University of Massachusetts Medical School

eScholarship@UMMS

GSBS Dissertations and Theses

Graduate School of Biomedical Sciences

2014-07-25

Early Folding Biases in the Folding Free-Energy Surface of βα-Repeat Proteins: A Dissertation

Robert P. Nobrega
University of Massachusetts Medical School

EARLY FOLDING BIASES IN THE FOLDING FREE-ENERGY SURFACE OF βα-REPEAT PROTEINS

A Dissertation Presented

By

ROBERT PAUL NOBREGA

Submitted to the Faculty of the University of Massachusetts Graduate School of

Biomedical Sciences, Worcester

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

July, 25 2014

Biochemistry and Molecular Pharmacology

EARLY FOLDING BIASES IN THE FOLDING FREE-ENERGY SURFACE OF βα-REPEAT PROTEINS

A Dissertation Presented

By

ROBERT PAUL NOBREGA

The signatures of the Dissertation Defense Committee signifies completion and approval as to the style and content of the Dissertation.

[redacted]

C. Robert Matthews, Ph.D., Thesis Advisor

[redacted]

Francesca Massi, Ph.D., Member of Committee

[redacted]

Scott Shaffer, Ph.D., Member of Committee

[redacted]

Konstantin Zeldovich, Ph.D., Member of Committee

[redacted]

Tobin Sosnick, Ph.D., Member of Committee

The Signature of the Chair of the Committee signifies that the written dissertation meets the requirements of the Dissertation Committee.

[redacted]

William Royer, Ph.D., Chair of Committee

The signature of the Dean of the Graduate School of Biomedical Sciences signifies that the student has met all graduation requirements of the school.

Anthony Carruthers, Ph.D.,
Dean of the Graduate School of Biomedical Sciences

Program in Biochemistry and Molecular Pharmacology

July, 25 2014

Dedication

To my parents, Bob and Cheryl, family, and friends. Without all of their welcomed distractions throughout the years, I may have never come this far.

My grandparents on my mother's side deserve special credit for the completion of this work. Although my grandfather often takes credit for my intelligence, everyone knows that I have my grandmother to thank for that. She had inspired myself and my mother to pursue our doctorates and I know that she would be very proud of my accomplishment. A copy of this work will be set beside her thesis and my mother's, hopefully inspiring future generations.

Acknowledgments

This thesis would not have been possible if it were not for the guidance of Dr. Matthews who has shared his expertise and intuition with me over the years. My thesis research advisor committee members, Dr. William Royer, Dr. Francesca Massi, Dr. Scott Shaffer, and Dr. Konstantin Zeldovich, have also helped me tremendously to progress my research towards a cohesive body of work.

The first chapter of this thesis could not have been completed if it were not for the efforts of Dr. Osman Bilsel and Dr. Sagar Kathuria. Without their help and the continuous-flow technology that they have developed, the fast folding reaction I report would not have been observable. I will be forever grateful to the two of them for allowing me to contribute to their technology development efforts, and from which I take with me a newfound interest that I hope to pursue further in my future research endeavors.

I owe many thanks to all of the collaborators that I've worked with throughout my research. Dr. Charlie Brooks III and Dr. Karunesh Arora have contributed excellent Gō-model simulations that have permitted a higher resolution view of the experimental data that I have collected on CheY and the permuted variants. Dr. Vijay Pande and his graduate students, T.J. Laine, and Jade Shi, have worked for a number of years to run MD simulations on CheY and sift through the immense volume of data that was produced. The insights from their efforts has undoubtedly clarified the intricacies of the folding mechanism of

CheY. Contributions for the SAXS work come from a number of people who I've had the pleasure to work with at the BIOCAT beamline at APS, including Dr. Raul Barerra, Dr. Rita Graceffa, Dr. Liang Guo, Dr. Srinivas Chakravarthy, and Dr. Thomas C. Irving.

The second chapter of my thesis could not have been completed if it were not for the contributions of Dr. Francesca Massi and, soon-to-be-Doctor Laura Deveau. The majority of the interesting insights have come from the NMR data of which Laura has collected and processed. The expertise that Francesca has contributed has helped to turn the data into an interesting and story.

The work in chapter III is the result of extensive collaborations with the BIOCAT beamline at APS at the Argonne National Labs in Argonne, IL, as well as with Blair Perot's group at UMass Amherst. Although these efforts a cited accordingly , the developments occur interdependently over the collaborations. Noah Cohen, and Kevin Halloran have also contributed to developments and continue to do so in their own work. Andrew Malaby and Brian Mackness have contributed to the SEC-SAXS work that was cited, but not explicitly discussed. Ideas generated from those developments, are part of the greater ongoing interdependent collaborative effort.

I owe Dr. Jill Zitzewitz a special thanks. Although I haven't had much of an opportunity to work with her directly on projects, her contributions to the lab are immensely valuable to every project, including my own. She has helped write the grants I work on, offered insights into the data that I've collected, and, most importantly, has always been able get Osman and Bob to look at the bigger picture with respect to the significance of our research. Without her I'm not sure any of us could get funded to do the research that we are so interested in doing.

I thank Dr. Scott Schaffer and Mrs. Karin Green who have worked with me for countless hours on projects that remain incomplete. Their time and effort has trained me well in Mass Spectrometry, a tool that I find immensely useful and will undoubtedly use throughout my career.

The past and present members of the Matthews lab have been an immensely valuable resource for me. I've gained experience in numerous biophysical techniques, and learned to analyze hugely diverse datasets with all of them by my side. My hands were trained working with Dr. Sagar Kathuria, Dr. Can Kayatekin, Dr. Zhenyu Gu, and Dr. Xiaoyan Yang, and I received this training alongside Dr. Gangadhara Dhara, Yvonne Chan, Kevin Halloran, Meme Tran, Brian Mackness, Noah Cohen, Dr. Selma Avendano-Vazquez, and Dr. Akash Saini. Each of them has helped me develop into a better scientist.

Along with the members from my lab, everyone in the biochemistry department has made UMass an amazing place to work and study. Specifically, Dr. Ryan Hietpas, Dr. Ben Roscoe, Dr. Brendan Hilbert, Carina Clingman and Kenny Lloyd who have all been incredibly helpful in providing edits and discussing results.

I've learned something from each person and so much from everyone.

Funding

This work was supported by the National Institutes of Health through the Center for Multi-Scale Modeling Tools for Structural Biology (RR012255), the National Science Foundation through the Center for Theoretical Biological Physics (PHY0216576), the Division of Molecular and Cellular Biosciences (MCB1121942) and the National Institutes of Health (GM098763).

Use of the Advanced Photon Source, an Office of Science User Facility operated for the U.S. Department of Energy (DOE) Office of Science by Argonne National Laboratory, was supported by the U.S. DOE under Contract No. DE-AC02-06CH11357.This project was supported by grants from the National Center for Research Resources (2P41RR008630-17) and the National Institute of General Medical Sciences (9 P41 GM103622-17) from the National Institutes of Health.

The content is solely the responsibility of the authors and does not necessarily reflect the official views of the National Center for Research Resources, National Institute of General Medical Sciences or the National Institutes of Health.

Abstract

Early events in folding can determine if a protein is going to fold, misfold, or aggregate. Understanding these deterministic events is paramount for *de* novo protein engineering, the enhancement of biopharmaceutical stabilities, and understanding neurodegenerative diseases including amyotrophic lateral sclerosis and Alzheimer's disease. However, the physicochemical and structural biases within high energy states of protein biopolymers are poorly understood.

A combined experimental and computational study was conducted on the small β/α-repeat protein CheY to determine the structural basis of its sub-millisecond misfolding reaction to an off-pathway intermediate. Using permutations, we were able to discriminate between the roles of two proposed mechanisms of folding; a nucleation condensation model, and a hydrophobic collapse model driven by the formation of clusters of isoleucine, leucine, and valine (ILV) residues. We found that by altering the ILV cluster connectivity we could bias the early folding events to either favor on or off-pathway intermediates.

Structural biases were also experimentally observed in the unfolded state of a *de novo* designed synthetic β/α-repeat protein, Di-III_14. Although thermodynamically and kinetically 2-state, Di-III_14 has a well structured unfolded state that is only observable under native-favoring conditions. This unfolded state appears to retain native-like structure, consisting of a hydrophobic

core (69% ILV) stabilized by solvent exposed polar groups and long range electrostatic interactions.

Together, these results suggest that early folding events are largely deterministic in these two systems. Generally, low contact order ILV clusters favor local compaction and, in specific cases, long range electrostatic interactions may have stabilizing effects in higher energy states.

Table of Contents

List of Figures

List of Tables

List of Third Party Copyrighted Material

Chapter II - Reprinted from The Proceedings of the National Academy of Sciences, Nobrega RP, Arora K, Kathuria SV, Graceffa R, Barrea RA, Guo L, Chakravarthy S, Bilsel O, Irving TC, Brooks CL 3rd, Matthews CR. " Modulation of frustration in folding by sequence permutation" 2014 June 13; 111(29): 10562–10567.

Chapter IV - Portions of this chapter were reprinted from:

1) *Lambright D, Malaby AW, Kathuria SV, Nobrega RP, Bilsel O, Matthews CR, "Complementary techniques enhance the quality and scope of information obtained from SAXS" Transactions American Crystallographic Association. 2013 July; Vol. 44, epub,*
2) *Kathuria SV, Kayatekin C, Barrea R, Kondrashkina E, Graceffa R, Guo L, Nobrega RP, Chakravarthy S, Matthews CR, Irving TC, Bilsel O "Microsecond Barrier-Limited Chain Collapse Observed by Time-Resolved FRET and SAXS" Journal of molecular biology 2014 May;426 (9), 1980-1994,*
3) *Graceffa R, Nobrega RP, Barrea RA, Kathuria SV, Chakravarthy S, Bilsel O, Irving TC. "Sub-millisecond time-resolved SAXS using a continuous-flow*

mixer and X-ray microbeam."Journal of Synchrotron Radiation. 2013 Nov;20(Pt 6):820-5,

4) *Kathuria SV, Chan A, Graceffa R, Nobrega RP, Robert Matthews C, Irving TC, Perot B, Bilsel O. "Advances in turbulent mixing techniques to study microsecond protein folding reactions" Biopolymers. 2013 Nov;99(11):888-96, and*

5) *Kathuria SV, Guo L, Graceffa R, Barrea R, Nobrega RP, Matthews CR, Irving TC, Bilsel O. "Minireview: structural insights into early folding events using continuous-flow time-resolved small-angle X-ray scattering" Biopolymers. 2011 Aug;95(8):550-8.* I have contributed to the data collection, analysis, and writing of these manuscripts.

List of Abbreviations

AA: Amino acid

BASiC: Branched aliphatic side chain

CD: Circular dichroism

CF: Continuous flow

CheY: Chemotaxis regulator protein Y

CO: Contact order

CPU: Central processing unit

Di: Designed ideal

ESI: Electrospray ionization

EX1: Exchange mechanism 1; $k_{close} << k_{int}$

EX2: Exchange mechanism 2; $k_{close} >> k_{int}$

FL: Fluorescence

FPOP: Fast photo oxidation of proteins

FRET: Förster resonance energy transfer

Gnd: Guanidine hydrochloride

GPU: Graphics processing unit

HDX: Hydrogen-deuterium exchange

HSQC: Heteronuclear single quantum coherence spectroscopy

I_{BP}: Burst-phase folding intermediate

ILV: Isoleucine, leucine, and valine

I_{OFF}: Off-pathway folding intermediate

I_{ON}: On-pathway folding intermediate

k_{Close}: Rate constant of local protein folding events

k_{int}: Rate constant intrinsic hydrogen exchange rates

k_{Obs}: Experimentally observed rate constant

k_{Open}: Rate constant of local protein unfolding events

MALDI: Matrix assisted laser desorption ionization

MD: Molecular dynamics

MS: Mass spectroscopy

MSM: Markov state model

N: Native state

NATA: N-acetyl tryptophanamide

NMR: Nuclear magnetic resonance spectroscopy

P(r): Pair-wise distance distribution

Q: Fractional native contacts in Gō-model simulations

Q-TOF: Quadrupole-time-of-flight

Q_{max}: Maximum momentum transfer of the Guinier region

Q_t: Quantum yeild-weighted average lifetime

R_e: Reynolds number

R_g: Radius of gyration

R_h: Hydrodynamic radius

SASA: Solvent accessible surface area

SAXS: Small angle x-ray scattering

SF: Stopped-flow

TCSPC: Time-correlated single photon counting

TSE: Transition state ensemble

U: Unfolded state ensemble

U_{Gnd}: Guanidine unfolded state

U_{RS}: Residual structure unfolded state

Preface

The work presented in Chapter II has been previously published as *Nobrega RP, Arora K, Kathuria SV, Graceffa R, Barrea RA, Guo L, Chakravarthy S, Bilsel O, Irving TC, Brooks CL 3rd, Matthews CR. " Modulation of frustration in folding by sequence permutation" PNAS. 2014 June 13; 111(29): 10562–10567.* This paper was the result of a collaborative effort. Dr. Karunesh Arora carried out the Gō-model simulations and analysis. I conducted the kinetic and equilibrium experiments and the corresponding analysis. Kathuria SV, Graceffa R, Barrea RA, Guo L, Chakravarthy S, Bilsel O, Irving TC were all instrumental in the CF-SAXS development that was required to acquire and analyze the data. Dr. C. Robert Matthews, Dr. Karunesh Arora, Dr. Charles Brooks III, and Dr. Sagar Kathuria all contributed to the writing and editing of the manuscript.

The work in Chapter III is in preparation for publication. This work is in collaboration with Laura Deveau. She has collected and analyzed the NMR data and I have contributed the thermodynamic and kinetic experiments and analysis. The manuscript is currently being written by myself, Dr. C. Robert Matthews, Laura Deveau, and Dr. Francesca Massi.

Chapter IV is a summary of work that I have contributed to in the technology development of the CF-SAXS apparatus at the BIO-CAT beamline at Argonne National Labs. These publications include **1)** *Lambright D, Malaby AW, Kathuria SV, Nobrega RP, Bilsel O, Matthews CR, "Complementary techniques enhance the quality and scope of information obtained from SAXS" Transactions*

American Crystallographic Association. 2013 July; Vol. 44, epub, **2)** *Kathuria SV, Kayatekin C, Barrea R, Kondrashkina E, Graceffa R, Guo L, Nobrega RP, Chakravarthy S, Matthews CR, Irving TC, Bilsel O "Microsecond Barrier-Limited Chain Collapse Observed by Time-Resolved FRET and SAXS" Journal of molecular biology 2014 May;426 (9), 1980-1994,* **3)** *Graceffa R, Nobrega RP, Barrea RA, Kathuria SV, Chakravarthy S, Bilsel O, Irving TC. "Sub-millisecond time-resolved SAXS using a continuous-flow mixer and X-ray microbeam."Journal of Synchrotron Radiation. 2013 Nov;20(Pt 6):820-5,* **4)** *Kathuria SV, Chan A, Graceffa R, Nobrega RP, Robert Matthews C, Irving TC, Perot B, Bilsel O. "Advances in turbulent mixing techniques to study microsecond protein folding reactions" Biopolymers. 2013 Nov;99(11):888-96, and* **5)** *Kathuria SV, Guo L, Graceffa R, Barrea R, Nobrega RP, Matthews CR, Irving TC, Bilsel O. "Minireview: structural insights into early folding events using continuous-flow time-resolved small-angle X-ray scattering" Biopolymers. 2011 Aug;95(8):550-8.* I have contributed to the data collection, analysis, and writing of these manuscripts.

Chapter I - Introduction

Brief History

The discovery of proteins dates to the early 19th century when, up to that point, these molecules were ambiguously grouped into the term "animal substance" which was used to describe complex animal tissues including muscle, dermis, and blood. In 1838 the distinguished Swedish chemist, Jön Jacob Berzelius coined the term "protein" to describe these isolated molecules, a reference to their significance in nutrition[1]. For the remainder of the 19th century protein research nebulously focused on the processes of what are now known as egg albumin aggregation and hemoglobin coagulation as both proteins were readily available in abundant quantities. It wasn't until 1899 that the process of albumin aggregation was proposed to be a 2-step process, the first being denaturation and the second being aggregation[2]. It was later shown in 1925, through thermal unfolding experiments in water, that the denaturation process was conformational and not compositional. The aggregation reaction was found to be absent of expected ammonium or other nitrogen containing degradation products, an expected consequence of the known nitrogen rich chemical composition of proteins [3]. This evidence was further substantiated with the finding that hemoglobin denaturation was reversible[3] thus leading to the modern fundamental understanding of protein folding within the context of conformational thermodynamics.

With the introduction of x-ray crystallography in the 1950's, the structure of natively folded proteins began to be well understood. In 1958 Kendrew et al[4] obtained the first x-ray crystal structure of a protein using myoglobin as the target molecule. Although the resolution of 6 Å was too poor to resolve individual sidechains, the chain conformation and length led to the postulations of secondary structure composition in terms of percent α-helix, and percent extended chain. Perhaps even more importantly, it began the dialogue suggesting that the structure and the function of a given protein are interdependent. The culmination of these findings led to what is generally regarded as the beginning of modern protein folding research, Anfinsen's dogma.

In 1973 Anfinsen showed that with the chemical denaturant, Urea, he could unfold Ribonuclease A and refold it by rapidly diluting the denaturant resulting in the recovery of biological function[5]. From this, Anfinsen proposed that the amino acid sequence encodes for the three-dimensional structure of the protein and contains all of the information necessary for the protein to fold from the unfolded state to the native state. However, as Levinthal had pointed out in 1968[6], the folding mechanism from the denatured state to the native state must have some physical bias as a stochastic search for the lowest energy conformer would require folding times far longer than what would be biologically relevant.

To better understand the physicochemical biases of a protein chain, polymer theory was applied to the folding problem using three-dimensional lattice models. In these models, a "beads on a string" approximation of a protein is placed within a three dimensional grid. Using simplified physical parameters of only hydrophobic-to-hydrophobic attractions, the beads are arranged to a minimum energy state in order to better understand the conformational constraints involved in protein folding[7,8]. Results from these experiments show that there are generally many optimal configurations, and thus the parameterization must be more complex to be representative of protein folding. These results eventually lead to modern Molecular Dynamics (MD) simulations in which sets of equations describing the molecular or quantum mechanics of the system are solved for a protein chain within a solvent until achieving the global free-energy minima using replica exchange and simulated annealing techniques.

β/α - repeat proteins as models for early events in folding

The β/α-repeat class of proteins encompasses one of the most diversified enzyme platforms, the triosphosphate isomerase (TIM)-barrel $(\beta/\alpha)_8$ motif[9], as well as the Leucine-rich repeat, Trefoil knot fold, Thioredoxin fold, Rossman fold, and the Flavodoxin fold motifs among others. The TIM-barrel family of proteins, which represents the majority of enzyme folds in the Protein Data Bank database of known structures[10], is estimated to be present in nearly 10% of all enzymes[11], and maintains high structural homology with low sequence homology[10]. The TIM-barrel motif consists of 8 repeating βα units, forming a ring with 8 parallel β-strands on the interior, surrounded by 8 outer helices. Folding studies have shown that off-pathway intermediates are populated during refolding of the TIM-barrel proteins *E. coli* alpha subunit of tryptophan synthase (αTS)[12], and *S.*

solfataricus indole-3-glycerol-phosphate synthase (sIGPS)[13,14]. These stable off-pathway structures are populated within milliseconds and limit access to productive folding via their unfolding rates.

The CheY-like superfamily of the second most abundant β/α-repeat motif, the Flavodoxin fold, has also demonstrated this behavior[15,16], implicating it as a common theme across the β/α-repeat class of proteins. Structurally, the flavodoxin fold approximates half of a TIM-barrel motif with an α/β/α-sandwich architecture in which a 5 parallel stranded central β-sheet resides between two sets of helices, 3 on one side, and 2 on the other. Thus the folding properties of the Flavodoxin fold may have implications for the much broader TIM-barrel motif.

It is unknown what the biological significance of an off-pathway folding intermediate is, or whether or not it is subject to evolutionary pressures. However the existence of these intermediates in relatively small Flavodoxin fold proteins provide the protein folding community with a model to understand complicated protein folding pathways and possibly illuminate sequence or structural properties of proteins that have a propensity to misfold. Ideally, insights from this model could be used to prevent misfolding of engineered proteins, optimize protein therapeutics, and provide insight into protein misfolding diseases such as Alzheimer's or ALS.

Folding mechanism of CheY

The small (129 AA) bacterial response regulator CheY has a well studied folding mechanism[15–20] and serves as a model protein for understanding off-pathway intermediates. Structurally, the N and C-termini of CheY reside in two helices on one side of the β-sheet, opposed by three helices on the alternate side[21] (Fig 1.1). An interesting feature of the central β-sheet is the strand sequence which contains an intercalated β1 strand between β2 and β3, yielding a β-sheet strand sequence of 2-1-3-4-5. This topology, although simple, appears to have complicated kinetic folding properties that manifest as an off-pathway intermediate[15].

The proposed kinetic folding mechanism[15] (Fig. 1.2) involves two parallel folding channels defined by the cis and trans isomers of the prolyl peptide bond between K109/P110. The unfolded protein in both the major trans (90%) and minor cis (10%) channels, U_t and U_c, sample an off-pathway sub-ms intermediate, I_{BPt} and I_{BPc}, prior to the rate-limiting isomerization reaction in the $I_{BPt} \rightarrow I_{BPc}$ step. I_{BPc} unfolds to the U_c state before accessing the productive TSE leading to the native conformation in the $U_c \rightarrow N_c$ step. Further complicating the mechanism is an on-pathway intermediate, I_{ON}, between U_c and N_c that has been observed by equilibrium NMR measurements[20] and in Gō-model simulations[22] but not by CD or FL experiments. Mutational analysis[18,19] has revealed a nucleation-condensation folding mechanism for CheY, in which the N-terminal subdomain (residues 1-70, (βα)1-2β3) serves as the nucleus for the subsequent condensation of the C-terminal subdomain (residues 70-129, α3(βα)4-5), providing structural details of the folding mechanis

Figure 1.1. Topology diagram of CheY. CheY is a βα-repeat, α/β/α sandwich motif. The central parallel β-sheet is flanked on both sides by helices. On one side sequence distal α1 and α5 pack onto the β-sheet, while the opposing side consists of consecutive helices. The strand order of the central β-sheet is an interesting feature as the β1 strand is intercalated between β2 and β3.

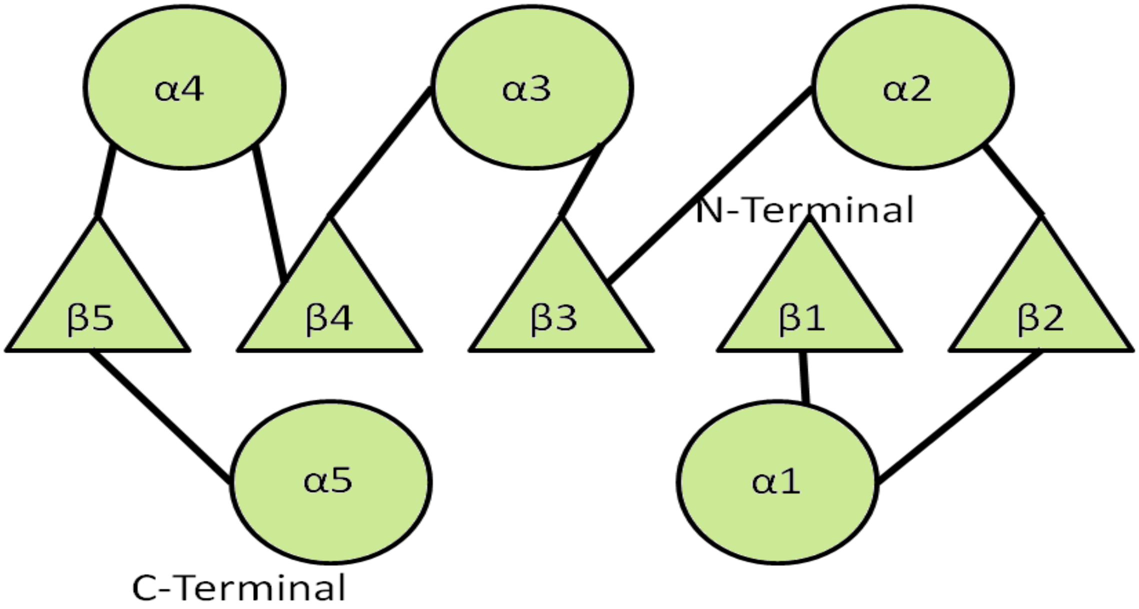


Figure 1.1. Topology diagram of CheY

Figure 1.2 Folding Mechanism of CheY**.** The folding mechanism of CheY is complicated by the K109/P110 trans->cis prolyl isomerization reaction occurring on the order of 100 seconds. This reaction splits the reaction mechanism into two parallel channels. Under strongly refolding conditions the dominant U_T state folds to I_{BPT}, an off-pathway intermediate. After the isomerization reaction I_{BPC} unfolds and then continues to fold rapidly to the N_C state. The folding reaction times of the dominant pathway are included for reference.

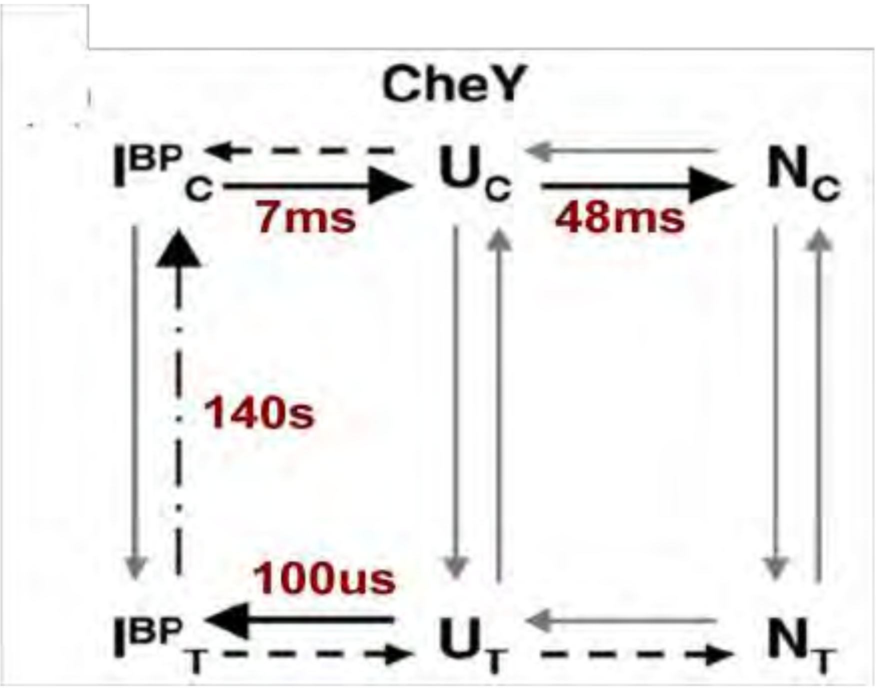


Figure 1.2. Folding Mechanism of CheY

Complexities and general principles of globular protein folding

The complexity of the protein folding problems stems from the chemical and physical diversity of the 21 naturally occurring amino acids. Generally, each amino acid side-chain can be categorized into one of 3 categories: polar charged, polar un-charged, or non-polar. Within these individual categories each side-chain has unique properties including thiol reactivity (i.e., redox potential), aromatic stacking ability, hydrogen bond character, acceptable phi/psi angles in backbone flexibility, degrees of freedom in side-chain flexibility, and pKa. All of these properties, within the context of folding, are susceptible to the influence of the solvent system including properties like hydrophobicity, pH, ionic strength, oxidation potential, temperature, and viscosity. Further complicating these general properties are protein specific considerations like disulfide scrambling[23], metal ion binding[24], ligation of prosthetic groups[25,26], and proline isomerization reactions[26].

As complicated as the exact interplay of all these properties are, there exists at least three major general principles of globular protein folding that have been used in simple computational models and can generally describe the folding phenomenon. First, helical structure occurs very rapidly in aqueous solvent, preceding hydrophobic collapse[27]. Second, side-chains mostly adhere to their partition coefficients preference for their local environments[7,28], such that in aqueous solvent hydrophobic residues have a preference to be buried within the core of a protein and polar residues prefer to be solvent exposed. Lastly, protein folding is a cooperative event[28].

Although these principles can describe the collapse of a polypeptide chain to some globule structure, this information content alone is not complete enough to capture the native state from a random-coil chain. Including the parameterization for polarization, torsion angles, and solvent effects greatly increases the computational expense of the calculations making full trajectory simulations of protein folding impossible just a few decades ago[29]. This impasse has led to the development of targeted computational models for both structure prediction algorithms, like Rosetta, and coarse-grained simulations, like Gō-models, such that the computational load could be decreased while still approximating a reasonable solution.

Rosetta was designed to predict the native fold of a given sequence. Initially, predictions were based on sequence-structure correlations[30] and grew into a process involving the comparison of sequence alignments across a database of known structures to identify structured fragments that could then be assembled and subsequently subjected to simulated annealing experiments[31]. This process was relatively successful and ultimately led to further developments resulting in RosettaDesign, which produced the first successful prediction of a completely engineered protein with a novel fold, Top7[32]. However, the limitations of predicting a native state from a sequence does little to explain the mechanism of folding. In fact, it was later shown experimentally that the folding mechanism for Top7 was exceedingly complex[33] despite the simple two-state thermodynamic characteristics that it was initially shown to have in the seminal design publication[32].

Conversely, Gō-model simulations were designed to gain insights into the folding mechanism while using simplified potential functions[8]. These models attempt to only describe the folding landscape and not the native structure. In fact, this particular modeling strategy is native-centric and relies on existing knowledge of the final structure. As in other simplified models, the protein chain is approximated to a beads-on-a-string model at the cα position. The chemical properties of the sidechains are parameterized at each position and the simplified potential drives each bead to the native tertiary contacts[22]. Although simplified and highly biased, these models allow for the observation of kinetic complexities like intermediates[14–16,34], and even non-native interactions[35]. Using coarse-grained potential functions decreases the computational load immensely, allowing users to calculate full trajectories of protein folding reactions in a matter of hours instead of months. The caveat, however, is the exact opposite of Rosetta's, in that changes to the protein sequence will not reveal changes to the final structure.

All atom simulations, although computationally expensive, offer the greatest insight into both the folding mechanism and the determination of the native state. In a fully parameterized all-atom MD simulation the computational power required to simulate the full folding trajectory of a protein approaching 100 amino acids in length now requires supercomputers[36] or large distributed computing networks[37]. Even then, the resulting data is immense and often difficult to interpret without the context of experimental data. Realistically, each full trajectory will be different, and in the case of the statistically assembled Markov State Models (MSM), the primary pathway for folding may not be obvious. Simply stated, even with our most sophisticated calculations, it is difficult to recapitulate the experimental responses which are generally described as simple exponentials.

Experimental and computational timescales and resolutions

A related challenge to the computational expense of folding simulations, and one approach towards better understanding the differences between computational and experimental responses, is the validation of computational results with experimental data. Unfortunately, the majority of atomistic folding simulations are on the timescale of ns to μs, with very few breaking the ms barrier. In contrast, for most naturally occurring proteins the folding times are on the ms to min timescales. Thus the results of the folding simulations are difficult to confirm as it is unknown whether or not the trajectory would ever reach the natively folded state. Additionally, there is generally an insufficient overlap of data between the two approaches to robustly sync the results of the simulations with the results of folding experiments. Verifying the similarities between the datasets is further complicated by the difference in resolutions across the two approaches. Thus as Moore's law lengthens the accessible simulation times, experimentalists need to also be approaching faster time scales to achieve reasonable agreement.

Recent advances in mixing techniques are improving the time resolution of protein folding experiments while simultaneously diversifying the compatible measurement techniques[38,39], improving the signal-to-noise with better detectors and analysis[40], and reducing the sample consumption[41,42]. Not only do these improvements increase the applicability of the experiments to a broader set of protein systems, but the improved time resolution and signal-to-noise ratios produce high-quality and simulation compatible data such as time resolved population distributions or pair-wise distances for direct comparison to the computational dataset.

With the exception of time resolved distance distributions, it remains challenging to combine the information content across the two platforms because of the differences in the resolution. There exist only few types of single molecule experiments; optical tweezer pulling[43], atomic force microscopy pulling[44], and single molecule Förster resonance energy transfer[45] which can be directly compared to appropriately designed simulations[46,47]. Comparison of simulations with ensemble experimental data sets is complicated by the reduction of the data down to only a few metrics. These metrics include ΔG, m-value, rate constants, and pairwise or global distance measurements. With the exception of Markov State Models, simulations are effectively a set of single molecule experiments (i.e. single molecule trajectories). Determining the appropriate weighting of the populations in simulations to recapitulate the simulation data remains an outstanding challenge. Further, calculating the expected raw data from a set of simulations for a direct comparison suffers from this same issue, even when using straight-forward calculations like pairwise distances.

Folding models: pathways and landscape theory

The observation that protein folding experiments generally exhibit simple exponential responses is consistent with Levinthal's supposition that the folding process must be pathway specific[6]. He proposed that through global energy minimization, a protein chain folds from an unfolded ensemble to the native state in a manner where specific structures are adopted in a specific sequence of increasing structure and decreasing free-energy[48]. If this were true then simple kinetic responses will be observed and, logically, must then describe the entire energy surface. In fact there are several models that stem from the simplicity of experimental observables.

In the *nucleation-condensation* folding model[49] it is postulated that the formation of continuous and local secondary structure will result in tertiary contacts that other secondary elements can then condense upon. However, this model fails to describe stable kinetic intermediates as it supposes that the rate-limiting step of folding is the nucleation event[48,50]. Other models suggest a less continuous accumulation of structure to better explain the presence of intermediate states.

In the *framework* and *diffusion-collision* models it is suggested that secondary structural elements are formed first and then collapse together to form a molten globule. In the *framework* model the rearrangements of tertiary contacts are rate-limiting[51], while in the *diffusion-collision* model, adhesion of secondary elements is rate-limiting[48,50], describing different mechanisms of kinetic traps.

Lastly, the *hydrophobic collapse* model supposes that hydrophobic collapse occurs which reduces the chain entropy and accelerates folding. Subsequent structural rearrangements, including topological and secondary structure rearrangements, are considered to be the rate-limiting step[48].

Although these models appear to be consistent with the pathway-centric experimental view, computational experiments demonstrate pathway independence as there are always multiple favorable pathways capable of traversing the same free-energy difference. These observations have led to the idea of *landscape theory* in which there are many micro-states that a protein chain can sample enroute to the lowest energy state. Conceptually, this idea is imagined as a funnel-shaped, continuous energy surface where conformational entropy describes the width of the funnel and the free-energy describes the height[52]. Depending on the protein system, the funnel will have a varying degree of ruggedness on the surface representing local energy minima and maxima. Thus the folding of a given protein sequence is free to sample any combination of consecutive points, guided by statistical thermodynamics, along a decreasing energy gradient. Within this context, the simplified kinetic responses overwhelmingly observed in experiments are rationalized to be the weighted average of the diverse sampling of the energy landscape such that experimentally observed kinetic intermediates represent statistically favorable units of structure that occur in a consistent temporal sequence.

The BASiC hypothesis

Each folding model presumes that sequence local contacts initiate folding. Recent work has correlated the stable core of a diverse set of globular proteins with the presence of Branched Aliphatic Side Chain (BASiC) residues: isoleucine, leucine, and valine (ILV)[53,54]. The basis of this model is that ILV residues are the most hydrophobic residues[55], and are capable of packing tightly and sliding over each other with minimal energetic penalty. These qualities permit a tightly packed yet malleable platform for protein folding events to build upon. The formation of these cores are implicated in early intermediates[15,22,56,57] and therefore may play a larger role in biasing the mechanism through which a given protein will fold.

Role of sequence and chain connectivity in protein folding

Within the context of landscape theory, it follows that the primary sequence could produce a statistical bias in the free-energy landscape through altering the local physicochemical properties of the chain[58]. In the context of point mutations[59–62] the folding mechanism is generally conserved along with the native topology. Therefore, the effects of small sequence perturbations manifest predominately in the modulation of the state interconversion rates. This phenomenon is so ubiquitously observed that it is the basis of phi value analysis, a method of protein engineering where point mutations are used to determine the presence of interactions at a given residue exist within the transition state[61].

In larger sequence perturbations, like structurally homologous proteins that can vary extensively in sequence, differences in state interconversion rates are also observed[15,63]. Within the context of the contact order model of folding proposed by Plaxco et al[64], these changes are suggested to be the result of differences in the structure of the transition state, much like the aforementioned point mutations, such that sequences that have a low contact order fold faster than those that have a high contact order. Extrapolating these findings to proteins with intermediate states suggests that depending on the relative contact order of the intermediates, the preferable order of folding events could be modulated by changes to the contact order.

