's $f_{1/2}$ values. At the other solution

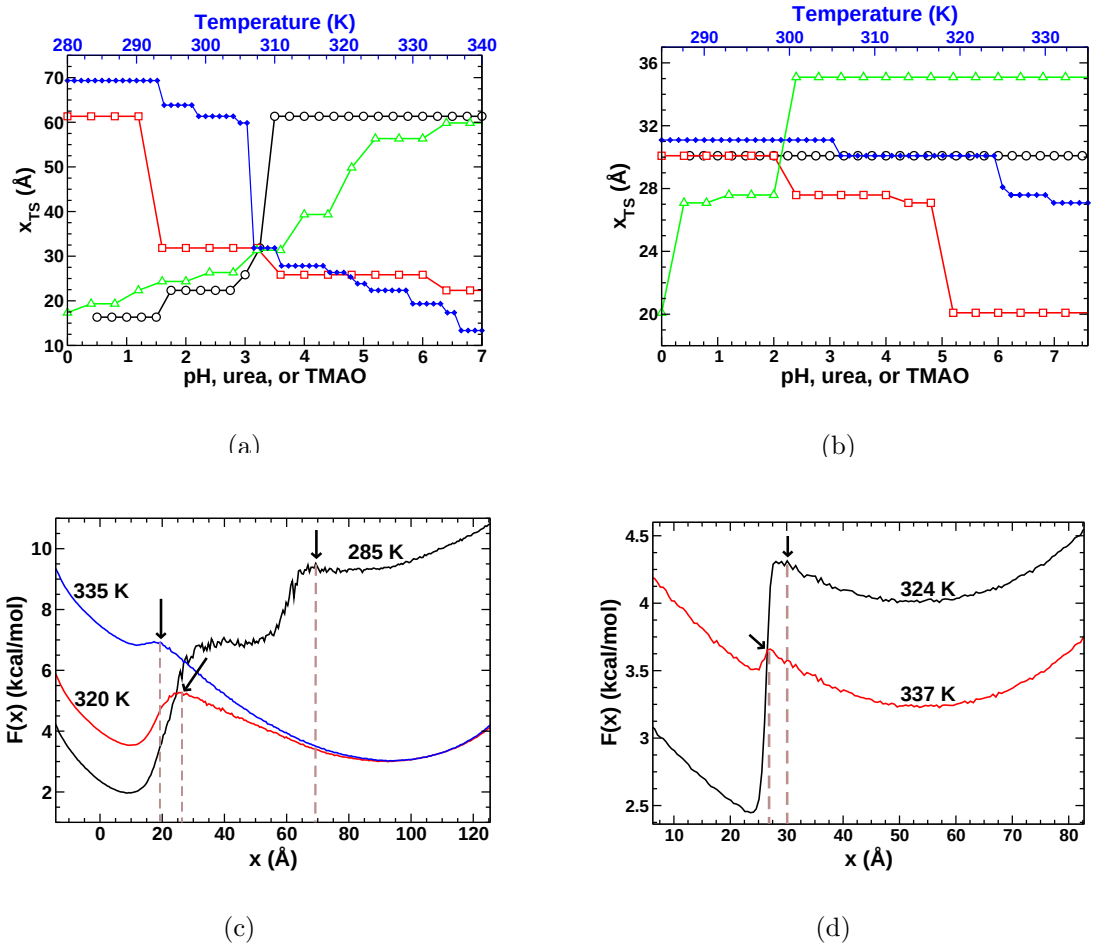


Figure 3.6: The transition state location versus pH, urea, or TMAO (lower abscissa) and temperature (upper abscissa) for (a) CI2, under a constant force $f = 8.34$ pN, and (b) protein G, at $f = 4.2$ pN. In both (a) and (b) Blue diamonds, red squares, black circles, and green triangles correspond, respectively to data for temperature, urea, pH and TMAO. The free energy profile as a function of x ($F(x)$) for CI2 at $f = 8.34$ is shown in (c) at several different temperatures and for protein G at $f = 4.2$ pN, $F(x)$ is shown in (d).

Table 3.2: Midpoint unfolding force ($f_{1/2}$) of protein G's structural elements under various solution conditions

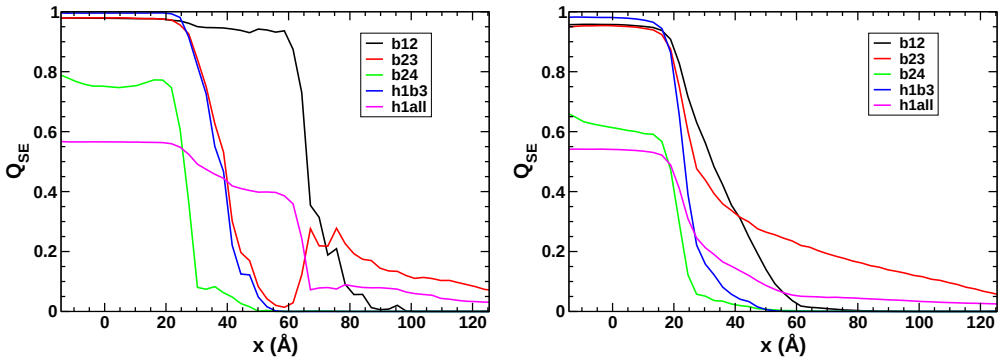
	$f_{1/2}$ (pN)				
Structural element	Standard[a]	317→325 K	2.3→7.0 pH	0→3.0 M urea	0→3.0 M TMAO
S_{12}	7.2	5.1	5.8	3.6	12.2
S_{14}	7.2	5.0	5.8	3.5	12.2
S_{34}	7.2	5.0	5.6	3.5	12.2
Helix (H)	7.4	5.2	4.5	3.9	12.4
$H - S_{12}$	7.2	5.1	5.9	3.6	12.2
$H - S_{34}$	7.2	5.1	5.8	3.6	12.2

[a]Solution conditions of T =317 K, pH=2.3, 0 M osmolytes.

Table 3.3: Midpoint unfolding force ($f_{1/2}$) of CI2's structural elements under various solution conditions

	$f_{1/2}$ (pN)				
Structural element	Standard[a]	302→325 K	3.5→1.0 pH	0→3.0 M urea	0→3.0 M TMAO
S_{12}	10.7	6.2	8.2	8.9	>13.9 [b]
S_{23}	10.8	6.3	Unf[c]	8.9	>13.9
S_{24}	10.5	5.6	Unf	8.5	>13.9
$H_1 - all$	10.7	6.2	8.1	8.9	>13.9
$H_1 - S_3$	10.7	6.1	7.6	8.9	>13.9

[a]Solution conditions of T =302 K, pH=3.5, 0 M osmolytes.
[b]This structural element remained folded in the range of pulling forces (0-13.9 pN) applied in this study.
[c]Unfolded under these solution conditions, i.e. $f_{1/2} = 0$ pN.



(a) (b)

Figure 3.7: The fraction of native contacts for various structural elements (Q_{SE}) of CI2 as a function of the distance between the N-terminus and C-terminus projected on to the x-axis (the direction of pulling) at (a) 280 K and (b) 320 K.

conditions listed in Table 3.2 $f_{1/2}^{H}$ is larger than all other SE $f_{1/2}$ values.

We find similar results for CI2, the rank ordering of the SE's remains essentially the same under all solution conditions (Table 3.3). One exception occurs when the pH is changed from a value of 3.5 to 1.0. In this instance, the $f_{1/2}$ values of SE's S_{23} and S_{24} are equal to 0, that is they are unfolded at all forces, under these solution conditions. At the other solution conditions listed in Table 3.3 these SE's have $f_{1/2}$ values that are similar in magnitude to the other SE's. These results suggest that the unfolding pathways of proteins under tension may show greater dispersity with pH changes than with changes in temperature or osmolyte concentration. Thus, while changing ξ can modify the mechanical stability of SE's, in most instances the relative rank ordering between them does not change.

Table 3.4: Urea and TMAO m-values at various forces ($m \equiv (\Delta G_{ND}([C]) - \Delta G_{ND}([0]))/[C]$)

	Urea m-value[a]		TMAO m-value	
Force (pN)	CI2[b]	protein G[c]	CI2[d]	protein G[e]
0.0	0.50	0.79	-1.31	-1.47
1.4	0.51	0.79	-1.34	-1.49
2.8	0.54	0.80	-1.39	-1.53
4.2	0.58	0.82	-1.44	-1.57
7.0	0.69	0.85	-1.49	-1.64
8.3	0.73	0.86	-1.47	-1.66

[a]m-values are in units of $kcal\ M^{-1}\ mol^{-1}$
[b]CI2's urea m-value was computed at T = 302 K and pH 3.5
[c]Protein G's urea m-value was computed at 317 K and pH 2.3
[d]CI2's TMAO m-value was computed at 340 K and pH 3.5
[e]Protein G's TMAO m-value was computed at 340 K and pH 2.3

3.2.7 The m-value increases with increasing f:

We now compute the m-values of CI2 and protein G at various values of f to test the hypothesis that the m-value will increase as a function of f. The results, listed in Table 3.4, clearly show that the urea m-value does indeed increase by as much as 46% and 9% for CI2 and protein G, respectively. This means that under tension, these proteins are more susceptible to chemical denaturation. On the other hand, TMAO m-values also increase in magnitude (Table 3.4) by as much as 13%, indicating that under tension these proteins are more susceptible to chemical renaturation. Plotting the total SASA (α_T) as a function of f (Fig. 3.8) shows that increasing f leads to greater SASA's in the DSE. CI2 exhibits much larger changes in α_T under applied tension than protein G (Fig. 3.8). The increase in SASA, according to the Tanford transfer model (TTM), explains the positive correlation between the m-values and f. According to the TTM, greater

SASA in the DSE leads to a greater interaction free energy between the protein and solution, while the NSE's SASA and interaction free energy are largely unchanged by f. Thus, the difference in the protein-solvent interaction free energy between the NSE and DSE increases with f. Therefore, proteins under higher tension will have larger changes in $\Delta\Delta G_{ND}(\Delta[C])$ than proteins under lower tension, i.e. $\Delta\Delta G_{ND}(\Delta[C], f_{High}) > \Delta\Delta G_{ND}(\Delta[C], f_{low})$, where $f_{High} > f_{low}$. From the perspective of preferential binding theories, this is equivalent to saying that the number of urea molecules bound to the DSE increases with increasing f due to an increase in the number of binding sites in the DSE, while the number of urea molecules bound to the NSE is largely unchanged.

3.3 Conclusions

We have shown that many of the effects of varying pH, osmolyte concentration, and temperature on proteins under tension ($f \neq 0$) can be *qualitatively* predicted based on knowledge of that protein's properties at $f = 0$. The stabilizing effect of TMAO was found to make mechanical unfolding more difficult by increasing $f_{1/2}$, while the denaturing effect of urea made mechanical unfolding easier. $f_{1/2}$ had a non-linear dependence on pH; acidic pH's increased $f_{1/2}$ for protein G and decreased $f_{1/2}$ for CI2. These results correlate with the behavior of ΔG_{ND} as a function of these solution conditions when $f = 0$. In addition, we have shown that the transition state location follows Hammond-Leffler behavior at all forces and solution conditions

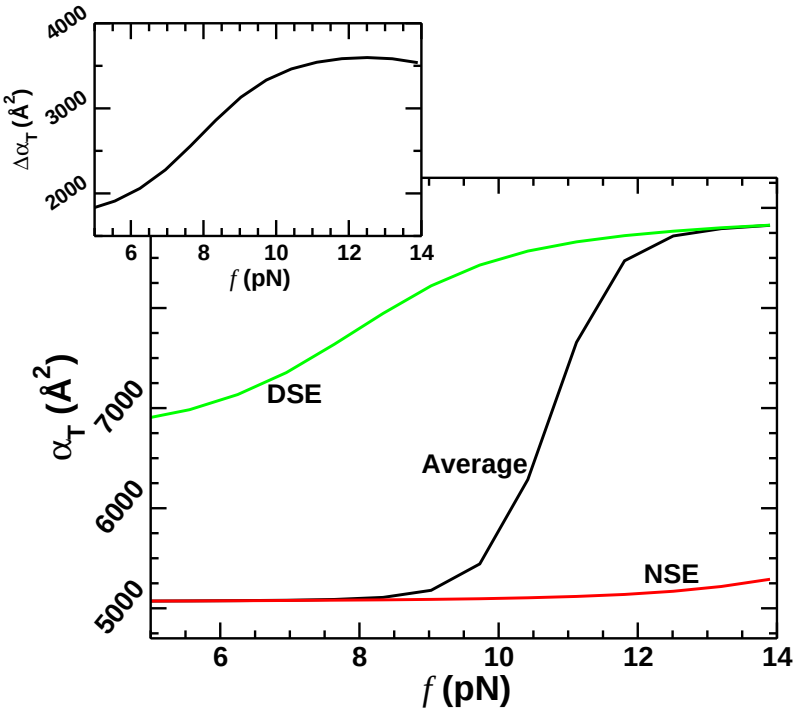


Figure 3.8: The total solvent accessible surface area (α_T) of CI2 versus the applied external force f. The average, DSE and NSE α_T are shown as black, green and red lines respectively. (Upper left panel) The difference in SASA between the NSE and DSE ($\Delta\alpha_T = \alpha_T^D - \alpha_T^N$) versus f. All data is at 302 K, pH 3.5, and 0 M cosolutes.

studied.

Perhaps one of the most surprising results is that the m-value can change significantly with f, increasing by as much as 46%. This large change in the m-value when $f > 0$ is due to a more extended DSE with greater SASA. We predict that larger proteins will exhibit even greater absolute changes in m-values since m-values, on average, are proportional to the number of amino-acids in a protein [101].

These results are relevant to cellular machinery such as chaperones, translocons, and proteosomes, which may mechanically unfold proteins [81, 82, 83, 84, 85, 86]. They suggest that low concentrations of urea (or another incompatible osmolyte) may make it easier for these machines to do their jobs, whereas the presence of counteracting osmolytes (such as TMAO) may make it harder to force unfold proteins in the cell.

We have also extended the capabilities of the MTM to model pH effects on proteins. We showed that the MTM can accurately reproduce the ΔG_{ND} vs. pH profile of CI2. The MTM is useful because it not only can predict quantities that can be directly compared to experiment but also offers a molecular level interpretation of these phenomena based on the simulation structures it utilizes.

3.4 Methods

3.4.1 CI2 and protein G models:

We model the 65 residue protein CI2 and 56 residue protein G using the C_α side chain model ($C_\alpha - SCM$) [59]. Details of the $C_\alpha - SCM$ have been published elsewhere [95], we briefly describe the model here. In the $C_\alpha - SCM$ each amino acid is represented as two interaction sites. One interaction site is located at the α carbon position of the backbone. If the amino acid has a side chain, the other interaction site is located at the side chain center-of-mass. The $C_\alpha - SCM$ is a Go model [6], side chains that are in contact or backbone groups that form hydrogen bonds in the crystal structure have attractive non-bonded Lennard-Jones interactions while all other non-bonded interactions are repulsive. Sequence dependent effects are modeled using non-bonded interaction parameters that are a function of the amino acid pairs that are interacting. In addition, the excluded volume of an amino acid side chain is proportional to its experimentally measured partial molar volume in solution. We use the crystal structures with PDB codes 2CI2 [102] and 1GB1 [103] for CI2 and protein G respectively.

3.4.2 The Molecular Transfer Model for osmolyte and pH effects on proteins under tension:

The MTM [95] utilizes protein conformations from the $C_\alpha - SCM$ simulations, experimentally measured, or theoretically computed, amino acid transfer free

energies and the Weighted Histogram Equations [37] to predict how changes in osmolyte type, osmolyte concentration, or pH effect the thermodynamic properties of a protein. The MTM equation is

$$\langle A([C_i], pH_2, T, f)\rangle \;=\; Z([C_i], pH_2, T, f)^{-1} \sum_{k=1}^{R} \sum_{t=1}^{n_k} \frac{A_{k,t} e^{-\beta E_P(k,t,[C_i],pH_2,f)}}{\sum_{m=1}^{R} n_m e^{F_m - \beta_m E_m(k,t,[0]) - f_m x(k,t)}}, \tag{3.3}$$

where $Z([C_i], pH_2, T, f)$ is the partition function under the solution condition of interest. A given solution condition is uniquely defined by the type of osmolyte (i) and its concentration ($[C_i]$) in solution and the pH and temperature (T) of solution. In Eq. 3.2, R is the number of independent simulated trajectories, n_k is the number of saved protein conformations from the k^{th} simulation, $A_{k,t}$ is the value of protein property A for the t^{th} conformation, $\beta = 1/k_B T$, where k_B is Boltzmann's constant and T is the temperature. The potential energy E_P of the t^{th} conformation from the k^{th} simulation in the presence of osmolyte i at concentration $[C_i]$, pH_2, and under external force f is $E_P(k,t,[C_i],pH_2,f) = E_P(k,t,[0],pH_1) + \Delta G_{tr}(k,t,[C_i],pH_1) + \Delta G_{tr}(k,t,[0],pH_2) - fx(k,t)$, where $E_P(k,t,[0],pH_1)$ is the potential energy of the system at 0 M osmolyte and pH_1, i.e. the osmolyte and pH conditions under which the simulations are carried out in this study. $\Delta G_{tr}(k,t,[C_i],pH_1)$ is the free energy cost of transferring the t^{th} conformation in the k^{th} simulation from 0 M to $[C_i]$ at pH_1, and $\Delta G_{tr}(k,t,[0],pH_2)$ is the free energy of transferring that conformation from pH_1 to pH_2 at 0 M osmolyte. f is the applied force constant and x is the end-to-end distance vector of the protein projected onto the x-axis - the direction of

the applied force. In the denominator of Eq. 3.2, n_m and F_m are, respectively, the number of conformations and the free energy of the m^{th} simulation. F_m is solved for self consistently at the simulated solution conditions as described in reference [37].

To estimate $\Delta G_{tr}(k, t, [C_i], pH_1)$ we use the Tanford Transfer Model (TTM) [32, 27] and assume that $\Delta G_{tr}(k, t, [C_i], pH_1)$ is independent of pH (i.e. $\Delta G_{tr}(k, t, [C_i], pH_1) \rightarrow \Delta G_{tr}(k, t, [C_i])$. In the TTM the free energy cost of transferring the l^{th} protein conformation from water to aqueous osmolyte solution at concentration [C] is written as

$$\Delta G_{tr}(l, [C_i]) = \sum_{k=1}^{N_S} \delta g_k^S([C_i]) n_k \alpha_{l,k}^S / \alpha_{k,G-k-G}^S + \sum_{k=1}^{N_B} \delta g_k^B([C_i]) n_k \alpha_{l,k}^B / \alpha_{k,G-k-G}^B \quad (3.4)$$

where the summations are over the side chain (S) and backbone (B) groups of different amino acid types in the protein, n_k is the number of amino acid residues of type k(= ala, gly, arg, etc.), and δg_k^S and δg_k^B are the transfer free energies of the side chain and backbone group of amino acid type k, respectively [26, 62]. The average solvent accessible surface areas of the side chain and backbone group of amino acid type k are $\alpha_{i,k}^S$ and $\alpha_{i,k}^B$, respectively, and $\alpha_{k,G-k-G}^S$ is the solvent accessible surface area of the side chain and backbone in the tripeptide $Glycine - k - Glycine$. The values $\alpha_{k,G-k-G}$ for the side chain and backbone groups (Eq. 3.4) are taken from [95]. For the osmolytes considered here (urea and TMAO) we use the experimentally measured δg_k data reported in [28]. To estimate δg_k at osmolyte concentrations that were not experimentally measured, we use a linear extrapolation [21, 77, 95].

To estimate $\Delta G_{tr}(k, t, [0], pH_2)$ we use a model developed by Tanford and coworkers [32] in which the free energy of transferring a titratible group k in con-

Table 3.5: pK$_a$ values of titratible side chains in the native and denatured states

CI2[a]				protein G[b]			
Number	Residue	pK_a^N	pK_a^D	Number	Residue	pK_a^N	pK_a^D
4	Glu	2.9	4.0	15	Glu	4.4	4.0
7	Glu	2.9	4.0	19	Glu	3.7	4.0
14	Glu	3.5	4.0	22	Asp	2.9	3.6
15	Glu	2.8	4.0	27	Glu	4.5	4.0
23	Asp	2.4	3.6	36	Asp	3.8	3.6
26	Glu	3.65	4.0	40	Asp	4.0	3.6
41	Glu	3.14	4.0	42	Glu	4.4	4.0
45	Asp	3.6	3.6	46	Asp	3.6	3.6
52	Asp	2.5	3.6	47	Asp	3.4	3.6
55	Asp	4.95	3.6	56	Glu	4.0	4.0

[a]Values taken from [99].
[b]Values taken from [97].

formation l of the protein from a solution at pH_1 to pH_2 is

$$\delta g_{k,l} = -k_B T ln \left[\frac{10^{pH_2} + \Theta_N(l)10^{pK_{N,k}} + \Theta_D(l)10^{pK_{D,k}}}{10^{pH_1} + \Theta_N(l)10^{pK_{N,k}} + \Theta_D(l)10^{pK_{D,k}}} \right], \tag{3.5}$$

where $\Theta_N(l)$ and $\Theta_D(l)$ are Heaviside step functions that identify a conformation l as being either native or denatured. $\Theta_N(l)$ ($\Theta_D(l)$) is one if conformation l is native (denatured) and zero otherwise. $pK_{N,k}$ and $pK_{D,k}$ are the pK_a values for group k in the native and denatured states respectively. We use $pK_{N,k}$ values that have been determined experimentally [99, 97] and list them in Table 3.5. Details on defining native and denatured conformations are given below. The second step is to sum up the $\delta g_{k,l}$ to compute $\Delta G_{tr}(k, t, [0], pH_2)(= \sum_{k=1}^{N_k} \delta g_{k,l})$.

3.4.3 Limitations of the MTM:

A number of assumptions underly the MTM including the temperature independence of $\Delta G_{tr}(k, t, [C_i], pH_1)$ and $\Delta G_{tr}(k, t, [0], pH_2)$, the pH (osmolyte) inde-

pendence of $\Delta G_{tr}(k, t, [C_i], pH_1)$ ($\Delta G_{tr}(k, t, [0], pH_2)$), and constant pK_a values for the NSE and DSE. While violations of these assumptions can lead to disagreement between MTM predictions and experiment, the excellent agreement observed previously [95] and in this study at specific solution conditions gives us confidence that, at a minimum, the MTM predictions will be in qualitative agreement with experiment. A number of assumptions are inherent to the Tanford models for osmolytes and pH, see references [26] and [32, 35] for a detailed discussion of these assumptions.

3.4.4 Simulation details:

We use Hamiltonian Replica Exchange (HREX) [36, 104] in the canonical (NVT) ensemble to obtain equilibrium simulations of CI2 and protein G under constant force applied in the positive x-direction to the C-terminal C_α bead of the protein. The N-terminal C_α bead is held fixed at the origin. In this HREX simulation, independent trajectories (replicas) are simulated at different temperatures and under different forces using Langevin dynamics [105] with a damping coefficient of 0.8 ps^{-1} and an integration time-step of 6 fs. We use the program CHARMM (version c33b2) to simulate the time evolution of the replicas [106]. Every 5,000 (7,000) integration time-steps CI2's (protein G's) system coordinates are saved for each replica and then exchanged, either between neighboring temperatures or between neighboring external forces (Hamiltonians) according to exchange criteria that preserve detailed balance [36]. 90,000 exchanges, alternating between temperature and force exchanges, were attempted. The first 10,000 exchanges were discarded to allow for equilibration.

For CI2, five temperature windows (300, 317, 330, 345, 380 K) and eight force constants ($f = 0.00, 0.35, 3.47, 8.68, 9.03, 9.38, 9.73, 10.42, 13.89$ pN) were used for a total of forty replicas. For protein G four temperature windows (310, 320, 330, 370 K) and ten force constants ($f = 0.00, 0.35, 1.60, 2.85$,$4.10, 5.35, 6.60, 7.85, 9.10, 10.42, 13.89$ pN) were used for a total of forty replicas. Swap acceptance ratios of between 10 and 40% were achieved in the MHREX runs.

3.4.5 Analysis:

A protein conformation is defined to be native if the root-mean-squared-distance (RMSD) of its C_α beads are within 5 Å, for protein G, or 11 Å, for CI2, of the corresponding C_α atoms in the crystal structure after a least squares minimization alignment is performed. A conformation is considered denatured if its RMSD $>$ 5 Å for protein G and $>$ 11 Å for CI2. CI2's larger RMSD cutoff is due to disordered random coil regions in the NSE (see Fig. 3.1A). The solvent accessible surface area of a conformation, used in Eq. 3.4, is computed analytically using a probe radius of 1.4 Å in the program CHARMM [106].

Two-dimensional native state stability diagrams (e.g. $\Delta G_{ND}(f,T)$, etc.) are computed by rewriting Eq. 3.2 into the probability of being folded as a function of f and T as

$$P_N(f,T) \;=\; Z([C_i],pH_2,T,f)^{-1}\sum_{k=1}^{R}\sum_{t=1}^{n_k}\frac{\Theta_N(k,t)e^{-\beta E_P(k,t,[C_i],pH_2,f)}}{\sum_{m=1}^{R} n_m e^{F_m-\beta_m E_m(k,t,[0])-f_m x(k,t)}}, \quad (3.6)$$

and using $\Delta G_{NU}(f,T) = -k_B T ln(P_N(f,T)/(1-P_N(f,T)))$. All terms in Eq. 3.6 are the same as in Eq. 3.2 except we use the Heaviside step function $\Theta_N(k,t)$, which is one if conformation (k,t) is native (see above) and zero otherwise. $f_{1/2}$ values

were determined by solving for the f value at which $P_N(f, \xi) \approx 0.5$.

Chapter 4
How accurate are polymer models in the analysis of FRET experiments on proteins?
4.1 Introduction

Much of our understanding of how proteins fold comes from experiments in which folding is initiated from an ensemble of unfolded molecules whose structures are hard to characterize [107]. In many experiments, the initial structures of the denatured state ensemble (DSE) are prepared by adding an excess amount of denaturants or by raising the temperature above the melting temperature (T_m) of the protein [3]. Theoretical studies have shown that folding mechanisms depend on the initial conditions, i.e. the nature of the DSE [108]. Thus, a quantitative description of protein folding mechanisms requires a molecular characterization of the DSE - a task that is made difficult by the structural diversity of the ensemble of unfolded states [53, 109].

In an attempt to probe the role of initial conditions on folding, single molecule FRET experiments are being used to infer the properties of unfolded proteins. The major advantage of these experiments is that they can measure the FRET efficiencies of the DSE under solution conditions where the native state is stable. The average denaturant-dependent FRET efficiency $\langle E \rangle$ has been used to infer the global properties of the polypeptide chain in the DSE as the external conditions are altered. The properties of the DSE are inferred from $\langle E \rangle$ by assuming a polymer model for the DSE, from which the root mean squared distance between two dyes

attached at residues i and j along the protein sequence ($R_{ij} = \langle|\mathbf{r}_i - \mathbf{r}_j|\rangle$), the distribution of the end-to-end distance $P(R)$ (where $R = |\mathbf{r}_N - \mathbf{r}_0|$), the root mean squared end-to-end distance ($R_{ee} = \langle\mathbf{R}^2\rangle^{1/2}$), the root mean squared radius of gyration ($R_g = \langle\mathbf{R}_g^2\rangle^{\frac{1}{2}}$), and the persistence length (l_p) of the denatured protein [110, 111, 55, 112, 113, 63, 114, 56, 64, 10] can be calculated.

In FRET experiments, donor (D) and acceptor (A) dyes are attached at two locations along the protein sequence [115, 53], and hence can only provide information about correlations between them. The efficiency of energy transfer E between the D and A is equal to $(1 + r^6/R_0^6)^{-1}$, where r is the distance between the dyes, and R_0 is the dye-dependent Förster distance [115, 53]. Because of conformational fluctuations, there is a distribution of r, $P(r)$, which depends on external conditions such as the temperature and denaturant concentration. As a result, the average FRET efficiency $\langle E\rangle$ is given by

$$\langle E\rangle \;=\; \int_0^\infty (1 + r^6/R_0^6)^{-1} P(r) dr, \tag{4.1}$$

under most experimental conditions, due to the central limit theorem [116]. If the dyes are attached to the ends of the chain, then $P(r) = P(R)$. Even if $\langle E\rangle$ is known accurately, the extraction of $P(R)$ from the integral equation (Eq. 4.1) is fraught with numerical instabilities. In applications to biopolymers, a functional form for $P(r)$ is assumed, and the parameters (a - the Kuhn length, l_p, or R_{ee}; see Table I) are adjusted to satisfy the equality in Eq. 4.1 as closely as possible. Typically, $P(R)$ is modeled using the Worm-like Chain (WLC) or Gaussian chain polymer models. For these models, and the Self-Avoiding Walk (SAW) chain model, the $P(R)$ distribution

functions are analytically known (see Table 4.1). Using this method (referred to as the "standard procedure" in this article), several researchers have estimated R_g and l_p as a function of the external conditions for protein L [63, 64], Cold Shock Protein (CspTm) [56], and Rnase H [115]. The justification for using homopolymer models to analyze FRET data comes from the anecdotal comparison of the R_g measured using X-ray scattering experiments and the extracted R_g from analysis of Eq. 4.1.

Here, we study an analytically solvable generalized Rouse model (GRM) [117] and the Molecular Transfer Model (MTM) for protein L [95] to assess the accuracy of using polymer models to solve Eq. 4.1. In the GRM, two monomers that are not covalently linked interact through a harmonic potential that is truncated at a distance c. The presence of the additional length scale, c, which reflects the interaction between non-bonded beads, results in the formation of an ordered state as the temperature (T) is varied. For the GRM, $P(R)$ can be analytically calculated, and hence the reliability of the standard procedure to solve Eq. 4.1 can be unambiguously established. We find that the accuracy of the polymer models in extracting the exact values in the GRM depends on the location of the monomers that are constrained by the harmonic interaction. Using coarse-grained simulations of protein L, we show that the error between the exact quantity and that inferred using the standard procedure depends on the property of interest. For example, the inferred end-to-end distribution $P(R)$ is in qualitative, but not quantitative agreement with the exact $P(R)$ distribution obtained from accurate simulations. In general, the DSE of protein L is better characterized by the SAW polymer model than the Gaussian chain model.

We propose that the accuracy of the popular Gaussian model can be assessed by measuring $\langle E \rangle$ with dyes attached at multiple sites in a protein [118, 119, 56]. If the DSE can be described by a Gaussian chain, then the parameters extracted by attaching the dyes at position i and j can be used to predict $\langle E \rangle$ for dyes at other points. The proposed self-consistency test shows that the Gaussian model only qualitatively accounts for the experimental data of CspTm, simulation results for protein L, and the exact analysis of the GRM.

