Sequence Determinants of the Folding Free-Energy Landscape of beta alpha-Repeat Proteins: A Dissertation

SEQUENCE DETERMINANTS OF THE FOLDING FREE-ENERGY LANDSCAPE OF βα-REPEAT PROTEINS

A Dissertation Presented

By

SAGAR VIRENDRA KATHURIA

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

June 16th 2010

Biochemistry and Molecular Pharmacology

SEQUENCE DETERMINANTS OF THE FOLDING FREE-ENERGY LANDSCAPE OF βα-REPEAT PROTEINS

A Dissertation Presented
By

SAGAR VIRENDRA KATHURIA

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

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

Daniel N. Bolon. Ph.D., Member of Committee

Lawrence J. Stern, Ph.D., Member of Committee

Heinrich Roder, 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.

William E. 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

Biochemistry and Molecular Pharmacology Program
June 16th , 2010

Dedication

To my family, for their unconditional love and support.

Acknowledgement

There are many who have made this thesis possible. Foremost among them is Dr. Matthews, for whose guidance and encouragement I am forever thankful. My thesis research advisory committee, Dr. William E. Royer, Dr. Celia Schiffer, Dr. Mary Munson, Dr. Lawrence J. Stern and Mr. James Evans, have helped me progress smoothly with my research. I would like to thank them and the other members of my thesis defense committee, Dr. Daniel N. Bolon and Dr. Heinrich Roder for their valuable inputs that have improved the quality of the work presented here.

I am grateful to Dr. Osman Bilsel, Dr. Jill Ann Zitzewitz and Dr. Louise A. Wallace for their guidance, practical training and assistance in writing the various manuscripts contained in this thesis. I have had several fruitful collaborations within the lab with Dr. Iain J. Day, Dr. Ramakrishna Vadrevu and Dr. Xiaoyan Yang, Dr. Zhenyu Gu, Ms. Komali Betha, Mr. Shaun Lavallee, Mr. Paul Nobrega and Ms. Ornella Bisceglia, and I am very grateful for their contributions. I am also fortunate to have shared the work space with some very talented, hardworking and brilliant people, Can, Anna, Ying, Agnita, Amanda, Ganga, Divya, Lori and Noah, and I would like to thank them all for enriching my experiences in the Matthews lab.

A special thanks to Dr. Charles Brooks III and Dr. Ronald D. Hills Jr. for their collaboration on the CheY project. Their simulations on CheY-like proteins were instrumental in elucidating the folding mechanism of these proteins. Their insights were indispensable in interpreting my experimental results.

I am greatly indebted to several members of the Schiffer lab, Aysegul, Seema, Keith, Rajintha and Madhavi Nallam, who have been great friends to me during my time at UMMS. I have also had the opportunity to interact with and learn from many members of the Biochemistry department and I would like to thank them all for shaping the course of my research; James Knapp, Nancy King, Hitesh, Jennifer Foulkes Murzycki, Moses Prabu, Yufeng Cai, Balaji Bhyravbhatla, John Gross, Barbara Evans, Karin Green, Ellen Nalivaika, Elizabeth Mandon, Nicholas Willis, Melonnie Fergusson, Steven Trueman, Shiven Shandilya, to name a few.

I bow to all the members of the Biochemistry department whose untiring efforts help maintain the excellent training environment that I have come to take for granted, Alan Lucia, Karen Logan, Karen Welch, Christine Pruitte, Maria and Irene and others. There are many who have mean much to me and have in some way or form given shape to my career, to all those unmentioned names I owe a deep gratitude.

No words can suffice, nor can any deed repay the sacrifices and the unconditional love and support of my family, my parents, my sister and my wife Madhavi, for whom I am and will forever be grateful.

Abstract

The most common structural platform in biology, the $\beta\alpha$-repeat classes of proteins, are represented by the $(\beta\alpha)_8$ TIM barrel topology and the $\alpha/\beta/\alpha$ sandwich, CheY-like topology. Previous studies on the folding mechanisms of several members of these proteins have suggested that the initial event during refolding involves the formation of a kinetically trapped species that at least partially unfolds before the native conformation can be accessed. The simple topologies of these proteins are thought to permit access to locally folded regions that may coalesce in non-native ways to form stable interactions leading to misfolded intermediates. In a pair of TIM barrel proteins, αTS and sIGPS, it has been shown that the core of the off-pathway folding intermediates is comprised of locally connected clusters of isoleucine, leucine and valine (ILV) residues. These clusters of Branched Aliphatic Side Chains (BASiC) have the unique ability to very effectively prevent the penetration of water to the underlying hydrogen bond networks. This property retards hydrogen exchange with solvent, strengthening main chain hydrogen bonds and linking tertiary and secondary structure in a cooperative network of interactions. This property would also promote the rapid formation of collapsed species during refolding. From this viewpoint, the locally connected topology and the appropriate distribution of ILV residues in the sequence can modulate the energy landscapes of TIM barrel proteins. Another sequence determinant of protein stability that can significantly alter the structure and stability of TIM barrels is the long-range main chain-side chain hydrogen bond. Three of these interactions have been shown to form the molecular underpinnings for the cooperative access to the native state in αTS.

Global analysis results presented in Chapter II and Chapter III, suggest that the off-pathway mechanism is common to three proteins of the CheY-like topology, namely CheY, NT-NtrC and Spo0F. These results are corroborated by Gō-simulations that are able to identify the minimal structure of kinetically trapped species during the refolding of CheY and Spo0F. The extent of transient, premature structure appears to correlate with the number of ILV side chains involved in a large sequence-local cluster that is formed between the central β-sheet and helices α2, α3 and α4. The failure of Gō-simulations to detect off-pathway species during the refolding of NT-NtrC may reflect the smaller number of ILV side chains in its corresponding hydrophobic cluster.

In Chapter IV, comparison of the location of large ILV clusters with the hydrogen exchange protected regions in 19 proteins, suggest that clusters of BASiC residues are the primarily determinants of the stability cores of globular proteins. Although the location of the ILV clusters is sufficient to determine a majority of the protected amides in a protein structure, the extent of protection is over predicted by the ILV cluster method.

The survey of 71 TIM barrel proteins presented in Chapter V, suggests that a specific type of long-range main chain-side chain hydrogen bond, termed "βα hairpin clamp" is a common feature in the βα-repeat proteins. The location and sequence patterns observed demonstrate an evolutionary signature of the βαβ modules that are the building blocks of several βα-repeat protein families.

In summary, the work presented in this thesis recognizes the role of sequence in modulating the folding free energy landscapes of proteins. The formation of off-pathway folding intermediates in three CheY-like proteins and the differences in the proposed

extent of structure formed in off-pathway intermediates of these three proteins, suggest that both topology and sequence play important and concerted roles in the folding of proteins. Locally connected ILV can clusters lead to off-pathway traps, whereas the formation of the productive folding path requires the development of long-range native-like topological features to form the native state. The ability of ILV clusters to link secondary and tertiary structure formation enables them to be at the core of this cooperative folding process. Very good correlations between the locations of ILV clusters and both strong protection against exchange and the positions of folding nuclei for a variety of proteins reported in the literature support the generality of the BASiC hypothesis. Finally, the discovery of a novel pattern of H-bond interactions in the TIM barrel architecture, between the amide hydrogen of a core ILV residue with a polar side chain, bracketing βαβ modules, suggests a means for establishing cooperativity between different types of side chain interactions towards formation of the native structure.

Table of Contents

List of Tables

List of Figures

List of Third Party Copyrighted Materials

Figure 1.2 Reprinted from Trends in Biochemical Sciences, Sheena E. Radford, Protein folding: progress made and promises ahead, Vol 25.12, Pages 611-618, 1 December 2000. With permission from Elsevier, License number 2486680767271, Aug 12 2010. http://www.sciencedirect.com/science/journal/09680004

Figure 1.4 Panels a) and b) Reprinted from Protein Science, Andrei Y. Istomin, Donald J. Jacobs and Dennis R. Livesay, *"On the role of structural class of a protein with two-state folding kinetics in determining correlations between its size, topology, and folding rate"*; Vol 16.11 Pages 2564-2569 2007. With permission from John Wiley and Sons, License number 2486681394942, Aug 12 2010. www.interscience.wiley.com

Figure 1.4 Panels c) and d) Reprinted from Journal of Molecular Biology, Kiyoto Kamagata, Munehito Arai, Kunihiro Kuwajima, *"Unification of the Folding Mechanisms of Non-two-state and Two-state Proteins"*; Vol 339.4, Pages 951-965, 11 June 2004. With permission from Elsevier, License number 2486690421147, Aug 12 2010. http://www.sciencedirect.com/science/journal/00222836

Figure 1.4 Panel e) Reprinted from Journal of Molecular Biology, Kiyoto Kamagata, Kunihiro Kuwajima, *"Surprisingly High Correlation between Early and Late Stages in Non-two-state Protein Folding"*; Vol 357.5, Pages 1647-1654, 14 April 2006. With permission from Elsevier, License number 2486690637110, Aug 12 2010. http://www.sciencedirect.com/science/journal/00222836

Chapter II Reprinted from Journal of Molecular Biology, Kathuria SV, Day IJ, Wallace LA, Matthews CR, *"Kinetic traps in the folding of $\beta\alpha$-repeat proteins: CheY* *initially misfolds before accessing the native conformation.”*; Vol 382.2, Pages 467-84, 3 Oct 2008. With permission from Elsevier, License number 2486690802731, Aug 12 2010. http://www.sciencedirect.com/science/journal/00222836

Chapter III Reprinted from Journal of Molecular Biology, *Hills RD Jr, Kathuria SV, Wallace LA, Day IJ, Brooks CL 3rd, Matthews CR. “Topological frustration in $\beta\alpha$-repeat proteins: sequence diversity modulates the conserved folding mechanisms of $\alpha/\beta/\alpha$ sandwich proteins.”*; Vol 398.2, Pages 332-50, Apr 30 2010. With permission from Elsevier, License number 2486691036319, Aug 12 2010. http://www.sciencedirect.com/science/journal/00222836

Chapter V Reprinted from PLoS One, *Yang X, Kathuria SV, Vadrevu R, Matthews CR. “$\beta\alpha$-hairpin clamps brace $\beta\alpha\beta$ modules and can make substantive contributions to the stability of TIM barrel proteins.”*; Vol 4.9:e7179, 29 Sep 2009. No permission required.

List of Abbreviations

1. ACO - Absolute Contact Order
2. αTS - alpha subunit of Tryptophan Synthase
3. BASiC - Branched Aliphatic Side Chains
4. BPTI - Bovine Panctreatic Trypsin Inhibitor
5. BSA - Buried Surface Area
6. CD - Circular Dichroism
7. CheY - Chemotaxis protein Y from *E. coli*
8. CI-2 - Chymotrypsin Inhibitor
9. CO - Contact Order
10. COREX - Correlation with hydrogen exchange protection factors
11. CSU - Contacts of Structural Units
12. DHFR - Dihydrofolate Reductase
13. eIGPS - Indole-3-Glycerol Phosphate Synthase from *E.coli*
14. FIRST - Floppy Inclusions and Rigid Substructure Topography
15. FL - Fluoroscence
16. FRET - Forster Resonance Energy Transfer
17. GNM - Gausian Network Model
18. HX - Hydrogen Exchange
19. IBP - Burst Phase Intermediate
20. IC - Intermediate with cis-proline
21. IGPS - Indole-3-Glycerol Phosphate Synthase

22. ILV - Isoleucine Leucine and Valine residues
23. Ioff - Off-pathway intermediate
24. IOLI - 278-residue TIM barrel protein of unknown function encoded by *the Bacillus subtilis iolI* gene
25. Ion - On-pathway intermediate
26. IT - Intermediate with trans-proline
27. LRO - Long Range Order
28. MC - Main Chain
29. N - Native State
30. Nc - Number of non-local contact clusters
31. NC - Native State with cis-proline
32. NMR - Nuclear Magnetic Resonance
33. NT - Native State with trans-proline
34. NT-NtrC - Amino-terminal reciever domain of the Nitrogen fixing protein NtrC from *Salmonella typhimurium*
35. pWT-CheY - pseudo Wild Type Chemotaxis protein Y
36. Qd - Number of sequence distant native pairs
37. RCO - Relative Contact Order
38. RNaseH - Ribonuclease H
39. SAB - Surface Area Buried
40. SAXS - Small Angle X-ray Scattering
41. SC - Side Chain

42. sIGPS - Indole-3-Glycerol Phosphate Synthase from *S. solfataricus*

43. Snase - Staphylococcal Nuclease

44. Spo0F - Sporulation protein F from *Bacillus subtilis*

45. TIM - Triose Phosphate Isomerase

46. TSE - Transition State Ensemble

47. U - Unfolded State

48. UC - Unfolded State with cis-proline

49. UT - Unfolded State with trans-proline

50. vdW - van der Waals

List of Electronic Files

Copies of the source code for the global analysis program and the cluster analysis program are provided in the enclosed disk.

Global analysis software

Project:

1. Global_Model.vbp
2. Global_Model.vbw

Forms:

1. AllowedChangeform.frm
2. Amplitude_Form.frm
3. Concentration.frm
4. Constant.frm
5. Dialog.frm
6. File_fitting.frm
7. Get_BP_param.frm
8. Get_List_amplitude.frm
9. Getrate.frm
10. Global_Form.frm
11. Linear.frm
12. Local_Form.frm
13. MDIForm1.frm
14. Offset.frm

15. Species_Signal.frm
16. file_fitting_linked.frm
17. fitting_form.frm
18. fitting_form_linked.frm
19. frmAbout.frm
20. get_node_name_connections.frm
21. getnumnodes.frm
22. getparam.frm
23. monitorprog.frm

Modules:

1. File_writer.bas
2. LM_linked.bas
3. LM_routine.bas
4. array_loader.bas
5. data_manipulate.bas
6. equilibrium_fit.bas
7. file_loader.bas
8. form_loader.bas
9. nlrcs.bas
10. variable_definition.bas

Cluster analysis software

Project:

1. CCSS.vbp
2. CCSS.vbw

Forms:

1. 2D.frm
2. ATOMType.frm
3. CCSSMain.frm
4. ClampClusterQuery.frm
5. Contactorder.frm
6. ControlFrm.frm
7. Locator.frm
8. PDBadd.frm

Modules:

1. GeneralDeclare.bas
2. BatchFileExport.bas
3. SCOPimport.bas
4. clustermap.bas
5. contactcalc1.bas
6. outputimporter.bas
7. pdbfileimport.bas
8. querybuilder.bas

Preface

The work presented in Chapter II has been published previously as *Kathuria SV, Day IJ, Wallace LA, Matthews CR. "Kinetic traps in the folding of βα-repeat proteins: CheY initially misfolds before accessing the native conformation." J Mol Biol. 2008 Oct 3;382(2):467-84.* This paper was the result of a collaborative effort. Dr. Louise A. Wallace carried out the equilibrium unfolding experiments, the unfolding and refolding kinetic experiments and the experiments on cyclophilin dependence of refolding. Dr. Iain J. Day carried out the NMR experiments. I repeated the equilibrium and kinetic experiments, performed the double jump refolding experiments, determined the temperature dependence of the refolding kinetics and performed the global analysis. Dr. C. Robert Matthews and I wrote the manuscript.

The work presented in Chapter III has been published previously as *Hills RD Jr, Kathuria SV, Wallace LA, Day IJ, Brooks CL 3rd, Matthews CR. "Topological frustration in βα-repeat proteins: sequence diversity modulates the conserved folding mechanisms of α/β/α sandwich proteins." J Mol Biol. 2010 Apr 30;398(2):332-50.* This paper was the result of collaboration between Dr. C.R. Matthews' group and Dr. C.L. Brooks' group. Dr. Ronald D. Hills Jr. performed and analyzed the Gō-simulations, Dr Louise A. Wallace carried out the equilibrium unfolding experiments, the unfolding and refolding kinetic experiments and the experiments on cyclophilin dependence of refolding for both NT-NtrC and Spo0F. Dr. Iain J. Day carried out the NMR experiments. I repeated the equilibrium and kinetic experiments on Spo0F and performed the global analysis on both NT-NtrC and Spo0F. Dr. C. Robert Matthews, Dr. Charles L. Brooks III, Dr. Ronald D. Hills Jr. and I wrote the manuscript.

The work presented in Chapter IV is being prepared for publication in JMB as *Kathuria SV, Matthews CR. "Clusters of Isoleucine, Leucine and Valine Side Chains Define Cores of Stability in Globular Proteins: Sequence Determinants of Structure, Stability and Folding."* The manuscript was conceived by Dr. C. Robert Matthews. I developed and analyzed the database. Dr. C Robert Matthews and I wrote the manuscript.

The work presented in Chapter V has been published previously as *Yang X, Kathuria SV, Vadrevu R, Matthews CR. "$\beta\alpha$-hairpin clamps brace $\beta\alpha\beta$ modules and can make substantive contributions to the stability of TIM barrel proteins." PLoS One. 2009 Sep 29;4(9):e7179.* This paper was the result of a collaborative effort. Dr. Xiaoyan Yang, Dr. Ramakrishna Vadrevu and Dr. C. Robert Matthews conceived and designed the experiments. Dr. Xiaoyan Yang performed the experiments. Dr. Xiaoyan Yang, analyzed the experimental data. I developed and analyzed the database. Dr. Xiaoyan Yang, Dr. Ramakrishna Vadrevu, Dr. C Robert Matthews and I wrote the manuscript.

The Equilibrium denaturation curves and the manual mixing refolding kinetics of pWT-CheY monitored by circular dichroism, shown in Chapter VI were collected and analyzed by Mr. R. Paul Nobrega. The section on the correlation of phi-analysis of CI-2 and ILV clusters was originally developed by Dr. C. Robert Matthews as a part of a manuscript in preparation.

Other published work: *Wu Y, Vadrevu R, Kathuria S, Yang X, Matthews CR. "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. 2007 Mar 9;366(5):1624-38*. Dr. Ying Wu, Dr. Ramakrishna Vadrevu and Dr. Xiaoyan Yang performed the experiments. I developed the algorithm for the analysis and categorization of hydrophobic clusters. Dr. Ying Wu, Dr. Ramakrishna Vadrevu and Dr. C. Robert Matthews wrote the manuscript.

Chapter I – Introduction

Protein folding

The mechanism by which an unstructured polypeptide chain finds its unique folded, native state has been studied by biophysicists for more than half a century. Early work in the field established that the functional native structure of a globular protein is its most stable form and that the information required to reach this native structure is encoded in the amino acid sequence of the protein.[1] Given the number of possible conformations that a small polypeptide chain can adopt, it became immediately apparent that the transition of unfolded chain to native structure is not a random event,[2] but occurs via a coordinated mechanism that takes place on a biologically relevant timescale. It follows that evolution selects for amino acid sequences that favor the formation of the native state rapidly and efficiently with minimally "frustrated" mechanisms.[3] In reality though, protein folding does not occur over a smooth free-energy landscape, as folding rate is only one of the determinants in the evolution of proteins. Sequences are designed to maximize several parameters, including protein function, native state stability, rapid access to the native state, and their ability to absorb mutations[4,5].

Relevance

Since the three-dimensional, functional form of a protein is a consequence of its sequence,[1] it should be possible to develop a generalized protein folding code, which can not only predict the native structure of any sequence, but also help in developing sequences for designer proteins. The development of a generalized protein folding code will help to convert the enormous sequence databases available to systems biologists into structural information. This information can add several layers of understanding to the complex networks of protein interactions in biological systems.[6] Further, an understanding of the sequence determinants of stability and dynamics of the native state can enhance the development of novel proteins for the biotechnology industry.

From a practical perspective, these problems have been approached using knowledge-based potentials, i.e. using the vast library of structures as models for the design of new proteins[7,8] and homology modeling to predict structures of newly sequenced genes.[9 11] These structures are then refined using physics-based potentials. With the development of fast screening methods[12] and the use of nonbiological materials,[13] the design of new proteins for practical applications is no longer limited by the understanding of the protein folding code. However, the rational design of proteins will certainly benefit from the development of such a code.[14,15]

Further, structure prediction in cases where no homologues exist is still a considerable challenge. The ultimate goal, *ab initio* fold prediction that uses only physics based models,[16,17] is extremely expensive in computer time and resources. For example, the 36mer villin headpiece that folds in the micro second timescale was computationally folded using 1000 cpu years.[16] Other techniques assume a hierarchical organization of protein structures and incorporate knowledge-based and physics-based potentials to define small structural units that are then used to reconstruct the larger whole.[18,19]

Insights into folding pathways from *ab initio* simulations and by experimental means will enhance our ability to efficiently and correctly predict protein structures.

The way proteins interact with their environments in the cell, their interactions with other proteins, and their propensities to misfold and aggregate are all closely related to their folding free energy landscapes. An understanding of protein folding and mechanisms can help identify the misfolding/aggregation competent species and design drugs or strategies for the alleviation of disease by altering the kinetic or thermodynamic access to these species in the cell.[20 22]

From an academic stand point, several interrelated questions are hotly debated despite the wealth of information available. Do the vast numbers of unfolded conformations fold to the native state via multiple routes? Is there a specific order of events in the folding of proteins? Are there biases in the unfolded state that predetermine the folding pathways? Are there defined folding cores that must be formed before the protein can reach the native state? What is the effect of native state complexity on folding free energy landscapes? How do mutations modulate folding free energy landscapes?

Folding free energy landscapes

Equilibrium denaturation of several proteins by chemical or thermal means revealed that protein folding is a cooperative process. Further, folding and unfolding kinetics are well described by simple exponential curves. These observations suggest that indeed only a handful of free energy states are populated during protein folding. A graphical representation of these free energy states is shown in Figure 1.1, where the reaction coordinate represents any parameter that measures "foldedness" of a given state and the height represents the free energy of that state. Thus the unfolded polypeptide chain (*U*) can fold to the native functional form (*N*), either directly (two-state folders) or via states that are intermediate (*I*) between the two end states (multi-state folders). Favoring the folding reaction are several enthalpic contributions, including hydrogen bonds, van der Waals (vdW) interactions, electrostatic interactions and a gain in solvent entropy. However, accumulation of structure along the reaction coordinate results in the reduction of conformational states available to the polypeptide chain. The set of structures where the conformational search for the native topology is solved is called the transition state ensemble (TSE or $I^{\ddagger}$). The structures within this ensemble with the lowest free energy determine the rate at which the extended polymer chains can fold to the native state.

Structurally, the conformational search for the native state has been interpreted in several ways (Fig. 1.2).[23] 1) In the framework model, local secondary structural elements form rapidly and eventually coalesce by simple diffusion collision, e.g. BPTI.[24,25] 2) In the nucleation-condensation model, local structural elements form sites of nucleation. Further structure develops from these nucleation sites in a hierarchical manner, combining tertiary interactions with secondary structure development, e.g. CI-2.[26] 3) In the hydrophobic collapse model, hydrophobic amino acid side chains drive the collapse of the polypeptide chain, thereby reducing the conformational search. The native

Figure 1.1 – Traditional representation of the folding free energy landscape. The reaction coordinate represents any parameter that measures the foldedness of a protein. The free energy difference between the native (N) and the unfolded (U) states represents the stability of the protein. The energetic barrier that forms the rate limiting step in the folding of the protein is the transition state ensemble ($I^{\ddagger}$ or TSE). **a)** Reaction coordinate of a two-state folder. Only the native and unfolded states are populated. The transition from U to N occurs is an all-or-none phenomenon. **b)** Reaction coordinate of a three-state folder. An additional obligate intermediate (I) is formed enroute to the native state. Such intermediates may help reduce the conformational space along the reaction coordinate. **c)** Reaction coordinate with off-pathway folding intermediate (I). While off-pathway intermediates are kinetically accessible to the unfolded state, they are trapped species whose structures are not productive and must be undone to form the native state.

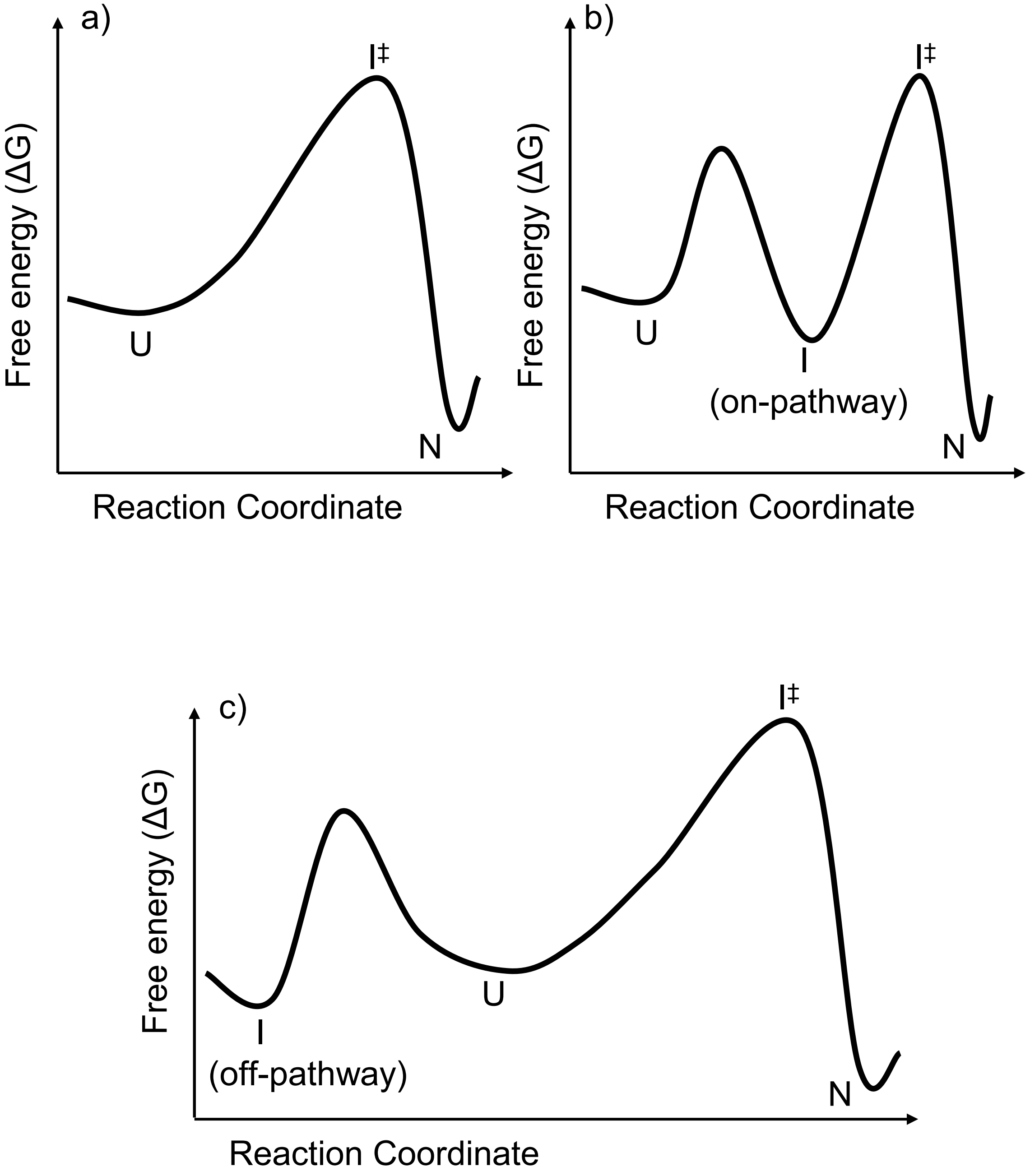

a)
I‡
Free energy (ΔG)
U
N
Reaction Coordinate
b)
I‡
Free energy (ΔG)
U
I
(on-pathway)
N
Reaction Coordinate
c)
I‡
Free energy (ΔG)
U
I
(off-pathway)
N
Reaction Coordinate

Figure 1.1

Figure 1.2 – Early models for mechanisms of folding. *
Structural interpretation of the folding models. Extreme cases are represented a) hydrophobic collapse occurs prior to secondary structure formation. b) Local secondary structures form rapidly and coallesce into the native structure by diffusion collision. c) Nucleation-condensation, where weak local structural elements can form nucleation sites for further secondary and tertiary structure formation in a coordinated manner. d) Each protein molecule can follow its own path to the native state, by any of the methods or by combinations of methods.