One way to perturb the contact order is through the use of circular permutations. Circular permutations are a special case of sequence diversity for a given protein because the physicochemical properties are nearly entirely conserved with the exception of local chain entropy, which is affected by the new chain connectivity. The effect of this perturbation is that local in sequence contacts can be made to be distal contacts, and vice versa. These modifications have been shown to change the transition state in folding[65,66] as well as populate low energy intermediates[67]. Therefore, permutations appear to differ from mutational perturbations in that they are more likely to change the location of early folding events on a sequence-local, and therefore contact order dependent basis. Notably, in most reported cases, permutations are surprisingly well tolerated in terms of achieving the native state regardless of changes to early structural biases[65,67,68].

At a glance, it is perhaps not surprising that the amino acid sequences of proteins are nearly universally malleable in terms of achieving the nominal native state. From an evolutionary perspective it is desirable that mutations are tolerated to permit sequence variations that can lead to improved fitness or novel functions. What is surprising is the conservation of the kinetic pathways across diverse sequence-space. Permutations, having the ability to significantly change the early folding events of a given polypeptide chain, may hold a key to understanding the sequence-structure relationship.

Scope of this thesis

This dissertation focuses on the early folding events of β/α-repeat proteins, specifically the chemical and physical origins of such events and their effect on the folding reaction. In Chapter II, the origins of sub millisecond folding intermediates in CheY are examined, in Chapter III early biases in folding via a compact unfolded state of a synthetic β/α-repeat protein are described, and in Chapter IV improvements in experimental mixing techniques are described with a proposed method of robust MSM analysis using experimental data.

In Chapter II the folding mechanism of the small Flavodoxin fold protein CheY is examined. CheY, a 129 residue chemotaxis signaling protein in *E. coli,* has been extensively studied in the protein folding field. Its folding mechanism is of particular interest because it populates a sub-ms off-pathway folding intermediate which is not typical of a protein of this size[15]. The folding mechanism, $I_{off} \rightleftharpoons N \rightleftharpoons U$, is further complicated by the single rate-limiting K109/P110 proline isomerization reaction, which results in parallel pathways separated by the prolyl-bond isomerization reaction, totaling 6 interconverting states (Fig. 1.2).

Structurally, experimental evidence suggests that CheY folds via a nucleation condensation reaction, where the N-terminal half of the protein nucleates the folding reaction followed by the condensation of the C-terminal domain upon the nucleated scaffold[19]. This order of events also correlates with the contact density of the two domains[69]. However little is currently known about the structural correlates of either intermediate state outside of what has been gleaned from previous Gō-model simulations[15,16]. These data suggest that the off-pathway intermediate results from the premature accumulation of native structure consisting of most of the secondary structural elements of the N-terminal domain and two subsequent elements of the C-terminal domain, α3 and β4[15]. This structural information, however, is limited by the native-centric bias of the Gō-model simulations in that non-native contacts will not be observed by the methods that have been employed .

Interestingly, the severity of the observed topological frustration, across a set of 3 structurally homologous proteins, is correlated with the calculated size of the native state stabilizing ILV clusters[15]. Application of the BASiC hypothesis to this observation suggests that early coalescence on ILV residues in a non-native configuration may be playing a role in the experimentally observed frustration. Initial attempts to resolve this information with hydrogen-deuterium exchange mass spectrometry were not capable of resolving significant structural details due to the relatively low stability of the intermediate state and the associated challenges involved with pulse-labeling a sub-ms intermediate.

In Chapter II we collaborate with two computational groups: the Brooks lab at Michigan who are experts with Gō-model simulations, and the Pande lab at Stanford who use the distributed computing platform *Folding@Home* to run large atomistic simulations. Along with our technology development efforts towards decreasing the dead time of continuous flow experiments and interfacing microfluidic mixing technology with SAXS experiments, we arrive at overlapping and comparable data that can be used to enhance the resolution of our classical folding studies. We make use of circular sequence permutations to directly change the sequence connectivity of the ILV clusters in CheY. The experimental kinetic, dimensional, and thermodynamic measurements are aligned with the results from the Gō-model simulations in order to directly test the application of the BASiC hypothesis in the early folding events of CheY and how those events relate to the off-pathway intermediate. Likewise, the fast folding events captured by SAXS are leveraged to identify the atomistic details of both on and off-pathway folding intermediates from the data obtained from the *Folding@Home* simulations which complement the Gō-models by including non-native contacts.

In Chapter III the focus shifts to even earlier events in folding, residual structure in the unfolded state, using the *de novo* designed β/α-repeat protein Di-III_14. This protein was engineered by the Baker group using the Rosetta suite of design tools[70]. Previous work from the Baker group in *de novo* design resulted in the successful design of Top7, a small protein that had a fold that has not been found in nature. This protein was cooperatively folded and thermodynamically 2-state. However, the later work showed that the kinetic folding mechanism was exceedingly complex and suggested that the non-natural development of the Top7 fold was absent of the natural selection for a smooth landscape[33].

In designing Di-III_14, the principles used for the design were loop length optimization to predict the native topology, optimizing hydrophobic packing in the interior of the protein, and positioning polar sidechains on the surface. Unlike Top7, the loop lengths were calculated for Di-III_14 such that the topology would mimic a naturally occurring fold. The Baker group was again successful in predicting the experimental native state and in designing a cooperatively folded thermodynamically 2-state protein. It was suggested that the energy landscape would be smooth because the fold was naturally occurring and the protein was optimized for sequence-local contact[70]. Work in Chapter III investigates this claim with thermodynamic and kinetic experiments. Native state hydrogen exchange observed by NMR was also used to gain further details of energy surface of Di-III_14.

In Chapter IV, technical advances in continuous flow mixing technology interfaced with small angle x-ray scattering is discussed. Improvements in the time resolution of continuous-flow mixing experiments permit a robust dataset, ranging from global to residue level resolutions, to be collected in the microsecond to millisecond time regime. These data can be can be directly compared to simulation data using easily calculated distance distributions.

We propose a method for model refinement based on the experimental data without direct mechanistic input for large MSM models of proteins with known complicated folding mechanisms. Simulations of proteins consisting of 100 amino acids or more and proteins containing slow phase folding complexities are still very challenging to approach with computational methods as they generate large datasets that can be difficult to interpret independently. Accounting for processes like proline isomerization or disulphide bridges during folding require assumptions to be incorporated into these models, biasing the results and complicating the analysis. Without introducing a mechanistic bias, experimental distance distributions can be used to refine the analysis of an MSM dataset, relying only on the synchronicity of the time domain. In this way high confidence and high resolution structural details of the folding process can be extracted.

Chapter V concludes this dissertation with a summary of the previous chapters. Future directions are also discussed.

Chapter II - Early Folding Events in CheY

This chapter has been published previously as:

***Nobrega RP**, Arora K, Kathuria SV, Graceffa R, Barrea RA, Guo L, Chakravarthy S, Bilsel O, Irving TC, Brooks CL 3rd, Matthews CR. " Modulation of frustration in folding by sequence permutation" PNAS. 2014 June 13; 111(29): 10562–10567*

The published work presented in this chapter was a collaborative effort. Gō-model simulations were performed and analyzed by Dr. Karunesh Aurora. All experimental work was performed by myself. Data interpretation and the writing on the manuscript was the work of myself and Dr. C. Robert Matthews.

This chapter has been expanded with data from another collaboration that I am currently involved with. The SAXS contributions are the work of myself and Dr. Sagar Kathuria, while the Molecular Dynamics simulations and analysis are the work of Jade Shi and T.J. Lane under the guidance of Dr. Vijay Pande. The manuscript is currently being written by Jade Shi and Vijay Pande.

Introduction

Highly-denatured states of globular proteins resemble statistical random coils when examined with low resolution techniques such as x-ray scattering[71] and hydrodynamic analyses[72]. A higher resolution view, provided by experimental models[27,73–75] and simulations[76], however, shows that the conformational ensemble is biased towards low contact order (CO) structures, e.g., α-helices, β-turns and β-hairpins, that form and melt in less than a few microseconds. During folding, these nascent structures presumably coalesce into higher order assemblies of ever-increasing free energy until reaching the transition-state ensemble (TSE) that leads to the native conformation. From another perspective, this assembly process mediates a global collapse of the chain in an unfavorable solvent[77]. Landscape theory[78] posits that, in the simplest scenario, native-like substructures appear and lead without pause to the TSE and the native conformation in an apparent 2-state fashion. However, simulations have found that topological frustration, e.g., the premature formation of sub-structure that impedes access to the productive TSE, can lead to the accumulation of intermediates that must unfold to some extent to successfully traverse the folding reaction coordinate[16,22,77]. Experimental and computational studies on the folding of the alpha subunit of Trp synthase[79,80], the response regulator CheY [15,22], a pair of apo-Flavodoxins[71,77,81] and tandem titan domains[82] revealed frustration in the form of off-pathway intermediates. Thus, yet unexplored aspects of sequence and structure can add complexity to folding reactions.

The observed inverse relationship between CO and folding rate constant[83] implies that elements of secondary structure that are near in sequence and near in space will associate preferentially over those that are distant in sequence. However, if such low CO substructures are not involved in the productive TSE, they could serve as sources of frustration. A case in point is CheY, a member of the very common Flavodoxin-fold family with its classic α/β/α-sandwich architecture. The (β/α)5 motif displays the α1 and α5 helices on one face of the parallel β-sheet and the α2, α3 and α4 helices on the opposing face (Fig. 2.1A,B). The proposed kinetic folding mechanism[15] (Fig. 2.1C) involves two parallel folding channels defined by the cis and trans isomers of the prolyl peptide bond between K109/P110. The unfolded protein in both the major trans (90%) and minor cis (10%) channels, U_t and U_c, sample an off-pathway sub-ms intermediate, I_{BPt} and I_{BPc}, prior to the rate-limiting isomerization reaction in the $I_{BPt}{\rightarrow}I_{BPc}$ step. I_{BPc} unfolds to the U_c state before accessing the productive TSE leading to the native conformation in the $U_c \rightarrow N_c$ step. Further complicating the mechanism is an on-pathway intermediate, I_{ON}, between U_c and N_c that has been observed by equilibrium NMR measurements[20] and in Gō-model simulations[22] but not by CD or FL experiments.

Figure 2.1. Topology, clusters, and subdomains of CheY. (A) Topology diagram of CheY. The N-terminal folding subdomain is highlighted in yellow, and the C-terminal folding subdomain is highlighted in blue. The effects of each permutation on the continuity of cluster 1 (blue), and cluster 2 (red) are shown. (B) Clusters of ILV residues are superimposed on the crystal structure of CheY (Protein Data Bank ID code: 3CHY). Cluster 1(blue) has a lower CO and resides on the α2/α3/α4 side of the central β-sheet. The larger cluster 2 (red) contains high-CO contacts and resides on the α1/α5 side of the β-sheet. (C) The folding mechanism of WT CheY. The major pathway is highlighted in red. The Uc →Nc step, involving the on-pathway intermediate, is designated by the triple dots.

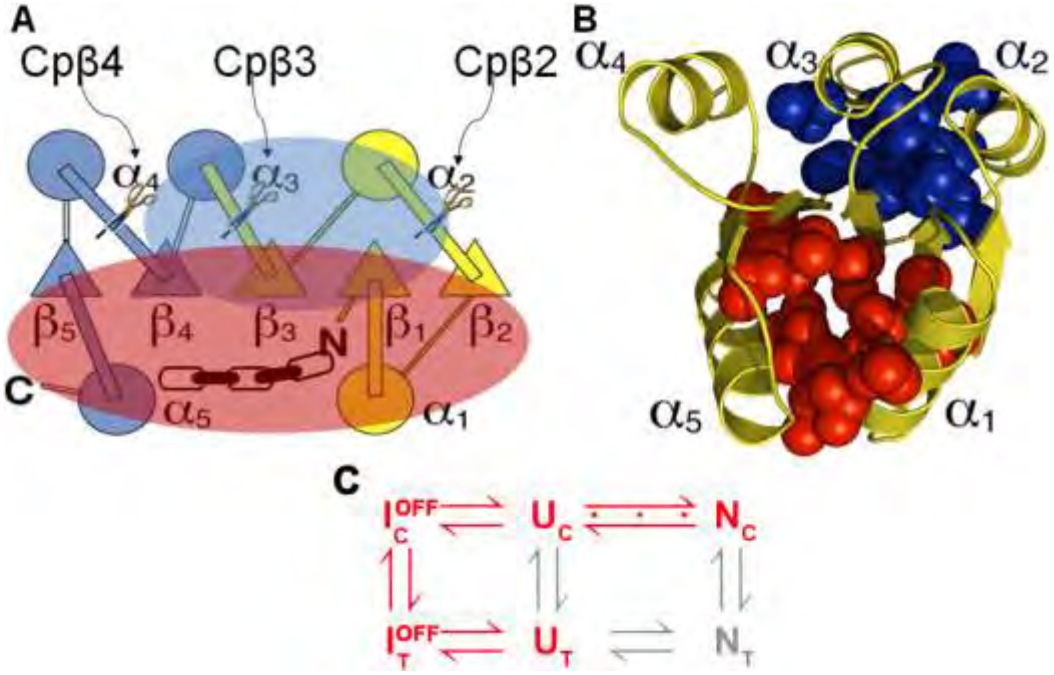


Figure 2.1. Topology, clusters, and subdomains of CheY

Mutational analysis[18,19] has revealed a nucleation-condensation folding mechanism for CheY, in which the N-terminal subdomain (residues 1-70, (βα)1-2β3) serves as the nucleus for the subsequent condensation of the C-terminal subdomain (residues 70-129, α3(βα)4-5) (Fig. 2.1A). However, native-centric simulations identified contacts between the N- and C-terminal subdomains, centered around (βα)3-4, early in folding that is incompatible with access to the productive TSE and that lead to frustration in the folding mechanism[22]. Another perspective is provided by the BASiC hypothesis, which supposes that large clusters of isoleucine, leucine and valine (ILV) side chains serve as cores of stability in folding intermediates[15,16]. Both of these clusters have been shown to have a high contact density[84]. CheY has two ILV clusters, each serving to fuse the surface helices to each other and to the central β-sheet (Fig. 2.1B). The smaller cluster (Cluster 1) contains 10 side chains and primarily links α2(βα)3β4 on one face of the β-sheet; the larger cluster (Cluster 2) contains 15 side chains and links the β-strands to α1 and α5. The sequence spanned by the smaller cluster agrees closely with the (βα)3-4 segment identified as the source of frustration in the simulations and, importantly, only involves low CO contacts. If Cluster 1 were to form early and, by sequestering β3, impede the development of the productive TSE in the N-terminal subdomain, (βα)1-2β3, the BASiC hypothesis would provide an alternative explanation for frustration in the folding of CheY.

Permutations in the sequence of CheY provide a means to compare the subdomain model and the ILV cluster model as explanations for the frustration in folding detected by simulations and experiments. By fusing the natural N- and C-termini with a short linker peptide (Gly-Ala-Gly) and inserting new termini in the loops after β2, β3 and β4, it is possible to cleave within the N-terminal subdomain, Cpβ2, between the subdomains, Cpβ3, and within the C-terminal subdomain, Cpβ4. Related to the ILV clusters, Cpβ2 cleaves Cluster 1 and leaves Cluster 2 largely intact, Cpβ3 cleaves both clusters and Cpβ4 only cleaves Cluster 2 (Fig. 2.1A). Our simulations and experiments on these permutants show that aspects of both models describe the relationships between sequence, structure and frustration in the folding of CheY. The results also show that frustration can be modulated by sequence permutations that can bias the initial stages of folding towards the productive TSE and away from kinetic traps.

Results

Permutations differentially affect the secondary structures of the folded state

We introduced sequence permutations into the F14N variant of CheY, denoted CheY*, to increase the stability of the platform and its tolerance for the introduction of the linker and the new termini; the folding mechanism for CheY* is unchanged from the WT protein[18]. The new N-termini for Cpβ2, Cpβ3, and Cpβ4 become D38, D64 and E89, respectively. An additional glycine residue at position -1 is a remnant of the cleaved 6-His affinity tag. The far-UV CD spectrum of Cpβ2 is markedly different from CheY* (Fig. 2.2), with the relative intensities of the double minima at ~210 and ~222 nm reversed from those of the CheY* protein. Unfortunately, the substantial perturbation of the secondary structure precluded Gō-model simulations that rely on a knowledge of the native structure. The CD spectra of Cpβ3 and Cpβ4 display the same relative double minima as CheY*; the spectrum of Cpβ4 is coincident with CheY* and Cpβ3 is reduced in amplitude by ~15% (Fig. 2.2). Although the secondary structure of Cpβ3 appears to fray to some extent, the basic β/α/β architecture is preserved. Therefore, both Cpβ3 and Cpβ4 were deemed to be good candidates for a combined experimental and computational analysis of their folding mechanisms.

Stability analysis of the permutants

The concerted disruption of secondary and tertiary structure with increasing concentrations of urea revealed an apparent 2-state process, $N \rightleftharpoons U$, for CheY* and Cpβ2 (Fig. 2.3, Table 2.1). Fits of the data to a linear dependence of free energy of folding on the denaturant concentration[85] showed that the stabilities varied from 2.11 kcal•mol^{-1} for Cpβ2 to 8.0 kcal•mol^{-1} for CheY* protein (Table 2.1). The denaturant dependence of the free energy of folding, the m-value (a measure of the change in buried surface area[86]), varied from 0.77 (kcal•mol^{-1})M_{urea}^{-1} for Cpβ2 to 1.99 (kcal•mol^{-1})M_{urea}^{-1} for CheY*.

Fig. 2.2. Secondary structure. The CD spectrum of each construct under native conditions is shown. CheY* (black) and Cpβ4 (red) are superimposable, Cpβ3 (blue) has a small decrease in ellipticity but maintains the same relative minima, and Cpβ2 (magenta) has a unique native CD spectrum.

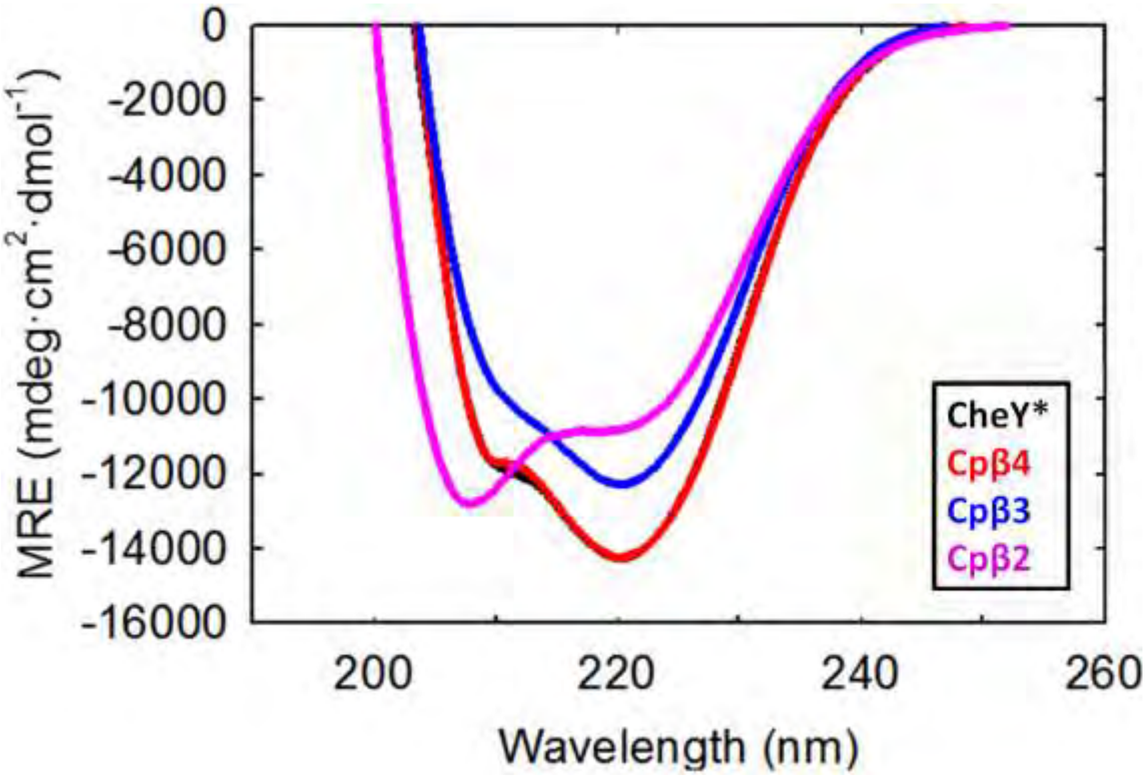


Figure 2.2. Secondary Structure of Permutants

Fig. 2.3. Analysis of N and I$_{BP}$ stability. Filled symbols display the urea melts derived from the ellipticity at 222 nm for CheY* and each of the permuted variants; the denaturant-induced unfolding reactions are fully reversible. The open symbols display the urea dependence of the ellipticity at 222 nm after 5 ms of refolding. With the exception of Cpβ4, the solid and dashed lines show the fits of these data to two-state equilibrium models. The data for Cpβ4 are fit to a three-state model.

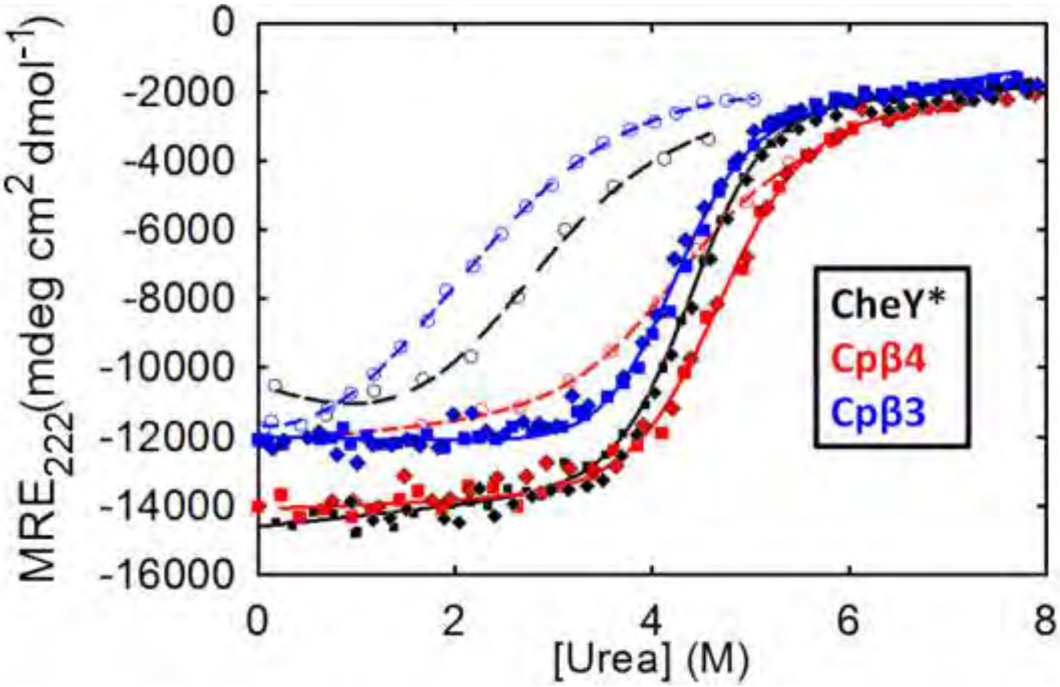


Figure 2.3. Analysis of N and I$_{BP}$ stability

Table 2.1. Thermodynamic parameters for CheY* and circular permutants.

CD (2-state)	$\Delta G^{\circ}_{N\rightleftarrows U}$ (kcal•mol^{-1})	$m_{N\rightleftarrows U}$ (kcal•mol^{-1} M$_{urea}$$^{-1}$)	$\Delta G^{\circ}_{IBP\rightleftarrows U}$ (kcal•mol^{-1})	$m_{IBP\rightleftarrows U}$ (kcal•mol^{-1} M$_{urea}$$^{-1}$)
CheY*	8.00±0.15	1.99±0.41	2.02±0.24	0.83±0.13
Cpβ4	6.19±0.07	1.32±0.02	4.31±0.19	0.99±0.04
Cpβ3	6.79±0.08	1.62±0.02	0.84±0.44	0.59±0.27
Cpβ2	2.11±0.08	0.77±0.02	--	--
CD (3-state)	$\Delta G^{\circ}_{N\rightleftarrows I}$ (kcal•mol^{-1})	$m_{N\rightleftarrows I}$ (kcal•mol^{-1}•Murea^{-1})	$\Delta G^{\circ}_{I\rightleftarrows U}$ (kcal•mol^{-1})	$m_{I\rightleftarrows U}$ (kcal•mol^{-1}•Murea^{-1})
Cpβ4	3.88±0.93	0.81±0.21	4.31±0.19	0.99±0.04
FL Fit	$\Delta G^{\circ}_{I\rightleftarrows U}$ (kcal•mol^{-1})	$m_{I\rightleftarrows U}$ (kcal•mol^{-1}•M$_{urea}$$^{-1}$)		
Cpβ3	4.14±0.06	1.34±0.02		

Notably, Cpβ3 displayed complex equilibrium unfolding reaction, with non-coincident CD and FL denaturation transitions (Fig. 2.4C). The lower Cm of the FL unfolding transition most likely reflects the introduction of the new N-terminus only a few residues downstream from the single tryptophan, W58. The stability estimated for the global unfolding reaction, indicated by the CD transition, is 6.79 kcal•mol^{-1}. Although the titration data for Cpβ4 could be fit to a 2-state model, kinetic analysis (see below) revealed the presence of a stable intermediate and dictated a 3-state model. The melting temperatures estimated from the heat capacities calculated by the simulations (Fig. 2.5) are in the same rank order as the midpoint points in the urea titrations (Fig. 2.3): Cpβ3 < CheY* < Cpβ4. Experimental thermal melts by both DSC and CD were irreversible and a reliable experimental measurement of Tm could not be obtained. Further experiments on Cpβ2 were not pursued.

Kinetic analysis of permutant folding

We monitored the dynamic responses of the permutants to rapid changes in the denaturant concentration in the micro-to-hundreds of seconds time range with a combination of continuous-flow (CF), stopped-flow (SF) and manual-mixing (MM) techniques interfaced to FL, CD and SAXS detection. For CheY*, a large amplitude FL phase occurs within the 25 μs dead time of CF-refolding, followed by a small amplitude, several hundred μs phase.

Fig. 2.4. Comparison of CD and fluorescence (FL) titrations. Titrations monitored by FL emission at 315 nm (open circles) and CD ellipticity (closed circles) are plotted for CheY*(A), Cpβ4 (B), and Cpβ3 (C).

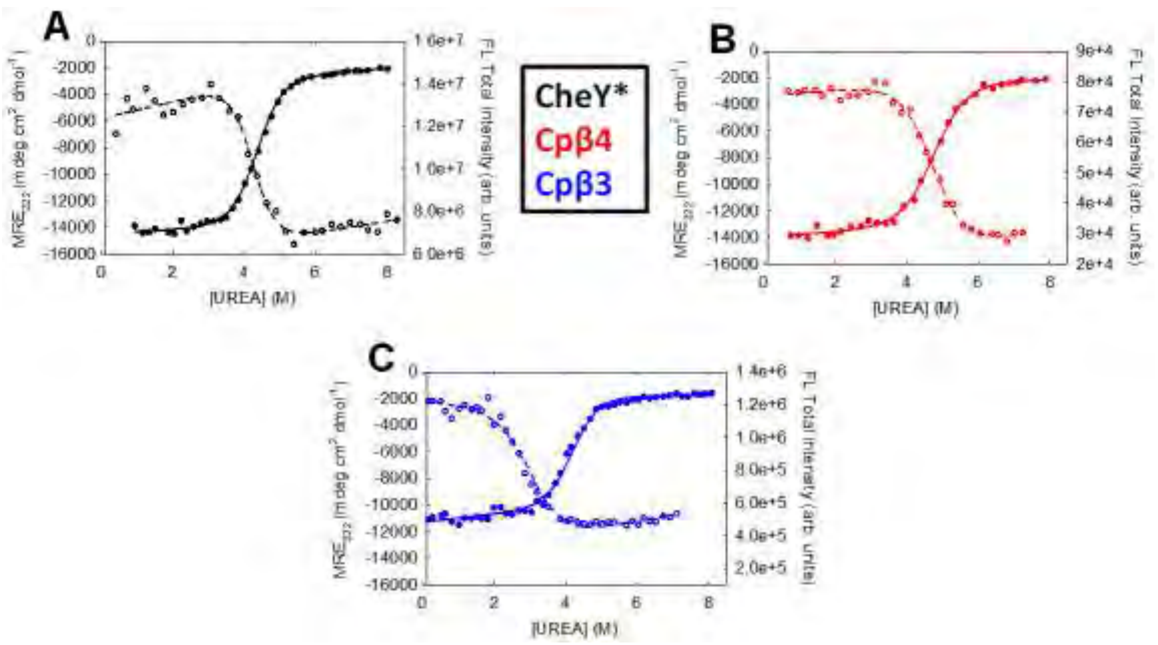


Figure 2.4. Comparison of CD and fluorescence (FL) titrations

Fig. 2.5. Heat capacities of folding transitions. Heat capacity as a function of temperature for CheY* (black), Cpβ4 (red), and Cpβ3 (blue), calculated from the coarse-grained simulation model.

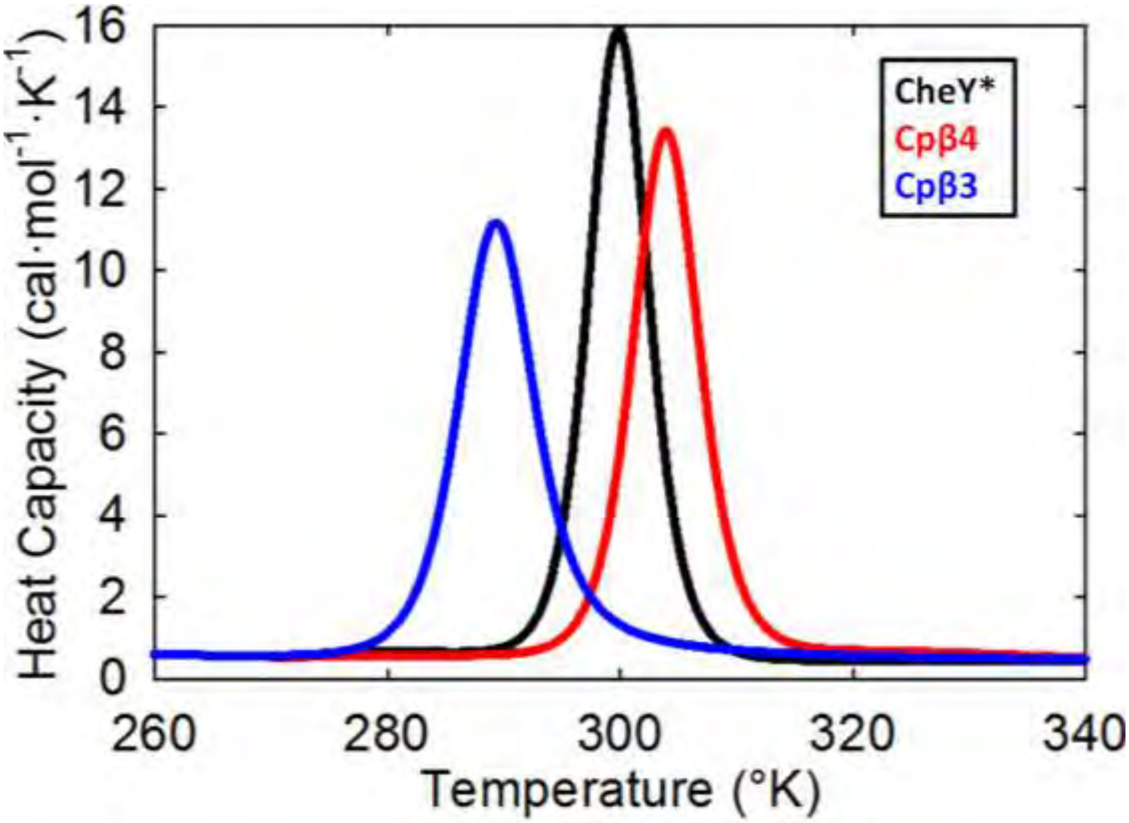


Figure 2.5. Heat capacities of folding transitions

The subsequent formation of the native state occurs in hundreds of seconds and has been attributed to the trans → cis isomerization of the K109-P110 peptide bond[15] (Fig. 2.6). Unfortunately, refolding along the cis channel for the permutants could not be resolved due to its small amplitude in direct refolding experiments and interrupted unfolding experiments. A pair of unfolding reactions were observed in the seconds to hundreds of seconds time range; the interconversion of the native cisP110 conformer to its trans counterpart, $N_c \rightarrow N_t$, controls unfolding in the transition zone and the direct unfolding of the native cisP110 to the unfolded cisP110, $N_c \rightarrow U_c$, controls unfolding at high denaturant concentrations. Similar overall responses were observed for Cpβ3 and Cpβ4, with the exception that the direct unfolding of the native cisP110 conformer was accelerated for Cpβ4.

Stability and secondary structure of I_{BP} states

The orders of magnitude in time separating the μs and 100's of s folding reactions for all three proteins enabled us to measure the stability of the product of the μs reaction, I_{BP}, and its CD spectrum. By plotting the ellipticity at 222 nm after 5 ms of refolding to varying final denaturant concentrations, the stability can be estimated by fitting the resulting titration curve to a 2-state model (Fig. 2.3). The I_{BP} species for Cpβ3 is significantly less stable than CheY*, 0.84 kcal•mol^{-1} vs. 2.02 kcal•mol^{-1}, and the m-value is also decreased (Table 2.1).

Fig. 2.6. Kinetic properties of folding reactions. Refolding (open symbols) and unfolding (closed symbols) relaxation times extracted from manual-mixing (MM), stopped-flow (SF), and continuous-flow (CF) kinetic experiments on CheY* (black), Cpβ4 (red), and Cpβ3 (blue). Ellipticities at 222 nm were monitored by MM and SF techniques (circles), and FL intensities were monitored by MM, SF, and CF techniques (triangles). Error bars reflect SD of n = 12 runs except for CF-FL experiments (n = 2).

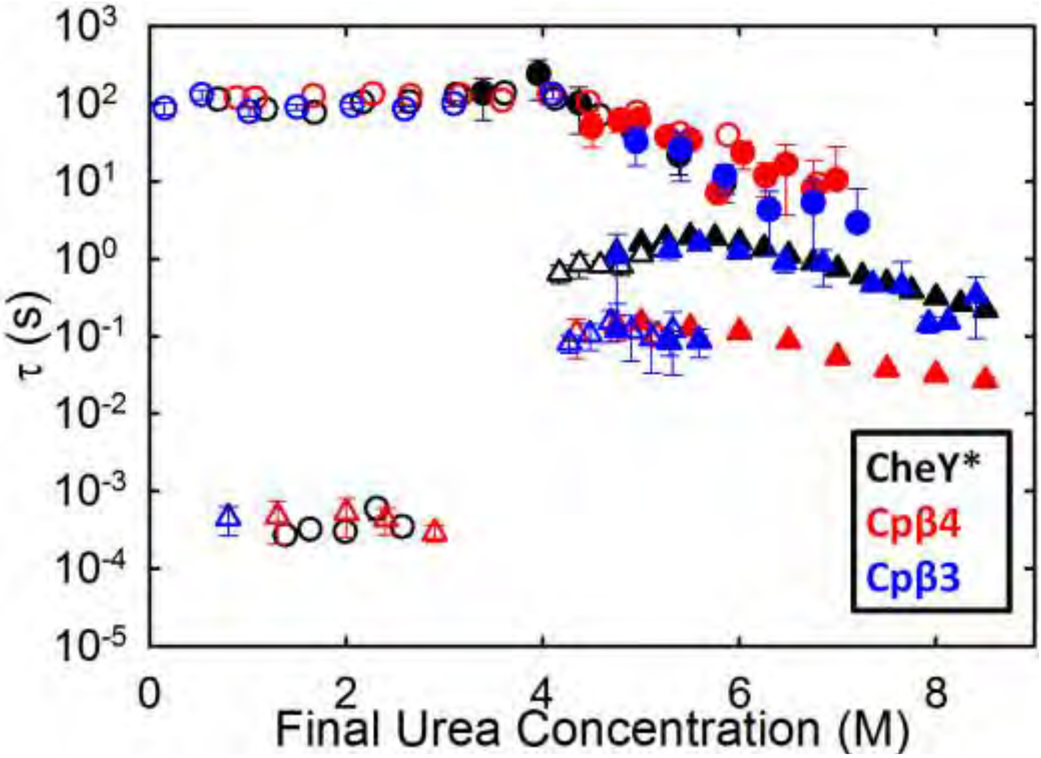


Figure 2.6. Kinetic properties of folding reactions

Very surprisingly, the stability of I_{BP} for Cpβ4 is much greater than CheY*, 4.31 kcal•mol^{-1}, and the m-value is also increased (Table 2.1). Comparison of the denaturation curves for folded Cpβ4 and its I_{BP} species (Fig. 2.3) shows that the two curves overlap between 3 and 5 M urea. By fixing the thermodynamic parameters for the I_{BP}↔U reaction to those extracted from the burst-phase titration data, the stability and m-value for the N↔U reaction could be estimated by fitting the equilibrium titration data for Cpβ4 to a 3-state model. The free energy difference between its native and unfolded forms is 8.19 kcal•mol^{-1} and the m-value is 1.80 (kcal•mol^{-1})•$M_{urea}{}^{-1}$, comparable to CheY*.