4.2 Results and Discussion

We present the results and discussion in three sections. In the first and second sections we examine the accuracy of the standard procedure in accurately inferring the properties of the denatured state of the GRM and protein L models. The third section presents results of the Gaussian Self-consistency Test applied to these models. We also analyze experimental data for CspTm to assess the extent to which the DSE deviates from a Gaussian chain.

4.2.1 GRM

The Generalized Rouse model (GRM) is a simple modification of the Gaussian chain with N bonds and Kuhn length a_0, which includes a single, non-covalent bond between two monomers at positions s_1 and s_2 (Fig. 4.1). The monomers at s_1 and s_2 interact with a truncated harmonic potential with spring constant k, with strength $\kappa = kc^2/2$, where c is the distance at which the interaction vanishes (Eq. 4.4). The GRM minimally represents a two state system, with a clear demarcation between

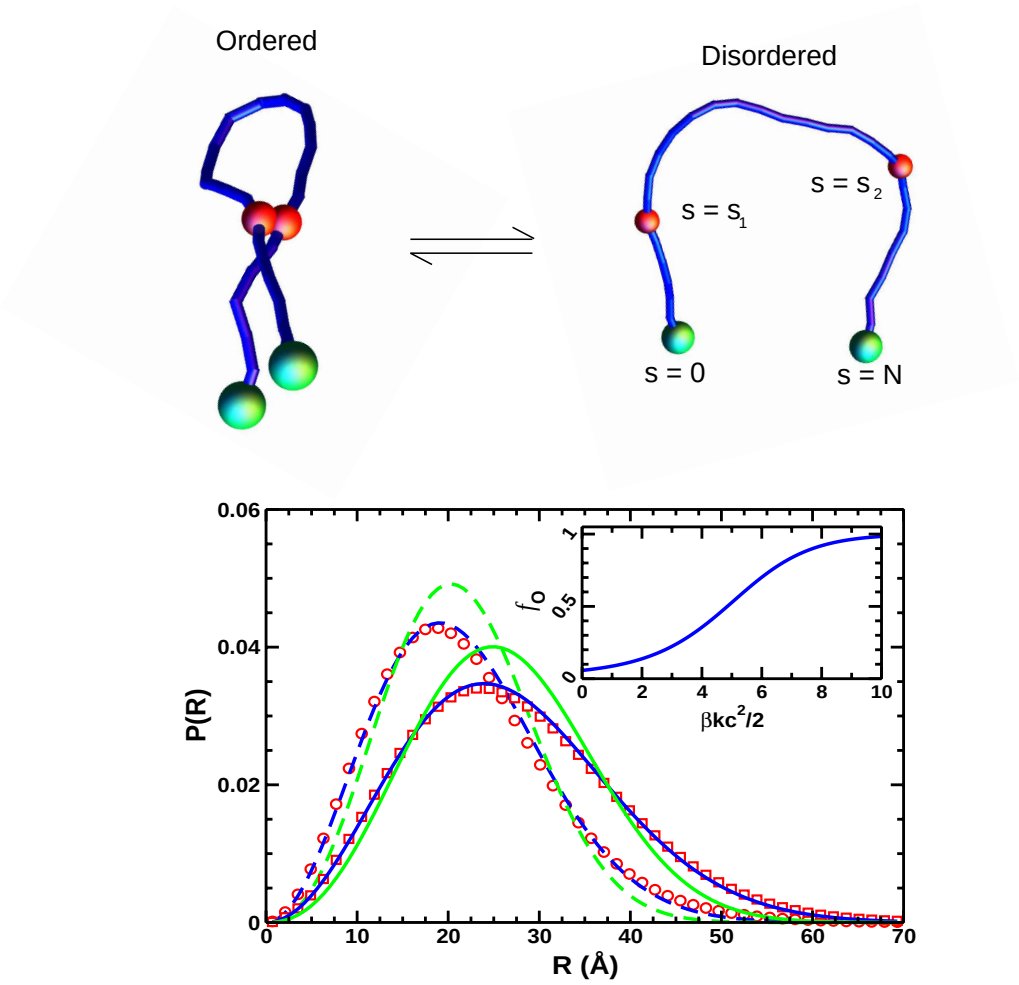


Figure 4.1: Top figures shows a schematic sketch of the GRM, with the donor and acceptor at the endpoints, represented by the green spheres, and the interacting monomers at s_1 and s_2 represented by the red spheres. In the ordered configuration, the monomers at s_1 and s_2 are tightly bound. The bottom figure shows the exact and the inferred end-to-end distribution functions $P(r)$ for interior interactions ($\Delta s = 31$). The blue lines correspond to the Gaussian chain model, light green lines to the SAW, and the symbols to the exact GRM distribution. Dashed lines and red circles are for $\beta\kappa = 6.6$, while solid lines and red squares correspond to $\beta\kappa = 2$. In the inset we show the fraction of ordered states as a function of $\beta\kappa$. Note that 75% of the structures are ordered at $\beta\kappa = 6.6$, yet the inferred Gaussian $P(r)$ is in excellent agreement with the exact result.

ordered (with $|\mathbf{r}(s_2) - \mathbf{r}(s_1)| \leq c$) and disordered (with $|\mathbf{r}(s_2) - \mathbf{r}(s_1)| > c$) states. Unlike other polymer models (see Table I), which are characterized by a single length scale, the GRM is described by a_0 and the energy scale κ. For $\beta\kappa \to 0$ (the high temperature limit, where $\beta = 1/k_B T$), the simple Gaussian chain is recovered (see Methods for details). By varying $\beta\kappa$, a disorder $\to$ order transition can be induced (see Fig. 4.1). The presence of the interaction between monomers s_1 and s_2 approximately mimics persistence of structure in the DSE of proteins. If the fraction of ordered states, f_O, exceeds 0.5 (Fig. 4.1 inset), we assume that the residual structure is present with high probability. The exact analysis of the GRM when $|\mathbf{r}(s_2) - \mathbf{r}(s_1)| \leq c$ allows us to examine the effect of structure in the DSE on the global properties of unfolded states.

Because $\langle E \rangle$ can be calculated exactly for the GRM (see Eq. 4.5), it can be used to quantitatively study the accuracy of solving Eq. 4.1 using the standard procedure [110, 113, 63, 56, 64]. Given the best fit for the Gaussian chain (Kuhn length a), WLC (persistence length l_p), and SAW (average end-to-end distance R_{ee}), many quantities of interest can be inferred ($P(R)$ or R_g, for example), and compared with the exact results for the GRM. The extent to which the exact and inferred properties deviate, due to the additional single energy scale in the GRM, is an indication of the accuracy of the standard procedure used to analyze Eq. 4.1.

4.2.1.1 $P(R)$ is accurately inferred using the Gaussian polymer model:

If the interacting monomers are located near the endpoints of the chain, the end-to-end distribution function is bimodal, with a clear distinction between the

ordered and disordered regions [117]. However, if the monomers s_1 and s_2 are in the interior of the chain, the two-state behavior is obscured because the distribution function becomes unimodal. In Fig. 4.1, we show the exact and inferred $P(R)$ functions for a chain with $N = 63$, $a_0 = 3.8$Å, $c = 2a_0$, and $|s_2 - s_1| = (N-1)/2 = 31$. We take the Förster distance (Eq. 4.1) $R_0 = 23\text{Å}\langle \mathbf{R}^2 \rangle_{\kappa=0}^{1/2}$ for the GRM. The distributions are unimodal for both weakly ($\beta\kappa = 2$) and strongly ($\beta\kappa = 6.6$) interacting monomers.

The strength of the interaction is most clearly captured with the fraction of conformations in the ordered state, f_O, with $f_O = 0.25$ for the weakly interacting chain and $f_O = 0.75$ for the strongly interacting chain (inset of Fig. 4.1). The inferred Gaussian distribution functions are in excellent agreement with the exact result. Because of the underlying Gaussian Hamiltonian in the GRM, the rather poor agreement in the inferred SAW distribution seen in Fig 4.1 is to be expected. We also note that the GRM is inherently flexible, so that the WLC and Gaussian chains produce virtually identical distributions.

4.2.1.2 The accuracy of the inferred R_g depends on the location of the interaction:

The two-state nature of the GRM is obscured by the relatively long unstructured regions of the chain, similar to the effect seen in laser optical tweezer experiments with flexible handles [117]. As a result, $P(R)$ is well represented by a Gaussian chain, with a smaller inferred Kuhn length, $a \leq a_0$ 4.2. For large $\beta\kappa$, where the ordered state is predominantly occupied and $\mathbf{r}(s_2) \approx \mathbf{r}(s_1)$, the end-to-end

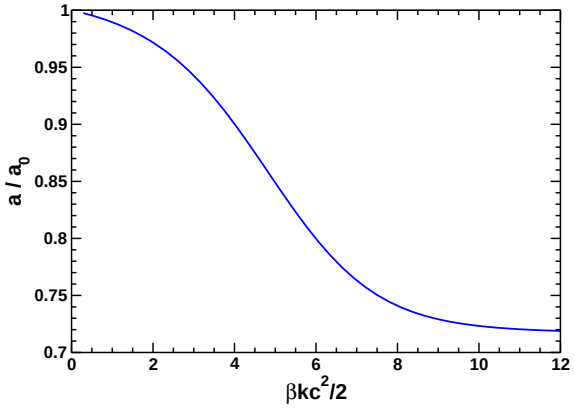


Figure 4.2: The inferred Kuhn length a as a function of $\beta\kappa$ for the GRM. R_{ee} monotonically decreases a function of the interaction strength, leading to the decrease in a/a_0. The Kuhn length a reaches its limiting value of $a \approx a_0\sqrt{1-\Delta s/N}$ when $f_O \approx 1$.

distribution function is well approximated by a Gaussian chain with $N^* = N - \Delta s$ bonds. Consequently, the single length scale for the Gaussian chain, decreases to $a \sim a_0\sqrt{1-\Delta s/N} \approx 0.71 a_0$ for large values of $\beta\kappa$ (Fig. 4.2).

Because the two-state nature of the chain is obscured for certain values of $|s_2 - s_1|$, the Gaussian chain gives an excellent approximation to the end-to-end distribution function. However, the radius of gyration R_g is not as accurately obtained using the Gaussian chain model, as shown in Fig. 4.3. The exact R_g for the GRM reflects both the length scale a_0 and the energy scale $\beta\kappa$, which can not be fully described by the single inferred length scale a in the Gaussian chain. For the GRM, R_g depends not only on the separation between the monomers Δs, but also explicitly on s_1 (i.e. where the interaction is along the chain; see Fig. 4.3 and the Methods section), which can not be captured by the Gaussian chain. If the interacting monomers are in the middle of the chain ($s_1 = (N+1)/4 = 16$ and $\Delta s = 31$), the inferred R_g is in excellent agreement with the exact result (Fig. 4.3). The relative

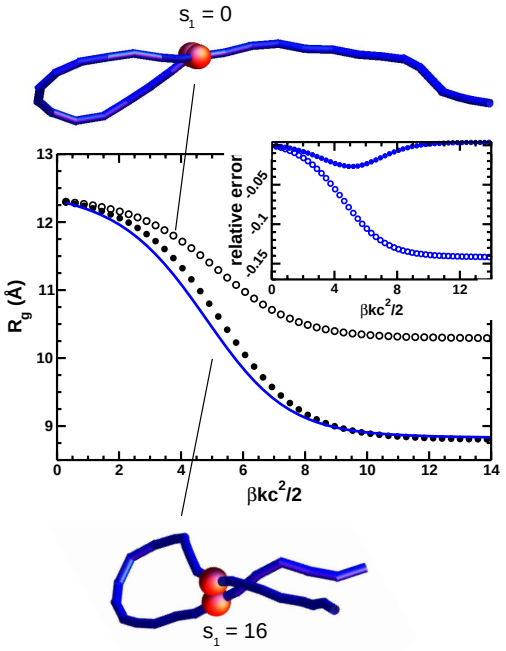
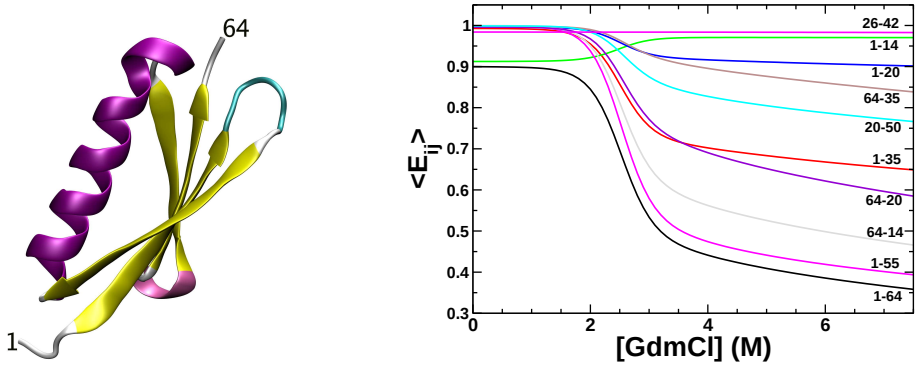


Figure 4.3: Comparison of the exact (symbols) and inferred (blue line) values of the radius of gyration (R_g) as a function of $\beta\kappa$ for $\Delta s = 31$. Shown are R_g's for the GRM with $s_1 = 0$ (open symbols) and $s_1 = 16$ (filled symbols) for $N = 63$. The structures in the ordered state are shown schematically. The R_g obtained using the standard procedure is independent of s_1, while the exact result is not. The inset shows the relative errors between the inferred and exact values of R_g.

error in R_g (the difference between the inferred and exact values, divided by the exact value) is no less than -2%. However, for interactions near the endpoint of the chain, with $s_1 = 0$ and the same $\Delta s = 31$, the relative error between the inferred and exact values of R_g is $\sim -14\%$. The large errors arise because the radius of gyration depends on the behavior of all of the monomers, so that the energy scale $\beta\kappa$ plays a much larger role in the determination of R_g than R_{ee}.

4.2.2 Protein L

Protein L is a 64 residue protein (Fig. 4.4A) whose folding has been studied by a variety of methods [71, 67, 65, 63, 64]. More recently, single molecule FRET experiments have been used to probe changes in the DSE as the concentration of GdmCl is increased from 0 to 7 M [63, 64]. From the measured GdmCl-dependent $\langle E \rangle$, the properties of the DSE, such as R_{ee}, $P(R)$, and R_g, were extracted by solving Eq. 4.1, and assuming a Gaussian chain $P(R)$ [63, 64]. To further determine the accuracy of polymer models in the analysis of $\langle E \rangle$, we use simulations of protein L in the same range of [C] as used in experiments [110, 112].



(a) (b)

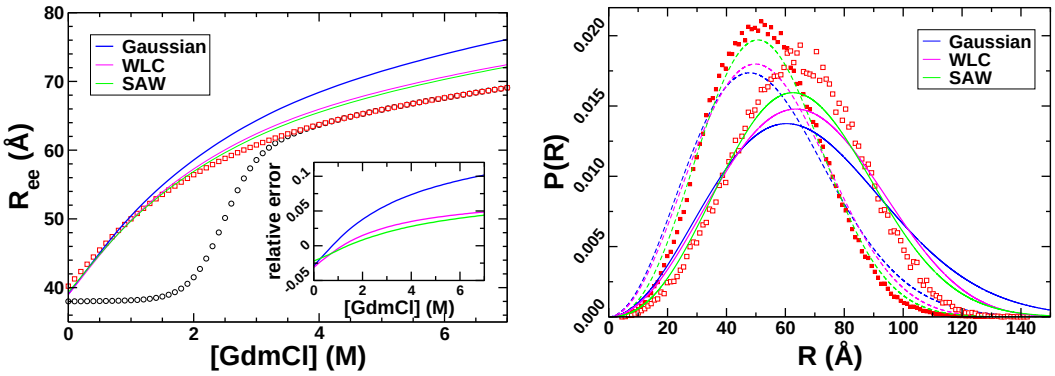
Figure 4.4: (a) A secondary structure representation of protein L in its native state. Starting from the N-terminus, the residues are numbered 1 through 64. (b) The average FRET efficiency between the various (i, j) residue pairs in protein L versus GdmCl concentration. The $\langle E_{ij} \rangle$ values, computed using MTM simulations, for each (i, j) pair is indicated by the two numbers next to each line. For example, the numbers '1-64' beneath the black line indicates that $i = 1$ and $j = 64$. The solid black line (lowest values of $\langle E \rangle$) is computed for the dyes at the endpoints.

4.2.2.1 The average end-to-end distance is accurately inferred from FRET data:

In a previous study [95], we showed that the predictions based on MTM simulations for protein L are in excellent agreement with experiments. From the calculated $\langle E \rangle$ with the dyes at the endpoints (solid black line in Fig. 4.4B), which is in quantitative agreement with experimental measurements [95], we determine the model parameter R_{ee} or l_p by assuming that the exact $P(R)$ can be approximated by the three polymer models in Table 4.1. Comparison of the exact value of R_{ee} to the inferred value R_F, obtained using the simulation results for $\langle E \rangle$, shows good agreement for all three polymer models (Fig. 4.5A). There are deviations between R_{ee} and R_F at [C] $>$ C$_m$, the midpoint of the folding transition. The maximum relative error (see inset of Fig. 4.5A) we observe is about 10% at the highest concentration of GdmCl. The SAW model provides the most accurate estimate of R_{ee} at GdmCl concentrations above C$_m$, with a relative error $\leq$ 0.05, and the Gaussian model gives the least accurate values, with a relative error $\leq$ 0.10 (Fig. 4.5A). Due to the relevance of excluded volume interaction in the DSE of real proteins, the better agreement using the SAW is to be expected.

4.2.2.2 Polymer models do not give quantitative agreement with the exact $P(R)$:

The inferred distribution functions, $P_F(R)$'s, obtained by the standard procedure at [C]=2 M and 6 M GdmCl differ from the exact results (Fig. 4.5B). Surprisingly, the agreement between $P(R)$ and $P_F(R)$ is worse at higher [C]. The range of R explored and the width of the exact distribution are less than predicted



(a) (b)

Figure 4.5: (a) The root mean squared end-to-end distance (R_{ee}) as a function of GdmCl concentration for protein L. The average R_{ee} (black circles) and the R for the sub-population of the DSE (red squares) from simulations are shown. The values of R_{ee} inferred by solving Eq. (1) by the standard procedure using the Gaussian chain, Worm Like Chain, and Self Avoiding polymer models are shown for comparison (solid lines). The inset shows the relative errors between the exact and the values inferred using the FRET efficiency for R_{ee} versus GdmCl concentration are shown. (b) Simulation results of the denatured state end-to-end distance distribution ($P(R)$) at 2.4 M GdmCl (solid red squares) and 6 M GdmCl (open red squares) and T=327.8 K are compared with $P(R)$s using the Gaussian chain, Worm Like Chain, and Self Avoiding Walk polymer models are also shown at 2.4 M GdmCl (dashed lines) and 6 M GdmCl (solid lines).

by the polymer models. The Gaussian chain and the SAW models account only for chain entropy, while the WLC only models the bending energy of the protein. However, in protein L (and in other proteins) intra-molecular attractions are still present even when [C]=6 M > C_m. As a result, the range of R explored in the protein L simulations is expected to be less than in these polymer models. Only at [C]/C_m >> 1 and/or at high T are proteins expected to be described by Flory random coils. Our results show that although it is possible to use models that can give a single quantity correctly (R_{ee}, for example), the distribution functions are less accurate. The results in Fig. 4.5B show that $P(R)$, inferred from the polymer

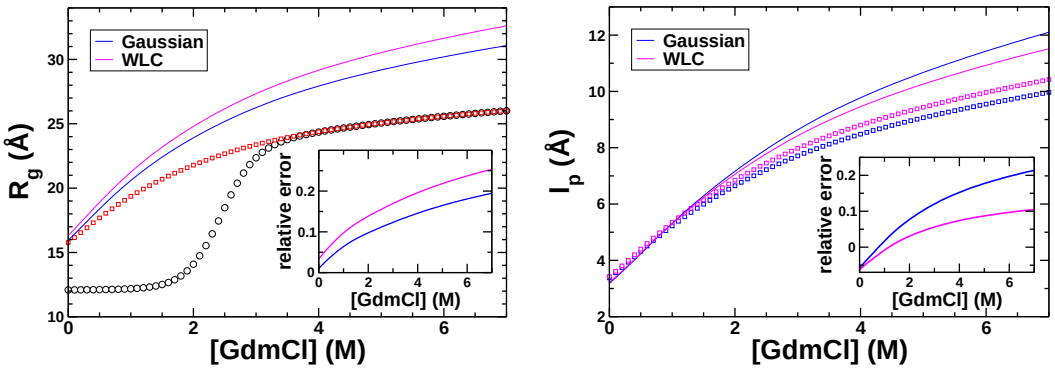
models, agrees only qualitatively with the exact $P(R)$, with the SAW model being the most accurate (Fig. 4.5B).

4.2.2.3 Inferred R_g and l_p differ significantly from the exact values:

The solution of Eq. 4.1 using a Gaussian chain or WLC model yields a and l_p, from which R_g can be analytically calculated (Table 4.1). Figs. 4.6A and 4.6B, which compare the FRET inferred R_g and l_p with the corresponding values obtained using MTM simulations, show that the relative errors are substantial. At high [C] values the R_g^F deviates from R_g by nearly 25% if the Gaussian chain model is used (Fig. 4.6A). The value of $R_g \approx 26$ Å at [C]= 8 M while R_g^F using the Gaussian chain model is ≈ 31 Å. In order to obtain reliable estimates of R_g, an accurate calculation of the distance distribution between all the heavy atoms in a protein is needed. Therefore, it is reasonable to expect that errors in the inferred $P(R)$ are propagated, leading to a poor estimate of internal distances, thus resulting in a larger error in R_g. A similar inference can be drawn about the persistence length obtained using polymer models (Fig. 4.6B). Plotting l_p^F as a function of [C] (Fig. 4.6B), against $l_p = R_{ee}/2L$, shows that l_p is overestimated at concentrations above 1 M GdmCl, with the error increasing as [C] increases. The error is less when the Gaussian chain model is used.

4.2.3 Gaussian Self-consistency test shows the DSE is non-Gaussian:

The extent to which the Gaussian chain accurately describes the ensemble of conformations that are sampled at different values of the external conditions (temperature or denaturants) can be assessed by performing a self-consistency test.



(a) (b)

Figure 4.6: (a) Comparison of R_g from direct simulations of protein L and that obtained by solving Eq. (1) using the Gaussian chain, and Worm Like Chain polymer models. The inset shows the relative errors as a function of GdmCl concentration. (b) Same as (a) except the figure is for l_p.

A property of a Gaussian chain is that if the average root mean square distance, R_{ij}, between two monomers i and j is known then R_{kl}, the distance between any other pair monomers k and l, can be computed using

$$R_{kl} = \sqrt{\frac{|k-l|}{|i-j|}} R_{ij}. \tag{4.2}$$

Thus, if the conformations of a protein (or a polymer) can be modeled as a Gaussian chain, then R_{ij} inferred from the FRET efficiency $\langle E_{ij} \rangle$ should accurately predict R_{kl} and the FRET efficiency $\langle E_{kl} \rangle$, if the dyes were to be placed at monomers k and l. We refer to this criterion as the Gaussian self-consistency (GSC) test, and the extent to which the predicted R_{kl} from Eq. 4.2 deviates from the exact R_{kl} reflects deviations from the Gaussian model description of the DSE.

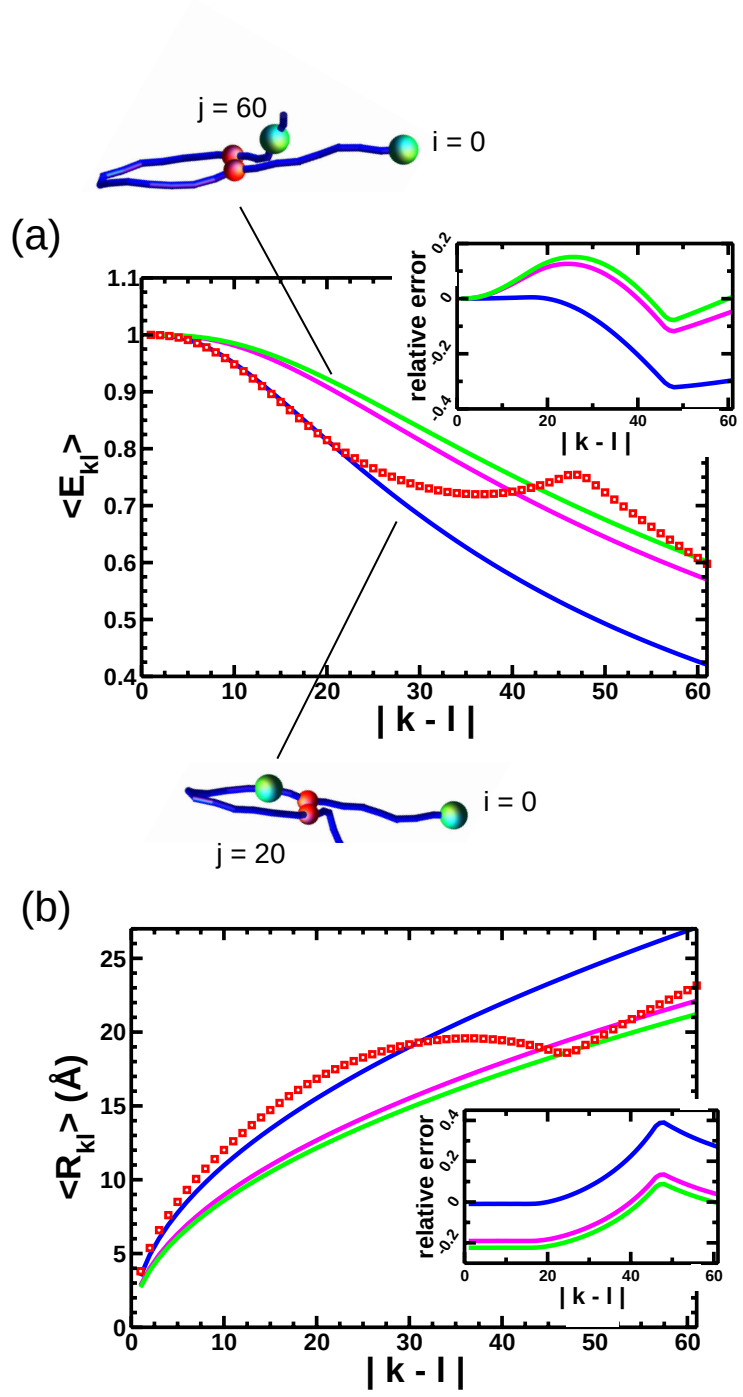
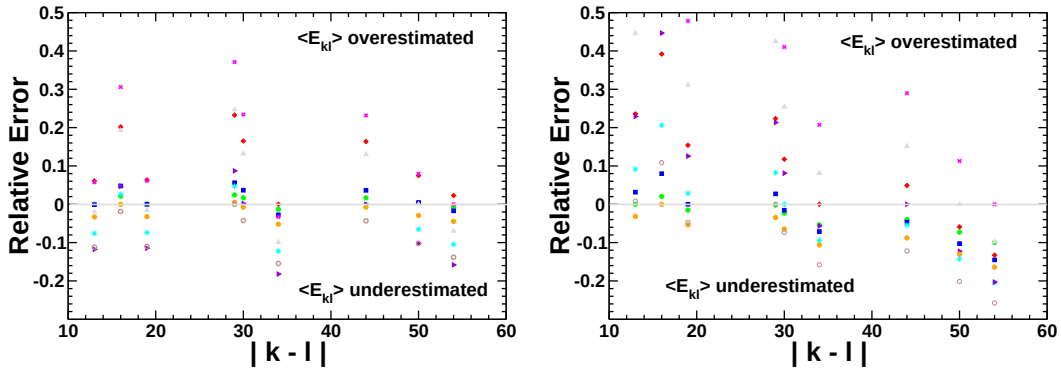


Figure 4.7: Gaussian Self-consistency test using (a) the FRET efficiency and (b) the average end-to-end distance for the GRM with $f_O = 0.75$ and interaction sites at $s_1 = 16$ and $s_2 = 47$. In both (a) and (b) the solid lines are the inferred properties and the open symbols are the exact values. In both (a) and (b), $j = 0$ and the blue, magenta, and green lines correspond to a dye at $i = 20, 40$, and 60, respectively. The insets show the relative error for $\langle E_{kl} \rangle$ and R_{kl}. Note that the relative error would be zero if the Gaussian chain accurately modeled the GRM.

4.2.3.1 GRM:

For the GRM, with a non-bonded interaction between monomers s_1 and s_2, we calculate $\langle E_{ij} \rangle$ using Eq. 4.8 with j fixed at 0 and for $i = 20, 40$, and 60. Using the exact results for $\langle E_{ij} \rangle$, the values of R_{ij} are inferred assuming that $P(r)$ is a Gaussian chain. From the inferred R_{ij} the values of $\langle E_{kl} \rangle$ and R_{kl} can be calculated using Eqs. 4.1 and 4.2, respectively. We first apply the GSC test to a GRM in which $f_O \approx 0.75$ due to a favorable interaction between monomers $s_1 = 16$ and $s_2 = 47$. There are discrepancies between the values of the Gaussian inferred (R_{kl}^G) and exact R_{kl} distances, as well as the inferred ($\langle E_{kl}^G \rangle$) and exact $\langle E_{ij} \rangle$ efficiencies when a Gaussian model is used (Fig. 4.7). The relative errors in the predicted values of the FRET efficiency and the inter-dye distances can be as large as 30-40%, depending on the choice of i and j (see insets in Fig. 4.7). The errors decrease as f_O decreases, with a maximum error of 20% when $f_O = 0.5$, and 10% when $f_O = 0.25$ (data not shown). By construction, the GRM is a Gaussian chain when $f_O = 0$ and therefore the relative errors will vanish at sufficiently small $\beta\kappa$ (Fig. 4.7 insets). These results show that even for the GRM, with only one non-bonded interaction in an otherwise Gaussian chain, its DSE cannot be accurately described using a Gaussian chain model. Thus, even if the overall end-to-end distribution $P(r)$ for the GRM is well approximated as a Gaussian (as seen in Fig. 4.1), the internal R_{kl} monomer pair distances can deviate from predictions of the Gaussian chain model.