*Reprinted from Trends in Biochemical Sciences, Vol 25.12, Sheena E. Radford, Protein folding: progress made and promises ahead, Pages 611-618, 1 December 2000, with permission from Elsevier (License number 2486680767271, Aug 12 2010). http://www.sciencedirect.com/science/journal/09680004

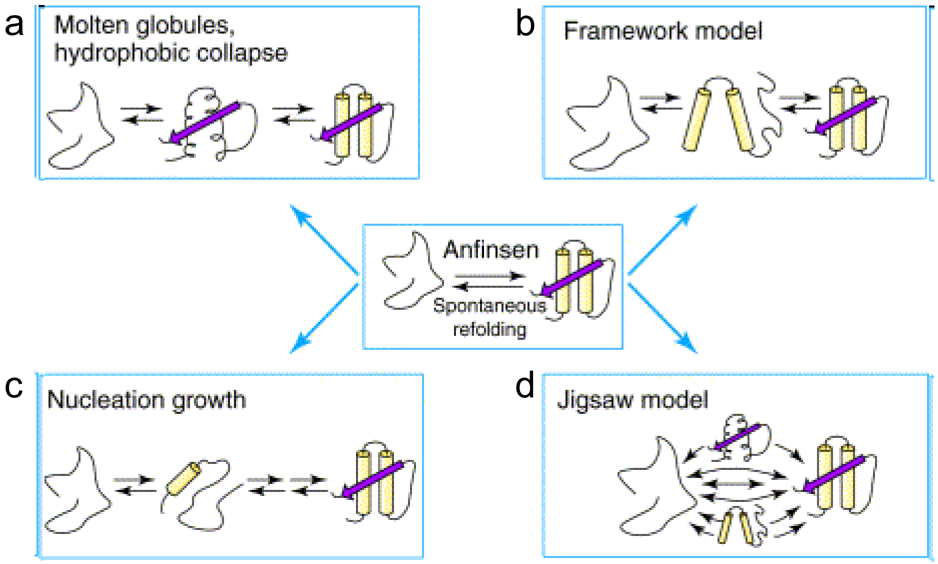

a
Molten globules,
hydrophobic collapse
b
Framework model
Anfinsen
Spontaneous
refolding
c
Nucleation growth
d
Jigsaw model

Figure 1.2

structure can then evolve from this intermediate by local rearrangements of tertiary interactions and formation of secondary structures, e.g. Barstar[27] and α-lactalbumin.[28] Variations on these basic models have been proposed to explain the range of experimental data within the context of these models.[29 31] However, it is highly unlikely that all proteins follow any one method of folding.[24]

An alternative three dimensional representation of the free energy landscape of protein folding was proposed by Dill and Wolynes,[3,32] wherein a multitude of folding paths are available to the unfolded ensemble (Fig. 1.3). Enthalpic (y-axis) and entropic (x-axis) contributions determine the accessibility of any given path. Intermediates are represented as energetic frustrations along the folding path, and the transition state ensemble is the saddle-point at which the conformational search problem is solved. The unfolded chain follows the path of least resistance to the central well, which represents the native state. The roughness of the energy surface within the native well represents the dynamic nature of the folded protein. Hidden intermediates that may form alternate native conformation on or off the main folding route may also be populated in the native well.

In the case of two-state folders, all non-native conformations of the protein are in rapid equilibrium with each other and transition to the native conformation with all-or-none cooperativity. Such proteins tend to be small, single domain proteins with marginal stability. On the other hand, for multi-state folders, a hierarchy of structure formation appears to be necessary in order to reach the native state. From an initial inspection of the multi-state free energy landscape (Fig. 1.1b), it appears that the inherent stability of

Figure 1.3 – The new view of the folding free energy landscape. The loss in chain entropy as a protein molecule reaches the native state is represented by the x-axis, and the gain in energy as native like contacts are formed is represented by the y-axis. This representation suggests that each unfolded protein molecule can reach the native state via an independent pathway. In reality however, biases in the unfolded state ensemble and frustrations along the energy landscape can channel the protein molecules to flow down specific pathways. The roughness of the native well illustrates the dynamic nature of the native state, and excursions from this state to hydrogen exchange competent species are represented as hidden intermediates.

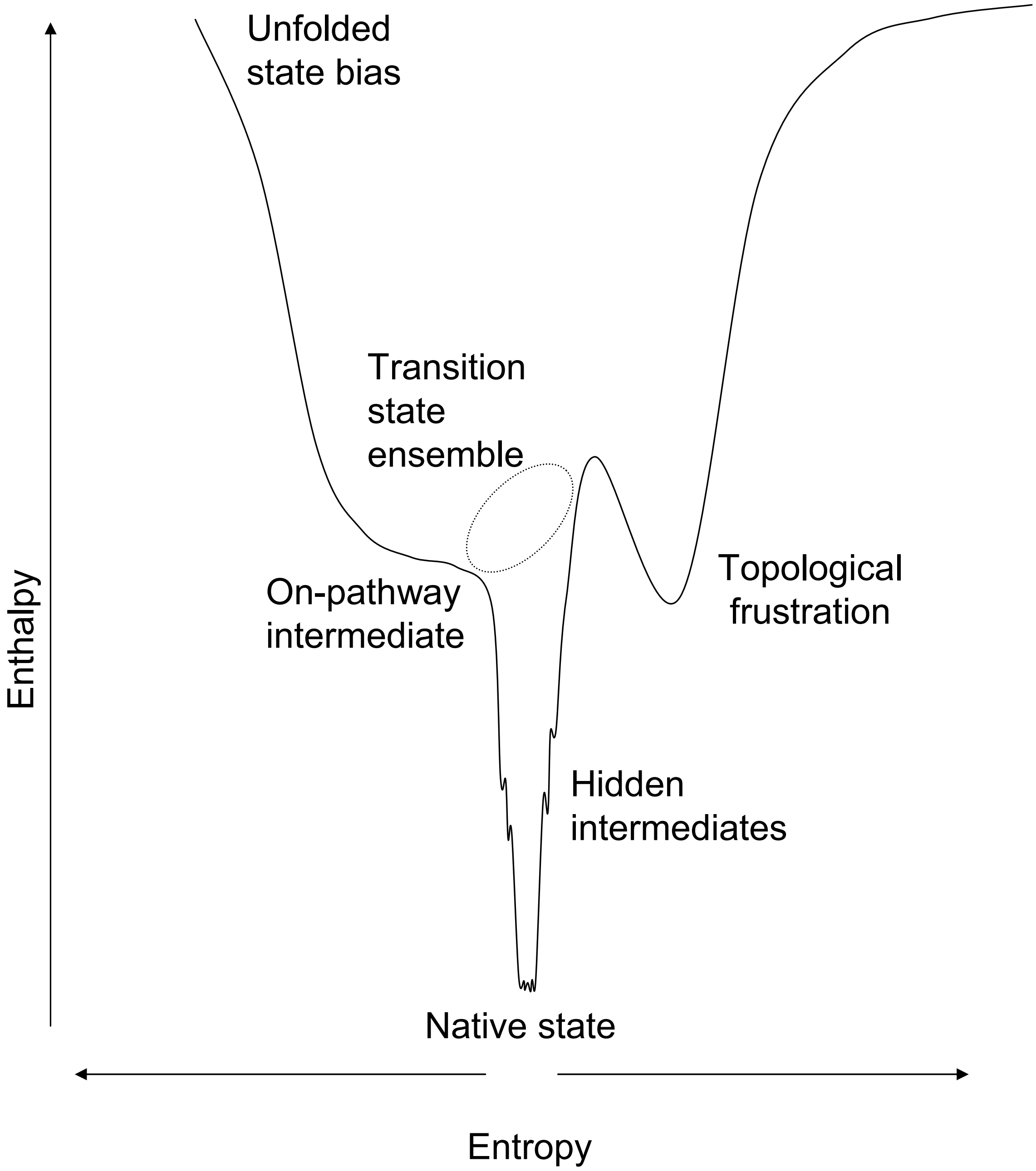

Unfolded
state bias
Transition
state
ensemble
On-pathway
intermediate
Topological
frustration
Hidden
intermediates
Native state
Enthalpy
Entropy

Figure 1.3

an intermediate will retard the rate at which the transition state is reached. However, it is important to note that on-pathway intermediates can serve to make the conformational search problem easier, and allow folding to progress in a step-wise manner. Further, a collapsed intermediate state can protect the protein from intermolecular reactions that may be kinetically accessible to the unfolded polypeptide. These intermediates may be further stabilized by non-native interactions that may simply rearrange on-route to the native structure.[33]

Another class of intermediates that are discussed in more detail in subsequent chapters is the off-pathway intermediates. Off-pathway intermediates may be non-native traps that are formed due to kinetic accessibility of alternate metastable structures, likely due to biases in the unfolded state and/or due to topological frustrations (Fig. 1.1c). These intermediates are not productive and do not help in the conformational search for the native state. They have been observed in several proteins of the βα-repeat class of proteins,[34 39] possibly due to the local connectivity of their secondary structure elements that may coalesce in non-natively collapsed states and cannot propagate productively on to the native structure.

While the basic phenomena involved in protein folding, namely, chain entropy, hydrophobic interactions, hydrogen bonding, electrostatic interactions, and solvent effects, are broadly understood, it has been a challenge to delineate the fine interplay of these effects during the protein folding process. As such, a generalized model to predict the mechanism of folding for a given amino acid sequence remains elusive.

The role of topology in protein folding

One very important determinant of protein folding rates is thought to be the structural complexity of the native state. In general, the more complex the native topology, the more difficult it will be to solve the conformational search problem and the rarer the productive folding events. Several groups in the past have proposed different metrics to measure the most important features of this structural complexity with respect to the folding rate of the protein.

The Contact Order (CO) parameter is one such metric,[40] wherein the chain separation of any two residues that are in contact with each other (defined as 6 Å from the center of the atom) is averaged over all the contacts within the protein. Short range contacts within local helical segments of the protein will reduce this metric, while longer range interactions that require closure of long loops will increase the value of the CO. Thus, the higher values of CO correspond to complex topologies with higher entropic costs of forming the native state. The CO metric in its original formulation was normalized for the length of the protein chain to obtain the Relative CO (RCO); however, because chain length is expected to increase the conformational space, this normalization was later discarded in favor of the Absolute CO (ACO).[41]

A good correlation between the log of folding rates of two state folders and their ACO has been reported when proteins are grouped into categories based on their secondary structure composition (Fig. 1.4a).[41,42] It is proposed that the inherent differences between the local contacts of residues in α-helices and the long range contacts in β-sheets contribute to the different responses observed in the different protein

Figure 1.4 – Correlations of folding rate with different topological parameters. a)[1] Correlation between folding rates and absolute contact order, ACO, the proteins are grouped into all-α (red), all-β (green) and mixed α+β (blue) **b)**[1] Correlation between folding rates and long-range order, LRO, color codes are the same as in (a) **c)**[2] Correlation between folding rates for the formation of the intermediates (non-two-state folders) and the number of sequence-distant native pairs Qd. **d)**[2] Correlation between folding rates for the formation of the native state (non-two-state folders) and Qd. **e)**[3] Correlation between folding rate constants for the formation of the intermediate (○) and native state (■), respectively, with the number of non-local contact clusters, N_c.

[1] Reprinted from Protein Science, Andrei Y. Istomin, Donald J. Jacobs and Dennis R. Livesay, *"On the role of structural class of a protein with two-state folding kinetics in determining correlations between its size, topology, and folding rate"*; Vol 16.11 Pages 2564-2569 – 2007. With permission from John Wiley and Sons, License number 2486681394942, Aug 12 2010. www.interscience.wiley.com

[2] Reprinted from Journal of Molecular Biology, Kiyoto Kamagata, Munehito Arai, Kunihiro Kuwajima, *"Unification of the Folding Mechanisms of Non-two-state and Two-state Proteins"*; Vol 339.4, Pages 951-965, 11 June 2004. With permission from Elsevier, License number 2486690421147, Aug 12 2010. http://www.sciencedirect.com/science/journal/00222836

[3] Reprinted from Journal of Molecular Biology, Kiyoto Kamagata, Kunihiro Kuwajima, *"Surprisingly High Correlation between Early and Late Stages in Non-two-state Protein Folding"*; Vol 357.5, Pages 1647-1654, 14 April 2006. With permission from Elsevier, License number 2486690637110, Aug 12 2010. http://www.sciencedirect.com/science/journal/00222836

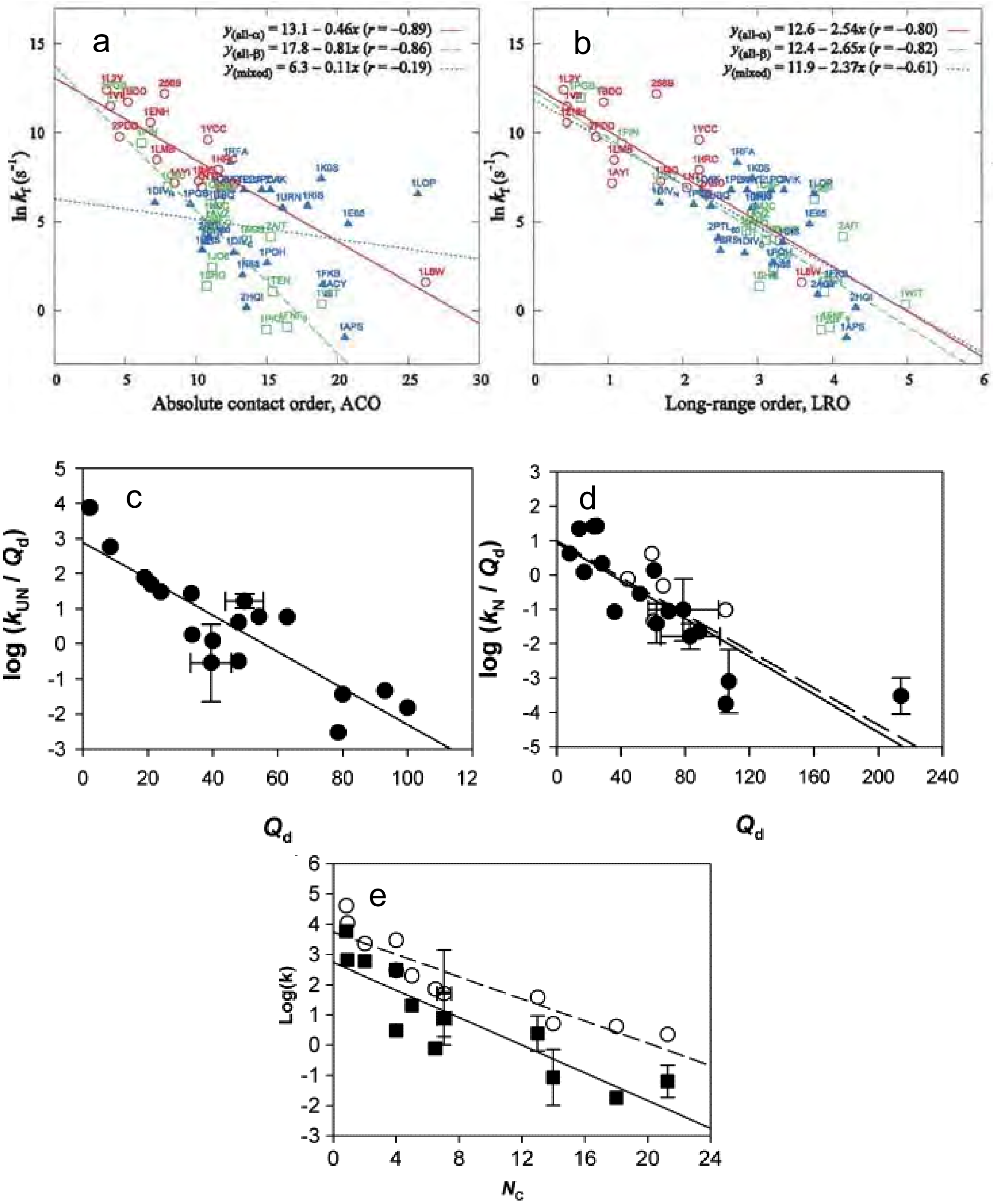

a
$y_{(all-\alpha)} = 13.1 - 0.46x$ ($r = -0.89$)
$y_{(all-\beta)} = 17.8 - 0.81x$ ($r = -0.86$)
$y_{(mixed)} = 6.3 - 0.11x$ ($r = -0.19$)
$\ln k_f$ (s^{-1})
Absolute contact order, ACO
0 5 10 15 20 25 30
15 10 5 0
1L2Y
1VII
1BDD
256B
1ENH
2PDD
1YCC
1LMB
1HRC
1RFA
1K8S
1LOP
1URN
1RIS
1E65
1POH
1FKB
1APS
2HQI
2AIT
1TEN
1WIT
1LBW
b
$y_{(all-\alpha)} = 12.6 - 2.54x$ ($r = -0.80$)
$y_{(all-\beta)} = 12.4 - 2.65x$ ($r = -0.82$)
$y_{(mixed)} = 11.9 - 2.37x$ ($r = -0.61$)
$\ln k_f$ (s^{-1})
Long-range order, LRO
0 1 2 3 4 5 6
15 10 5 0
1L2Y
1PGB
1BDD
256B
1ENH
2PDD
1FIN
1YCC
1LMB
1AYI
1HRC
1RFA
1K8S
1LOP
1E65
2AIT
1LBW
2HQI
1APS
1WIT
c
log (k_{UN} / Q_d)
Q_d
0 20 40 60 80 100 12
5 4 3 2 1 0 -1 -2 -3
d
log (k_N / Q_d)
Q_d
0 40 80 120 160 200 240
3 2 1 0 -1 -2 -3 -4 -5
e
Log(k)
N_c
0 4 8 12 16 20 24
6 5 4 3 2 1 0 -1 -2 -3

Figure 1.4

categories. The Long Range Order (LRO) parameter which ignores contributions from local interactions was proposed to overcome the different correlations observed for the different classes of proteins (Fig. 1.4b). The LRO is defined as the number of non-local contacts (Cα distance < 8 Å and sequence separation between contacting residues > 12) normalized for the chain length.[43]

A related term Qd, where the Cα distance of 6 Å was used and not normalized for chain length, was developed by Kuwajima.[44] This parameter was shown to correlate with the folding rates of both two-state (U → N) and multi-state folders (I → N) (Fig. 1.4c and d). Thus it appears that the rate limiting step in all proteins is the development of a transition-state ensemble with a native-like topology. Further, it was observed by the same group that in multi-state folders there is a strong correlation between the rate of folding of the intermediate (U → I) and the rate of folding of the native state (I → N). This observation indicates that the structure of the intermediate state is also dictated by the native state topology, albeit with a less well packed interior. Thus a new parameter (Nc) that counts the non-local contact clusters in the native state is shown to be correlated with the folding rates of intermediates and those of the native state (Fig. 1.4e).[45]

It follows that the sequence of events leading up to the formation of the transition state ensemble may also be determined by the native state topology. Coarse-grained, native-centric Gō-simulations have been used to demonstrate that even in the absence of sequence information, topological biases can prejudice folding cores and that at least some region of the conformational space sampled will not lead to productive folding.[46,47]

This phenomenon has been termed topological frustration and has been shown to exist even in simple two-state folders.[46,47]

Proteins are generally very robust and can withstand a variety of mutations, as is evident from the vast sequence diversity in several protein folds and even the same proteins from different organisms. On the other hand, single point mutations that are not expected to change the topology of the protein can result in large changes in the folding free energy landscape, as is evident from the large number of protein misfolding diseases identified to date.[48] Such variations within the same topology make it difficult to reduce this complex problem to a single parameter.

More recently, sequence flavored versions of Gō-models have been successful in differentiating between structural homologues with small differences in side-chain packing.[47] However, the pathways predicted by native centric Gō-simulations underestimate the contributions made by non-native conformations to the folding free-energy surface.

The role of sequence in protein folding

The largest contribution to the stability of proteins is that of main chain-main chain hydrogen bonds that form the framework for secondary structure elements of the protein.[49] While amino acid residues have preferred φ and ψ angles and distributions that can be used to predict the secondary structure elements of the protein (http://www.expasy.ch/tools/#proteome), this information is insufficient to define the final structure of the protein or its folding mechanism. The specificity for the final structure of the protein and for the cooperative events that lead to its formation must be stored in the side chains and the way they pack against each other. These interactions can be divided into two major types; electrostatic interactions and vdW interactions. Further there may be hydrogen bond interactions involving side chains. Elements of these packing interactions contribute to the topological order parameters described above to different extents.

Electrostatic interactions are unlikely to be the cause of such specificity as they generally tend to be fewer and distributed on the surfaces of proteins. Further, mutations in the protein exterior seldom alter the folding free energy landscapes of the protein. Side chain hydrogen bonds are also predominantly distributed on the surface of proteins and do not contribute significantly to the protein folding mechanisms (see below for exceptions).

The interiors of globular proteins however tend to be packed densely with hydrophobic residues,[50] and these vdW interaction energies contribute significantly to the stability of their resident proteins.[51] Their specific patterns of interactions are necessary to form the final tertiary packing in protein structures and thus, modulate the folding free energy landscapes.[52] Mutational phi-analyses developed by the Fersht group[53] and hydrogen exchange studies to identify the cores of protein stability, developed in the 1970’s and 80’s by the Wüthrich, Woodward and Englander groups,[54 56] have been used to demonstrate that proteins with the same topology can have widely differing folding mechanisms and stability cores.[57 59] A common feature of these experiments is the preponderance of non-polar residues in the interior of proteins that contribute to the folding cores of globular proteins. [60] Further these residues tend to be clustered together and not uniformly across the protein's interior.

Recent work from the Matthews laboratory on two members of the TIM barrel $(\beta\alpha)_8$-repeat group, alpha subunit of tryptophan synthase (αTS)[59] and Indole 3 glycerol phosphate synthase from *S. solfataricus* (sIGPS)[58] have suggested that the cores of stability for these two structurally homologous proteins are different in the two proteins. Both these proteins fold via off-pathway kinetic intermediates, and again the location and extent of structure of these species are different for the two proteins. Thus, the common topology of these proteins allows rapid access to locally frustrated landscapes; however, the differences in location and extent of structure in the off-pathway intermediates are a function of their variable sequences. In conclusion, while a protein's topology can be used to define the gross features of its folding mechanism, a subset of the native interactions in the core of a protein can significantly modulate the folding free energy landscape of a protein.

Scope of this thesis

This thesis focuses on two members of the $\beta\alpha$-repeat family of proteins, specifically the TIM barrel proteins (Fig. 1.5a) and the CheY-like proteins of the flavodoxin fold (Fig. 1.5b). This group of proteins represents the largest class of enzymes and signaling proteins. The TIM barrel architecture $(\beta\alpha)_8$ is represented in every class of enzymes, and the CheY-like proteins $(\beta\alpha)_5$ are the predominant receiver domains of the bacterial two-component signal transduction systems. Both these groups

Figure 1.5 – Topology of βα-repeat proteins a) TIM barrel $(\beta\alpha)_8$ repeat architecture. The 8 β-strands are arranged in a parallel β-sheet that forms the central barrel. Arranged in an antiparallel fashion to the central β-strands are the eight α-helices that form the outer shell of the barrel. Hydrophobic packing interactions between these two layers provide the stability typical of these proteins. **b)** CheY-like proteins. The 5 β-strands are arranged as a parallel β-sheet that forms the central layer of the $(\beta\alpha)_5$ repeat sandwich topology. α-helices 1 and 5 pack on one side of the central β-sheet and α-helices 2, 3 and 4 pack on the opposing side.

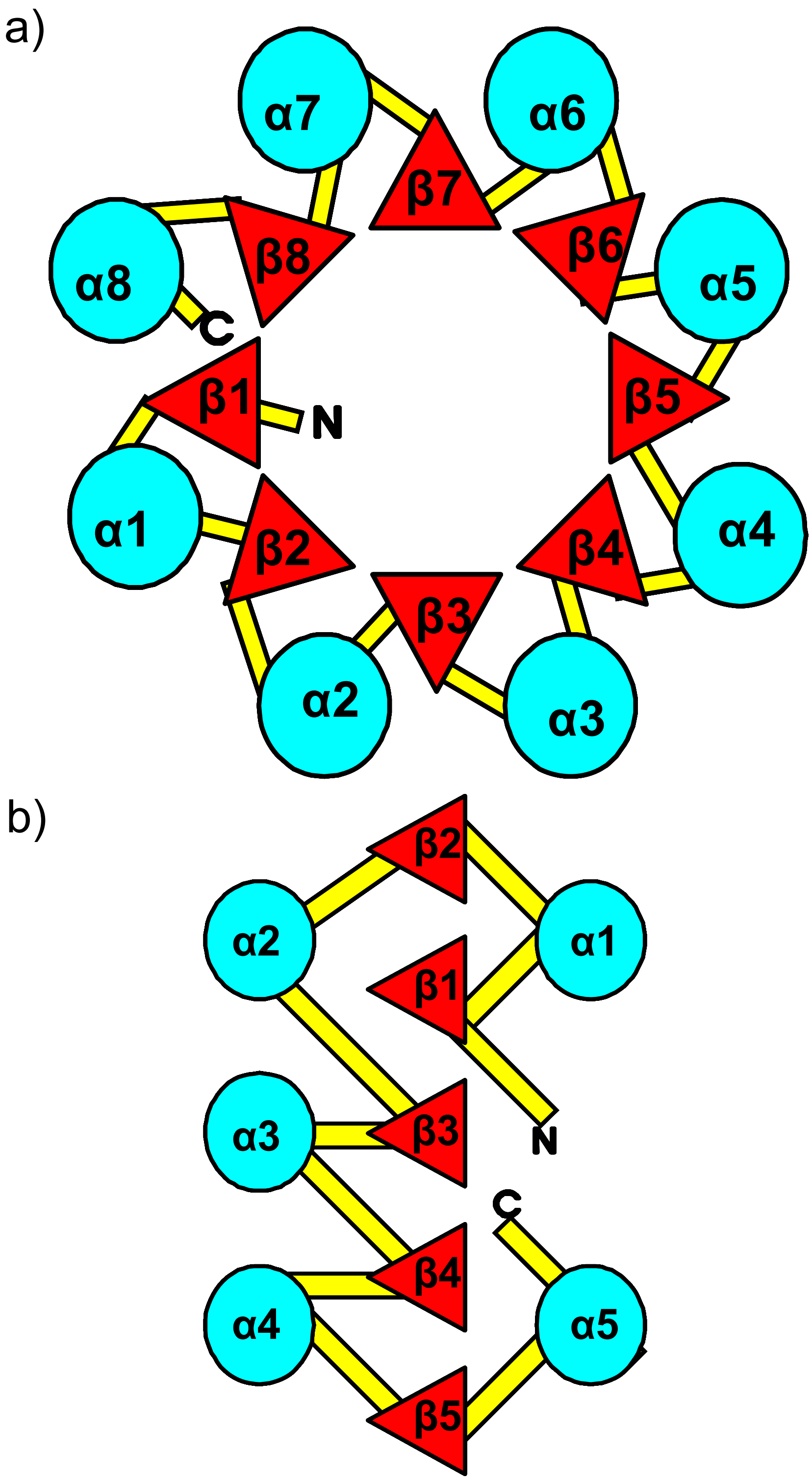

a)
α7
α6
β7
β8
β6
α8
α5
C
β1
β5
N
α1
β2
β4
α4
β3
α2
α3
b)
β2
α2
α1
β1
α3
β3
N
C
β4
α4
α5
β5

Figure 1.5

of proteins are thought to be derived from a common ancestral $(\beta\alpha)_4$-repeat protein.[61] As mentioned earlier, off-pathway folding intermediates appear to be a common feature of TIM barrel proteins. Similar off-pathway intermediates have been suggested to occur in the folding mechanism of the *E.coli* chemotaxis response regulatory protein, CheY.

In the subsequent chapters, I have addressed the following questions: 1) Do proteins with similar topology fold by similar mechanisms? 2) What are the sequence determinants for the differences in folding free energy landscapes of proteins with similar topology?