We obtained the CD spectra of the I_{BP} species by refolding jumps to the same final urea concentration in the folded baseline and varying the detection wavelengths in the far-UV range. The I_{BP} species for CheY*, Cpβ3, and Cpβ4 recover ~85%, ~80% and ~90% of their native ellipticities at 222 nm within 5 ms (Fig. 2.3). The subtle but significant differences previously observed between the I_{BP} and native states of CheY*[15] (Fig 2.7) indicate that the aromatic side chains have not yet attained their native packing. In contrast, the very similar shapes of the spectra for the I_{BP} and native states of Cpβ4 and Cpβ3 show that an exciton coupling, likely between the side chains in a cluster of phenylalanines on the α1/α5 side of the β-sheet (Fig. 2.7C,D), is present in the I_{BP} state for both permutants.

Fig. 2.7. Far-UV CD spectra of I$_{BP}$. The far-UV CD spectra for natively folded CheY* (black) (A), Cpβ4 (red) (B), and Cpβ3 (blue) (C) are shown as solid lines. The spectra for native CheY* and Cpβ4 are superimposable. Spectra for the corresponding IBP species are shown as dashed lines. Note that in the difference spectrum (dotted lines) in A, CheY* (black) has a negative inflection between 215 and 230 nm indicating a perturbation in exciton coupling. The exciton coupling is not observed in the permuted variants (B and C).

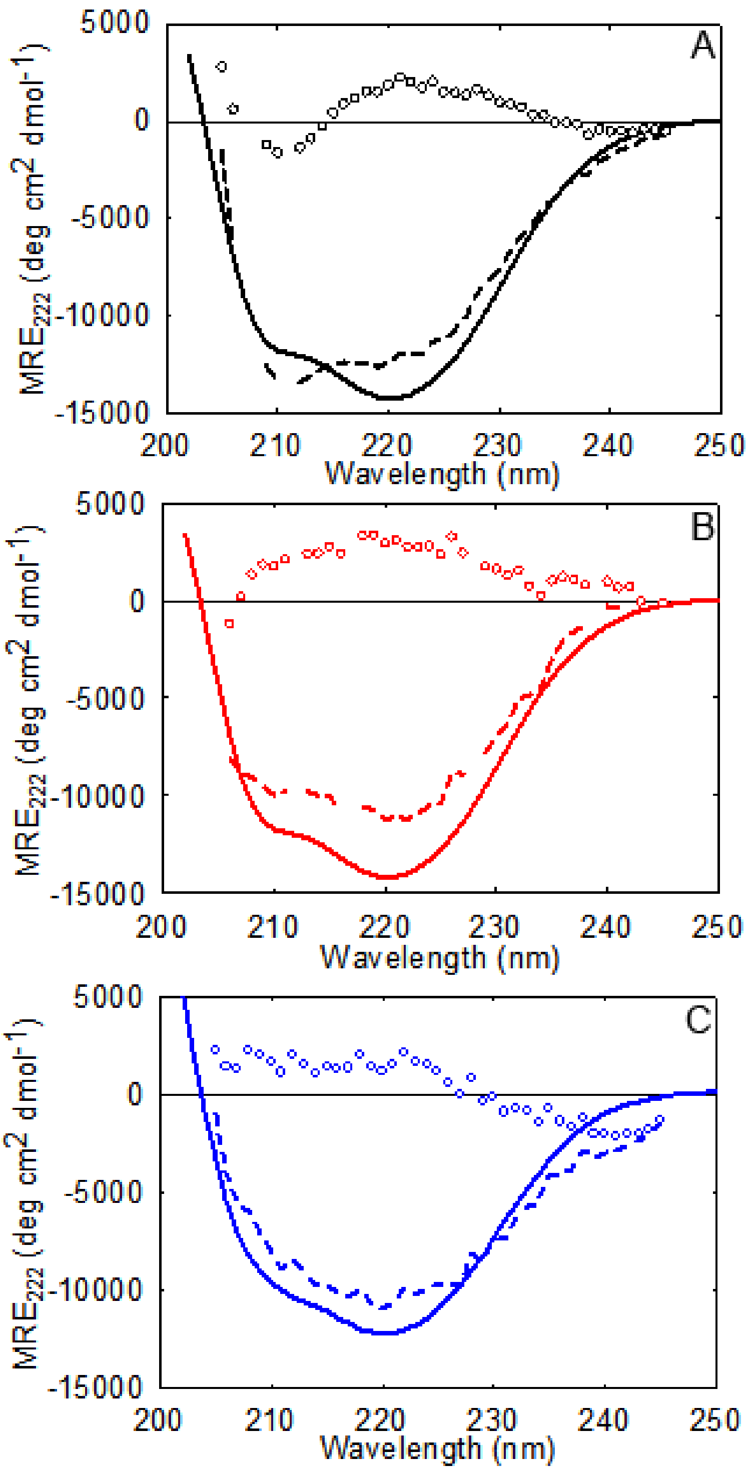


Figure 2.7. Far-UV CD spectra of I_{BP}

Compaction of CheY and Cpβ4 by CF-SAXS*

The very surprising increases in the stability and the apparent compaction for the I_{BP} species for Cpβ4, the latter implied by the increased m-value for its urea melt, motivated us to measure its radius of gyration (R_g) in the ~100 μs to 1 ms time range by CF-SAXS. The urea-denatured states of CheY* and Cpβ4 display R_g's of ~35 Å, slightly smaller than predicted for space filling random coils of 129 amino acids, 38 Å[87]. CheY* collapses to an apparent R_g of ~25 Å within the ~100 μs dead time, experiences a further compaction to ~23 Å by 1 ms, and ultimately contracts to an R_g of 15 Å in the native conformation (Fig. 2.8). In distinct contrast, Cpβ4 collapses to a near-native R_g, ~18 Å, within ~100 μs and remains unchanged after 2.4 ms before contracting to the 15.5 Å R_g of the native state (Fig. 2.8A). Although the change in connectivity does not have a discernible effect on the size of the unfolded ensemble, the cleavage of the chain after β4 and the fusion of the natural N- and C-termini cause Cpβ4 to collapse more rapidly to a near-native radius of gyration.

Topological frustration by simulations

The significant differences in the stabilities of the I_{BP} species of these proteins are surprising given the similarity of the kinetic responses observed. Unfortunately, the small amplitude of the refolding reaction along the cis-channel precluded the use of global analysis to resolve the folding mechanism of the permutant proteins.

Fig. 2.8. Dimensional analysis of CheY* and Cpβ4 during folding by SAXS and simulations. The radius of gyration for CheY* (black) and Cpβ4 (red) from CFSAXS (A) and the average R_g from Gō-model simulations (CheY*: n = 46; Cpβ4: n = 32) in which the intermediate was observed (B) as a function of folding time. Statistical analysis of the simulations finds the intermediate to be highly populated within the average time values of the first and last occurrences (green box; see Table 2.2 for details).The unweighted R_g values of I_{ON} and I_{OFF} species from simulations are shown as dotted lines. Arrows indicate the R_g values and their estimated uncertainties under equilibrium conditions for the folded and the unfolded states (A). Ninety-three points were collected within the mixer channel from 142–2,400 μs and averaged over 20 scans. After low-quality data points were removed, the remaining data were binned into two parts, 142–959 μs and 1,055–2,400 μs. CheY* R_g = 25.3 ± 2.2 Å (n = 11, 142–791 μs) and 22.6 ± 2.0 Å (n = 15 1,223–1,944 μs). Cpβ4 R_g = 18.0 ± 0.7 Å (n = 21, 142–959 μs) and 17.8 ± 0.7 Å (n = 33 1,055–2,304 μs).

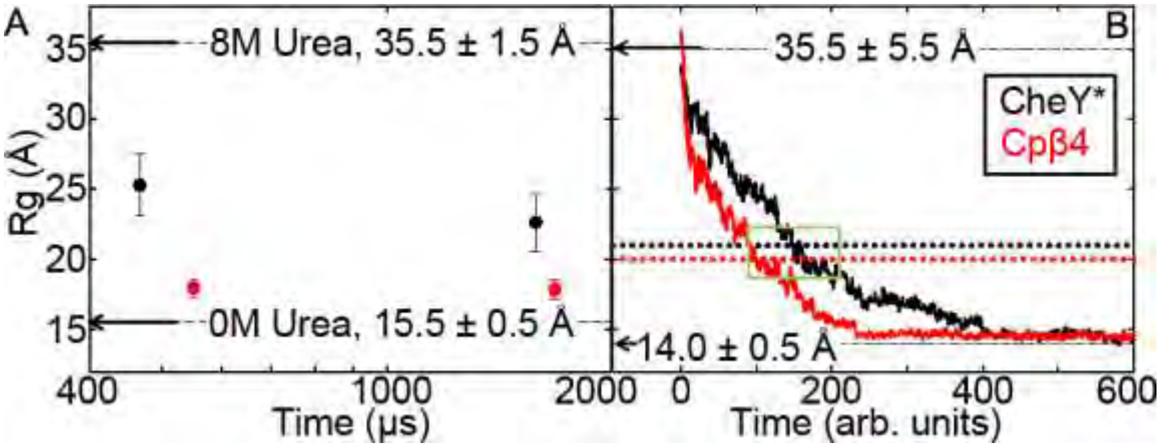


Figure 2.8. Dimensional analysis of CheY* and Cpβ4 during folding by SAXS and simulations

We have therefore used Gō-model simulations to resolve the underlying structural basis of the differences in the I_{BP} stability and inform the kinetic model that is most consistent with the experimental observables. Previous experimental work has concluded that the off-pathway intermediate is not a consequence of the proline isomerization reaction[15,19]. Likewise, in computational work where the trans geometry was enforced via harmonic restraints, CheY was still able to access the folded state from the unfolded configurations. Although, the folded state is destabilized by 2.1 kcal•mol^{-1} relative to flexible Pro110[69], the relative energy landscapes of the cis and trans channels in the native and intermediate states are similar[16,22].

Although we employ a model in which native interactions are predominantly favored, the model can capture frustration arising from the formation of native interactions in an incorrect order[88]. Figure 2.9 shows the influence of chain connectivity on the topological frustration as deduced from folding simulations of CheY* and its circular permutants. Our results are consistent with those reported earlier[22] and show that the folding of CheY* proceeds with significant frustration that arises from the competition of interactions between N-terminal, C-terminal and interfacial native contacts. At Q_{total} = 0.4 local unfolding or backtracking of interfacial contacts (negative slope) between the N-and C-termini coincides with the sudden increase in the contacts of the N-terminal subdomain (Fig. 2.9A,D). These prematurely formed contacts in the C-subdomain partially unfold before folding proceeds to the native conformation. Similar results are observed for Cpβ4, however, the interfacial frustration is markedly reduced (Fig. 2.9B,E). In Cpβ3 this interfacial frustration is absent (Fig. 2.9C,F) because the new termini disrupt the frustrated region. These results are also consistent with the ILV cluster model for folding in that the WT connectivity is driven to fold to the off-pathway I_{BP} species by the premature formation of Cluster 1 spanning the interfacial contacts[15]. The novel local connectivity of the larger cluster (Cluster 2) in Cpβ4 enables it to out-compete the formation of Cluster 1.

Notably, a minor restructuring event in the N-terminal subdomain is observed late in folding at $Q_{total} = 0.6$ in all connectivities. This second event corresponds to the loosening of structure that is routinely observed in the folding of alpha helices prior to final maturation of the tertiary structure and is not comparable to early frustration[88].

Kinetic simulations

More detailed structural insights into the folding mechanisms are gleaned through simulations from the time evolution of R_g and the corresponding time courses of the mean fraction of secondary structure contacts formed for the representative folding trajectories of CheY* and permutants. Cpβ4 collapses faster than CheY* (Fig. 2.8B) before both approach a common R_g of ~14 Å in their respective native conformations. Examination of individual trajectories for

Fig. 2.9. Frustration observed in Gō-model simulations. (A–C) Ensemble averaged fractional contacts of the N-terminal subdomain (dark red), C-terminal subdomain (blue), and subdomain interface (green) are plotted as a function of fractional total native contacts for CheY* (A), Cpβ4 (B), and Cpβ3 (C). (D–F) The interfacial region is dissected in D–F where β3– β4 contacts are shown in magenta, α2–α3 contacts are shown in black, and α5–C-terminal contacts are shown in green. The C-terminal subdomain is dissected into fragments of β4–β5 contacts (gold) and α3–α4 contacts (blue) for CheY* (D), Cpβ4 (E), and Cpβ3 (F).

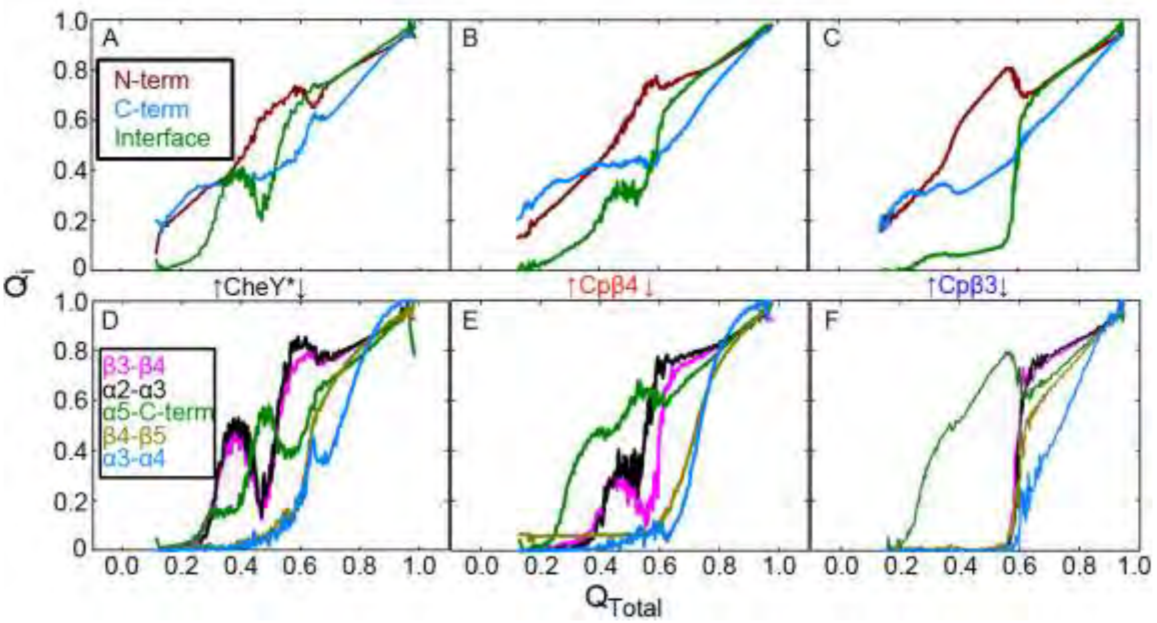


Figure 2.9. Frustration observed in Gō-model simulations

CheY* and Cpβ4 (Fig. 2.10) revealed pauses, reflecting the transient occupancy of partially folded states with discrete R_g values. Of 100 kinetic trajectories, only about half pass through this intermediate and persist long enough to be observable. Therefore the intermediate can be regarded as a non-obligate on-pathway intermediate (I_{ON}). Statistical analysis suggests that the intermediate for Cpβ4 is slightly more compact, 20.2 vs. 21.3 Å, appears earlier, 86 vs. 97 time units, and disappears sooner, 104 vs. 151 time units, than its CheY* counterpart (Fig. 2.8B; Table 2.2). The R_g's for these intermediates are in remarkably good agreement with those observed by SAXS after 1 ms of folding, ~23 Å for CheY* and ~18 Å for Cpβ4 (Fig. 2.8A). The differences in the folding kinetics of CheY* and the permutants may reflect the extent of frustration that arises during folding of each system.

To structurally characterize the intermediates, we extracted structures sampled during kinetic folding simulations that fall within 20 Å < R_g < 22 Å and measured the probability of forming native contacts in this ensemble. The results for CheY* are consistent with previous work[22] where the N_{heptad} was identified as the structured region encompassing the first 7 elements (βα)1-3β4 (Fig. 2.11A). Further, a subsection of the N_{heptad} with the highest probability of contact formation is apparent at the subdomain interface, β3-β4, a region previously described as an area where topological frustration is present[74]. Through a similar analysis of Cpβ4 (Fig. 2.11B) and Cpβ3 (Fig. 2.11C), the differences in chain connectivity were found to have structural repercussions on the early folding intermediates. The probability of forming contacts in the frustrated region is diminished as the N_{heptad} is lengthened to include α5/β5.

Fig. 2.10. Representative simulation trajectories. Time evolution of R_g (A and C) and fractional contact formation (B and D) from representative folding trajectories for CheY* (A and B) and Cpβ4 (C and D).Shown are the time courses of contacts formed in the N terminus (red), C terminus (green), and between β3 and β4 (blue), between α2 and α3 (magenta), between α5 and rest of the protein (cyan), and between α1 and β2 (yellow). For clarity, kinetic traces are shown as moving averages of 10 successive snapshots.

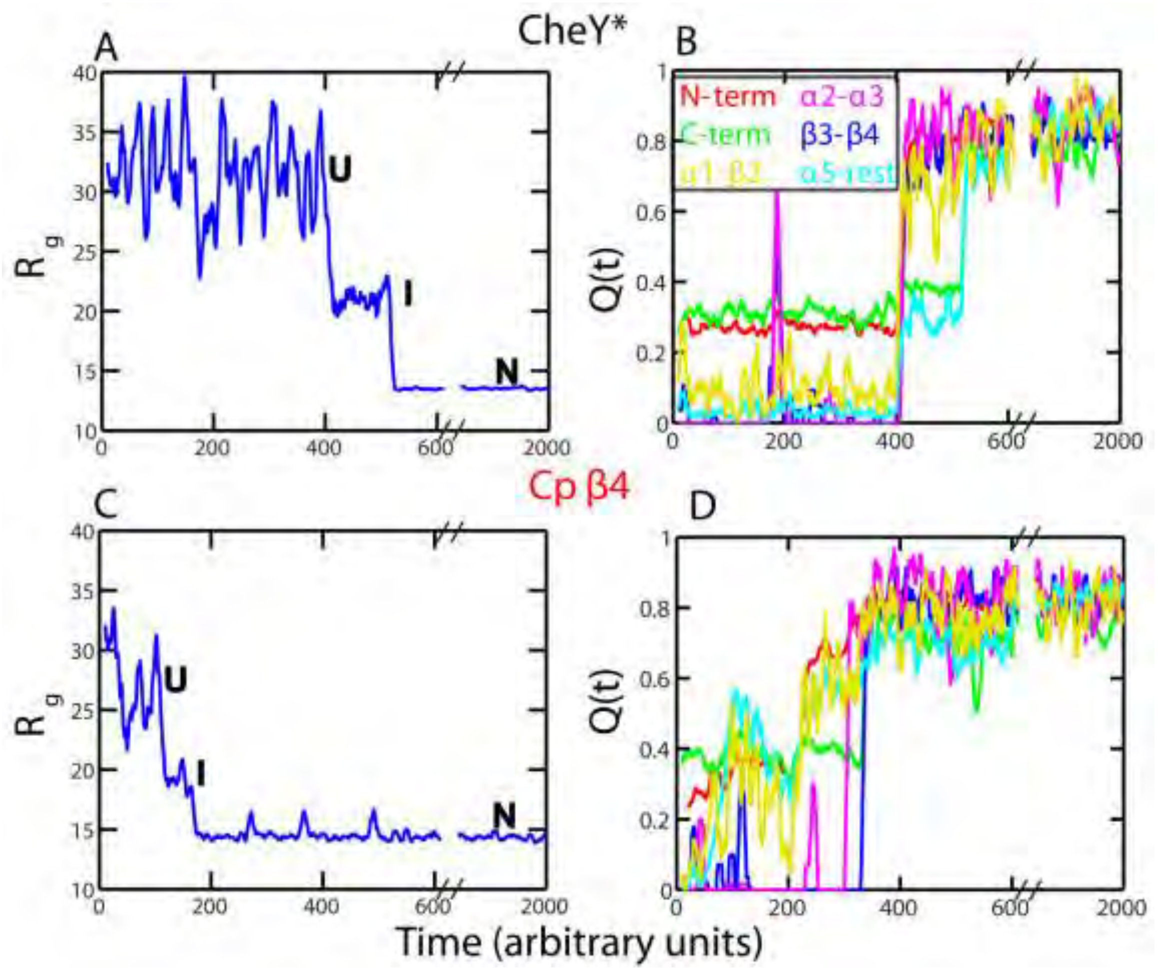


Figure 2.10. Representative simulation trajectories

Fig. 2.11. Probability of contact formation. Probability for pairwise contacts for conformations in kinetic simulations with 20 Å < R_g < 22 Å for CheY* (A) and for conformations with 19 Å < R_g < 21 Å for Cpβ4 (B) and Cpβ3 (C) above the diagonal. Different colors in the contact map indicate different probabilities as quantified by the color scale on the right. The contact maps for CheY*, Cpβ3, and Cpβ4 in their native states are shown below the diagonals. The elements of secondary structure are indicated on the ordinate, with α-helices in green and β-strands in purple. The green box in all three panels indicates the location of the N_{heptad} (I_{ON}), and the smaller green box in the upper left quadrant of B and C indicates the expansion of the Nheptad to include the β1α1/β5 contacts in the permuted proteins. The red ellipse in all three panels indicates the location of the contacts in cluster 1, the region of topological frustration defined as I_{OFF} in CheY*.

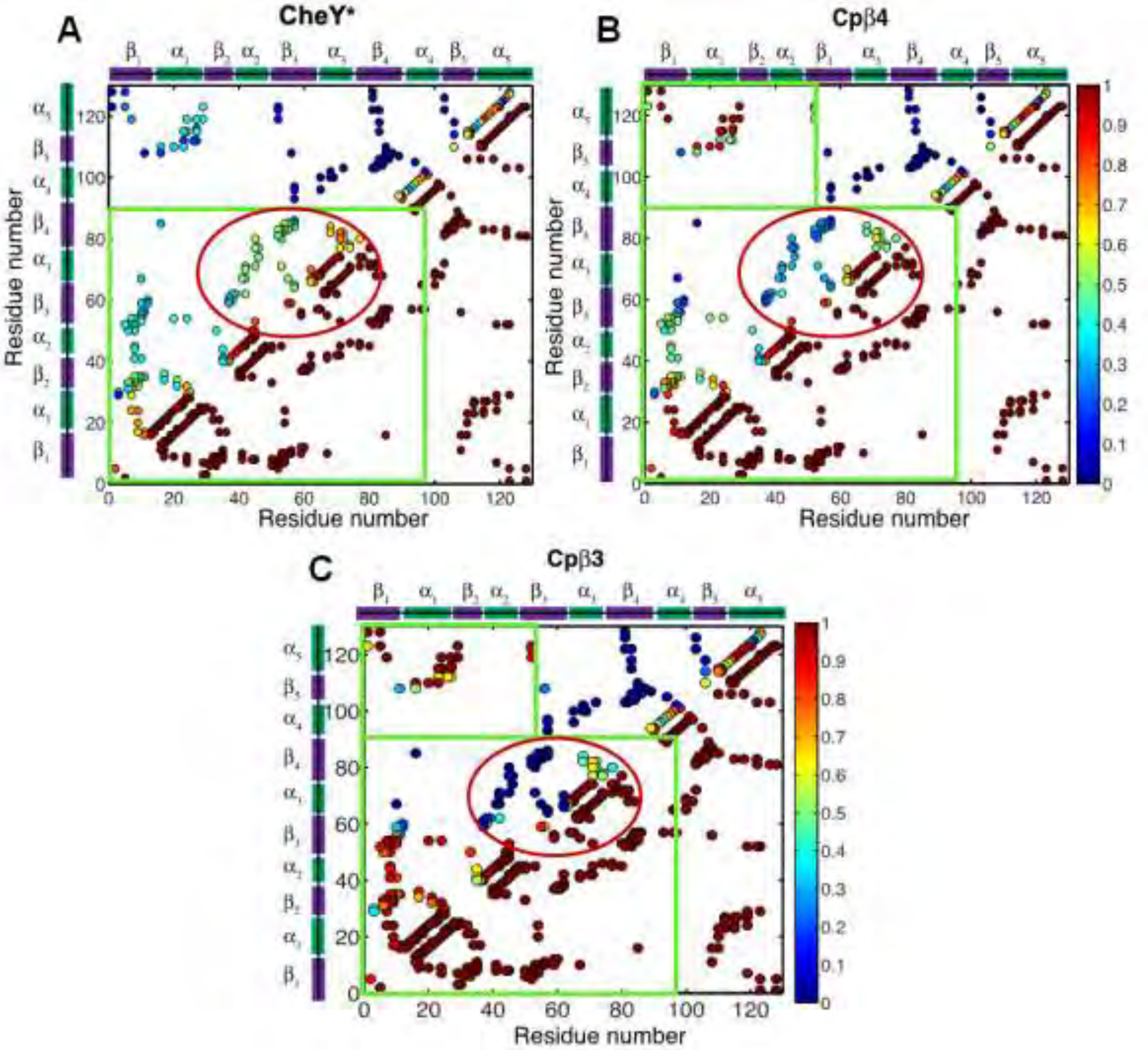


Figure 2.11. Probability of contact formation

Folding@Home Simulations

A landmark 42 milliseconds of all-atom, implicit solvent simulation was collected for CheY*. To account for the proline isomerization reaction the transition probabilities for the isomerization event were calculated using the experimental rate constants. The linked *trans-cis* model yielded a slowest (non-equilibrium) implied timescale of 69 seconds, within the same order of magnitude as the experimentally observed relaxation time of ~100 seconds.

Using Markov State Models (MSMs) to analyze this simulation data, (Jade et. al, unpublished data) we identified a putative structure for the kinetic intermediate of CheY*, which was then compared with small-angle X-ray diffraction data. Theoretical SAXS scattering intensity profiles were calculated for each microstate and the Kratky profiles were compared with the 5-ms experimental SAXS data by fitting a sixth-order polynomial function to the experimental Kratky profile. A two-step method was then used to compare the theoretical Kratky profiles with the experimental data. First, the least squares deviations between the experimental and individual microstate Kratky profiles were computed, and the microstate with minimum least squares deviation was chosen as an initial guess for the structure of the sub-millisecond kinetic intermediate. Then Ensemble Refinement of SAXS (EROS) was used, which is a simulated annealing technique regularized with a prior of maximum entropy, to obtain an optimized ensemble-averaged representation of the intermediate. A similar approach was taken to identify the on-pathway intermediate of Cpβ4.

The top-weighted structures in $I_{BP/OFF}$ are generally similar to the native state. All five native helices and a central β-sheet are formed, and helices α2-4 are packed against the β-sheet in a native-like manner. However, there are also clear non-native motifs present. Namely, the central β-sheet contains a non-native strand packing of β2 against β3 in contrast to the native ordering β2β1β3 (Fig. 2.12A). This excludes the N-terminal strand β1 from the central sheet. As a result, both β1 and α1 cannot dock to the structured part of the protein and are very mobile. Also, the C-terminus is relatively unstructured with α5 frequently being loose and not packed natively.

In contrast to the CheY* results, β1 is in its native position between β2 and β3 in the top-weighted structures of Cpβ4 (Fig 2.12B), and the non-native β2-3 motif of $I_{BP/OFF}$ is not observed. Also, there is a similar degree of native-like packing of α2 and a greater amount of native-like packing of α1 and α5, while α3 is packed in an almost orthogonal orientation to the central sheet and α4 is completely unfolded.

Fig. 2.12. Predicted high probability structures of I_{BP} from atomistic models. CheY* (A) shows non-native strand order of the central β-sheet. In the native arrangement β1 (dark yellow) is intercalated between β2 and β3, in $I_{BP/OFF}$ β1 is not assembled into the β-sheet and α5 (yellow) is misplaced. In Cpβ4 (B) the strand order of the β-sheet is native like along with the packing of most helices. Notably α4 is not folded or packed onto the rest of the structure.

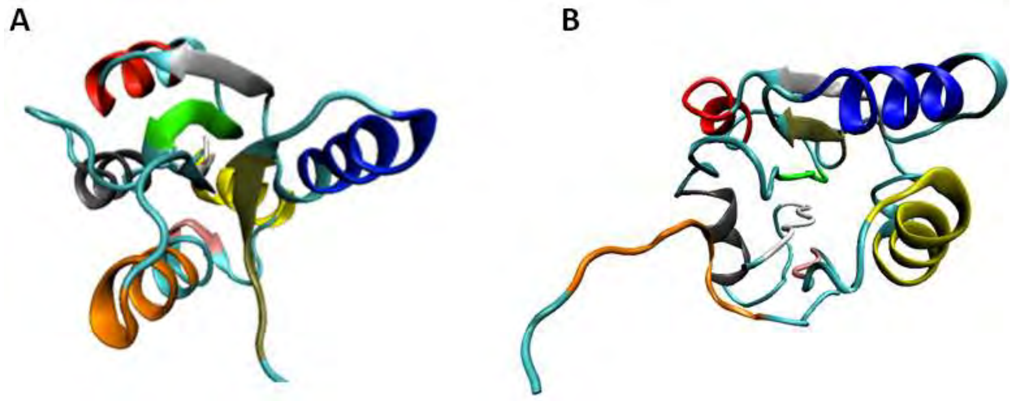


Figure 2.12. Predicted high probability structures of I_{BP} from atomistic models

Discussion

CheY is a member of the Flavodoxin fold family of proteins whose α/β/α sandwich architecture represents one of the more common motifs in biology. Unlike the Flavodoxins, CheY has a conserved cis proline that controls the access to the native conformation[15,18]. Like CheY, however, a pair of homologous Flavodoxins sample a kinetic trap before successfully traversing the productive TSE[71,77]. Elucidating the molecular basis for the frustration in folding for CheY has implications for an entire motif.

CheY*: The results presented here on the F14N variant of CheY are consistent with previous experimental and computational work on the WT protein[15,22]. By Gō-model simulations, topological frustration arises at the subdomain interface before partially unfolding to resume folding from the N-to-C terminus. This result is consistent with experimental data that show non-native Phe packing in the I_{BP} species (Fig. 2.7). The non-native packing of I_{BP} along with the backtracking observed by simulations[15] (Fig. 2.6A,D) and the negative m-value observed through global analysis of experimental data[15] argues that CheY populates an off-pathway kinetic trap, I_{OFF}. Mechanistic details gleaned from the simulations suggest that low CO ILV contacts in Cluster 1 drive early folding events and lead to the premature formation of the subdomain interface. Atomistic simulations suggest that this frustration may also be due to the incorrect strand ordering of the β-sheet (Fig. 2.12A).

Cpβ2: By introducing new termini between β2 and α2, the Cpβ2 permutant cleaves the N-terminal subdomain while leaving Cluster 1 essentially intact and Cluster 2 discontinuous. The observation that Cpβ2 is incapable of adopting the CheY* fold demonstrates an essential role for the intact N-terminal subdomain because Cluster 2 is discontinuous in CheY* and all three permutants. This conclusion is consistent with the results of a previous mutational analysis, where the N-terminal subdomain was found to be a central feature of the productive TSE[19].

Cpβ3: The introduction of new termini between β3 and α3 leaves the two subdomains intact but cleaves both ILV clusters and the N_{heptad}. Notably, the FL and CD titrations are non-coincident, suggesting that multiple species are present prior to the global unfolding reaction. However, because the kinetic response is similar to CheY* under strongly unfolding conditions, Cpβ3 transverses the same barriers as CheY* (Fig. 2.6). The additional faster phase in unfolding may reflect a small fraction of the protein moving through a parallel channel in a limited range of unfolding conditions.

Although the amplitude of the CD spectrum of the I_{BP} species for Cpβ3 is only decreased by ~15% from its CheY* counterpart, the stability is markedly reduced from 2.02 kcal•mol^{-1} for I_{BPt} in CheY* to 0.84 kcal•mol^{-1} in Cpβ3 (Table 2.1) and the m-value is reduced from 0.83 to 0.59 kcal•mol^{-1}•M^{-1}. We attribute the decreased stability of I_{BPt} to the cleavage of Cluster 1, postulated to be a key stabilizing component of the I_{BPt} species for WT CheY[15]. Interestingly, the loss in stability is accompanied by native-like packing of the Phe cluster on the α1/α5 face of the β-sheet (Fig. 2.3D).

Simulations show the elimination of the interfacial frustration of the subdomains for Cpβ3, expected if β3 and β4 are segregated to opposite ends of the chain. The absence of early frustration in the Cpβ3 simulations may reflect the marginal stability of the I_{BP} species, as has been observed previously for a CheY homolog, NT-NtrC[15]. In contrast to CheY*, frustration in Cpβ3 arises late in folding around the β1α1/β5α5 interface on the opposite face of the β-sheet (Fig. 2.13). The high number of native contacts, [Q] values, where this frustration occurs is not consistent with the small m-value for the I_{BP} species for Cpβ3 and likely reflects annealing reactions often seen in the late stages of folding in Gō-model simulations when helix repacking often occurs.

The structural basis for the altered folding properties in Cpβ3 can also be visualized in 2D contact maps derived from the simulations (Fig. 2.11C). For its I_{OFF} species, CheY* has a high probability of contacts in the α2(βα)3β4 region, while Cpβ3 does not. Indeed, the region of high probability of native contacts in Cpβ3 shifts to the β1α1 and β5α5 segments that are covalently linked by permutation of the sequence.

Fig. 2.13. Frustration observed in kinetic simulations of Cpβ3. The mean fraction of contacts formed, Qi, is shown as function of the fraction of native contacts formed in the entire protein, Qtotal, for contacts within α1 (gold), within α5 (blue), between α1 and β5 (dark green), and between α1 and β2 (orange).

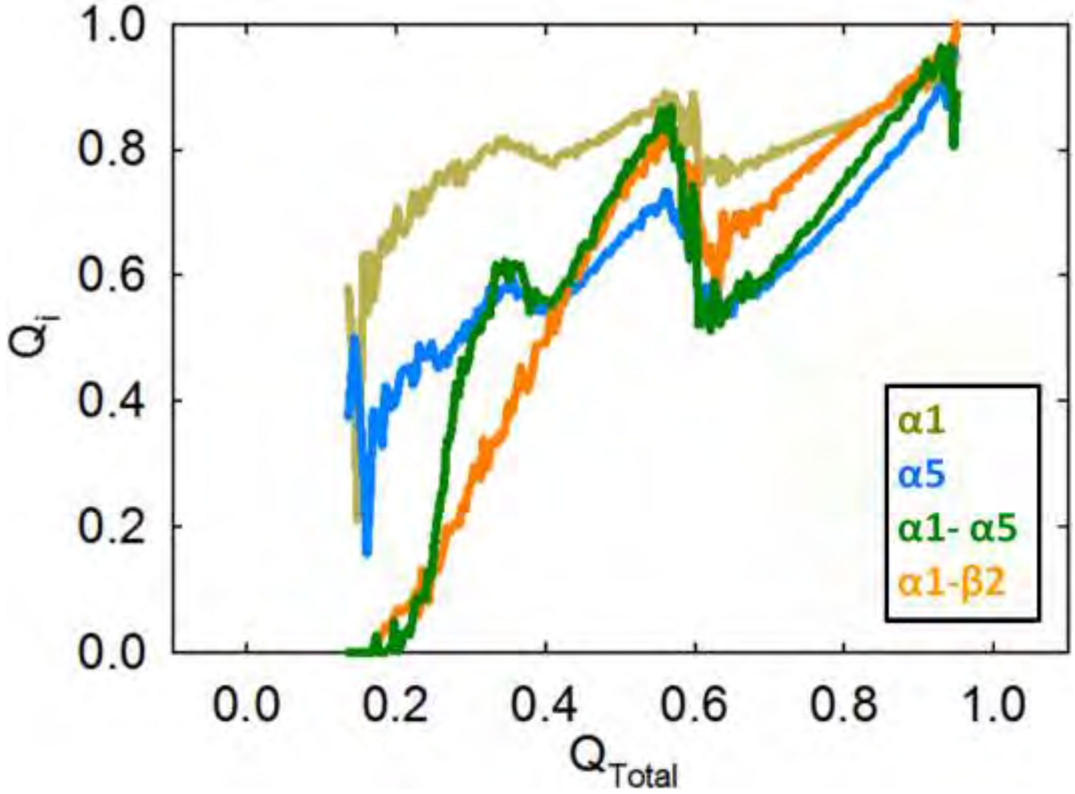


Figure 2.13. Frustration observed in kinetic simulations of Cpβ3

Cpβ4: The introduction of new termini between β4 and α4 in Cpβ4 cleaves the C-terminal subdomain while leaving Cluster 1 intact and Cluster 2 discontinuous. The coincidence of the far-UV CD spectra of CheY* and Cpβ4 (Fig. 2.1) shows that an intact C-terminal subdomain is not essential for proper folding and in agreement with the view that the C-terminal subdomain forms after the TSE[19]. The resultant I_{BP} species folds more rapidly, is both more stable and more compact than CheY* and has native-like packing of its phenylalanine cluster. The increased stability of I_{BP} provides a logical explanation of the accelerated unfolding reaction, via the Hammond effect (Fig. 2.14), and argues for its assignment as an on-pathway intermediate. These surprising experimental results are in very good agreement with the predictions of decreased frustration from an off-pathway intermediate and a more compact on-pathway intermediate including β1, α1, β5 and α5 in the Gō-models and the atomistic simulations.