(a) (b)

Figure 4.8: The Gaussian self consistency test applied to simulated DSE $\langle E_{ij} \rangle$ data of protein L using the (i, j) pairs listed in Fig. 4.4B. Shown are the relative errors at (a) 2.0 M GdmCl and (b) 7.5 M GdmCl. In both (a) and (b), green circles correspond to $|i-j| = 13$, orange circles to $|i-j| = 16$, blue squares to $|i-j| = 19$, brown circles to $|i-j| = 29$, cyan $*$ to $|i-j| = 30$, red diamonds to $|i-j| = 34$, violet triangles to $|i-j| = 44$, grey triangles to $|i-j| = 50$, and magenta x's to $|i-j| = 54$.

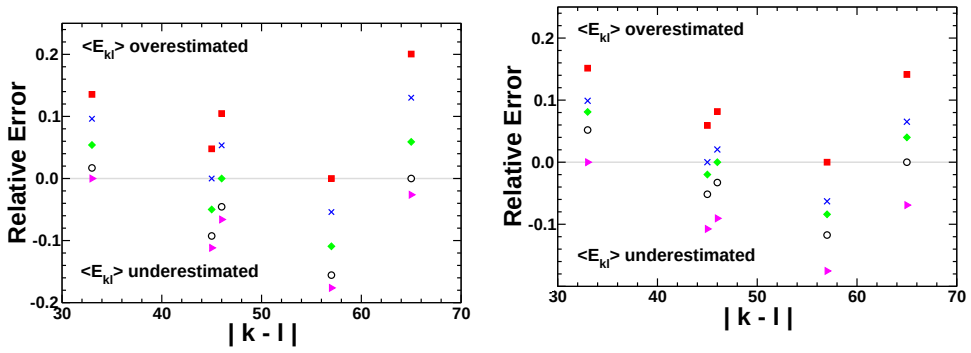
4.2.3.2 Protein L:

We apply the GSC test to our simulations of protein L at GdmCl concentrations of [C]=2.0 M (below C_m=2.4M) and [C]=7.5 M (well above C_m). While our simulations allow us to compute the DSE $\langle E_{ij} \rangle$ for all possible (i, j) pairs, we examine only a subset of $\langle E_{ij} \rangle$ as a function of GdmCl concentration (Fig. 4.4B). We use this subset of $\langle E_{ij} \rangle$ in the GSC test. The results are shown in Figs. 4.8A and 4.8B. Relative errors in $\langle E_{kl} \rangle$ as large as 36% at 2.0 M GdmCl and 50% at 7.5 M GdmCl are found. In addition, the number of data points that underestimate $\langle E_{kl} \rangle$ increases as [C] is changed from 7.5 M to 2.0 M for $|k-l| < 20$. Despite these differences, the gross features in Figs. 4.8A and 4.8B are concentration independent. Because the error does not vanish for all (k, l) pairs (Figs. 4.8A and 4.8B), we conclude that

the DSE of protein L cannot be modeled as a Gaussian chain.

4.2.3.3 The GSC test applied to experimental data:

In an interesting single molecule experiment, Schuler and coworkers have measured FRET efficiencies by attaching donor and acceptor dyes to pairs of residues at five different locations of a CspTm [56]. They analyzed the data by assuming that the DSE properties can be mimicked using a Gaussian chain model. We used the GSC test to predict $\langle E_{kl} \rangle$ for dyes separated by $|k-l|$ along the sequence using the experimentally measured values $\langle E_{ij} \rangle$.



(a) (b)

Figure 4.9: The Gaussian Self-consistency test applied to experimental data from CspTm. One dye was placed at one endpoint, and the location of the other was varied. We show relative error of the predicted $\langle E \rangle$, using Eqs. 4.1 and 4.2, versus the distance between the dyes ($|k-l|$) for [C]=2M (a) and 5M (b). In both (a) and (b), triangles correspond to $|i-j| = 33$, x's to $|i-j| = 45$, diamonds to $|i-j| = 46$, squares to $|i-j| = 57$, and circles to $|i-j| = 65$. The trends in Figs. (7) and(8) are similar.

The relative error in $\langle E_{kl} \rangle$ (Eq. 4.2) should be zero if CspTm can be accurately modeled as a Gaussian chain. However, there are significant deviations (up to 17%) between the predicted and experimental values (Fig. 4.9). The relative error is fairly

insensitive to the denaturant concentration (compare Figs. 4.9A and 4.9B). It is interesting to note that the trends in Fig. 4.9 are qualitatively similar to the relative errors in the GRM at $f_O > 0$. Based on these observations we conclude tentatively that whenever the DSE is ordered to some extent (i.e., when there is persistent residual structure) then we expect deviations from a homopolymer description of the DSE of proteins. At the very least, the GSC test should be routinely used to assess errors in the modeling of the DSE as a Gaussian chain.

4.3 Conclusions

In order to assess the accuracy of polymer models to infer the properties of the DSE of proteins from measurement of FRET efficiencies, we studied two models for which accurate calculations of all the equilibrium properties can be carried out. Introduction of a non-bonded interaction between two monomers in a Gaussian chain (the GRM) leads to an disorder-order transition as the temperature is lowered. The presence of 'residual structure' in the GRM allows us to clarify its role in the use of the Gaussian chain model to fit the accurately calculated FRET efficiency. Similarly, we have used the MTM model for protein L to calculate precisely the denaturant-dependent $\langle E \rangle$ from which we extracted the global properties of the DSE by solving Eq. 4.1 using the $P(R)$'s for the polymer models in Table I. Quantitative comparison of the exact values of a number of properties of the DSE (obtained analytically for the GRM and accurately using simulations for protein L) and the values inferred from $\langle E \rangle$ has allowed us to assess the accuracy with which polymer models can be used to analyze the experimental data. The major findings and implications of our

study are listed below.

(1) The polymer models, in conjunction with the measured $\langle E \rangle$, can accurately predict values of R_{ee}, the average end-to-end distance. However, $P(R)$, l_p, and R_g are not quantitatively reproduced. For the GRM, R_g is underestimated, whereas it is overestimated for protein L. The simulations show that the absolute value of the relative error in the inferred R_g can be nearly 25% at elevated GdmCl concentration.

(2) We propose a simple self consistency test to determine the ability of the Gaussian chain model to correctly infer the properties of the DSE of a polymer. Because the Gaussian chain depends only on a single length scale, the FRET efficiency can be predicted for varying dye positions once $\langle E \rangle$ is accurately known for one set of dye positions. The GSC test shows that neither the GRM, simulations of protein L, nor experimental data on CspTm can be accurately modeled using the Gaussian chain. The relative errors between the exact and predicted FRET efficiencies can be as high as 50%. For the GRM, we find that the variation in the FRET efficiency as a function of the dye position changes abruptly if one dye is placed near an interacting monomer. Taken together these findings suggest that it is possible to infer the structured regions in the DSE by systematically varying the location of the dyes.

(3) The properties of the DSE inferred from Eq. 4.1 become increasingly more accurate as [C] decreases. At a first glance this finding may be surprising, especially considering that stabilizing intra-peptide interactions are expected to be weakened at high GdmCl concentrations [C], and therefore the protein should be more "polymer-like." The range of R-values sampled at low [C] is much smaller than at high [C]. Protein L swells as [C] is increased, as a consequence of the increase in the solvent

quality. It is possible that [C]≈2.4 M might be close to a Θ-solvent (favorable intrapeptide and solvent-peptide interactions are almost neutralized), so that $P(R)$ can be approximated by a polymer model. The inaccuracy of polymer models in describing $P(R)$ at [C]=6 M suggests that only at much higher concentrations does protein L behave as a random coil. In other words, T=327.8 K and [C]=6 M is not an athermal (good) solvent.

(4) It is somewhat surprising that polymer models, which do not have side chains or any preferred interactions between the beads, are qualitatively correct in characterizing the DSE of proteins with complex intramolecular interactions. In addition, even [C]=6 M GdmCl is not an athermal solvent, suggesting that at lower [C] values the aqueous denaturant may be closer to a Θ-solvent. A consequence of this observation is that, for many globular proteins, the extent of collapse may not be significant, resulting in the nearness of the concentrations at which collapse and folding transitions occur, as shown by Camacho and Thirumalai [120] some time ago. We suggest that only by exploring the changes in the conformations of polypeptide chains over a wide range of temperature and denaturant concentrations can one link the variations of the DSE properties (compaction) and folding (acquisition of a specific structure).

4.4 Theory and computational methods

4.4.1 GRM model:

In order to understand the effect of a single non-covalent interaction between two monomers along a chain, we consider a Gaussian chain with Kuhn length a_0 and

N bonds, with a harmonic attraction between monomers $s_1 \leq s_2$, which is cutoff at a distance c. The Hamiltonian for the GRM is

$$\beta H = \frac{3}{2a^2}\int_0^N ds\ \dot{\mathbf{r}}^2(s) + \beta V[\mathbf{r}(s_2) - \mathbf{r}(s_1)] \tag{4.3}$$

$$\beta V[\mathbf{r}] = \begin{cases} k\mathbf{r}^2/2 & |\mathbf{r}| < c \\ \\ kc^2/2 & |\mathbf{r}| \geq c \end{cases}, \tag{4.4}$$

where k is the spring constant that constrains $\mathbf{r}(s_2) - \mathbf{r}(s_1)$ to a harmonic well. The Hamiltonian in Eq. 4.3 allows the exact determination of many quantities of interest. Defining $\mathbf{x} = \mathbf{r}(s_2) - \mathbf{r}(s_1)$ and $\Delta s = s_2 - s_1$, we can determine most averages of interest for the GRM using

$$\langle \cdots \rangle = \frac{\int d^3\mathbf{r}_1 d^3\mathbf{x} d^3\mathbf{r}_N (\cdots) G(\mathbf{x}, \mathbf{r}_N; \Delta s, N)}{\int d^3\mathbf{r}_1 d^3\mathbf{x} d^3\mathbf{r}_N\ G(\mathbf{x}, \mathbf{r}_N; \Delta s, N)} \tag{4.5}$$

$$G(\mathbf{x}, \mathbf{r}_N; \Delta s, N) = \exp\Big(- \frac{3\mathbf{x}^2}{2\Delta s\, a^2} - \frac{3(\mathbf{r}_N - \mathbf{x})^2}{2(N - \Delta s)a^2} - \beta V[\mathbf{x}] \Big). \tag{4.6}$$

4.4.2 C_α-SCM protein model and GdmCl denaturation:

We use the coarse-grained C_α-side chain model (C_α-SCM) to model protein L (for details see the supporting information in [95]). In the C_α-SCM each residue in the polypeptide chain is represented using two interaction sites, one that is centered on the α-carbon atom and another that is located at the center-of-mass of the side chain [59]. Langevin dynamics simulations [105] are carried out in the underdamped limit at zero molar guanidinium chloride. Simulation details are given in [95].

We model the denaturation of protein L by GdmCl using the molecular transfer model (MTM) [95]. MTM combines simulations at zero molar GdmCl with experimentally measured transfer free energies, using a reweighting method [40, 37, 76]

to predict the equilibrium properties of proteins at any GdmCl concentration of interest.

4.4.3 Analysis:

4.4.3.1 GRM:

The average squared end-to-end distance can be computed directly from Eq. 4.5, using $\langle \mathbf{R}_{ee}^2 \rangle = N a_0^2 + (\langle \mathbf{x}^2 \rangle - \Delta s \, a_0^2)$. The exact expression for $\langle \mathbf{x}^2 \rangle$ is easily determined, but somewhat lengthy, and we omit the explicit result here. Also of interest is the end-to-end distribution function, $P(\mathbf{R}) = \langle \delta[\mathbf{r}_N - \mathbf{R}] \rangle$, which can be obtained from Eq. 4.5. In order to determine the probability of an interior bond being in the 'ordered' state (i.e. the fraction of residual structures, see the inset for Fig. 4.1a), we compute the interior distribution, $P_I(\mathbf{X}) = \langle \delta[\mathbf{x} - \mathbf{X}] \rangle$, so that $f_O = \int_{|\mathbf{x}| \leq c} d^3\mathbf{x} \, P_I(\mathbf{x})$. The radius of gyration requires a more complicated integral than the one found in Eq. 4.5, but we find

$$R_g^2 = \frac{N a_0^2}{6} + (\langle \mathbf{x}^2 \rangle - \Delta s \, a_0^2) \Big[\frac{\Delta s}{3N} + \frac{s_1}{N} - \Big(\frac{\Delta s}{2N} + \frac{s_1}{N} \Big)^2 \Big] \tag{4.7}$$

Note that, unlike the average end-to-end distance, the radius of gyration depends not only on Δs, but also on s_1.

The FRET efficiency for a system with dyes attached to $\mathbf{r}(j = 0) = \mathbf{0}$ and $\mathbf{r}(i)$, $\langle E \rangle = \langle [1 + (|\mathbf{r}(i)|/R_0)^6]^{-1} \rangle$, is determined from Eq. 4.5 as

$$E(i) = \begin{cases} E^G(i) & 0 \leq i \leq s_1 \\ \frac{\int_0^\infty dx dr \, g_1(x,r;\{s_i\})/[1+(r/R_0)^6]}{\int_0^\infty dx dr \, g_1(x,r;\{s_i\})} & s_1 < i < s_2 \\ \frac{\int_0^\infty dx dr \, g_2(x,r;\{s_i\})/[1+(r/R_0)^6]}{\int_0^\infty dx dr \, g_2(x,r;\{s_i\})} & s_2 \leq i \leq N \end{cases} \tag{4.8}$$

where $E^G(i)$ is the FRET efficiency for a Gaussian chain with i bonds, and

$$g_1(x,r;\{s_i\}) = xr\ \sinh\left(\frac{3(i-s_1)xr}{\lambda a_0^2}\right)e^{-3(ix^2+\Delta sr^2)/2\lambda a_0^2-\beta V[x]} \quad (4.9)$$

$$g_2(x,r;\{s_i\}) = xr\ \sinh\left(\frac{3xr}{(i-\Delta s)a_0^2}\right)e^{-3x^2/2\Delta sa_0^2-3(x^2+r^2)/2(i-\Delta s)a_0^2-\beta V[x]} \quad (4.10)$$

$$\lambda = (s_2+s_1)i - s_1^2 - i^2 \quad (4.11)$$

This result allows us to compute the Gaussian Self-consistency test, after a numerical integral over r.

4.4.3.2 Protein L simulations:

Averages and distributions were computed using the MTM [95] which combines experimentally measured transfer free energies [28], converged simulations and the WHAM equations [40, 37, 76]. The WHAM equations use the simulation time-series of potential energy and the property of interest at various temperatures and gives a best estimate of the averages and distributions of that property. The native state ensemble (NSE) and DSE subpopulations were defined as having a structural RMSD (root mean squared deviation), after least squares minimization, of less than or greater than 5 Å relative to the crystal structure for the NSE and DSE respectively. The exact values of l_p are computed using the average R from simulations and the relationships listed in Table 4.1.

4.4.3.3 Notation:

Throughout this chapter, exact values of all quantities are reported without superscript or subscript. For the GRM, exact values are analytically obtained or calculated by performing a one-dimensional integral numerically. For convenience,

exact results for protein L refer to converged simulations. While these simulations have residual errors, the simplicity of the MTM has allowed us to calculate all properties of interest with arbitrary accuracy. The use of subscript or superscript is, unless otherwise stated, reserved for quantities that are extracted by solving Eq. 4.1 using the polymer models listed in Table 4.1.

Table 4.1: Polymer models and their properties

	Property		
Polymer Model	End-to-end distribution $P(R)$[a]	Radius of gyration R_g	Persistence length l_p
Gaussian	$4\pi R^2 \left(\frac{3}{2\pi N a^2}\right)^{3/2} \exp\left(\frac{-3R^2}{2Na^2}\right)$	$a\sqrt{N/6}$	$\frac{Na^2}{2L} = \frac{a}{2}$
Worm-like Chain[b]	$\frac{4\pi R^2 C_1}{L(1-(R/L)^2)^{9/2}} \exp\left(\frac{-3L}{4l_p(1-(R/L)^2)}\right)$	$\frac{L}{6C_2} + \frac{1}{4C_2^2} + \frac{1}{4LC_2^3} - \frac{1-\exp(-L/l_p)}{8C_2^4 L^2}$	$R_{ee}^2 = 2l_pL - 2l_p^2 - 2l_p^2 \exp(-\frac{L}{l_p})$[c]
Self Avoiding Polymer[d]	$\frac{a}{R_{ee}}(\frac{R}{R_{ee}})^{2+\theta} \exp(-b\left(\frac{R}{R_{ee}}\right)^{\delta})$	N/A	N/A

[a] The average end-to-end distance $R_{ee} = \left(\int R^2 P(R) dR\right)^{1/2}$

[b] L and l_p are the contour length and persistence length respectively. $C_1 = (\pi^{3/2} e^{-\alpha} \alpha^{-3/2}(1+3\alpha^{-1}+\frac{15}{4}\alpha^{-2}))^{-1}$ where $\alpha = 3L/(4l_p)$. $C_2 = 1/(2l_p)$.

[c] Using the simulated $\langle R^2 \rangle$, l_p was solved for numerically using this equation.

[d] θ and δ equal 0.3 and 2.5, respectively. The constants a and b are determined by solving the integrals of the zeroth and second moment of $\int P(R)dr = \int R^2 P(R)dr = 1$, resulting in values of $a = 3.67853$ and $b = 1.23152$.

Chapter 5
Thermodynamic basis of the dock-lock growth mechanism of amyloid fibrils
5.1 Introduction

Proteins and peptides, that are unrelated by sequence or structure, form morphologically similar fibrillar structures upon aggregation [121]. The emergence of a global cross β-structure, that is the characteristic of all fibril forming proteins including those that are associated with distinct strains, suggests that their growth processes must be similar. Experiments on Aβ amyloid forming protein [122] have found that the process of monomer addition to an elongating amyloid fibril (Fig. 5.1) is kinetically complex, and can be approximately described using two distinct timescales [123, 124]. Based on early kinetic experiments, Lee and Maggio [123] envisioned that the growth of fibrils occurred by a sequential process involving two distinct steps that were pictorially described as the dock-lock growth mechanism. On a relatively fast timescale a monomer reversibly binds (or docks) to the fibril surface. A second slower timescale is associated with the lock process, which presumably involves structural rearrangements within the monomer leading to a greater binding affinity for the fibril [123, 124]. Upon completion of the lock process the monomer adopts the β-strand conformation that is commensurate with the underlying fibril structure.

Such a pictorial description is simplistic because the locked phase has numerous intermediate species [124]. The plausibility for a dock-lock mechanism of

fibril growth comes solely from bulk experiments that suggest that the kinetics of monomer dissociation from a fibril can be fit by a sum of two or three exponentials [123, 124]. In addition, there is little direct evidence for the structural rearrangements that are hypothesized to occur within the fibril-bound monomer in the docked to locked transition [123, 124, 125]. Measuring such conformational changes is hampered by the inherently low concentration of fibril-bound monomer in the docked phase, and the length scale of the structural rearrangements involved in the dock-lock transition [123, 124].

Molecular simulations are ideally suited for providing structures, energies, and dynamics of the process of monomer addition to a fibril [126, 127, 128, 129, 130, 131]. For example, in several previous computational studies we investigated many aspects of the early events of amyloid formation, including monomer addition to preformed structured oligomers [128], and the effect of urea on these species [126]. Results from these studies suggest that even oligomer growth can be described by a dock-lock mechanism. More recently, we have shown using lattice models that fibril growth occurs by a dock-lock process [132].

Here, we use simulations to provide a thermodynamic basis for the global dock-lock mechanisms of fibril growth. Although growth is an inherently kinetic process the clear separation in the time scales between the dock and lock process allows us to examine free energy and structural changes as the monomer interacts with the template fibril surface. In order to illustrate the thermodynamics of the addition of a monomer to an elongating fibril we consider a disordered Aβ peptide that is added to a preformed fibril surface. The availability of molecular structures [133]

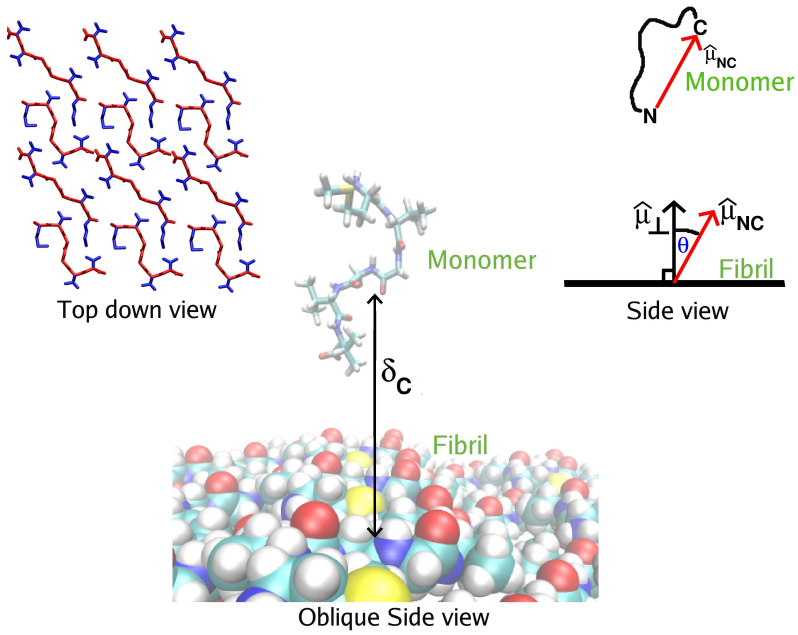


Figure 5.1: Monomer addition to an amyloid fibril. 'Top down view' shows the peptides in the fibril surface from above. The peptides are displayed as sticks with backbone atoms in red and side chain atoms in blue. 'Oblique side view' shows the fibril surface in a van der Walls representation while the unincorporated monomer is shown in a stick representation. 'Side view' offers a simple geometric perspective of the fibril surface and monomer from the side to illustrate the calculation of the θ-angle (see Methods). The vector normal to the fibril surface is shown as a black arrow, while the monomers N-to-C termini vector is shown as a red arrow. θ is the angle formed by these two vectors. The $cos(\theta)$ term used in Fig. 5.4B is equal to $\hat{\mu}_{NC} \cdot \hat{\mu}_{\perp}/(|\hat{\mu}_{NC}||\hat{\mu}_{\perp}|)$.

enables us to monitor the energetic and structural changes in the monomer as it attaches to the fibril. Using Multiplexed Hamiltonian Replica Exchange (MhREX) simulations [36, 134], we sample the reversible association/dissociation of a peptide (MVGGVV) from the Aβ protein to a fibril whose structure has recently been determined at atomic resolution [133]. The use of an implicit solvent model and enhanced sampling methods allows us to fully characterize the thermodynamics of the process under a variety of solution conditions.

Our simulations reveal a number of novel features of the thermodynamics

of amyloid growth. We find that the dock-lock mechanism is manifested as three basins in a free energy profile, which monitors the reversible work related to bringing a monomer to the fibril surface. The three basins correspond to two substate basins of the docked monomer and a locked phase in which the monomer adopts an extended anti-parallel conformation with modest β-strand content. The dock $\rightarrow$ lock transition is a disorder to order transition that involves an increase in the end-to-end distance of the monomer that is driven by the favorable peptide-fibril interactions. The free energy barrier separating the docked and locked phases arises largely from the loss of favorable intra-peptide interactions of the monomer that is deposited onto the surface.

To further shed light on the energetics governing monomer addition we have used simulations probe the influence of cosolvents (urea and Trimethylamine N-oxide (TMAO)) and molecular crowders on the free energy profiles. A modest concentration (0.75 M) of urea or TMAO stabilizes the locked phase, while molecular crowding only marginally stabilizes the docked phase. A measure of the free energy of stability of the fibril structure is the critical monomer concentration, C_R, which is the concentration of soluble monomer that is in equilibrium with the amyloid fibril. We show that C_R is strongly temperature dependent, and weak cosolvent dependence. Our study provides a conceptual framework for interpreting the thermodynamics of fibril elongation.

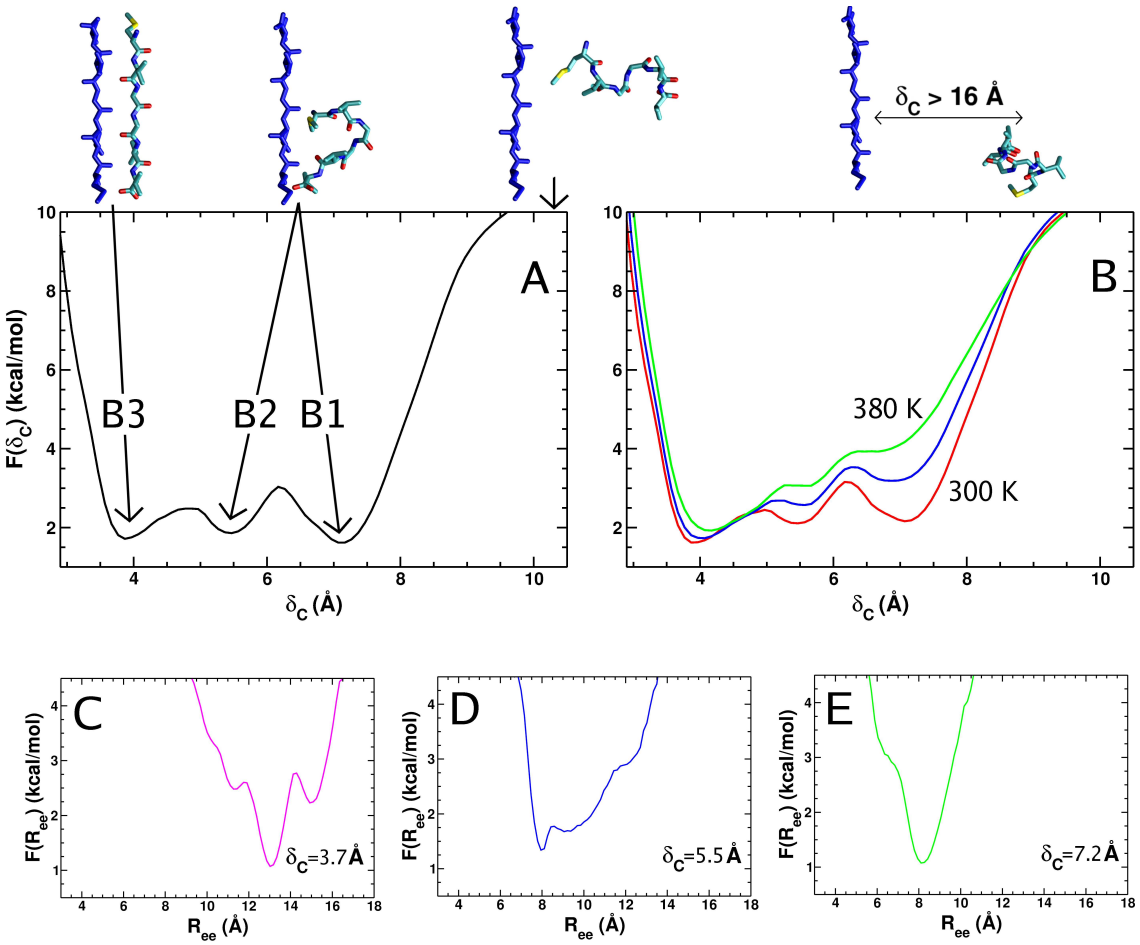


Figure 5.2: The free energy profile ($F(\delta_C) = -k_B T ln[Z(\delta_C)/Z]$) of monomer addition as a function of δ_C. (A) The temperature is 280 K. (B) The curves correspond are at temperatures of 300 K (red line), 340 K (blue line), and 380 K (green lines). Representative structures in the free energy basins B1, B2 and B3 labeled in (A) are shown. In addition, two monomer-fibril configurations that have $\delta_C > 10$ Å are also shown. A peptide in the fibril surface is shown in blue, while the docking monomer is displayed in non-blue colors. The free energy profile as a function of the monomer end-to-end distance at a specified δ_C ($F(R_{ee}|\delta_C) = -k_B T ln[Z(R_{ee}|\delta_C)/Z(\delta_C)]$) is shown for δ_C = 3.7, 5.5, and 7.2 Å (i.e. for basins B1, B2 and B3 in (A)) in (C), (D), and (E) respectively.

5.2 Results and Discussion

5.2.1 The PMF of monomer addition to the fibril surface has multiple basins of attraction:

The PMF, $F(\delta_C)$, that gives the reversible work required to bring the monomer to a distance δ_C above the fibril surface (Fig. 5.1), shows multiple basins of attraction as T is changed from 280 K to 380 K (Fig. 5.2). There are three distinct basins at temperatures below 340 K (Figs. 5.2A and 5.2B). The minimum in the first basin (B_1 in Fig. 5.2A) is at $\delta_C = 7.1$ Å, and the other two basins (B_2 and B_3) are at 5.5 and 3.9 Å, respectively. At 280 K the free energy barrier separating B_1 and B_2 is ∼1.2 kcal/mol. At higher temperatures the barriers decrease, and at 380 K there is virtually no free energy barrier separating the basins. When $\delta_C < 3.9$ Å the PMF increases due to the unfavorable steric interactions between the monomer and the fibril surface. The PMF also increases sharply at $\delta_C > 9$ Å, where there are very few contacts between the monomer and the fibril.

In order to associate the features in the PMF with the dock-lock picture it is necessary to examine the structural transitions that occur as δ_C changes. If the basins observed in the $F(\delta_C)$ profile correspond to the docked and locked phases, we expect structural changes in the monomer when δ_C decreases from 7.1 Å to 3.9 Å. The fibril-bound monomer undergoes a global expansion, with an increase in R_g from 4.6 Å→ 5.7 Å, as it goes from B_1 to B_3 (Fig. 5.3). The end-to-end distance also dramatically increases from 8 Å to 13 Å (Fig. 5.3B). It is significant that the maxima in the derivatives of $\frac{dR_g}{d\delta_C}$ and $\frac{dR_{ee}}{d\delta_C}$, in the range of $9 > \delta_C > 3.9$ Å, occur at $\delta_C = 6.2$ and 4.7 Å (computed using Fig. 5.3, data not shown). The positions

of these maxima coincide with the locations of the free energy barriers in the PMF (Fig. 5.2A), which suggests that the barriers arise during the process of expansion of the monomer as it interacts with the fibril surface.