In Chapter II, I have used global analysis of equilibrium and kinetic data to demonstrate that the folding intermediate of CheY is most likely an off-pathway, non-natively packed species that must unfold before the productive transition state is accessed. This study was supported by Gō-simulation results from the Brooks laboratory, which identified the elements of secondary structure that participate in the off-pathway intermediate.[62]

In Chapter III, again in collaboration with the Brooks laboratory, the folding mechanism of CheY is compared with that of two more proteins of the CheY-like family, the N-terminal domain of the nitrogen regulatory protein, NTRC, from *S. typhimurium* (NT-NtrC) and the sporulation response protein, Spo0F, from *Bacillus subtilis*. A close inspection of the structures of these three proteins, suggests that self-sufficient hydrophobic clusters that may be kinetically accessible to locally connected secondary structure elements may form the cores of these intermediates. Such topologically frustrated cores may also be stabilized by non-native interactions to yield kinetically trapped species that cannot productively transition to the native state. Gō-simulations corroborate the experimental findings that the CheY-like topology is indeed frustrated and that non productive local interactions are formed between certain elements of secondary structure. A good correspondence is observed between the location of this frustration and the locally connected clusters of hydrophobic residues previously mentioned.

The sequence of these locally connected hydrophobic clusters tends to be dominated by the Branched Aliphatic Side Chain residues; Isoleucine, Leucine and Valine residues. The hydrophobic packing of these residues can not only direct the early collapse reaction but also provide stability to the native structure. The BASiC hypothesis, therefore supposes that the side chains of isoleucine, leucine and valine (ILV) residues form hydrophobic clusters that are very effective in preventing solvent access to their underlying hydrogen bond networks. The linkage between the secondary and tertiary structures enables these ILV clusters to serve as cores of stability in high-energy states.

In Chapter IV, I compare the locations of these ILV clusters in 19 proteins with their respective hydrogen exchange protected cores. The clusters of BASiC residues are sensitive to the location of protected amides, however they tend to overestimate their numbers. An improvement in the quality of the predictions may be possible with the inclusion of simple filters, such as distance from the protein exterior, hydrogen bond potentials, etc. Nevertheless, this BASiC hypothesis based method provides the means to incorporate the role of sequence in predicting the cores of protein stability.

Another group of protected main chain amides in TIM barrel proteins are hydrogen-bonded with side-chain carbonyl groups. This surprising finding by Dr. Xiaoyan Yang, and Dr. Ramakrishna Vadrevu from the Matthews laboratory led to the study reported in Chapter V. Non-local main chain-side chain hydrogen bonds appear to be an architectural principle of TIM barrel proteins. These hydrogen bonds that appear to connect the C-terminus of one α-helix to the N-terminus of the preceding β-strand are responsible for bracing βαβ modules that form the building blocks of the TIM barrel architecture.

In the final chapter, Chapter VI, I report preliminary results of the fast folding kinetics of NT-NtrC, which are consistent with the off-pathway folding mechanism. I also discuss future experiments that will provide conclusive proof for the off-pathway nature of the CheY folding pathway and also provide insights into the structure of the off-pathway intermediate. The possibility of developing a BASiC hypothesis driven metric for predicting folding rates of proteins is also discussed.

Chapter II – Kinetic Traps in the Folding of βα-Repeat Proteins: CheY Initially Misfolds before Accessing the Native Conformation

This chapter has been published previously as *Kathuria SV, Day IJ, Wallace LA, Matthews CR. "Kinetic traps in the folding of beta alpha-repeat proteins: CheY initially misfolds before accessing the native conformation." J Mol Biol. 2008 Oct 3;382(2):467-84.*

The work presented in the following chapter was a collaborative effort. Dr. Louise A. Wallace carried out the equilibrium unfolding experiments, the unfolding and refolding kinetic experiments and the experiments on cyclophilin dependence of refolding. Dr. Iain J. Day carried out the NMR experiments. I repeated the equilibrium and kinetic experiments, performed the double jump refolding experiments, determined the temperature dependence of the refolding kinetics and performed the global analysis. Dr. C. Robert Matthews and I wrote the manuscript.

Abstract

The βα-repeat class of proteins, represented by the $(\beta\alpha)_8$ barrel and the α/β/α sandwich, are among the most common structural platforms in biology. Previous studies on the folding mechanisms of these motifs have revealed or suggested that the initial event involves the sub-millisecond formation of a kinetically trapped species that must at least partially unfold before productive folding to the respective native conformation can occur. To test the generality of these observations, CheY, a bacterial response regulator, was subjected to an extensive analysis of its folding reactions. Although earlier studies had proposed the formation of an off-pathway intermediate, the data available were not sufficient to rule out an alternative on-pathway mechanism. A global analysis of single- and double-jump kinetic data, combined with equilibrium unfolding data, was used to show that CheY folds and unfolds though two parallel channels defined by the state of isomerization of a prolyl peptide bond in the active-site. Each channel involves a stable, highly-structured, folding intermediate, whose kinetic properties are better described as an off-pathway species. Both intermediates subsequently flow through the unfolded state ensemble and adopt the native *cis* prolyl isomer prior to forming the native state. Initial collapse to off-pathway folding intermediates is a common feature of the folding mechanisms of βα-repeat proteins, perhaps reflecting the favored partitioning to locally-determined substructures that cannot directly access the native conformation. Productive folding requires the dissipation of these prematurely-folded substructures as a prelude to forming the larger-scale transition state that leads to the native conformation. Results

from Gō-modeling studies in the accompanying paper elaborate on the topological frustration in the folding free energy landscape of CheY.

Introduction

The complex conformational changes that accompany the conversion of space-filling random coils representing unfolded proteins to unique native structures of proteins are well-described by surprisingly simple exponential responses. This evident simplicity is likely the result of a strong bias for native or native-like substructures and the avoidance of off-pathway traps for most proteins.[3,63] For small proteins (<~100 amino acids) that fold via a 2-state kinetic mechanism, the transition states mapped by experiment and theory sequester a substantial fraction of their ultimate buried surface area from solvent and favor a subset of native contacts.[64] Larger proteins often display more complex responses, and the transient and stable intermediates detected highlight the cores of stability that usually guide the reaction directly to the native conformation. In support of this view, it has been observed that (1) chemically-denatured states of proteins often reveal preferences for secondary structure in nascent helical segments,[65 68] (2) fragments of proteins often adopt native-like structures in the absence of their complementary sequences[69,70] and (3) native-centric Gō-model simulations of simple Cα chain models of proteins have had some success in predicting the folding mechanisms of monomeric and dimeric proteins.[71] Although non-native interactions have been observed in some cases,[33,72,73] the structural differences with the native fold are rather modest and the corresponding intermediates are kinetically on-pathway. Thus, it is surprising and notable that the initial events in the mechanisms of two very common structural motifs involve off-pathway species.

The alpha subunit of tryptophan synthase from *Escherichia coli* (αTS), indole-3-glycerol-phosphate synthase from *Sulfolobus solfataricus* (sIGPS), and IOLI, a protein of unknown function from yeast, all members of the ubiquitous TIM barrel class of protein structures, have been shown to fold within milliseconds to stable, partially-folded, kinetically off-pathway species whose unfolding reactions limit access to subsequent on-pathway intermediates.[34,36,58,74,75] Two members of the very large flavodoxin fold family, CheY[76,77] and apo-flavodoxin itself,[35] also collapse within milliseconds to an intermediate possessing significant stability and secondary structure. Experimental analyses of the folding mechanisms of these proteins provided definitive evidence favoring the off-pathway intermediate for apo-flavodoxin,[35] but left the on- vs. off-pathway issue unresolved for CheY.[78]

What distinguishes both the TIM barrel and flavodoxin motifs from others that have on-pathway folding intermediates[79] is their βα-repeat architecture. The barrels or sheets are comprised of parallel β-stands alternating in sequence with helices oriented in an anti-parallel sense to and docked directly on the strands. Both motifs are thought to be derived from the same evolutionary progenitor, comprising of a half barrel with four βα-repeats.[80] In the canonical TIM barrel motif, eight strands pack sequentially in a cylindrical format with β8 docking on β1; the amphipathic helices dock on the hydrophobic surface of the barrel. The strand order for the smaller flavodoxin fold is β2-β1-β3-β4-β5, with α1 and α5 on one side of the sheet and α2-α3-α4 on the opposing side to form a α/β/α sandwich (Fig. 2.1). The relative contact order[41] for both motifs is small, typically 8% to 9%, reflecting the topological simplicity, i.e., the dominance of sequence-local interactions, of the βα-repeat.

An example of a kinetic trap during the folding of the α/β/α sandwich architecture is provided by CheY,[78] a bacterial two-component response regulator involved in signal transduction.[81] CheY (Fig. 2.1) is a member of the CheY-like superfamily of proteins, itself a member of the larger group of the flavodoxin-like fold (α/β/α class) of small (~120 amino acids), single-domain proteins. Numerous members of the response-regulator family have been identified (>14000 Pfam[82]) and dozens of their structures have been elucidated (21 in SCOP[83 85] and 59 non-identical in CATH[86,87]). Earlier studies on the folding mechanism of CheY[88] have shown that its urea-induced equilibrium denaturation is well described by a two-state process involving only the native and denatured species. The kinetic refolding reaction of CheY is dominated by the formation of a sub-millisecond highly-structured intermediate, i.e. the stopped-flow burst-phase intermediate ($\tau < 5$ ms), and a subsequent slow isomerization reaction of the K109 P110 peptide bond from the *trans* to the native *cis* conformation. Mutational analysis[78,89] and Gō-modeling studies[77] on CheY have suggested that the transient intermediate is comprised of a native-like core region β1 β3 and a prematurely folded region comprised of β3 α5. The C-terminal region of this intermediate would have to unfold before the folding reaction can access the N-terminal transition state leading to the native structure. However, on the basis of the data available, it was not possible to choose between an on-pathway or an off-pathway role for this crucial species.

Figure 2.1 – A ribbon diagram of the crystal structure of CheY reported by Voltz et al.[90] The α-helices are colored cyan; the β-strands, magenta, and the loop regions, salmon. The aromatic side chains are shown as stick figures, tryptophan (W58) in green, tyrosines (Y51 and Y106) in blue, and phenylalanines (F8, F14, F30, F53, F111 and F24) in pink. Four adjacent phenylalanine residues (F8, F30, F53 and F124), which may contribute to the exciton coupling observed in the sub-millisecond intermediate, are connected with broken lines. The proline residue (P110) that contributes to the cis-prolyl bond (K109 P110) is also shown as a stick figure and is highlighted in red.

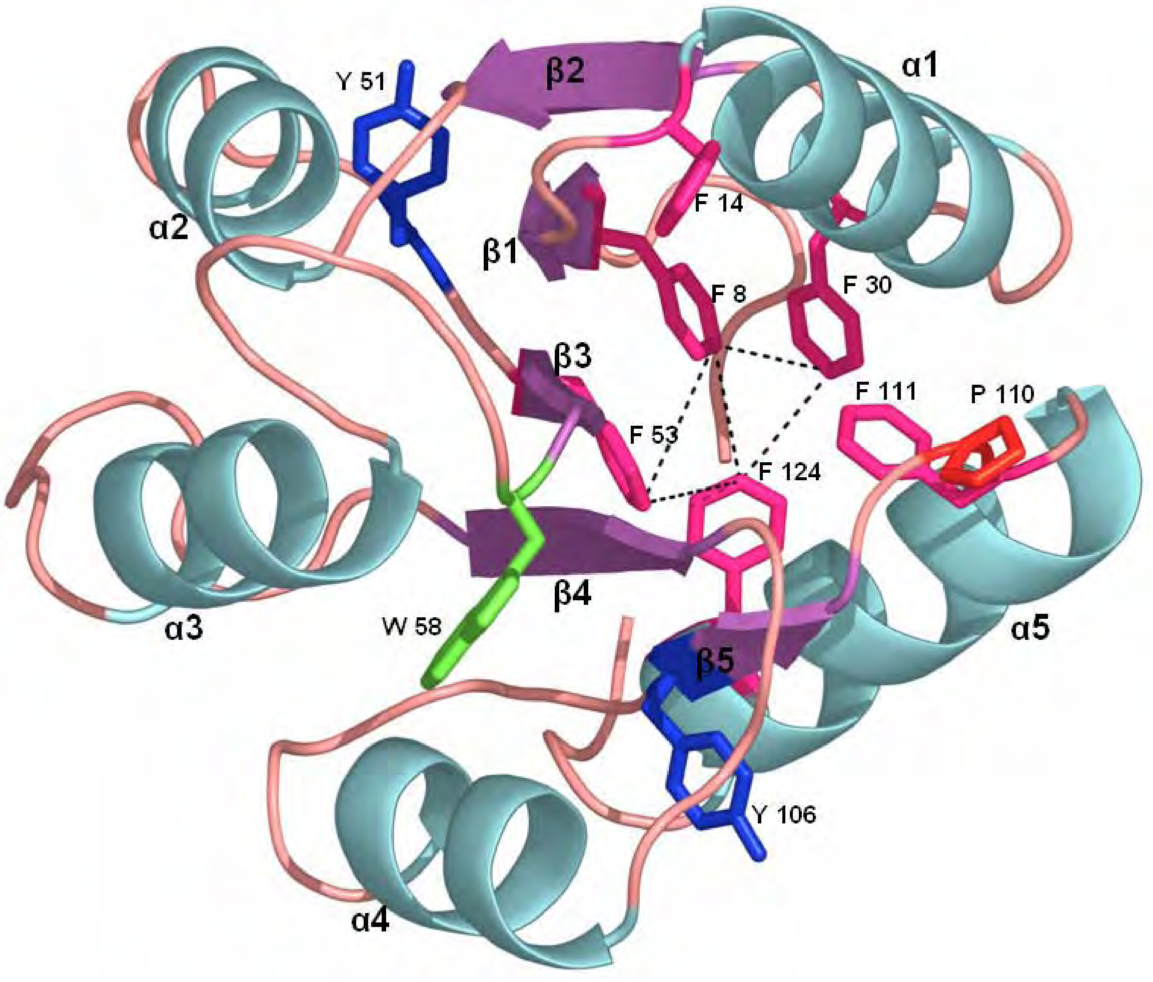

Y 51
β2
α1
α2
F 14
β1
F 30
F 8
β3
F 111
P 110
F 53
F 124
α3
β4
α5
W 58
β5
Y 106
α4

Figure 2.1

In the present study, a global analysis of a comprehensive set of thermodynamic and kinetic data for wild type *E. coli* CheY was used to determine its folding mechanism and the associated microscopic rate constants. The most likely folding model places the burst-phase species in an off-pathway position linked to the unfolded state. These kinetically-trapped species must at least partially unfold before productively folding along parallel channels defined by the *cis* and *trans* isomers of the K109 P110 peptide bond. Gō-modeling studies in the accompanying paper corroborate the early formation of a non-productive species during the folding of CheY.[62] The prevalence of kinetic traps in βα-repeat motifs may reflect their propensity to readily access local minima in the folding free energy surface.

Results

Equilibrium studies

Structural analysis

Far-UV CD and tryptophan fluorescence spectroscopy were used to monitor the loss of secondary and tertiary structure of CheY in the presence of the chemical denaturant urea. The CD spectrum of the native conformation is typical of βα-sheet proteins, with the characteristic α-helical signature of two minima at ~210 and ~222 nm modulated by the β-sheet signature at ~218 nm (Fig. 2.2a). The far-UV CD spectrum of CheY measured in the presence of 6 M urea resembles that of a random coil, indicating a complete loss of secondary structure upon denaturation (Fig. 2.2a). The fluorescence emission spectra of native and denatured CheY (Fig. 2.2b) are dominated by the

Figure 2.2 – CD and FL spectra of CheY (a) CD spectra of native CheY (solid line), chemically denatured CheY in the presence of 6.25 M urea (dotted line), and the kinetic burst-phase intermediate detected in 5 ms (△). The difference spectrum between the native and burst-phase species is also shown (×). **(b)** Fluorescence emission spectra of native CheY (solid line) and chemically denatured CheY in the presence of 6.5 M urea (dotted line). Buffer conditions: 10 mM potassium phosphate at pH 7.0 and 25 °C.

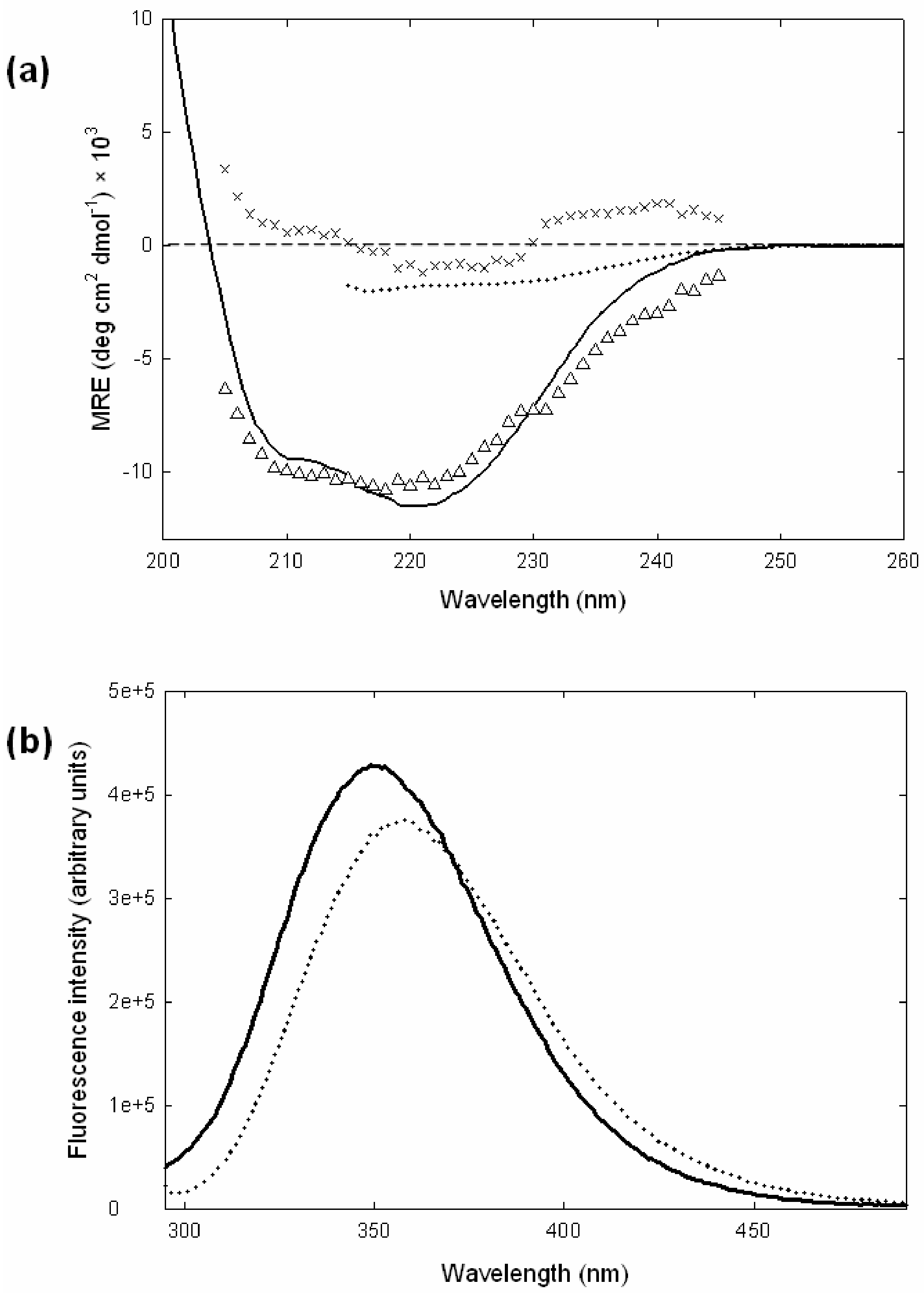

(a)
10
5
0
-5
-10
MRE (deg cm2 dmol-1) × 103
200
210
220
230
240
250
260
Wavelength (nm)
(b)
5e+5
4e+5
3e+5
2e+5
1e+5
0
Fluorescence intensity (arbitrary units)
300
350
400
450
Wavelength (nm)

Figure 2.2

contribution from the single tryptophan residue at position 58. The red shift in the emission maximum from 350 nm to 357 nm and the decrease in intensity upon unfolding reflects the exposure of the buried tryptophan, W58, to solvent and the accompanying increase in the efficiency of fluorescence quenching.

Equilibrium unfolding free energy surface

The urea-induced equilibrium unfolding transitions monitored by CD at 222 nm and the fluorescence emission at 315 nm show a single sigmoidal transition (Fig. 2.3a). While the normalized CD and FL equilibrium transition curves are coincident within error (Fig. 2.3b), the small and consistent shift to higher urea concentration for the FL curve may reflect the presence of an intermediate. A folding intermediate has been reported to be populated at ~3 M urea by Garcia *et al.*[91] at very high protein concentrations (100 - 200 μM). The excellent reversibility of the urea denaturation reaction was demonstrated by coincident unfolding and refolding CD transitions (Fig. 2.3b). A global analysis of the CD and FL spectral changes with urea concentration, yielded values for the Gibbs free energy of unfolding from N to U, ΔG^{o}, calculated in the absence of urea, of 5.37 ± 0.21 kcal mol ^{1}, the dependence of ΔG^{o} on the denaturant concentration, the m-value, of 1.59 ± 0.06 kcal mol ^{1} M ^{1}, and the mid point of the transition, C_{m}, of 3.33 M urea at pH 7.0 and 25 ºC. These results indicate that the native (N) and unfolded (U) states dominate the population of states at all urea concentrations and at the concentrations of protein used in this study (<100 μM), i.e., the unfolding transition is highly cooperative. These thermodynamic parameters are comparable to and within error of those reported by Filimonov *et al.*[88] under the same conditions.

Figure 2.3 – Equilibrium denaturation of CheY (a) Equilibrium unfolding of CheY. The CD signal at 222 nm is plotted as a function of denaturant concentration (○) and fit to a two-state model (broken and dotted line). The burst-phase amplitude measured by stopped-flow refolding CheY from 5.2 M urea is plotted as a function of final urea concentration (△) and fit to a two-state model (thick broken line). The baselines for the native state (solid line) and the unfolded state (dotted line) predicted from the two-state model are also shown. The magnitude of the burst-phase amplitude under strongly refolding conditions (0.52 M urea) is represented by the double-headed arrow. The fluorescence intensity at 315 nm is plotted as function of urea concentration (□) and fit to a two-state model (thin dashed line). (b) Apparent unfolded fraction (Fapp) of CheY as a function of denaturant concentration. The unfolding transition monitored by CD at 222 nm (○), the refolding transition monitored by CD at 222 nm (●), and unfolding transition monitored by tryptophan fluorescence (□) are shown. The confidence limits of the global fit to a two-state model are also shown (broken and dotted lines). The buffer conditions are described in the caption for Figure 2.2

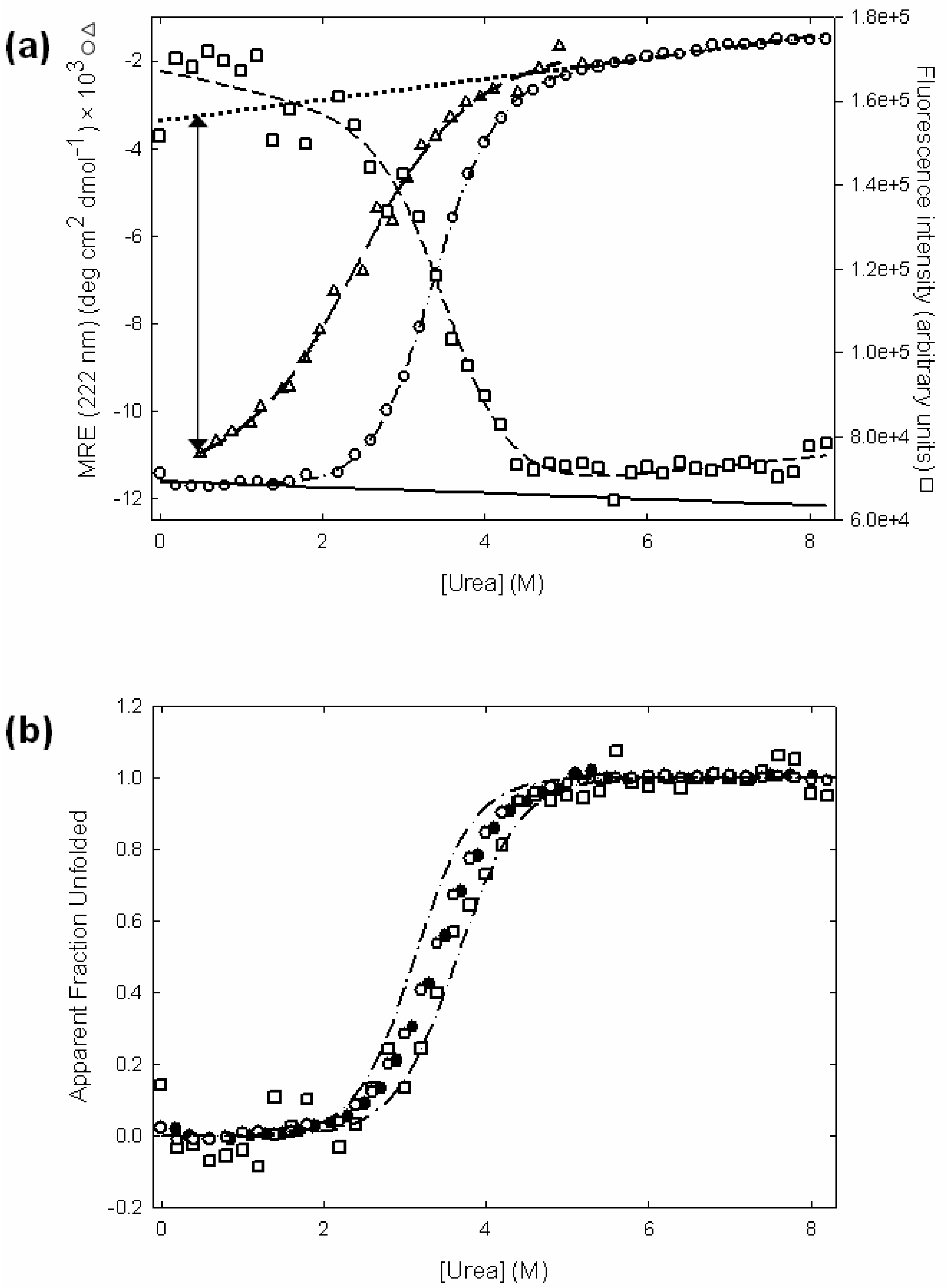

(a)
MRE (222 nm) (deg cm^2 dmol^{-1}) × 10^3 ○△
-2
-4
-6
-8
-10
-12
Fluorescence intensity (arbitrary units) □
1.8e+5
1.6e+5
1.4e+5
1.2e+5
1.0e+5
8.0e+4
6.0e+4
0
2
4
6
8
[Urea] (M)
(b)
Apparent Fraction Unfolded
1.2
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
0
2
4
6
8
[Urea] (M)

Figure 2.3

A more refined view of the equilibrium free energy surface for CheY can be obtained by recognizing the role of *cis/trans* proline isomerization in the unfolded state. Although the K109 P110 peptide bond adopts the *cis* isomeric conformation in the native state, it is expected to equilibrate to a mixture of both isomers in the absence of secondary and tertiary structure in the unfolded state.[92] An estimate of the ratio of the isomers in the unfolded state of CheY can be obtained by examining the 1D proton NMR spectrum of the Ac-VKPFT-NH_2 penta-peptide containing the key prolyl peptide bond (Fig. 2.4). The dominant *trans* isomer (whose presence was demonstrated in the denatured full-length protein, see below) comprises ~90% of the population, consistent with a free energy difference of 1.36 kcal mol ^{1} between the two isomers. A re-fit of the CD and FL equilibrium unfolding data to a three-state model assuming a urea-independent free energy difference of 1.36 kcal mol ^{1} between the two unfolded states yields free energy differences for the $N_C \leftrightarrows U_T$ and $N_C \leftrightarrows U_C$ transitions of 5.42 ± 0.21 and 6.78 ± 0.21 kcal mol ^{1}, respectively. The subscripts indicate the isomeric state of the K109 P110 peptide bond. The m-value for the $N_C \leftrightarrows U_C$ transition in this model is 1.59 ± 0.06 kcal mol ^{1} M ^{1}, consistent with the behavior expected for a globular protein of 129 amino acids.[93]

Kinetic studies

A stable burst-phase intermediate is formed during refolding

The formation and disruption of secondary structure in CheY during folding and unfolding reactions was monitored by the changes in ellipticity at 222 nm and fluorescence intensity above 320 nm using stopped-flow and manual-mixing methods.