The 2D contact map of the Cpβ4 folding intermediate reveals an intact N_{heptad} and a high probability for contacts between the covalently-connected β1α1 and β5α5 sequences. The linkage of the natural termini leads to the preferential formation and stabilization of a species that corresponds to the I_{ON} for CheY*. The decreased frustration for Cpβ4 likely reflects both the destabilization of the C-terminal subdomain via cleavage and the increased competition from the more rapidly forming and stable extended N_{heptad}, including the β1α1/ β5α5 complex, which is also suggested to be the case in the atomistic models (Fig. 2.12B).

Fig. 2.14. Structures of intermediates and the simplified folding free-energy surfaces. The sequence of events in folding is indicated by the arrows. The proline isomerization step, occurring between the cis and trans I_{BP} species, is not shown. Structured components of each species as determined by Gō-model simulations. Elements in gray are not yet formed; colored elements [A: black, CheY*; B: red, Cpβ4; C: blue, Cpβ3] are significantly structured; elements implicated in topological frustration are orange. (D) Reaction coordinate diagrams for CheY* (black), Cpβ3 (blue), and Cpβ4 (red). The barrier heights were estimated using the Kramer's formalism with a prefactor of 1 μs, and m-values were calculated from equilibrium and kinetic experiments, when available. Each permutant would have a unique unfolded ensemble, but the free energies have been aligned for direct comparison.

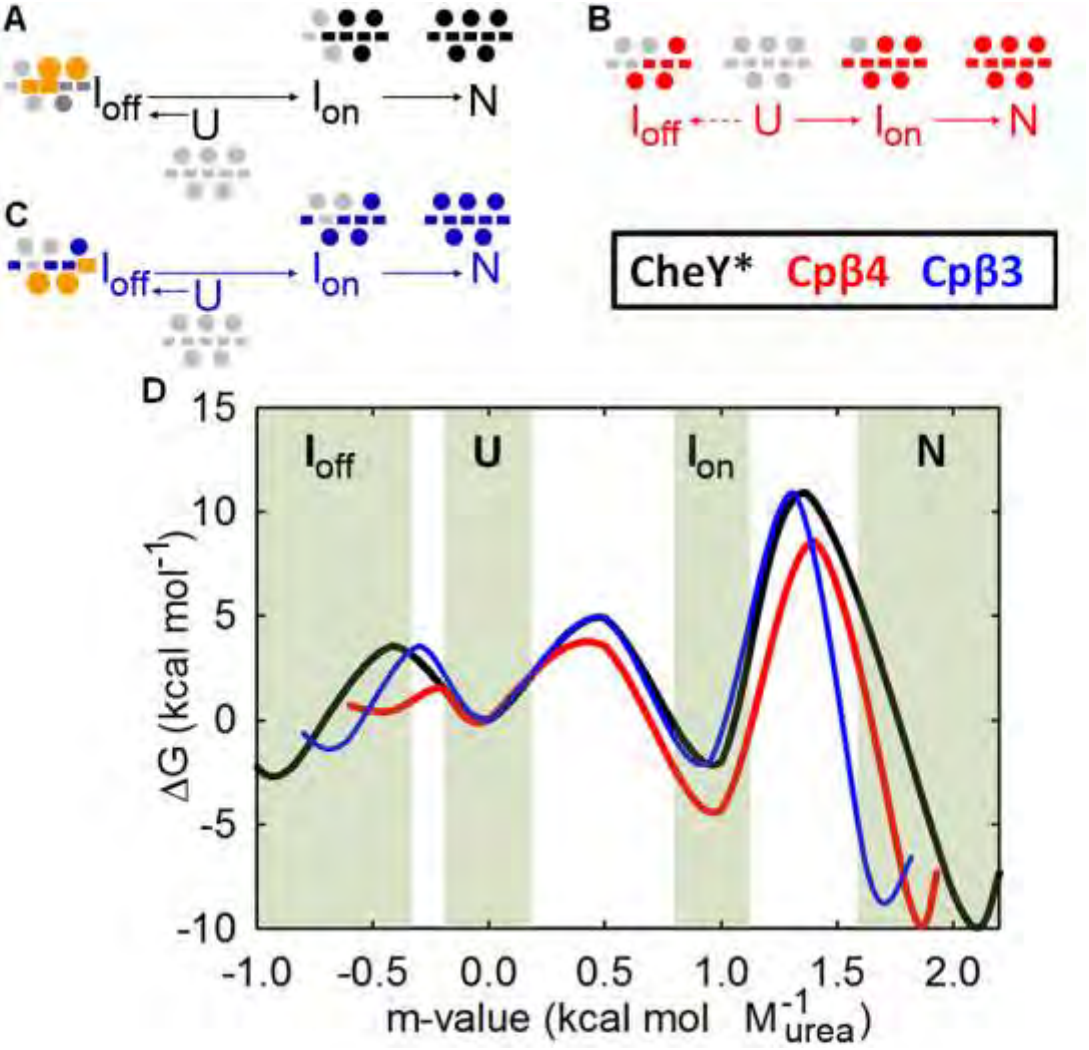


Figure 2.14. Structures of intermediates and the simplified folding free-energy surfaces

Early folding events by CF-SAXS, simulations and CF-FL

The faster collapse of unfolded Cpβ4 observed by CF-SAXS (Fig. 2.8A) and simulations (Fig. 2.8B) is not reflected in the CF-FL data, where essentially identical relaxation times were found for Cpβ4 and CheY* (Fig. 2.6). The discrepancy can be traced to the small m-value for the 300 μs phase and the implied small change in buried surface area accompanying this reaction. The commonality of the relaxation time of this phase for Cpβ3, Cpβ4 and CheY* strongly suggests a local folding event at the single Trp residue that does not reflect the global collapse monitored by CF-SAXS and simulations.

Modulation of the folding landscape by permutations

Both experiments on and simulations of CheY*, Cpβ3 and Cpβ4 reveal that the initial events in the folding are dictated by the connectivity of the chain. In another case, Cpβ2, altering the chain connectivity leads to a distinctly different but well-defined thermodynamic state. The combined results for those sequences that can attain the wild-type native conformation can be displayed on a reaction coordinate diagram shown in Figure 2.14D; the proposed structured elements for the various species are shown in Figure 2.14A-C.

The path from the unfolded state to the respective intermediate for CheY*, Cpβ3 and Cpβ4 is controlled by preferred interactions between low CO elements of secondary structure. The varying structures, stabilities and buried surface areas for these partially-folded states can be understood in terms of the thermodynamic compulsion to minimize the chain entropy penalty and maximize the participation of their resident aliphatic side chains in one of 2 ILV clusters located on either face of the central β-sheet. For CheY*, Cluster 1 forms early and stabilizes I_{off}. For Cpβ3, Cluster 1 is cleaved and a fraction of Cluster 2 drives the formation of a poorly-folded fragile I_{off}. For Cpβ4, the C-terminal elements of Cluster 2 reinforce the N_{heptad}, resulting in a remarkably stable I_{on}. Thus, the folding free energy surface of CheY and its attendant frustration in folding can be modulated either by the destabilization of the off-pathway intermediate, Cpβ3, or by the stabilization of an on-pathway intermediate, Cpβ4. Although the initial sources of frustration for these permuted sequences are quite different, all can achieve essentially the same native conformation.

Subdomain vs. ILV cluster model for the folding of CheY

The totality of the results suggests that the ILV cluster model provides the more parsimonious and complete description of the early events in folding but that the subdomain model better captures the crucial TSE required to access the proper native fold. In other words, low CO clusters of ILV residues can strongly influence the early stages of folding before subdomain and global cooperativity engage expanding portions of the sequence to reach the native conformation.

Perspective

Chain entropy plays a crucial role in defining the energies and structures of partially-folded states on the folding free energy surface of CheY. Thus, frustration can be modulated and productive folding favored by altering the sequence connectivity and, thereby, the local chain entropy. The local-in-sequence local-in-space topology of βα-repeat proteins, including the Rossmann-fold, TIM barrels and the Flavodoxin/CheY folds, make them prime candidates for frustration in the early stages of folding. The associated partially-folded states may not only impede the folding reaction, but also may serve to nucleate aggregation reactions in pathological sequence variants. Recognition of the early events in folding and the partially-folded structures that they produce provides a rational basis for the design of small molecules that might inhibit aggregation by binding at the interfaces of these nascent kernels of structure.

Methods

Protein expression and purification

All CheY variants were engineered with an N-terminal hexahistidine tag and an intervening Tev Protease site (GenScript), ligated into the expression plasmid pGS-21a, and transfected into the *E. coli* strain BL21 Codonplus® (DE3)RIL for expression. All proteins were isolated from inclusion bodies by dissolving the insoluble fraction of the cell lysate in 8 M urea and refolding into 10 mM potassium phosphate buffer at pH 7.0. Precipitates were removed using centrifugation and the soluble fraction was bound to a nickel resin overnight at room temperature. The nickel resin was then thoroughly washed and the His-tagged protein was eluted with a step gradient of 10 mM, 25 mM, and 300 mM imidazole. Pure fractions were pooled and dialyzed into 50 mM tris, 1 M urea, pH 7.8. Overnight incubation with His-tagged Tev Protease was used to cleave off the His-tag from CheY. The reaction product was bound to a nickel resin overnight and the cleaved protein was washed off of the resin in 10 mM potassium phosphate buffer at pH 7.0. Protein was concentrated and applied to a Q Sepharose column and eluted using a salt gradient from 0 to 400 mM NaCl. The purity was confirmed (> 98%) using a Waters Q-TOF ESI mass spectrometer.

Protein sequences

CheY* (129 AA): G A D K E L K F L V V D D N S T M R R I V R N L L K E L G FN N V E E A E D G V D A L N K L Q A G G Y G F V I S D W N M P N M D G L E L L K T I R A D G A M S A L P V L M V T A E A K K E N I I A A A Q A G A S G Y V V K P F T A A T L E E K L N K I F E K L G M

Cpβ2 (132 AA): G D G V D A L N K L Q A G G Y G F V I S D W N M P N M D G L E L L K T I R A D G A M S A L P V L M V T A E A K K E N I I A A A Q A G A S G Y V V K P F T A A T L E E K L N K I F E K L G M G A G A D K E L K F L V V D D N S T M R R I V R N L L K E L G F N N V E E A E

Cpβ3 (131 AA): G D G L E L L K T I R A D G A M S A L P V L M V T A E A K K E N I I A A A Q A G A S G Y V V K P F T A A T L E E K L N K I F E K L G M G A G A D K E L K F L V V D D N S T M R R I V R N L L K E L G F N N V E E A E D G V D A L N K L Q A G G Y G F V I S D W N M P N

Cpβ4 (132 AA): G E A K K E N I I A A A Q A G A S G Y V V K P F T A A T L EE K L N K I F E K L G M G A G A D K E L K F L V V D D N S T M R R I V R N L L K E L G F N N V E E A E D G V D A L N K L Q A G G Y G F V I S D W N M P N M D G L E L L K T I R A D G A M S A L P V L M V T A

Native state Analysis

Circular Dichroism Structure Analysis

Far-UV CD spectra were collected on a JASCO model J810 CD spectrophotometer . All samples were buffered with 10 mM potassium phosphate at pH 7.0 and 25^o C. Measurements were taken in a 0.5 cm path length cuvette with a bandwidth of 2.5 nm and a step size of 0.5 nm at a protein concentration of approximately 6 μM from 202 nm to 260 nm, with the exception of Cpβ2 which was recorded from 198 nm to 260 nm. Three buffer subtracted spectra were collected and averaged for each protein with a total averaging time of 3 s per wavelength.

Equilibrium unfolding and refolding

Unfolded (9 M urea) and folded (buffer) stocks of protein were diluted with varying concentrations of 10 mM potassium phosphate buffered urea at pH 7.0 to achieve a final protein concentration of approximately 6 μM protein and a range of final urea concentrations from 0 M to 8.25 M. Each sample was thoroughly mixed and left to equilibrate overnight at room temperature. CD spectra were collected on a JASCO model J810 spectrophotometer under the same conditions as the wavelength scan, and by intrinsic tryptophan fluorescence using Horiba

Flourolog3 with a slit width of 5 nm, excitation at 295 nm with a 0.5 cm path length, and emission spectra collected above 310 nm with a 1 cm path length at 25º C. The fluorescence and CD data were globally fit to a two state N⇌U model as a function of urea using our in-house Savuka software package.

I_{BP} Analysis

I_{BP} CD Wavelength Scans

Urea denatured protein in 10 mM potassium phosphate at pH 7.0 was refolded with a 10-fold dilution of 10 mM potassium phosphate pH 7.0 at 25º C. Measurements were recorded at each wavelength from 205 nm to 240 nm using a 0.2 cm cuvette at a final protein concentration of approximately 10 µM with a bandwidth of 2.5 nm in triplicate. Buffer subtraction of the data and exponential extrapolation to 0 time provided the amplitude of the I_{BP} refolding reaction at each wavelength. The difference of U→N and U→I_{BP} amplitudes is plotted as a function of wavelength in order to construct the CD spectra of the I_{BP} species.

Continuous-flow intrinsic tryptophan fluorescence

The all-quartz mixer was custom made by Translume, Inc. with channels 50 µm wide and 100 µm high. Nanoport connectors for use with 1/32 inch outer diameter PEEK tubing (Upchurch) were attached using UV curing epoxy by the manufacturer (Translume, Inc.). Trp fluorescence utilized 292 nm excitation from the tripled-output of a Ti:sapphire laser. A blank and NATA control were acquired for each trace to correct for background fluorescence (typically ~1%)

and variations in excitation intensity along the channel (typically < 5%), respectively. Computer controlled scanning of the channel was accomplished by an x-y translation stage (Biopoint 2, Ludl Scientific).

Dimensional analysis

Equilibrium small angle x-ray scattering

Equilibrium measurements were collected as previously described [89]. The protein concentration was 1.5 mg·mL^{-1} in 10 mM potassium phosphate buffer at pH 7.0 and 25º C.

Continuous-flow small angle x-ray scattering

Continuous-flow SAXS measurements were made as previously described[90]. The total flow rate was 20 mL·min^{-1} using a 1:10 dilution of the unfolded protein for a final protein concentration of 1.5 mg·mL^{-1} in 10 mM potassium phosphate buffered 8 M urea at pH 7.0 and refolding with 0 M urea buffer.

Small Angle X-Ray Scattering Data Analysis

Radial averaging of the raw SAXS data images was accomplished using IGOR Pro (WaveMetrics) macros written by the BioCAT staff at APS. The exported scattering profiles were imported into in-house software for further analysis. The R_g values were obtained based on the Guinier approximation within the Guinier region ($R_gQ_{max} \leq 1.3$).

Gō-model simulations

System preparation and model

Cpβ4 and Cpβ3 were modeled on the crystal structure of WT CheY from *E. coli* (PDB ID: 3CHY)[21]. Models of both permutants were constructed by joining together the N and C termini of the CheY* with a Gly-Ala-Gly peptide and cleaving the bond between residues 63 and 64 and residues 88 and 89 for Cpβ3 and Cpβ4 permutant, respectively. The protein folding simulations were performed with an unrestrained prolyl-bond geometry using a coarse-grained model developed by Karanicolas and Brooks[91]. All simulations were performed using coarse grained Gō-like model that has been previously successfully applied to study protein-folding mechanisms of several proteins[91]. In the model, the protein backbone is represented as a string of beads connected by virtual bonds. Each bead represents a single amino acid and is centered at its alpha carbon position. Adjacent beads on a string are held together with the potential encoding bond angle and bond length constraints. Specifically, bond length is kept fixed while bond angle interactions are harmonically restrained. In addition, dihedral angles are subject to potentials to mimic backbone chirality and Ramachandran conformational preferences. Nonbonded interactions are represented using a model in which only residues that are in contact in the native state interact favorably. Backbone hydrogen bonds and side-chain pairs with non-hydrogen atoms separated by less than 4.5 Å interact via a modified Lennard-Jones potential that consists of energy well and a small desolvation barrier to ensure folding cooperativity. The sequence dependence is accounted for by weighing the strength of side-chain contacts according to their abundance in the Protein Data Bank (PDB)[92]. All non-native interactions experience volume-exclusion repulsion. The detailed description of this energy function can be found elsewhere[91,93].

Molecular dynamics protocol

Molecular dynamics simulations were performed using the CHARMM macromolecular mechanics package[94]. All models were evolved through Langevin dynamics, by using a friction coefficient of 1.36 ps^{-1} and a molecular dynamics time step of 22 fs. The virtual bond lengths were kept fixed using the SHAKE algorithm. For each permutant, 100 independent folding simulations were each performed for 2×10^8 dynamics steps at 0.87 T_f, where T_f is the folding transition temperature estimated as a temperature corresponding to the peak in the specific heat curve, C_v (T) (See Fig. 2.5).

Replica exchange simulations

For thermodynamic characterization, specifically to estimate heat capacity as a function of temperature for all systems (CheY*, Cpβ3 and Cpβ4), we performed two-dimensional replica exchange simulations in which each one of total 28 replicas was restrained in the chosen temperature (0.87, 0.97, 1.08, or 1.20 T_f) and radius of gyration value, R_g, (1.0, 1.1, 1.2, 1.3, 1.5, 1.7, and 2.0 R^0_g, where R^0_g is the radius of gyration of the native folded state), with force constants 0.5, 5.0, 5.0, 5.0, 4.0, 0.8, and 0.5 kcal/mol/$Å^2$ respectively. Conformational exchanges between temperature windows and restraints were attempted every 40,000 dynamics steps. All structures from different temperatures were combined and unbiased using weighted histogram (WHAM) method[95]. Finally, heat capacity was obtained from fluctuations of the potential energy and plotted as a function of temperature for all systems (See Fig. 2.7).

Molecular dynamics protocol for kinetic simulations

First, unfolded starting structures for the folding simulations were generated by equilibration dynamics at 1.5 T_f (note that the T_f is different for WT CheY, Cpb3 and Cpb4) for 10^7 molecular dynamics steps starting from randomly assigned initial velocities. Following, kinetic folding simulations were performed for 2×10^8 dynamics steps at 0.87 T_f and protein coordinates were saved every 10^5 dynamics steps. The fraction of native contacts formed, Q, was used to monitor the folding progress. Each contact was considered formed if its residue pair was within a cutoff distance chosen such that the given contact is satisfied 85% of the time in native-state simulations at 0.83 $T_{f.}$ The Gō-models were built from the PDB structure in which peptide bond is in cis configuration. We performed unrestrained kinetic simulations without placing any specific harmonic restraint on the dihedral to enforce sampling of the cis configuration.

Chapter III - Residual Structure in the Unfolded State of Di-III_14

This chapter is a body of work that is being prepared for publication. The data herein are the results of my own work as well as the work of Laura Deveau, a Ph.D. candidate working under the supervision of Dr. Francesca Massi. Laura has contributed the NMR experiments and analysis. I have contributed the thermodynamic and kinetic characterization. The data interpretation and manuscript preparation is the work of myself and Dr. C. Robert Matthews, Dr. Francesca Massi, and Laura Deveau.

Introduction

The unfolded state ensemble is the least accessible and most poorly understood region of the protein folding free-energy landscape. Statistical thermodynamics describes this state as largely dynamic with few, if any, persistent features[87]. In the current understanding, a protein heteropolymer is often approximated to be a statistical random coil due to the assumed high conformation entropy of the state. Although this appears to remain true in the presence of chemical denaturants[96], it is unlikely to be true in aqueous (i.e., native-favoring) conditions. Protein heteropolymers consists of diverse sidechain chemistries and are composed largely (40-50%) of hydrophobic character in soluble globular proteins[97]. The effect of these chemistries within the primary sequence manifests in the phenomenon that protein chains in aqueous solvent behave as a polymer within a poor solvent, favoring intra-polymer contacts over polymer-solvent interactions[77]. Therefore, it stands to reason that a given protein in dynamic equilibrium will populate a collapsed and structurally biased unfolded state under native-favoring conditions. This biologically accessible unfolded state is arguably the biologically relevant unfolded state that may lead to protein misfolding and aggregation, contributing to such diseases as amyotrophic lateral sclerosis[98] and Alzheimer's disease[99]. Unfortunately direct observation of residual structure in the unfolded state under native conditions is limited due to technical difficulties associated with making such measurements, namely the short lifetime of the species, relatively long experimental acquisition timescales, and limited experimental resolution. However, a recently designed unnatural protein, Di-III_14[70], exhibits intrinsic properties that make these measurements accessible, leading to new insights towards what structural elements may persist in the unfolded state under native conditions.

The rational design of unnatural proteins is made possible by the ever increasing computational power and the continual development of efficient algorithms for structure prediction[32]. However, little is known about energy surfaces that are not subjected to eons of evolution, or even if the free-energy surface itself is subject to evolutionary pressure. A case in point is the design of Top7 by the Baker group[32], the first *de novo* designed novel protein fold. In the seminal publication this protein was demonstrated to be thermodynamically 2-state, and very stable ($\Delta G = -13.2$ kcal mol^{-1}). However, in a subsequent study the folding kinetics were examined and found to be multi-state and exceedingly complicated[33]. The folding mechanism was also suggested to be less cooperative than previously thought due to elements that are capable of folding independently[100]. The sum of these results obviate the caveat of structure-based design; structure does not predict folding kinetics.

Further work from Baker and colleagues implemented kinetic folding trajectories in order to formulate design rules to favor 5 naturally occurring βα-repeat fold motifs. The basis of this approach was that successfully biasing the kinetics would produce a smooth folding landscape[70]. Similar to previous work, the proteins again appear to be thermodynamically 2-state. The small chain lengths, less than 100 amino acids, of these novel proteins combined with their kinetic optimization through simulations suggests that they will fold via a 2-state mechanism. Proteins containing less than 100 amino acids typically demonstrate 2-state folding behavior, presumably due to the small size not being amenable to the formation of independently stable subdomains[101]. However the experimental exploration of the energy surface by way of folding kinetic experiments on these designed proteins remains to be accomplished.

The folding dynamics of a small β1α1β2α2β4β3 protein, Di-III_14, was designed to mimic a natural βαβ fold. It's small size, at 89 amino acids, low contact order, and single hydrophobic core should eliminate the presence of kinetic complexities[15,64] including domain competition[100,102], and competing hydrophobic cores[15,103]. The topology of Di-III_14 consists of a 4-stranded β-sheet with α1 and α2 paired on the same side of the sheet (Fig 3.1A). A single closely packed hydrophobic core of 9 tertiary isoleucine, leucine, and valine (ILV) contacts connects the two structural features (Fig 3.1B). On the solvent exposed surface and central to these contacts on the β-sheet side, there exist 3 salt bridges connecting the two central stands of the β-sheet. The intercalated β4 strand is highly cationic with 3 positively charged solvent exposed residues. These residues establish the 3 salt bridges with the adjacent and compensatory anionic central β2 strand (Fig 3.1C), suggesting an engineered mode of N state stabilization through a reduction in chain entropy that was not previously claimed to be rationally engineered.

Comparison of the results of chemical denaturation with those of native-state hydrogen exchange experiments (HDX) reveal a complex response consistent with the retention of native-like topology in the unfolded state of D_III-14 in the absence of denaturant. Although not explicitly a feature of the design criteria, the unfolded state of Di-III_14 appears to be significantly structurally biased towards the native state to which it folds to in tens of microseconds. The conservation of the native topology of the sequence before folding begins demonstrates its crucial role throughout the folding reaction coordinate and establishes Di-III_14 as a model protein and engineering platform that completely avoids kinetic traps.

Fig 3.1. Structural elements of Di-III_14. Di-III_14 is a βα-repeat protein with a terminal β-hairpin. (A) The topology of Di-III_14 consists of a β-sheet with 2 α-helices packed onto one side. (B) Calculated ILV clusters bind these elements of secondary structure together. (C) Salt bridges are exposed on the solvent accessible side of the β-sheet. The negatively charged residues (black) are well segregated in sequence from the positively charged residues (red) in the central β-strands.

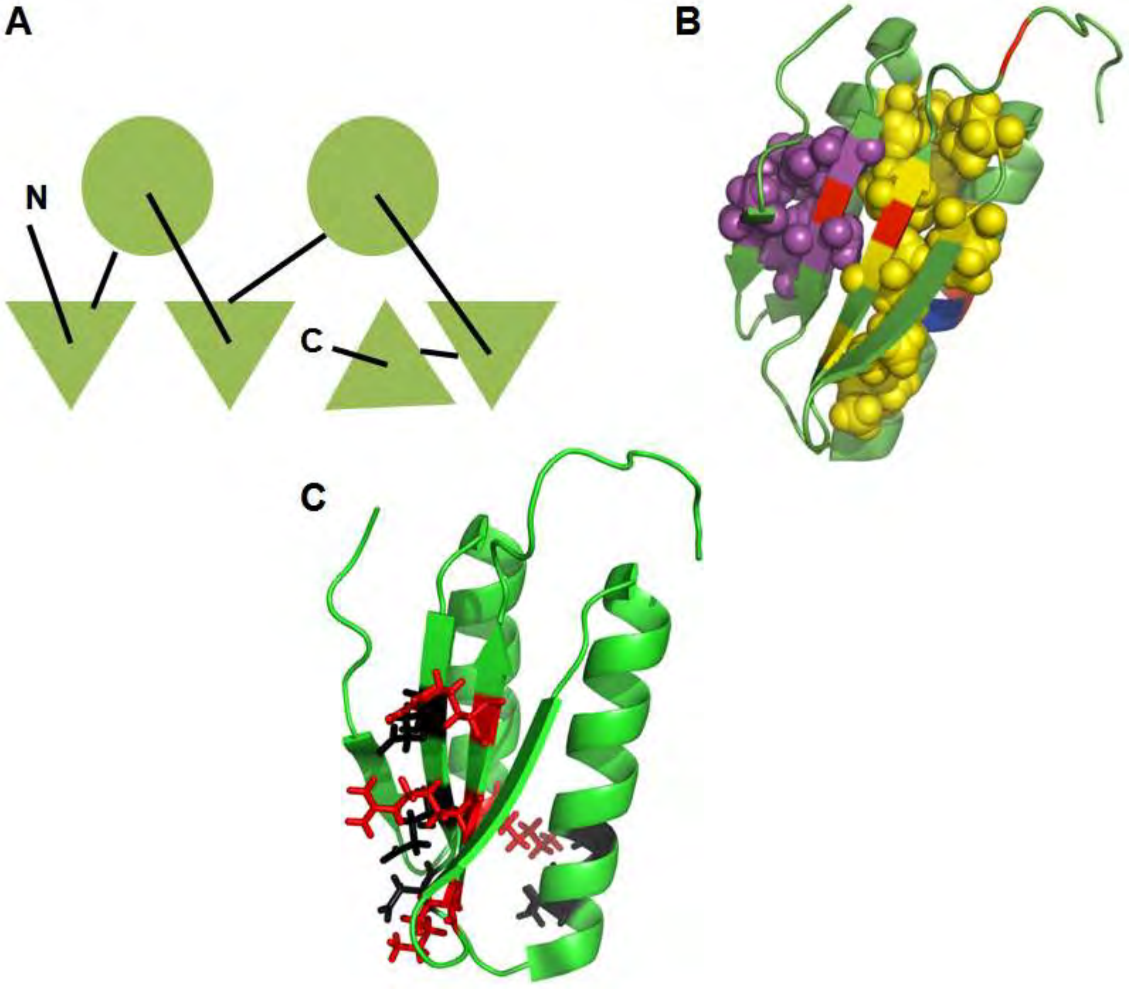


Figure 3.1. Structural elements of Di-III_14

Results

Di-III_14 tagless construct

Previous work by *Koga et. al.*[70] correctly predicted the structure of Di-III_14 computationally without a hexahistidine-tag (his-tag) however the demonstrated 2-state behavior of Di-III_14 obtained by guanidine denaturant (Gnd) melts were conducted on a construct with an additional C-terminal his-tag. We redesigned Di-III_14 to have a cleavable N-terminal his-tag to determine the importance of the tag in the perceived stability. The tagless variant has a global stability of 5.43 ± 0.36 kcal mol^{-1} and an m-value of 2.34 ± 0.13 kcal mol^{-1} M_{Gnd}^{-1} and appears to be unaffected across the pH range of pH 6.0 to 7.4 (Fig 3.2). Compared to the N-tagged variant, there is no observable effect on the 2-state global stability of the protein (Table 3.1). Unfortunately, upon cleavage of the tag, the protein readily aggregates upon reconstitution of lyophilized product to high concentrations (~400 μM) precluding HDX NMR studies.

Fig 3.2. Gnd Denaturation melts of tagless Di-III_14 at 222nm by CD. Gnd denaturation melts of the tagless construct are pH independent from pH 6.0 (blue) to pH 7.4 (black). The data was globally fit to a 2-state model (solid lines).

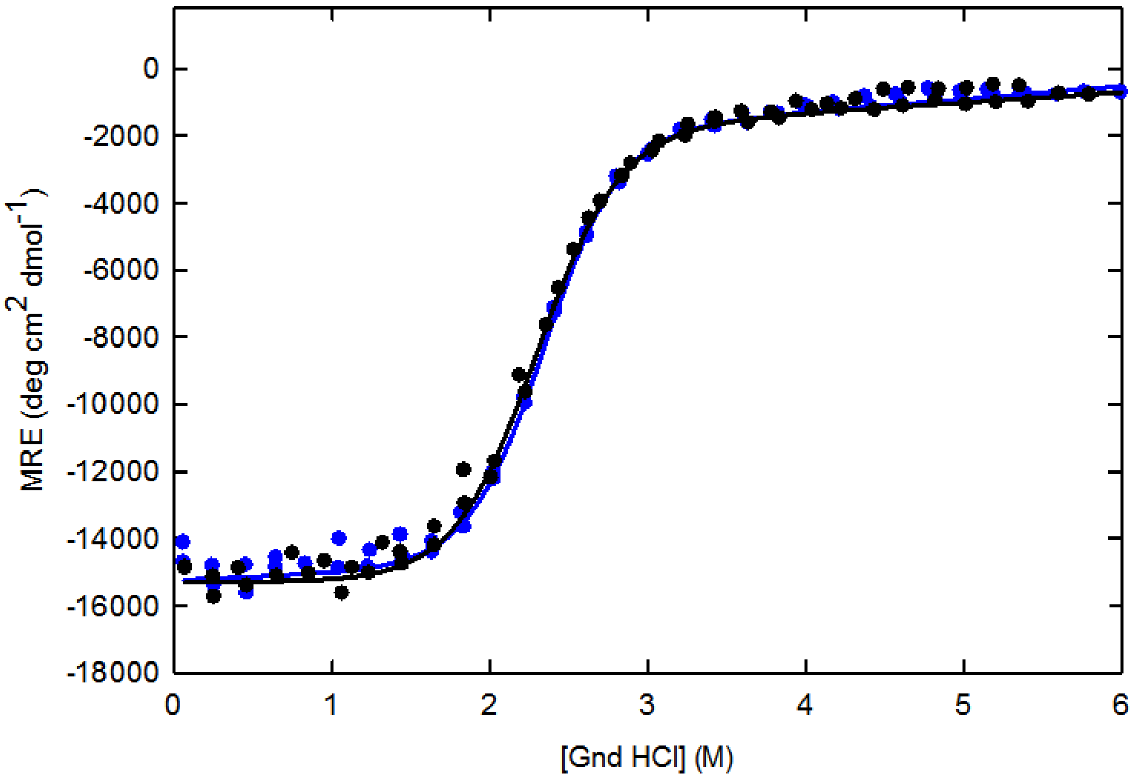
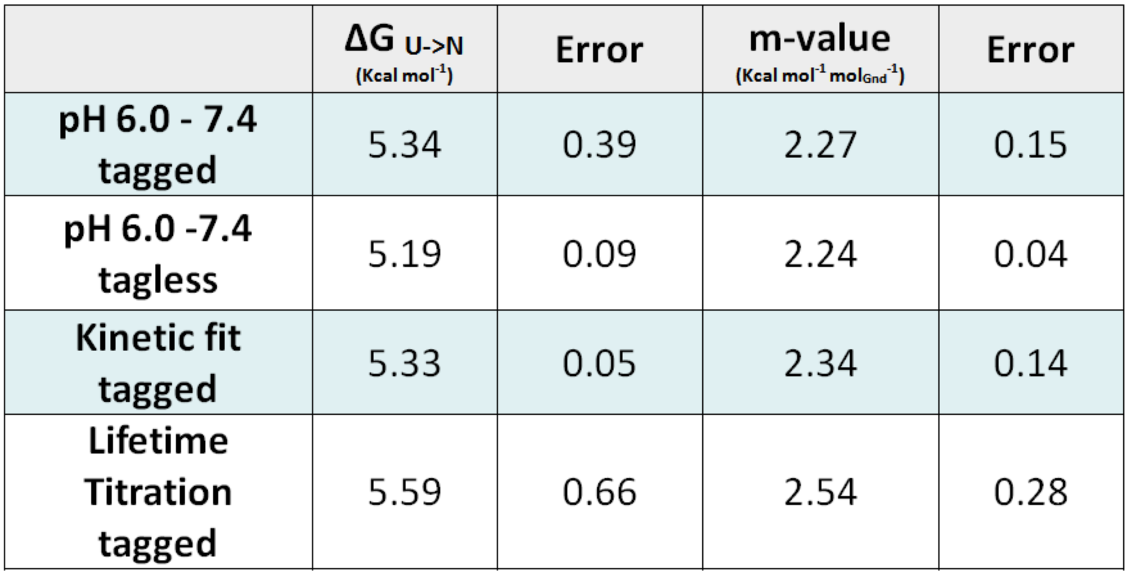


Figure 3.2. Gnd Denaturation melts of tagless Di-III_14 at 222nm by CD

Table 3.1. Thermodynamic parameters



	$\Delta G_{U\text{->}N}$ (Kcal mol^{-1})	Error	m-value (Kcal mol^{-1} mol$_{Gnd}^{-1}$)	Error
pH 6.0 - 7.4 tagged	5.34	0.39	2.27	0.15
pH 6.0 -7.4 tagless	5.19	0.09	2.24	0.04
Kinetic fit tagged	5.33	0.05	2.34	0.14
Lifetime Titration tagged	5.59	0.66	2.54	0.28

Di-III_14 is thermodynamically 2-state

Using the N-terminal tagged construct, Gnd melts were performed for thermodynamic characterization at pH 6.0, 6.5, and 7.4 prior to HDX NMR. There is no apparent pH dependence of the stability of the protein (Fig. 3.3) as the denaturant melts are virtually coincident at pH 6.0 and pH 7.4. The global stability across this pH range, yields an average global stability of 5.52 ± 0.16 kcal mol^{-1} and an m-value of 2.35 ± 0.07 kcal mol^{-1} M$_{Gnd}^{-1}$ (Table 3.1). A CD wavelength scan at 0 M denaturant reveals that the secondary structure of the Native state (N) is likewise unaffected by pH within this range (Fig. 3.4).

Gnd melts were also surveyed by total intensity fluorescence (FL) using the single intrinsic tryptophan of the construct. These results also appear to demonstrate 2-state behavior and again appear to be virtually coincident at all three pHs surveyed (Figure 3.3B). Unfortunately the titrations measured by FL total intensity cannot be reliably fit owing to the steep native baselines. Further analysis of the data by center-of-mass shows a loss of coincidence due to a wavelength shift in the pH 7.4 data (Figure 3.3C) precluding comparable fits across the pH range. The thermodynamic stabilities reported herein reflect the CD measurements and are consistent with the previous CD measurements made on the original construct with the C-terminal his-tag[70].

Unfolding of the protein by Gnd reveals a midpoint at 2.34 M Gnd. Generally the corresponding urea melt should have a midpoint at approximately twice (~4.68M) the molar concentration[104]. Surprisingly, Urea has little apparent effect on the Native state of Di-III_14 from 0-8.5 M (Fig. 3.5), suggesting that salt is contributing significantly to the Gnd denaturation profile. However, repeating the urea titration from 0-8 M at the solubility limit of 2.4 M NaCl is insufficient to perturb the structure further (Fig. 3.5).