We examine the free energy profile of R_{ee} at fixed δ_C values ($F(R_{ee}|\delta_C)$) in Figs. 5.2C, 5.2D, and 5.2E. These figures show that additional complexity (basins) is present in the free energy surface that is not seen when projected on to the one-dimensional order parameter δ_C. For example, at δ_C= 3.7 Å and δ_C= 5.5 Å (basins B_1 and B_2) $F(R_{ee}|\delta_C)$ exhibits two or more basins. Thus, there are numerous metastable states in the dock-lock process. Based on the global structural changes in R_g and R_{ee}, we tentatively designate the monomer as unbound if $\delta_C > 9$ Å, docked if the monomer is in the range of $9 > \delta_C > 5$ Å, and locked when $\delta_C < 5$ Å.

5.2.2 Free energy landscape during the growth process:

Surprisingly, an additional structural transformation in the monomer, that is not evident in $F(\delta_C)$, is suggested by the $R_g(\delta_C)$ and $R_{ee}(\delta_C)$ profiles. In the range of $16 > \delta_C > 9$ Å the monomer is 'stretched', with R_g and R_{ee} values close to that found in an extended β-strand (Fig. 5.3). Examination of the backbone-backbone contacts that occur between individuals residues of the monomer and the fibril (Fig. 5.4) shows that the N-terminal methionine residue contacts the fibril surface when δ_C is between 12 and 14 Å. The favorable interaction of the N-terminal residue with the fibril surface leads to the chain expansion observed in $R_g(\delta_C)$ and $R_{ee}(\delta_C)$ when $\delta_C \sim 12$ Å (Fig. 5.3).

To examine the global orientation of the monomer, as it interacts with the

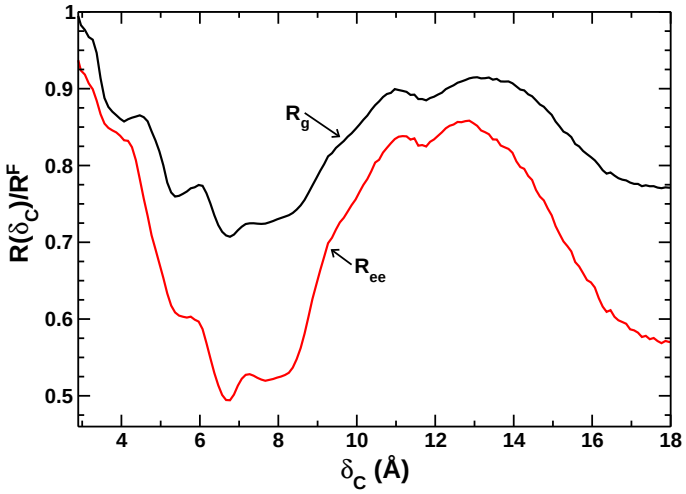
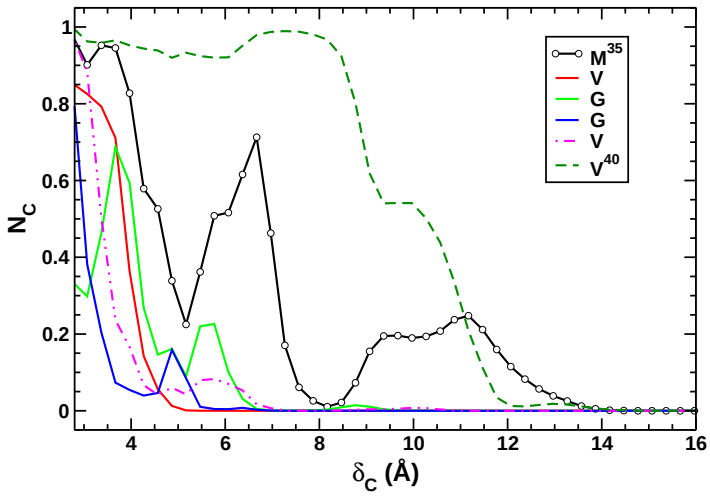
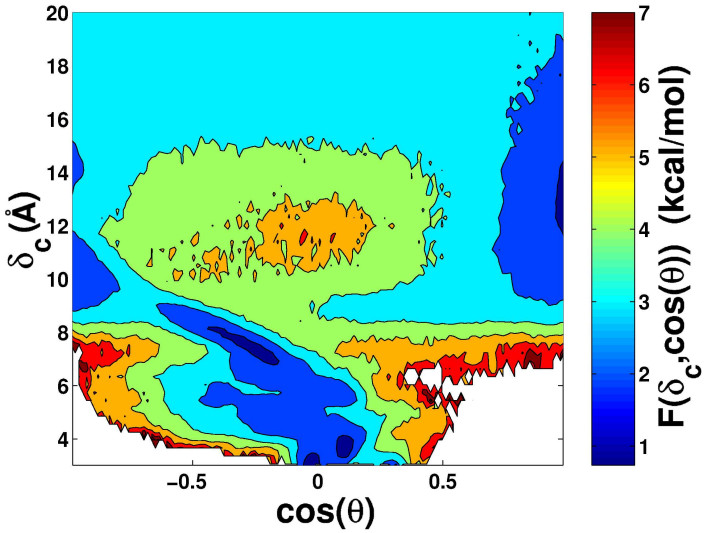


Figure 5.3: The radius-of-gyration (R_g) of the monomer, scaled by its average value of 7.2 Å in the fibril surface (R_g^F), as a function of δ_C in bulk is shown in black. The red curve shows R_{ee} of the monomer, scaled by its average value of 15.4 Å in the fibril surface (R_{ee}^F), as a function of δ_C. The temperature is 300 K.

fibril surface, we show, in Fig. 5.4B, the free energy surface ($F(\delta_C, cos(\theta))$) as a function of δ_C and $cos(\theta)$. θ is the angle formed between a vector normal to the fibril surface and the N to C-termini vector of the monomer (see Fig. 5.1). When $cos(\theta) = -1$ (1) the monomer is oriented towards (away from) the surface (see Fig. 5.1). A value of $cos(\theta) = 0$ implies that the monomer is parallel to the fibril surface. At the farthest distances from the fibril ($\delta_C > 19$ Å), the orientation of the monomer is randomly distributed (Fig. 5.4B) as indicated by the lack of a dominant free energy basin in $F(\delta_C, cos(\theta))$. However, at $\delta \approx 12$ Å and $cos(\theta) = 1$ there is a basin, indicating that the monomer's N to C-termini vector is pointing away from the fibril surface (Fig. 5.4B). Thus, the N-terminus is closest to the fibril surface in the 'stretched' state. As δ_C decreases to 4 Å, basins with values of $cos(\theta) \approx 0$ are favored, which shows that the monomer is aligned parallel to the fibril surface.

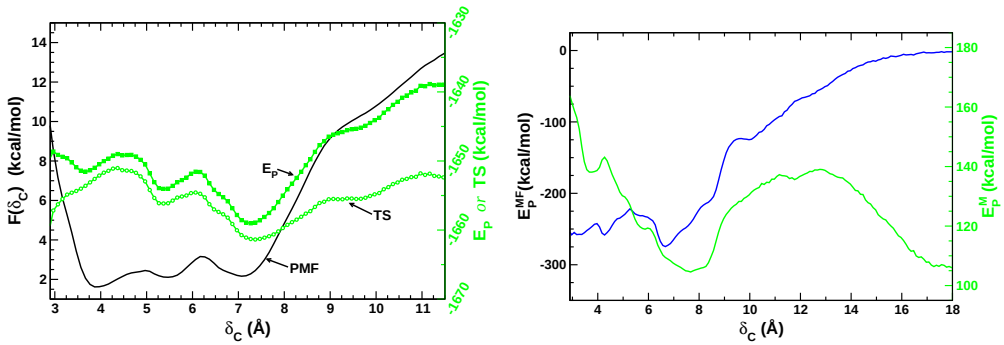


(a)



(b)

Figure 5.4: (A) The number of anti-parallel in-register backbone-backbone contacts between the monomer and fibril as a function of δ_C. The symbols for the various residues starting from the N-terminal methionine are shown in the legend. (B) The free energy surface ($F(\delta_C, cos(\theta)) = -k_B T ln[Z(\delta_C, \theta)/Z(\delta_C)]$) as a function of δ_C and $cos(\theta)$ at 340. θ is the angle formed by a vector normal to the plane of the fibril surface and the N to C terminal vector of the monomer (see Fig. 5.1).



(a) (b)

Figure 5.5: (A) Deconvolution of $F(\delta_C)$ into entropic ($TS(\delta_C)$) and energetic ($E_P(\delta_C)$) components as a function of δ_C at 300 K. The δ_C-profile of each term is indicated on the graph. (B) The interaction energy between the monomer and the fibril (E_P^{MF}, blue lines), and monomer's intrapeptide interaction (E_P^M) as a function of δ_C at 300 K. Bulk ($\phi_C = 0.00$) and crowded ($\phi_C = 0.14$) conditions are shown as solid and dashed lines respectively.

5.2.3 Monomer deposition to the fibril surface results in multiple structural transitions:

We characterize the structural changes that the monomer undergoes while interacting with the fibril using the number of peptide-fibril contacts, and the number of in-register peptide-fibril backbone contacts. In the range of $16 > \delta_C > 9$ Å, the monomer makes a few contacts with the fibril surface (data not shown). Several non-specific peptide-fibril contacts are made that are energetically favorable (Fig. 5.5B). Because δ_C is large (relative to R_g) the peptide must extend to make contact with the fibril, as evidenced by the increase in R_g and R_{ee} (Fig. 5.3). In the extended conformations there is a significant decrease in the favorable intra-peptide interactions (Fig. 5.5B). Upon reducing δ_C in the range of $9 > \delta_C > 6.2$ Å, the monomer docks onto the fibril surface (Fig. 5.2). In the process, the monomer

undergoes a dramatic reduction in R_g from a maximum of 5.7 Å, when $\delta_C > 9$ Å, to 4.6 Å (Fig. 5.3). The reduction in R_g is accompanied by an increase in favorable enthalpic interactions both within the monomer and between the monomer and the fibril (Fig. 5.5). When hopping between basins B_1 and B_2 in the docked phase, the monomer undergoes only small structural rearrangements, as measured by $R_g(\delta_C)$ and $R_{ee}(\delta_C)$ (Fig. 5.3). There are fewer in-register backbone contacts in the docked phase as compared to the locked phase (Figs. 5.4A and 5.4B).

The monomer undergoes a large scale structural rearrangement as it locks onto to the fibril surface ($5.0 > \delta_C > 3.0$ Å). In addition to an increase in R_g (Fig. 5.3), favorable intra-peptide interactions are lost (Fig. 5.5B), and are replaced by peptide-fibril contacts and interactions (Figs. 5.4A and 5.5B). The monomer forms antiparallel in-register backbone contacts (Fig. 5.4A) in agreement with the monomer orientation in the crystal structure [133]. In Fig. 5.2 we show monomer-fibril configurations corresponding to the unbound, docked and locked phases. Note that the 'stretched' conformation shown in Fig. 5.2 correlates with the expanded R_g in Fig. 5.3.

We analyze the secondary structural content as a function of δ_C using the STRIDE program [135]. For $\delta_C > 9$ Å the monomer is unstructured and is dominated by random coil ($> 60\%$) with moderate turn content ($< 40\%$). In the docked phase, turn content dominates ($\approx$60%) and the coil content drops to $\approx$35%. In the locked phase the peptide is predominantly a random coil ($\approx$80%), turn content is around 10%, and β-bridge content around 10%. Thus, in contrast to the structure of a peptide in the fibril crystal structure the β-strand content in the simulated

monomer is small even after locking is complete. There are two possible reasons why the β-strand is not stable in the locked phase in our simulations. First, the width of fibrils is finite consisting of just a few β-sheets whereas in our simulations the fibril surface is essentially infinite. As a result, a single monomer can bind to multiple sites on the surface leading to an increase in the binding entropy that can compensate for the energy gain that arises from forming an in-register β-sheet with another monomer in the fibril. As a result the free energy of the added monomer can be minimized by making multiple out-of-register backbone contacts with different strands in the fibril - leading to small β-strand content . The second possibility is that the GBSW implicit solvent model is inaccurate. Nevertheless, we show below the critical concentration calculated from these simulations exhibit realistic changes with solution conditions, which suggests that the present simulations capture qualitatively the complexity of the dock-lock mechanism.

5.2.4 The free energy barrier separating the docked from locked phases is largely enthalpic:

To determine the origin of the free energy barriers separating the docked and locked phases (Fig. 5.2) we compute the potential energy (E_P) and entropic (TS) contributions to $F(\delta_C)$. The profiles of $E_P(\delta_C)$ and $TS(\delta_C)$ (Fig. 5.5A) have maxima at the same locations as the basins in $F(\delta_C)$ (Fig. 5.2). At $\delta_C \approx 6$ Å, the maximum in E_P is greater than in TS, indicating that potential energy gives rise to the free energy barriers separating the docked and locked phases in $F(\delta_C)$. Interestingly, the monomer gains entropy upon reaching the top of the barrier from the docked phase in

$F(\delta_C)$ (Fig. 5.5A). However, the monomer loses entropy upon locking on to the fibril surface as indicated by $T\Delta S_{D\rightarrow L} \equiv T(S_L - S_D)$, where $S_i = \sum_{\delta_{i,l}}^{\delta_{i,u}} S(\delta_i) e^{-\beta F(\delta_i)}$ and $\delta_{i,u}$ and $\delta_{i,l}$ correspond to the upper (u) and lower (l) bounds in δ_C that separate the docked and locked basins in the $F(\delta_C)$ profile. At 300 K, $T\Delta S_{D\rightarrow L} = -2.6$ kcal/mol, and at 380 K $T\Delta S_{D\rightarrow L}$ =-6.7 kcal/mol.

To determine the molecular origin of the barriers in $F(\delta_C)$ we deconvolute the $E_P(\delta_C)$ profile into contributions from the monomer internal energy ($E_M(\delta_C)$), that is the interaction energy of the monomer with itself, and the monomer-fibril interaction energy ($E_{MF}(\delta_C)$). These profiles (Fig. 5.5B) clearly show that the docked phase is energetically stabilized by internal monomer interactions and monomer-fibril interactions, while in the locked phase favorable internal monomer interactions are lost and replaced by monomer-fibril interactions. Consequently, it is the interplay of these two energies as the monomer undergoes conformational changes that contributes to the potential energy barrier separating the docked and locked phases.

5.2.5 Urea and TMAO stabilize the fibril-bound monomer:

The cellular environment, besides containing large biomolecules, also contains small organic molecules known as osmolytes that can dramatically effect protein function [14], stability [14, 136], and amyloid formation [137, 138, 139]. Naturally occurring osmolytes, such as TMAO and urea, can be found in a variety of organisms at concentrations from 0 to 6 M [14, 140]. Therefore, to carry out simulations at physiologically relevant concentrations, we simulate the process of monomer addition in aqueous urea and TMAO solution at 0.75 M using a coarse grained model for

urea (Eq. 5.3).

Urea and TMAO increase the stability of the locked phase to a much greater extent relative to bulk. In contrast, the unbound and docked states are destabilized. The force-field employed here shows that urea stabilizes the locked phase to a lesser extent than TMAO. This is due to the stronger interaction of urea (see Methods section) with the peptide and the fibril. For example, R_g of the unbound monomer is greater in urea than in TMAO (Fig. 5.3B), due to the stronger attraction between the urea molecules and the peptide. The greater affinity is reflected in the radial distribution function (RDF) between the cosolutes and the peptide groups (O=C-N-H) of the protein backbone. The first peak in the RDF, located at $r = 7.5$ Å, indicates that urea preferentially interacts with the backbone and binds more strongly than TMAO (data not shown).

To contrast the effect of specific interactions between urea and TMAO on the stability of the docked and locked states we also carried out simulations to probe the influence of small crowding particles on their stabilities. The interaction between the crowding particle and the peptide atom is mimicked using Eq. 5.3 with $\epsilon_{ij} = \epsilon = 0.1$ kcal/mol and $\lambda = 0$. At a crowder volume fraction of $\phi_C = 0.14$ (concentration of 0.75 M), the PMF indicates that inert-crowding particles only marginally stabilize the docked phase and destabilize the locked and unbound phases.

5.2.6 The effect of TMAO and urea on the critical concentration C_R:

When amyloid fibrils reach equilibrium (when the rate of monomer addition to the fibril equals the rate of dissociation) some number of monomers remain unbound

in solution. The equilibrium concentration of the soluble unbound monomers is the critical concentration, C_R. One can relate C_R to the equilibrium constant of dissociation of a monomer from the fibril [141]. Wetzel and coworkers have used this observation to map the regions in $A\beta_{1-40}$ that harbor amyloidogenic tendencies [141, 125]. The free energy profiles computed in our simulations allow us to calculate the relative changes in C_R as the cosolvent concentration or temperature is varied. From the density of monomer a distance δ_C from the fibril surface, C_R can be calculated using

$$C_R = C \int_{\delta_M}^{\delta_U} e^{-\beta F(\delta_C)} d\delta_C, \tag{5.1}$$

where C is the bulk density of peptide in solution, $\beta = 1/k_B T$, where k_B is Boltzmann's constant, T is the temperature of solution condition, and $F(\delta_C)$ is the PMF. We assume that the monomer is unbound if $\delta_C > \delta_M = 9$ Å. We took $\delta_U = 22$ Å.

The relative change in C_R, upon a change in solution conditions (altering cosolute concentration of temperature) can be computed without determining C in Eq. 5.1, using

$$R = \frac{C_{R,j}}{C_{R,i}} = \frac{\int_{\delta_M}^{\delta_U} e^{-\beta_j F_j(\delta_C)} d\delta_C}{\int_{\delta_M}^{\delta_U} e^{-\beta_i F_i(\delta_C)} d\delta_C}, \tag{5.2}$$

where $C_{R,j}$ ($C_{R,i}$) is the value of C_R in solution condition j (i) and F_j (F_i) is the corresponding free energy profile.

At 300 K we find that when crowder, TMAO or urea is added to solution $R(= C_R(cosolute)/C_R(Bulk))$ equals 0.3, 0.5, and 0.2, respectively. Thus, addition of the cosolutes decreases C_R, which is in accord with the stabilization observed of

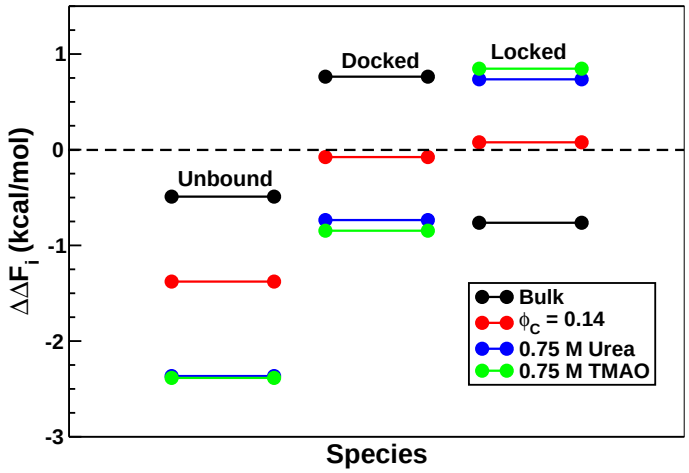


Figure 5.6: The impact of a temperature change (T_1 = 300 K$\rightarrow$ T_2 = 380 K) on the relative free energy ($\Delta\Delta F_i = -k_B T_2 ln[Z_i(T_2)/Z(T_2)] + k_B T_1 ln[Z_i(T_1)/Z(T_1)]$) of the unbound, docked and lock species in bulk, crowder and osmolyte solutions as indicted in the legend.

the fibril-bound monomer. An increase in temperature from 300 K to 380 K leads to $R(= C_R(380K)/C_R(300K))$ values of 38.5, 160, 638 and 526 in bulk, crowder, TMAO and urea solutions respectively. Thus, increasing temperature increases C_R under all solution conditions. Interestingly, the locked phase can still be stabilized despite the increase in C_R. For example, in bulk solution we find that increasing the temperature from 300 to 380 K results in a stabilization of the unbound and locked phases by 0.4 and 0.6 kcal/mol respectively (Fig. 5.6). The docked phase on the other hand is destabilized by $\approx$ 0.9 kcal/mol. This result is important because it illustrates that C_R only measures the equilibrium constant of monomer association and cannot measure the equilibrium constants of the monomer in the docked and locked phases. Thus, increases in C_R do not always indicate destabilization of the locked phase. These predictions are amenable to experimental tests.

5.3 Conclusions

By exploiting the large separation in the time scales of the two major events (dock and lock) in the growth of amyloid fibrils, we have provided a thermodynamic interpretation of addition of a monomer to a fully formed fibril. Although the results have been obtained by examining the addition ^{35}MVGGVV^{40} to a template fibril, the framework is expected to be of general validity. Because the structure of the unbound monomer is usually not commensurate with the fibril it follows that the monomer must undergo a cascade of structural transitions. Our simulations show that, surprisingly, even a small peptide can adopt a diverse set of conformations prior to locking onto the fibril. Because there is a great deal of structural diversity in the docked state it follows that the subsequent lock process must be dynamically heterogeneous. The diversity ini the locking state, and hence in the growth of amyloid fibrils, can be assessed using single molecule experiments.

From a computational perspective, we have provided a method for computing interactions between cosolvents for use in implicit solvent simulations. Using this methodology, we showed that small concentrations of urea and TMAO, that are known to have opposing effects on protein stability, increase the stability of the locked phase. The use of implicit cosolute models and the free energy profiles may be particularly useful in the computation of C_R, the critical monomer concentration that is in equilibrium with the fibril. The C_R values [141] can be used to predict qualitatively the relative (with respect to a reference condition) stability of the fibril bound monomer under varying solution conditions. The sensitivity of C_R to

mutations and cosolutes can be used to map regions of proteins or peptides that harbor amyloidogenic tendency.

5.4 Computational Methods

5.4.1 Fibril model:

To illustrate the structural transformations in a monomer interacting with a fibril, we chose a six residue peptide (^{35}MVGGVV^{40}) fragment from the Aβ [122] protein that forms amyloid-like fibrils *in vitro*, and whose fibrillar structure is known to 2 Å resolution [133]. We select a cross-section of this fibril's crystal structure (PDB code 2OKZ), two-by-three unit cells wide, made up of a total of twelve peptides, that lies perpendicular to the long fibril axis (Fig. 5.1). This leads to an approximately rectangular surface that is ∼48 Å long by ∼45 Å wide, and has the peptide backbones fully exposed to solvent.

The unit cell of the amyloid fibril crystal is monoclinic with angles α, β, γ of 90°, 96.9°, and 90°, respectively [133]. The unit cell distances a, b, c are 15.148, 9.58, and 23.732 Å, respectively [133]. We carry out simulations on a fibril surface that uses monoclinic periodic boundary conditions with the same α, β, γ angles as in the crystal and a, b, c values of 45.444, 125.0, 47.464 Å. This results in a fibril surface that has no lateral edges because it is infinite in the xz-plane (Fig. 5.1). Consequently, our simulations only probe monomer association to the surface of the fibril that is perpendicular to the long fibril axis. This is justified based on the experimental observations that, under certain conditions, soluble monomers deposit largely on the backbone exposed surface of the fibril. The probability of

lateral association of protofilaments is likely to be small during the late stages of fibril elongation [142, 143].

5.4.2 Solvent Model:

The CHARMM 22 force-field [144] in conjunction with grid based correction maps (CMAP) to the backbone dihedral angles [145], that reproduce *ab initio* computed Ramachandran plots of dipeptides, is used to model bonded and non-bonded protein interactions. It is difficult to carry out all-atom explicit solvent simulations that adequately sample the equilibrium conformational space. Hence, we use an all-atom representation of the protein and include the effects of solvent using a Generalized Born implicit solvent model (GBSW) [146]. With this simplification the simulation times can be greatly extended allowing us to obtain converged results for various thermodynamic quantities.

5.4.3 Mimics of cosolvents Urea and TMAO for use in implicit solvent simulations:

For use in implicit solvent simulations we introduce a novel way to model interactions involving cosolvents and proteins. We model urea and TMAO as spherical particles that interact with atoms in the peptide via

$$V(r_{ij}) \;=\; 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \lambda\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right], \tag{5.3}$$

where r_{ij} is the distance (in Å) between a protein atom i and a cosolute molecule j, ϵ_{ij} is the interaction strength between them, $\lambda = 1$, and the values of σ_{ij} are computed using the Lorentz-Brethlot mixing rules [106]. Interactions between the osmolyte molecules are repulsive (i.e. $\lambda = 0$ and $\epsilon_{ij} = 0.1$). The size of urea and

TMAO is $\sigma_i = 7$ Å. The larger collision diameter, compared to molecular volume and partial molar volume estimates of TMAO and urea molecules [147], approximately accounts for the ordered first solvation shell of water surrounding the cosolutes.

The ϵ_{ij} values in Eq. 5.3 are chosen so that $\delta g_{tr,E}(0M \rightarrow 1M) \approx \delta g_{tr,C}(0M \rightarrow 1M)$, where $\delta g_{tr,C}(0M \rightarrow 1M)$ is the computed free energy of transferring the individual protein groups (backbone or side chain) from pure water to aqueous osmolyte solution at 1 M, and $\delta g_{tr,E}(0M \rightarrow 1M)$ is the experimentally measured value [28, 62]. We calculated $\delta g_{tr,C}$ using the Widom particle insertion technique [148], where $\delta g_{tr,C} = -k_B T ln \int ds_{N+1} < exp(-\Delta U/(k_B T)) >_N$, with ΔU being the non-bonded interaction energy (i.e. the Lennard-Jones energy) between a system containing N TMAO (or urea) molecules and a randomly inserted protein group. The quantity $exp(-\Delta U/(k_B T))$ is averaged over all system configurations of the cosolutes. Because we are using an implicit solvent model and solutions at fairly low osmolyte concentrations (0.75 M) we are able to obtain converged $\delta g_{tr,C}$ data [149]. Typically, 10^5 insertion attempts were necessary to obtain $\delta g_{tr,C}$ values that had a standard error of less than 10^{-5} $kcal/mol$ [148]. Thus, many ϵ_{ij} parameters were tested until $\delta g_{tr,C}$ was within 0.5 cal/mol of $\delta g_{tr,E}$. The ϵ_{ij} parameters are listed in Table 5.1.

5.4.4 Simulation Details:

To enhance sampling efficiency low friction Langevin simulations [105], with a damping constant of 1.2 ps^{-1}, were carried out in conjunction with Multiplexed Hamiltonian Replica Exchange (MhREX) [36, 134]. In an MhREX run, multiple

independent trajectories (replicas) are simulated at different temperatures and with different Hamiltonians [36]. Periodically, the coordinates between the replicas are swapped according to a set of rules that preserve detailed balance [36].

We used three temperature windows (280, 325, and 380 K) and twelve different Hamiltonians (denoted $H_{i=1,..,12}$). For each temperature-Hamiltonian pair two independent trajectories are generated simultaneously. Thus, a total of 72 replicas are simulated in one MhREX run. The Hamiltonians differ only in the potential energy term $E_{U,i} = 0.5K_{U,i}(\delta_C - \delta_C^i)^2$, that restrains the center-of-mass of the monomer, defined using the C_α atoms of the monomer backbone, to a distance δ_C^i (in Å) along the y-axis from the fibril surface (Fig. 5.1). $K_{U,i}$ is the force constant (in $kcal/\text{Å}^2$) in the i^{th} Hamiltonian. The (δ_C^i,$K_{U,i}$) pairs for $i = 1, .., 12$ are (1.75,3.00), (3.00,3.50), (4.50,2.50), (6.00,2.5), (7.50,3.5), (9.00,3.25), (10.0,2.75), (11.0,1.5), (13.5,1.2), (15.0,1.0),(17.0,1.0),(19.0,1.0), respectively. We alternate the swaps between temperatures and Hamiltonians. Random shuffling between replicas at the same temperature and Hamiltonian are carried out at each swapping attempt. Every 143 integration time-steps swapping of system coordinates between temperatures or between Hamiltonians is attempted. In all, 55,000 swaps are attempted, with the first 5,000 discarded to allow for equilibration. The swapping acceptance ratio's were between 10% and 40%. Trajectories are simulated in the canonical (NVT) ensemble, and the equations of motion are integrated with a 2 fs time-step. The total simulation time per replica is 14.3 ns and the sum total simulation time (over all replicas) is 1.03 μs. We use the CHARMM software package (version c33b2) to generate the trajectories [106]. An in-house perl script was written to run MhREX.

5.4.5 Potential-of-mean-force (PMF) and Structural probes:

Thermodynamic properties of the system are computed using the WHAM equations [40, 37]. The PMF is computed as $F(\delta_C) = -k_B T ln[P(\delta_C)]$, where $P(\delta_C)$ is the probability of finding the monomer at a distance δ_C from the fibril surface. We used STRIDE to compute the secondary structure content of the monomer [135]. To examine the global orientation of the monomer, relative to the fibril surface, we compute the two-dimensional free energy surface ($F(\delta_C, cos(\theta)) = -k_B T ln[Z(\delta_C, cos(\theta))/Z(\delta_C)]$ as a function of δ_C and θ, the angle formed between a vector normal to the fibril surface and a vector connecting the C_α atoms of the N-terminus and C-terminus (Fig. 5.1). Backbone contacts between the added monomer and the peptides on the surface of the fibril are assumed to be formed if the C_α atoms between peptides are within a distance of 6 Å. Numbering each residue in a peptide from 1 to 6, starting from the N-termini, in-register parallel backbone contacts occur if residue i, the residue number, of strand j is in contact with residue k of strand l and $i = k$. Similarly, in-register anti-parallel contacts occur between strands j and l if i and k are in contact and $k = 7 - i$.

Table 5.1: Lennard-Jones parameters for urea and TMAO particle interactions with peptide atoms used in $4\epsilon_{ij}[(\sigma_{ij}/r_{ij})^{12} - (\sigma_{ij}/r_{ij})^{6}]$.

Atom Type i^a	$\epsilon_{i,urea}$	$\epsilon_{i,TMAO}$
CT1	0.0924875	0.074005
CT2	0.0924875	0.074005
CT3	0.091795	0.085255
CT4^b	0.0859	0.08521
CT5	0.0924875	0.074005
C	0.09162	0.059
O	0.09162	0.059
NH1	0.09162	0.059
H	0.022905	0.01475
HB	0.022905	0.01475
S	0.098	0.10725
HA	0.02294875	0.02131375

[a]Atom names, unless otherwise indicated, are the same as in the CHARMM 22 force-field [106]. Lorentz-Brethlot mixing rules are used for all other atoms [106].

[b]Atoms CT4 and CT5 are new atom types added to the CHARMM 22 force-field. CT4 and CT5 have the exact same properties as atoms CT2 and CT3, respectively, except for the Lennard-Jones parameters listed in this table. CT4 replaces CT2 in the valine residue. CT5 replaces CT3 in the methionine side chain.