Figure 2.4 – 1D proton spectrum The amide region of the 1D proton NMR spectrum of the Ac-VKPFT-NH2 penta-peptide containing the K109-P110 peptide bond. The peaks associated with the cis and trans isomers are highlighted.

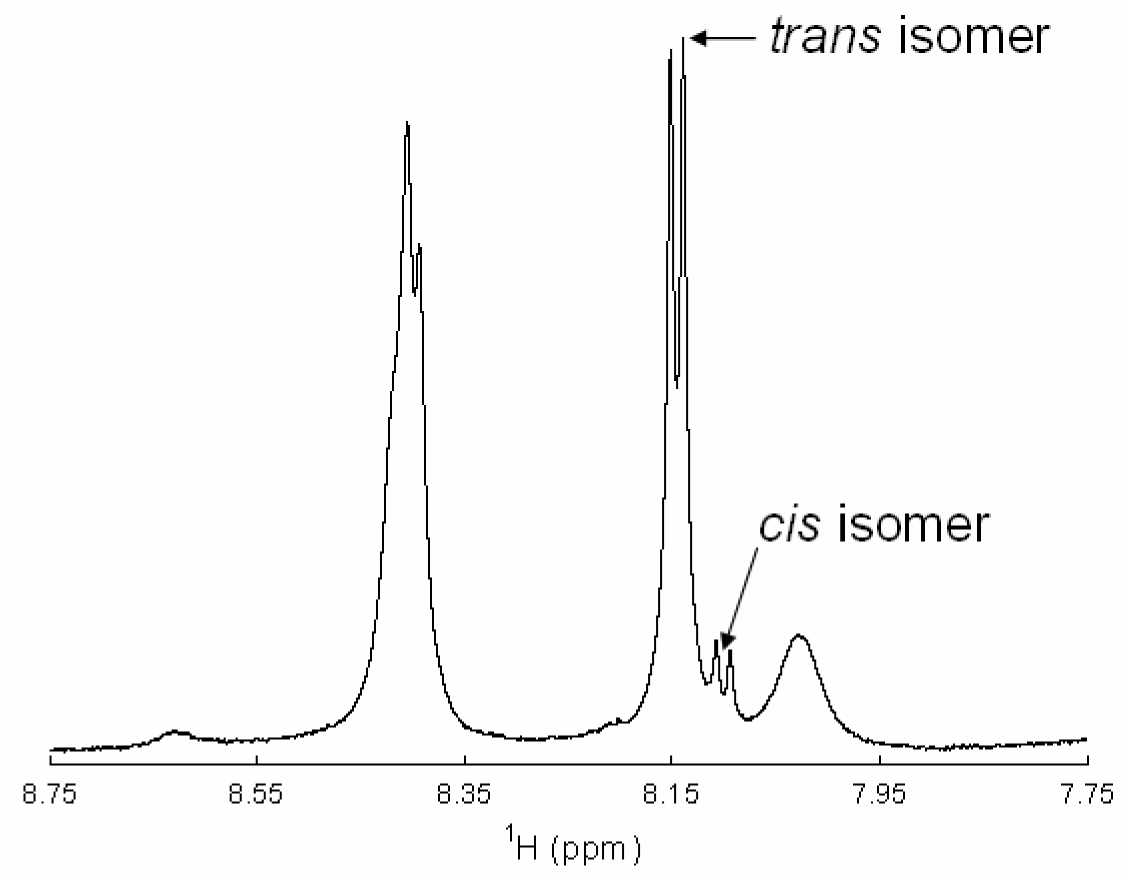


Figure 2.4

Approximately 95% of the native CD (Fig. 2.2a) signal at 222 nm is recovered within the dead-time of the instrument, ~5 ms, during refolding from the urea-denatured state. Serrano and colleagues have previously reported a similar burst-phase intermediate, I^{BP} that accounts for 94% of the native CD signal and ~80% of the fluorescence amplitude.[76] These investigators postulated that the equilibrium intermediate detected by NMR spectroscopy is similar in structure to this burst-phase species.[91]

The CD spectrum of I^{BP} was obtained by measuring the amplitude of the burst-phase signal for the refolding reaction at a series of wavelengths between 205 nm and 245 nm (Fig. 2.2a). The spectrum of the intermediate is similar to that of the native state; however, the small but significant differences near 220 nm and above 230 nm suggest that the packing of aromatic side chains might not be identical. The difference spectrum between the N and I^{BP} species (Fig. 2.2a), negative between ~215 and 230 nm, a cross-over at 230 nm and positive above 230 nm, is indicative of the perturbation of an exciton coupling between two aromatic side chains in vdW[94] in native CheY. Inspection of the structure (Fig. 2.1) shows a quartet of closely-packed phenylalanine residues between the C-termini of $\alpha1$ and $\alpha5$ and the N-termini of $\beta1$ and $\beta3$. A looser or altered packing of one or more of these residues in I^{BP} could be responsible for this effect. Neither of the two tyrosine residues nor the single tryptophan is in close contact with another aromatic side chain (Fig. 2.1). By implication, the secondary structures of N and I^{BP} are indistinguishable.

The apparent stability of I^{BP} was determined by monitoring the amplitude of the CD burst-phase reaction at 222 nm as a function of the final urea concentration in refolding (Fig. 2.3a). The sigmoidal decrease in amplitude with increasing urea concentration indicates a cooperative loss in secondary structure accompanying the chemical denaturation of I^{BP}. A fit of the data to a two-state model yields estimates of $\Delta G°$ 2.3 ± 0.4 kcal mol^{1} for its stability, an m-value of 0.92 kcal mol^{1} M^{1} and a C_m of 2.49 M urea. Although the secondary structure content is similar for N and I^{BP}, the lower m-value implies either that I^{BP} is not as well-packed as N,[93] or that additional intermediates are populated during the urea-induced unfolding process. The putative perturbation of an exciton coupling between aromatic side chains in I^{BP} favors the former explanation.

Proline isomerization is the rate limiting step for formation of native structure

The remaining 5% of the native CD signal at 222 nm and the FL signal above 320 nm at 0.5 M urea was recovered by an exponential process with a relaxation time of 150 s at pH 7.0 and 25 °C. The very slow rate of this reaction and its minimal dependence on the final urea concentration (Fig. 2.5a) suggest that it might reflect the *trans* → *cis* isomerization reaction of the K109 P110 peptide bond. To test this possibility, the refolding reaction was performed in the presence of increasing concentrations of cyclophilin, a prolyl isomerase. Refolding of CheY from 6 M urea to 1 M urea at 15 °C is accelerated in the presence of cyclophilin (Fig. 2.6a), consistent with its assignment to a prolyl isomerization reaction. Further support for this conclusion was provided by measuring the temperature dependence of this refolding reaction (Fig. 2.6b). The observed activation energy, 20.8 ± 1.1 kcal mol^{1}, is typical of prolyl isomerization reactions.[92,95] Because the K109 P110 peptide bond preferentially forms the *trans*

Figure 2.5 – Chevron analysis of CheY. (a) The recovery of the native signal upon refolding from high denaturant concentration occurs by single exponential kinetics. The associated relaxation time determined by CD (×) and by fluorescence (●) are shown. Two phases were observed in unfolding kinetics; the relaxation times determined by CD (✚, slow phase and **–**, fast phase) and by fluorescence (■, slow phase and ▲, fast phase) are also shown. The relaxation times determined by refolding the protein after a 2 s exposure to unfolding conditions (6.5 M urea), i.e., the refolding double-jump experiments, are represented by open symbols; a fast refolding phase (○), and a slower phase with negative amplitude (□) are observed. **(b)** Amplitudes associated with the relaxation times determined by fluorescence. The symbols used are the same as in panel A. A single refolding phase is observed in refolding kinetics (●). Unfolding to moderate urea concentrations is dominated by the slow phase (■) below 4.2 M urea. The faster unfolding channel (▲) becomes dominant above 4.2 M urea. The buffer conditions are described in the caption for Figure 2.2

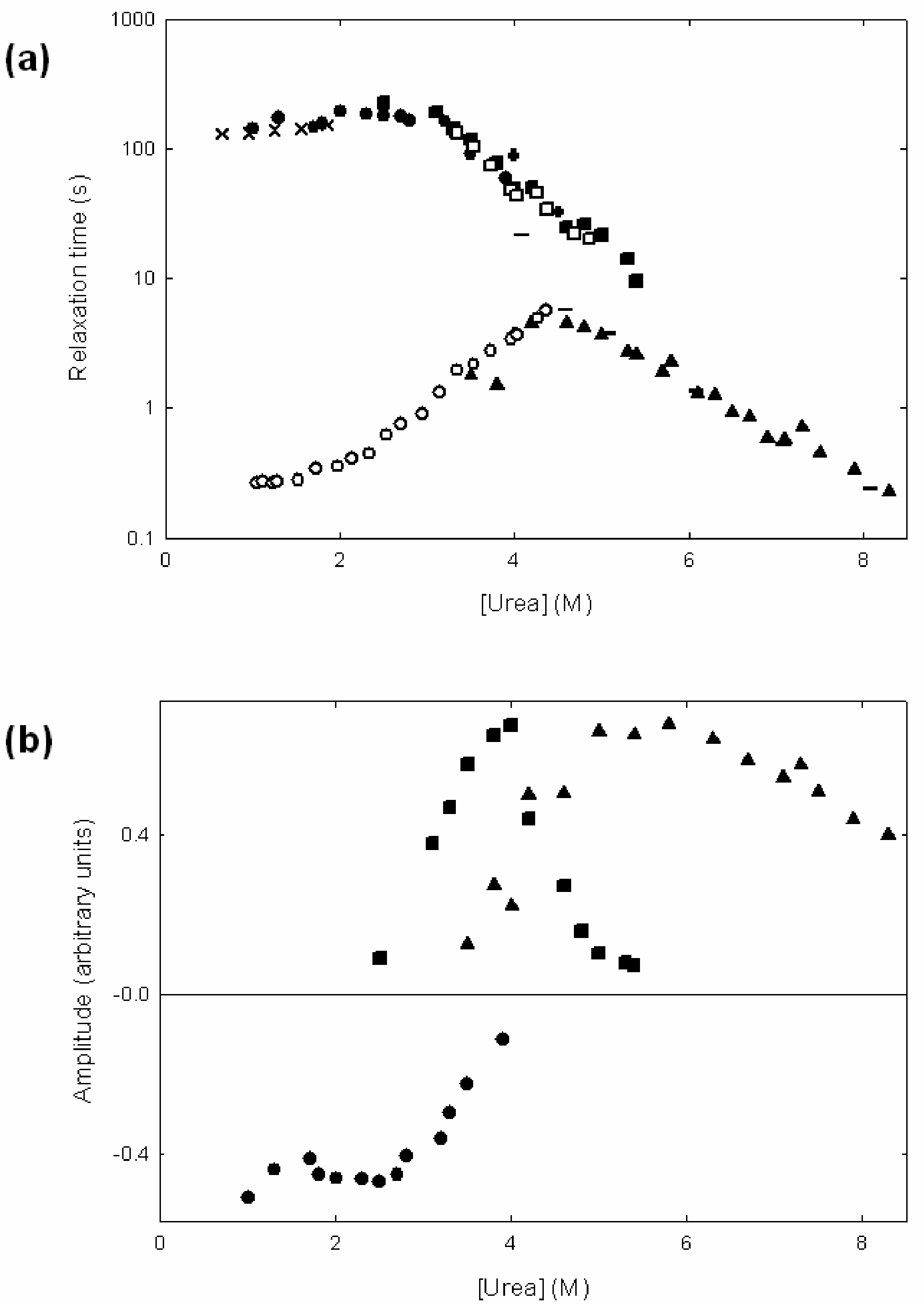

(a)
1000
100
10
1
0.1
Relaxation time (s)
0
2
4
6
8
[Urea] (M)
(b)
0.4
-0.0
-0.4
Amplitude (arbitrary units)
0
2
4
6
8
[Urea] (M)

Figure 2.5

Figure 2.6 – Proline isomerization in CheY (a) Dependence of the slow refolding relaxation time on the cyclophillin concentration. Refolding to native conditions from 6.5 M urea was performed by 10-fold dilution with refolding buffer at 15 °C, and the change in signal was monitored by CD at 222 nm. **(b)** Arrhenius plot for the slow refolding rate constant for refolding jumps from 6.5 M urea to 1 M urea observed by fluorescence. The data were fit to the Arrhenius equation, $\ln k = \ln A - \frac{E_a}{RT}$, and the fit is represented by the solid line.

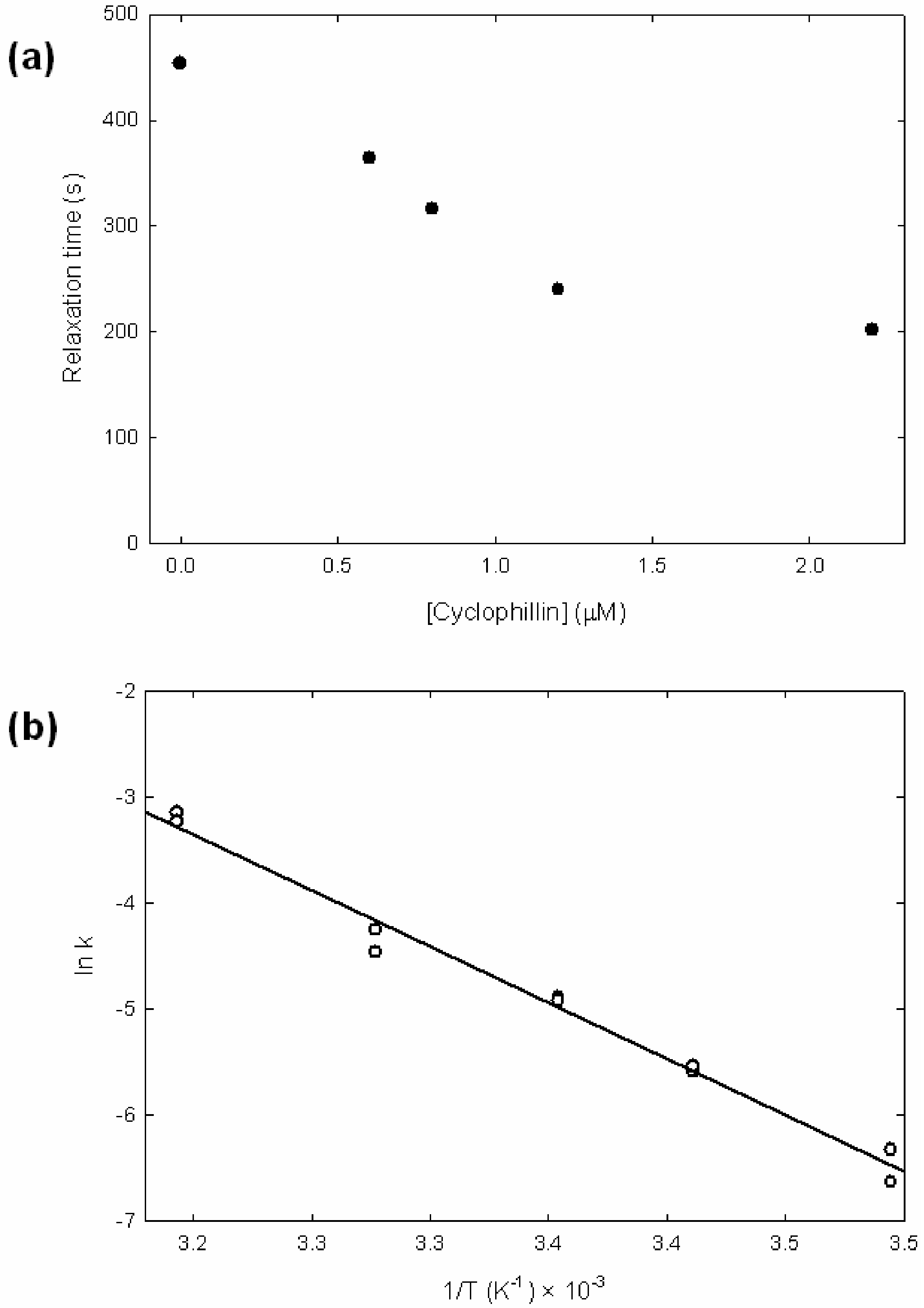

(a)
Relaxation time (s)
500
400
300
200
100
0
0.0
0.5
1.0
1.5
2.0
[Cyclophillin] (μM)
(b)
ln k
-2
-3
-4
-5
-6
-7
3.2
3.3
3.3
3.4
3.4
3.5
3.5
1/T (K-1) × 10-3

Figure 2.6

isomer in the unfolded state (Fig. 2.4), these data support the assignment of the rate-limiting step in folding to the formation of the *cis* K109 P110 peptide bond found in the native conformation. This assignment is also consistent with a previous mutational analysis of CheY in which the replacement of proline 110 with glycine eliminated this slow folding phase.[76] A faster refolding phase that is coincident with the *cis*-channel refolding chevron (see below) is observed at temperatures below 15 °C (data not shown).

Complex unfolding reaction

The unfolding reaction of CheY as observed by FL emission shows a single slow phase below 3.5 M urea, biphasic kinetics between 3.5 M and 5.4 M urea and a single phase above 5.4 M urea (Fig. 2.5a). Similar results were obtained with a more limited investigation of the far-UV CD signal (Fig. 2.5a). The unfolding relaxation time of the slower phase decreases exponentially from 192 s at 3.1 M urea to 10 s at 5.4 M urea. At lower concentrations of urea the relaxation time of the slow unfolding phase is urea independent and coincident with that for the slow refolding phase. The relaxation time of the faster unfolding phase also decreases exponentially with urea, from 11 s at 3.5 M urea to 0.34 s at 7.9 M urea. The amplitude of the slow unfolding phase increases to a maximum at ~4 M urea and subsequently decreases at higher urea concentrations (Fig. 2.5b). The amplitude of the fast unfolding phase increases proportionally as that for the slow phase decreases at higher urea concentrations; above 5.4 M urea, the slow unfolding phase entirely disappears.

Double-jump refolding kinetics

A pair of double-jump refolding experiments were performed on CheY to obtain assignments for the pair of unfolding reactions, to examine the refolding reaction from the unfolded state containing the native *cis* isomer at the K109 P110 peptide bond, and to monitor the rate of the *cis* → *trans* isomerization reaction in the unfolded state. The first two objectives could be met with a double-jump refolding experiment in which the native state was rapidly unfolded in 6.5 M urea for 2 s and then refolded to a series of final urea concentrations in the native baseline region. By limiting the unfolding time to a value sufficient to unfold CheY but not to allow the *cis* K109 P110 peptide bond to isomerize, 2 s (Fig. 2.5a), it was possible to monitor the urea dependence of the refolding reaction from the U_C state. The relaxation times for the major refolding phase near 4 M urea, i.e., from the U_C state, merge smoothly with those for the faster unfolding phase (Fig. 2.5a). The chevron shape for these data are those expected for the reversible transition between two thermodynamic states,[96] demonstrating that the faster unfolding phase reflects the unfolding of the N_C state. Although the refolding relaxation time for this phase accelerates exponentially with decreasing urea concentration down to ~2 M urea, it becomes independent of the urea concentration below that point. It is notable that an additional small phase of opposite amplitude, observed between 3.3 and 4.8 M urea, has a urea-dependent relaxation time that is coincident with the slow unfolding phase (Fig. 2.5a).

The continuous change in the relaxation times for the slow unfolding and refolding phases and the assignment of the slow refolding phase to the isomerization of the K109 P110 peptide bond make it plausible to assign the slow unfolding phase to that for the N_T state. The switching of the unfolding amplitudes from favoring the slow phase to favoring the fast phase at increasing urea concentrations (Fig. 2.5b) suggests that the N_C state switches between unfolding via the N_T state at moderate urea concentrations and via the direct unfolding route to the U_C state at high urea concentrations.

The third and final objective, determination of the rate constant for the isomerization of the K109 P110 peptide bond in the unfolded state, can be met by a complementary double-jump refolding experiment. If the N_C state is unfolded for various lengths of time and then refolded to the same strongly folding conditions, the time dependence of the disappearance of the amplitude of the fast refolding phase will correspond to the relaxation kinetics of the *cis* peptide bond in denaturing conditions.[92] As shown in Figure 2.7, the relaxation time for the re-equilibration of the *cis/trans* isomer ratio for the K109 P110 peptide bond at 6.5 M urea, τ, is 20.0 ± 1.3 s at 25 ºC. Combined with the estimate of the equilibrium constant derived from the penta-peptide containing the prolyl peptide bond, K   [*trans*]/[*cis*]   ~9.0, the individual microscopic rate constants for the inter-conversion of the *cis* and *trans* isomers can be determined from τ^{1} $k_{CT} + k_{TC}$ and K k_{CT}/k_{TC}. The values for k_{CT} and k_{TC} are 4.5×10^{2} s^{1} and 5.0×10^{3} s^{1}, respectively.

A significant and surprising feature of the double-jump refolding data was the observation that the amplitude of the fast refolding reaction approaches 80% of the expected signal change at very small delay times (Fig. 2.7). If N_C were the only species in solution in the absence of denaturant prior to the unfolding jump, the extrapolated

Figure 2.7 – Double jump refolding Fluorescence amplitude associated with the cis refolding channel as a function of unfolding time in the double-jump refolding experiment. The protein was allowed to denature for varying lengths of time at 6.5 M urea prior to refolding at 1.1 M urea. All the traces were fit simultaneously to globally-linked relaxation times for the *cis* and *trans* channels. The fit to a single exponential is represented by the solid line.

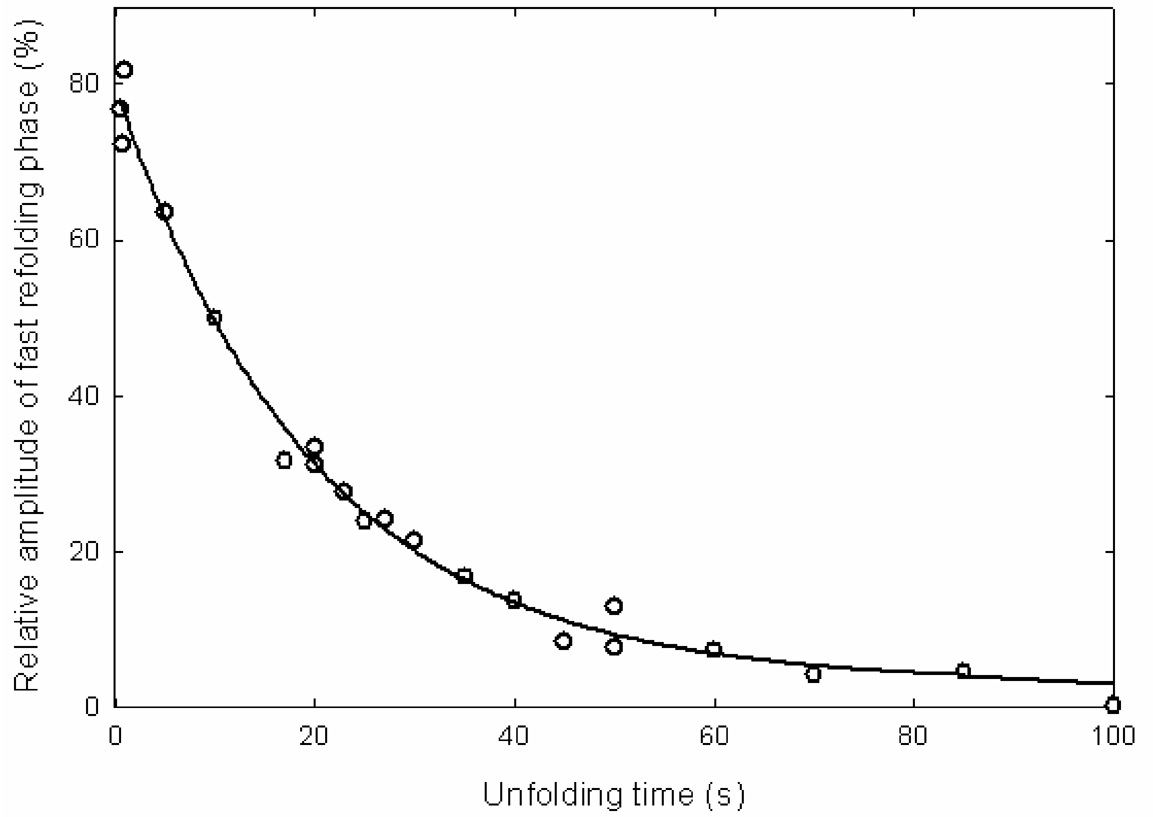

Relative amplitude of fast refolding phase (%)
0
20
40
60
80
Unfolding time (s)
0
20
40
60
80
100

Figure 2.7

amplitude should have been 100%. Inspection of the double-jump refolding data over longer time ranges reveals a second slow refolding phase whose relaxation time is coincident with that for the slow phase observed in single-jump refolding experiments (Fig. 2.7). This result implies that the U_T state becomes partially populated even upon unfolding for 500 ms (the shortest delay time interval used in the experiment). Because this very short time interval between unfolding and refolding jumps is insufficient for the prolyl isomerization reaction to occur in the unfolded state, the N_T state must be partially-populated under native conditions.

Global analysis to elucidate kinetic folding mechanism

The kinetic folding model for CheY must be consistent with the following observations:

1. The N_C native state, containing the *cis* isomer for the K109 P110 peptide bond, can unfold either through its native counterpart with the *trans* peptide bond, N_T, to the U_T state or directly to the unfolded state with the *cis* peptide bond, U_C, depending on the final urea concentration.
2. The N_T state is measurably populated, ~20%, under native conditions.
3. The unfolded state with the *trans* isomer, U_T, dominates the population at high urea concentrations and is the primary starting point for the refolding reaction. The absence of an observable refolding phase from the U_C state implies that it must represent less than 10% of the population, the detection limit of the CD instrument.
4. In less than a few milliseconds, the U_T state folds to an intermediate state with native-like secondary structure, altered packing of the phenylalanine side chains and

substantial stability, $I^{BP}{}_{T}$. The *trans* peptide bond isomerizes to the native *cis* isomer either in the $I^{BP}{}_{T}$ state to $I^{BP}{}_{C}$ or in the U_T state to U_C, where it is at least partially accessible to the solvent and the prolyl isomerase.

5. Direct folding from the U_C to the N_C state becomes limited by a denaturant-independent process under strongly folding conditions.

Two parallel channel kinetic folding models involving 6 states, N_C, N_T, $I^{BP}{}_{C}$, $I^{BP}{}_{T}$, U_C and U_T, were tested for their consistency with these observations (Fig. 2.8a and 2.8d). In Model 1, the burst-phase intermediate states are placed on the direct folding pathway from the unfolded states to the native states. In Model 2, these intermediates are placed off the direct folding pathway, in rapid reversible equilibrium with the unfolded states. The kinetic unfolding and refolding traces, including those from the refolding double-jump experiment, were fit to a global model whose initial estimates were based on 1) experimentally determined equilibrium properties (Fig. 2.3a), 2) the microscopic rate constants and their urea-dependences obtained from the chevron plot of the dependences of the relaxation times on the final denaturant concentrations (Fig. 2.5a), and 3) the experimentally determined rate of prolyl isomerization (Fig. 2.7). The initial estimates of these parameters were enhanced by an in-house algorithm, Chevron Fitter, which enables their manual adjustment for direct comparison of the predicted eigenvalues for a chosen mechanism with the observed relaxation times. Amplitudes associated with each observable rate constant and the steady state distributions of species as a function of denaturant concentration are also predicted using the same software. The procedure is described in detail in Materials and Methods. Although the rate constants for the burst-

Figure 2.8 – Global analysis. (a) The on-pathway model. The folding mechanism occurs via parallel channels based on the isomerization state of the K109 P110 peptide bond. The burst-phase species, I^{BP} is placed on-pathway, between the unfolded and native states along either channel. **(b)** Predicted chevron from the on-pathway model. The predicted observable relaxation times are shown as solid black lines; the microscopic rate constants determined by the model are shown as broken green lines for the *trans* channel and dotted green lines for the *cis* channel. The isomerization relaxation times in each state are shown as broken and dotted lines, blue for the native states, red for the unfolded states, and magenta for the burst-phase intermediates. **(c)** The equilibrium population of species predicted by the on-pathway model. The native states are represented by blue lines, the unfolded states by red lines and the intermediate states by magenta lines. The dotted lines represent the species in the *cis* state while the broken lines represent those in the *trans* state. The sum of the *cis* and *trans* channel populations of each species is represented by the solid lines. **(d)** The off-pathway model. The burst-phase species, I^{BP} is placed off-pathway from the unfolded states in both channels. **(e)** Predicted chevron from the off-pathway model. The legend is the same as in panel b. **(f)** The equilibrium population of species predicted by the off-pathway model. The legend is the same as in panel c.