Fig 3.3. Gnd titrations of N-tagged Di-III_14. Gnd denaturation melts of the tagless construct are pH independent at pH 6.0 (blue), pH 6.5 (red), and pH 7.4 (black). The data was globally fit to a 2-state model (solid lines; see Table 3.1). (A) Gnd melts observed by CD at 222nm show coincident responses at all 3 pHs. (B) Total intensity fluorescence measurements reveal a steep native baseline that precludes accurate fits. (C) The center of mass analysis of the fluorescence data shows a wavelength shift at high molar Gnd preventing comparable fits across the dataset.

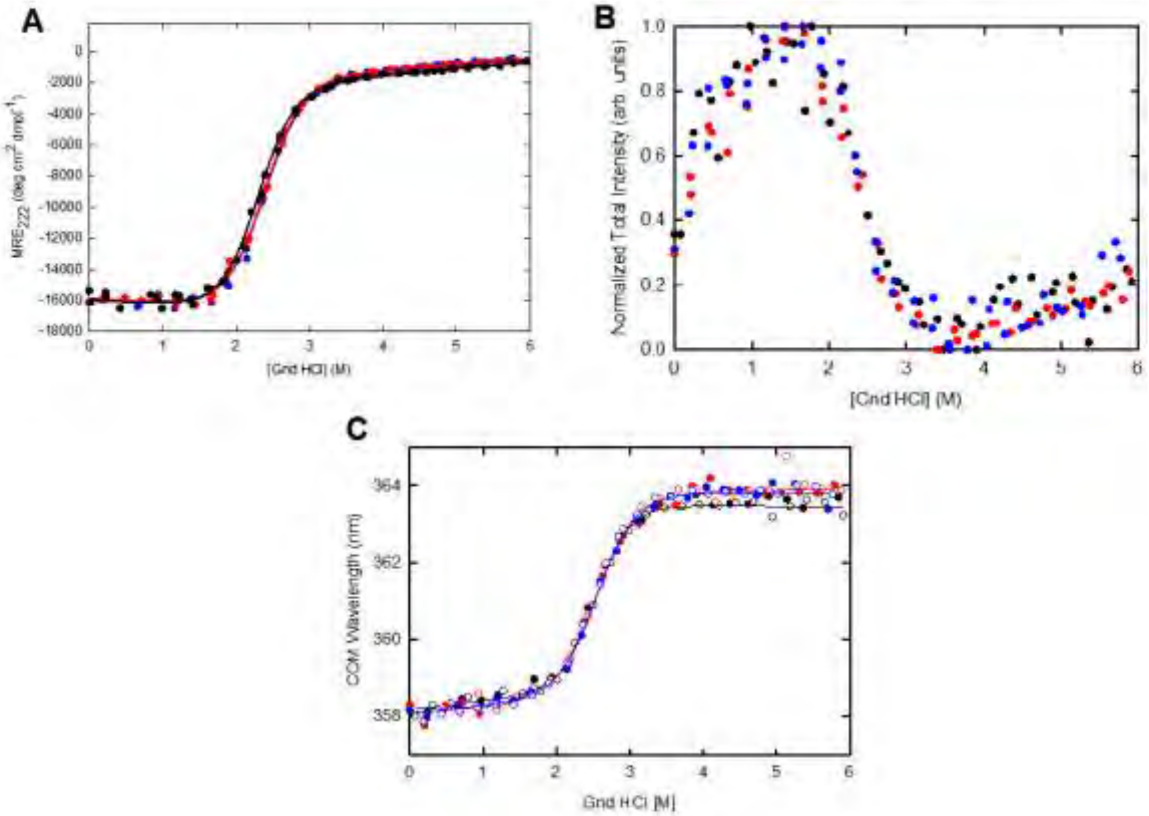


Figure 3.3. Gnd titrations of N-tagged Di-III_14

Fig 3.4. CD wavelength scans of Native Di-III_14. There is no apparent pH dependence on the Native state of Di-III_14. Scans at pH 6.0 (blue), pH 6.5 (red), and pH 7.4 (black) are superimposable.

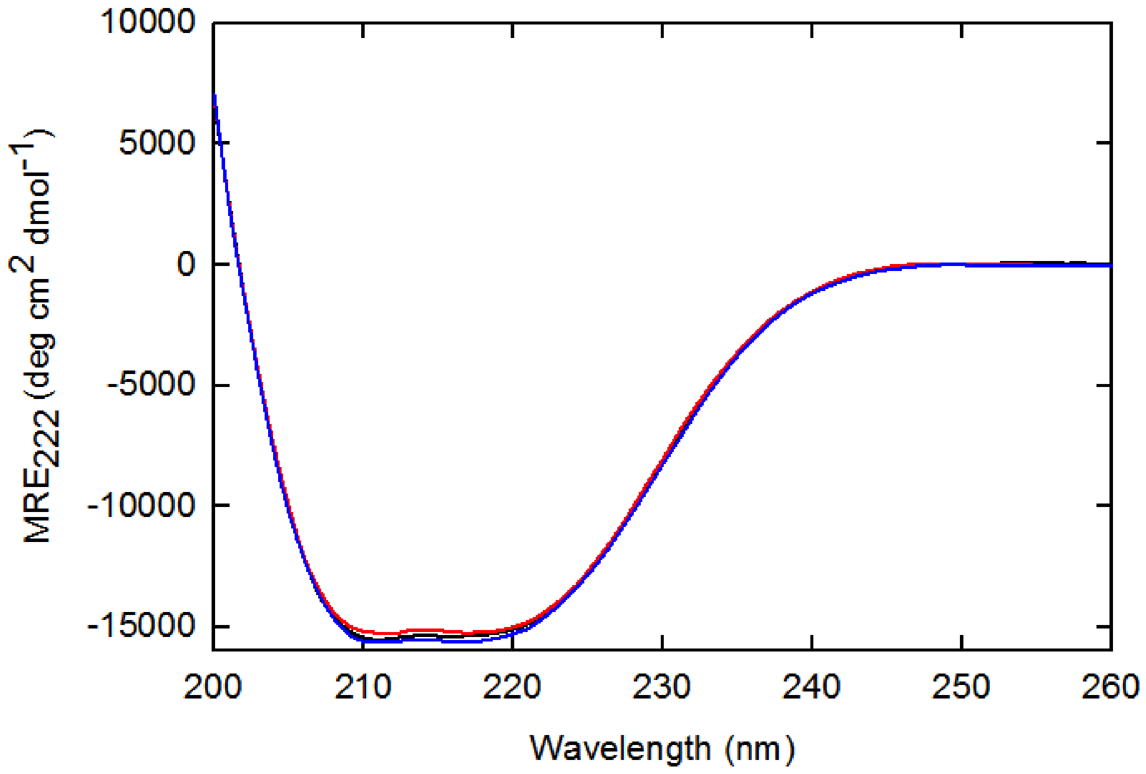


Figure 3.4. CD wavelength scans of Native Di-III_14

Fig 3.5. NaCl and urea sensitivity of Di-III_14 by CD wavelength scan. The native structure of Di-III_14 is superimposable at native conditions (black and red dashed lines) with the 2.4 M NaCl condition (red solid line). 8 M urea (black solid line) disrupts the native structure but remains more structured than the midpoint signal of the Gnd melts (blue dashed line). In 8 M Gnd + 2.4 M NaCl (green solid line) there is no further disruption to the structure as compared to the 8 M condition.

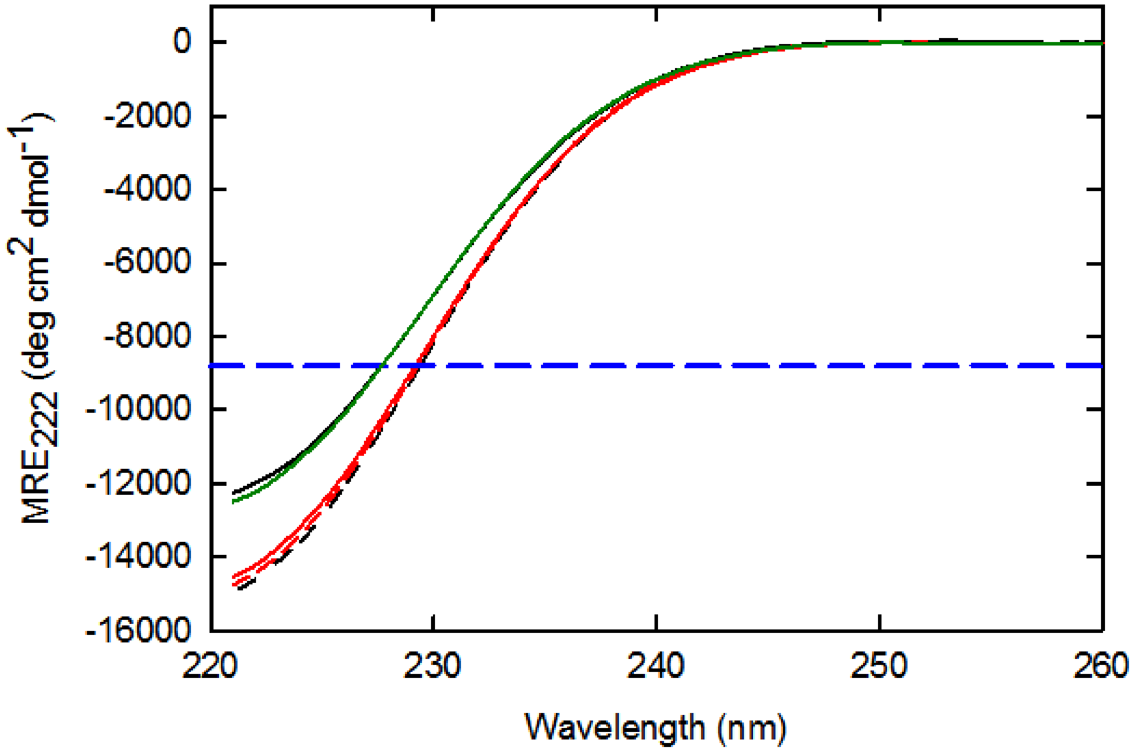


Figure 3.5. NaCl and urea sensitivity of Di-III_14 by CD wavelength scan

Di-III_14 is kinetically 2-state

The kinetic unfolding and refolding experiments with respect to the Gnd denatured state occur faster than can be observed using a conventional stopped-flow CD with a dead-time of 5 ms at 222 nm. Although all of the secondary structure is formed within this dead time, even when jumping to the midpoint, tertiary changes are observable by stopped-flow and continuous-flow mixing experiments observed by FL. Notably, Di-III_14 demonstrates very fast folding and unfolding kinetics with an extrapolated refolding relaxation time of 48.0 μs and unfolding relaxation time of 479 ms at 0 M denaturant (Fig 3.6). These equilibrium dynamics place the protein within the EX2 regime of hydrogen exchange at 25 °C where the average exchange rate of ~9.4 ms is approximately 196 times slower than the refolding rate, as calculated using SPHERE[105] for this sequence at pH 7.4 and 25 °C. The chevron plot of the kinetic relaxation times is consistent with the thermodynamic analysis as it demonstrates no roll-over behavior and fitting the chevron to a 2-state model, with reference to the Gnd unfolded state ($U_{Gnd} \rightleftharpoons N$), yields a comparable Gibbs free-energy of 5.33 ± 0.05 kcal mol^{-1} and an m-value of 2.34 ± 0.14 kcal mol^{-1} M_{Gnd}^{-1} (Table 3.1, Fig 3.6). The coincidence of the kinetic data at pH 6.0 and 7.4 and with the tagless variant suggests a pH and histag independence of the kinetics which is consistent with the observed pH independence of the thermodynamics (Fig 3.6).

Using time-correlated single photon counting (TCSPC) intrinsic FL lifetime measurements were made to determine if there is a detectable burst-phase intermediate. Equilibrium titration by this method describes a 2-state process with a comparable free-energy and m-value to the previous equilibrium and kinetic studies (Fig. 3.7). Calculation of the quantum yield averaged lifetime from the 0.6 M Gnd jump describes a two state process as the lifetimes correspond to the N and U_{Gnd} states observed at equilibrium (Table 3.1).

Fig 3.6. Di-III_14 kinetic analysis. Rate constants acquired from folding kinetic experiments are plotted against final denaturant concentrations of the kinetic jump. The kinetics of tagless (open circles) and N-tagged (closed circles) are within error of each other. There is no apparent pH dependence of the data as the pH 7.4 data (black) is within error of the pH 6.0 data (blue). Fitting of the data to a two state model (solid black line) is consistent with the thermodynamic data (see Table 3.1).

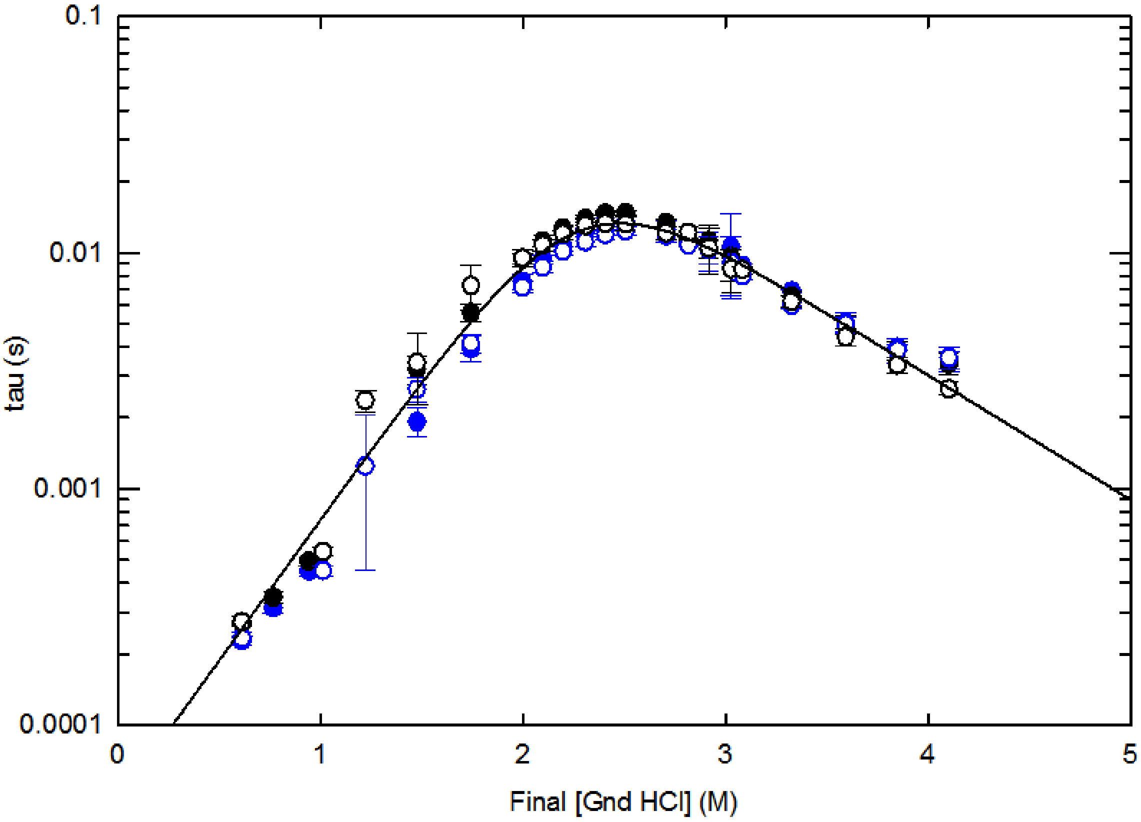


Figure 3.6. Di-III_14 kinetic analysis

Fig 3.7. N-tagged Di-III_14 lifetime analysis. The quantum yield weighted average lifetimes for the equilibrium Gnd titration measured by TCSPC (black dots) are fit to a 2-state model (black line). The fitting parameters are consistent with CD and FL kinetic estimates (see Table 3.1). Fitting the CF-TCSPC 0.6 M refolding jump yields lifetimes consistent with the equilibrium measurements at both 6 M Gnd and 0.6 M (dashed red lines), suggesting that the kinetics are entirely 2-state.

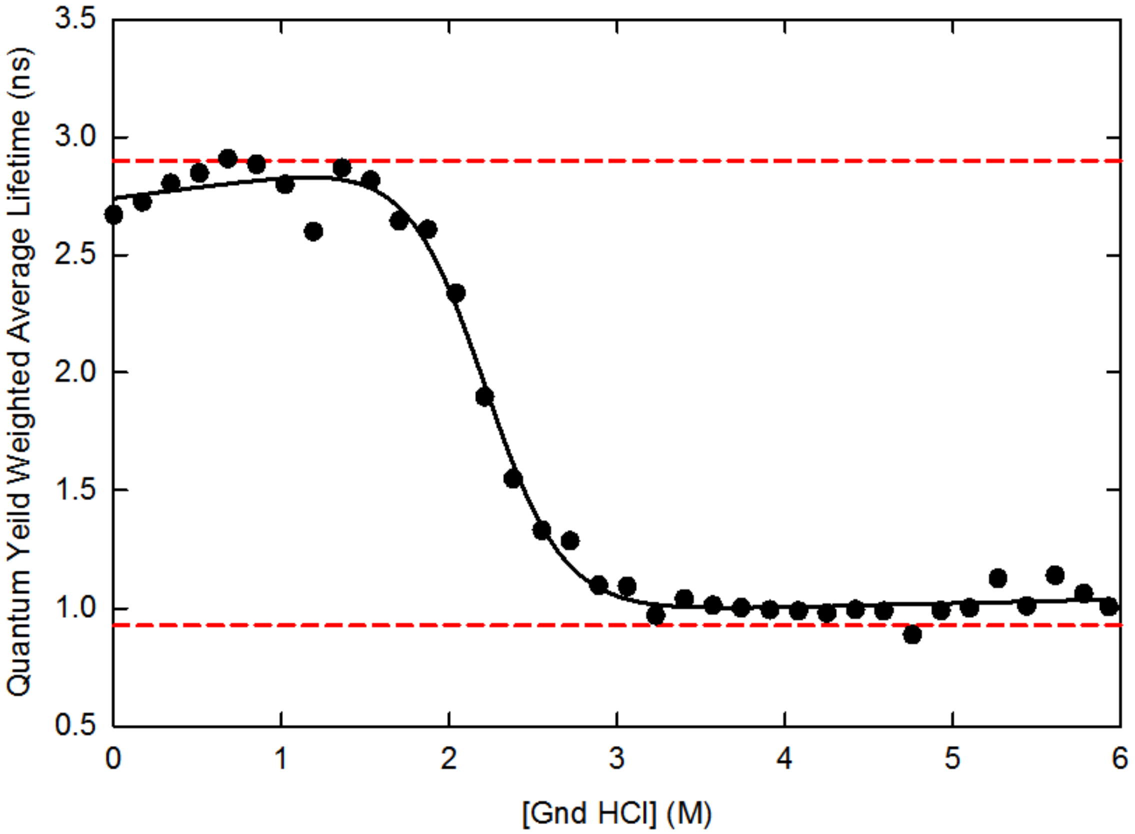


Figure 3.7. N-tagged Di-III_14 lifetime analysis

Native state HDX reveals residual structure in the U state

To better survey the "smoothness" of the energy landscape, native state HDX experiments observed by NMR were conducted at pH 6.0, 6.5, and 7.4. The exchange of 33 out of the total 89 amino acids are observable with a range of exchange rates from 9 minutes to 26 hours. Exchange of two amino acids, V31 and I70 on β-strands 2 and 4, occurs too slow to be reliably fit over a collection time of 3 days. Comparison of the observed exchange rates (k_{obs}) across the different pHs reveals a symmetric correlation with a slope approximating unity along the diagonal (Fig. 3.8), suggesting that exchange is pH independent, and thus exchange occurs through an EX1 mechanism (k_{close}<<k_{int})[106]. Notably, the pH 7.4 data, although approximating EX1 behavior, appears to be demonstrate slightly faster exchange rates than the pH 6.0 data. Approximating a full exchange response as three times the relaxation time, the kinetic response of the protein would suggest that in both EX1 or EX2 (k_{close}>>k_{int}) exchange regimes all residues should be fully exchanged well within the 24 min acquisition time of the experiment, 1.5 s and 3.3 ms. Observed rates therefore suggest an apparent super-protection that is indicative of an intermediate or residually structured unfolded state (U_{RS}). Because 63% of the residues are fully exchanged prior to the first acquisition and there is no significant correlation between solvent accessible surface area (SASA) of the exchangeable amide and exchange rate (Fig. 3.9), this residual structure is likely to reside on the unfolded side of the barrier.

Fig 3.8. Residual structure is slow to exchange. Exchange rates at pH 6.5 (red) and pH 7.4 (black) are plotted against the rates observed at pH 6.0. The slope of 1 on the diagonal suggests that U_{RS} is exchanging via an EX1 mechanism. EX2 reference lines for pH 6.0 (red) and pH 7.4 (black) are shown as dashed lines.

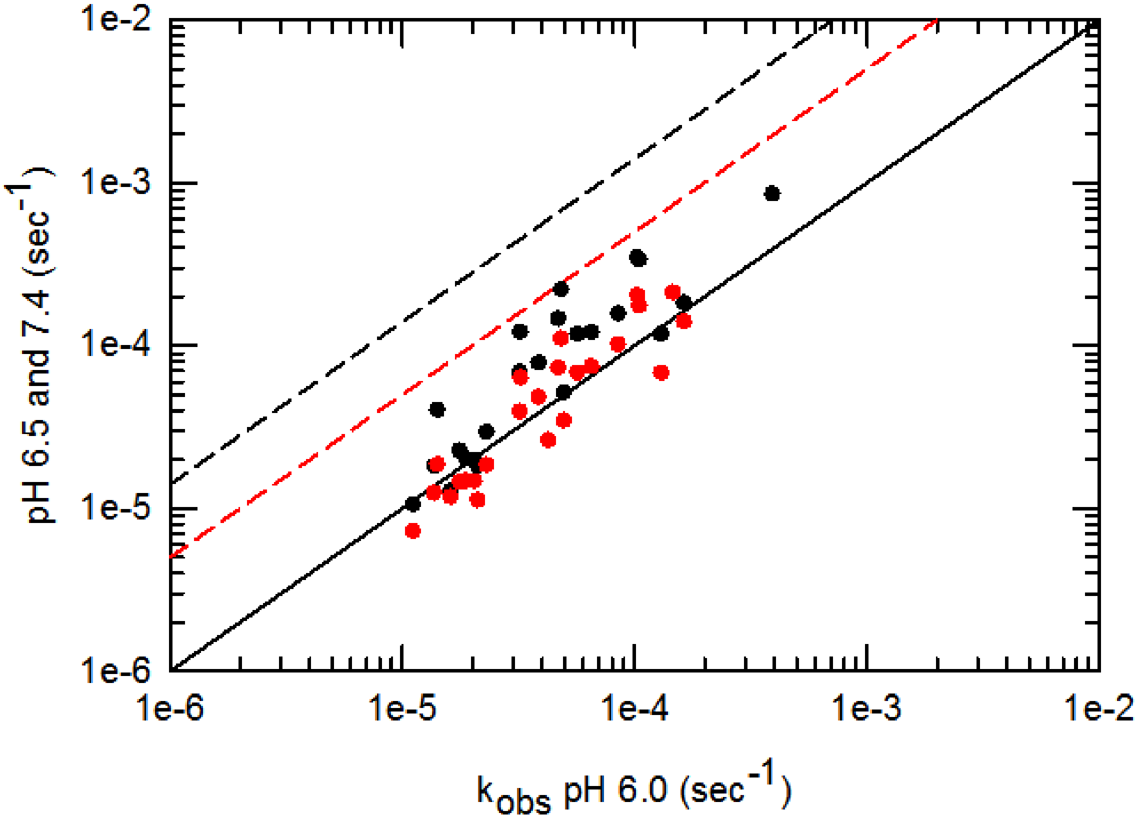


Figure 3.8. Residual structure is slow to exchange

Fig 3.9. Solvent accessible surface area of the amide nitrogen. The SASA of the amide nitrogen, as calculated by Chimera from the NMR structure, are plotted by amino acid position (blue). The residues persistent in U_{RS} are superimposed (red circles). Although many protected residues have no SASA, many that are not protected and also have no appreciable SASA. (B) The exchange rates are not correlated to SASA .

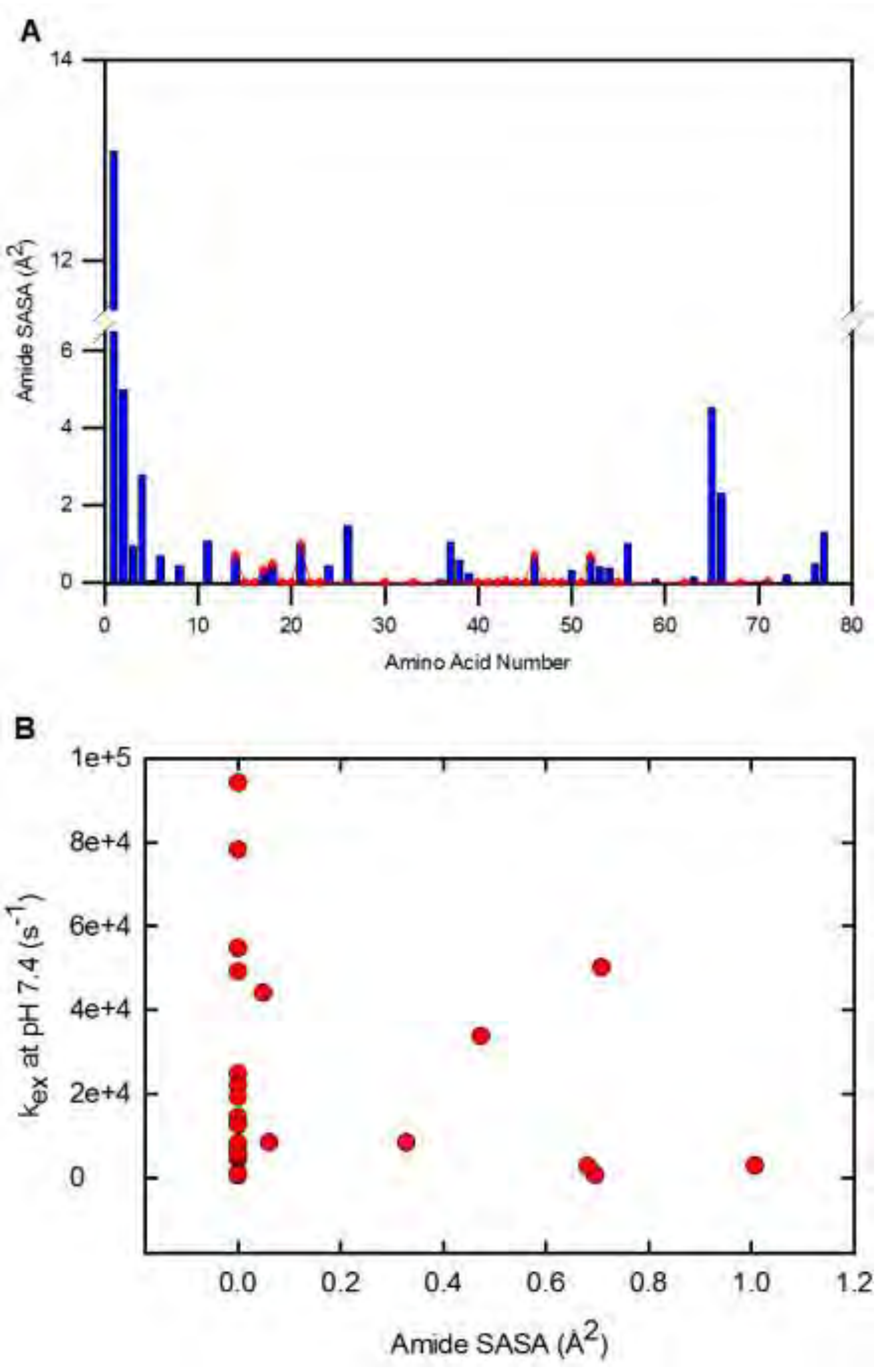


Figure 3.9. Solvent accessible surface area of the amide nitrogen

Mapping the observed exchange rates to the 3-D structure reveals a stable hydrophobic core surrounded by solvent exposed polar groups (Fig 3.10). In fact 14 of the 33 super-protected residues are hydrophobic with 10 being ILV residues calculated to be part of a larger predicted isoleucine, leucine, and valine (ILV) cluster (Kathuria et. al, unpublished data). Structural contributions are from both helices as well as the 2 central β-strands (Fig. 3.10).

Interestingly, of the 8 predicted (ESBRI) salt bridges[107], 5 span 28 or more residues in sequence and connect adjacent secondary structure elements. The most notable of these are the three contacts between K67 to D34, K69 to E32, and R71 to E30, which all connect the C-terminal β4 strand to the intercalated β2 strand in the center of the β-sheet (Fig. 3.1C). These β-strands may be stabilized by their resident salt bridges as the slow exchange rates are enriched within these elements (Fig. 3.10). Notably the slowest rates are at the center of these two strands, directly adjacent to E30 and R71. The slowly exchanging V31 and I70 are not only flanked two salt bridges in sequence, but are also hydrogen bonded to one another.

Residual structure in the Unfolded state is Sensitive to Guanidine

Repeating the HDX NMR experiments in lower molar Gnd demonstrates decreasing exchange times with increasing denaturant concentrations (Fig. 3.11A). These experiments are conducted in low molar Gnd concentrations that are well within the native baseline, 0.5 and 1.0 M. At these concentrations the rate of

Fig 3.10. Structural correlates of exchange rates. All exchange rates are superimposed on the NMR structure of Di-III_14 (top left). Only polar residues (bottom left), only ILV residues (bottom right), and only hydrophobic residues (top right) are shown as spheres. Elements in grey exchange too quickly to be observable. The black elements have observable exchange over a 3 day collection window, but are too slow to reliably fit. The scale from red to blue is every 20% of the observable range of the rates.

Figure 3.10. Structural correlates of exchange rates

Fig 3.11. NaCl and Gnd sensitivity of exchange rates. (A) In response to increasing concentrations of Gnd, a slight decrease in exchange rates is typically observed before acceleration, implicating low molar Gnd in slightly stabilizing interactions. (B) The residues with observable exchange rates at 0.5 M Gnd are mapped onto the NMR structure. All hydrophobic residues are in orange and all polar residues are in blue. (C) Increasing concentrations of NaCl generally increases the exchange rates. NaCl at 0.5 M is more perturbing than 0.5 M Gnd and 1M Gnd (see A). (D) The residues with observable exchange rates at 0.5 M NaCl are mapped onto the NMR structure. All hydrophobic residues are in orange and all polar residues are in blue. NaCl generally disrupts the β-sheet more significantly than Gnd.

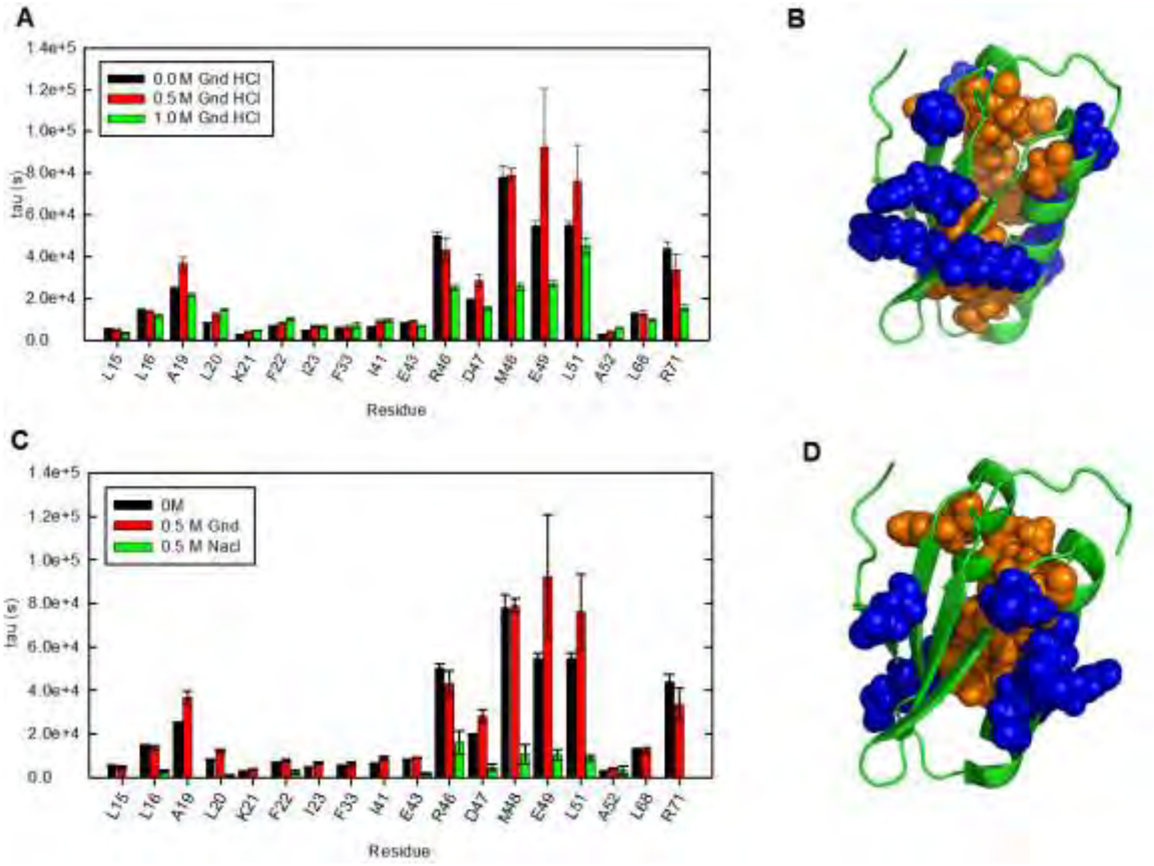


Figure 3.11. NaCl and Gnd sensitivity of exchange rates

refolding at equilibrium differs 10-fold between 0 M and 1 M denaturant and 5-fold for the rate of unfolding, based on the equilibrium rates. However, the observed difference in rates varies only 1.5 to 4-fold across the set of residues, excluding residues 31 and 70. These results suggest that the residual structure observed in the HDX NMR experiment is perturbed by low concentrations of Gnd, in a way that is inconsistent with the kinetic rates. The different rates at 0 M Gnd suggests multiple transitions from U_{RS} to the exchange competent unfolded state (U_{EX}). Therefore the exchange reaction is limited by barriers that are not observed in the $U_{Gnd} \rightleftharpoons N$ thermodynamic and kinetic studies.

The residual structure is mapped onto the NMR structure (Fig. 3.11B). These elements suggest a large hydrophobic contribution between the helices and the β-sheet within the residual structure. Surrounding polar groups may also be contributing to the stability of the hydrophobic core.

Residual structure in the Unfolded state is Sensitive to Salt

The Gnd dependence of exchange rates considered with inability for urea to sufficiently disrupt native structure of Di-III_14 (Fig 3.11A) suggests that the ionic strength of the solution may be destabilizing. Under equilibrium conditions, there appears to be no effect on the native state at concentrations of 2.4 M NaCl (Fig. 3.5). The solubility limit of 2.4 M NaCl with 8M urea preludes a determination of the effect of NaCl on the stability of the native state, as no significant salt dependence can be extracted between 0 M and 8 M with NaCl concentrations as high as 2.14 M.

Repeating the HDX experiments in the presence of 0.5 and 1.0 M NaCl show an acceleration of exchange rates at 0.5 M NaCl that are faster than the accelerated exchange rates observed at 0.5 and 1M Gnd (Fig 3.11A,C). At both concentrations of NaCl, aggregation severely limits the signal intensity such that no meaningful data could be resolved at 1 M concentrations. This observation suggests that the unfolding of U_{RS} leads to aggregation prone species, which is accelerated upon the observed destabilization of U_{RS}. This acceleration of exchange rates also suggests that, although both Gnd and NaCl are both destabilizing, the Gnd has semi-compensatory stabilizing affect at low molar concentrations that modulates the effect of the ionic contributions. This effect is perhaps most pronounced in Figure 3.11A where residues grouped from R46 to L51 and A19 appear to show an increase in exchange time from 0 M to 0.5 M Gnd. Previous work on Cytochrome C has observed a similar stabilizing affect at low molar urea and Gnd concentrations[108]. Work from Farber and colleagues[109] and Kumar et. al[110] also demonstrate that Gnd can bind to the protein backbone, increasing rigidity, supporting this observation.

Mapping the protected residues that persist at 0.5 M NaCl onto the NMR structure shows a significant loss in residue contributions from the β-sheet of both hydrophobic and polar character (Fig. 3.11D). Comparison of the residues that remain in 1M Gnd (Fig. 3.11B) to those that remain in 0.5 M Nacl (Fig. 3.11D) suggest that the salt bridges (Fig. 3.1C) are disrupted by NaCl but not by

Gnd. This potential disruption would increase the dynamics of the β-sheet and therefore increase the exchange rates of the local residues.