Chapter 6

Factors governing helix formation in peptides confined to carbon nanotubes

6.1 Introduction

There is great interest in studying protein folding and dynamics in confined spaces because of their possible relevance to a variety of biological problems [150, 151, 152, 153, 154, 155, 156]. These include the fate of newly synthesized proteins as they exit the nearly 100 Å long and approximately cylindrical ribosome tunnel [150, 153], the effect of encapsulation of substrate proteins in the central cavity of the chaperonin GroEL [152], and the translocation of peptides across pores [157, 158, 159, 160]. Understanding the factors that determine the stability of confined proteins is also relevant in biotechnology applications [161]. The effect of being localized in the cylindrical tunnel of the ribosome, or the GroEL cavity, on peptide and protein stability is hard to predict because of the interplay of a number of energy and length scales [162, 163, 164, 165, 166, 167, 168, 169, 170]. They include the decrease, with respect to bulk, in conformational entropy of the ensemble of unfolded and native states, and the residue-dependent solvent-averaged interaction between the substrate protein with the interior of the confining pore. For example, the ribosome tunnel is lined with RNA near the peptidyl transfer center (PTC),

and proteins closer to the the exit tunnel. As a result, the interaction of a nascent peptide with the walls of the tunnel varies as it traverses from the PTC towards the exit [155]. Thus, the formation of α-helical structure in the tunnel, that is observed in experiments [153], not only depends on the sequence but also on where the peptide is localized inside the ribosome [156, 153].

A number of factors contribute to the changes in the stability of a peptide upon confinement to a nanotube. The simplest scenario is the entropic stabilization mechanism (ESM) [171, 162, 163, 164], which postulates that in confined spaces the number of allowed conformations is restricted compared to the bulk. As a result, the free energy change ΔF_U of the denatured state ensemble (DSE) and the ΔF_N in the native state ensemble (NSE) both increase. If the native state is not significantly altered in the confined space then $\Delta F_U >> \Delta F_N$. Hence, confinement entropically stabilizes the native state relative to the DSE. The stabilization of polypeptide chains suggested by ESM holds good only when D, the diameter of the nanotube, exceeds a threshold value, because the entropy cost of confinement of the ordered (α-helical) conformation is prohibitive when D is small [166]. If water mediated interactions involving proteins are altered by confinement then it may be possible for $\Delta F_N > \Delta F_U$ [151, 164, 167, 169, 172]. In this case, the native state can be destabilized in nanotubes. More generally, if specific interactions between the polypeptide and the walls of the pore are relevant, as appears to be the case in certain regions of the ribosome tunnel, the diagram of states of a confined polypeptide or protein can be rich [173].

Here, we study the changes in stabilities of a number of peptide sequences that

form helices to varying extents in bulk. By varying D, the strength of interaction, λ (see Eq. B.7 in Appendix B), between the hydrophobic residues and the carbon nanotube, and the polypeptide sequence we show that an interplay of a number of factors determines the stability of helical states of peptides confined to nanotubes. We find that the helix is entropically stabilized when D is small and the interaction between peptides and nanotube is weak. As λ increases the peptide can adsorb onto the wall of the nanotube. Interestingly, adsorption results in stabilization of the helix for an amphiphilic sequence, and destabilization for a polyalanine sequence. If the wall of the nanotube is decorated with patches that are 'hydrophobic' the helical stability can increase for the polyalanine. Thus, a very rich diagram of states of helix forming sequences is envisioned upon confinement in a nanotube.

6.2 Methods

In order to explore a wide range of possibilities we consider several helix forming sequences. The sequences are GDLDDLLKKLKDLLKG (an amphiphilic sequence denoted by AS) [174, 175], polyasparagine N_{16} (a polar sequence denoted PN) [176, 177], and polyalanine A_{16} (a hydrophobic sequence denoted PA) [178]. Each sequence is 16 residues long, which is close to the average helix length of $\sim$14 found in globular proteins [179]. We use three variations of AS to probe the effects of varying the bulk peptide properties (the nature of the DSE and NSE) on confinement. The parameters of sequence AS_1 (see Table 6.1) renders the helical state unstable in the bulk ($D \to \infty$). Sequences AS_2 and AS_3 are modeled so that

they form stable helices in the bulk. The changes in the intra-peptide interactions (see Table 6.1 in Appendix B) between the hydrophobic residues in AS_2 and AS_3 accounts for differences in ϵ_{BB} (Eq. B.6 in Appendix B) that can arise by adding cosolvents (see Appendix B for details).

We use the Honeycutt-Thirumalai (HT) [180] model for the polypeptide chain. In the HT model, each amino-acid is represented by one bead located at the C_α-carbon position. A three letter code is used to classify the twenty naturally occurring amino acids; L for hydrophilic residues, B for hydrophobic residues, and N for neutral residues. The potential energy of a conformation of a polypeptide with M residues, and coordinates $r_i(i = 1, 2, ..., M)$ in the HT representation is $V = V_B + V_A + V_D + V_{NB} + V_{HB}$, where V_B, V_A, and V_D are the bond-stretch, bond-angle, and the dihedral potentials respectively. The stability of the helices in the bulk can be altered by tuning the interaction, V_{NB}, between non-covalently linked beads, as well as the hydrogen bond potential V_{HB}. Details on the functional form, and the parameters of the energy function are provided in Appendix B.

In order to enhance the sampling of the conformational space of the peptide we use underdamped Langevin dynamics [105] with a friction coefficient of 0.016 ps^{-1}, and an integration time-step of 15 fs. Simulations are prepared and simulated in the NVT ensemble at 300 K using the CHARMM software package (version c32b2) [106].

Helical basin (HB): A given peptide conformation is classified as helical using two order parameters. They are the end-to-end distance (R_{ee}), and the number of helical triads (N_{HT}). We define helical triads as three consecutive dihedral angles

that are in the helical region ($35^\circ \leq \phi \leq 75^\circ$). A polypeptide with 16 residues has a total of eleven helical triads. In a completely helical conformation $N_{HT} = 11$, while $N_{HT} = 0$ corresponds to a completely random coil conformation. A conformation is deemed to be in the HB if 21.25 Å $< R_{ee} <$ 28.75 Å and $8 \leq N_{HT} \leq 11$. The two order parameters R_{ee} and N_{HT} separate the helical and denatured basins into distinct regions (see the inset in Fig. 6.1C).

6.3 Results and Discussion

For sequence AS_1 the probability of being in the HB (P_{HB}) is 0.17 in bulk. The values of P_{HB} for AS_2, AS_3, PA, and PN, are between 0.40-0.50 in the bulk (Table 6.1).

6.3.1 Helices are entropically stabilized in narrow and weakly hydrophobic nanotubes

If the attractive interaction between the hydrophobic residues and the nanotube is weak ($\lambda < 0.4$) then confinement enhances helix stability of all sequences provided $D < D^*$, where D^* depends on the sequence (Fig. 6.1) and is greater than or equal to 20 Å for the sequences studied here. For example, when AS_3, PA and PN are in a nanotube with $D = 14.9$ Å and $\lambda = 0.01$, the helix is stabilized by 0.71, 0.68 and 0.49 kcal/mol, respectively (computed using the data from Fig. 6.1). The enhanced helix stability at $D < D^*$ and $\lambda < 0.4$ can be explained using polymer arguments [166], from which it follows that when D is small enough the

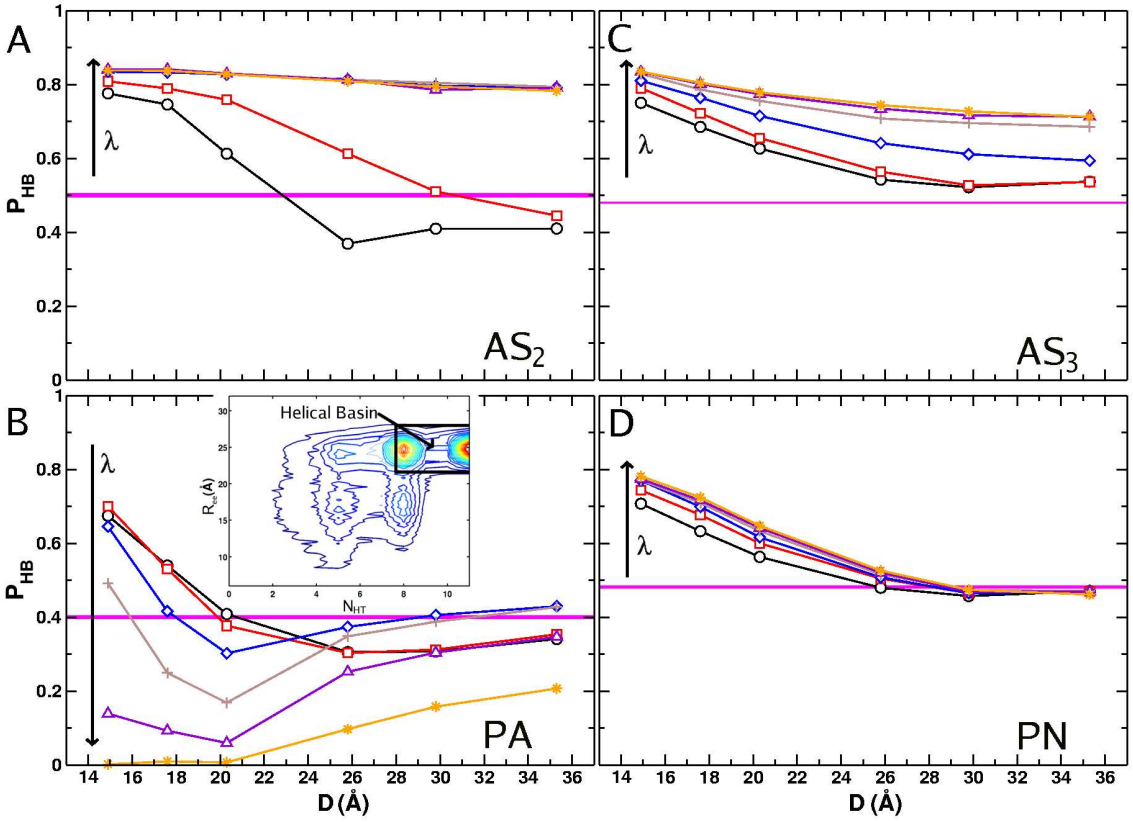


Figure 6.1: The probability of being in the HB as a function of nanotube diameter for the sequences AS_2 (A), PA (B), AS_3 (C) and PN (D) at various λ values (λ= 0.01 (black circles), 0.1 (red squares), 0.3 (blue diamonds), 0.5 (brown plus signs), 0.7 (purple triangles) and 1.0 (orange stars)). The horizontal magenta colored line, in each graph, corresponds to the probability of being helical in bulk, and the width corresponds to the standard error of P_{HB}^{B}. We characterized a given peptide conformation as helical using two order parameters, the end-to-end distance (R_{ee}) and the number of backbone dihedral angles that are helical ('Helical Triads') (see the inset in (C)). A peptide conformation is helical if 21.25 Å $< R_{ee} <$ 28.75 Å and $8 \leq N_{HT} \leq 11$.

helical basin is entropically stabilized. Fig. 6.1 shows that the helical content of AS_3 and PN increases for all D. While for AS_2 and PA P_{HB} increases only below $D < D^* \sim (20-22)$ Å. The sequence-dependent values of D^* are difficult to predict using polymer theory alone. Interestingly, for AS_2 and PA we find that P_{HB} changes non-monotonically as D decreases (Figs. 6.1A and 6.1B). Such a behavior is also mirrored in the variation of $\langle R_{ee} \rangle$ as D is changed (data not shown), in agreement with theoretical predictions [181].

For small $\lambda (\sim 0.01)$, we expect that the effect of confinement can be described by the difference in entropy changes in the DSE and the HB. We estimate confinement-induced free energy changes using

$$\begin{aligned} \Delta\Delta G(D, \lambda \sim 0.01) &\approx -T[k_B ln(\alpha_{HB}(D)) - k_B ln(\alpha_{DSE}(D))] \\ &\approx -T[\Delta S_{HB}(D) - \Delta S_{DSE}(D)], \end{aligned} \tag{6.1}$$

where $\Delta S_{HB}(D)$ and $\Delta S_{DSE}(D)$ are the changes in entropy upon confinement of the helix and DSE, respectively. The volume fraction accessible to the HB ($\alpha_{HB}(D)$) and DSE ($\alpha_{DSE}(D)$), are calculated numerically using the Widom particle insertion method (see Appendix B for details). The similarity (Fig. 6.2) in the values of $\Delta\Delta G(D)$ computed using $\alpha_{HB}(D)$ and $\alpha_{DSE}(D)$ and that obtained directly from $P_{HB}(D)$ (Fig. 6.1) shows that the helix formed by AS_3 is entropically stabilized for all D. In contrast, $\Delta S_{DSE}(D) > \Delta S_{HB}(D)$ for AS_2 and PA when $D > D^* \sim 20$ Å which leads to destabilization of the helix upon confinement. Thus, the differences in the intrapeptide interaction strength between sequences AS_2 and AS_3 can change the nature of the DSE and HB, and can result in either helix stabilization (for AS_3)

or helix destabilization (for AS_2) when $D > 20$ Å. The differing behavior of AS_2 ($\epsilon_{BB}/k_BT \approx 3$) and AS_3 ($\epsilon_{BB}/k_BT \approx 0.9$) shows that the nature of the conformations explored in the bulk affects confinement-induced stability. In principle, ϵ_{BB} can be altered in experiments by addition of cosolvents or by changing temperature.

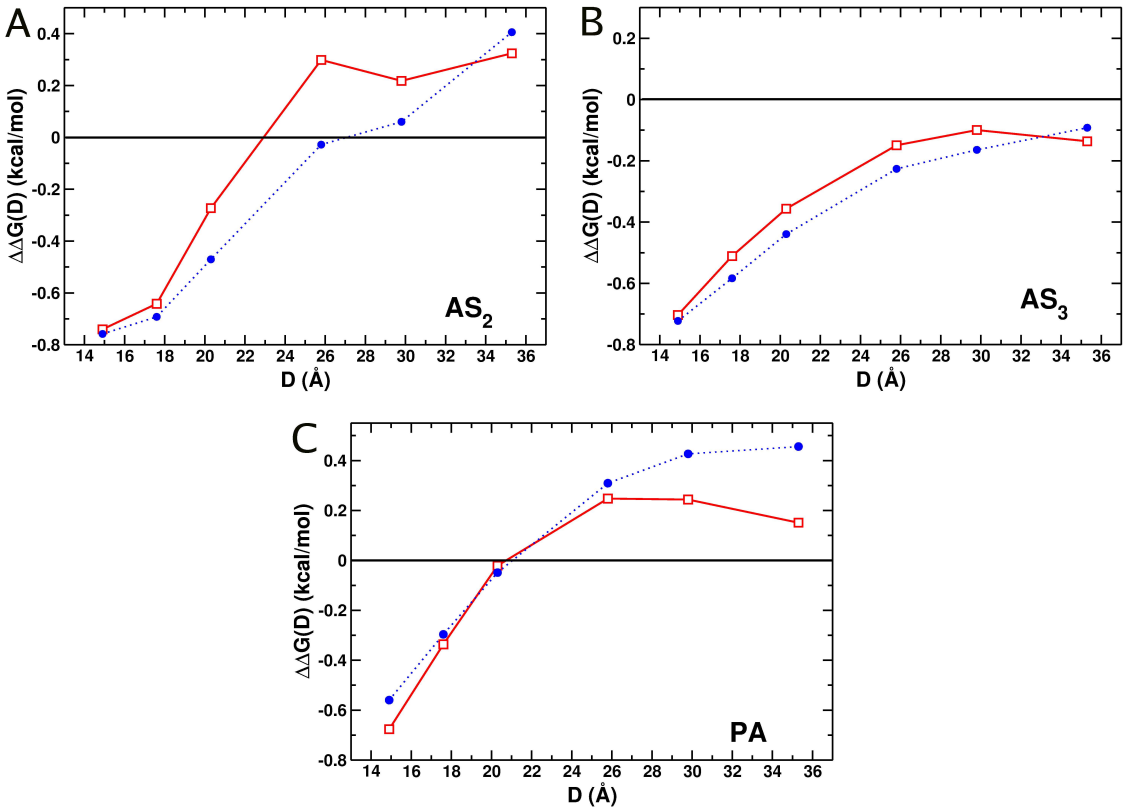


Figure 6.2: The change in free energy ($\Delta\Delta G(D) = \Delta G(D) - \Delta G(B)$) of the HB, relative to the DSE, upon nanotube confinement as a function of D. The free energy difference in the bulk ($D \to \infty$) is given by $\Delta G(B)$. $\Delta\Delta G(D)$ computed from $P_{HB}(D)$ ($\Delta\Delta G(D) = -k_BTln\left[\frac{P_{HB}(D)P^B_{HB}}{(1-P_{HB}(D))(1-P^B_{HB})}\right]$ and $\alpha(D)$ (see Eq. 6.1) are shown as red squares and blue circles, respectively. Lines are to guide the eye. The results in panels (A), (B), and (C) are for AS_2, AS_3, and PA respectively.

6.3.2 Hydrophobic residues are pinned to the nanotube as λ increases

We expect that increasing λ should result in sequences containing hydrophobic residues to adsorb onto the nanotube wall. The probability density of finding a residue i at a distance r_i from the long nanotube axis, shows all sequences sample the interior of the nanotube at $\lambda = 0.01$ (Fig. 6.3). As a result, we expect that confinement-induced helix stabilization should be largely determined by entropy considerations. However, as λ increases, sequences containing hydrophobic residues (PA, AS_1, AS_2, and AS_3) can be pinned to the wall, as indicated by the greater probability density of peptide residues near the nanotube surface (Figs. 6.3A and 6.3B). In the case of the amphiphilic sequence, the peptide sticks to the wall (Fig. 6.3A) and forms a helix (Figs. 6.1A and 6.1C). The spatial distribution of residues in the HB corresponds well with the probability density plotted for $\lambda = 1.0$ (Fig. 6.3A). The results in Fig. 6.3, which show that hydrophobic residues are pinned to the wall, while polar residues are more likely to be sequestered in the interior of the nanotube, suggests that a 'phase separation' occurs on the molecular length scale between hydrophobic and polar peptide residues.

The distribution functions in Fig. 6.3 shows that for an amphiphilic sequence, the stability of helices should be determined by the opposing tendency of hydrophobic residues to be pinned to the wall of the nanotube and the preference of the polar residues to be localized in the interior. Indeed, we find that for AS_1, AS_2, and AS_3 the helical content increases as λ increases (Fig. 6.4). The effect of increasing λ is most dramatic for AS_1 ($\epsilon_{BB} = 0$), for which P_{HB} increases dramatically from below

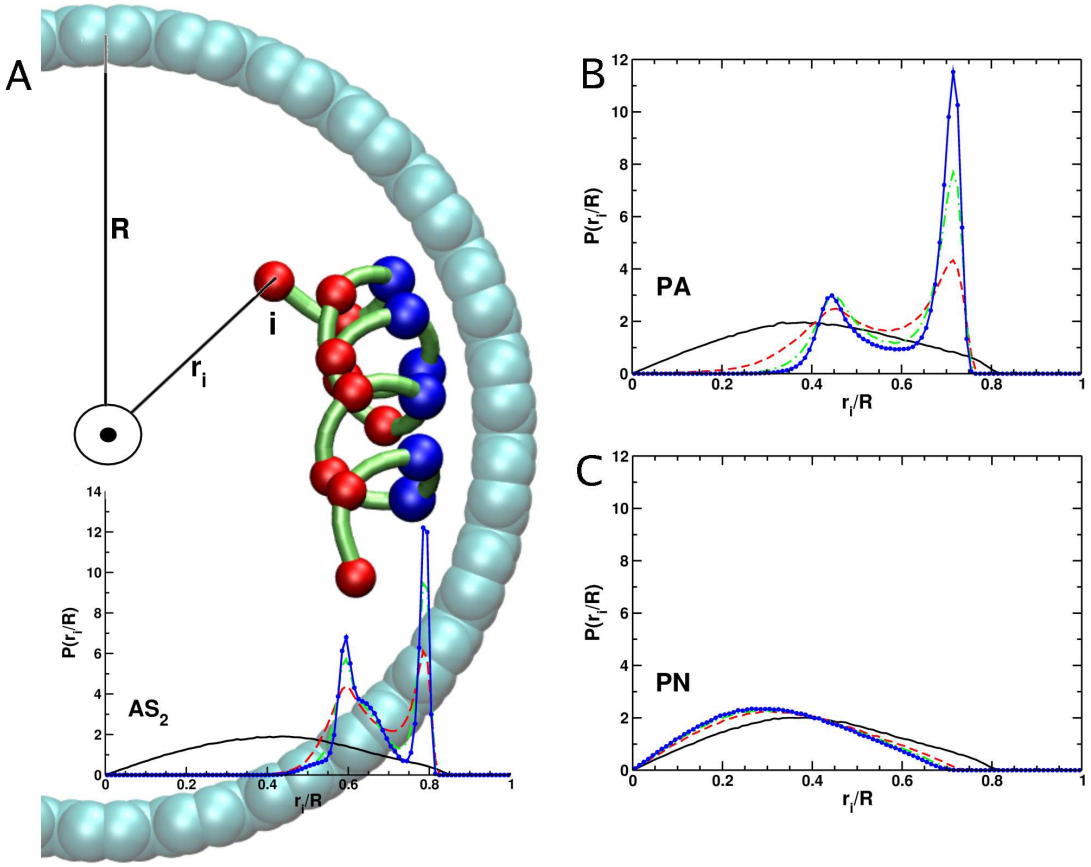


Figure 6.3: The probability density of finding a residue i at a distance r_i/R (R, the nanotube radius, is 14.9 Å in (A), $R = 12.9$ Å in (B) and (C)) from the long nanotube axis at different λ values for AS_2 (A), PA (B) and PN (C). Four different values of λ are plotted, $\lambda = 0.01$ (solid black line), 0.3 (dashed red line), 0.7 (dash-dot green line) and 1.0 (solid blue line with circles). The image in the background of (A) is on the same scale as the graph overlaying it. The spatial distribution of the residues in the image correspond well with the probability density at $\lambda = 1.0$. In the image hydrophobic residues are shown in blue, and polar residues are in red.

the bulk value of $P_{HB}^{B} \approx 0.17$ (Fig. 6.4A). For AS_1, the helix is greatly stabilized by the favorable interactions between the hydrophobic residues and the nanotube. In the case of AS_2, increasing λ maximizes the attractive interactions between B (hydrophobic) beads with the nanotube without compromising the intra-peptide BB interactions in the HB. Similarly, P_{HB} increases (Fig. 6.4B) for AS_3 ($\epsilon_{BB} = 0.5$ $kcal/mol$) as λ increases although the changes in P_{HB} occur over a wider range of λ compared to AS_2 ($\epsilon_{BB} = 2.125$ $kcal/mol$) (Fig. 6.4B).

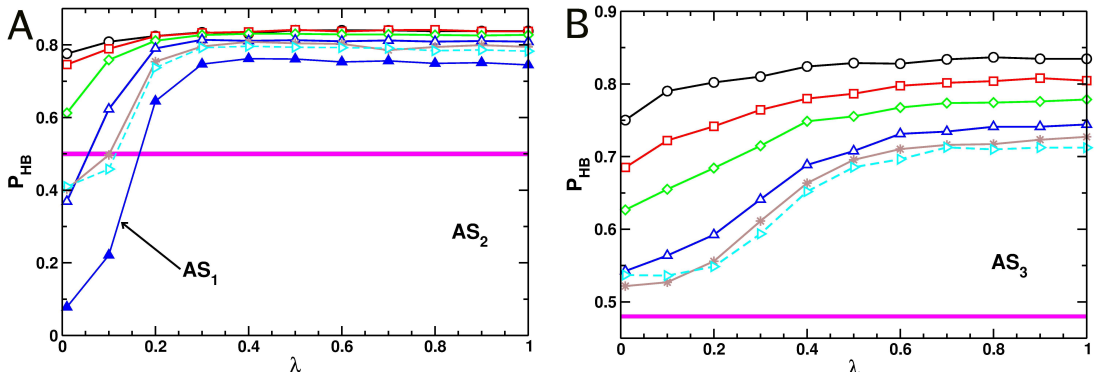


Figure 6.4: Probability of being in the HB as a function of λ in different diameter nanotubes for the three variations of the amphiphilic sequence. The graphs show that AS_1, AS_2, and AS_3 tend to be stabilized by increasing the strength of the hydrophobic interactions with the nanotube. P_{HB} versus λ is shown for the two amphiphilic sequences AS_2 (A) and AS_3 (B) for different nanotube diameters ($D =$ 35.3 Å - cyan triangles, 29.8 Å -brown stars, 25.8 Å - blue triangles, 20.3 Å - green diamonds, 17.6 Å - red squares, and 14.9 Å - black circles). Results for AS_1 with $D = 25.8$ Å are shown as blue filled triangles in (A).

When the amphiphilic sequence is in the HB, all of the hydrophobic residues are aligned on one side of the helix while the polar residues are exposed on the other side (Fig. 6.3A). Thus, for all variations of AS the HB is stabilized because it maximizes the hydrophobic interaction between the hydrophobic face of the helix and the

hydrophobic surface of the nanotube. If the helical pitch (p) is commensurate with the distance between the carbon atoms (R_{CC}) along the long axis of the nanotube, we expect that the interactions between the hydrophobic residues and the nanotube can be maximized without compromising the helical structure. Conversely, if p and R_{CC} are incommensurate it is likely that the helix may be denatured. Thus, besides the sequence, the relative positions of the hydrophobic residues in the helix are also important determinants of stability in a nanotube, especially as λ increases.

6.3.3 Diagram of states of polyalanine in a carbon nanotube is rich:

The interplay between the strength of the hydrophobic interactions and the entropy of confinement results in a rich phase diagram in the (λ, D) plane for PA (Fig. 6.5A). The stability of the HB decreases as λ increases as long as $D <\sim 20$ Å (see points 1, 5, and 6 in Fig. 6.5A). The effect is most dramatic in the narrowest tube ($D = 14.9$ Å in Fig. 6.5B) in which P_{HB} nearly vanishes as λ approaches unity. In larger nanotubes ($D > 20$ Å), P_{HB} increases by about (7-10)% as λ increases from $\lambda = 0.01$, reaches a maximum at $\lambda \sim 0.4$ and then decreases upon further increase in λ (Fig. 6.5B and see points 2, 3, and 4 in Fig. 6.5A). This modest helix stabilization occurs because the peptide weakly binds to the wall of the nanotube as λ increases (Figs. 6.3B and 6.5A, point 3), resulting in preferential alignment of the peptide along the long axis of the nanotube (Fig. 6.5B point 3 and Fig. B.2A in Appendix B). At $\lambda \approx 0.4$ and $D > 20$ Å, the interaction with the nanotube is not strong enough to overcome the internal peptide energies which favor the helix.

As a result, the nanotube-peptide interactions are maximized when the peptide is in the HB. As λ is further increased, hydrophobic interactions with the wall cause the helical content to decrease (Fig. 6.5B). In the largest nanotube ($D \approx 35$ Å), as λ approaches unity P_{HB} decreases because the peptide gets splayed out along the interior of the nanotube surface (Fig. 6.5A, point 4). For nanotubes with $D \approx 20$ Å, increasing λ stabilizes a 'broken' helix (Fig. 6.5, point 5) that does not align along the long nanotube axis (Fig. B.2A in Appendix B), but instead binds to the nanotube perpendicular to the nanotube axis (Fig. B.2B in Appendix B). For the smallest diameter nanotubes, increasing λ stabilizes a coiled peptide that coats the interior surface of the nanotube (Fig. 6.5A, point 6) but has no helical dihedral angles.

Taken together these results show that the effect of varying the hydrophobic character of the nanotube on helix stability is subtle for the PA. For the largest nanotube diameters there is an optimal hydrophobic strength which stabilizes the helix modestly. For smaller nanotube diameters divergent behavior is observed. Weakly hydrophobic nanotubes ($\lambda < 0.4$) stabilize the helix as D gets smaller. In contrast, destabilization of the helix occurs when $\lambda > 0.6$.