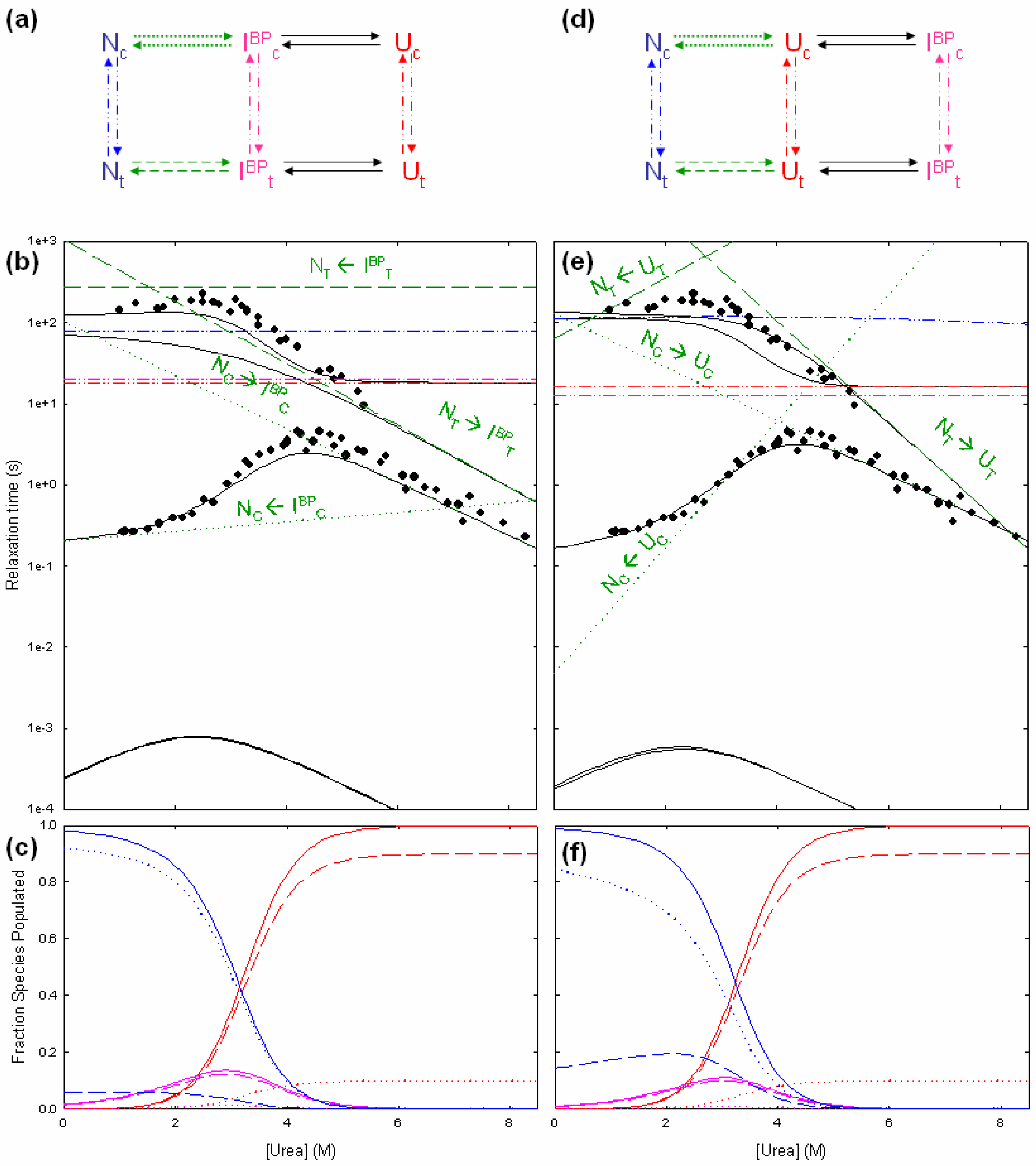

(a)
N_c
I^BP_c
U_c
N_t
I^BP_t
U_t
(d)
N_c
U_c
I^BP_c
N_t
U_t
I^BP_t
(b)
N_T ← I^BP_T
N_C → I^BP_C
N_T → I^BP_T
N_C ← I^BP_C
(e)
N_T ← U_T
N_C → U_C
N_T → U_T
N_C ← U_C
Relaxation time (s)
1e+3
1e+2
1e+1
1e+0
1e-1
1e-2
1e-3
1e-4
(c)
(f)
Fraction Species Populated
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
[Urea] (M)

Figure 2.8

phase folding reaction cannot be measured by stopped-flow mixing, the stability of the $I^{BP}{}_{C}$ and $I^{BP}{}_{T}$ species and their urea dependences were fixed to the values obtained from fitting the urea-dependence of the CD burst-phase amplitude (Fig. 2.3a). The folding rate constants for the burst-phase intermediates were assumed to be $>10^4$ s ^{1} to account for their appearance within 5 ms. As a first-approximation, the urea dependences of the unfolding and refolding rate constants for the $I^{BP}{}_{T}$ and $I^{BP}{}_{C}$ species, $m^{BP}{}_{T}$ and $m^{BP}{}_{C}$, were each assigned to be half of the m-value for the equilibrium unfolding reaction, 0.92 kcal mol ^{1} M ^{1}.

A set of 98 unfolding, refolding and double-jump kinetic traces under a variety of final conditions were then fit to both models, and the parameters were optimized using the Levenberg-Marquardt algorithm.[34,97] The microscopic rates, kinetic m-values, and the Z-values (relative signal contribution from each species normalized to the difference between the signals for the native and unfolded species) were modeled globally,[34] while the total signals of the native (N_C) and unfolded (U_C) were allowed to vary for each kinetic trace.

The quality of the fits to the two models was assessed by comparison of their reduced chi-squared values and by visual comparisons of the predicted chevrons and amplitudes for the globally minimized parameters. The equilibrium population of the intermediates in either model is sufficiently low at all urea concentrations as to remain undetectable, and the resulting equilibrium denaturation plot is consistent with the experimentally observed two-state behavior. Although either model provides credible fits of the kinetic data, the reduced chi-square value for the on-pathway model is 30%

higher than that obtained from the fit for the off-pathway model (number of degrees of freedom ~10000). Additional support for the off-pathway model is provided by inspection of the urea-dependence of the microscopic rate constants for both models (Fig. 2.8b and 2.8e) and the predicted populations of the N_T state in the absence of denaturant (Fig. 2.8c and 2.8f). In the on-pathway model, the nearly urea-independent refolding rate constants in both the *cis* and *trans* channels implies that the intervening transition states have a buried surface area that is very similar to the corresponding I^{BP} states. This result would be contrary to observations on the folding kinetics of many proteins,[98] which typically find that the buried surface area of transition states is more similar to that of the more well-folded state in a 2-state reaction.[64] By contrast, the typical urea-dependences for the $U_C \rightarrow N_C$ and $U_T \rightarrow N_T$ steps in the off-pathway model (Fig. 2.8e) demonstrate the burial of a significant fraction of exposed surface area in both transition states. The validity of the off-pathway model is also supported by the better agreement of the predicted fraction of the N_T state in the absence of urea, 14.5% ± 0.4%, compared to the on-pathway model (6.3% ± 0.5%), (Fig. 2.8f) with the observed fraction from the refolding double-jump experiment under the same conditions, ~20% (Fig. 2.7). Representative refolding, double jump and unfolding traces along with their fits using the parameters from the global analysis of the off-pathway model are shown in Figure 2.9. The superior statistical fit, the more reasonable urea dependence for the microscopic rate constants and the better agreement with the population of the N_T state all support the off-pathway kinetic folding model for CheY. The microscopic rate constants and their

Figure 2.9 – Global analysis fits. (a) Representative kinetic refolding traces of CheY after a 2 s exposure to unfolding conditions (6.5 M urea), i.e., the refolding double-jump experiments. The individual traces (open circles) and their fits to the off-pathway model obtained from the global analysis (solid lines) are shown. **(b)** Representative kinetic traces from the refolding experiments (filled circles), unfolding experiments (open circles), and their fits to the off-pathway model obtained from the global analysis (refolding: solid lines and unfolding dashed lines) are shown. The buffer conditions are described in the caption for Figure 2.2, and the final urea concentration for each trace is shown.

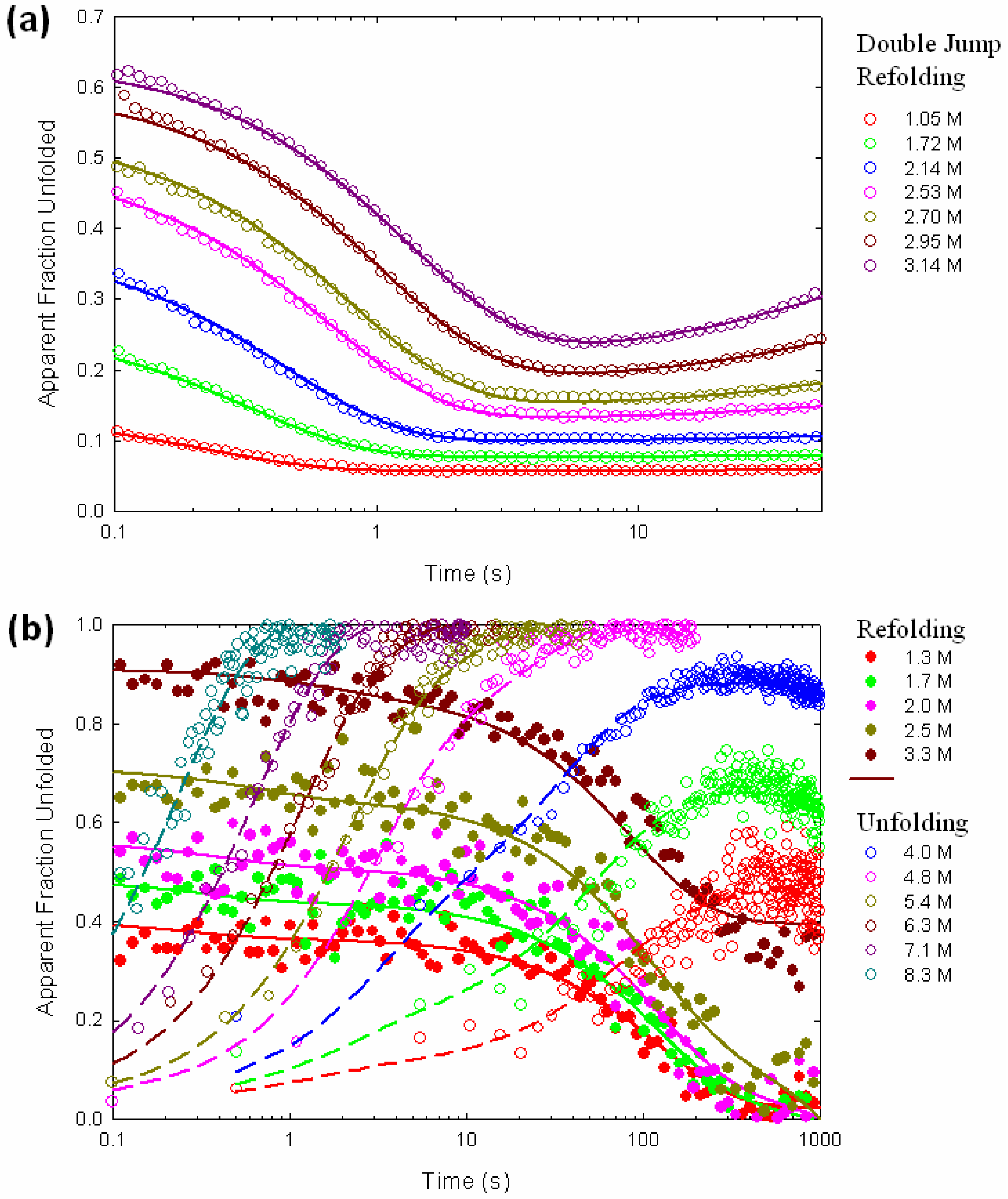

(a)
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Apparent Fraction Unfolded
0.1
1
10
Time (s)
Double Jump
Refolding
○ 1.05 M
○ 1.72 M
○ 2.14 M
○ 2.53 M
○ 2.70 M
○ 2.95 M
○ 3.14 M
(b)
1.0
0.8
0.6
0.4
0.2
0.0
Apparent Fraction Unfolded
0.1
1
10
100
1000
Time (s)
Refolding
● 1.3 M
● 1.7 M
● 2.0 M
● 2.5 M
● 3.3 M
Unfolding
○ 4.0 M
○ 4.8 M
○ 5.4 M
○ 6.3 M
○ 7.1 M
○ 8.3 M

Figure 2.9

associated m-values derived from the on- and off-pathway models are shown in Table 1a and Table 1b, respectively.

For completeness, a third model, similar to the folding mechanism of apo-flavodoxin as described by Sancho *et al.*,[35] was also tested (results not shown). This model allows the intermediates to be either on- or off- pathway, i.e., a triangular model in each folding channel, and the initial values for the parameters were derived from the on- and off-pathway models. The rate constants for the on-pathway and off-pathway reaction were allowed to float and a chi-square minimization routine was performed on the data. Although the reduced chi-square values are marginally better for this model, the on-pathway route is too slow for folding or unfolding to occur via this pathway at any urea concentration. In effect, the triangular model defaults to Model 2, the off-pathway burst-phase intermediate model.

Definitive validation of the off-pathway folding mechanism for the I^{BP} species in CheY requires direct observation of the sub-millisecond reaction with micro-fluidic techniques. These experiments are in progress.

Discussion

In an attempt to determine the structure of both the transition state and the burst-phase intermediate formed during the refolding of CheY, Lopez-Hernandez *et al.*[78] constructed a series of mutants of CheY that stabilized individual helical elements of secondary structure. The resulting stabilization of I^{BP} and/or the transition state, relative to the native state, was used to assess the role of helical segments in the formation of I^{BP}

Table 2.1 – Microscopic rate constants and their associated urea-dependences.[A]

[A]Obtained from a global analysis of 15 refolding, 35 unfolding and 48 double-jump refolding kinetic traces. The errors reported are standard errors from the global fits of the data obtained by standard propagation methods.
[B]The lower limits of the rate constants for the burst phase reactions are reported.

a) Model 1: on-pathway intermediate

Microscopic step	k (s^{-1})	m (kcal mol^{-1} M^{-1})
$U_C \rightarrow U_T$	$4.91\times10^{2} \pm 1.50\times10^{4}$	0
$U_T \rightarrow U_C$	$5.47\times10^{3} \pm 1.67\times10^{4}$	0
$U_C \rightarrow I^{BP}_C$ [B]	4.13×10^{3}	0.46
$I^{BP}_C \rightarrow U_C$ [B]	1.01×10^{2}	-0.46
$U_T \rightarrow I^{BP}_T$ [B]	4.03×10^{3}	0.46
$I^{BP}_T \rightarrow U_T$ [B]	9.82×10	-0.46
$I^{BP}_C \rightarrow I^{BP}_T$	$4.45\times10^{2} \pm 5.49\times10^{3}$	$0 \pm 2.18\times10^{2}$
$I^{BP}_T \rightarrow I^{BP}_C$	$4.96\times10^{3} \pm 1.08\times10^{3}$	$0 \pm 3.87\times10^{2}$
$I^{BP}_C \rightarrow N_C$	$4.91 \pm 4.87\times10^{2}$	$0.08 \pm 9.97\times10^{4}$
$N_C \rightarrow I^{BP}_C$	$9.68\times10^{3} \pm 1.94\times10^{4}$	$-0.45 \pm 2.17\times10^{3}$
$I^{BP}_T \rightarrow N_T$	$3.61\times10^{3} \pm 6.84\times10^{4}$	$0 \pm 2.90\times10^{2}$
$N_T \rightarrow I^{BP}_T$	$9.51\times10^{4} \pm 1.62\times10^{4}$	$-0.52 \pm 2.49\times10^{2}$
$N_T \rightarrow N_C$	$1.19\times10^{2} \pm 4.51\times10^{3}$	$0 \pm 7.17\times10^{2}$
$N_C \rightarrow N_T$	$8.00\times10^{4} \pm 2.49\times10^{4}$	$-0.01 \pm 6.00\times10^{2}$

b) Model 2: off-pathway intermediate

Microscopic step	k (s^{-1})	m (kcal mol^{-1} M^{-1})
$U_C \rightarrow U_T$	$5.57\times10^{2} \pm 2.97\times10^{4}$	0
$U_T \rightarrow U_C$	$6.21\times10^{3} \pm 9.04\times10^{5}$	$0 \pm 5.47\times10^{3}$
$U_C \rightarrow I^{BP}_C$ [B]	4.86×10^{3}	0.46
$I^{BP}_C \rightarrow U_C$ [B]	1.48×10^{2}	-0.46
$U_C \rightarrow N_C$	$2.10\times10 \pm 1.59$	$1.06 \pm 9.36\times10^{4}$
$N_C \rightarrow U_C$	$7.71\times10^{3} \pm 1.22\times10^{4}$	$-0.45 \pm 1.95\times10^{3}$
$U_T \rightarrow I^{BP}_T$ [B]	5.49×10^{3}	0.46
$I^{BP}_T \rightarrow U_T$ [B]	1.48×10^{2}	-0.46
$U_T \rightarrow N_T$	$1.59\times10^{2} \pm 7.89\times10^{4}$	$0.51 \pm 1.29\times10^{2}$
$N_T \rightarrow U_T$	$3.07\times10^{5} \pm 1.82\times10^{6}$	$-0.85 \pm 1.33\times10^{2}$
$I^{BP}_C \rightarrow I^{BP}_T$	$7.15\times10^{2} \pm 7.70\times10^{3}$	$0 \pm 2.13\times10^{2}$
$I^{BP}_T \rightarrow I^{BP}_C$	$7.08\times10^{3} \pm 4.90\times10^{4}$	$0 \pm 1.48\times10^{2}$
$N_T \rightarrow N_C$	$7.57\times10^{3} \pm 8.63\times10^{4}$	$0.04 \pm 2.90\times10^{2}$
$N_C \rightarrow N_T$	$1.29\times10^{3} \pm 1.33\times10^{4}$	$-0.11 \pm 2.41\times10^{2}$

Table 2.1

and/or the transition state. The result of this study suggested that all five helices are formed in the burst-phase intermediate. The striking similarity between the far-UV CD spectra of the burst-phase species and the native state (less the contribution from the presumed exciton coupling), reported in this paper, is consistent with this finding and also provides strong evidence that all of the elements of secondary structure in the native state are formed within the dead time of the stopped-flow instrument. A Gō-model simulation performed by Clementi *et al.*[77] also came to this same conclusion.

Both Lopez-Hernandez *et al.*[78] and Clementi *et al.*[77] highlighted a crucial role for native-like structure in the N-terminus but not the C-terminus, in the transition state,[89] suggesting that the extent of structure formed in the transition state is substantially less than that observed in the intermediate. At least partial unfolding of I^{BP} would thus be essential in order to access this transition state, as suggested by the results of the global analysis reported in this paper. While the near zero m-value of refolding in the on-pathway model reported here, is not as convincing as a negative value reported in a previous study,[78] it is important to note that the model used in the earlier study was a stable mutant (F14N) of CheY. The preferential stabilization of the burst-phase intermediate over the transition state in this mutant could yield a negative m-value for the refolding reaction.

The results of a more recent Gō-model simulation containing sequence information, described in the accompanying paper,[62] provides additional structural insights for I^{BP} and also suggests that the early formation of native-like contacts between the interface of the N- (β1 β3) and C-terminal (α3 α4) regions of the protein must be disrupted before the productive folding reaction can occur. Premature folding in the C-terminus thus impedes progress in attaining the native conformation and the off-pathway intermediate formed must at least partially unfold for the formation of the productive transition state.

Mechanistic analysis of CheY folding

Equilibrium unfolding mechanism

The equilibrium unfolding mechanism of CheY is reasonably well-described by a two-state model that invokes significant populations of only the native conformation and the unfolded conformation and ignores the role of proline isomerization in defining the equilibrium energy surface. A comprehensive kinetic analysis of the dynamic folding mechanism, including information on the role of proline isomerization in both folded and unfolded states, however, revealed additional complexities that enable a detailed description of a more elaborate folding free energy surface. The "native" state is comprised of two slowly-interconverting conformers containing the *cis*, ~80%, and the *trans*, ~20%, isomers that unfold independently through parallel channels. The "unfolded" thermodynamic state is comprised of two conformers containing the *trans*, ~90%, and *cis*, ~10%, prolyl isomers.

The off-pathway sub-millisecond kinetic intermediate with the *trans* isomer is also populated to ~10% under equilibrium conditions near the midpoint of the unfolding transition, 3 M urea (Fig. 2.8f). An intermediate with similar properties has previously been inferred to exist in the unfolding transition region as an associated species at the high concentrations (100 - 200 μM) required for NMR methods.[91] The less efficient packing of the I^{BP} species, implied by its smaller m-value for unfolding, might provide opportunities for non-specific intermolecular interactions.

Kinetic folding and unfolding mechanisms

In less than the 5 ms dead time of the stopped-flow instrument, the U_T and U_C conformers for CheY collapse to their respective off-pathway intermediate forms, $I^{BP}{}_T$ and $I^{BP}{}_C$, the former of which has modest stability and native-like secondary structure. Although the properties of $I^{BP}{}_C$ could not be directly measured, the presence of the native *cis* prolyl isomer makes it reasonable to suppose that its stability and structure would be at least comparable to those for $I^{BP}{}_T$. Given the presence of 10% of the molecules in the *cis* confomer in the unfolded state, the nearly native-like burst-phase CD signal in the absence of denaturant further strengthens the assumption that $I^{BP}{}_C$ is comparable to $I^{BP}{}_T$ in the amount of secondary structure formed. The difference CD spectrum between the native conformers and the burst-phase conformers implies that an exciton coupling between aromatic side chains in vdW contact in the native conformers is altered or lost in the burst-phase intermediates. Along with the reduced m-value for the urea denaturation of the burst-phase species, these data imply that the cluster of 4 phenylalanines and other buried nonpolar side chains are not as well packed in the burst-phase intermediates as those in the native conformations. This less efficient packing, however, does not appear to preclude the complete development of secondary structure. The extent of secondary structure formed and the decrease in side-chain packing in the intermediate are consistent with a molten globule-like state.[31,99,100] Mutational analysis of the core residues of CheY is required to test this hypothesis (see below).

The subsequent K109 P110 peptide bond isomerization reaction probably occurs between the I^{BP}_{T} and I^{BP}_{C} conformers because they represent 98% of the population of molecules within milliseconds after folding begins. The looser packing would presumably enhance isomerization and access to cyclophilin when present. This urea-independent isomerization reaction serves as the rate-limiting step in folding because the partial unfolding of I^{BP}_{C} and the subsequent refolding to N_C are fast steps under strongly folding conditions (Table 1b and Fig. 2.10a). Refolding along the *trans* proline channel is impeded by the very slow folding of U_T to N_T, approximately three orders of magnitude slower than the corresponding U_C to N_C reaction (Table 1b).

On initiation of refolding from the U_C state, revealed by the double-jump refolding experiment, the refolding of I^{BP}_{C} to N_C via the U_C state becomes apparent. Because the burst-phase intermediate, I^{BP}_{C}, (relative to the U_C state) and the $U_C \rightarrow N_C$ transition state are nearly identical in the amount of surface area buried, the refolding reaction along the *cis*-channel becomes urea-independent below 2 M urea (Fig. 2.10b). As I^{BP}_{C} becomes destabilized at increasing urea concentrations, the urea dependence of the $U_C \rightarrow N_C$ transition state becomes apparent above 2 M urea (Fig. 2.10c). By implication the surface area buried by the transition state of the $U_C \rightarrow I^{BP}_{C}$ reaction is smaller than (and hence more readily accessible via conformational searching) that of the $U_C \rightarrow N_C$ transition.

The complex behavior observed for the unfolding of CheY (Fig. 2.5a) reflects the direct unfolding of the minor N_T conformer and the urea-sensitive partitioning of the major N_C conformer between the $N_C \rightarrow N_T$ isomerization reaction and the $N_C \rightarrow U_C$

Figure 2.10 – Folding mechanism of CheY. Mechanism for refolding (panel a, b and c) and unfolding (panels d, e and f) of CheY predicted by the off-pathway model at different concentrations of urea. The progress of the reaction is shown as thick arrows, while the reactions not accessible under the different conditions are represented by thin gray arrows. The rate limiting reactions are shown as broken and dotted lines, and the minor channels are shown as broken lines. **(a)** Refolding to 0 M urea from strongly denaturing conditions. The unfolded state with the K109 P110 bond in the *trans* isomer is the dominant population, which collapses rapidly to an off-pathway intermediate. Isomerization of the prolyl bond to the *cis*-state in this intermediate is the rate limiting step for completion of folding. The minor refolding phase along the *trans* channel is represented by the broken line. **(b)** Refolding along the *cis*-channel at 0 M urea (observed in the double jump experiments). The dominant unfolded form is the U_C state, which rapidly equilibrates with the $I^{BP}{}_C$ state. The refolding from U_C to N_C is the rate limiting reaction. **(c)** Refolding along the *cis*-channel at 2 M urea. The $I^{BP}{}_C$ state is destabilized, and refolding now proceeds directly from U_C to N_C. The minor unfolded state, U_T, represented by the broken lines, follows the same refolding path as described in panel a. **(d)** Unfolding reaction at 3 M urea. The U_C state is kinetically accessible; however, the equilibrium is driven towards N_T. Isomerization occurs primarily in the native state, prior to unfolding to U_T. **(e)** Unfolding reaction at 4 M urea. The destabilization of N_C causes rapid equilibration with U_C. Isomerization occurs both in the native and unfolded states. **(f)** Unfolding reaction under strongly denaturing conditions, 8 M urea. The native state is destabilized, and unfolding proceeds directly from N_C to $U_{C;}$ isomerization occurs only in the unfolded state. The small fraction of molecules that exist in the N_T state unfold directly to the U_T state, as represented by the broken line.