Discussion

Di-III_14 has an electrostatically dependent, native-like structured unfolded state

The Gnd and NaCl dependent apparent super-protection observed by HDX experiments suggests that electrostatic charges play a major role in establishing the residual structure in the unfolded state. The failure of urea to unfold the protein, with and without 2.4 M NaCl (Fig. 3.5), demonstrates that electrostatics are not the primary mode of stabilization of the N state. Conversely, the stability of the U_{RS} is largely dependent on electrostatic interactions as the exchange rates can be significantly modulated by the presence of either Gnd or NaCl (Fig. 3.11). Although the major modes of stability are different between these two states, the chemical shifts of the 33 slow exchanging amino acids are comparable to the Native state as measured by heteronuclear single quantum coherence spectroscopy (HSQC) (Fig. 3.12). The residual structure of the unfolded state, therefore, maintains native like structure that biases the folding towards native. This potential conservation of topology in a addition to the low contact order may explain the microsecond refolding rates at 0 M denaturant[64].

Fig 3.12. Native structure in U_{RS}. The chemical shift differences (left panel) are all less than 1ppm in response to pH (top), Gnd (middle), and NaCl (bottom). The Corresponding ^{15}N-HSQC overlays (right panel) with pH 7.4 reference data show significant agreement of all cross-peaks. (credit: Laura Deveau)

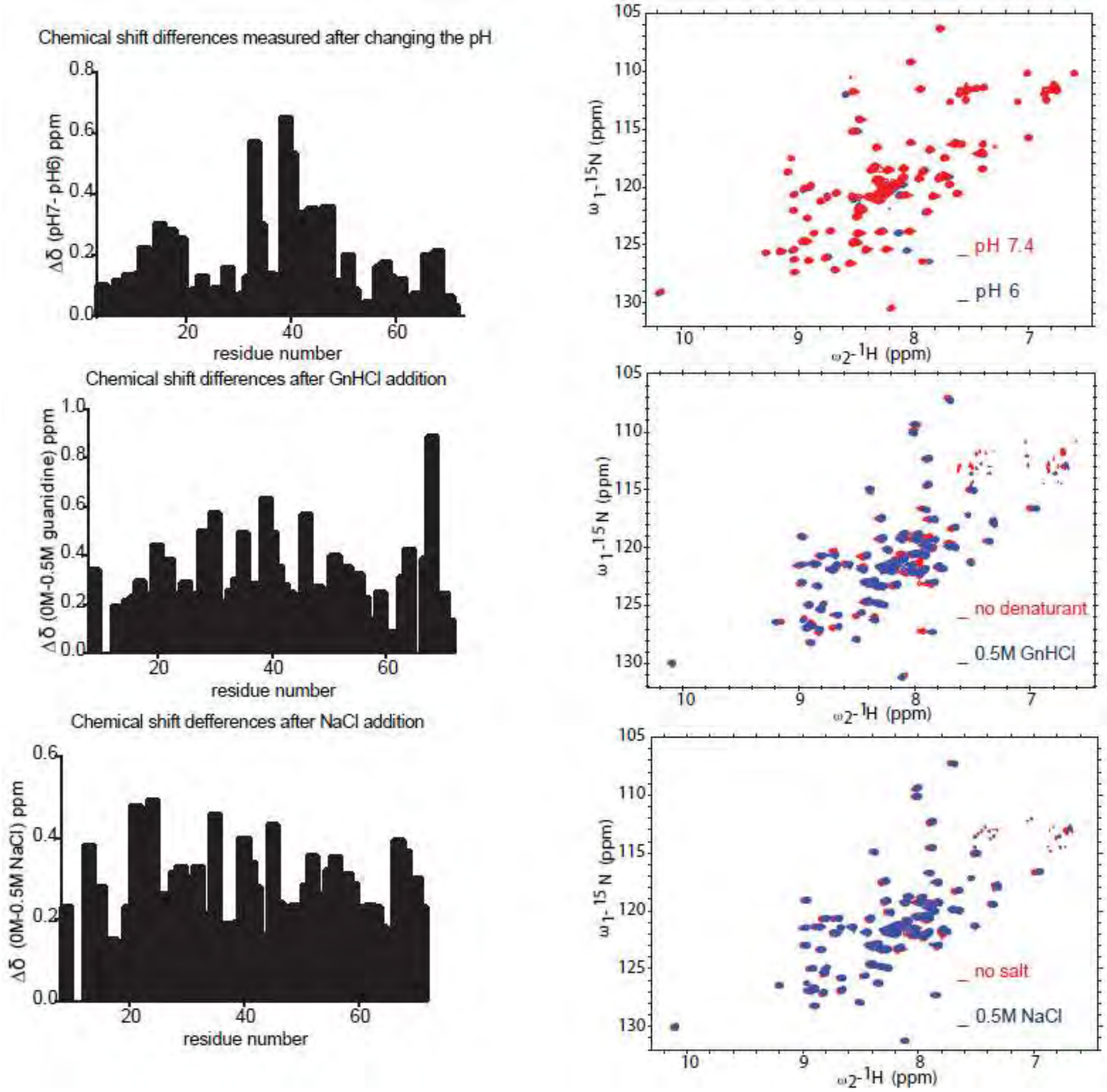


Figure 3.12. Native structure in U_{RS}

The ionic sensitivity of the structure of Di-III_14 is potentially explained by the unnatural sequence composition.   Unlike naturally occurring globule proteins, Di-III_14 is enriched for polar charged amino acids by 10-15%, typical of intrinsically disordered protein (IDP) sequences[111]. Like IDPs, Di-III_14 can be characterized as a polyampholyte with a kappa value of 0.25, indicating a significant segregation of opposite charges and therefore an intrinsic ability to collapse under native conditions[112,113]. The charge segregated, and sequence distal salt bridges between β2 and β4 are illustrative of this point.

Although the his-tag contributes to the charge segregation, the presence of a his-tag has previously been reported to be correlated with a reduction of the hydrodynamic radius (R_h) of IDPs as a function of total sequence representation[114], further supporting this point. The N-terminal cleavable his-tagged construct consists of 89 residues, making a hexahistag 6.7% of the total sequence. According to Marsh and Forman-Kay's correlation[114], the expected compaction from the his-tag would be around 20%. In the case of Di-III_14, 33% of the residues are implicated in the residual structure and is thus in probable agreement with this estimate.

The description of the U_{RS} from the sum of these studies suggest that the residual structure in the unfolded state is native like, compact, and stabilized by long range electrostatic interactions. Further, the implication of the his-tag in compaction of the structured unfolded state qualitatively agrees with the noted increased aggregation propensity of the tagless construct. Structural correlates gleaned from the NMR data further describe this structure as consisting of a hydrophobic core between the two central β-strands and the two alpha helices. Solvent exposed polar residues flank this core and long range salt bridges span the central beta-strands which also share main chain hydrogen bonds, further stabilizing the structure.

Modeling the free-energy landscape of Di-III_14

Di-III_14 unfolds to a structured intermediate under Native favoring conditions

The data presented here suggests evidence for a compact unfolded state (U_{RS}) that retains native like character (Fig. 3.12), presumably due to a combination of long range electrostatic interactions (Fig. 3.1C) and low CO hydrophobic contacts (Fig. 3.1B, 3.10, 3.11B). Based on the equilibrium dynamics extracted from the kinetic experiments the exchange mechanism is expected to be EX2 and therefore occur entirely within the first acquisition by NMR. Approximately 67% of the protein follows this expectation in a manner that is not dependent on the SASA of the amide nitrogen (Fig. 3.9), suggesting that the remaining 33% of slow exchanging residues is on the unfolded side of the barrier (Fig. 3.13).

Fig 3.13. Free-energy landscape of Di-III_14. Folding from the Gnd unfolded state to N is kinetically a 2-state process (dashed line). Under native-favoring conditions, the native state unfolds to a structured unfolded state (U_{RS}). U_{RS} unfolds to the exchange competent unfolded state (U_{EX}) through at different rates, suggesting a structurally diverse unfolded state that is more structured than the guanidine unfolded state (U_{Gnd}).

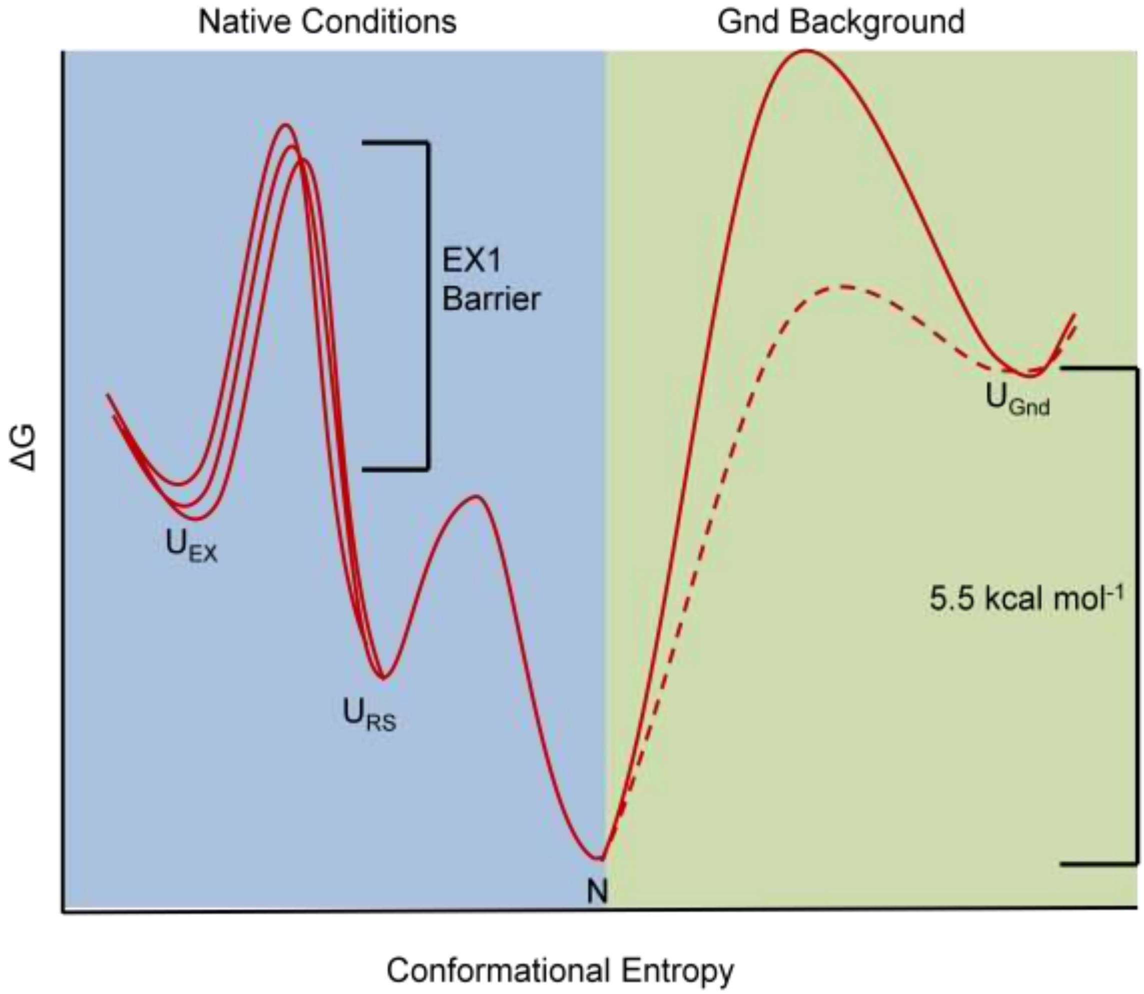


Figure 3.13. Free-energy landscape of Di-III_14

U_{RS} is not populated in refolding to N from U_{Gnd}

A compact unfolded intermediate, such as the one observed under native-favoring conditions, is not observed in equilibrium or kinetic experiments by CD or FL. The deviance from a 2-state process under these conditions suggests significant differences in the free-energy landscape on the unfolded side of the barrier which is dependent on the presence of ionic strength of Gnd and NaCl at concentrations that have no observable effect on the Native structure (Fig. 3.5). None of these barriers are experimental observed under equilibrium conditions observed by CD or FL, Kinetic experiments observed by total intensity FL, or in the FL lifetime measurements which describe complete refolding from $U_{Gnd} \rightarrow N$ (Fig. 3.6). In agreement with this data is the m-value which is as expected for an 89 residue protein[86].

The U_{Gnd}, U_{EX} and U_{RS} states are distinct unfolded state populations

Interestingly, the EX1 exchange rates are not homogeneous (Fig. 3.11). In EX1 exchange the unfolding rate is much slower than the rate of exchange such that[106] $k_{ex} = k_{open}$. Non-homogenous EX1 rates imply that the barriers between U_{EX} and U_{RS} are also variable. This observation implies that multiple structurally unique U_{EX} states must exist that are not in rapid equilibrium with one another. Structure in the U_{EX} implies that this state must also be distinct from the U_{Gnd} which is likely to be random coil[87]. If this interpretation is correct, it is the first experimental evidence, to the best of my knowledge, that may recapitulate the model of the unfolded state described by hub models of folding produced by MSM analysis of molecular dynamic simulations[37].

Designing against U_{RS}

Koga et. al.[70] used MD simulations to develop design rules to bias folding to the N state from a random coil. Comparison of their predictions with the data taken in the Gnd background yields perfect agreement, while data taken under native-favoring conditions reveals unexpected complexities on the unfolded side of the barrier. These complexities are likely to be largely due to the charge segregation of the sequence[112], and specifically to the salt bridges within the β-sheet. Salt bridges are also seen in the Di_I-5 construct, the only other construct to have helices on only one side of the β-sheet. These salt bridges are, likewise, between internal strands (1 and 3) in the center of the β-sheet. Unlike DI-III_14, the charges alternate on each strand, which may provide a test for the hypothesis that the charge segregation is contributing significantly to the U_{RS} structure. Alternatively mutations within these two β-stands within Di-III_14 may be able to decrease the complexities on the unfolded side of the barrier.

Perspective

The successful rational *de novo* design of proteins has been a long-sought after goal since the 1980's[115]. The success of this idea promises to define the fundamentals of the physicochemical basis of protein folding while also providing customizable engineered protein therapeutic platforms, and rational stabilization methods. Previous work has been promising but has ultimately fallen short of simplifying the kinetic mechanism[33] of folding such that misfolding and aggregation reactions are avoided. Here we provide evidence to support the successful rational engineering of the protein Di-III_14 which is not only stable but also maintains a smooth energy landscape over biologically relevant time scales, a result of incorporating folding kinetics into the design. However, this accomplishment comes at the cost of obfuscating the naturally occurring physicochemical basis of protein folding by enriching the sequence to an unnaturally high composition of polar charged amino acids (40.5%) for globular proteins (typically 25-30%)[97]. The result is a protein that retains native-like structure in a structured unfolded state, completely biasing it's folding trajectory and minimizing the degrees of freedom through which it is permitted to traverse the folding free-energy landscape. Complications arise on the unfolded side of the barrier in perceived roughness of in the high energy region of the landscape, although access is kinetically limited. Therefore, although these results may not be entirely generalizable to natural sequences, the success of this design may prove to be a valuable model for the collapsed unfolded states as well as a platform for the development of conformationally and kinetically stable biologics.

Methods

Protein expression and purification

Di-III_14 was engineered with an N-terminal hexahistidine tag and an intervening Tev Protease site (GenScript), ligated into the expression plasmid pGS-21a, and transfected into the *E. coli* strain BL21 Codonplus® (DE3)RIL for expression. All proteins were isolated from inclusion bodies by dissolving the insoluble fraction of the cell lysate in 6 M Gnd and refolding into 10 mM potassium phosphate buffer at pH 7.4. Precipitates were removed using centrifugation and the soluble fraction was bound to a nickel resin overnight at room temperature. The nickel resin was then thoroughly washed and the His-tagged protein was eluted with a step gradient of 10 mM, 25 mM, and 300 mM imidazole. Pure fractions were pooled and dialyzed into the working buffer. For the tagless variant, the protein was dialyzed into 50 mM tris, 1 M urea, pH 7.8. Overnight incubation with His-tagged Tev Protease was used to cleave off the His-tag from Di-III_14. The reaction product was bound to a nickel resin overnight and the cleaved protein was washed off of the resin in 10 mM potassium phosphate buffer at pH 7.4. Protein was concentrated and applied to a Q Sepharose column and eluted using a salt gradient from 0 to 400 mM NaCl. The purity was confirmed (> 98%) using a Waters Q-TOF ESI mass spectrometer.

For NMR experiments, protein was dialyzed in to 10 mM ammonium bicarbonate. The concentrations of the fully dialyzed samples were determined using Beer's law with the UV absorbance at 280 nm and the calculated extinction coefficients of 5500 M^{-1} cm^{-1} for the tagless construct and 6990 M^{-1} cm^{-1} for the N-tagged construct. Appropriate volumes for reconstitution of 600 µM of a 1 mL volume were aliquotted and lyophilized, typically from stock concentrations of 100 µM.

Protein sequences

Di-III_14 [N-terminal tag] (89 AA):

H H H H H H S S D I E N L Y F Q G L T R T I T S Q N K E E L L E I A L K F I S Q G L D L E V E F D S T D D K E I E E F E R D M E D L A K K T G V Q I Q K Q W Q G N K L R I R L K G

Di-III_14 [tagless] (73 AA):

G L T R T I T S Q N K E E L L E I A L K F I S Q G L D L E V E F D S T D D K E I E E F E R D M E D L A K K T G V Q I Q K Q W Q G N K L R I R L K G

Native state analysis

Circular dichroism structure analysis

Far-UV CD spectra were collected on a JASCO model J810 CD spectrophotometer . All samples were buffered with 100 mM NaCl, 5.6 mM Na_2HPO_4, 1.1 mM KH_2PO_4 at pH 7.4, 6.5, and 6.0 at 25° C. Measurements were taken in a 0.5 cm path length cuvette with a bandwidth of 2.5 nm and a step size of 0.5 nm at a protein concentration of approximately 5 µM from 202 nm to 260 nm. Three buffer subtracted spectra were collected and averaged for each protein with a total averaging time of 3 s per wavelength.

Equilibrium unfolding and refolding

Unfolded (7 M Gnd) and folded (buffer) stocks of protein were diluted with varying concentrations of 100 mM NaCl, 5.6 mM Na_2HPO_4, 1.1 mM KH_2PO_4 with and without Gnd at pH 7.4, 6.5, or 6.0 to achieve a final protein concentration of approximately 5 µM protein and a range of final Gnd concentrations from 0 M to 6.0 M. Each sample was thoroughly mixed and left to equilibrate overnight at room temperature. CD spectra were collected on a JASCO model J810 spectrophotometer under the same conditions as the wavelength scan, and by intrinsic tryptophan fluorescence using Horiba Flourolog3 with a slit width of 5 nm, excitation at 295 nm with a 0.5 cm path length, and emission spectra collected above 310 nm with a 1 cm path length at 25º C. The fluorescence and CD data were globally fit to a two state N⇌U model as a function of Gnd using our in-house Savuka software package.

Kinetic studies

Stopped-Flow intrinsic tryptophan fluorescence

Measurements were collected on a Applied Photophysics model SX.18MV stopped flow. Di-III_14 was injected from stock concentrations of 40 µM and mixed with a 1:10 dilution to 4 µM with 100 mM NaCl, 5.6 mM Na_2HPO_4, 1.1 mM KH_2PO_4 to a the final experimental Gnd concentration. Data was collected for 0.25 seconds with a logarithmic point density totaling 4000 individual points. Each kinetic jump was repeated 20 times.

Continuous-flow intrinsic tryptophan time-correlated single photon counting

The all-quartz mixer was custom made by Translume, Inc. with channels 50 μm wide and 100 μm high. Nanoport connectors for use with 1/32 inch outer diameter PEEK tubing (Upchurch) were attached using UV curing epoxy by the manufacturer (Translume, Inc.). Trp fluorescence utilized 292 nm excitation from the tripled-output of a Ti:sapphire laser. A blank and NATA control were acquired for each trace to correct for background fluorescence (typically ~1%) and variations in excitation intensity along the channel (typically < 5%), respectively. Computer controlled scanning of the channel was accomplished by an x-y translation stage (Biopoint 2, Ludl Scientific).

Time-correlated single photon counting data analysis

Lifetime data was fit using a 2 exponential fit with the software *DecayFit* (Fluorescence Decay Analysis Software 1.4, FluorTools, www.fluortools.com). The quantum yield weighted averages were calculated with equation 3.1,

$$Q_\tau = \frac{\sum_{n=1}^{x}(\alpha_n \, \tau_n^2)}{\sum_{n=1}^{x}(\alpha_n \, \tau_n)} \qquad \text{(Eqn. 3.1)}$$

where Q_τ is the quantum yield weighted average lifetime, *x* is the number of lifetimes the data is fit to, *α* is the amplitude of the lifetime, and *τ* is the lifetime decay.

Hydrogen exchange NMR

Isotopic Labeling

Labeling with ^{15}N was performed by growing cells in isotopically enriched M9 medium, 1g $^{15}NH_4Cl$ per liter. The sample was lyophilized after purification s described above.

2D ^{1}H-^{15}N Heteronuclear Single Quantum Coherence

2D ^{1}H-^{15}N heteronuclear single quantum coherence (HSQC) spectra were collected using samples of U-^{15}N Di-III_14 in a 90% H_2O/10%D_2O buffer solution of 100 mM NaCl, 5.6 mM Na_2HPO_4, 1.1 mM KH_2PO_4 at pH 7.4, pH 6.5, pH 6, 0.5M NaCl, and 0.5M and 1M guanidine. Assignments were transferred using previously assigned spectrum taken from BMRB[70].

NMR for exchange

Exchange was initiated by dissolving the lyophilized protein in ^2H_2O buffer, which had been prepared at the required pH and buffer conditions. All pH values reported are corrected meter readings. Upon addition of ^2H_2O buffer, the sample was immediately transferred to a 5 mm NMR tube (Wilmad LabGlass, Vineland, NJ) and placed in the spectrometer at 25 °C; the protein concentration was ~0.6 mM. The time between the initiation of exchange, the transfer to the NMR tube, placement in the spectrometer, tuning and shimming and the beginning of data collection averaged 6 min. 2D ^{15}N–^{1}H HSQC spectra were recorded over a period of hours to days, and the sample remained in the spectrometer for the entire course of the exchange reactions. All NMR experiments were recorded on a Varian 600-MHz spectrometer, and the spectra were processed in NMRPipe[116] and analyzed with Sparky.

Chapter IV - Advances in Continuous Flow Mixing and μSAXS to Approach Shorter Experimental Time Scales and Compatibility with Computational Studies

This chapter is a combination of a single photon counting methods manuscript that is currently being prepared for publication as well as content from publications listed below. For all publications, I contributed to the content through technical developments and/or data acquisition and analysis.

1. *Lambright D, Malaby AW, Kathuria SV,* ***Nobrega RP****, Bilsel O, Matthews CR, "Complementary techniques enhance the quality and scope of information obtained from SAXS" Transactions American Crystallographic Association. 2013 July; Vol. 44, epub*

2. *Kathuria SV, Kayatekin C, Barrea R, Kondrashkina E, Graceffa R, Guo L,* ***Nobrega RP****, Chakravarthy S, Matthews CR, Irving TC, Bilsel O "Microsecond Barrier-Limited Chain Collapse Observed by Time-Resolved FRET and SAXS" Journal of molecular biology 2014 May;426 (9), 1980-1994*

3. *Graceffa R,* ***Nobrega RP****, Barrea RA, Kathuria SV, Chakravarthy S, Bilsel O, Irving TC. "Sub-millisecond time-resolved SAXS using a continuous-*

flow mixer and X-ray microbeam."Journal of Synchrotron Radiation. 2013 Nov;20(Pt 6):820-5

4. *Kathuria SV, Chan A, Graceffa R,* ***Nobrega RP****, Robert Matthews C, Irving TC, Perot B, Bilsel O. "Advances in turbulent mixing techniques to study microsecond protein folding reactions" Biopolymers. 2013 Nov;99(11):888-96.*

5. *Kathuria SV, Guo L, Graceffa R, Barrea R,* ***Nobrega RP****, Matthews CR, Irving TC, Bilsel O. "Minireview: structural insights into early folding events using continuous-flow time-resolved small-angle X-ray scattering" Biopolymers. 2011 Aug;95(8):550-8*

Introduction

Experimentally, early events in protein folding reactions are difficult to observe due to heterogeneous conformational ensembles, marginal stabilities inherent to high energy states, and the transient lifetimes of structures. As such, many techniques are limited by time or measurement resolutions. Kinetic protein folding experiments by dilution, although widely used, are limited by time resolution, and are unable to provide high resolution structural insights unless coupled to other techniques, such as hydrogen-deuterium exchange (HDX) or fast photochemical oxidation of proteins (FPOP). Recent advances in mixing technology have brought the time resolution of refolding experiments into the microsecond time regime while maintaining acceptable rates of sample consumption[42] (Fig. 4.1). Interfacing efficient mixing strategies with improving detection methods can provide residue specific, pair-wise, and direct global measurements. Combining measurements at varying resolutions over comparable timescales offers the advantage of detailed global analyses[117], as well as the direct and comprehensive comparison to high resolution simulations.

Computational approaches to the protein folding problem have been successfully implemented since 1975[118]. Full understanding and validation of computational models requires that they be benchmarked with experimental data. Successful simulations are therefore capable of significantly enhancing the resolution of folding models only when significant agreement between the

Fig 4.1. Time line of mixer development. Reported mixing times of turbulent mixers (green filled circles), laminar mixers (red filled circles), and chaotic mixers (blue filled circles) are plotted as a function of the publication year. All three mixing techniques can access a mixing time of few tens of microseconds. The sample consumption rate of these mixers is also reported (open triangles) with the same color scheme. The exponential decay in the sample consumption of turbulent mixers is represented by the green line (fit of the reported flow rates). While sample consumption of laminar and chaotic mixers is an order of magnitude smaller than their modern turbulent counterparts, the sample concentrations required is correspondingly an order of magnitude higher in the laminar mixers. In the case of SAXS and CD there has been a significant improvement in the interfacing technology and a concomitant decrease in the reported mixing times. Improvements over the last 50 years have primarily been associated with a reduction in sample consumption more than a reduction in mixing times. (Credit: Sagar Kathuria)

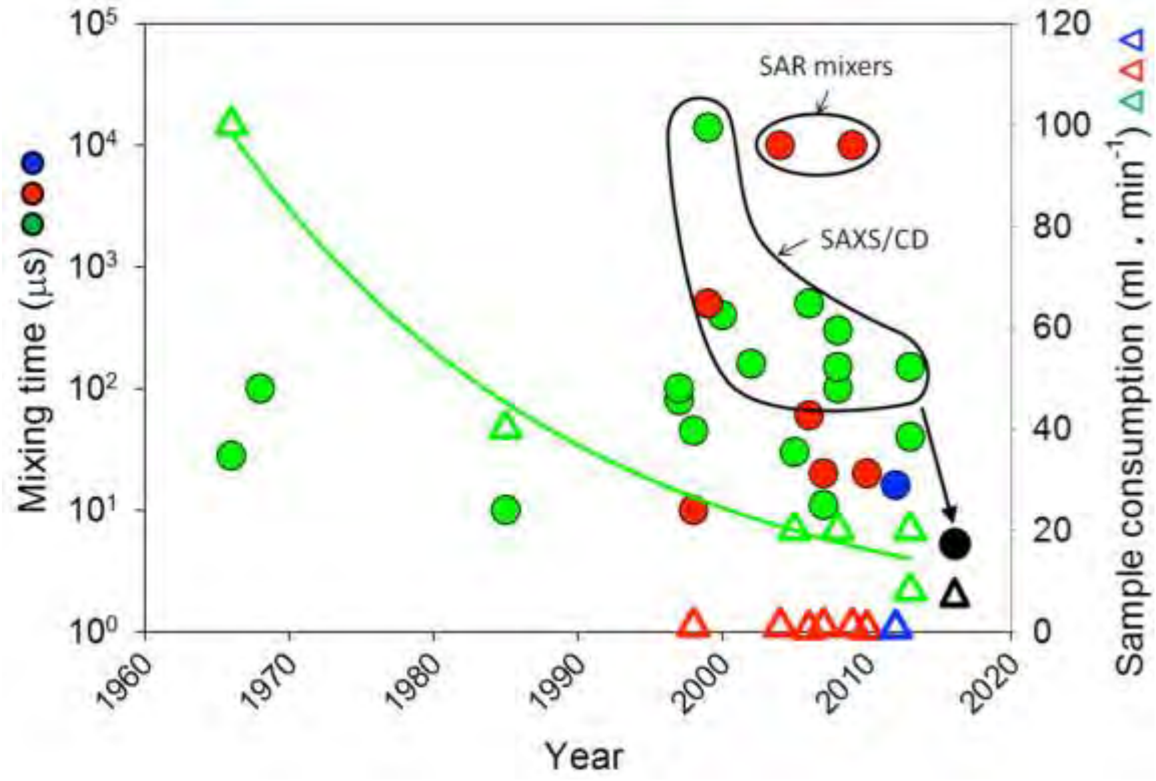


Figure 4.1. Time line of mixer development

experimental and simulation datasets exists. However, direct comparisons between simulated folding trajectories and experimental data are complicated by several factors. Primarily, until recently119 atomistic simulations could not achieve millisecond timescale resolution. Although a minority of small proteins can fold completely in the microsecond timescale27, most biological proteins fold on the millisecond to minute timescales. Computational expense requirements for atomistic simulations not only limit the number of proteins amenable to this approach, but also the number of full (U→N) folding trajectories that can be obtained that may not properly represent the broad experimental ensemble.

To overcome limiting computational requirements, especially in the field's infancy, coarse-grained simulations were employed[8]. Modern implementations of coarse-grained simulations offer the acquisition of numerous full trajectories. Although the data density and the number of applicable proteins increases with this approach, assumptions are made in the parameterization that incorporate reasonable doubt for positional accuracy at a residue level within the generated models, reducing the effective resolution of the computational dataset.

Recent advances with graphics processing units (GPU) and distributed computing have increased the timescales of atomistic simulations to tens of ms[119] (see also chapter 2) such that single trajectory simulations can now be performed for milliseconds of folding time. However, these simulations are still limited statistically by sample size (1 ms trajectories on Anton: n = 1-4), reducing the comparability with experimental data[120]. The use of Markov State Model (MSM) analysis, a method by which a large number of short trajectories are combined to describe the folding landscape, allows for a more complete sampling of the entire free-energy surface (see review for details: Pande VS et. al, [37]) and can therefore be readily compared to bulk experimental measurements (see Chapter 2).

Advances in experimental techniques are as necessary as advances in computational strategies if the convergence of timescales between the two methods is to occur with significant resolution to obtain meaningful agreement. Successful comparisons require agreement at all resolutions; residue-specific, pair-wise, and global, across the micro-to-millisecond timescales. Robust datasets spanning the microsecond-to-millisecond timescales will complement the breadth of data that can be obtained at slower time scales by conventional stopped-flow and manual mixing techniques at ms-to-hour timescales while also converging with growing simulation times. Here, advances in continuous flow (CF) mixing technology and the adaptability of such technologies to a range of techniques to provide both high and low resolution data for direct comparison to simulations are discussed. The analysis strategy described herein is a proposed guideline for achieving detailed structural information of folding events through the combination of MSM models and CF-mixing data.

Advantages of mixing techniques

Kinetic folding techniques and the unfolded state

Several techniques are capable of resolving early protein folding events. Laser-induced temperature jump (t-jump)[121] and pressure jump (p-jump)[122] experiments are examples of bulk kinetic experiments that have equivalent or better time resolution than CF-mixing experiments (Fig. 4.2). However, these techniques fail to be as generally applicable and unbiased as CF-mixing experiments with reference to the chemically denatured state. The chemically denatured state generally obeys random coil statistics by SAXS measurements[87] while temperature and pressure jump experiments reference the cold denatured and pressure denatured states, respectively. These states are highly biased by secondary structure and the hydrophobic effect owing to the native favoring conditions of the solvent[123,124]. Referencing the biased unfolded states either ignores or misrepresents higher energy states that could lend insight towards the fundamental basis of early folding events.

Fig 4.2. Experiment and folding timescales. Accessible timescales of biophysical experiments and simulations (above axis) compared to timescales of protein folding processes (below axis). Continuous flow mixing experiments overlap significantly with MD simulations and can observe nearly the entire process of tertiary contact formation. (Credit: Osman Bilsel)

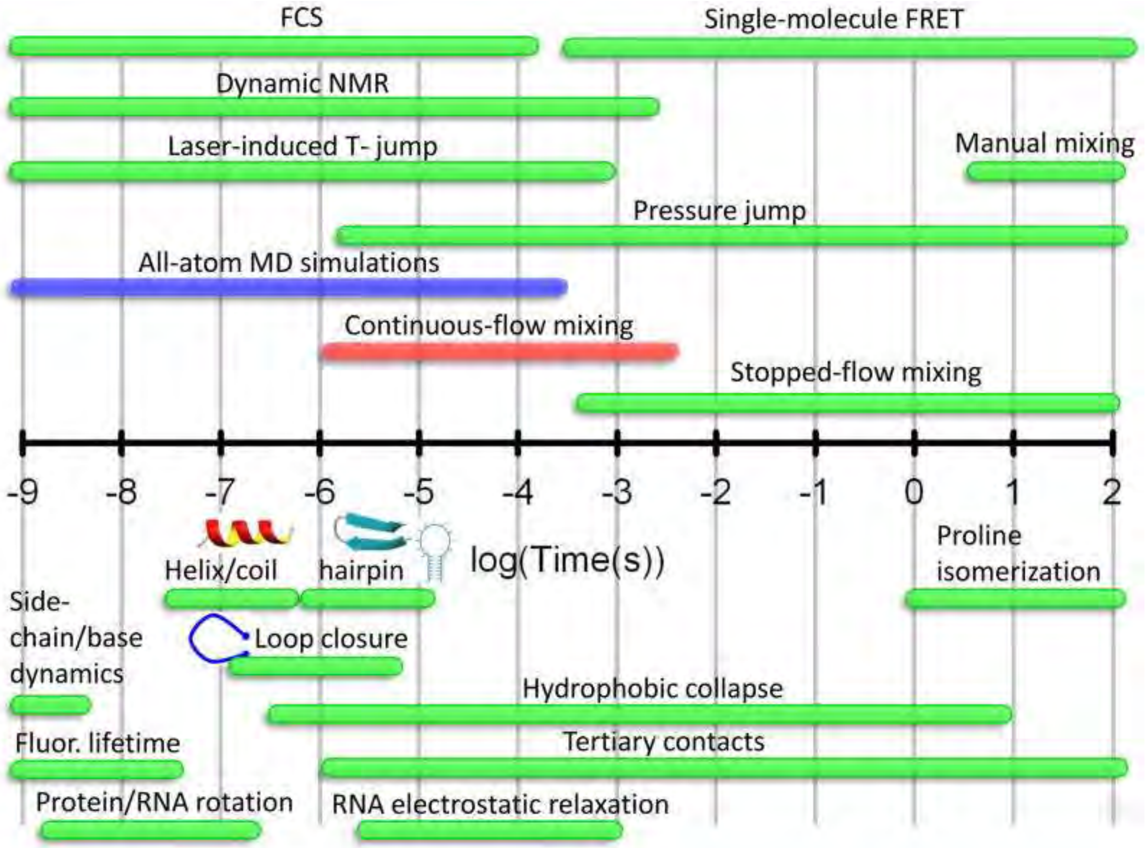


Figure 4.2. Experiment and folding timescales

Physical considerations for continuous flow mixing experiments

Mixing experiments rely on the dilution of one solvent into another resulting in rapid exchange of the solvent of the experimental system. The mixing time, *t* (sec), of this rapid dilution is dependent on the distance by which molecules must diffuse, *r* (cm), and the translational diffusion coefficient, *D* (cm^2 s^{-1}), appropriate for the size of the molecule (Eqn. 4.1)[73].

$$t = r^2/D \quad \text{(Eqn. 4.1)}$$

The diffusion distance is ultimately governed by the characteristics of mixing that are described by the dimensionless Reynolds number (Re) quantity. This value is calculated as the ratio of inertial forces to viscous forces (Eqn. 4.2)[125]

$$Re = \rho v d/\eta \quad \text{(Eqn. 4.2)}$$

where ρ is the density (g cm^{-3}), *v* is the flow velocity (cm s^{-1}), *d* describes the dimension scale of the channel (cm), and η is the viscosity (cm^2 s^{-1}) of the fluid. Under ideal mixing conditions, the lowest possible dead time is the diffusion time appropriate for the viscosity limited interspersion of the solutions. At an experimentally achievable diffusion length scale, i.e. 0.1 μm with microfluidic mixers, the fastest achievable dead time of mixing for biological macromolecule (when $D = 10^{-7}$ cm^2 s^{-1}) is 10 μs[39]. The two most common approaches towards achieving near-diffusion limited dead times with continuous flow microfluidics are laminar mixing with hydrodynamic focusing and turbulent mixing.