6.3.4 Hydrophobic patches lining the nanotube affect P_{HB} of PA:

To mimic the chemical heterogeneity of the groups in the ribosome tunnel, which has small hydrophobic patches from proteins (such as L4, L17 and L39 in the ribosome of eukaryotes [153] surrounded by hydrophilic patches from RNA [182]),

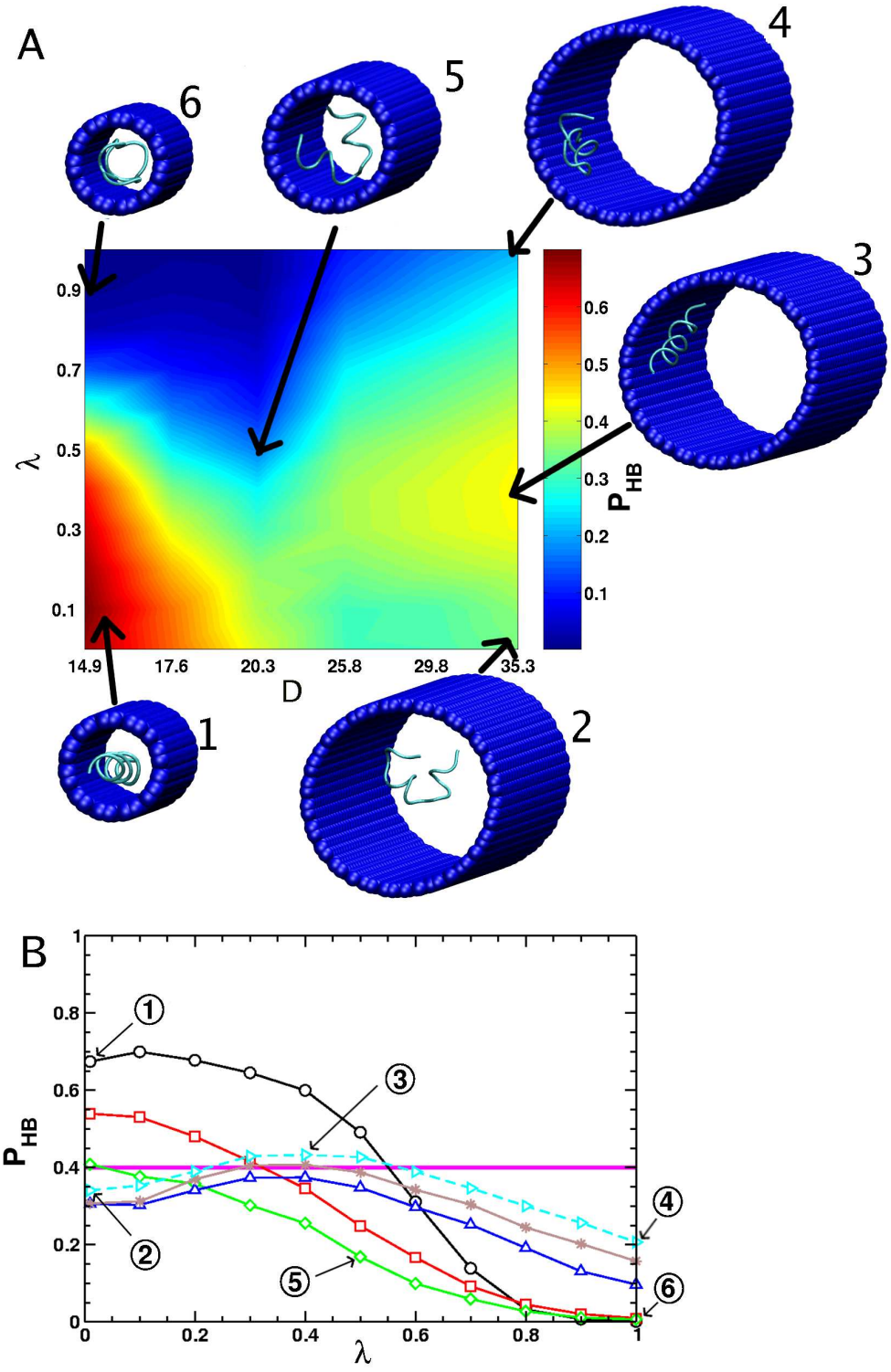


Figure 6.5: The probability of being in the HB as a function of D and λ for PA. (A) Phase diagram in the (λ, D) plane. Representative structures are shown in the images labeled 1 through 6. (B) The dependence of P_{HB} on λ for various D. See Fig. 6.4 for explanation of the symbols. The points labeled 1 through 6 correspond to the structures labeled in (A).

we created different size hydrophobic patches that line the nanotube (Fig. 6.6A). The desired heterogeneity is achieved by assigning hydrophilic character to subsets of nanotube atoms that run parallel to the long nanotube axis, and hydrophobic behavior to the rest of the nanotube atoms (see Methods section for details). With $\lambda = 0.9$, we vary the size of the hydrophobic patch. The fraction of hydrophobic surface area f_H varies from 0 to 1. Surprisingly, we find that the helical stability of PA, whose helical content is negligible at $\lambda = 0.9$ and $f_H = 1$ for all D (Fig. 6.1), increases as f_H decreases (Fig. 6.6B). In the smallest nanotube ($D = 14.9$ Å), P_{HB} increases monotonically as f_H decreases, with the smallest hydrophobic patch imparting the greatest helix stability. In larger nanotubes, P_{HB} as a function of f_H is nonmonotonic. Thus, there is an optimal f_H, between 0.08 and 0.15, in these larger nanotubes that maximizes P_{HB} for PA.

6.4 Conclusions:

The effect of nanotube confinement on the stability of the helical states depends on the sequence, the tube diameter, nanotube-peptide interactions as well as the chemical heterogeneity of the the nanotube. The remarkably complex behavior of peptides in nanotubes illustrates that it is possible to control confinement-induced helix stability by altering a number of variables. The substantial diversity in the stability as a function of (D, λ), even for a specific sequence (Fig. 6.5A), shows that solvent-mediated peptide-nanotube interactions (parameterized by λ) can either stabilize or destabilize the HB depending on D. Our results show that it would

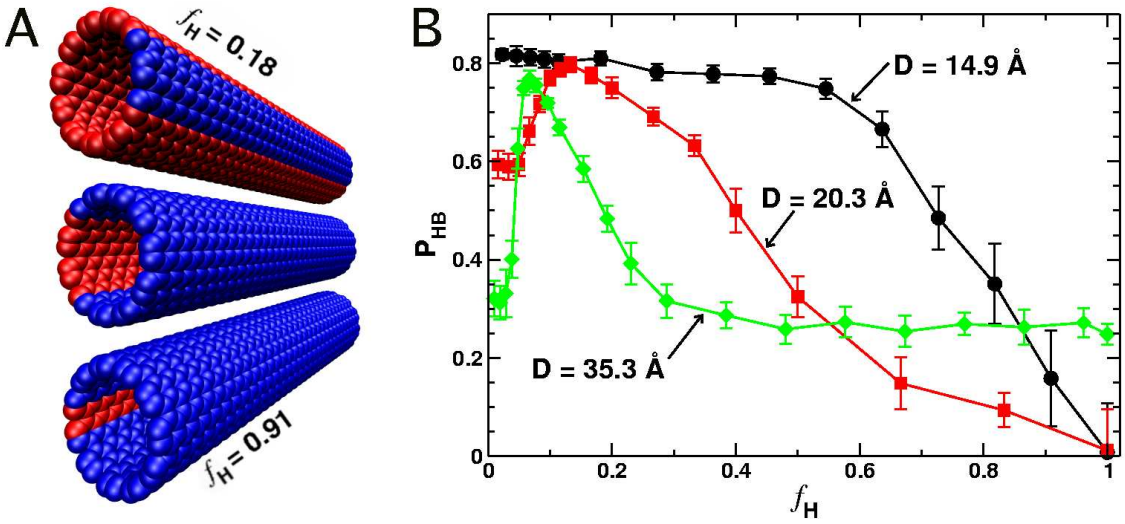


Figure 6.6: Changes in P_{HB} in a chemically heterogeneous nanotube. (A) The size of the hydrophobic patch lining the nanotube (nanotube atoms with hydrophobic character are shown in blue, while those with hydrophilic character are shown in red). The value of D is 14.9 Å, and the fraction of nanotube hydrophobic surface area, f_H, is 0.18, 0.73 and 0.91 for the top, middle and bottom nanotubes. (B) The probability of being in the HB as a function of f_H with $D = 14.9$, 20.3 and 35.3 Å, and $\lambda = 0.9$. For the smallest nanotube, a homogeneous hydrophobic environment ($f_H = 1$) destabilizes the helix, while the smallest hydrophobic patches maximize helix stability. For larger D there is an optimal hydrophobic patch size that maximizes helix stability.

be erroneous to draw general conclusions [169] based on the study of a single sequence in a nanotube with various values of D.

A key prediction of this study is that confinement-induced helix stability can be dramatically altered by varying the intra-peptide interactions, or by changing the interaction strength between the peptide and the nanotube. The changes in the stability of the HB of the amphiphilic sequence (AS_1, AS_2, and AS_3) most vividly illustrate the effects of λ, ϵ_{BB}, and D (Fig. 6.1). The variations in ϵ_{HB} and ϵ_{BB}, which distinguish AS_1, AS_2, and AS_3, can be realized by varying cosolvent

conditions. The differences in their stabilities upon confinement in AS_1, AS_2, and AS_3 is due to substantial changes in the DSE. The finding that the stability of a polyalanine sequence can be greatly altered by changing λ and D (see Fig. 6.5A) can be experimentally tested. The changes in λ can be achieved by varying the solvent density in the nanotube.

A prediction of plausible relevance to peptide folding in the ribosome is the demonstration that helix stability also depends strongly on the size of the hydrophobic patch lining the nanotube. If the entire interior of the nanotube is hydrophobic ($f_H = 1$), the HB of the polyalanine peptide is completely destabilized when the interaction between the peptide and the nanotube is $\lambda = 0.9$. However, as the patch takes up a smaller percentage of the surface area of the nanotube, the stability of the polyalanine helix increases. In the nanotube diameter range comparable to the ribosome tunnel ($D \approx 15$ Å), we find that the smallest size hydrophobic patches maximizes the helix stability. As a result, we predict that helix stability can increase in regions of the ribosome tunnel where small hydrophobic patches exist. Clearly, the extent of stabilization in the ribosome tunnel will depend on the sequence.

Table 6.1: Models and simulation details

Sequence	Label	ϵ_{HB}[a]	ϵ_{BB}[b]	D (Å)	Time (μs)[c]	P_{HB}^{B}[d]
GDLDDLLKKLKDLLKG[e]	AS_1[f]	0.00	2.125	25.8	2.0	0.17
	AS_2[g]	1.75	2.125	all[h]	3.3	0.50
	AS_3[i]	2.75	0.50	all	3.3	0.48
A_{16}	PA	2.75	0.50	all	3.3	0.40
N_{16}	PN	2.50	0.50	all	3.3	0.48

[a]The implicit hydrogen bonding energy in $kcal/mol$, see Eq. B.6 in Appendix B.

[b]The Lennard-Jones well-depth between hydrophobic residues in $kcal/mol$, see Eq. B.5 in Appendix B.

[c]The total simulation time per nanotube diameter.

[d]The probability of being in the HB in bulk.

[e]One letter code is used for amino acids.

[f]Original parameter set of Guo and Thirumalai [183].

[g]Modified Dihedral Potential (see Table 1 in Appendix B) and V_{HB} term (see Eq. B.6 in Appendix B).

[h]'all' indicates that nanotubes with D= 35.3, 29.8, 25.8, 20.3, 17.6, and 14.9 Å were studied.

[i]Same Parameter Set as AS_2 except $\epsilon_{BB} = 0.5\ kcal/mol$.

Chapter A
Appendix for Chapter 1
A.1 C_α-SCM for polypeptide chains:

The general formalism for obtaining thermodynamic averages, described in the text, is applicable to any model for which adequate sampling of conformational space can be performed. In general, it is difficult to use all atom molecular dynamics simulations to sample the protein conformations sufficiently to obtain reliable values for thermodynamic quantities such as the free energy. In order to circumvent this sampling problem, we use coarse-grained models that have proven to be successful in providing insights into protein folding mechanisms. In the present study, we used the C_α-SCM model [59] to represent protein L and CspTm.

In terms of the coordinates of the C_α and side chain (SC) interaction sites, the potential energy of the C_α-SCM is $E_P = E_A + E_{HB} + E^N_{NB} + E^{NN}_{NB}$. For the angular potential, E_A, we used

$$E_A = \sum_{i=1}^{N_A} K_A(\theta_i - \theta_{i,0})^2 + \sum_{i=1}^{N_D}\sum_{j=1}^{3} K_{D_j}(1 + cos(n_j\phi_i - \delta_{ij})) + \sum_{i=1}^{N_{Ch}} K_{Ch}(\psi_i - \psi_{i,0})^2 \quad \text{(A.1)}$$

The sum of the hydrogen bond potential (E_{HB}) and the non-bonded potential (E^N_{NB}), between sites that are in contact in the crystal structure, was taken to be

$$\begin{aligned} E_{HB} + E^N_{NB} = & \sum_{i=1}^{N_{HB}} \epsilon_{HB}\left[\left(\frac{r^o_{HB,i}}{r_i}\right)^{12} - 2\left(\frac{r^o_{HB,i}}{r_i}\right)^{6}\right]. \\ & + \sum_{i=1}^{N_N} \epsilon^N_i\left[\left(\frac{r_{min,i}}{r_i}\right)^{12} - 2\left(\frac{r_{min,i}}{r_i}\right)^{6}\right] \end{aligned} \quad \text{(A.2)}$$

For non-bonded interaction sites that are not in contact in the crystal structure,

$$E_{NB}^{NN} = \sum_{i=1}^{N_{NN}} \epsilon_i^{NN} \left[\left(\frac{r_{min,i}}{r_i} \right)^{12} \right]. \quad \text{(A.3)}$$

Detailed explanation of the various terms in the force field for the C_α-SCM representation of polypeptide chains are given below. The values of the parameters in Eqs. A.1-A.3 are given in Table A.3.

Angular and Chiral potentials: Backbone bond lengths were set to the C_α atom distances found in the crystal structure. Bond lengths between C_α's and side chains correspond to the distance between the C_α atom and the side chain (SC) center-of-mass. We fixed the bond lengths using the SHAKE algorithm [184]. We used three angular restraints per residue to enforce the values of the bond angles. These angles are defined by the sets $\{C_{\alpha,i-1},C_{\alpha,i},C_{\alpha,i+1}\}$, $\{C_{\alpha,i-1},C_{\alpha,i},SC_i\}$, and $\{SC_i,C_{\alpha,i},C_{\alpha,i+1}\}$, where $C_{\alpha,j}$ and SC_j correspond to the C_α atom and the SC site of the j^{th} residue, respectively. The angles were harmonically restrained around their values, $\theta_{i,0}$, found in the crystal structure (first term in Eq. A.1). To model the restricted rotation around backbone bonds we used a dihedral potential which has three minima (second term in Eq. A.1). The most enthalpically favorable minimum corresponds to $\delta_{i1} + 180^o$, where δ_{ij} is the value of the j^{th} dihedral potential energy term of the i^{th} dihedral angle in the crystal structure. We enforced chirality, around the α-carbon atoms, using an improper dihedral angle, ψ, whose equilibrium angle, $\psi_{i,0}$, was set to that found in the crystal structure (third term in Eq. A.1).

Hydrogen bond potential: We modelled backbone hydrogen bonding using a Lennard-Jones interaction applied only between residues forming a hydrogen bond

in the crystal structure (first term in Eq. A.2). We used $\epsilon_{HB} = 0.75\ kcal/mol$, and $r^o_{HB,i}$ was set to the C_α- C_α native state distance between i^{th} pair of residues that are identified by Stride [135] as forming backbone hydrogen bonds.

Non-bonded potentials: We divided the non-bonded potential into contributions arising from native interactions, E^N_{NB} (second term in Eq. A.2), and that due to non-native interactions, E^{NN}_{NB} (Eq. A.3). These two potentials are only applied to interaction sites separated by four or more covalent bonds. We assumed that native non-bonded interactions between $C_\alpha - SC$ and $SC - SC$ pairs were present if the distance between any heavy atoms of the two groups was less than 4.5 Å in the crystal structure. These interactions were modelled using an attractive Lennard-Jones interaction (second term in Eq. A.2). In the case of $SC - SC$ pair interactions the well depth, ϵ^N_i, was taken to be proportional to the energy terms in the Miyazawa-Jernigan statistical potential [185] (for additional details see the footnote in Table A.3). For the $C_\alpha - SC$ interactions we set $\epsilon^N_i = -0.37$ kcal/mol. The interactions between non-bonded pairs that do not satisfy the criterion for native contacts were assumed to be purely repulsive (Eq. A.3). For all the non-native $C_\alpha - C_\alpha$, $C_\alpha - SC$ and $SC - SC$ pairs $\epsilon^{NN}_i = 10^{-12}$ kcal/mol and $r_{min,i}$ was set to 2.74 Å for the C_α interaction site. The sequence dependent side chain values of $r_{min,i}$ are listed in Table A.4.

Transfer free energies for SC and BB: The values of $\delta g^{SC}_{tr,k}([C])$ and $\delta g^{BB}_{tr}([C])$ were fit to the experimental transfer free energy data using

$$\delta g^{SC}_{tr,k}([C]) = m_k[C] + b_k \tag{A.4}$$

and

$$\delta g_{tr}^{BB}([C]) = m_{BB}[C] + b_{BB}. \tag{A.5}$$

For all cosolutes, except GdmCl, $b_k = 0$ and $b_{BB} = 0$. The values of b_k for GdmCl are listed in paranthesis in Table A.2. Experimental values of m_k for Ser, Asp, Glu, and Lys in GdmCl are unavailable. For Ser, Asp, and Glu we used the m_k values for Thr, Asn, and Glu respectively. The values of m_k for Lys in GdmCl was taken to be three times that for urea. The values for m_k (in units of $cal\ mol^{-1}\ M^{-1}$) and b_k ($cal\ mol^{-1}$) for all cosolutes are in Table A.2. The m_k and b_k values for GdmCl were extracted from [46]. Parameters for all other cosolutes were taken from [28].

A.2 Simulations:

The equilibrium simulations at zero osmolyte concentration were carried out using Multiplexed-Replica Exchange (MREX) [134] in conjunction with low friction Langevin dynamics [105]. MREX simulates multiple independent trajectories (referred to as replicas) at each temperature. MREX uses the conventional replica exchange acceptance/rejection criteria for swapping replicas between temperatures [186], but in addition it allows swapping between replicas at the same temperature [134].

In the MREX simulations, we used eight to nine temperature windows. For protein L, replicas at 315, 335, 350, 355, 360, 365, 380, 400 K were simulated, while for CspTm an additional replica at 450 K was included. At each temperature we generated four independent trajectories simultaneously, for a total of 32 or 36 replicas. Every 5,000 integration time-steps the system configurations were saved for analysis

and random shuffling occurred between replicas at the same temperature with 50% probability. Exchanges between neighboring temperatures were then attempted using the standard replica exchange acceptance criteria [186]. We attempted 90,000 exchanges for each protein, with the first 10,000 discarded to allow for equilibration. We used Langevin dynamics in the under damped limit to simulate the time evolution of each replica [105]. A damping coefficient of 1.0 ps^{-1} was used, with a 5 fs integration time-step. All trajectories were simulated in the canonical (NVT) ensemble.

A.3 Data Analysis

Solvent Accessible Surface Area: The solvent accessible surface area (SASA) of a backbone ($\alpha_{i,k}^{BB}$) or side chain ($\alpha_{i,k}^{SC}$) group in residue k of the i^{th} simulated protein conformation was computed using the CHARMM program. CHARMM computes the analytic solution for the SASA. A probe radius of 1.4 Å, equivalent to the size of a water molecule, was used.

Tripeptide $\alpha_{k,Gly-k-Gly}$***:*** To determine the SASA of residue k in the tripeptide $Gly-k-Gly$, for use in Eq. 3.2, we modelled the twenty tripeptides using the CHARMM 22 force field. A systematic search in the (ϕ, ψ) backbone dihedral space was carried out, and the SASA of the backbone and side chain groups of residue k at each (ϕ, ψ) point computed. We increased ϕ and ψ in 10° increments, starting from $\phi = \psi = 0°$. A total of 1,369 unique ϕ, ψ pairs and SASA measurements were generated. The values of $\alpha_{k,Gly-k-Gly}$, listed in Table A.1, correspond to the maximum SASA found during the (ϕ, ψ) search.

Radius-of-gyration (R_g): The radius-of-gyration (R_g) was computed using

$$R_g^2 = \frac{1}{N}\left\langle \sum_{i=1}^{N}(r_i - r_{CM})^2 \right\rangle \tag{A.6}$$

where r_i is the position of interaction site i, and $r_{CM} = 1/N \sum_{i=1}^{N} r_i$ is the mean position of the N interaction sites of the protein. The histogram of R_g values was taken to be the probability distribution $P(R_g)$. The average R_g is denoted by $\overline{R_g}$.

Native Contacts (Q): The fraction of native contacts (Q) of a given protein conformation was calculated using,

$$Q_i = \sum_{j}^{N-4} \sum_{k=j+4}^{N} \frac{\Theta(R_C - d_{jk})}{C_i} \tag{A.7}$$

where Q_i is the fraction of native contacts corresponding to either the entire protein or some substructure of the protein. In protein L, i, in Eq. A.7, can represent the set of native contacts within the helix ($i = H$), or between β-strands 1 and 2 ($i = s12$), 1 and 4 ($i = s14$), 3 and 4 ($i = s34$), or between strands 1, 2 and the helix ($i = H - s12$), strands 1, 3 and the helix ($i = H - s13$), strands 3, 4 and the helix ($i = H - s34$). The cutoff distance $R_C = 8$ Å, and d_{jk} is the distance between interaction sites j and k, $\Theta(R_C - d_{jk})$ is the Heaviside step function. In protein L, strand 1 ($s1$) corresponds to residues 4-11, $s2$ between 17-24, $s3$ corresponds to 47-52, $s4$ between 57-62, and H spans residues 26-44. In Eq. A.7, C_i is the maximum number of native contacts in the set i. In CspTm, $s1$ corresponds to residues 3-9, $s2$ to 14-19, $s3$ to 24-27, $s4$ to 44-52, and $s5$ to residues 57 through 65.

Fret Efficiency (E): The FRET efficiency (E) for a given protein confor-

mation was computed as

$$E = \frac{1}{1 + R_{ee}^6/R_o^6}, \tag{A.8}$$

where R_o was set to 55 Å for both proteins [63, 64]. In principle R_o should depend on the denaturant concentration. The variations in the FRET efficiency are relatively small when a denaturant-dependent R_o is used. Given the uncertainty among the different experimental measurements for protein L and CspTm (see Fig. 2.1) we did not include the changes in $\langle E \rangle$ due to changes in R_o.

Root Mean Square Deviation (Δ): The root mean square deviation (Δ) between two structures was computed using the CHARMM program. Least squares fitting was initially carried out to align a given protein conformation with the crystal structure before the RMSD was computed.

Computation of thermodynamic properties of the NSE and DSE: The NSE and DSE are differentiated using Δ as an order parameter. For protein L (CspTm) we defined the native basin as conformations with a $\Delta \leq 5$ Å (13.5 Å) and denatured conformations as $\Delta > 5$ Å (13.5 Å). We used this criterion to calculate the thermodynamic properties of the NSE and DSE using

$$\langle A^l \rangle = \frac{1}{\sum_t \Theta_l} \sum_t A(t)\Theta_l \tag{A.9}$$

where $\langle A^l \rangle$ is the average of any thermodynamic property A in the DSE or NSE. The superscript l denotes either the DSE or NSE. The sum is over the time series for property A. Θ_l is the Heaviside step function that, for protein L, is equal to $\Theta(5-\Delta(t))$ when l =NSE and $\Theta(5+\Delta(t))$ when l =DSE. For CspTm, $\Theta(13.5-\Delta(t))$ is used when l =NSE and $\Theta(13.5+\Delta(t))$ when l =DSE.

The free energy surface, plotted in terms of E_P and Δ, for the protein CspTm at [C]=0, shows a basin of attraction that is intermediate between the NBA and the DSE (see Fig. 2.4E). From a structural perspective, the population of this intermediate represents disorder in one of the β-strands (green strand in Fig. 2.1A) with the rest of the native structure intact as described in Chapter 2. Because the experiments analyze denaturant-induced transitions in CspTm using a two-state approximation we included the structures in the intermediate as a part of the NBA. In order to determine the value of Δ (denoted Δ_B) that gives boundary between the NBA and the DSE we solved $\int_0^{\Delta_B} P(\Delta, T_m)d\Delta = 0.5$, where T_m is the temperature at which the specific heat is a maximum (Fig. 2.4). For CspTm we obtain $\Delta_B = 13.5$ Å, which is large only because the NBA includes the native-like intermediate. Except for the disruption of the strand shown in green in Fig. 2.1A, the rest of the structure is native-like. We obtain a much smaller value for protein L which is much better described as a two-state folder. We should emphasize that in obtaining these values no fit to experimental data was made.

Distribution functions $P(R_{ee}^{DSE})$ and $P(R_g^{DSE})$: The finding that $\overline{R_g^{DSE}} \sim a_D([C], T)N^{\nu}$ [68] ($\nu \sim 0.6$) implies that proteins can be described as random coils at high denaturant concentrations. In order to show that the conformations are solely determined by excluded volume interactions between the monomers, as implied by the Flory scaling, it is also important to analyze the distribution $P(R_{ee}^{DSE})$ of the end-to-end distance R_{ee}^{DSE}. For a self-avoiding polymer (negligible intrapeptide attractive

interactions) $P(y)$ ($y = R_{ee}^{DSE}/\overline{R_{ee}^{DSE}}$) acquires a universal shape given by,

$$P(y) = c_1 y^{2+\theta} exp(-c_2 y^{1/(1-\nu)}) \tag{A.10}$$

where the des Clouieax exponent $\theta = (\gamma - 1)/\nu$ with γ being the susceptibility exponent. The constants c_1 and c_2 are determined using the conditions $\int P(y)dy = \int y^2 P(y)dy = 1$ [187]. We also expect Eq. A.10 to be satisfied for $y = R_g^{DSE}/\overline{R_g^{DSE}}$ because only one length scale, namely the size of the protein, determines the distribution function at high [C]. However, in practice we find for a self-avoiding polymer (N. Toan, unpublished) and for proteins (Fig. A.3) that Eq. A.10 is not as accurate for $P(R_g^{DSE})$ as it is for $P(R_{ee}^{DSE})$.

Table A.1: Solvent accessibility of the backbone and side chain groups of residue k in the tripeptide $Gly - k - Gly$ ($\alpha_{k,Gly-k-Gly}$)

	$\alpha_{k,Gly-k-Gly}$ (Å^2)	
k	Backbone	Side chain
Ala	62.5336	108.259
Met	50.2648	164.683
Arg	46.1820	185.982
Gln	52.0904	155.429
Asn	55.6039	138.647
Gly	84.9817	0.000
Tyr	47.3327	179.916
Asp	56.6455	133.722
Trp	43.7780	198.715
Phe	48.3400	174.605
Cys	57.7220	128.640
Pro	56.8578	132.713
Lys	48.3400	174.605
Hsd[a]	51.3509	159.160
Hse	51.3509	159.160
Hsp	51.3509	159.160
Ser	60.9227	114.940
Thr	56.2249	135.721
Val	53.8055	147.128
Ile	50.2648	164.683
Glu	53.0339	150.786
Leu	50.2648	164.683

[a]Hsd - Neutral histidine, proton on ND1 atom. Hse - Neutral histidine, proton on NE2 atom. HSP - Protonated histidine.

Table A.2: Values of m_k, b_k, and m_{BB} and b_{BB} (Eqs. A.4-A.5).

	Osmolyte							
Residue	GdmCl	urea	betaine	proline	sucrose	sarcosine	sorbitol	TMAO
Gly	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Ala	-7.20(-2.28)	-4.69	4.77	-0.07	22.05	10.91	16.57	-14.64
Val	-41.77(-23.41)	-21.65	-19.63	7.96	33.92	29.32	24.65	-1.02
Leu	-75.99(-41.42)	-54.57	-17.73	4.77	37.11	38.33	39.07	11.62
Ile	-68.10(-37.93)	-38.43	-1.27	-2.72	28.12	39.98	36.90	-25.43
Met	-85.86(-42.76)	-48.34	-14.16	-35.12	-6.66	8.18	20.97	-7.65
Phe	-124.57(-61.12)	-83.11	-112.93	-71.26	-96.35	-12.64	26.38	-9.32
Pro	-50.86(-27.76)	-17.65	-125.16	-63.96	-73.02	-34.23	-4.48	-137.73
Ser	-18.75(-31.25)	-20.56	-41.85	-33.49	-2.79	-27.98	-1.58	-39.04
Thr	-18.75(-31.25)	-22.09	0.33	-18.33	20.82	-7.54	13.20	3.57
Asn	-102.03(-65.73)	-38.79	33.17	-17.71	-28.28	-40.93	-21.21	55.69
Hln	-56.90(-57.07)	-54.81	7.57	-32.26	-40.87	-10.19	-23.98	41.41
Tyr	-123.19(-78.71)	-45.08	-213.09	-138.41	-78.41	-26.37	-53.50	-114.32
Trp	-196.25(-138.75)	-141.46	-369.93	-198.37	-215.27	-113.03	-67.23	-152.87
Asp	-102.03(-65.73)	3.55	-116.56	-90.51	-37.17	-14.20	-83.88	-66.67
Glu	-56.90(-57.07)	0.62	-112.08	-89.17	-41.65	-12.61	-70.05	-83.25
His	-65.00(-85.00)	-50.51	-35.97	-45.10	-118.66	-20.80	-42.45	42.07
Hsd	-65.00(-85.00)	-50.51	-35.97	-45.10	-118.66	-20.80	-42.45	42.07
Lys	-67.95(0.00)	-22.76	-171.99	-59.87	-39.60	-27.42	-32.47	-110.23
Arg	42.34(0.00)	-21.17	-109.45	-60.18	-79.32	-32.24	-24.65	-109.27
BB	-39.21(-31.86)	-39.00	67.00	48.00	62.00	52.00	35.00	90.00

Table A.3: Parameters used in C_α-SCM (Eqs. A.1-A.3).

Parameter	value[a]
K_A	30
K_{D_1}	0.70
K_{D_2}	0.00
K_{D_3}	0.35
n_1[b]	1
n_2	2
n_3	3
K_{Ch}	18.013 (25.73)[c]
ϵ_{HB}	0.75 (1.5)[d]
ϵ_i^N	NB[e]
ϵ_i^{NN}	10^{-12}

[a]The basic unit of energy is $kcal/mol$.

[b]n_j is the dimensionless period of the cosine function of Eq. A.1.

[c]For protein L $K_{Ch} = 18.013\ kcal\ mol^{-1}\ degree^{-2}$. For CspTm $K_{Ch} = 25.73\ kcal\ mol^{-1}\ degree^{-2}$

[d]Residue pairs that make just one backbone hydrogen bond are assigned an $\epsilon_{HB} = 0.75$. For pairs that make two hydrogen bonds $\epsilon_{HB} = 1.5$.

[e]The statistical potential of Miyazawa-Jernigan [185] formed the basis for choosing ϵ_i^N values for $SC - SC$ interactions. Values reported in Table 5 of [185] were subtracted by 1.2 so that all pair energies would be negative. To obtain protein melting temperatures above 300 K we scaled the resultant values by multiplying them by 0.7, in the case of protein L, and by 1.0 in the case of CspTm. The resulting values were assigned to ϵ_i^N based on the amino acids forming the native contact. For $C_\alpha - SC$ $\epsilon_i^N = 0.37\ kcal\ mol^{-1}$

Table A.4: van der Waals radius of the side chain beads for various amino-acids based on measured partial molar volumes [1].

Residue	Radius (Å)
Ala	2.52
Cys	2.74
Asp	2.79
Glu	2.96
Phe	3.18
Gly	2.25
Hsd [a]	3.04
Ile	3.09
Lys	3.18
Leu	3.09
Met	3.09
Asn	2.84
Pro	2.78
Gln	3.01
Arg	3.28
Ser	2.59
Thr	2.81
Val	2.93
Trp	3.39
Tyr	3.23

[a]The same value of the radius was used regardless of the protonation state.