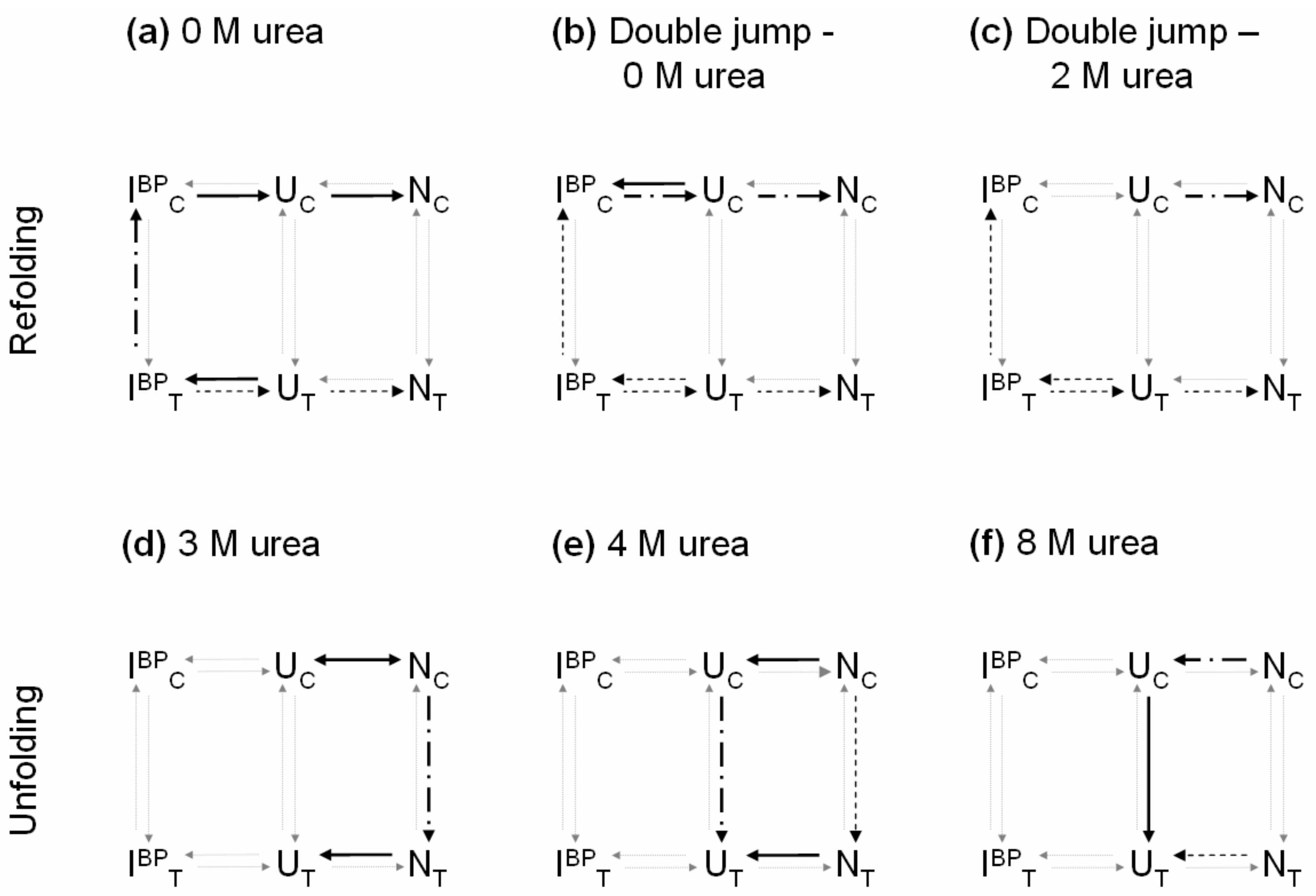

Refolding
(a) 0 M urea
(b) Double jump - 0 M urea
(c) Double jump – 2 M urea
Unfolding
(d) 3 M urea
(e) 4 M urea
(f) 8 M urea
$I^{BP}{}_{C}$
U_C
N_C
$I^{BP}{}_{T}$
U_T
N_T

Figure 2.10

unfolding reaction. The modest changes in buried surface area to reach the transition state for the $N_C \rightarrow N_T$ isomerization reaction and the substantial changes for the $N_C \rightarrow U_C$ and $N_T \rightarrow U_T$ unfolding reactions result in very different dependences of the associated rate constants on the urea concentration (Table 1b and Fig. 2.8e).

Under native conditions the K109 P110 bond is predominantly in the *cis* isomeric state (~80% N_C). Unfolding from these conditions to mild denaturant concentrations, 3 M urea, slightly shifts the equilibrium of the native confomers to the N_T species (Fig. 2.8f). The slow isomerization reaction is then followed by unfolding along the *trans* channel (Fig. 2.10d). Between 3 and 4 M urea, the observed relaxation time shifts from reflecting the urea-independent $N_C \rightarrow N_T$ reaction to the urea-dependent $N_T \rightarrow U_T$ unfolding reaction. Although the $N_C \rightarrow U_C$ reaction is faster than either of these competing reactions, CheY does not unfold significantly through the *cis* channel below 4 M urea because the N_C/U_C equilibrium favors the N_C state (the midpoint for the N_C/U_C equilibrium is 4.2 M urea). At higher urea concentrations, above 4 M urea, both the native states are destabilized, and the two observed phases reflect the unfolding of the N_T and N_C conformers in parallel reactions (Fig. 2.10e); both of these reactions are faster than the $N_C \rightarrow N_T$ isomerization reaction under these conditions. Under strongly unfolding conditions, above 5.5 M urea, the contribution along each channel corresponds to the initial ratio of *cis* and *trans* states (Fig. 2.10f). The model predicts that the unfolding rate constants of N_T and N_C are very similar under strongly denaturing conditions and the relaxation times of these two phases merge to produce a single observed kinetic phase that accelerates with increasing urea concentration. The interplay between an inter-channel isomerization phase and parallel unfolding reactions reflects the comparable magnitudes of their rate constants and their contrasting dependences on the urea concentration. The latter property undoubtedly reflects the structural consequences, i.e., changes in buried surface area, of a local conformational change, isomerization of a single peptide bond, and a global unfolding reaction.

Molecular basis for premature folding reactions in βα-repeat proteins

In a recent study of the folding of the alpha subunit of tryptophan synthase (αTS),[59] a βα-repeat protein of the TIM barrel class, the formation of a sub-millisecond off-pathway intermediate was attributed to a collapse around a tightly-packed hydrophobic cluster whose boundaries were not suitable docking sites for the further propagation of structure throughout the barrel. Off-pathway early intermediates have also been reported for two other TIM barrel proteins of very low sequence identity.[58,74] While the specific locations of the structure formed in these prematurely-folded intermediates vary between these proteins, the structures are dominated by hydrophobic clusters of isoleucine, leucine and valine (ILV) residues. It is noteworthy that TIM barrel proteins have a higher representation of ILV residues than other protein families;[101] however, the special role of ILV residues appears to be a function of their unique and similar chemical properties.

The crucial role for <u>b</u>ranched <u>a</u>liphatic <u>si</u>de <u>c</u>hains in stabilizing partially-folded states of TIM barrel proteins has been coined the BASiC hypothesis.[59] Its molecular basis resides in the uniquely unfavorable hydration free energy of aliphatic side chains[102,103] and in their capacity to exclude solvent from the underlying peptide bond.[104,105] In a recent simulation study by Liu *et al.*,[106] a dewetting transition has been observed to precede hydrophobic collapse at the subunit interface of the melittin tetramer. Each subunit of the melittin tetramer contributes 10 nonpolar side chains to the interface; all but a lone tryptophan are isoleucine (3), leucine (4) or valines (2). Relevant to the unfavorable hydration free energies of the aliphatic side chains, the process of dewetting is reported to take place in the manner of a liquid to vapor phase transition. This observation gains in significance when compared to a similar simulation on the docking of two domains in 2,3-dihydroxy-biphenyl dioxygenase.[106 108] By contrast with melittin, the hydrophobic composition of the interface of the dioxygenase domains is more diverse, incorporating aromatic side chains in addition to the BASiC residues, and the hydrophobic collapse occurs without a distinct dewetting transition.

It is noteworthy that the average volume of peptidic linkages reduces by ~10% (5.2 Å^2) in the absence of solvent.[104] Thus, the unfavorable free energy of hydration of ILV side chains would favor the formation of hydrophobic clusters and the resulting exclusion of solvent from the underlying backbone may serve to decrease the volume of the peptidic linkage. The decrease in volume of the backbone would, in turn, be expected to tighten the packing of the attached side chains and further decrease the access of solvent to the backbone. The synergy between backbone hydrogen bonding patterns and clusters of large aliphatic side chains provides a molecular mechanism for coupling the secondary and tertiary structure to produce highly cooperative reactions.

Two large ILV-dominated hydrophobic clusters are observed in CheY, one on either side of the central β-sheet (Fig. 2.11). An elaborate network of sequence-local

Figure 2.11 – Cluster analysis of CheY. (a) A ribbon diagram of the crystal structure of CheY reported by Voltz et al[90] ILV residues that bury greater than 10 Å^2 by contacting other ILV residues are highlighted, and the VDW surfaces of the heavy atoms of these residues are shown as dots. The two major clusters of such residues are colored red (Cluster 1) and blue (Cluster 2). **(b)** Contact map of ILV residues of CheY. The contacts in Cluster 1 are colored red and those in Cluster 2 are colored blue. The residue numbers are represented on both axes, and the corresponding elements of secondary structure are also shown. The dashed red box highlights the contacts in Cluster 1, which are closer in sequence than those in Cluster 2.

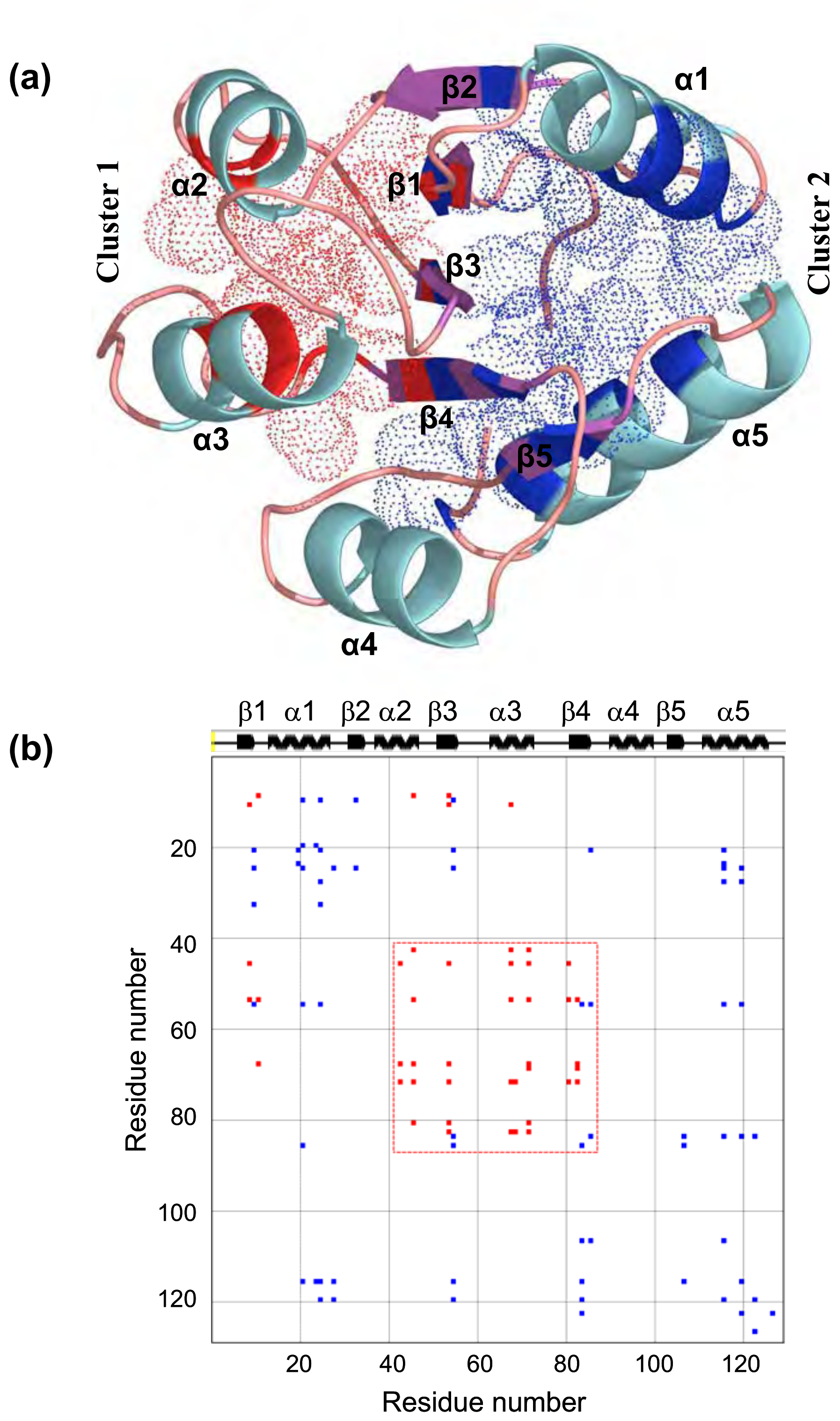

(a)
β2
α1
Cluster 1
α2
β1
β3
Cluster 2
α3
β4
α5
β5
α4
(b)
β1 α1 β2 α2 β3 α3 β4 α4 β5 α5
20
40
60
80
100
120
Residue number
20
40
60
80
100
120
Residue number

Figure 2.11

vdW interactions between 10 ILV side chains in β1, β3 and β4 and α2 and α3 comprise a tightly-packed ILV core; this collection of side chains is designated Cluster 1. Another network of vdW interactions between 15 ILV side chains on the opposing side of the β-sheet encompasses all five strands in the β-sheet and α1 and α5; this collection of side chains is designated Cluster 2. Detailed comparisons of these two clusters shows that Cluster 1 is dominated by well-packed ILV side chains (7 residues with buried surface area > 50 Å^2); the buried surface area is 607.9 Å^2 (see Materials and Methods for details). The 15 side chains in Cluster 2 bury 837.7 Å^2 of surface area, and 8 of these residues are deeply buried within the cluster (buried surface area > 50 Å^2). The remaining 7 ILV side chains in Cluster 2 are interspersed with phenylalanine and methionines, including the cluster of 4 phenylalanines proposed to be the source of the exciton coupling in native CheY. The perturbed exciton coupling in I^{BP}, however, suggests that the side chain packing in the larger Cluster 2 cannot achieve native like tertiary structure. A rationale for dominance of the smaller Cluster 1 in driving the off-pathway reaction requires an elaboration of the BASiC hypothesis.

The BASiC hypothesis in its original formulation[59] supposed that a cluster of 12 or more ILV side chains in a TIM barrel protein was required to drive its early folding reaction. The initial predictions of stable ILV-clusters did not quantitatively account for the chain entropy penalties entailed in forming clusters from side chains that are distal in the sequence. The more sequence-local contributions of residues in Cluster 1 and the more disperse contributions of residues for Cluster 2 motivated an examination of the potential role of chain entropy in the formation of ILV clusters.

Detailed inspection of these two clusters revealed:

1. The average surface area buried by a contact between two side chains within Cluster 1, 27.6 Å^2, is marginally greater than for Cluster 2, 26.2 Å^2. The average surface area buried per residue by the respective cluster is marginally higher in Cluster 1, 60.8 Å^2, as compared to 55.9 Å^2 in Cluster 2. The former metric represents more efficient packing in Cluster 1, and the latter metric represents the more compact nature of Cluster 1.
2. The residues involved in Cluster 1 are closer in sequence than those in Cluster 2 (Fig. 2.11b). This is readily apparent from the calculated Absolute Contact Order (ACO)[40,41] for the ILV residues in the two clusters. Cluster 1 has an ACO 21.64, reflecting sequence-local contacts, while Cluster 2 has an ACO 36.09, reflecting longer range contacts.

Cluster 1 thus packs more efficiently and with a smaller chain entropy penalty than Cluster 2. The good agreement between the elements of secondary structure involved in Cluster 1, β1, β3, β4, α2 and α3, and those identified as a kinetic trap by Gō-model simulation in the accompanying paper,[62] β3, β4, α2 and α3, suggests that the smaller ILV cluster is crucial to the off-pathway reaction. Propagation of structure through the β-sheet to Cluster 2 would provide a platform for the loose association of the peripheral helices. An intermediate with similar properties has been reported by Wu *et al.*[59] for αTS. The partial unfolding of this intermediate for CheY would then correspond to the disruption of structure in the two prematurely-formed clusters of nonpolar side chains. The native-centric nature of Gō-like simulations presumably highlight the central core of Cluster 1, as reported in the accompanying paper,[62] but not the looser and possibly non-native associations in the remaining structure. The altered packing of the quartet of phenylalanines at the periphery of Cluster 2 and the lower m-value of I^{BP} are consistent with this hypothesis and a molten globule-like-structure for the second cluster. A mutational analysis of the branched nonpolar side chains in both hydrophobic clusters is required to test the relative packing densities of the two clusters in I^{BP}.

Overview on off-pathway folding reactions in proteins

Recent all-atom, non-Gō-like simulations on the refolding of three-helix bundle proteins[109,110] points to the possibility that formation of off-pathway intermediates during refolding may be a common feature of other classes of proteins. Disruption of interactions in these states is required for the proper folding to the native state. The mechanism by which these intermediates form and the extent of specific structure they contain may be a function of the topological and energetic frustration of the particular protein.

The observations of fast-forming off-pathway folding intermediates in a pair of proteins with the flavodoxin fold, apo-flavodoxin itself[35] and CheY, and three TIM barrel proteins of very low sequence identity, αTS,[34] sIGPS[58] and IOLI,[74] imply a similar mechanism for these diverse members of the βα-repeat family. The local-in-sequence/local-in-space nature of the βα-repeat motif would favor the conformational search-limited formation of stable hydrophobic clusters of nonpolar side chains formed by the association of consecutive βα hairpins. Either by prematurely forming native-like structure as demonstrated in the accompanying paper[62]or by having a densely-packed native-like core with loosely packed boundaries,[59] these sub-millisecond intermediates are sufficiently stable that they cannot anneal to progressively form the native conformation. Rather, at least partial unfolding of these species appears to be required to initiate a productive folding reaction.

Materials and Methods

Protein expression and purification

The expression plasmid (pET 21) with the gene encoding CheY was obtained from David Wemmer at UC Berkeley; the DNA sequence was confirmed at the UC Davis sequencing facility. The *Escherichia coli* strain BL21 Codonplus®(DE3)RIL was used for expression, and the method described by Filimonov *et al.*[88] was used for purification. Further purification was done using a Sephadex® G-75 gel filtration column in 10 mM potassium phosphate at pH 7.0. The identity and purity (> 98%) was confirmed using nano-spray mass-spectrometry at the Proteomics Facility at UMMS, Worcester. An extinction coefficient of 8250 M ^{1}cm ^{1} was used (as determined by Filimonov *et al.*,[88] by the method of Gill and von Hippel[111]) to determine the protein concentration.

Stability analysis

Samples of 5 μM protein in 10 mM potassium phosphate at pH 7.0 were equilibrated over-night with 0 M to 8 M urea at concentration increments of 0.2 M urea. The far-UV CD and the steady-state tryptophan fluorescence emission spectra of each sample at 25 °C using a 1cm cuvette in a Peltier-style thermostatted sample compartment were recorded on a JASCO model J810 CD spectrophotometer and a T-format Horiba fluorolog fluorimeter, respectively. The CD spectra were recorded between 205 nm and 260 nm, with a band width of 2.5 nm, and a step size of 0.5 nm, integrated for 1 s and averaged over three traces. The emission spectra after excitation at 290 nm were recorded between 295 nm and 500 nm at a 1 nm interval and averaged over three traces. The measurements were repeated twice and the reversibility of the reaction was confirmed by coincidence of the equilibrium transition curve obtained by starting from the unfolded state in 8 M urea.

After buffer correction, the transition curves at 222 nm for CD signal change and 315 nm for change in fluorescence emission were plotted as a function of urea concentration and fitted to a two state model,

$$\mathrm{N} \leftrightarrows \mathrm{U}$$

where N is the native form of the protein and U is its denatured form. The free energy change associated with unfolding in the absence of denaturant was determined by assuming a linear dependence of the apparent free-energy change on the denaturant concentration,[112,113]

$$\Delta G^{\circ}_{[\mathrm{Urea}]} = \Delta G^{\circ}_{H_2O} - m[\mathrm{Urea}] = -RT\ln\left(K_{\mathrm{eq}[\mathrm{Urea}]}\right) \tag{2.1}$$

where $\Delta G^{\circ}_{(H_2O)}$ is the standard unfolding free energy change in the absence of urea, $\Delta G^{\circ}_{[\mathrm{Urea}]}$ is the standard unfolding free energy change at any urea concentration [Urea], and m is its dependence on the concentration of urea.[112,113] A nonlinear regression analysis module of the software Savuka[34] was used to fit the data to this model. Fitting to a three-state model did not improve the fit significantly. The two-state fit was confirmed by globally fitting the fluorescence and CD data across all wavelengths using singular value decomposition (SVD) vectors (for description, see Ionescu *et al.*[114] and Gualfetti *et al.*[115] and references therein). No significant contribution was observed after the second vector.

The occurrence of the *cis* isomer of the K109 P110 bond (~10%) in the unfolded state motivated the fit of the data to a three-state model

$$\mathrm{N} \leftrightarrows \mathrm{U_C} \leftrightarrows \mathrm{U_T}$$

where the subscripts refer to the conformational state of the K109 P110 peptide bond. The equilibrium ΔG° between the unfolded states was kept fixed at 1.36 kcal mol ^{1} (corresponding to the U_T/U_C 90/10 ratio observed in the peptide model) and the m-value between the unfolded states was fixed at 0 kcal mol ^{1}M ^{1} because isomerization of the peptide bond in the unfolded state is not expected to be associated with surface area burial.

NMR

NMR experiments were performed using a Varian Unity INOVA 600, operating at a ^{1}H frequency of 599.7 MHz and equipped with an inverse triple-resonance cryogenically-cooled probe and preamplifier. The samples were 2.5 mM solutions of the Ac V108 K109 P110 F111 T112 NH_2 penta-peptide, obtained from New England Peptides Inc., dissolved in 10 mM potassium phosphate buffer at pH 7.0 containing 10% D_2O. Spectra were obtained using a simple pulse-acquire sequence with excitation sculpting for solvent suppression.[116] 16,000 complex data points were recorded over a spectra width of 8 kHz. All data were apodized with 0.5 Hz exponential line broadening prior to Fourier transformation. All data were processed using the NMRpipe software package.[117]

Kinetics

Tryptophan fluorescence

The change in fluorescence emission associated with refolding or unfolding was monitored using an Applied Photophysics SX 17MV instrument (dead time 2 ms). The excitation wavelength was 290 nm and emission was monitored using a 320 cut-off filter. The relaxation times and the amplitudes associated were calculated by fitting the kinetic data to the equation

$$A(t) = A(\infty) + \sum_{i=1}^{n} A_i \exp\left(-t/\tau_i\right) \qquad \textbf{(2.2)}$$

where $A(\infty)$ is the observed signal at infinite time, $A(t)$ is the observed signal at time t, A_i is the signal and τ_i is the relaxation time associated with phase (i) and n is the number of exponentials. The kinetic data were fit to a series of exponentials using a non-linear least squares fitting program, Savuka.[34] The logarithm of the relaxation time was plotted as a function of final denaturant concentration.

Refolding: The protein was equilibrated overnight in 6.5 M urea and 10 mM potassium phosphate at pH 7.0 and at 25 °C and refolded by rapid mixing into refolding buffer and varying final concentrations of denaturant (1 M urea to 4 M urea), and 8.5 μM protein.

Unfolding: The protein was unfolded by rapid mixing into high concentration of urea buffered with 10 mM potassium phosphate at pH 7.0 and 25 °C, to final urea concentrations ranging from 2.5 M to 8.3 M, and 8.5 μM protein.

Double jump refolding kinetics: The protein was denatured by rapid mixing with urea buffered in 10 mM potassium phosphate at pH 7.0 and 25 °C. After denaturing in 6.5 M urea for varying lengths of time (1 s to 100 s) in a 118 μl delay line, the protein was refolded by rapid dilution into buffer to a final protein concentration of 5 μM and 1.1 M urea. A dead time of 50 ms was used to compensate for mixing artifacts from the aging loop. The kinetic data were fit to exponentials to determine the relaxation times by the method described above, and the amplitude associated with the slow relaxation time was plotted as a function of the delay time. The decay in signal amplitude with respect to delay time was then fit to exponentials. The experiment was also performed with refolding jumps to varying concentrations of urea to a final concentration of 1.1 M to 4.5 M. The unfolding was carried out at 6.5 M urea and the delay time was kept constant at 2 s. The relaxation times and associated amplitudes were determined as described earlier and plotted as a function of final urea concentration.

Activation energy of the slow refolding phase: Refolding of denatured protein (6.5 M urea) was repeated at different temperatures ranging from 13 °C to 39 °C and the fluorescence emission at 315 nm was monitored after excitation at 290 nm on a T-format Horiba fluorolog fluorimeter. The initial conditions were identical to those described for the refolding kinetics; the protein was incubated at the respective temperature for one hour before the experiment and rapidly mixed into refolding buffer in a 1cm quartz cuvette in a Peltier-style thermostatted sample compartment. The final concentration of protein was 4 μM at 1 M urea. The relaxation times and associated amplitudes for each kinetic trace were determined by the method described above. The natural logarithm of the inverse of the slow relaxation time was plotted as a function of the inverse of the temperature and fit to the Arrhenius equation,[118]

$$\ln k = \ln A - \frac{E_a}{RT} \tag{2.3}$$

where the rate constant k 1 / τ (the relaxation time), ln A is the y-intercept, and E_a is the activation energy for the reaction. The fitting was performed using Savuka.[34]

Circular dichroism

Refolding and unfolding were also monitored by the far-UV CD ellipticity at 222 nm using an AVIV model 202 stopped-flow CD spectrophotometer (dead time 5 ms) and a JASCO model J810 CD spectrophotometer (manual mixing dead time ~10 s). The conditions for the experiments were as described above, and the data were fitted by the same method used for fluorescence emission.

Effect of Cyclophilin on the slow refolding phase: The refolding experiments were repeated by rapid mixing of denatured protein (6.5 M urea) into refolding buffer (10 mM potassium phosphate at 15 °C and pH 7.0) containing varying amounts of cyclophilin to provide final concentrations of 0 μM to 2.2 μM. The final protein concentration was 8.5 μM in 1 M urea, 10 mM potassium phosphate, and at 15 °C and pH 7.0. The relaxation times and related amplitudes were determined by the method described above and plotted as a function of cyclophilin concentration.

Stability of burst-phase intermediate: The refolding kinetic traces monitored by CD at 222 nm were buffer-corrected and extrapolated to 0 s to determine the signal associated with the burst-phase intermediate. The amplitude of the signal was then plotted against the final denaturant concentration. The sigmoidal unfolding curve was then fitted to a two-state model

$$I^{BP} \leftrightarrows U$$

where I^{BP} is the burst-phase intermediate and U is the denatured form. The free energy change associated with the unfolding of the intermediate in 0 M urea and its dependence on urea concentration were determined by the method described above.

CD spectrum of the burst-phase intermediate: Refolding was induced by rapid mixing to strongly native conditions (1 M urea), and the kinetics monitored at wavelengths from 205 nm to 260 nm at intervals of 1 nm. The initial conditions, the final protein concentration, the buffer concentration, the temperature and pH were the same as for the refolding kinetics. Each kinetic trace was then extrapolated back to 0 s, and the signal was plotted as a function of the wavelength. The signal at infinite time was also plotted as a function of the wavelength (data not shown) to confirm its coincidence with that of the native protein.

Global analysis

Both the CD and the fluorescence kinetic traces from the refolding, unfolding and double jump experiments were fit globally to several different models. For every model with *n* number of species, a $n \times n$ systems matrix of the form

$$\mathbf{S}^0 = \begin{pmatrix} -\sum_{\mathrm{j=1}}^{\mathrm{n}} k_{1j}^0 & k_{21}^0 & \cdots & k_{n1}^0 \\ k_{12}^0 & -\sum_{\mathrm{j=1}}^{\mathrm{n}} k_{2j}^0 & \cdots & k_{n2}^0 \\ \vdots & \vdots & \cdots & \vdots \\ k_{1n}^0 & k_{2n}^0 & \cdots & -\sum_{\mathrm{j=1}}^{\mathrm{n}} k_{nj}^0 \end{pmatrix}$$

was constructed, where $k°_{\mathrm{ij}}$ is the microscopic rate of formation of species *j* from species *i* in the absence of urea. The microscopic rates between species that were not connected directly in the model, and $k°_{\mathrm{ii}}$ were set to 0. The microscopic rate constants are rendered urea dependent by the following equation

$$k_{\mathrm{ij}} = k_{\mathrm{ij}}^0 \exp\left(\frac{-m_{\mathrm{ij}}[\mathrm{Urea}]}{RT}\right) \tag{2.4}$$

where m_{ij} is the urea dependence of the microscopic rate and k_{ij} is the microscopic rate in the presence urea. Similarly the matrix $\mathbf{S}^0$ is rendered urea dependent by replacing matrix elements by those calculated at the given urea concentration to obtain a matrix $\mathbf{S}^{[\mathrm{Urea}]}$.