Laminar mixing is achieved at low Reynolds number flows (R_e < 2300). A protein refolding experiment under these conditions consists of a protein sample in high molar denaturant being injected between two sheathing flows containing the low denaturant refolding buffer (Fig. 4.3). Increasing the flow rates of the sheathing flows relative to the sample flow will result in narrower dimensions of the sample flow. Diffusion time across the distance of half of the central channel is the limiting factor resulting in the mixing dead time. Using this method, dead times as low as 10-20 μs have been reported[126,127] with protein flow rates as low as 0.5 nL s^{-1} at concentrations of 100 nM to 100 μM protein[128] for Förster resonance energy transfer (FRET) and tryptophan fluorescence experiments, respectively. Unfortunately this method of mixing is not as easily interfaced with techniques where water can contribute to background noise, such as small angle x-ray scattering (SAXS) (Fig 4.3). Additionally, at low Reynolds numbers friction between the solvent and the channel walls are not overcome by the flow velocity resulting in channel flow with a parabolic leading edge, that is, the flow at the center of the channel moves faster than at the walls. This phenomenon, as well as the diffusion gradient of the sample, must be corrected for to define the reaction time axis and protein concentrations throughout the channel.

Under high Reynolds number flows ($R_e > 4000$), mixing occurs in the turbulent regime. Turbulent mixing relies on fast flow rates and shear flows to create eddies, the dimensions of which limit the mixing efficiencies. Under ideal flow conditions many homogenous and small eddies would be created to approach diffusion limited mixing times. The advantages of turbulent flow over laminar flow mixing is that upon completion of mixing the entire channel is

Fig 4.3. Comparison of turbulent/chaotic and laminar mixing strategies. Turbulent and chaotic mixing (left panel) produces a homodispersed channel which is advantageous for SAXS detection methods. Laminar mixing with hydrodynamic focusing (right panel) maintains the sample in a narrow dimension flow, which is a disadvatage for SAXS measurements because the beam focus is larger than the sample flow, contributing to background noise.

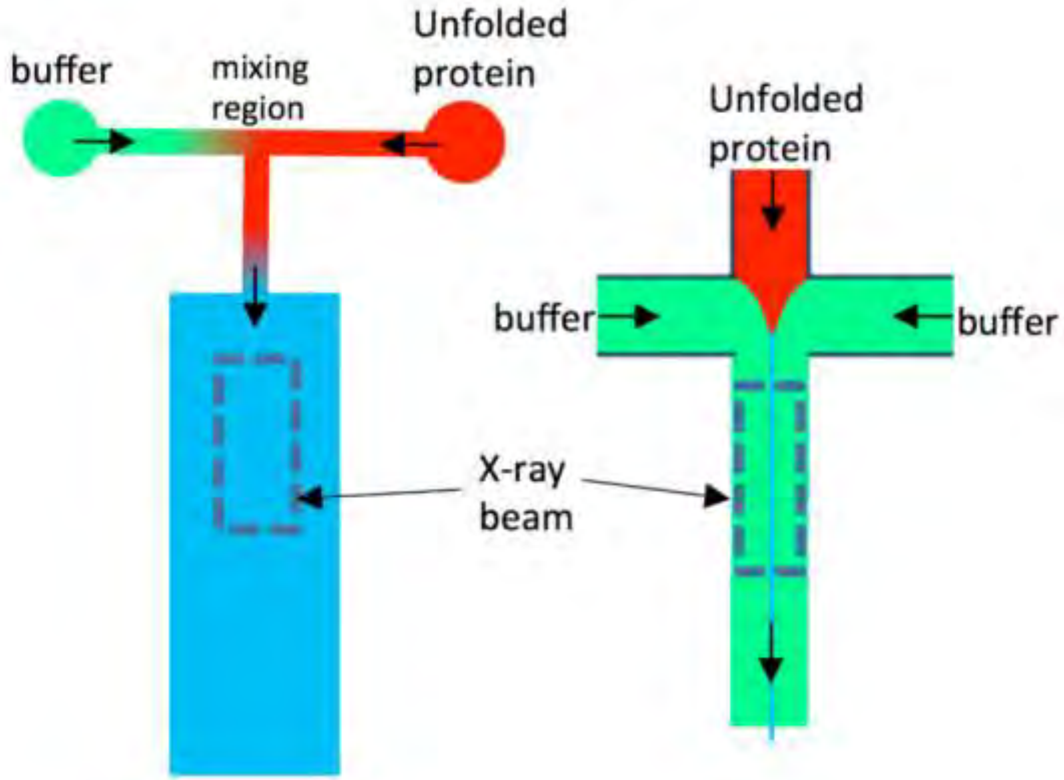


Figure 4.3. Comparison of turbulent/chaotic and laminar mixing strategies

homodisperse and the flow rates overcome the solvent to wall interactions creating a plug flow. Under these conditions the reaction time axis is simply proportional to the length of the flow channel with respect to the flow velocities and corrected for the mixing time. The homogeneity of the channel makes this mixer type ideal for increasing the signal to noise ratio of SAXS experiments, as well as reducing the operable concentrations of fluorescence and FRET experiments. Unfortunately, this mixing strategy relies on fast flow rates on the scale of mL min^{-1} which results in much higher sample consumption compared to laminar mixing.

A third and less explored approach to continuous flow mixing is chaotic flow which occurs in the transitional flow regime ($2300 < R_e < 4000$). In this flow regime both lamination of flow streams and turbulent character, such as vortices or eddies, are created which accelerate the mixing across the flow interface. Optimizing mixing in this regime is best accomplished by using mixer geometries to introduce inertial mixing through dramatic changes in flow direction (Fig. 4.4). Chaotic flow is advantageous because it retains the advantages of turbulent flow (i.e., homodisperse channel, and near plug flow), while also operating at lower flow rates to achieve lower sample consumption. The major drawback of chaotic mixing is the sacrifice of time resolution compared to turbulent mixing.

Fig 4.4. Design of a chaotic mixer for SAXS. (A) 3-d CAD drawing of the channel, sharp turns are included to induce inertial mixing. (B) COMSOL simulation of a 2 mL min^{-1} 1:10 dilution shows that pressures in the observation channel are atmospheric. (C) COMSOL simulation of the channel highlighting the velocities. The change in sharp turns show rapid changes in momentum that improve mixing efficiencies.

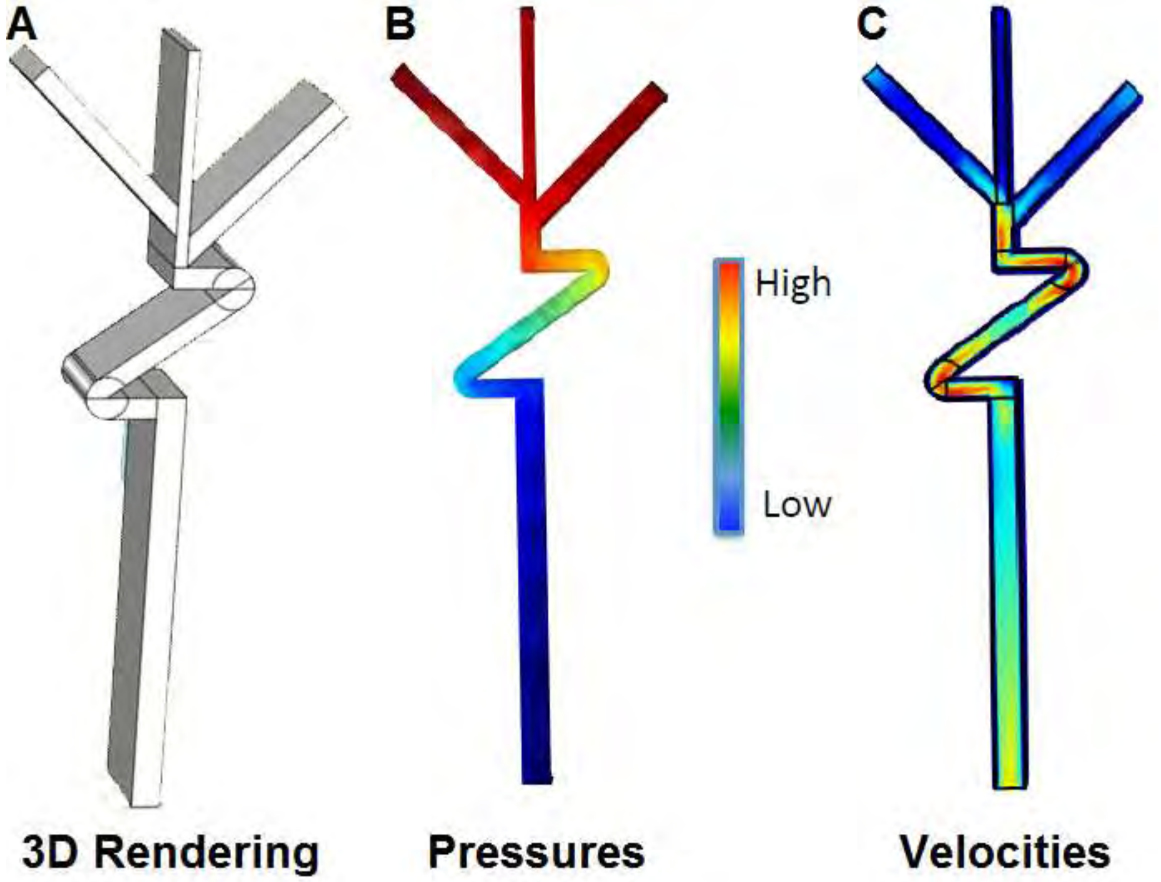


Figure 4.4. Design of a chaotic mixer for SAXS

CF-SAXS Development at BIOCAT

Challenges and considerations

Physical differences in mixing strategies are perhaps most pronounced with CF-SAXS measurements. X-ray diffraction patterns are produced when x-rays influence the oscillations of electrons which produces coherent scattered waves. As all electrons in the sample produce a scattered wave, the interference of these waves creates a diffraction pattern[129]. In homodispersed solutions of dilute biopolymers in aqueous solvent, resolving meaningful diffraction patterns are challenging due to the scattering contributions of the bulk water. In solution scattering experiments, the difference of concentration-weighted electron density of the solute must exceed that of the solvent for the scattering particles to be observable. Protein scattering experiments at concentrations as high as 2-3 mg mL^{-1} are generally low contrast. As mentioned above, the contrast suffers further in protein folding experiments when laminar mixing is used as the majority of the sample exposed to the beam contains no protein making turbulent and chaotic mixers ideal mixing strategies for obtaining these measurements.

Although working at higher concentrations can significantly improve the signal-to-noise ratio of the experiment due to the coherent nature of the scattered waves, it is not an ideal strategy for protein refolding experiments. During protein refolding hydrophobic residues and partially structured states are more accessible to intermolecular interactions, leading to aggregation. Aggregation obscures the Guinier region of the scattering profile and biases the R_g

measurement because long pairwise distances have a disproportionately large influence on the root-mean-square average over all pairwise distances. In the absence of aggregation and at high concentrations, typically in excess of 5 mg mL^{-1}, concentration-dependent interparticle effects become apparent and distort the low q region to bias the R_g determination towards lower R_g values.

Enhancing the signal-to-noise ratio without relying solely on the protein concentration can be accomplished by increasing the flux (photon counts sec^{-1} m^{-2}), which can be accomplished by using brighter synchrotron sources or narrowing the focus of the beam. In CF applications, narrowing the focus of the beam is advantageous because the mixer geometries can be correspondingly reduced, increasing the mixing efficiency and decreasing the dead time of the mixer. However, care must be taken in high flux measurements as radiation damage can occur due to the heat and free-radicals produced from the incident beam[130]. In chaotic and turbulent mixing this is often not a problem as the flow rates are generally sufficient to reduce exposure time and maintain the temperature of the sample.

Ongoing Developments

The development and optimization of the CF-SAXS apparatus at the Biophysics Collaborative Access Team (BioCAT) beamline 18ID at the Advanced Photon Source, Argonne National Laboratory focused on achieving a high flux micro-focused beam, ultra-fast mixing, and decreased sample consumption. The initial developments consist of implementing a Kirkpatrick-Baez (KB) mirror system[131] to micro-focus the source beam down to a 20 µm x 5 µm a turbulent mixer with post-mixing observation channel dimensions of 100 µm x 400 µm x 2 cm (Height x Depth x Length), a pilatus 100k pixel array detector, and a translation stage for the mixer. With this configuration the duty cycle is optimized (~85%) by scanning the mixer across the beam and concurrently acquiring detector images. Up to 90 points within the observation channel can be detected, with a dead time of 150 µs, and a total acquisition time of 2-3 s per point for every 20 mL of 2-3 mg mL^{-1} of injected protein. The large sample requirements of this setup have motivated further developments.

A major consideration in signal quality is the flux loss acquired by implementing the KB based micro-focus. The maximum acceptance of the KB mirror system is 0.5 mm x 0.5 mm, which excludes 90% of the available flux[90]. We have since removed the KB mirrors and replaced them with an upstream compound refractive lens (CRL). With a focal distance of 2 m the CRL has a comparably reduced beam divergence from the 0.5 m distance of the KB micro-focus. The resulting compact cross-sectional area of the beam at the detector extends the observable small q range to achieve a comparable q range to conventional equilibrium experiments, while increasing the flux and maintaining the focused beam dimensions at the sample (20 µm x 5µm).

Advances in chaotic mixing are also being made. Using the finite element analysis software COMSOL, we are designing new mixer geometries and testing them *in silico* (Fig. 4.4). These efforts are focusing on the development of chaotic mixers that should reduce the sample requirements 4-8 fold while maintaining an equivalent signal-to-noise ratio at a minimal cost of ~30% increase in dead time. While the dead times of turbulent and chaotic mixers have approached the diffusional limits of mixing over the last 60 years, reduction of sample consumption has been slower to progress and remains to be nearly an order of magnitude higher than that of laminar mixers, in most cases (Fig. 4.1). Successful implementation of these designs, in conjunction with careful volume handling, is expected to reduce the sample requirements from ~200 mg a scan to ~50 mg with a dead time of ~300 μs.

Further developments in data processing are also underway to make the process more streamlined including the automation of the masking step required for data reduction. Acquiring 50 scans of data (25 for protein and 25 for blank), each consisting of 90 detections, produces 4500 tif image files that must be analyzed in the correct order for each kinetic jump. The first step in processing the image files is to mask out the beamstop and parasitic scattering flares, prior to circularly averaging the diffraction pattern. As the majority of SAXS users do not change the window position of the observation cell while collecting this step has historically been done manually. Translating the mixer window requires a mask at each position where data is collected, requiring 90 manual maskings. During the initial CF-SAXS development a python/Fit2D script was written to automatically mask and circularly average the data using intensity thresholding. The exported data was then imported into a custom in-house GUI-based software package, CF-SAXS-bot, for further manipulation. CF-SAXS-bot maintained the bookkeeping of the data while manual or automated manipulations were applied, including: removal of low quality data, averaging, subtractions, simple transformations, SVD, and Guinier analysis. In this way the data could quickly and easily be manipulated on a point-by-point or scan-by-scan basis. Although successful, the software is being replaced by the DELA software package being developed by Lambright et al[40]. DELA contains more functions and is suited for flexible and large analyses with its scriptable platform. Further developments in automated masking are also being pursued to exploit the low electronic noise of the photon counting Pilatus detectors to enforce Poisson distributions at each q (Fig 4.5) that will remove electronic noise, the beamstop, and parasitic scattering in a model independent manner and without thresholding.

Global Measurements with CF-SAXS

Monitoring global measurements of a folding reaction define the context for the interpretation of higher resolution measurements. SAXS is a particularly useful as the dimensional data can provide both an R_g as well as low resolution structural models[132] . The R_g values are extracted from the low q region of the scattering curve using the Guinier approximation. The simplest implementation of the Guinier approximation is to first transform the I(q) versus q data of the

Fig 4.5. Improving CF-SAXS data analysis. Comparison of the distribution of intensities at two frames (0.1 ms [top] and 1.9 ms [bottom]) show significant deviation from the expected Poisson distribution. The intensity distribution at low Q for the two frames with either buffer or protein is shown with a fit of a Poisson distribution around the main peak. The outliers are distinctly different between the two frames, indicating the need for treating each frame to a separate mask. (Credit: Sagar Kathuria)

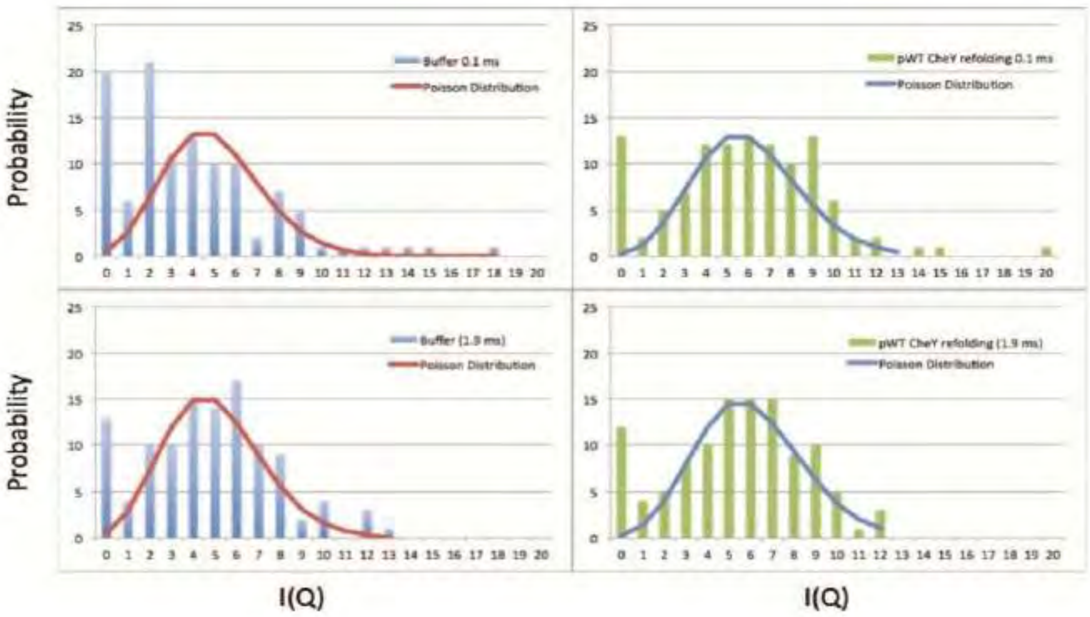


Figure 4.5. Improving CF-SAXS data analysis

scattering profile to Ln[I(q)] versus q^2 and fit the low q^2 region to the linearized expansion of the Guinier approximation in the form of y = mx+b (Eqn 4.3)[129],

$$\mathrm{Ln}[\mathrm{I}(\mathrm{q})] = -\frac{\mathrm{R}_\mathrm{g}^2}{3}\,(\mathrm{q}^2) + \mathrm{Ln}[\mathrm{I}_0] \quad \text{(Eqn. 4.3)}$$

where the slope of the line is described by $Rg^2/3$, I(q) is the scattered intensity, q is the magnitude of the scattering vector, I_0 is the zero angle intensity, and R_g is the radius of gyration. Using the Guinier approximation with good quality data yields high quality time resolved dimensional data (Fig. 4.6) that can describe the global folding kinetics. However, comparison of R_g measurements to simulation data sets is not particularly useful because pairwise conformational constraints are absent.

A useful metric for comparison with vast computational datasets is the pair-distance (P(r)) distribution, which describes the pair-wise distances across all pairs. The P(r) distribution can be fit to a given scattering profile and is sensitive to small conformational changes. Comparison of experimental P(r) distributions to calculated P(r) distributions of simulated subpopulations of structures on equivalent time-scales provides a more detailed comparison of the two datasets than using the singular R_g values. A time resolved comparison can be used to streamline the analysis of the simulation data and identify high probability structures from the simulations from which structural insights can be observed.

Fig 4.6. CF-SAXS data of Cytochrome C during sub-millisecond refolding. (a) The radius of gyration obtained by Guinier analysis, R_g (circles), during refolding from 4.5 M GdnHCl to 0.45 M GdnHCl at pH 7.0 in the presence of 0.2 M imidazole and modeling of R_g^2 versus time (continuous line) to a double exponential with fixed time constants of 45 μs and 650 μs are shown. The final protein concentration was 2 mg mL−1 and the total flow rate was 20 mL min−1. (b) The zero-angle scattering intensity obtained from the Guinier analysis in (a) shows that Cytochrome C is monomeric throughout folding.

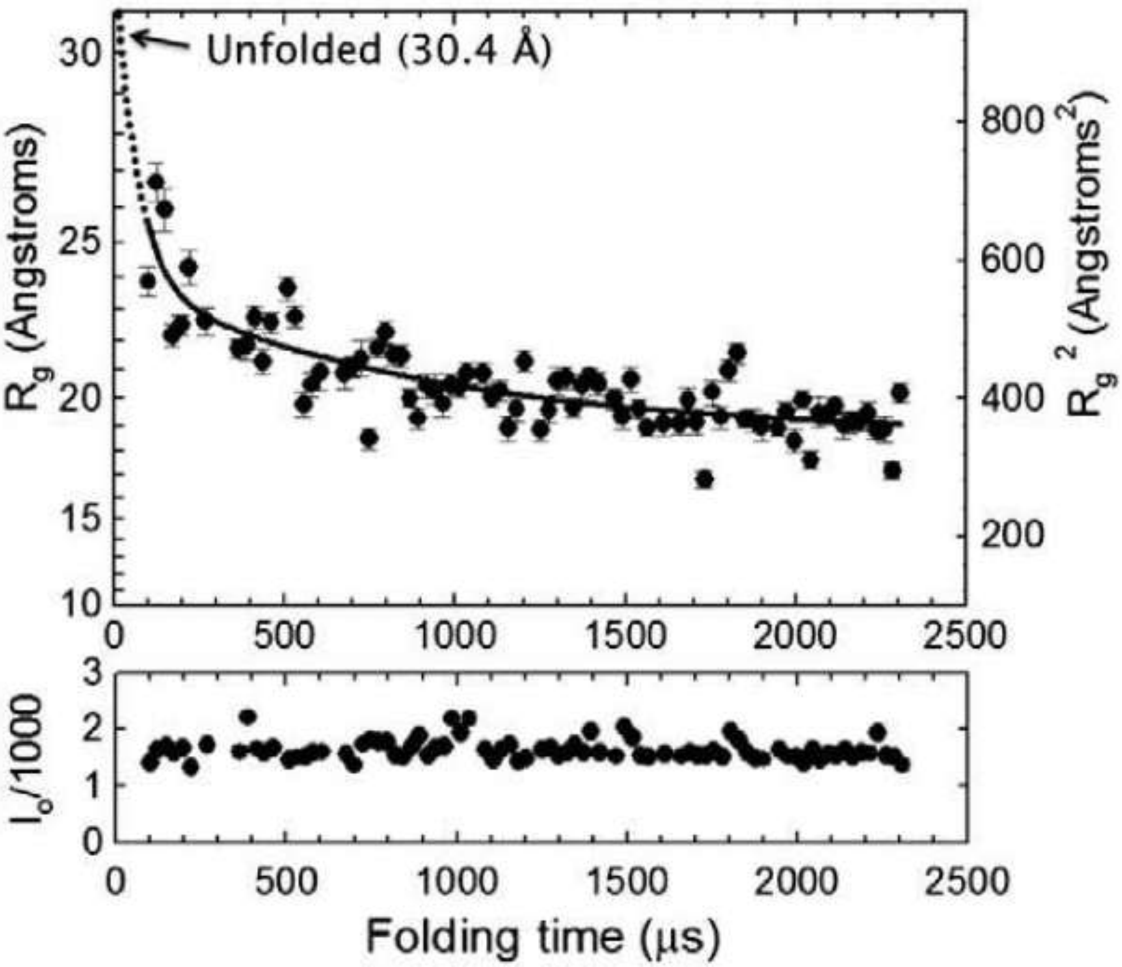


Figure 4.6. CF-SAXS data of Cytochrome C during sub-millisecond refolding

Pair-wise measurements using CF-SAXS and CF-TCSPC

Pair-wise SAXS measurements have been made successfully with large DNA macromolecules labeled with gold nanoparticles[133,134]. Application of this approach to globular proteins presents challenges with labeling efficiencies and distance resolution as the sizes of proteins are of the same scale as the nanoparticles. Experiments are currently underway to achieve efficient labeling of 2 gold labels per protein molecule using gold nano-crystals, which are monodispersed at sizes approaching 1.5 nm (unpublished data; Kevin T, Halloran, et al.).

Time correlated single photon counting (TCSPC), like SAXS, can be conducted label-free in the case of intrinsic tryptophan fluorescence experiments (see chapter 3), or with labels for monitoring changes in FRET. The basis of this approach relies on low-probability single photon emissions that are sampled over a large number of excitation events where each photon has a known emission time and arrival time. The constructed time-correlated photon distributions from these experiments reveal the fluorescence lifetime decay of the probe. Observing the decay of the probe is advantageous because, unlike total intensity measurements, the lifetime measurements are concentration independent such that normalization of the data is not required for comparison across datasets. Additionally, the measurements have a high signal-to-noise ratio, and changes to individual lifetime amplitudes can be observed, providing separate measurements that can yield more descriptive data than a single readout.

However, the most noteworthy advantage to TCSPC measurements, when using a FRET system, is the ability to fit the photon count distribution to a P(r) distribution, which can then be directly compared to computational and global SAXS measurements of the same system achieving domain specific distance information (Fig. 4.7). The disadvantage of this technique over a comparable SAXS approach is that the distance dependence of FRET is weaker as it scales with $1/r^6$, limiting observations to distances smaller than the length scale of most biomolecules[135]. In comparison, pairwise SAXS distances scale as $1/r^2$, permitting accurate measurements of longer distances[40] on the length scale of most biomolecules. However, in a combinatorial analysis short distance measurements can provide significant clarification of contextual data[117]. In conjunction with computational datasets, short distance measurements can be used to orient a structure within a low resolution structural model as well as refine structures within computational groupings.

Fig 4.7. SsIGPS CF-TCSPC FRET. SsIGPs E63W- R238C trp-EADANS FRET pair is refolded from 70 µs to 1.07ms and binned every 166 µs (DA1 through DA6). Fitting the data to a bimodal distribution reveals two structural intermediates with populations changing over the reaction time. (Credit: Kevin Halloran)

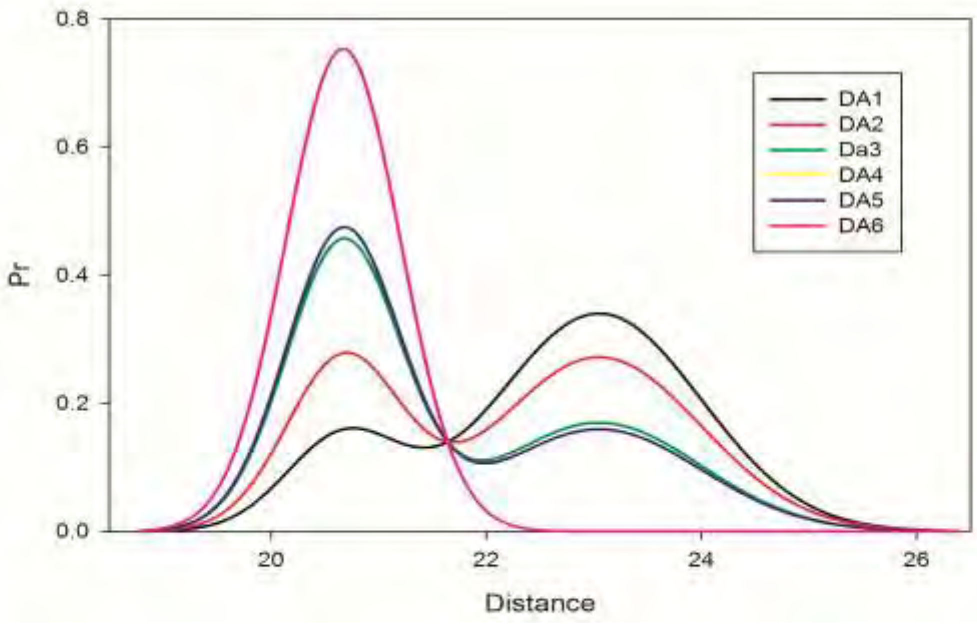


Figure 4.7. SsIGPS CF-TCSPC FRET

Residue specific labeling with FPOP

Residue specific resolution of folding processes is difficult to obtain experimentally. The highest throughput approaches for experimentally observing residue specific information of protein folding events are hydrogen-deuterium exchange (HDX)[136], and fast photochemical oxidation of proteins (FPOP)[137,138]. Using these techniques, the folding process is pulsed at variable time points for approximately 1 µs with a chemical label. These labels probe the solvent exposure of each residue at the pulse time.

The reactions are subsequently quenched and the proteins are then analyzed by mass spectrometry or NMR.

Labeling partially folded states of marginal stability by HDX has been accomplished with labeling events occurring as fast as 90 µs, labeling at a pH 9.8 in a competition format[136]. Ideally, refolding would occur for a period of time before applying a pH pulse label, labeling all non-structured elements at the pulse time point. In a competition format, the refolding reaction occurs at basic pH such that the folding events occurring faster than the rate of labeling are protected and slower events are labeled. Application of this technique within the microsecond time regimes is exceedingly challenging due to the high pH of the pulse required for labeling on that timescale, which may perturb the fragile structures being probed. Further complicating the data acquisition is the occurrence of back-exchange that occurs post-quench. Back-exchange decreases the signal of the experiment requiring carefully controlled conditions for reproducible and accurate results to be obtained[139].

An analogous technique, fast photochemical oxidation of proteins (FPOP), uses the photochemical production of hydroxyl radicals to label amino acid sidechains. The covalent nature of the oxidative labeling eliminates the complication of back-exchange, simplifying the post-quench sample handling conditions compared to HDX. Recent studies have successfully labeled proteins in the continuous flow time regime[140,141] using laminar mixers. However, the current studies have demonstrated low labeling efficiency with good time resolutions[141] or good labeling efficiencies with poor time resolution[140]. In the former study, the mixer efficiency provided an excellent time resolution but poor labeling efficiency due to the low flux of their doubled Argon ion laser. The latter had a better labeling efficiency from the higher flux of a KrF excimer laser but lower time resolution due to the beam width and mixer inefficiencies. Experiments using a synchrotron source to produce hydroxide radicals through the radiolysis of water that appear to efficiently label amino acid side chains due to the higher flux[142]. Future development work at the BIOCAT beamline will exploit the micro-focus capabilities used in the CF-SAXS measurements to produce high flux and high time resolution FPOP labeling. Further increases in flux will be made by removing the monochrometer from the beamline apparatus and by employing a cylindrically hydrodynamically focused laminar mixer such that entire sample volume is within the beam focus. After collection, the exposed protein will be analyzed using proteolytic digests with tandem mass spectroscopy to identify the labeled residues and provide complementary high resolution structural data of the same kinetic process observed in the scattering experiment.

By themselves, HDX and FPOP data can be challenging to accurately compare to simulation data as they indirectly describe the solvation of the labile nucleus. It is possible to calculate the buried surface area or local dynamics of the labeling sites for a given structure[143] so that those metrics can be correlated to the experimental exchange rates, however an unguided analysis would be computationally challenging. Having first refined the computational data with lower resolution experimental data, the buried surface area or dynamic calculations would be far more targeted and accessible. If this process were successful then the kinetic trajectory of the simulation would be validated by experimental measurements and further structural details of the folding process could reliably be extracted.

Discussion

Experimentally, folding kinetics are observed as simple exponential responses. Generally, computational studies show a much more complicated view of protein folding through extensive potential pathways in MSM models and variations between individual trajectories in full trajectory MD. In broad agreement with folding simulations, landscape theory[52] describes a complex free-energy landscape that implies many favorable routes through which a protein could traverse the free-energy difference from the unfolded to folded state to the native basin. Presumably, experiments resolve a weighted average of events, with observations being most sensitive to the highest populations, or dominant folding pathways.

Supercomputers like Anton[120] are capable of millisecond timescale full trajectory simulations. However, the computational time to conduct these simulations is still taxing, limiting the number of full trajectories used in a single analysis. Without significant sampling it is difficult to determine the statistically significant features of each different trajectory in the absence of experimental data. Therefore computational approaches that rely on a relatively small number of full folding trajectories are not ideal for comparison to bulk measurement experiments.

An alternative computational approach, MSM modeling, uses a large number of independent short trajectories and models an energy landscape based on the interconversion rates between microstates[37]. Comparison of MSM data to experiments is ideal because MSM data represents a more complete view of the entire energy landscape, and thus is directly compatible with comparisons to bulk measurements. Although there exist methods to determine the major pathways of such models[144,145], the analysis becomes more challenging with larger datasets and when complicated folding mechanisms are involved (i.e., proline isomerization[146], disulfide bridges[23], etc.) such that assumptions and secondary processes must be employed to generate a complete dataset (see chapter 3). Furthermore, implicit solvents and other assumptions are often used in long folding time simulations to reduce the computational time. The concessions made produce an expectedly imperfect dataset such that absolute quantitative agreement is never expected. In these cases a guided analysis with experimental data is an ideal compromise between computational time and model independence.

Assuming the time axes of the experiment and simulation are reasonably aligned, the application of experimental data for MSM model refinement as described above can be applied with little bias. Beginning with global distance distributions, the high probability MSM states can be refined by changing the boundaries, within reasonable physical limitations, to maximize agreement with the experimental data through comparison of distance distributions. Once a reasonable microstate or grouping of microstates is found with reasonable representation in the experimental data, further refinement can be made by optimizing the boundary conditions to best fit all of the global and pair-wise distributions simultaneously. Residue specific data could then be used to assign a weighting factor to each structure based on the likelihood that it is significantly contributing to the experimental signal. A final round of refinement could then be conducted to optimize the weighting factors with the simultaneous optimization of all distance distributions.

With this approach to model refinement, the experimental mechanism is not enforced upon the interpretation of the simulation. Instead, the simulation time axis is effectively refined to match the experimental time axis with high structural agreement. If the calculated distance distributions are in good agreement with the experimental distributions then the simulated structures are highly likely to be representative of the experimental ensemble. High resolution structural details can then reliably be observed, and subsequent experiments based on these observations can further validate the agreement between the two datasets.

Chapter V - Discussion

Summary

Physicochemical properties that bias early folding events

Results form Chapter II provided evidence to suggest that ILV clusters contribute significantly to the formation of the off-pathway intermediate in CheY. Altering the chain connectivity, relative to the calculated ILV clusters, via circular permutation is a significant enough perturbation that the populations of the on- and off-pathway intermediates can be modulated in a seemingly rational way. We suggest that this result is due to both the size and low CO of the ILV clusters. In the WT connectivity the smaller but sequence local ILV cluster (cluster 1) initiates the earliest folding events as it outcompetes the larger cluster (cluster 2) that is split in half by the location of the N and C termini. Changing the chain connectivity by linking the larger cluster together in contiguous sequence, permits the larger low CO cluster to outcompete the smaller cluster and collapse first. This prevents the frustration imparted on the folding reaction by the early formation of cluster 1 and disfavors formation of the off-pathway intermediate.

Events driven by the collapse of cluster 1 result in frustration of the folding reaction, seen by Gō-models as the impedance of forming native contacts through structuring regions that must come apart before other native contacts can form. At a glance, this frustration seems to be founded on physical principles that are not specific to protein folding. However, Gō-models are particularly well suited for low CO folding[69] and therefore are useful tools for describing low CO protein folding events at low resolutions. In fact, the atomistic models run on CheY* and Cpβ4 are in general agreement with the findings from the Gō-model simulations.

The higher resolution details from atomistic models offer a glimpse into other features that may be playing a role in observed frustration. The most obvious structural defect that may be impeding folding is the non-native order of the central β-sheet. In the native configuration β1 is intercalated into the center of the β-sheet, between β2 and β3. The atomistic simulations suggest that the off-pathway intermediate has not yet incorporated β1 in the correct strand ordering, requiring an un-structuring event before productive folding can continue. Structurally, both sets of simulations describe the subdomain interface as the region which must remodel for folding to continue, which according to the atomistic models, includes correcting the strand ordering of the β-sheet. In atomistic models of the Cpβ4 connectivity, the strand ordering is not observed to be non-native and therefore supports both the experimental and Gō-model data in describing the off-pathway intermediate as less favorable.

Results in Chapter III suggest that ILV residues and long range electrostatic interactions give rise to a compact unfolded state under native-favoring conditions. The coincidence of the HSQC cross-peaks of this compact unfolded state with the native state suggest that the residual structure on the unfolded side of the barrier is native-like and therefore structural insights can be gained from superimposing results onto the known structure. In this way we observe an intact hydrophobic core of mostly ILV residues flanked by solvent exposed polar residues and a β-sheet stabilized in the correct strand order by salt bridges inclusive of the intercalated terminal β-strand. The correct strand order is therefore maintained in the readily accessible unfolded state, which is effectively imprinted with the native topology, driving a fast and smooth folding reaction towards the native state.