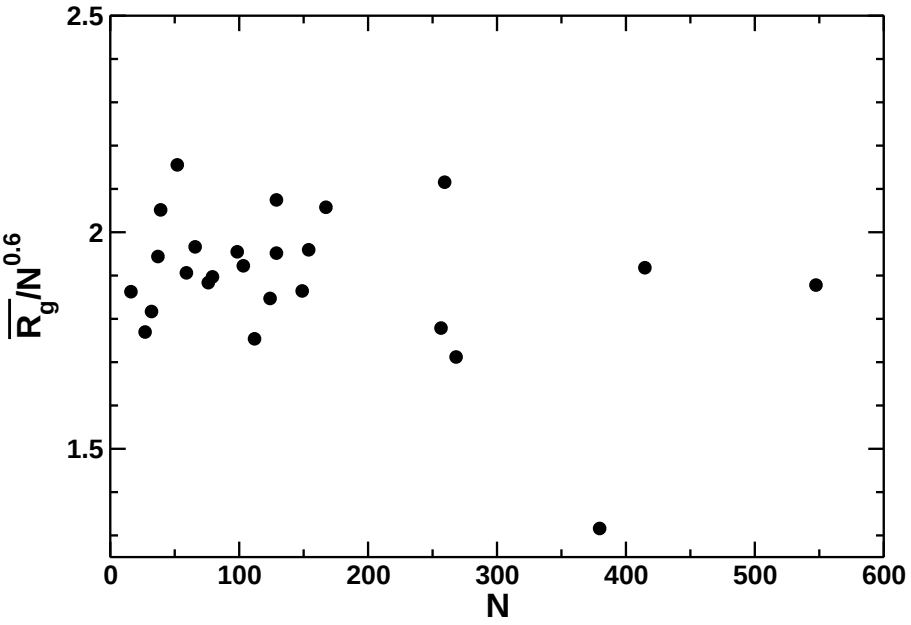


Figure A.1: A log-log plot of $\overline{R_g}$ versus N for a number of proteins confirms the predictions of the Flory theory [68]. The average Kuhn length was found to be $a_D \sim 2$Å [68], which gives the impression that a_D is independent of denaturant concentration, pH, temperature etc. If the universal aspect of Flory theory is obeyed then the effective Kuhn length for every protein can be extracted using $a_D([C], T) = \overline{R_g}/N^{\nu}$. The plot here shows the effective Kuhn length as a function of N for proteins for which $\overline{R_g}$ are listed in [68]. The dispersion seen here is within the range for protein L and CspTm (see Inset of Fig. 2.2). Although the dispersion in $a_D([C], T)$ is relatively small, it can result in significant errors when computing absolute values of $\overline{R_g}$ for a given N. The changes in $a_D([C], T)$ have to be taken into account when obtaining accurate values of $\overline{R_g}$ from SAXS or FRET experiments.

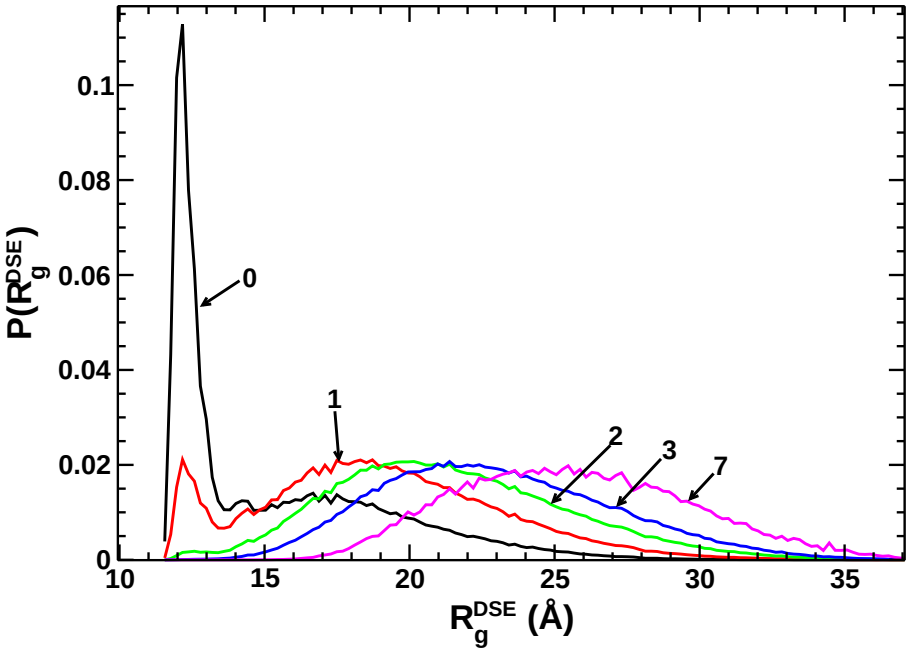


Figure A.2: Distribution of $P(R_g^{DSE})$ for protein L at various GdmCl concentrations at 328 K. The concentration ([C]) of GdmCl in molar units are shown in the curves. As [C] decreases, the maximum value of R_g^{DSE} sampled decreases.

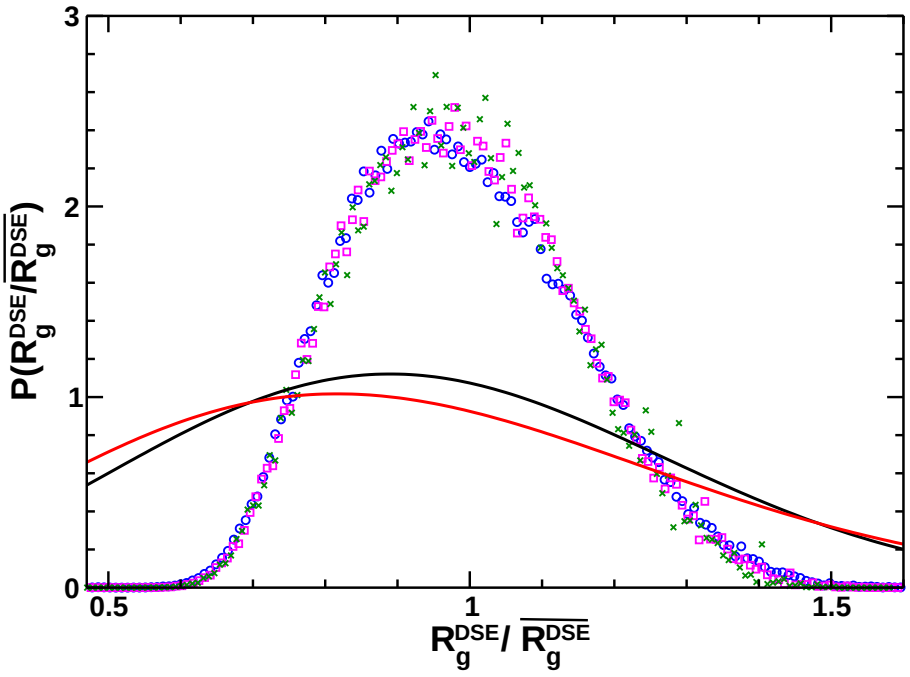


Figure A.3: The DSE distribution $P(R_g/\overline{R_g})$ for protein L in 5, 7, and 9 M GdmCl at 328 K. The solid black line represents the expected universal shape for a self avoiding polymer. In addition, we plot the universal shape for a Gaussian chain shown as the solid red line. For a Gaussian polymer $P(y) = c_1 y^2 exp(-c_2 y^2)$, where c_1 and c_2 are 4.2 and 1.5 respectively. Thus, only at high [C] values the random coil nature of proteins is manifested. In general [C] has to be far greater than $[C_m] + 0.5\Delta C_m$ where ΔC_m is width of the transition region. For protein L $C_m = 2.4$ M, $\Delta C_m = 1$ M, obtained from the derivative of f_{NBA} with respect to [C].

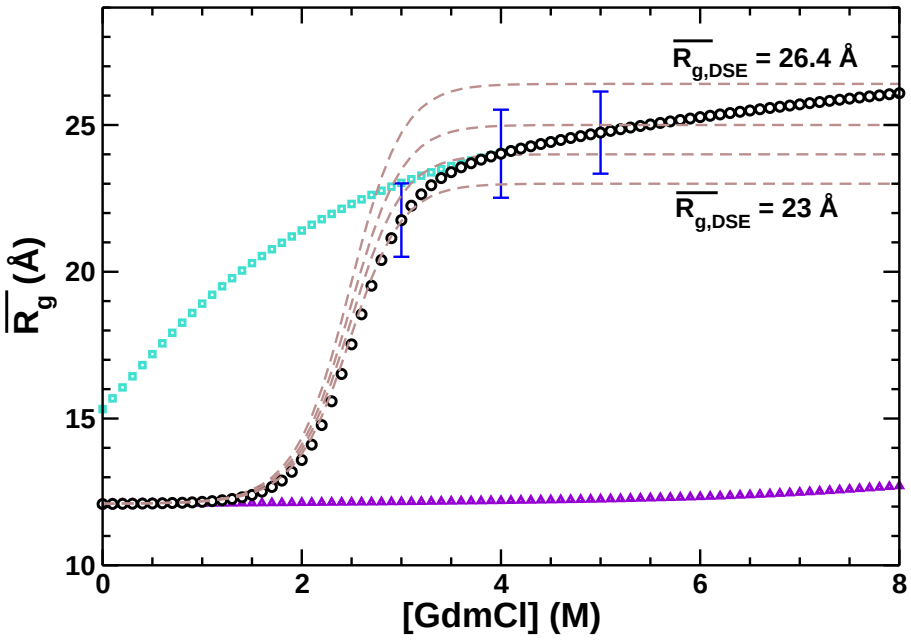


Figure A.4: Computation of $\overline{R_g}$ protein L versus GdmCl concentration at 328 K. Notation is the same as Fig. 2.2C. Gray lines represent predictions using the two state equation $\overline{R_g} = f_{NBA}\overline{R_g^{NSE}} + (1 - f_{NBA})\overline{R_g^{DSE}}$, where f_{NBA} is taken from Fig. 2.1B, and $\overline{R_g^{NSE}}$ = 12.1 Å. In principle, $\overline{R_g^{DSE}}$ should depend on [C]. The gray lines are constructed using different $\overline{R_g^{DSE}}$ values (= 23.0, 24.0, 25.0, 26.4 Å). The black circle is obtained using two-state assumption, and the [C]-dependent $\overline{R_g^{DSE}}$ (in cyan), and $\overline{R_g^{NSE}}$ (purple). The error bars at 3, 4, and 5 M correspond to the experimental error associated with SAXS measurements of $\overline{R_g}$ in [65]. The plots show that if the variations in $\overline{R_g^{DSE}}$ with [C] are not taken into account the exact simulation results (black circles) cannot be reproduced.

Chapter B

Appendix for Chapter 6

Computational Methods:

Protein Model. We use the coarse graining procedure introduced by Honeycutt and Thirumalai (HT) [180] to model the polypeptide chain. In the HT model, each amino-acid is represented by one bead located at the C_α-carbon position along the protein backbone. A three letter code is used to classify the twenty naturally occurring amino acids; L for hydrophilic residues, B for hydrophobic residues and N for neutral residues. The potential energy of a conformation of a polypeptide with M residues and coordinates $r_i (i = 1, 2, ..., M)$ in the HT representation is

$$V = V_B + V_A + V_D + V_{NB} + V_{HB} \tag{B.1}$$

where V_B accounts for chain connectivity between residues i and $i+1$, and is given by

$$V_B = \frac{K_b}{2} \sum_{i}^{M-1} (r_\circ - |r_i - r_{i+1}|)^2. \tag{B.2}$$

The equilibrium distance between neighboring C_α atoms $r_\circ = 3.81$ Å, and the spring constant $K_b = 100\,\mathrm{kcal/Å^2}$. The term V_A restricts the bond angle formed by residues $i, i+1, i+2$ using

$$V_A = \frac{K_\theta}{2} \sum_{i}^{M-2} (\theta_\circ - \theta_i)^2, \tag{B.3}$$

where $\theta_\circ = 105°$, and $K_\theta = 12.5\ kcal/rad^2$. The local secondary structure preferences of the backbone are accounted for by the dihedral angle potential,

$$V_D = \sum_{i}^{M-3} \sum_{j}^{3} A_j(1 + cos(n_j\phi_i - \delta_j)), \tag{B.4}$$

where A_j is a constant, n_j is the period, ϕ_i is the dihedral angle defined by residues $i, i{+}1, i{+}2, i{+}3$ and δ_j is the phase shift. The parameters used in Eq. B.4 are listed in Table B.1. For sequence AS_1 (see below) we use the Guo-Thirumalai [183] parameter set (see Table B.1), and for all other sequences we devised a new parameter set. The non-bonded potential V_{NB} between non-covalently linked ($|i-j| \geq 4$) residues is taken to be,

$$V_{NB} = \sum_{i=1}^{M-4} \sum_{j=i+4}^{M} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right]. \tag{B.5}$$

For all sequences, except AS_3, PA and PN (see below), if i and j are both hydrophobic (B) then $\epsilon_{ij} = 2.125\ kcal/mol$ (see Table 6.1), and $\sigma_{ij} = 3.8$ Å. For AS_3, PA and PN $\epsilon_{ij} = 0.5\ kcal/mol$. When i is type L or N and j is either B, L or N then $\epsilon_{ij} = 10^{-12}\ kcal/mol$ and $\sigma_{ij} = 40.47$ Å, which approximates a short range repulsive interaction (Fig. B.1A).

The potential for the backbone-backbone hydrogen bond is taken to be

$$V_{HB} = 4\epsilon_{HB} \sum_{i}^{M-3} \left[\left(\frac{\sigma_{HB}}{r_{i,i+3}} \right)^{12} - \left(\frac{\sigma_{HB}}{r_{i,i+3}} \right)^{6} \right], \tag{B.6}$$

where $\sigma_{HB} = 4.63$ Å, which results in an r_{min} of 5.2 Å, corresponding to the distance between the the C_α-atoms of residues i and $i+3$ in an α-helix. Hydrogen bond potential V_{HB}, is used only between residues separated by 3 covalent bonds. Except

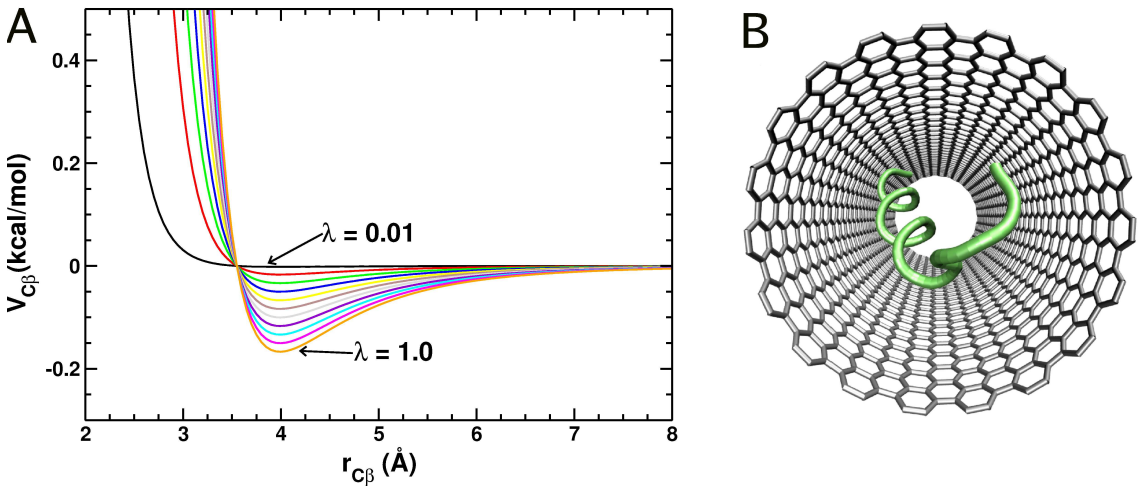


Figure B.1: (A) The strength of the hydrophobic effect between the nanotube and peptide is varied by scaling the Lennard-Jones interaction (see Eq. B.7) between the nanotube atoms and peptide residues. Eleven different interaction strengths were studied, ranging from $\lambda = 1$ (strongly hydrophobic) to $\lambda = 0.01$ (weakly hydrophobic). (B) A peptide in a carbon nanotube.

for AS_1, for which $\epsilon_{HB} = 0$, we adjust the value of ϵ_{HB} such that, in bulk, the probability of being helical at 300 K is between 40-50% for all sequences.

Peptide sequences: The role of sequence in determining confinement-induced stability probed using PA, PN, and AS. We consider three variations of the amphiphilic sequence AS that are generated by modifying the parameters used in the energy function (Eq. B.1). In the first parameter set, that generates the sequence AS_1, the original Guo-Thirumalai (GT) model [183] was used. We modified the GT dihedral potential (see Table B.1), and set $\epsilon_{HB} = 1.75\ kcal/mol$ (see Table 6.1) to generate AS_2. The sequence AS_3 is the same as AS_2 except $\epsilon_{BB} = 0.5\ kcal/mol$, and $\epsilon_{HB} = 2.75\ kcal/mol$. The changes in ϵ_{BB} and ϵ_{HB} can account for variations in external conditions (eg. pH, cosolvents, or temperature). Because these changes

can potentially alter the nature of the denatured state ensemble, comparison of the results for AS_2 and AS_3 allows us to assess the role external solution conditions play in affecting helix stability upon confinement. The PA and PN sequences also use a variation of the GT potential, that includes the modified dihedral angle potential, and non-zero ϵ_{HB}. Additional details on the sequences, and the associated values of the parameters are in Table 6.1.

Carbon Nanotube Confinement and Initial Conditions: Confinement of the polypeptides in infinitely long single walled carbon nanotubes is carried out by using one-dimensional periodic boundary conditions. Lennard-Jones parameters for the nanotube are taken from Steel and coworkers [188]. The primary cell is 75.975 Å in length. The diameters of the six nanotubes are 35.3, 29.8, 25.8, 20.3, 17.6, and 14.9 Å with graphite lattice indices [189] of (26,26), (22,22), (19,19), (15,15), (13,13) and (11,11), respectively. Initial structures are prepared by randomly inserting a denatured peptide conformation (prepared in bulk simulations at 600 K) into the nanotube, and retaining only those conformations that fit within the volume enclosed by the nanotube (see Fig. B.1B). The starting structures are heated inside the nanotube at 600 K for 3.8 ns, and then equilibrated at 300 K for 75 ns. At least five independent trajectories are generated for each sequence, at each D and λ (see below), resulting in a total of 1,353 independent trajectories. The duration of each trajectory is 675 ns.

Peptide-nanotube interactions: The hydrophobic interaction between the nanotube and the polypeptide (Fig. B.1A) is modeled using the Lennard-Jones potential between peptide residues and the carbon atoms in the nanotube (Fig. B.1B). At

$\lambda = 0.01$ (Fig. B.1B), there is no net attractive interaction between peptide hydrophobic residues and the carbon nanotube. The interactions between polar and neutral protein residues and the nanotube are purely repulsive. We vary the strength of the interaction between the polypeptide and the nanotube using λ, which scales the Lennard-Jones interactions between the nanotube atoms (C) and the peptide (β) residues,

$$V_{C\beta} = 4\epsilon_{C\beta}\lambda \left[\left(\frac{\sigma_{C\beta}}{r_{C\beta}} \right)^{12} - \delta \left(\frac{\sigma_{C\beta}}{r_{C\beta}} \right)^{6} \right], \tag{B.7}$$

For all $\beta =$ B, L or N, $\epsilon_{C\beta} = 0.345\ kcal/mol$, and $\sigma_{C\beta} = 3.6$ Å (see Fig. B.1). For $\beta =$ B the value of δ is one, and is zero otherwise.

We also examine the effect of chemical heterogeneity in the nanotube by creating different size 'hydrophobic patches' which run parallel to the long axis of the nanotube (Fig. 6.6). Each hydrophobic patch has a width that is equal to the number of rows of nanotube atoms that make up the patch. Nanotube atoms within the patch interact with protein atoms as defined in Eq. B.7. All other interactions involving the nanotube atoms are purely repulsive. As a quantitative measure of the size of the patch, we use the fraction of surface area, f_H ($0 < f_H \leq 1$), of the hydrophobic patch along the interior of the nanotube.

Computation of entropy of DSE and HB: The volume fractions accessible to the center-of-mass of the helical ensemble (α_{HB}) and denatured ensembles (α_{DSE}) are computed from the time-series of the saved peptide structures. The accessible volume fraction, as a function of D, is calculated separately for the HB and DSE using the Widom particle insertion method [190, 191]. If all interactions are

Table B.1: Parameters in the dihedral angle potential, $V_D = \sum_i \sum_j A_j(1+cos(n_j\phi_i - \delta_j))$

Potential	j^a	A_j	n_j	${\delta_j}^b$
GT^c	1	2.00	3	-112.5
	2	1.53	1	0.0
	3	2.47	0	0.0
Modified	1	0.90	3	-66.5
	2	2.27	2	-68.1
	3	2.91	1	-37.3

[a]Notation described following Eq. 4.
[b]In degrees.
[c]Taken from Guo and Thirumalai [183].

hard core then the volume fraction accessible to a given peptide conformation is the ratio of successful peptide insertions divided by the number of attempts. A successful insertion is one in which no peptide residues overlap with the nanotube atoms upon randomly placing the center of mass of the peptide within the nanotube and randomly orienting the peptide. The overlap occurs when a carbon nanotube atom is within 2.75 Å of a peptide residue. The overlap criteria is based on the Lennard-Jones potential between peptide residues and the nanotube atoms used in our simulations at $\lambda = 0.01$ (Fig. B.1B). We computed the accessible volume, for a given peptide conformation, from the number of insertion attempts necessary to attain 2,500 successful insertions; α_{HB} and α_{DSE} are simple averages over all the accessible volume fractions of the peptide conformations in the HB and DSE, respectively. Computation of α_{HB} and α_{DSE} allows us to estimate the entropy changes upon confinement (see Eq. 6.1).

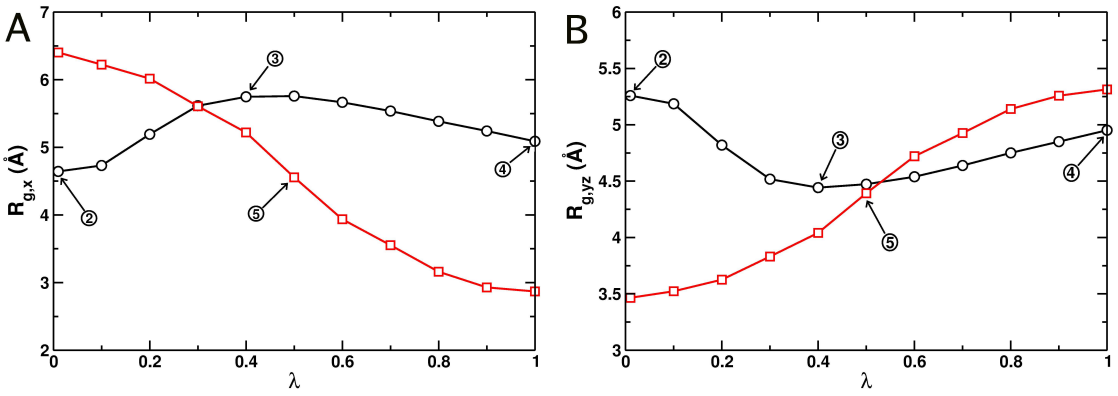


Figure B.2: The radius-of-gyration for PA as a function of λ in two different diameter nanotubes, $D = 35.3$ Å (black circles) and $D = 20.3$ Å (red squares). Decomposing R_g in one ($R_{g,x}$) and two-dimensions ($R_{g,yz}$) shows that, in a nanotube with $D =$ 35.3 Å, as λ is increased to a value of 0.4 (point 3), the peptide binds to the interior surface of the nanotube and aligns preferentially along the x-axis (A), leading to a decrease in $R_{g,yz}$ (see point 3 in (B)). When D =20.3 Å, increasing λ causes the peptide to contract along the long axis of the nanotube (x-axis) (A), due to the concomitant binding of the peptide perpendicular to the long-axis. As a result the peptide expands along $R_{g,yz}$ (B, see also point 5 of Fig. 6.5A).

Bibliography

[1] A. A. Zamyatnin. Amino-acid, peptide, and protein volume in solution. *Ann. Rev. Biophys. Bioeng.*, 13:145–165, 1984.

[2] R. B. Simpson and W. Kauzmann. The kinetics of protein denaturation .1. the behavior of the optical rotation of ovalbumin in urea solutions. *J. Am. Chem. Soc.*, 75(21):5139–5152, 1953.

[3] A. R. Fersht. *Structure and Mechanism in Protein Science: A guide to enzyme catalysis and protein folding*. W. H. Freeman and Company, New York, 2nd edition, 1999.

[4] C. Anfinsen and E. Haber. Studies on reduction and re-formation of protein disulfide bonds. *J. Biol. Chem.*, 236(5):1361–1363, 1961.

[5] C. B. Anfinsen. Principles that govern folding of protein chains. *Science*, 181(4096):223–230, 1973.

[6] N. Go. Theoretical-studies of protein folding. *Ann. Rev. Biophys. Bioeng.*, 12:183–210, 1983.

[7] A. M. Labhardt and R. L. Baldwin. Recombination of s-peptide with s-protein during folding of ribonuclease-s .1. folding pathways of the slow-folding and fast-folding classes of unfolded s-protein. *J. Molec. Biol.*, 135(1):231–244, 1979.

[8] D. Wetlaufer. Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc. Natl. Acad. Sci. USA*, 70:691–701, 1973.

[9] K. A. Dill and H. S. Chan. From levinthal to pathways to funnels. *Nat. Struc. Biol.*, 4(1):10–19, 1997.

[10] D.K. Klimov and D. Thirumalai. Deciphering the timescales and mechanisms of protein folding using minimal off-lattice models. *Curr. Opin. Struc. Biol.*, 9(2):197–207, 1999.

[11] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes. Funnels, pathways, and the energy landscape of protein-folding - a synthesis. *Prot. Struc. Func. Gene.*, 21(3):167–195, 1995.

[12] H. Wako and N. Saito. Statistical mechanical theory of protein conformation .1. General considerations and application to homopolymers. *J. Phys. Soc. Jpn.*, 44(6):1931–1938, 1978.

[13] H. Wako and N. Saito. Statistical mechanical theory of protein conformation .2. folding pathway for protein. *J. Phys. Soc. Jpn.*, 44(6):1939–1945, 1978.

[14] P. H. Yancey, M. E. Clark, S. C. Hand, R. D. Bowlus, and G. N. Somero. Living with water-stress - evolution of osmolyte systems. *Science*, 217(4566):1214–1222, 1982.

[15] M. J. Gething and J. Sambrook. Protein folding in the cell. *Nature*, 355(6355):33–45, 1992.

[16] I. H. Madshus. Regulation of intracellular ph in eukaryotic cells. *Biochem. J.*, 250(1):1–8, 1998.

[17] A. Kurkdjian and J. Guern. Intracellular ph - measurement and importance in cell-activity. *Ann. Rev. Plant Physiol. Plant Molec. Biol.*, 40:271–303, 1989.

[18] F. A. Gallagher and et. al. Magnetic resonance imaging of pH in vivo using hyperpolarized c-13-labelled bicarbonate. *Nature*, 453(7197):940–944, 2008.

[19] F. Chiti and C. M. Dobson. Protein misfolding, functional amyloid, and human disease. *Ann. Rev. Biochem.*, 75:333–336, 2006.

[20] C. Tanford and P. K. De. Unfolding of beta-lactoglobulin at ph 3 by urea, formamide, and other organic substances. *J. Biol. Chem.*, 236(6):1711–, 1961.

[21] D. W. Bolen and M. M. Santoro. Unfolding free-energy changes determined by the linear extrapolation method .1. unfolding of phenylmethanesulfonyl alpha-chymotrypsin using different denaturants. *Biochemistry*, 27(21):8063–8068, 1988.

[22] M. M. Santoro, Y. F. Liu, S. M. A. Khan, L. X. Hou, and D. W. Bolen. Increased thermal-stability of proteins in the presence of naturally-occurring osmolytes. *Biochemistry*, 31(23):5278–5283, 1992.

[23] Y. Nozaki and C. Tanford. Solubility of amino acids and related compounds in aqueous urea solutions. *J. Biol. Chem.*, 238(12):4074, 1963.

[24] C. Tanford. Protein denaturation. part a and b. *Advan. Prot. Chem.*, 23:121–282, 1968.

[25] G. I. Makhatadze and P. L. Privalov. Protein interactions with urea and guanidinium chloride. *J. Molec. Biol.*, 226:491–505, 1992.

[26] C. Tanford. Isothermal unfolding of globular proteins in aqueous urea solutions. *J. Am. Chem. Soc.*, 86(10):2050, 1964.

[27] M. Auton and D. W. Bolen. Predicting the energetics of osmolyte-induced protein folding/unfolding. *Proc. Natl. Acad. Sci. USA*, 102(42):15065–15068, 2005.

[28] M. Auton and D. W. Bolen. Additive transfer free energies of the peptide backbone unit that are independent of the model compound and the choice of concentration scale. *Biochemistry*, 43(5):1329–1342, 2004.

[29] M. Auton and D. W. Bolen. Application of the transfer model to understand how naturally occuring osmolytes affect protein stability. *Meth. Enzym.*, 428:397–418, 2007.

[30] M. L. Groves, N. J. Hipp, and T. L. McMeekin. Ffect of ph on the denaturation of beta-lactoglobulin and its dodecyl sulfate derivative. *J. Amer. Chem. Soc.*, 73(6):2790–2793, 1951.

[31] C. Tanford, J. G. Buzzell, D. G. Rands, and S. A. Swanson. The reversible expansion of bovine serum albumin in acid solutions. *J. Amer. Chem. Soc.*, 77(24):6421–6428, 1955.

[32] C. Tanford. Protein denaturation: Part c. theoretical models for the mechanism of denaturation. *Adv. Prot. Chem.*, 24:1–95, 1970.

[33] J. Wyman. Linked functions and reciprocal effects in hemoglobin - a 2nd look. *Advan. Prot. Chem.*, 19(223):223, 1964.

[34] K. C. Aune and C. Tanford. Thermodynamics of the denaturation of lysozyme by guanidine hydrochloride. i. dependence on pH at 25○. *Biochemistry*, 8(11):4579–4585, 1969.

[35] D. Bashford and M. Karplus. Multiple-site titration curves of proteins - an analysis of exact and approximate methods for their calculation. *J. Phys. Chem.*, 95(23):9556–9561, 1991.

[36] Y. Sugita, A. Kitao, and Y. Okamoto. Multidimensional replica-exchange method for free-energy calculations. *J. Chem. Phys.*, 113(15):6042–6051, 2000.

[37] S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg. The weighted histogram analysis method for free-energy calculations on biomolecules .1. the method. *J. Comp. Chem.*, 13(8):1011–1021, 1992.

[38] M. S. Cheung, J. M. Finke, B. Callahan, and J. N. Onuchic. Exploring the interplay between topology and secondary structural formation in the protein folding problem. *J. Phys. Chem. B*, 107(40):11193–11200, 2003.

[39] U. H. E. Hansmann. Parallel tempering algorithm for conformational studies of biological molecules. *Chem. Phys. Lett.*, 281(1):140–150, 1997.