The time dependence of the species concentration was expressed as follows

$$\frac{\mathrm{d}}{\mathrm{d}t}\vec{\mathrm{c}} = \mathbf{S}^{[\mathrm{Urea}]} \cdot \vec{c} \tag{2.5}$$

where $\mathbf{S}^{[\mathrm{Urea}]}$ is the systems matrix generated for the final concentration of urea, and $\vec{c}$ is a column vector of the form

$$\vec{c} = \begin{pmatrix} c_1 \\ c_2 \\ c_i \\ \vdots \\ c_n \end{pmatrix}$$

where c_i is the concentration of the i^{th} species. The concentration of all species at time t was determined by integrating equation 5 to yield:

$$\vec{c}(t) = \exp\left(\mathbf{S}^{[\text{Urea}]} \cdot t\right) \vec{c}^{\,0} \tag{2.6}$$

where $\vec{c}(t)$ is a column vector of the concentration of each species at time t, $\vec{c}^{\,0}$ is a vector of the initial concentrations of each species, and $\mathbf{S}^{[\text{Urea}]}$ is the systems matrix generated for the final concentration of urea. The matrix exponential exp($\mathbf{S}\cdot t$) was solved using Eigenvalue decomposition:

$$\exp(\mathbf{S} \cdot t) = \mathbf{V} \exp(\Lambda \cdot t) \mathbf{V}^{-1} \tag{2.7}$$

where **V** is the matrix of Eigenvectors of **S** with associated Eigenvalues in the diagonal matrix **Λ**. $\mathbf{V}^{1}$ was determined using singular value decomposition of **V**. The initial concentration matrix $\vec{c}^{\,0}$ is solved by applying the law of mass conservation in the equation

$$\vec{c}^{\,0} = \mathbf{S}_{\text{eq}}^{-1} \cdot \vec{c}_{\text{eq}} \tag{2.8}$$

where $\mathbf{S}_{\text{eq}}$ is the systems matrix generated for the initial concentration of urea with the first row elements set to 1, and $\vec{c}_{\text{eq}}$ is a column vector with n rows and all elements except the first set to 0. The first element is set to the sum of the concentrations of all species, which should equal the total protein concentration. For the double jump experiments the $\vec{c}^{\,0}$ for the refolding jump was replaced with the $\vec{c}(t)$ at the end of the unfolding jump, where t is the delay time.

The observed rates for such a system are determined by Eigenvalue decomposition of the system matrix

$$\mathbf{SV} = \lambda \mathbf{V} \tag{2.9}$$

where the vector of Eigenvalues, λ, represents the macroscopic rates (-λ) and the matrix of Eigenvectors, **V**, may be used to determine the amplitude, A_i, associated with the i^{th} macroscopic rate k_i as follows

$$A_i = \mathbf{V}\Delta\mathbf{V}^{-1}\,\mathrm{E}\,\vec{c}^{\,0} \tag{2.10}$$

where Δ is a zero matrix with only the (*i*,*i*) element set to 1, $\vec{c}^{\,0}$ is a vector of the initial concentration of each species and E is a matrix of the optical property associated with each species.

The experimentally determined equilibrium and kinetic parameters were used as starting points for the microscopic rate constants and their associated urea dependence. The starting values of all parameters were further refined by visually fitting the modeled macroscopic rates (Eigenvalues) and the associated relative amplitudes to the experimental data.

The kinetic traces were then reconstructed using the equation

$$\mathrm{Y}(t) = \sum_{\mathrm{i=1}}^{\mathrm{n}}((\varepsilon_{\mathrm{i}} + (\mathrm{m_i}\,[\mathrm{Urea}])) \times c_{\mathrm{i}}(t)) \tag{2.11}$$

where Y(*t*) is the total signal at time *t*, ε_i is the optical property of the i^{th} species at 0 M urea, m_i is the urea dependence of the optical property, and $c_i(t)$ is the concentration of the i^{th} species at time *t*. The optical property of each species was linked across the kinetic traces as a normalized value, Z, relative to the optical properties of the first and the last species in the model

$$Z_i = \frac{\varepsilon_i - \varepsilon_1}{\varepsilon_n - \varepsilon_1} \qquad \textbf{(2.12)}$$

where Z_i is the normalized Z value of the i^{th} species, ε_1, ε_i and ε_n are the optical properties of the 1^{st}, i^{th} and n^{th} species respectively.

The Levenberg-Marquart method[97] was then used to obtain the best fit to the kinetic data. The agreement with experimental data was determined using the reduced Chi-square statistic.

The equilibrium concentrations of each species at concentrations ranging from 0 to 10 M urea were also calculated using the above method, and the total signal at every urea concentration was determined. The deviation of this equilibrium model from the experimental data was then determined by calculating the reduced chi-square statistic.

Analysis of hydrophobic clusters

The contact surface area between atoms was calculated using the CSU software developed by Sobolev *et al.*[119] The contacts between side chain carbon atoms of the BASiC residues, Isoleucine, Leucine and Valine residues (ILV) were selected for the analysis. The total surface area buried by each pair of ILV residues, was calculated as the sum of the individual contributions from the side chain atoms. ILV residues that buried at least 10 Å^2 of a given ILV residue were considered to be spatially contiguous with it. A list of such ILV residues was used to define two BASiC hydrophobic clusters of spatially contiguous residues. The total surface area buried by ILV residues for each ILV side chain was used to calculate the total and average surface area buried by each residue in the two clusters. Pair-wise measurements were used to calculate the average surface area buried by contacts between two side chains in each cluster. The same measurement was used to determine the Absolute Contact Order[41] (ACO) for each cluster by the following equation,

$$\mathrm{ACO} = {}^{1}\!/_{\mathrm{n}} \times \sum_{1}^{\mathrm{n}} \mathrm{d}_{\mathrm{pq}} \tag{2.13}$$

where d_{pq} is the sequence distance between ILV residues p and q that are in contact with each other and n is the number of ILV contacts in each cluster.

Acknowledgements

The authors would like to thank Dr. David Wemmer for generously sharing the expression plasmid with the gene encoding CheY, Dr. Charles Brooks III and Mr. Ronald D. Hills Jr. for insightful discussions and help with understanding the results of their Gō-modeling simulations, Dr. Jill Zitzewitz for valuable discussions and help in preparation of the manuscript, and Dr. Osman Bilsel for his insights into the global analysis and for help with development of the Chevron Fitter. This work was supported by the National Science Foundation through grants MCB0327504 and MCB0721312 to C.R.M.

Chapter III – Topological Frustration in βα-Repeat Proteins: Sequence Diversity Modulates the Conserved Folding Mechanisms of α/β/α Sandwich Proteins

This chapter has been published previously as *Hills RD Jr, Kathuria SV, Wallace LA, Day IJ, Brooks CL 3rd, Matthews CR. "Topological frustration in beta alpha-repeat proteins: sequence diversity modulates the conserved folding mechanisms of alpha/beta/alpha sandwich proteins." J Mol Biol. 2010 Apr 30;398(2):332-50.*

The work presented in the following chapter was a collaborative effort. Dr. Ronald D. Hills Jr. performed and analyzed the Go-simulations, Dr Louise A. Wallace carried out the equilibrium unfolding experiments, the unfolding and refolding kinetic experiments and the experiments on cyclophilin dependence of refolding for both NT-NtrC and Spo0F. Dr. Iain J. Day carried out the NMR experiments. I repeated the equilibrium and kinetic experiments on Spo0F and performed the global analysis on both NT-NtrC and Spo0F. Dr. C. Robert Matthews, Dr. Charles L. Brooks III, Dr. Ronald D. Hills Jr. and I wrote the manuscript.

Abstract

The thermodynamic hypothesis of Anfinsen postulates that structures and stabilities of globular proteins are determined by their amino acid sequences. Chain topology, however, is known to influence the folding reaction, in that motifs with a preponderance of local interactions typically fold more rapidly than those with a larger fraction of non-local interactions. Together, the topology and sequence can modulate the energy landscape and influence the rate at which the protein folds to the native conformation. To explore the relationship of sequence and topology in the folding of $\beta\alpha$ repeat proteins, which are dominated by local interactions, a combined experimental and simulation analysis was performed on two members of the flavodoxin-like, $\alpha/\beta/\alpha$ sandwich fold. Spo0F and the N-terminal receiver domain of NtrC (NT-NtrC) have similar topologies but low sequence identity, enabling a test of the effects of sequence on folding. Experimental results demonstrated that both response-regulator proteins fold via parallel channels through highly structured sub-millisecond intermediates before accessing their *cis* prolyl peptide bond-containing native conformations. Global analysis of the experimental results preferentially places these intermediates off the productive folding pathway. Sequence-sensitive Gō-model simulations conclude that frustration in the folding in Spo0F, corresponding to the appearance of the off-pathway intermediate, reflects competition for intra-subdomain vdW contacts between its N- and C-terminal subdomains. The extent of transient, premature structure appears to correlate with the number of isoleucine, leucine and valine (ILV) side-chains that form a large sequence-local cluster involving the central β-sheet and helices $\alpha 2$, $\alpha 3$ and $\alpha 4$. The failure to

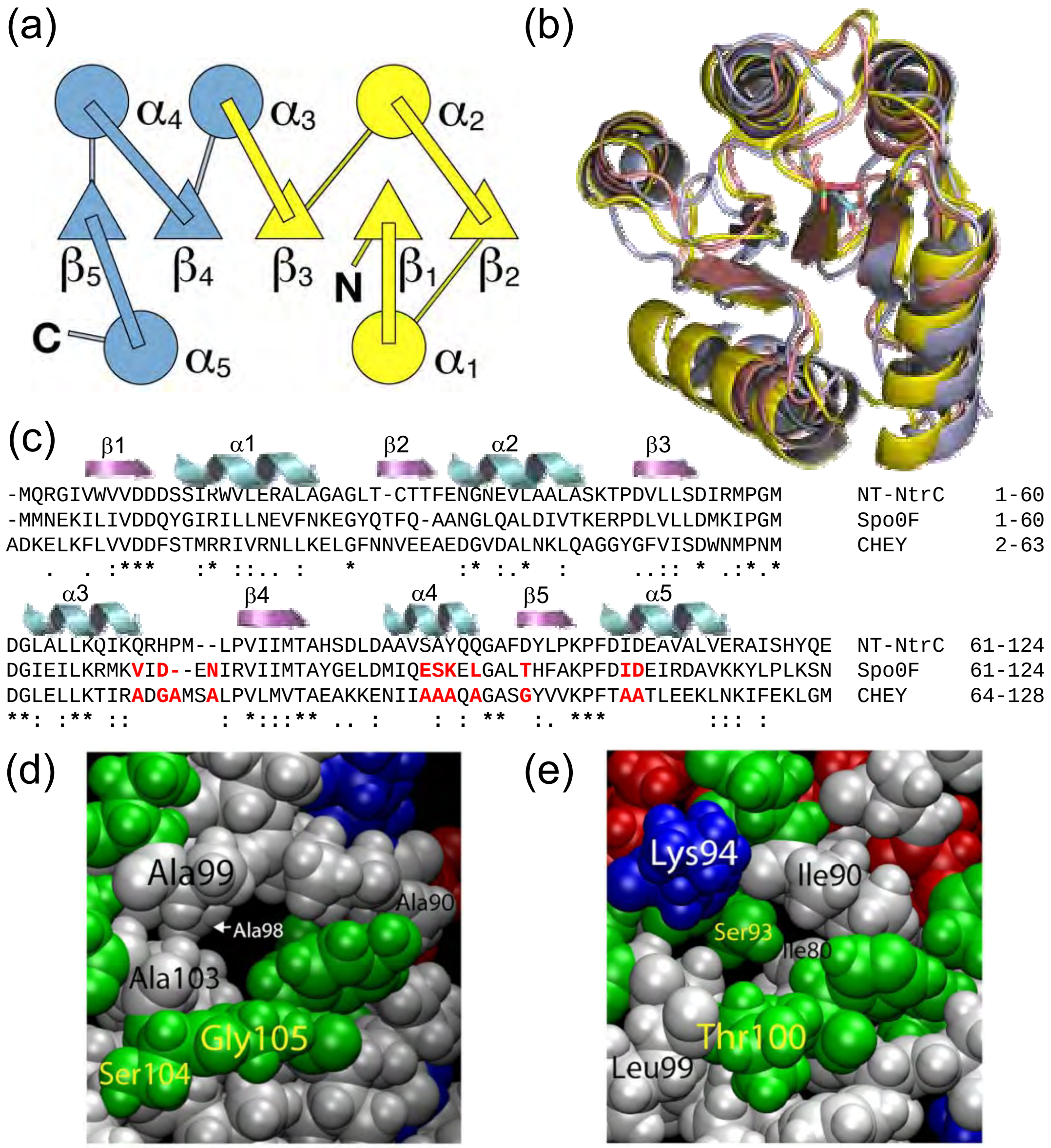
detect the off-pathway species in the simulations of NT-NtrC may reflect the reduced number of ILV side-chains in its corresponding hydrophobic cluster. The location of the hydrophobic clusters in the structure may also be related to the differing functional properties of these response regulators. Comparison with the results of previous experimental and simulation analyses on the homologous CheY argues that prematurely-folded unproductive intermediates are a common property of the βα-repeat motif.

Introduction

Although it is well accepted that the native conformation of a protein represents its global free energy minimum,[1] an understanding of the dynamic process by which the sequence is decoded into its three-dimensional structure on a biologically-feasible time scale remains elusive. Landscape theory[3] posits the view that sequences have evolved not only to be stable but also to have the capacity to rapidly and efficiently access the native conformation via a funnel-shaped energy surface biased towards the formation of the native structure. The explicit role for conformational entropy in determining the shape of the energy surface and correlations between folding rate constants and metrics for chain topology[40,41,45] argue for the importance of topology in folding reactions. However, the role of the sequence remains evident in the results of mutational analyses on dozens of proteins,[98] where single amino acid replacements can significantly alter the stability and the folding kinetics. Also, structural homologs with diverse sequences have been observed to fold at very different rates[57,120] or via different mechanisms.[121] Thus, deciphering the folding information contained in the amino acid sequence of a protein remains a major challenge in biophysics.

We have adopted a combined experimental and computational approach towards the elucidation of the relative contributions of the sequence and the topology to the folding mechanisms of three members of the CheY-like family of response-regulator proteins: the bacterial chemotaxis protein CheY from *Escherichia coli*,[90] the N-terminal receiver domain of nitrogen regulation protein NtrC from *Salmonella typhimurium* (NT-NtrC)[122] and the sporulation response regulatory protein Spo0F from *Bacillus subtilis*.[123] These small repeat-structure proteins, $(\beta\alpha)_5$, typically contain ~125 amino acids arranged as a $\alpha/\beta/\alpha$ sandwich. The five β-strands form a central parallel β-sheet, β2β1β3β4β5, with helices α1 and α5 docking on one face of the β-sheet and helices α2, α3 and α4 docking on the opposing face (Fig. 3.1a). The pair-wise RMSD values for CheY:NT-NtrC, CheY:Spo0F and NT-NtrC:Spo0F are 2.57 Å, 1.85 Å and 2.44 Å (Fig. 3.1b), respectively, and the pair-wise sequence alignment scores calculated using ClustalW[124] are 30%, 25% and 33%, respectively (Fig. 3.1c). Based upon the results of previous studies on other folds with similar structures but very different sequences, [57,74,75,120,121,125,126] we hypothesize that common features in the folding mechanisms will reflect the topology while the differences in the mechanisms and the perturbations in the kinetic and thermodynamic properties will reflect the variable sequences.

Figure 3.1 – Sequence and structural homology of CheY-like proteins. (a) Topology of CheY-like proteins. The central β-sheet comprises 5 parallel β strands in the order β2β1β3β4β5 and forms an α/β/α sandwich with helices α1 and α5 on one face of the sheet and helices α2, α3 and α4 on the other. The N-terminal (yellow) and C-terminal (blue) folding subdomains of CheY[127] are comprised of β1α1β2α2β3 and α3β4α4β5α5, respectively. **(b)** Structural alignment of NT-NtrC (blue), Spo0F (red) and CheY (yellow). The pair-wise RMSD values for CheY:NT-NtrC, CheY:Spo0F and NT-NtrC:Spo0F are 2.57 Å, 1.85 Å and 2.44 Å respectively. The catalytic aspartic acid residue, 54 in NT-NtrC, 54 in Spo0F and 57 in CheY, is at the beginning of the loop connecting the two subdomains and is shown as sticks. The PDB codes used were NT-NtrC: 1DC7, [122] Spo0F: 1SRR[123] and CheY: 3CHY.[90] **(c)** Sequence alignment of NT-NtrC, Spo0F and CheY using ClustalW.[124] The pair-wise sequence alignment scores for CheY:NT-NtrC, CheY:Spo0F and NT-NtrC:Spo0F are 30%, 25% and 33%, respectively. The elements of secondary structure are indicated above the aligned sequences, and the sequence conservation is indicated below, (*****) identical, (**:**) conserved and (**.**) semi-conserved. Residues highlighted in red font denote either C-terminal alanines and glycines in CheY or corresponding residues in Spo0F with bulkier side-chains. **(d)** An alanine-rich cavity resides between α4 and β4β5 in the inactive CheY structure. **(e)** The same region in Spo0F is filled with bulkier residues.



```
-MQRGIVWVVDDDSSIRWVLERALAGAGLT-CTTFENGNEVLAALASKTPDVLLSDIRMPGM      NT-NtrC     1-60
-MMNEKILIVDDQYGIRILLNEVFNKEGYQTFQ-AANGLQALDIVTKERPDLVLLDMKIPGM      Spo0F       1-60
ADKELKFLVVDDFSTMRRIVRNLLKELGFNNVEEAEDGVDALNKLQAGGYGFVISDWNMPNM      CHEY        2-63
   .  . :***   :* ::.. :   *       :* :.*  :    ..:: * .:*.*

DGLALLKQIKQRHPM--LPVIIMTAHSDLDAAVSAYQQGAFDYLPKPFDIDEAVALVERAISHYQE  NT-NtrC   61-124
DGIEILKRMKVID--ENIRVIIMTAYGELDMIQESKELGALTHFAKPFDIDEIRDAVKKYLPLKSN  Spo0F     61-124
DGLELLKTIRADGAMSALPVLMVTAEAKKENIIAAAQAGASGYVVKPFTAATLEEKLNKIFEKLGM  CHEY      64-128
**: :** ::       : *:::** .. :   : : **  :. ***        ::: :
```

Figure 3.1

As a basis for this comparative analysis, a recent experimental study of CheY[37] found that folding initiates with the appearance of an off-pathway partially-folded state in the sub-millisecond time range. This kinetically-trapped species must at least partially unfold before the protein can access the productive transition state ensemble (TSE) and fold to the native conformation. The companion coarse-grained Gō-simulation[62] came to a similar conclusion and predicted that the premature folding of the α2β3α3β4 tetrad towards the C-terminus of CheY was responsible for the kinetic trap. A sequence-local cluster of isoleucine, leucine and valine side-chains in precisely the same C-terminal segments identified in the Gō-model simulations was hypothesized to provide the core of stability in the off-pathway folding intermediate.[37] A previous Gō-model simulation of CheY reported the premature folding of the five α-helices, but did not describe a role for the β-strands.[77] The dissipation of structure in the α2β3α3β4 tetrad in the kinetically-trapped species allowed the formation of the productive TSE involving the same N-subdomain containing the β1α1β2α2β3 elements of secondary structure identified in the mutational analysis of CheY (Fig. 3.1a).[127] The C-subdomain containing the α3β4α4β5α5 elements of secondary structure is unstructured in this TSE (Fig. 3.1a).

The variations in sequence for these three proteins also allow an exploration of the relationship between the structural characteristics of their folding reactions and their functional properties. Phosphorylation of an aspartic acid residue, D57, in the loop connecting β3 and α3 of CheY causes the α4β5 surface to undergo a conformational rearrangement that enable binding of CheY to its downstream target, the flagellar motor protein, FliM.[128 130] The flexibility of the α4β5 surface of the protein has been attributed to the lack of a strong N-capping residue in α4 and an alanine-lined cavity[127,131] (Fig. 3.1d) between this helix and the rest of the protein.[132,133] Analogous to the conformational dynamics of CheY, phosphorylation of D54 in NT-NtrC induces a structural rearrangement in β5 and α4 of the receiver domain that transmits the signal to the C-terminal DNA-binding domain.[122,134] NMR relaxation measurements and atomistic molecular dynamics simulations have identified flexibility in α4 in the inactive state,[135,136] and results of a recent combined NMR and X-ray analysis,[137] support a population-shift activation mechanism as has also been suggested for CheY.[128,130,138,139] By contrast to CheY and NT-NtrC, the most significant conformational rearrangement in Spo0F during phosphorylation of D54 occurs in α1. This rearrangement enables Spo0F to interact directly with its immediate downstream partner in the phosphor-relay signaling pathway, Spo0B.[123,140 142] Unlike CheY and NT-NtrC, the cavity between helix α4 and the β-sheet of Spo0F is filled with bulky side-chains, some of which participate in a large hydrophobic cluster (Fig. 3.1c, 3.1d and 3.1e).

The complementary insights into the energetic and structural aspects of the folding free energy surfaces for NT-NtrC and Spo0F, provided by a combined experimental and computational analysis, reveal a significant role for the sequences in modulating their folding reactions. Prematurely-folded intermediates, stabilized by local-in-sequence clusters of aliphatic side-chains[143] appear to be a common feature in the folding of CheY-like proteins.

Results

Experimental Analysis

Equilibrium folding reactions

Equilibrium unfolding free-energy surface: Far-UV CD and fluorescence spectroscopy (FL) were employed to monitor the loss of secondary and tertiary structure of NT-NtrC and Spo0F in the presence of the chemical denaturant urea (Fig. 3.2). The urea-induced equilibrium unfolding transitions, monitored by CD at 222 nm and by FL emission at 315 nm for NT-NtrC and 305 nm for Spo0F, show single sigmoidal transitions for both proteins (Fig. 3.3a and 3.3d). The normalized CD and FL equilibrium transition curves are coincident within error (Fig. 3.2c and 3.2f), and the reversibility of the urea denaturation reaction was demonstrated by the coincidence of the unfolding and refolding CD transitions (Fig. 3.3a and 3.3d). Assuming a two-state equilibrium model for the unfolding reaction, a global analysis of the CD and FL spectral changes with urea concentration yielded values for the Gibbs free energy of unfolding from the native, N, to the unfolded, U, state in the absence of urea, ΔGº (H_2O), the dependence of ΔGº on the denaturant concentration, the *m*-value, and the mid-point of the transition, Cm, as shown in Table 3.1. For comparison, the ΔGº (H_2O), the *m*-value and the C_m for a two-state fit of the urea-induced equilibrium unfolding reaction for CheY are also shown.[37]

Kinetic folding reactions

Burst-phase reaction: The formation of secondary structure during the refolding of NT-NtrC and Spo0F induced by stopped-flow mixing methods was monitored by the changes in ellipticity at 222 nm. Over 50% (66% for NT-NtrC and 52% for Spo0F) of

Figure 3.2 – CD and FL spectra of NT-NtrC and Spo0F. **(a)** The CD spectrum of the native conformation of NT-NtrC (continuous line) is typical of βα-repeat proteins, with the characteristic α-helical minima at ~208 and ~222 nm modulated by the β-sheet minimum at ~218 nm. The far-UV CD spectrum measured in the presence of 8 M urea (dotted line) resembles that expected for a space-filling random coil, indicating a global disruption of structure upon denaturation. **(b)** The FL emission spectra of native (continuous line) and denatured NT-NtrC in the presence of 8 M urea (dotted line) are dominated by the contribution of two tryptophan residues; W7 and W17. The red shift in the emission maximum from 346 to 357 nm and the reduced intensity upon unfolding reflects the solvent exposure of the buried tryptophan residues. **(c)** Apparent fraction unfolded (Fapp) of NT-NtrC as a function of denaturant concentration. The unfolding transitions monitored by CD at 222 nm (○) and by tryptophan FL 315 nm (□) are shown. The confidence limits for the global fit of the combined CD and FL data to a two-state model are also shown (broken and dotted lines). **(d)** The CD spectrum of the native conformation of Spo0F (continuous line) is similar to that of NT-NtrC. Global disruption of structure is observed in Spo0F upon denaturation in the presence of 8 M urea (dotted line). **(e)** In the absence of an intrinsic tryptophan residue, the FL emission spectrum of Spo0F (continuous line) is dominated by the contribution of four tyrosine residues; Y13, Y28, Y84 and Y118. The quenching of the emission peak at 305 nm with increasing urea concentration is indicative of the exposure of the tyrosine residues to solvent upon denaturation (dotted line). **(f)** Apparent fraction unfolded (Fapp) of Spo0F as a function of denaturant concentration. The unfolding transitions monitored by CD at 222 nm (○) and by tyrosine FL at 305 nm (□) are shown. The confidence limits for the global fit of the combined CD and FL data to a two-state model are also shown (broken and dotted lines). Buffer conditions: 10 mM potassium phosphate at pH 7.0 and 25 ºC.

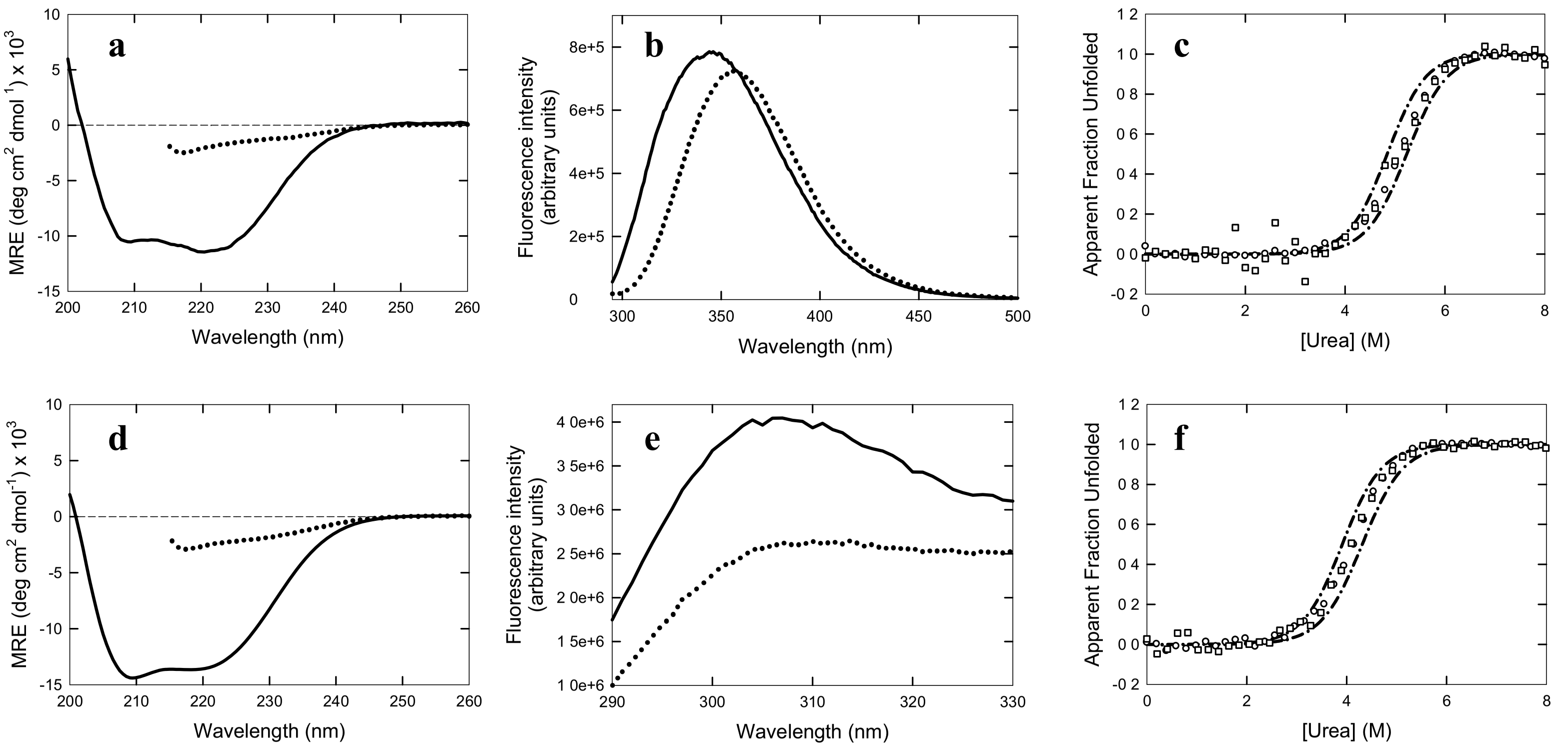

a
MRE (deg cm^2 dmol ^{1}) x 10^3
10
5
0
-5
-10
-15
200
210
220
230
240
250
260
Wavelength (nm)
b
Fluorescence intensity
(arbitrary units)
8e+5
6e+5
4e+5
2e+5
0
300
350
400
450
500
Wavelength (nm)
c
Apparent Fraction Unfolded
1 2
1 0
0 8
0 6
0 4
0 2
0 0
-0 2
0
2
4
6
8
[Urea] (M)
d
MRE (deg cm^2 dmol^{-1}) x 10^3
10
5
0
-5
-10
-15
200
210
220
230
240
250
260
Wavelength (nm)
e
Fluorescence intensity
(arbitrary units)
4 0e+6
3 5e+6
3 0e+6
2 5e+6
2 0e+6
1 5e+6
1 0e+6
290
300
310
320
330
Wavelength (nm)
f
Apparent Fraction Unfolded
1 2
1 0
0 8
0 6
0 4
0 2
0 0
-0 2
0
2
4
6
8
[Urea] (M)

Figure 3.2

Figure 3.3 – Equilibrium and kinetic experimental analyses of NT-NtrC and Spo0F. **(a)** Equilibrium unfolding and refolding of NT-NtrC. The equilibrium denaturation is completely reversible as is seen by the coincidence of the unfolding (○) and refolding (●) CD signal at 222 nm plotted as a function of denaturant concentration. The fit to a two-state model is shown (broken and dotted line). The baselines for the native state (continuous line) and the unfolded state (dotted line) predicted from the two-state model are also shown. The burst-phase amplitude measured by stopped-flow CD refolding of NT-NtrC from 6 M urea is plotted as a function of final urea concentration (Δ) and fit to a two-state model (thick broken line). The magnitude of the burst-phase amplitude under strongly refolding conditions (0.6 M urea) is represented by the double-headed arrow. The FL intensity at 315 nm is plotted as a function of urea concentration (□) and fit to a two-state model (thin dashed line). **(b)** Chevron analysis of NT-NtrC. The recovery of the native signal upon refolding from high denaturant concentration occurs by bi-exponential kinetics. The relaxation times determined by CD [(×), slow phase; (+), fast phase] and by FL [(○), slow phase; (□), fast phase] are shown. The amplitudes associated with the fast refolding phase by both CD and FL at denaturant concentrations > 4 M urea and the amplitudes associated with the slow refolding phase by CD at denaturant concentrations < 2 M urea were too small to obtain accurate relaxation times and were thus excluded from the chevron analysis. A single phase is observed in unfolding kinetics; the relaxation times determined by FL (●) are also shown. **(c)** Amplitudes associated with the relaxation times determined by FL. The symbols used are the same as in (b). **(d)** Equilibrium unfolding and refolding of Spo0F. The symbols used are the same as in (a). The reversibility of the equilibrium denaturation is seen by the coincidence of the unfolding (○) and refolding (●) CD signal at 222 nm plotted as a function of denaturant concentration. The burst-phase amplitude (Δ) is measured by stopped-flow CD refolding of Spo0F from 6 M urea. The FL intensity (□) at 305 nm is plotted as a function of urea concentration. **(e)** Chevron analysis of Spo0F. The refolding relaxation times determined by CD [(×), slow phase; (+), fast phase] and by FL [(○), slow phase; (□), fast phase], and the unfolding relaxation times determined by CD (|) and by FL [(●), fast phase; (▲), slow phase] are shown. **(f)** Amplitudes associated with the relaxation times determined by FL. The symbols used are the same as in (e). Buffer conditions: 10 mM potassium phosphate at pH 7.0 and 25 ºC.