Additional features on the of the free-energy landscape of Di-III_14 at higher energies are presumably also resulting from the electrostatic contributions to the stability of the structured unfolded state. These complexities are not seen in the data obtained from of Gnd denatured state and access to the exchange competent state can be accelerated at higher ionic strengths. Interestingly, this response suggests that even the exchange competent unfolded states appears to contain residual structure. The general implications of this result are obfuscated by the unnatural sequence composition of Di-III_14, which we assume is largely responsible for the result. However, a few hydrophobic residues, M48, L16, F22, L51, and A52, maintain observable exchange rates at 0.5 M NaCl, suggesting that hydrophobic interactions may still play a significant role in biasing the unfolded state under native conditions for non-polyampholyte sequences.

Increasing the resolution of early events in folding

Chapter IV discusses past and present technological developments for improving mixing strategies and detection techniques to enhance experimental time resolution. Overall, the best option for improving resolution appears to be generating high confidence atomistic simulations. Apart from theoreticians improving the simulation strategies, experimentalists can focus on improving time resolution of various kinetic experiments and obtain large datasets of easily comparable data. This chapter suggests a strategy for using distance distributions to refine large MSM models to obtain high structural confidence so that high resolution details could be obtained.

Chapters II and III had each incorporated computational components that suggest that, for the most part, the details we get from simulations are consistent with experimental data. Chapter II was computationally intensive and, overwhelmingly, the most challenging part of combining experimental and computational data aligning the both on the same time axis. For instance, in the absence of the CF-SAXS data on CheY and Cpβ4 we would have no reasonable way to determine agreement between the two data sets. Only after aligning the relative R_g measurements were we able to make sense of the Gō-models relative time scale and leverage the simulations for higher resolution details. Although this single metric, by itself, is not a confidence inspiring amount of significant overlap between the datasets, previous experiments on CheY using a similar combinatorial approach[15] further suggests that comparisons across the datasets are appropriate.

The MSM models, briefly discussed in Chapter II, are currently still being analyzed. Finding agreement between MSM models and experimental data is far more straightforward than coarse grained or single trajectory simulations, for 3 reasons. First, the time axis can be independently calculated with a level of confidence suitable for comparison to experimental timescales. This particular point is a huge advantage because this cannot be done accurately with coarse grained models, including Gō-models, and it is capable of immediately aligning experimental and computation data with some degree of confidence. Second, the positional accuracy of amino acids is much better which makes comparisons to experimental data as straightforward as using simple calculations like distance distributions. And lastly, MSM models, unlike full trajectory simulations, recapitulate a broad energy surface, much like bulk experiments, so comparisons can be refined with population statistics to enhance structural accuracies.

The ongoing analysis of the Folding@Home CheY simulations is nonetheless still challenging. The large size of the dataset is very difficult to work with and assumptions had to be made to make the simulation reasonable, like using implicit solvent and modeling the proline isomerization reaction, which complicate the downstream analysis. Current efforts are being made to improve agreement with the experimental data, which include refining the model with SAXS data. The advantage of using this approach with a broad model like an

MSM, is that refinement can be done independent of a model. In other words, refining a time segment of an MSM to fit an experimental distance distribution does not consider the connectivity of states that are being weighted. Therefore, as long as the MSM analysis is originally reasonable, the time axes are in agreement, and the experimental data is accurate, fitting to the experimental data will only improve the quality of the MSM model by unbiasedly removing states and therefore pathways that are not represented in reality.

Future directions

The CheY system

Results from Chapter II suppose that ILV clusters play a large role in the early folding events, causing one cluster to out compete the other depending on the connectivity. These inferences are made from the results of the computational models, and therefore it would be prudent to test this hypothesis with further experiments. One approach that we are actively pursuing is a multi-color FRET experiment to observe correlated motions during folding. In this way we will be able to determine if the clusters are, in fact, in competition with one another. Sagar Kathuria is currently developing an split-intein strategy to accomplish this. By splitting CheY and the permutants into sections that can be independently labeled we will not have to rely on stochastic labeling methods and can therefore be assured of the site specific labels, while also increasing the labeling efficiency and simplifying the purification process.

The supposed on-pathway intermediate is also an interesting experimental target. Experimentally this intermediate has been optically silent across the set of permutants. The only experimental data consistent with it is from the Serrano group[20], which was conducted by NMR at equilibrium. Single molecule fluorescence correlated spectroscopy (sm-FCS) may therefore be an appropriate experiment to confirm the presence of this intermediate. In this experiment, low concentrations of CheY labeled with a high quantum yield FRET system would be observed at equilibrium. At sufficiently low concentrations and low focal volumes the experiment would be statistically observing single molecules. The fluctuations of the emission intensities would reflect different distances between the dyes that are sampled at equilibrium. If the FRET system is sensitive to the on-pathway intermediate, then integrating the dataset across the distance domain should reveal distance distributions with an observable on-pathway state, analogous to the analysis of force puling experiments[147].

The "Di" set of proteins

Experiments of the *de novo* designed protein Di-III_14 reveal promising results that are largely consistent with the design expectations, and notably improved from previous attempts[32]. However unexpected complexities have still been observed by experiment that must be understood to further advance *de novo* design strategies. In the original design[70], general principles for design rules were applied with seemingly good success. Therefore identification of general principles and application to further design may avoid these complexities in the future. Specifically, the two major findings from Chapter III suggest that electrostatics play a large role in the complexities by forming salt bridges and perhaps by favoring high energy structure due to the charge segregation across the sequence.

Developing the design rules for salt bridges from the Di-III_14 construct requires straight forward mutational experiments. Mutating out a few or all of the salt bridges and testing the effects on the kinetics and the observed structured unfolded state would serve to identify their contributions to "smoothing" the energy landscape, and essentially provide a test for the reoccurring strand-order hypothesis of this thesis. From the current data set, it would be expected that reducing the number of salt bridges would lower the barrier between the structured unfolded state and the exchange competent unfolded state. Likewise, the response of removing all salt bridges is expected to completely remove the structural complexities on the unfolded side of the barrier while introducing complexities on the native side. It is possible that the reduction to some ideal number of salt bridges would yield a smooth energy landscape and would therefore serve as a basis for further design rules.

The charge segregation of the Di-III_14 sequence may be leading to a compact unfolded state that resembles an IDP-like structure under native conditions. Scrambling the charge orders of the β2-β4 salt bridges should be an effective way to test if the charge segregation is significantly driving the structure. Alternatively, identifying permutations that would reduce the charge segregation would provide different means for the test while isolating the results from the independent effect of the salt bridges.

Further studies across the current Di set of proteins are also warranted. The result of Di-III_14 are promising, but it is possible that these results are not representative of the set. As mentioned in Chapter III, the Di_1-IV protein is the only other protein in the set that does not have helices packed onto both sides of the β-sheet, and the salt bridge content is significantly different. If the salt bridge hypothesis we propose is correct then Di_1-IV would be expected to have a smooth energy surface on both sides of the barrier. If this turns out to not be the case then other general design principles might be illuminated that are more generalizable. Additionally, having helices on both sides of the β-sheet would imply that there are other mechanisms than salt bridges that can be used to enforce the correct strand order, identifying further design principles. Although the results from Chapter II suggest that ILV clusters may a play a role in this, it is yet to be seen if it applies to the Di set of proteins with similar topologies, or even if the Di proteins with more complicated topologies are experimental 2-state.

Conclusion

The results from Chapters II and III discuss early events in folding with an emphasis on the content of isoleucine, leucine and valine residues (i.e, ILV clusters). Previous research has proposed that ILV clusters contribute significantly to the native state stability of globular proteins[15,56,57,103]. These clusters have also been implicated in the formation of kinetic traps in the βα-repeat proteins CheY, NtrC, and Spo0f[15]. Although ILV clusters appear to be important, they are perhaps more of a general principle to all globular proteins due to their physicochemical properties. Specific to βα-repeat proteins, the strand organization of the β-sheet appears to be largely deterministic of the roughness of the energy surface. In both CheY and Di-III_14 terminal β-strands are intercalated within the β-sheet. In Chapter II we observe differences in folding related to β-sheet formation where efficient formation reduces the tendency for off-pathway species to be populated. Likewise, in Chapter III, the correct strand order is enforced with salt-bridges, driving a smooth 2-state folding reaction. Although the exact contributions of both ILV clusters and electrostatics in mitigating the complications of the strand ordering are yet to be described in detail, our results implicate both in being capable of significantly resolving the energetic of the strand ordering events in early folding reactions. In the βα-repeat proteins that have the β-sheet organized topologically between sets of α-helices, ILV clusters are likely to contribute to the strand ordering events, while in simpler topologies where α-helices are packed on to one side of the β-sheet, salt bridges on the solvent exposed side may prove to be efficient solutions to the strand order organization.

General perspective

The βα-repeat motif is found extensively across known structures of biological proteins. These proteins represent a large fraction of known enzymes[10] and signaling proteins. Understanding the folding of these proteins towards appreciable design principles may someday result in efficient methods for rationally stabilizing commercial enzymes and the efficient *de novo* design of novel enzymes. More generally, fundamental principles representative of such a large representative population of proteins may be applicable to other fold motifs. General principles found across motifs have the potential to produce efficient assumptions of folding processes that could be used to increase the efficiency of computational and predictive modeling. Further, physicochemical principles that govern misfolding events and biases in the unfolded state under native conditions can potential provide insights towards the underlying processes and causes of aggregation diseases. A better understanding of the thermodynamic processes can potentially lead to the rational development of molecular chaperone-type drugs.

The data presented within this work suggests that a small subset of 3 of the 21 amino acids, ILV, are significantly represented in early misfolding events and biases in the unfolded state under native-conditions. Similarly, salt bridges and charge segregation may have similar effects. Further studies are required to elucidate the how general these themes are to develop efficient design principles and progress the general understanding of the protein folding phenomenon.

References

1. Hartley, H. Origin of the word "Protein."*Nature* **168,** 244 (1951).

2. Hardy, W. B. On the structure of cell protoplasm. *J. Physiol.* **24,** 158–210 (1899).

3. Anson, M. I. & Mirsky, A. E. On some general properties of proteins. *J. Gen. Physiol.* 169–179 (1925).

4. Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G. & Wyckoff, H. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* **181,** 662–666 (1958).

5. Anfinsen, C. B. Principles that govern the folding of protein chains. *Science (80-.).* **181,** 223–30 (1973).

6. Levinthal, C. Are there pathways for protein folding? *Extr. du J. Chim. Phys.* **65,** 44 (1968).

7. Lau, K. F. & Dill, K. A. A lattice statistical mechanics model of the conformational and sequence space of proteins. *Macromolecules* **22,** 3986–3997 (1989).

8. Ueda, Y., Taketomi, H. & Go, N. Studies on protein folding, unfolding, and fluctuations by computer simulation. A three-dimensional lattice model of lysozyme. *Biopolymers* **17,** 1531–1548 (1978).

9. Anantharaman, V., Aravind, L. & Koonin, E. V. Emergence of diverse biochemical activities in evolutionarily conserved structural scaffolds of proteins. *Curr. Opin. Chem. Biol.* **7,** 12–20 (2003).

10. Wierenga, R. K. The TIM-barrel fold: a versatile framework for efficient enzymes. *FEBS Lett.* **492,** 193–8 (2001).

11. Farber, G. K. An alpha/beta-barrel full of evolutionary. *Curr. Opin. Structual Biol.* **3,** 409–412 (1993).

12. Bilsel, O., Zitzewitz, J. a, Bowers, K. E. & Matthews, C. R. Folding mechanism of the alpha-subunit of tryptophan synthase, an alpha/beta barrel protein: global analysis highlights the interconversion of multiple native, intermediate, and unfolded forms through parallel channels. *Biochemistry* **38,** 1018–29 (1999).

13. Forsyth, W. R. & Matthews, C. R. Folding hmechanism of Indole-3-glycerol Phosphate Synthase from Sulfolobus solfataricus: A test of the conservation of folding mechanisms hypothesis in (βα)8 barrels. *J. Mol. Biol.* **320,** 1119–1133 (2002).

14. Gu, Z., Rao, M. K., Forsyth, W. R., Finke, J. M. & Matthews, C. R. Structural analysis of kinetic folding intermediates for a TIM barrel protein, indole-3-glycerol phosphate synthase, by hydrogen exchange mass spectrometry and Gō model simulation. *J. Mol. Biol.* **374,** 528–46 (2007).

15. Kathuria, S. V, Day, I. J., Wallace, L. A. & Matthews, C. R. Kinetic traps in the folding of beta alpha-repeat proteins: CheY initially misfolds before accessing the native conformation. *J. Mol. Biol.* **382,** 467–84 (2008).

16. Hills, Jr, R. D. *et al.* Topological frustration in beta alpha-repeat proteins: sequence diversity modulates the conserved folding mechanisms of alpha/beta/alpha sandwich proteins. *J. Mol. Biol.* **398,** 332–50 (2010).

17. Lacroix, E., Bruix, M., López-Hernández, E., Serrano, L. & Rico, M. Amide hydrogen exchange and internal dynamics in the chemotactic protein CheY from Escherichia coli. *J. Mol. Biol.* **271,** 472–87 (1997).

18. Munoz, V., Lopez, E. M., Jager, M. & Serranot, L. Kinetic Characterization of the Chemotactic Protein from Escherichia coli, CheY, Kinetic Analysis of the inverse hydrophobic effect. *Biochemistry* **33,** 5858–5866 (1994).

19. López-Hernández, E. & Serrano, L. Structure of the transition state for folding of the 129 aa protein CheY resembles that of a smaller protein, CI-2. *Fold. Des.* **1,** 43–55 (1996).

20. Garcia, P., Serrano, L., Rico, M. & Bruix, M. An NMR view of the folding process of a CheY mutant at the residue level. *Structure* **10,** 1173–1185 (2002).

21. Volz, K. & Matsumura, P. Crystal structure of Escherichia coli CheY refined at 1.7-A resolution. *J. Biol. Chem.* **266,** 15511–9 (1991).

22. Hills, Jr, R. D. & Brooks, III, C. L. Subdomain competition, cooperativity, and topological frustration in the folding of CheY. *J. Mol. Biol.* **382,** 485–95 (2008).

23. Wang, X., Kumar, S. & Singh, S. K. Disulfide scrambling in IgG2 monoclonal antibodies : Insights from molecular dynamics simulations. *Pharm. Res.* **28,** 3128–3144 (2011).

24. Li, W., Zhang, J., Wang, J. & Wang, W. Metal-coupled folding of Cys 2 His 2 zinc-finger. *J. Am. Chem. Soc.* 892–900 (2008).

25. Pierce, M. M. & Nall, B. T. Fast folding of Cytochrome C. *Protein Sci.* **6,** 618–627 (1997).

26. Pierce, M. M. & Nall, B. T. Coupled kinetic traps in Cytochrome c folding : His-Heme misligation and proline isomerization. *J. Mol. Biol.* **298,** 955–969 (2000).

27. Kubelka, J., Hofrichter, J. & Eaton, W. A. The protein folding “speed limit.”*Curr. Opin. Struct. Biol.* **14,** 76–88 (2004).

28. Dill, K. a *et al.* Principles of protein folding--a perspective from simple exact models. *Protein Sci.* **4,** 561–602 (1995).

29. Levitt, M. Protein conformation, dynamics, and folding by computer simulation. *Annu. Rev. Biophysocs Bioeng.* 251–71 (1982).

30. Bystroff, C., Simonst, K. T., Han, K. F. & Baker, D. Local sequence-structure correlations in proteins. *Curr. Opin. Biotechnol.* **7,** 417–421 (1996).

31. Simons, K. T., Bonneau, R., Ruczinski, I. & Baker, D. Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins* **176,** 171–176 (1999).

32. Kuhlman, B. *et al.* Design of a novel globular protein fold with atomic-level accuracy. *Science* **302,** 1364–8 (2003).

33. Watters, A. L. *et al.* The highly cooperative folding of small naturally occurring proteins is likely the result of natural selection. *Cell* **128,** 613–24 (2007).

34. Huang, L. & Shakhnovich, E. I. Is there an en route folding intermediate for Cold shock proteins? *Protein Sci.* **21,** 677–85 (2012).

35. Zarrine-afsar, A. *et al.* Theoretical and experimental demonstration of the importance of specific nonnative interactions in protein folding. *Proc. Natl. Acad. Sci.* **105,** 9999–10004 (2008).

36. Lindorff-larsen, K., Piana, S., Dror, R. O. & Shaw, D. E. How fast-folding proteins fold. *Science (80-.).* **334,** 518–520 (2011).

37. Pande, V. S., Beauchamp, K. & Bowman, G. R. Everything you wanted to know about Markov State Models but were afraid to ask. *Methods* **52,** 99–105 (2010).

38. Kane, A. S. *et al.* Microfluidic mixers for the investigation of rapid protein folding kinetics using synchrotron radiation circular dichroism spectroscopy. *Anal. Chem.* **80,** 9534–9541 (2008).

39. Kathuria, S. V *et al.* Minireview: structural insights into early folding events using continuous-flow time-resolved small-angle X-ray scattering. *Biopolymers* **95,** 550–8 (2011).

40. Lambright, D. *et al.* Complementary techniques enhance the quality and scope of information obtained from SAXS. *ACA Trans.* 1–12 (2013).

41. Gambin, Y., Simonnet, C., VanDelinder, V., Deniz, A. & Groisman, A. Ultrafast microfluidic mixer with three-dimensional flow focusing for studies of biochemical kinetics. *Lab Chip* **10,** 598–609 (2010).

42. Kathuria, S. V *et al.* Advances in turbulent mixing techniques to study microsecond protein folding reactions. *Biopolymers* **99,** 888–96 (2013).

43. Shank, E. A., Cecconi, C., Dill, J. W., Marqusee, S. & Bustamante, C. The folding cooperativity of a protein is controlled by its chain topology. *Nature* **465,** 637–40 (2010).

44. CARRION-VAZQUEZ, M. *et al.* Mechanical and chemical unfolding of a single protein : A comparison. *Biophysics (Oxf).* **96,** 3694–3699 (1999).

45. Schuler, B. & Eaton, W. a. Protein folding studied by single-molecule FRET. *Curr. Opin. Struct. Biol.* **18,** 16–26 (2008).

46. Zhang, Y. & Lou, J. The Ca(2+) influence on calmodulin unfolding pathway: a steered molecular dynamics simulation study. *PLoS One* **7,** e49013 (2012).

47. Hinczewski, M., Gebhardt, J. C. M., Rief, M. & Thirumalai, D. From mechanical folding trajectories to intrinsic energy landscapes of biopolymers. *Proc. Natl. Acad. Sci.* **110,** 4500–4505 (2013).

48. Kim, P. S. & Baldwin, R. L. Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding. *Annu. Rev. Biochem.* **51,** 459–89 (1982).

49. Wetlaufer, D. B. Nucleation, Rapid Folding, and Globular Intrachain Regions in Proteins. *Proc. Natl. Acad. Sci.* **70,** 697–701 (1973).

50. Daggett, V. & Fersht, A. R. Is there a unifying mechanism for protein folding? *Trends Biochem. Sci.* **28,** 18–25 (2003).

51. Ptitsyn, O. B. & Rashin, A. A. A model of myoglobin self-organization. *Biophys. Chem.* **3,** 1–20 (1975).

52. Dill, K. A. & Sun, C. H. From Levinthal to pathways to funnels. *Nature* **4,** 10–19 (1997).

53. Smith, C. K., Withka, J. M. & Regan, L. A thermodynamic scale for the beta-sheet forming tendencies of the amino acids. *Biochemistry* **33,** 5510–7 (1994).

54. Jones, B. E. & Matthews, C. R. Early intermediates in the folding of dihydrofolate reductase from Escherichia coli detected by hydrogen exchange and NMR. *Protein Sci.* **4,** 167–77 (1995).

55. Radzicka, A. & Wolfenden, R. Comparing the Polarities of the Amino Acids: Side-Chain Distribution Coefficients between the Vapor Phase, Cyclohexane, 1 -0ctano1, and Neutral Aqueous Solutiont. *Biochemistry* **27,** 1664–1670 (1988).

56. Gu, Z., Zitzewitz, J. A. & Matthews, C. R. Mapping the structure of folding cores in TIM barrel proteins by hydrogen exchange mass spectrometry: the roles of motif and sequence for the indole-3-glycerol phosphate synthase from Sulfolobus solfataricus. *J. Mol. Biol.* **368,** 582–94 (2007).

57. Wu, Y., Vadrevu, R., Kathuria, S., Yang, X. & Matthews, C. R. A tightly packed hydrophobic cluster directs the formation of an off-pathway sub-millisecond folding intermediate in the alpha subunit of tryptophan synthase, a TIM barrel protein. *J. Mol. Biol.* **366,** 1624–38 (2007).

58. Fersht, a R. Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications. *Proc. Natl. Acad. Sci.* **92,** 10869–73 (1995).

59. Matthews, C. R., Crisanti, M. M., Manz, J. T. & Gepner, G. L. Effect of a single amino acid substitution on the folding of the alpha subunit of tryptophan synthase. *Biochemistry* **22,** 1445–52 (1983).

60. Beasty, A. M. *et al.* Effects of the Phenylalanine-22 Leucine, Glutamic Acid-49, Methionine,Glycine-234, Aspartic Acid, and Glycine-234 Lysine mutations on the folding and stability of the alpha subunit of Tryptophan Synthase from Escherichia coli. *Biochemistry* **25,** 2965–2974 (1986).

61. Fersht, A. R., Andreas, M. & Serrano, L. The folding of an enzyme. Theory of protein engineering analysis of stability and pathway protein folding. *J. Mol. Biol.* **224,** 771–782 (1992).

62. Burton, R. E., Huang, G. S., Daugherty, M. a, Fullbright, P. W. & Oas, T. G. Microsecond protein folding through a compact transition state. *J. Mol. Biol.* **263,** 311–22 (1996).

63. Kragelund, B. B. *et al.* Fast and one-step folding of closely and distantly related homologous proteins of a four-helix bundle family. *J. Mol. Biol.* **256,** 187–200 (1996).

64. Plaxco, K. W., Simons, K. T. & Baker, D. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277,** 985–94 (1998).

65. Viguera, A. R., Serrano, L. & Wilmanns, M. Different folding transition states may result in the same native structure. *Nature* **3,** 874–880 (1996).

66. Lindberg, M., Tångrot, J. & Oliveberg, M. Complete change of the protein folding transition state upon circular permutation. *Nat. Struct. Biol.* **9,** 818–22 (2002).

67. Ivarsson, Y., Travaglini-Allocatelli, C., Brunori, M. & Gianni, S. Engineered symmetric connectivity of secondary structure elements highlights malleability of protein folding pathways. *J. Am. Chem. Soc.* **131,** 11727–33 (2009).

68. Lindberg, M. O. *et al.* Folding of circular permutants with decreased contact order: general trend balanced by protein stability. *J. Mol. Biol.* **314,** 891–900 (2001).

69. Hills, Jr, R. D. in *Protein Dyn. Methids Protoc. Methods Mol. Biol.* (Livesay, D. R.) **1084,** 123 – 140 (Humana Press, 2014).

70. Koga, N. *et al.* Principles for designing ideal protein structures. *Nature* **491,** 222–7 (2012).

71. Bollen, Y. J. M., Kamphuis, M. B. & van Mierlo, C. P. M. The folding energy landscape of apoflavodoxin is rugged: hydrogen exchange reveals nonproductive misfolded intermediates. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 4095–100 (2006).

72. Nozaki, Y., Schechter, N. M., Reynolds, J. A. & Tanford, C. Use of gel chromatography for the determination of the Stokes radii of proteins in the presence and absence of detergents. A reexamination. *Biochemistry* **15,** 3884–90 (1976).

73. Roder, H., Maki, K. & Cheng, H. Early events in protein folding explored by rapid mixing methods. *Chem. Rev.* **106,** 1836–61 (2006).

74. Eaton, W. A., Thompson, P. A., Chan, C. K., Hage, S. J. & Hofrichter, J. Fast events in protein folding. *Structure* **4,** 1133–9 (1996).

75. Hofmann, H. *et al.* Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 16155–60 (2012).

76. Lindorff-Larsen, K., Trbovic, N., Maragakis, P., Piana, S. & Shaw, D. E. Structure and dynamics of an unfolded protein examined by molecular dynamics simulation. *J. Am. Chem. Soc.* **134,** 3787–91 (2012).

77. Fernandez-Recio, J., Genzor, C. G. & Sancho, J. Apoflavodoxin Folding Mechanism : An R / Protein with an Essentially Off-Pathway Intermediate. *Biochemistry* **40,** 15234–15245 (2001).

78. Onuchic, J. N. & Wolynes, P. G. Theory of protein folding. *Curr. Opin. Struct. Biol.* **14,** 70–5 (2004).

79. Wu, Y., Kondrashkina, E., Kayatekin, C., Matthews, C. R. & Bilsel, O. Microsecond acquisition of heterogeneous structure in the folding of a TIM barrel protein. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 13367–72 (2008).

80. Finke, J. M. & Onuchic, J. N. Equilibrium and kinetic folding pathways of a TIM barrel with a funneled energy landscape. *Biophys. J.* **89,** 488–505 (2005).

81. Nabuurs, S. M., Westphal, A. H. & van Mierlo, C. P. M. Extensive formation of off-pathway species during folding of an alpha-beta parallel protein is due to docking of (non)native structure elements in unfolded molecules. *J. Am. Chem. Soc.* **130,** 16914–20 (2008).

82. Borgia, M. B. *et al.* Single-molecule fluorescence reveals sequence-specific misfolding in multidomain proteins. *Nature* **474,** 662–5 (2011).

83. Plaxco, K. W., Simons, K. T., Ruczinski, I. & Baker, D. Topology , Stability , Sequence , and Length : Defining the Determinants of Two-State Protein Folding Kinetics. *Biochemistry* **39,** 11177 – 11183 (2000).

84. Hills, RD, J. & Brooks, CL, I. Insights from coarse-grained Gō models for protein folding and dynamics. *Int. J. Mol. Sci.* **10,** 889–905 (2009).

85. Pace, C. Determination and analysis of urea and guanidine hydrochloride denaturation curves. *Methods Enzymol.* **131,** 266–80 (1986).

86. Myers, J. K., Pace, C. N. & Scholtz, J. M. Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding. *Protein Sci.* **4,** 2138–48 (1995).

87. Kohn, J. E. *et al.* ranodom-coil behavior and the dimensions of chemically unfolded proteins. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 12691–12696 (2004).

88. Clementi, C., Jennings, P. A. & Onuchic, J. N. Prediction of folding mechanism for circular-permuted proteins. *J. Mol. Biol.* **311,** 879–90 (2001).

89. Kathuria, S. V *et al.* Microsecond Barrier-Limited Chain Collapse Observed by Time-Resolved FRET and SAXS. *J. Mol. Biol.* (2014). doi:10.1016/j.jmb.2014.02.020

90. Graceffa, R. *et al.* Sub-millisecond time-resolved SAXS using a continuous-flow mixer and X-ray micro-beam. *J. Synchrotron Radiat.* **20,** 1–6 (2013).

91. Karanicolas, J. & Brooks, III, C. L. The origins of asymmetry in the folding transition states of protein L and protein G. *Protein Sci.* **11,** 2351–2361 (2002).

92. Miyazawa, S. & Jernigan, R. L. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* **256,** 623–44 (1996).

93. Karanicolas, J. & Brooks, C. L. Improved Gō-like Models Demonstrate the Robustness of Protein Folding Mechanisms Towards Non-native Interactions. *J. Mol. Biol.* **334,** 309–325 (2003).

94. Brooks, B. R. *et al.* CHARRM: The Biomolecular Simulation Program. *J Comput Chem* **30,** 1545–1614 (2009).

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

96. Mizukami, T., Xu, M., Cheng, H., Roder, H. & Maki, K. Nonuniform chain collapse during early stages of staphylococcal nuclease folding detected by fluorescence resonance energy transfer and ultrarapid mixing methods. *Protein Sci.* **22,** 1336–48 (2013).

97. McCaldon, P. & Argos, P. Oligopeptide biases in protein sequences and their use in predicting protein coding regions in nucleotide sequences. *Proteins* **4,** 99–122 (1988).

98. Martins, D. & English, A. M. SOD1 oxidation and formation of soluble aggregates in yeast: Relevance to sporadic ALS development. *Redox Biol.* **2,** 632–9 (2014).

99. Knowles, T. P. J., Vendruscolo, M. & Dobson, C. M. The amyloid state and its association with protein misfolding diseases. *Nat. Rev. Mol. Cell Biol.* **15,** 384–96 (2014).

100. Zhang, Z. & Chan, H. S. Native topology of the designed protein Top7 is not conducive to cooperative folding. *Biophys. J.* **96,** L25–7 (2009).

101. Xu, D. & Nussinov, R. Favorable domain size in proteins. *Fold. Des.* **3,** 11–7 (1998).

102. Packer, L. E., Song, B., Raleigh, D. P. & McKnight, C. J. Competition between intradomain and interdomain interactions: A buried salt bridge is essential for Villin Headpiece folding and actin-binding. *Biochemistry* **50,** 3706–3712 (2012).

103. Nobrega, R. P. *et al.* Modulation of frustration in folding by sequence permutation. *Proc. Natl. Acad. Sci.* 2–7 (2014). doi:10.1073/pnas.1324230111

104. Smith, J. S. & Scholtz, J. M. Guanidine hydrochloride unfolding of peptide helices: separation of denaturant and salt effects. *Biochemistry* **35,** 7292–7 (1996).

105. Zhang, Y.-Z. *Protein and peptide structure and interactions studied by hydrogen exchange and NMR*.

106. Krishna, M. M. G., Hoang, L., Lin, Y. & Englander, S. W. Hydrogen exchange methods to study protein folding. *Methods* **34,** 51–64 (2004).

107. Costantini, S., Colonna, G. & Facchiano, A. M. Bioinformation ESBRI : A web server for evaluating salt bridges in proteins bioinformation. *Bioinformation* **3,** 137–138 (2008).

108. Bhuyan, A. K. Protein stabilization by urea and guanidine hydrochloride. *Biochemistry* **41,** 13386–94 (2002).

109. Dunbar, J., Yennawar, H. P., Banerjee, S., Luo, J. & Farber, G. K. The effect of denaturants on protein structure. *Protein Sci.* **6,** 1727–33 (1997).

110. Jha, S. K. & Marqusee, S. Kinetic evidence for a two-stage mechanism of protein denaturation by guanidinium chloride. *Proc. Natl. Acad. Sci.* **111,** 4856–61 (2014).

111. Lee, R. Van Der *et al.* Classification of intrinsically disordered regions and proteins. *Chem. Rev.* **114,** 6589–6631 (2014).

112. Das, R. K. & Pappu, R. V. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc. Natl. Acad. Sci.* **110,** 13392–13397 (2013).

113. Meng, W., Lyle, N., Luan, B., Raleigh, D. P. & Pappu, R. V. Experiments and simulations show how long-range contacts can form in expanded unfolded proteins with negligible secondary structure. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 2123–8 (2013).

114. Marsh, J. a & Forman-Kay, J. D. Sequence determinants of compaction in intrinsically disordered proteins. *Biophys. J.* **98,** 2383–90 (2010).

115. Degrado, W. F., Kezdy, F. J. & Kaiser, E. T. Design, Synthesis and Characterization of a Cytotoxic Peptide with Melittin-Like Activity. *J. Am. Chem. Soc.* **103,** 679–681 (1981).

116. Delaglio, F. *et al.* NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6,** 277–93 (1995).

117. Kathuria, S. V *et al.* Microsecond barrier-limited chain collapse observed by time-resolved FRET and SAXS. *J. Mol. Biol.* **426,** 1980–94 (2014).

118. Levitt, M. & Warshel, A. Computer simulation of protein folding. *Nature* **253,** 694–698 (1975).

119. Voelz, V. a, Bowman, G. R., Beauchamp, K. & Pande, V. S. Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1-39). *J. Am. Chem. Soc.* **132,** 1526–8 (2010).

120. Lindorff-Larsen, K., Piana, S., Dror, R. O. & Shaw, D. E. How fast-folding proteins fold. *Science* **334,** 517–20 (2011).

121. Gruebele, M., Sabelko, J., Ballew, R. & Ervin, J. Laser temperature jump induced protein refolding. *Acc. Chem. Res.* **31,** 699–707 (1998).

122. Prigozhin, M. B. *et al.* Misplaced helix slows down ultrafast pressure-jump protein folding. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 8087–92 (2013).

123. Dill, K. a & Shortle, D. Denatured states of proteins. *Annu. Rev. Biochem.* **60,** 795–825 (1991).

124. Klein-Seetharaman, J. *et al.* Long-range interactions within a nonnative protein. *Science* **295,** 1719–22 (2002).

125. Roder, H., Maki, K., Cheng, H. & Shastry, M. C. R. Rapid mixing methods for exploring the kinetics of protein folding. *Methods* **34,** 15–27 (2004).

126. Knight, J., Vishwanath, A., Brody, J. & Austin, R. Hydrodynamic focusing on a silicon chip: mixing nanoliters in microseconds. *Phys. Rev. Lett.* **80,** 3863–3866 (1998).

127. Lapidus, L. J. *et al.* Protein hydrophobic collapse and early folding steps observed in a microfluidic mixer. *Biophys. J.* **93,** 218–24 (2007).

128. Waldauer, S. a, Wu, L., Yao, S., Bakajin, O. & Lapidus, L. J. Microfluidic mixers for studying protein folding. *J. Vis. Exp.* 1–9 (2012). doi:10.3791/3976

129. Putnam, C. D., Hammel, M., Hura, G. L. & Tainer, J. a. X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. *Q. Rev. Biophys.* **40,** 191–285 (2007).

130. Meisburger, S. P. *et al.* Breaking the radiation damage limit with Cryo-SAXS. *Biophys. J.* **104,** 227–36 (2013).

131. Eng, P. J., Newville, M., Rivers, M. L. & Sutton, S. R. X-Ray Micro-Focusing Optics. *SPIE* **3449,** 145–156 (1998).

132. Grant, T. D. *et al.* Small angle X-ray scattering as a complementary tool for high-throughput structural studies. *Biopolymers* **95,** 517–30 (2011).

133. Mathew-Fenn, R. S., Das, R. & Harbury, P. a B. Remeasuring the double helix. *Science* **322,** 446–9 (2008).

134. Mastroianni, A. J., Sivak, D. a, Geissler, P. L. & Alivisatos, a P. Probing the conformational distributions of subpersistence length DNA. *Biophys. J.* **97,** 1408–17 (2009).

135. Schuler, B., Lipman, E. A. & Eaton, W. A. Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy. *Nature* **419,** 743–748 (2002).

136. Fazelinia, H., Xu, M., Cheng, H. & Roder, H. Ultrafast hydrogen exchange reveals specific structural events during the initial stages of folding of cytochrome c. *J. Am. Chem. Soc.* **136,** 733–40 (2014).

137. Hambly, D. M. & Gross, M. L. Laser flash photolysis of hydrogen peroxide to oxidize protein solvent-accessible residues on the microsecond timescale. *J. Am. Soc. Mass Spectrom.* **16,** 2057–63 (2005).

138. Gau, B. C., Sharp, J. S., Rempel, D. L. & Gross, M. L. Fast photochemical oxidation of protein footprints faster than protein unfolding. *Anal. Chem.* **81,** 6563–71 (2009).

139. Walters, B. T., Ricciuti, A., Mayne, L. & Englander, S. W. Minimizing back exchange in the hydrogen exchange-mass spectrometry experiment. *J. Am. Soc. Mass Spectrom.* **23,** 2132–9 (2012).

140. Vahidi, S., Stocks, B. B., Liaghati-Mobarhan, Y. & Konermann, L. Submillisecond protein folding events monitored by rapid mixing and mass spectrometry-based oxidative labeling. *Anal. Chem.* **85,** 8618–25 (2013).

141. Wu, L. & Lapidus, L. J. Combining ultrarapid mixing with photochemical oxidation to probe protein folding. *Anal. Chem.* **85,** 4920–4 (2013).

142. Maleknia, S. D., Brenowitz, M. & Chance, M. R. Millisecond radiolytic modification of peptides by synchrotron X-rays identified by mass spectrometry. *Anal. Chem.* **71,** 3965–3973 (1999).

143. Petruk, A. A., Defelipe, L. A., Marti, M. A. & Turjanski, A. G. Molecular Dynamics simulations provide atomistic insight into hydrogen exchange Mass Spectrometry experiments. *J. Chem. Theory Comput.* **9,** 658–669 (2013).

144. Bowman, G. R., Huang, X. & Pande, V. S. Using generalized ensemble simulations and Markov state models to identify conformational states. *Methods* **49,** 197–201 (2009).

145. Beauchamp, K. a *et al.* MSMBuilder2: Modeling conformational dynamics at the picosecond to millisecond scale. *J. Chem. Theory Comput.* **7,** 3412–3419 (2011).

146. Faller, C. E., Reilly, K. a, Hills, R. D. & Guvench, O. Peptide backbone sampling convergence with the adaptive biasing force algorithm. *J. Phys. Chem. B* **117,** 518–26 (2013).

147. Jagannathan, B. & Marqusee, S. Protein folding and unfolding under force. *Biopolymers* **99,** 860–9 (2013).