[40] A. M. Ferrenberg and R. H. Swendsen. Optimized monte-carlo data-analysis. *Phys. Rev. Lett.*, 63:1195–1198, 1989.

[41] D. Thirumalai, D. K. Klimov, and R. I. Dima. Emerging ideas on the molecular basis of protein and peptide aggregation. *Curr. Opin. Struc. Biol.*, 13(2):146–159, 2003.

[42] J. N. Onuchic and P. G. Wolynes. *Curr. Opin. Struc. Biol.*, 14(1):70–75, 2004.

[43] D. Thirumalai and C. Hyeon. RNA and protein folding: Common themes and variations. *Biochemistry*, 44(13):4957–4970, 2005.

[44] R. D. Schaeffer, A. R. Fersht, and V. Daggett. Combining experiment and simulation in protein folding: closing the gap for small model systems. *Curr. Opin. Struc. Biol.*, 18:4–9, 2008.

[45] R. L. Baldwin. Energetics of protein folding. *J. Molec. Biol.*, 371(2):283–301, 2007.

[46] C. N. Pace. Determination and analysis of urea and gunidine hydrochloride denaturation curves. *Methods in Enzymology*, 131:266–280, 1986.

[47] J. M. Scholtz, D. Barrick, E. J. York, J. M. Steward, and R. L. Baldwin. Urea unfolding of peptide helices as a model for interpreting protein unfolding. *Proc. Natl. Acad. Sci. USA*, 92(1):185–189, 1995.

[48] J. C. Lee and S. N. Timasheff. The stabilization of proteins by sucrose. *J. Biol. Chem.*, 256(14):7193–7201, 1981.

[49] D. R. Robinson and W. P. Jencks. Effect of compounds of urea-guanidinium class on activity coefficient of acetyltetraglycine ethyl ester and related compounds. *J. Am. Chem. Soc.*, 87(11):2462–2469, 1965.

[50] J. A. Schellman. Protein stability in mixed solvents: A balance of contact interaction and excluded volume. *Biophys. J.*, 85(1):108–125, 2003.

[51] P. R. Wills and D. J. Winzor. Thermodynamic analysis of "preferential solvation" in protein solutions. *Biopolymers*, 33:1627–1629, 1993.

[52] J. Klein-Seetharaman, M. Oikawa, S. B. Grimshaw, J. Wirmer, E. Duchardt, T. Ueda, T. Imoto, L. J. Smith, C. M. Dobson, and H. Schwalbe. Long-range interactions within a nonnative protein. *Science*, 295(5560):1719–1722, 2002.

[53] B. Schuler and W. A. Eaton. Protein folding studied by single-molecule fret. *Curr. Opin. Struc. Biol.*, 18:160–26, 2008.

[54] B. Schuler, E. A. Lipman, and W. A. Eaton. Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy. *Nature*, 419(17):743–747, 2002.

[55] E. Rhoades, M. Cohen, B. Schuler, and G. Haran. Two-state folding observed in individual protein molecules. *J. Am. Chem. Soc.*, 126(45):14686–14687, 2004.

[56] A. Hoffman, A. Kane, D. Nettels, D. E. Hertzog, P. Baumgartel, J. Lengefeld, G. Reichardt, D.A. Horsley, R. Seckler, O. Bakajin, and B. Schuler. Mapping protein collapse with single-molecule fluorescence and kinetic synchrotron radiation circular dichroism spectroscopy. *Proc. Natl. Acad. Sci. USA*, 104(1):105–110, 2007.

[57] J. D. Honeycutt and D. Thirumalai. Metastability of the folded states of globular-proteins. *Proc. Natl. Acad. Sci. USA*, 87(9):3526–3529, 1990.

[58] D. K. Klimov and D. Thirumalai. Linking rates of folding in lattice models of proteins with underlying thermodynamic characteristics. *J. Chem. Phys.*, 109(10):4119–4125, 1998.

[59] D. K. Klimov and D. Thirumalai. Mechanisms and kinetics of beta-hairpin formation. *Proc. Natl. Acad. Sci. USA*, 97(6):2544–2549, 2000.

[60] C. Clementi, H. Nymeyer, and J. N. Onuchic. Topological and energetic factors: What determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? an investigation for small globular proteins. *J. Molec. Biol.*, 298(5):937–953, 2000.

[61] D. Tobi, R. Elber, and D. Thirumalai. The dominant interaction between peptide and urea is electrostatic in nature: A molecular dynamics simulation study. *Biopolymers*, 68(3):359–369, 2003.

[62] M. Auton, L. M. F. Holthauzen, and D. W. Bolen. Anatomy of energetic changes accompanying urea-induced protein denaturation. *Proc. Natl. Acad. Sci. USA*, 104(39):15317–15322, 2007.

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

[64] 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. USA*, 104(5):1528–1533, 2007.

[65] D. E. Kim, C. Fisher, and D. Baker. A breakdown of symmetry in the folding transition state of protein l. *J. Molec. Biol.*, 298(5):971–984, 2000.

[66] D. Perl, C. Welker, T. Schindler, K. Schroder, M. A. Marahiel, R. Jaenicke, and F. X. Schmid. Conservation of rapid two-state folding in mesophilic, thermophilic and hyperthermophilic cold shock proteins. *Nature*, 5(3):229–235, 1998.

[67] K. W. Plaxco, I. S. Millett, D. J. Segel, S. Doniach, and D. Baker. Chain collapse can occur concomitantly with the rate-limiting step in protein folding. *Nature Struc. Biol.*, 6(6):554–556, 1999.

[68] J. E. Kohn, I. S. Millett, J. Jacob, B. Zagrovic, T. M. Dillon, N. Cingel, R. S. Dothager, S. Seifert, P. Thiyagarajan, T. R. Sosnick, M. Z. Hasan, V. S. Pande, I Ruczinski, S. Doniach, and K. W. Plaxco. Random-coil behavior and the dimensions of chemically unfolded proteins. *Proc. Natl. Acad. Sci. USA*, 101(34):12491–12496, 2004.

[69] S. E. Jackson and A. R. Fersht. Folding of chymotrypsin inhibitor-2 .1. evidence for a 2-state transition. *Biochemistry*, 30(43):10428–10435, 1991.

[70] L. Pradeep and J. B. Udgaonkar. Osmolytes induce structure in an early intermediate on the folding pathway of barstar. *J. Biol. Chem.*, 279(39):40303–40313, 2004.

[71] 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, 2007.

[72] R. I. Dima and D. Thirumalai. Asymmetry in the shapes of folded and denatured states of proteins. *J. Phys. Chem. B*, 108(21):6564–6570, 2004.

[73] K. W. Plaxco, I. S. Millett, D. J. Segel, S. Doniach, and D. Baker. Chain collapse can occur concomitantly with the rate-limiting step in protein folding. *Nat. Struc. Biol.*, 6(6):554–556, 1999.

[74] M. Arai, E. Kondrashkina, C. Kayatekin, C. R. Matthews, M. Iwakura, and O. Bilsel. Microsecond hydrophobic collapse in the folding of escherichia coli dihydrofolate reductase an alpha/beta-type protein. *J. Molec. Biol.*, 368:219–229, 2007.

[75] M. E. Fisher. Shape of a self-avoiding walk or polymer chain. *J. Chem. Phys.*, 44:616–622, 1966.

[76] J. Shea, Y. D. Nochomovitz, Z. Guo, and C. L. Brooks. *J. Chem. Phys.*, 109(7):2895–2903, 1998.

[77] G. I. Makhatadze. Thermodynamics of protein interactions with urea and guanidinium hydrochloride. *J. Phys. Chem. B*, 103(23):4781–4785, 1999.

[78] J. M. Fernandez and H. Li. Force-clamp spectroscopy monitors the folding trajectory of a single protein. *Science*, 303(5664):1674–1678, 2004.

[79] A. Irback, S. Mitternacht, and S. Mohanty. Dissecting the mechanical unfolding of ubiquitin. *Proc. Natl. Acad. Sci. USA*, 102(38):13427–13432, 2005.

[80] C. Hyeon and D. Thirumalai. Measuring the energy landscape roughness and the transition state location of biomolecules using single molecule mechanical unfolding experiments. *J. Phys-Cond. Matter*, 19(11), 2007.

[81] D. Thirumalai and G. H. Lorimer. Chaperonin-mediated protein folding. *Ann. Rev. Biophys. Biomolec. Struc.*, 30:245–269, 2001.

[82] Z. Lin and H. S. Rye. Expansion and compression of a protein folding intermediate by GroEL. *Molec. Cell*, 16(1):23–34, 2004.

[83] Z. Lin, D. Madan, and H. S. Rye. GroEL stimulates protein folding through forced unfolding. *Nat. Struc. Molec. Biol.*, 15(3):303–311, 2008.

[84] Monitoring protein conformation along the pathway of chaperonin-assisted folding. *Cell*, 133(1):142–153, 2008.

[85] J. A. Kenniston, T. A. Baker, and J. M. Fernandez and. Linkage between atp consumption and mechanical unfolding during the protein processing reactions of an aaa(+) degradation machine. *Cell*, 114(4):511–520, 2003.

[86] S. H. Huang, K. S. Ratliff, and A. Matouschek. Protein unfolding by the mitochondrial membrane potential. *Nat. Struc. Biol.*, 9(4):301–307, 2002.

[87] D. R. Robinson and W. P. Jencks. Effect of compounds of urea-guanidinium class on activity coefficient of acetyltetraglycine ethyl ester and related compounds. *J. Am. Chem. Soc.*, 87(11):2462–2470, 1965.

[88] S. N. Timasheff and G. F. Xie. Preferential interactions of urea with lysozyme and their linkage to protein denaturation. *Biophys. Chem.*, 105:421–448, 2003.

[89] C. Tanford and R. Roxby. Interpretation of protein titration curves - application to lysozyme. *Biochemistry*, 11(11):2192–2198, 1972.

[90] A. S. Yang and B. Honig. On the ph-dependence of protein stability. *J. Molec. Biol.*, 231(2):459–474, 1993.

[91] J. H. Cho, S. Sato, J. C. Horng, B. Anil, and D. P. Raleigh. Electrostatic interactions in the denatured state ensemble: Their effect upon protein folding and protein stability. *Arch. Biochem. Biophys.*, 496(1):20–28, 2008.

[92] C. L. Chyan, F. C. Lin, H. B. Peng, J. M. Yuan, C. H. Chang, S. H. Lin, and G. L. Yangy. Reversible mechanical unfolding of single ubiquitin molecules. *Biophys. J.*, 87(6):3995–4006, 2004.

[93] Y. Cao and H. Li. How do chemical denaturants affect the mechanical folding and unfolding of proteins? *J. Molec. Biol.*, 375:316–324, 2008.

[94] C. Hyeon and D. Thirumalai. Mechanical unfolding of rna hairpins. *Proc. Natl. Acad. Sci. USA*, 102(19):6789–6794, 2005.

[95] E. P. O'Brien, G. Ziv, G. Haran, B. R. Brooks, and D. Thirumalai. Denaturant and osmolyte effects on proteins are accurately predicted using the molecular transfer model. *Proc. Natl. Acad. Sci. USA*, 2008.

[96] S. E. Jackson, M. Moracci, N. Elmasry, C. M. Johnson, and A. R. Fersht. Effect of cavity-creating mutations in the hydrophobic core of chymotrypsin inhibitor-2. *Biochemistry*, 32(42):11259–11269, 1993.

[97] D. Khare, P. Alexander, J. Antosiewicz, P. Bryan, M. Gilson, and J. Orban. pK(a) measurements from nuclear magnetic resonance for the b1 and b2 immunoglobulin g-binding domains of protein g: Comparison with calculated values for nuclear magnetic resonance and x-ray structures. *Biochemistry*, 36(12):3580–3589, 1997.

[98] S. Lindman, S. Linse, F. A. A. Mulder, and I. Andre. pK(a) values for side-chain carboxyl groups of a pgb1 variant explain salt and pH-dependent stability. *Biophys. J.*, 92(1):257–266, 2007.

[99] Y. J. Tan, M. Oliveberg, B. Davis, and A. R. Fersht. Perturbed pk(a)-values in the denatured states of proteins. *J. Molec. Biol.*, 254(5):980–992, 1995.

[100] J. E. Leffler. Parameters for the description of transition states. *Science*, 117(3039):340–341, 1952.

[101] J. K. Myers, C. N. Pace, and J. M. Scholtz. Denaturant m values and heat capacity changes: Relation to changes in accessible surface areas of protein unfolding. *Prot. Sci.*, 4:2138–2148, 1995.

[102] C. A. McPhalen and M. N. James. Crystal and molecular structure of the serine proteinase inhibitor CI-2 from barley seeds. *Biochemistry*, 26:261–269, 1987.

[103] A. M. Gronenborn, D. R. Filpula, N. Z. Essig, A. Achari, M. Whitlow, P. T. Wingfield, and G. M. Clore. A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein g. *Science*, 253:657–661, 1991.

[104] M. Kouza, C. K. Hu, and M. S. Li. New force replica exchange method and protein folding pathways probed by force-clamp technique. *J. Chem. Phys.*, 128:045103, 2008.

[105] T. Veitshans, D. Klimov, and D. Thirumalai. Protein folding kinetics: Timescales, pathways and energy landscapes in terms of sequence-dependent properties. *Fold. Des.*, 2(1):1–22, 1997.

[106] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. *J. Comp. Chem.*, 4:187–217, 1983.

[107] S. E. Jackson. How do small single-domain proteins fold? *Folding and Design*, 3:R81–R91, 1998.

[108] D. K. Klimov and D. Thirumalai. Symmetric connectivity of secondary structure elements enhances the diversity of folding pathways. *J. Molec. Biol.*, 353:1171–1186, 2005.

[109] G. Haran. Single-molecule fluorescence spectroscopy of biomolecular folding. *J. Phys. Cond. Mat.*, 15:R1291–R1317, 2003.

[110] A. A. Deniz, T. A. Laurence, G. S. Beligere, M. Dahan, A. B. Martin, D. S. Chemla, P. E. Dawson, P. G. Schultz, and S. Weiss. Single-molecule protein folding: Diffusion fluorescence resonance energy transfer studies of the denaturation of Chymotrypsin Inhibitor 2. *Proc. Natl. Acad. Sci. USA*, 97(10):5179–5184, 2000.

[111] A. Navon, V. Ittah, P. Landsman, H. A. Scheraga, and E. Haas. Distributions of intramolecular distances in the reduced and denatured states of bovine pancreatic ribonuclease a. folding initiation structures in the c-terminal portions of the reduced protein. *Biochemistry*, 40(1):105–118, 2001.

[112] K. K. Sinha and J. B. Udgaonkar. Dependence of the size of the initially collapsed form during the refolding of barstar on denaturant concentration: evidence for a continuous transition. *J. Molec. Biol.*, 353:704–718, 2005.

[113] E. V. Kuzmenkina, C. D. Heyes, and G. U. Nienhaus. Single-molecule forster resonance energy transfer study of protein dynamics under denaturing conditions. *Proc. Natl. Acad. Sci. USA*, 102(43):15471–15476, 2005.

[114] A. M. Saxena, J. B. Udgaonkar, and G. Krishnamoorthy. Characterization of intra-molecular distance and site-specific dynamics in chemically unfolded barstar: Evidence for denaturant-dependent non-random structure. *J. Molec. Biol.*, 359:174–189, 2006.

[115] G. U. Nienhaus. Exploring protein structure and dynamics under denaturing conditions by single-molecule fret analysis. *Macro. Biosci.*, 6(11):907–922, 2006.

[116] I. V. Gopich and A. Szabo. Single-macromolecule fluorescence resonance energy transfer and free-energy profiles. *J. Phys. Chem. B*, 107(21):5058–5063, 2003.

[117] C. Hyeon, G. Morrison, and D. Thirumalai. Force-dependent hopping rates of RNA hairpins can be estimated from accurate measurement of the folding landscapes. *Proc. Natl. Acad. Sci*, 105:9604–9609, 2008.

[118] C. Magg, J. Kubelka, G. Holtermann, E. Haas, and F. X. Schmid. Specificity of the initial collapse in the folding of cold shock protein. *J. Molec. Biol.*, 360:1067–1080, 2006.

[119] K. K. Sinha and J. B. Udgaonkar. Dissecting the non-specific and specific components of the initial folding reaction of barstar by multi-site fret measurements. *J. Molec. Biol.*, 370:385–405, 2007.

[120] C. J. Camacho and D. Thirumalai. Kinetics and thermodynamics of folding in model proteins. *Proc. Natl. Acad. Sci. USA*, 90:6369–6372, 1993.

[121] R. Tycko. Progress towards a molecular-level structural understanding of amyloid fibrils. *Curr. Opin. Struc. Biol.*, 14(14):96–103, 2004.

[122] M. Citron, T. Oltersdorf, C. Haass, L. McConlogue, A. Y. Hung, P. Seubert, C. Vigopelfrey, I. Lieberburg, and D. J. Selkoe. Mutation of the beta-amyloid precursor protein in familial alzheimers-disease increases beta-protein production. *Nature*, 360:672–674, 1992.

[123] W. P. Esler, E. R. Stimson, J. M. Jennings, H. V. Vinters, J. R. Ghilardi, J. P. Lee, P. W. Mantyh, and J. E. Maggio. Alzheimer's disease amyloid propagation by a template-dependent dock-lock mechanism. *Biochem.*, 39(21):6288–6295, 2000.

[124] M. J. Cannon, A. D. Williams, R. Wetzel, and D. G. Myszka. Kinetic analysis of beta-amyloid fibril elongation. *Anal. Biochem.*, 328(1):67–75, 2004.

[125] R. Wetzel. Kinetics and thermodynamics of amyloid fibril assembly. *Acc. Chem. Res.*, 39:671–679, 2006.

[126] D. K. Klimov, J. E. Straub, and D. Thirumalai. Aqueous urea solution destabilizes a beta(16-22) oligomers. *Proc. Natl. Acad. Sci. USA*, 101(41):14760–14765, 2004.

[127] B. Tarus, J. E. Straub, and D. Thirumalai. Probing the initial stage of aggregation of the a beta(10-35)-protein: Assessing the propensity for peptide dimerization. *J. Molec. Biol.*, 345(5):1141–1156, 2005.

[128] P. H. Nguyen, M. S. Li, G. Stock, J.E. Straub, and D. Thirumalai. Monomer adds to preformed structured oligomers of a beta-peptides by a two-stage dock-lock mechanism. *Proc. Natl. Acad. Sci. USA*, 104(1):111–116, 2007.

[129] D. K. Klimov and D. Thirumalai. Dissecting the assembly of a beta(16-22) amyloid peptides into antiparallel beta sheets. *Structure*, 11(3):295–307, 2003.

[130] G. Bellesia and J. E. Shea. Self-assembly of beta-sheet forming peptides into chiral fibrillar aggregates. *J. Chem. Phys.*, 126(24), 2007.

[131] F. Massi, D. Klimov, D. Thirumalai, and J. E. Straub. Charge states rather than propensity for beta-structure determine enhanced fibrillogenesis in wild-type alzheimer's beta-amyloid peptide compared to e22q dutch mutant. *Prot. Sci.*, 11(7):1639–1647, 2002.

[132] M. S. Li, D. Klimov, J. Straub, and D. Thirumalai. *J. Chem. Phys.*, 2008.

[133] M. R. Sawaya, S. Sambashivan, R. Nelson, M. I. Ivanova, S. A. Sievers, M. I. Apostol, M. J. Thompson, M. Balbirnie, J.J.W. Wiltzius, H.T. McFarlane, A.O. Madsen, C. Riekel, and D. Eisenberg. Atomic structures of amyloid cross-beta spines reveal varied steric zippers. *Nature*, 447(7143):453–457, 2007.

[134] Y. M. Rhee and V. S. Pande. Multiplexed-replica exchange molecular dynamics method for protein folding simulation. *Biophys. J.*, 84(2):775–786, 2003.

[135] D. Frishman and P. Argos. Knowledge-based protein secondary structure assignment. *Proteins*, 23(4):566–579, 1995.

[136] T. Arakawa and S. N. Timasheff. The stabilization of proteins by osmolytes the stabilization of proteins by osmolytes. *Biophys. J.*, 47(3):411–414, 1985.

[137] D. S. Yand, C. M. Yip, T. H. J. Huang, A. Chakrabartty, and P. E. Fraser. Manipulating the amyloid-beta aggregation pathway with chemical chaperones. *J. Biol. Chem.*, 274(46):32970–32974, 1999.

[138] D. Hamada and C. M. Dobson. A kinetic study of beta-lactoglobulin amyloid fibril formation promoted by urea. *Prot. Sci.*, 11(10):2417–2426, 2002.

[139] T. P. J. Knowles, W. M. Shu, G. L. Devlin, S. Meehan, S. Auer, C. M. Dobson, and M. E. Welland. Kinetics and thermodynamics of amyloid formation from direct measurements of fluctuations in fibril mass. *Proc. Natl. Acad. Sci. USA*, 104(24):10016–10021, 2007.

[140] S. D. Wolff and R. S. Balaban. Regulation of the predominant renal medullary organic solutes in vivo. *Annu. Rev. Physiol.*, 52:727–746, 1990.

[141] A. D. Williams, S. Shivaprasad, and R. Wetzel. Alanine scanning mutagenesis of a beta(1-40) amyloid fibril stability. *J. Molec. Biol.*, 357(3):1283–1294, 2006.

[142] M. R. Nichols, M. A. Moss, D. K. Reed, W. L. Lin, R. Mukhopadhyay, J. H. Hoh, and T. L. Rosenberry. Growth of beta-amyloid(1-40) protofibrils by monomer elongation and lateral association. characterization of distinct products by light scattering and atomic force microscopy. *Biochem.*, 41(19):6115–6127, 2002.

[143] S. R. Collins, A. Douglass, R. D. Vale, and J. S. Weissman. Mechanism of prion propagation: Amyloid growth occurs by monomer addition. *PLOS Biol.*, 10(2):1582–1590, 2004.

[144] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B.*

[145] A. D. Mackerell, M. Feig, and C. L. Brooks. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comp. Chem.*, 25(11):2004, 1400–1415.

[146] W. Im, M.S. Lee, and C.L. Brooks III. Generalized born model with a simple smoothing function. *J. Comput. Chem.*, 24:1691–1702, 2003.

[147] G. T. Weatherly and G. J. Pielak. Second virial coefficients as a measure of protein-osmolyte interactions. *Prot. Sci.*, 10:12–16, 2001.

[148] D. Frenkel and B. Smit. *Understanding Molecular Simulation: From Algorithms to Applications.* Academic Press, San Diego, Ca, 2nd edition, 2002.

[149] D. Frenkel, G. C. A. M. Mooij, and B. Smit. Novel scheme to study structural and thermal-properties of continuously deformable molecules. *J. Phys. Cond. Mat.*, 4(12):3053–3076, 1992.

[150] P. Nissen, J. Hansen, N. Ban, P. B. Moore, and T. A. Steitz. The structural basis of ribosome activity in peptide bond synthesis. *Science*, 289:920–929, 2000.

[151] D. K. Eggers and J. S. Valentine. Molecular confinement influences protein structure and enhances thermal protein stability. *Prot. Sci.*, 10(2):250–261, 2001.

[152] D. Thirumalai and G. H. Lorimer. Chaperonin-mediated protein folding. *Ann. Rev. Biophys. Biomol. Struc.*, 30:245–269, 2001.

[153] C. A. Woolhead, P. J. McCormick, and A. E. Johnson. Nascent membrane and secretory proteins differ in FRET-detected folding far inside the ribosome and in their exposure to ribosomal proteins. *Cell*, 116(5):725–736, 2004.

[154] M. Groll, M. Bochtler, H. Brandstetter, T. Clausen, and R. Huber. Molecular machines for protein degradation. *Chembiochem*, 6(2):222–256, 2005.

[155] J. Lu and C. Deutsch. Secondary structure formation of a transmembrane segment in kv channels. *Biochem.*, 44:8230–8243, 2005.

[156] J. Lu and C. Deutsch. Folding zones inside the ribosomal exit tunnel. *Nat. Struc. Molec. Biol.*, 12(12):1123–1129, 2005.

[157] S. C. Hinnah, R. Wagner, R. Sveshnikova, R. Harrer, and J. Soll. The chloroplast protein import channel TOC75: pore properties and interaction with transit peptides. *Biophys. J.*, 83:899–911, 2002.

[158] C. Muro, S. M. Grigoriev, D. Pietkiewicz, K. W. Kinnally, and M. L. Campo. Comparison of the TIM and TOM channel activities of the mitochondrial protein import complexes. *Biophys. J.*, 84:2981–2989, 2003.

[159] T. Hessa, H. Kim, K. Bihlmaier, C. Lundin, J. Boekel, H. Anderson, I. Nilsson, S. H. White, and G. voh Heijne. Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature*, 433:377–381, 2005.

[160] M. M. Mohammad, S. Prakash, A. Matouschek, and L. Movileanu. Controlling a single protein in a nanopore through electrostatic traps. *J. Am. Chem. Soc.*, 130(12):4081–4088, 2008.

[161] C. Dekker. Solid-state nanopores. *Nat. Nano.*, 2(4):209–215, 2007.

[162] M. R. Betancourt and D. Thirumalai. Exploring the kinetic requirements for enhancement of protein folding rates in the groel cavity. *J. Molec. Biol.*, 287(3):627–644, 1999.

[163] H. X. Zhou and K. A. Dill. Stabilization of proteins in confined spaces. *Biochemistry*, 40(38):11289–11293, 2001.

[164] D. K. Klimov, D. Newfield, and D. Thirumalai. Simulations of β-hairpin folding confined to spherical pores using distributed computing. *Proc. Natl. Acad. Sci. USA*, 99(12):8019–8024, 2002.

[165] A. Baumketner, A. Jewett, and J. E. Shea. Effects of confinement in chaperonin assisted protein folding: Rate enhancement by decreasing the roughness of the folding energy landscape. *J. Molec. Biol.*, 332(3):701–713, 2003.

[166] G. Ziv, G. Haran, and D. Thirumalai. Ribosome exit tunnel can entropically stabilize α-helices. *Proc. Natl. Acad. Sci. USA*, 102(52):18956–18961, 2005.

[167] M. S. Cheung, D. Klimov, and D. Thirumalai. Molecular crowding enhances native state stability and refolding rates of globular proteins. *Proc. Natl. Acad. Sci. USA*, 102(13):4753–4758, 2005.

[168] M. S. Cheung and D. Thirumalai. Nanopore-protein interactions dramatically alter stability and yield of the native state in restricted spaces. *J. Molec. Biol.*, 357(2):632–643, 2006.

[169] E. J. Sorin and V. S. Pande. Nanotube confinement denatures protein helices. *J. Am. Chem. Soc.*, 128(19):6316–6317, 2006.

[170] A. H. Elcock. Molecular simulations of cotranslational protein folding: Fragment stabilities, folding cooperativity, and trapping in the ribosome. *PLOS Comp. Biol.*, 2(7):824–841, 2006.

[171] A. P. Minton. Confinement as a determinant of macromolecular structure and reactivity. *Biophys. J.*, 63(4):1090–1100, 1992.

[172] H. X. Zhou. Helix formation inside a nanotube: Possible influence of backbone-water hydrogen bonding by the confining surface through modulation of water activity. *J. Chem. Phys.*, 127(24):245101, 2007.

[173] W. F. Degrado and J. D. Lear. Indcution of peptide conformation at apolar/water interfaces. *J. Am. Chem. Soc.*, 107:7684–7689, 1985.

[174] S. P. Ho and W. F. Degrado. Design of a 4-helix bundle protein - synthesis of peptides which self-associate into a helical protein. *J. Am. Chem. Soc.*, 109(22):6751–6758, 1987.

[175] H. Xiong, B. L. Buckwalter, H. M. Shieh, and M. H. Hecht. *Proc. Natl. Acad. Sci. USA*, 92:6349–6353, 1995.

[176] J. Ziegler, H. Sticht, U. C. Marx, W. Muller, P. Rosch, and S. Schwarzinger. CD and NMR studies of Prion Protein (PrP) Helix 1. *J. Biol. Chem.*, 278(50):50175–50181, 2003.

[177] R. I. Dima and D. Thirumalai. Probing the instabilities in the dynamics of helical fragments from mouse prpc. *Proc. Natl. Acad. Sci. USA*, 101(43):15335–15340, 2004.

[178] S. Williams, T. P. Causgrove, R. Gilmanshin, K. S. Fang, R. H. Callender, W. H. Woodruff, and R. B. Dyer. Fast events in protein folding: Helix melting and formation in a small peptide. *Biochemistry*, 35(3):691–697, 1996.

[179] S. Kumar and M. Bansal. Geometrical and sequence characteristics of helices in globular proteins. *Biophys. J.*, 75:1935–1944, 1998.

[180] J. D. Honeycutt and D. Thirumalai. Metastability of folded states of globular proteins. *Proc. Natl. Acad. Sci. USA*, 87:3526–3529, 1990.

[181] G. Morrison and D. Thirumalai. The shape of a flexible polymer in a cylindrical pore. *J. Chem. Phys.*, 122:194907, 2005.

[182] N. R. Voss, M. Gerstein, T. A. Steitz, and P. B. Moore. The geometry of the ribosomal polypeptide exit tunnel. *J. Molec. Biol.*, 360:893–906, 2006.

[183] Z. Guo and D. Thirumalai. Kinetics and thermodynamics of folding of a de novo designed four-helix bundle protein. *J. Molec. Biol.*, 263(2):323–343, 1996.

[184] W. F. van Gunsteren and H. J. C. Berendsen. Algorithms for macromolecular dynamics and constraint dynamics. *Molec. Phys.*, 34(5):1311–1327, 1977.

[185] S. Miyazawa and R. L. Jernigan. Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. *Prot. Struc. Func. Gene.*, 34(1):49–68, 1999.

[186] Y Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chemical Phys. Lett.*, 314(1):141–151, 1999.

[187] J. P. Valleau. Distribution of end-to-end length of an excluded-volume chain. *J. Chem. Phys.*, 104(8):3071–3074, 1996.

[188] W. A. Steele. The physical interaction of gases with crystalline solids. *Surf. Sci.*, 36(1):317–352, 1973.

[189] A. Maiti. Carbon nanotubes: Bandgap engineering with strain. *Nature Materials*, 2:440–442, 2003.

[190] P. Bolhuis and D. Frenkel. Numerical study of the phase-diagram of a mixture of spherical and rodlike colloids. *J. Chem. Phys.*, 101(11):9869–9875, 1994.

[191] B. Widom. Some topics in theory of fluids. *J. Chem. Phys.*, 39(11):2808–2812, 1963.