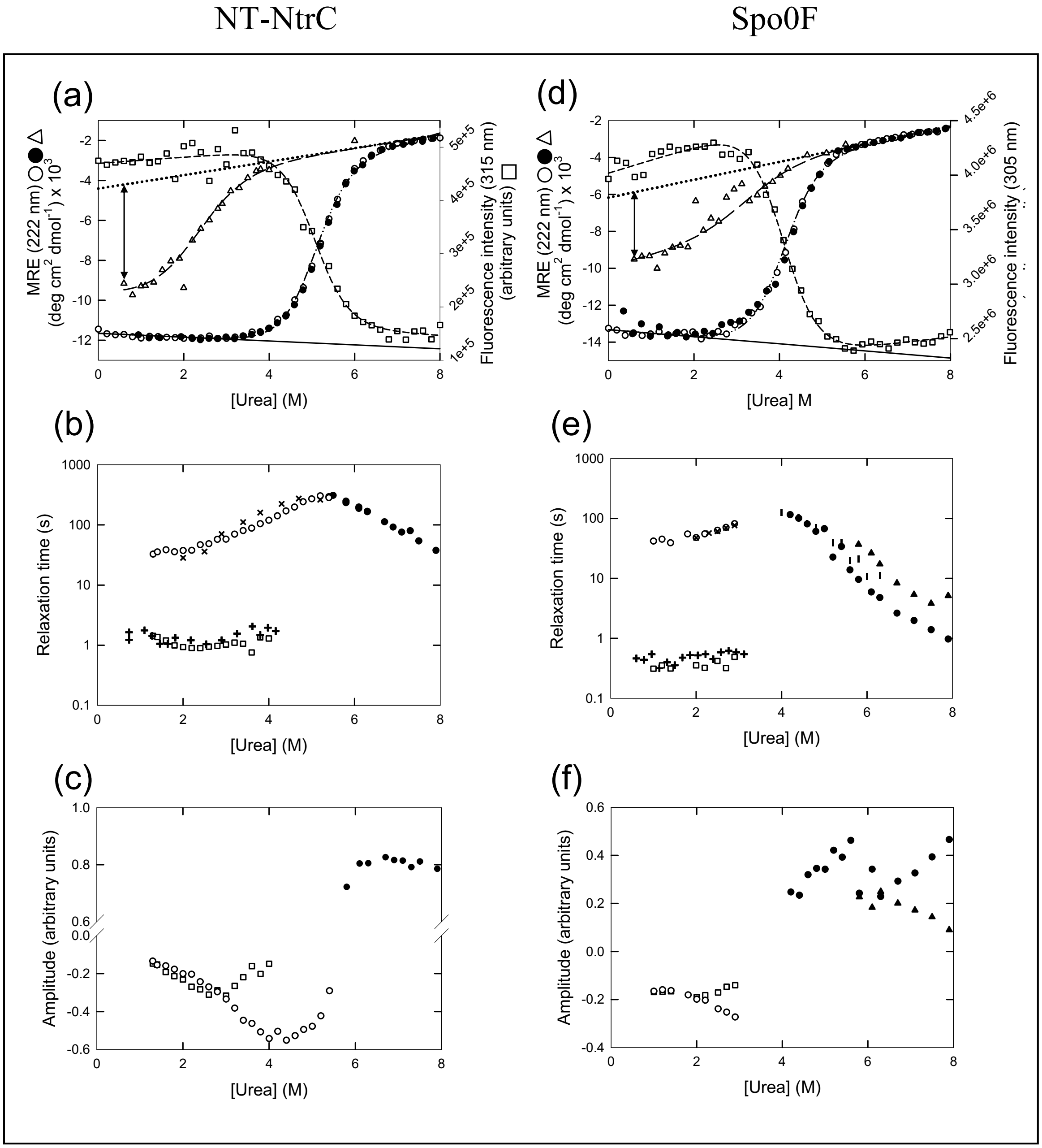

NT-NtrC
Spo0F
(a)
MRE (222 nm) ○●△ (deg cm^2 dmol^{-1}) x 10^3
-2
-4
-6
-8
-10
-12
Fluorescence intensity (315 nm) (arbitrary units) □
5e+5
4e+5
3e+5
2e+5
1e+5
0
2
4
6
8
[Urea] (M)
(d)
MRE (222 nm) ○●△ (deg cm^2 dmol^{-1}) x 10^3
-2
-4
-6
-8
-10
-12
-14
Fluorescence intensity (305 nm)
4.5e+6
4.0e+6
3.5e+6
3.0e+6
2.5e+6
[Urea] M
(b)
Relaxation time (s)
1000
100
10
1
0.1
[Urea] (M)
(e)
Relaxation time (s)
1000
100
10
1
0.1
[Urea] (M)
(c)
Amplitude (arbitrary units)
1.0
0.8
0.6
0.0
-0.2
-0.4
-0.6
[Urea] (M)
(f)
Amplitude (arbitrary units)
0.6
0.4
0.2
0.0
-0.2
-0.4
[Urea] (M)

Figure 3.3

the native ellipticity at 222 nm was recovered in the dead-time of the stopped-flow instrument (~5 ms) for both proteins (Fig. 3.3a and 3.3d), compared to the 95% signal recovered in the same time frame for CheY.[37,76] The apparent thermodynamic properties of the burst-phase species were estimated by measuring the amplitude of the burst-phase CD reaction as a function of the final denaturant concentration in refolding. The sigmoidal loss in the ellipticity at 222 nm for both NT-NtrC and Spo0F at increasing final urea concentrations (Fig. 3.3a and 3.3d) is consistent with the cooperative disruption of secondary structure in a stable partially-folded state, I^{BP}. Fitting the urea-dependence of the burst-phase amplitude to a two-state model provided an estimate of the stability for the I_{BP} species in the two proteins (Table 3.1) along with estimated stability for the burst-phase intermediate in CheY.[37] Compared to the native state, the decreased stabilities and *m*-values for the intermediates suggest that the interiors of these early partially-folded states are less well packed than their native counterparts. For CheY[37] and, to a lesser extent, for NT-NtrC and Spo0F, a substantial amount of secondary structure appears early in the folding process.

Slow refolding reactions: Subsequent to the burst-phase reaction, the remainder of the CD and FL signal for the refolding of NT-NtrC is recovered by bi-exponential kinetics whose relaxation times are shown in a chevron plot in Figure 3.3b. The two phases are nearly equal in amplitude (Fig. 3.3c), and both relaxation times are independent of denaturant concentration under strongly folding conditions. Because both refolding phases at low urea concentration are accelerated modestly in the presence of a prolyl isomerase, cyclophilin (Fig. 3.4a), the cis/trans isomerization of an Xaa-Pro

Table 3.1 – Apparent thermodynamic properties of CheY-like proteins and their respective sub-millisecond intermediates [a,b]

[a] The native state stability and *m*-values are obtained by a global analysis of refolding and unfolding urea denaturation curves at multiple wavelengths monitored by both FL and CD spectroscopy. The stability of the intermediate and its dependence on urea concentration is determined by fitting the amplitude of the burst-phase reaction, monitored by CD at 222 nm, to a two-state model. The errors reported are standard errors from the global fits.

[b] These thermodynamic parameters are regarded as apparent because the two-state fits ignore the contribution by the cis/trans isomerization of the native cis-prolyl peptide bond K104-P105 in the unfolded state.

Protein	NT-NtrC		Spo0F		CheY[37]	
State	Native	Intermediate	Native	Intermediate	Native	Intermediate
ΔG° (H_2O) kcal mol^{-1}	7.52 ± 0.14	2.36 ± 0.40	5.99 ± 0.12	2.98 ± 0.23	5.37 ± 0.21	2.30 ± 0.40
m-value kcal mol^{-1} M^{-1}	1.49 ± 0.03	1.04 ± 0.09	1.47 ± 0.03	0.93 ± 0.07	1.59 ± 0.06	0.92 ± 0.34
C_m M	5.05 ± 0.2	2.27 ± 0.58	4.08 ± 0.16	3.20 ± 0.49	3.33 ± 0.26	2.5 ± 1.30

Table 3.1

Figure 3.4 – Cyclophilin dependence of the refolding kinetics of NT-NtrC and Spo0F. (a) The relaxation times determined for the refolding of 8.5 μM NT-NtrC at 1.1 M urea, with varying final concentrations of cyclophilin and monitored by FL, are shown [(○) slow phase; (□) fast phase]. **(b)** The relaxation times determined for the refolding of 6.7 μM Spo0F at 1 M urea, with varying final concentrations of cyclophilin and monitored by FL, are shown [(○) slow phase; (□) fast phase]. Buffer conditions: 10 mM potassium phosphate at pH 7.0 and 15 ºC.

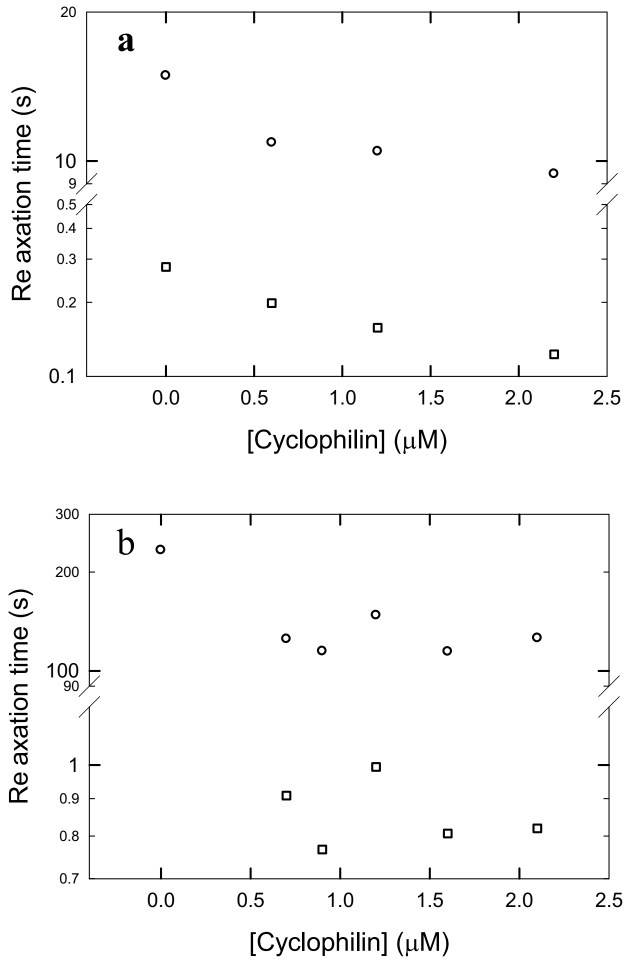

a
Re axation time (s)
20
10
9
0.5
0.4
0.3
0.2
0.1
0.0
0.5
1.0
1.5
2.0
2.5
[Cyclophilin] (μM)
b
Re axation time (s)
300
200
100
90
1
0.9
0.8
0.7
0.0
0.5
1.0
1.5
2.0
2.5
[Cyclophilin] (μM)

Figure 3.4

peptide bond must limit folding to some extent. This behavior has been observed previously for the single slow refolding phase in CheY[37] and reflects the presence of a cis prolyl peptide bond (K104-P105) in the native conformation. When these response regulators are unfolded, the trans isomer becomes dominant and must convert to the cis isomer via a rate-limiting reaction that is coupled to folding. The weak cyclophilin dependence of this rate implies limited accessibility to the prolyl bond, which suggests that the isomerization reaction occurs in a partially folded state.

Similar kinetic results are seen for Spo0F (Fig. 3.3e and 3.3f). Refolding in the presence of cyclophilin modestly decreased the relaxation time for the slow phase but had no discernible effect on the fast folding phase under strongly refolding conditions (Fig. 3.4b).

Unfolding reactions: The unfolding reaction of NT-NtrC monitored by fluorescence and CD is well-described by a single exponential phase, whose relaxation time decreases exponentially above 5.8 M urea (Fig. 3.3b). The amplitude of the unfolding phase accounts for the entire ellipticity change expected from the equilibrium unfolding profile, eliminating the possibility of rapid unfolding reactions (Fig. 3.3c).

The unfolding reaction of Spo0F is more complex (Fig. 3.3e). Above 5.5 M urea concentration, unfolding occurs by bi-exponential kinetics; the urea dependence of the major, fast phase is collinear with the single unfolding phase observed below 5.5 M urea. Only the faster phase is observed when unfolding of Spo0F is monitored by CD. The CD amplitude associated with this unfolding phase is within error of that expected from equilibrium measurements (data not shown); the absence of the slower phase detected by fluorescence may reflect the lower signal to noise ratio in the CD experiment.

Global analysis

The largely comparable equilibrium and kinetic responses of NT-NtrC and Spo0F with those of CheY led to a test of the hypothesis that NT-NtrC and Spo0F also fold via prolyl isomer-dictated parallel channels with either early on- (Model 1 - Fig. 3.5a) or off-pathway (Model 2 - Fig. 3.5b) intermediates, as has been done previously for CheY.[37] In these models, N_C and N_T correspond to the native conformation with cis and trans isomers at the native cis prolyl peptide bond, U_C and U_T correspond to the unfolded states with the respective prolyl isomers and $I^{BP}{}_C$ and $I^{BP}{}_T$ correspond to the burst-phase intermediate with cis and trans prolyl isomers.

A comprehensive set of unfolding and refolding traces were fit globally to these kinetic models with initial estimates for the parameters based on (1) experimentally determined equilibrium properties (Table 3.1), (2) the microscopic rate constants, *k*, and their urea dependences, $m^{\ddagger}$, obtained from the chevron plot of the dependences of the relaxation times (*k*   1/ τ) on the final denaturant concentrations and an in-house algorithm, Chevron Fitter,[37] and (3) the experimentally determined distribution of prolyl isomers in the unfolded state for each protein obtained from penta-peptide models (Fig. 3.6a and 3.6b). The sequence identity adjacent to the cis prolyl residue in NT-NtrC, Spo0F and CheY, Lys-Pro-Phe, resulted in trans:cis distributions that are equivalent within error, 90:10. The equilibrium analysis provided the stability and denaturant dependence of the major species observed, N_C relative to U_T, and the starting and final

Figure 3.5 – Global analysis of NT-NtrC and Spo0F. **(a)** Model 1: The on-pathway model. The folding mechanism occurs via parallel channels based on the isomerization state of the K104 P105 peptide bond. The burst-phase species, I^{BP} is placed on-pathway, between the unfolded and native states along either channel. **(b)** Model 2: The off-pathway model. The burst-phase species, I^{BP} is placed off-pathway from the unfolded states in both channels. **(c)** Predicted chevron for NT-NtrC from the on-pathway model. The predicted observable relaxation times are shown as continuous black lines; the microscopic rate constants determined by the model are shown as dotted green lines for the cis channel. The isomerization relaxation times in each state are shown as broken and dotted lines, blue for the native states, red for the unfolded states, and magenta for the burst-phase intermediates. The chevron analysis from Fig. 2b is shown as open circles for comparison. **(d)** Predicted chevron for NT-NtrC from the off-pathway model. The legends are the same as in (c). **(e)** Predicted chevron for Spo0F from the on-pathway model. The legends are the same as in (c) and the microscopic rate constants for the trans channel are shown as broken green lines. The chevron analyses from Fig. 2e are shown as open circles for comparison. **(f)** Predicted chevron for Spo0F from the off-pathway model. The legends are the same as in (e).

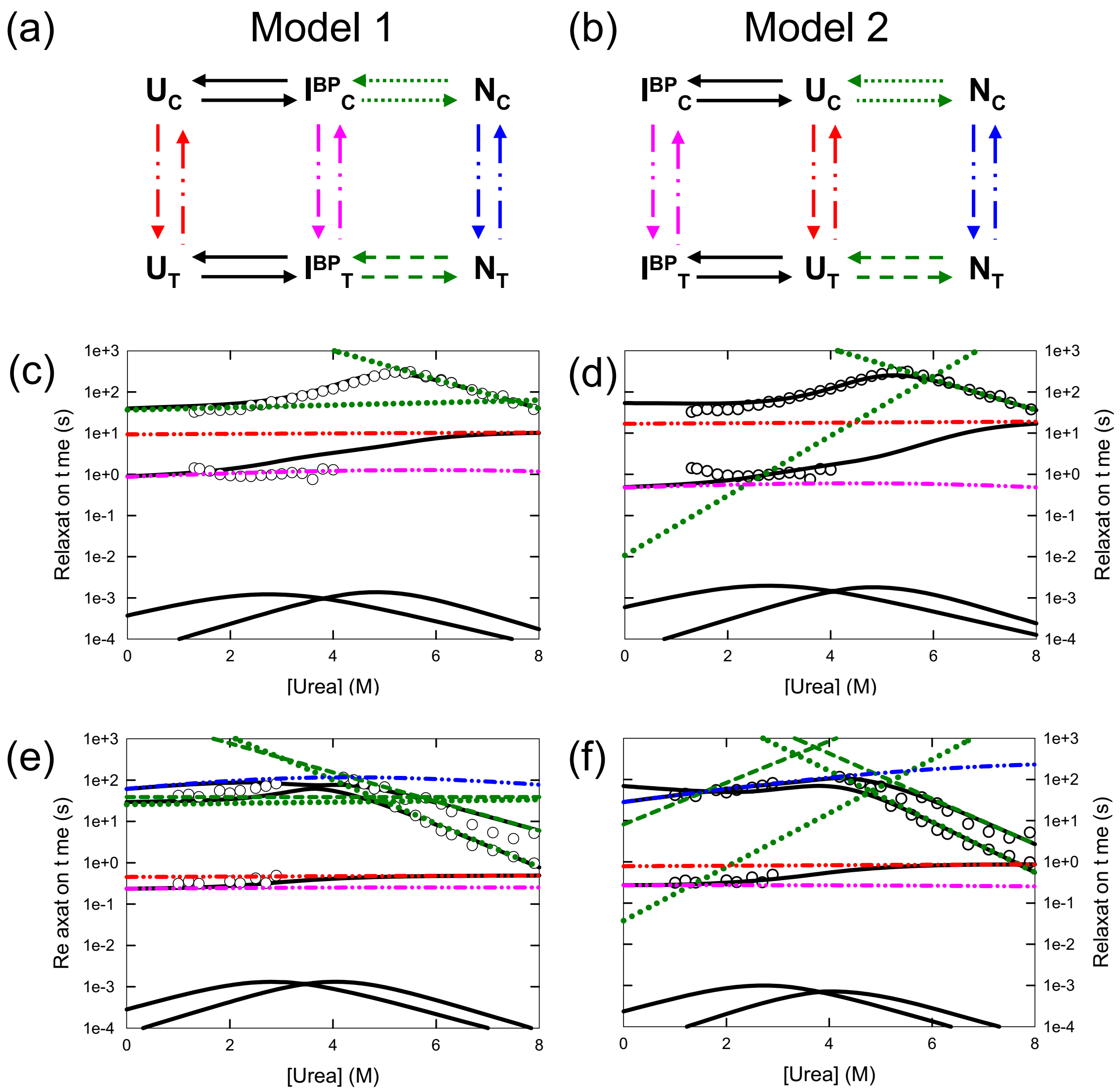
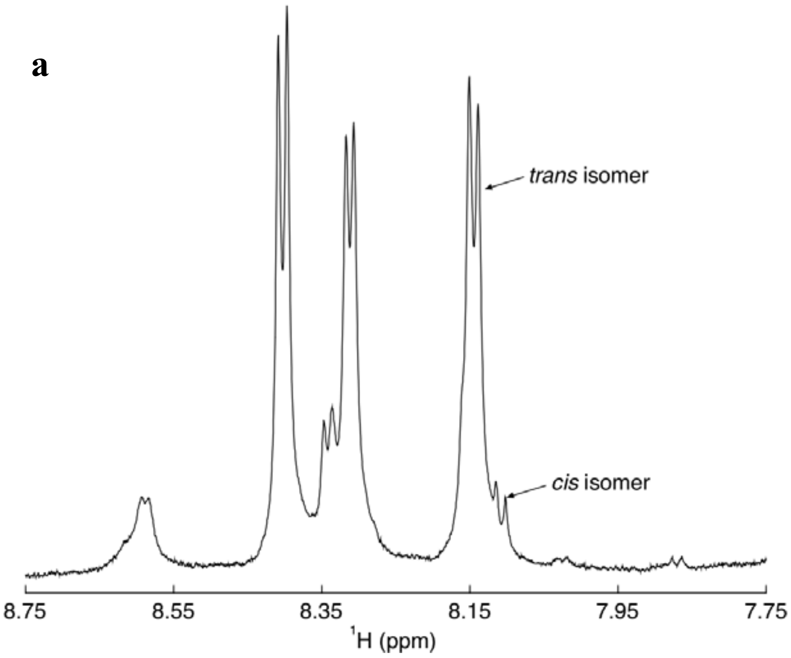

(a)
Model 1
U_C
I^BP_C
N_C
U_T
I^BP_T
N_T
(b)
Model 2
I^BP_C
U_C
N_C
I^BP_T
U_T
N_T
(c)
(d)
(e)
(f)
Relaxat on t me (s)
Re axat on t me (s)
1e+3
1e+2
1e+1
1e+0
1e-1
1e-2
1e-3
1e-4
0
2
4
6
8
[Urea] (M)

Figure 3.5

Figure 3.6 – 1D Proton NMR. The amide region of the 1D proton NMR spectra of the penta-peptide containing the K-P peptide bond in NT-NtrC and Spo0F. (**a**) The NT-NtrC, K104-P105 peptide bond contained in the penta-peptide Ac-PKPFD-NH_2. (**b**) The Spo0F, K104-P105 peptide bond contained in the penta-peptide Ac-AKPFD-NH_2. The peaks associated with the *cis* and *trans* isomers are highlighted.


a
trans isomer
cis isomer
8.75
8.55
8.35
8.15
7.95
7.75
^{1}H (ppm)

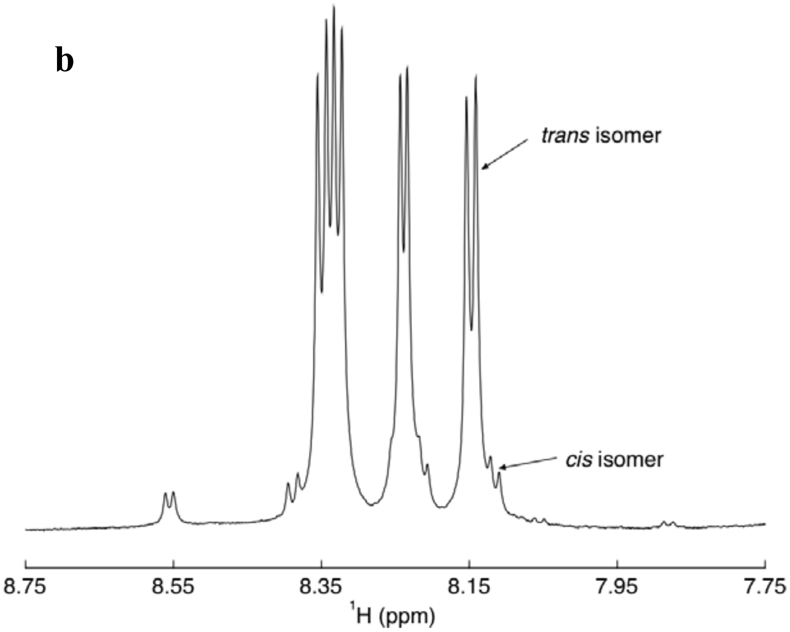

b
trans isomer
cis isomer
8.75
8.55
8.35
8.15
7.95
7.75
^{1}H (ppm)

Figure 3.6

amplitudes for the kinetic traces. Although the rate constants for the burst-phase refolding reaction cannot be determined by stopped-flow mixing, the equilibrium constants for the $I^{BP}{}_{C}/U_{C}$ and the $I^{BP}{}_{T}/U_{T}$ reactions were assumed to be equal and were obtained by fitting the urea dependence of the CD burst-phase amplitude to a two-state model (Fig. 3.3a and 3.3d). The folding rate constants for the burst-phase intermediates were assumed to be $>10^{4}$ s^{-1} to account for their appearance within 5 ms. Because the rates of formation of the intermediates are at least $>10^{5}$ faster than the observed rate constant for the appearance of the native conformation, these two processes do not kinetically couple with each other. However, simulations show that the significant stabilities and the nonzero m-values of the intermediates (Figs. 3.3a and 3.3d) in rapid pre-equilibrium with the unfolded states result in a significant impact on the observed relaxation time (Fig. 3.7). As a first approximation, the urea dependences of the refolding and unfolding rate constants for the $I^{BP}{}_{T}$ and $I^{BP}{}_{C}$ species, $m^{\ddagger}{}_{r}{}^{BP}{}_{T}$ and $m^{\ddagger}{}_{u}{}^{BP}{}_{T}$ and $m^{\ddagger}{}_{r}{}^{BP}{}_{C}$ and $m^{\ddagger}{}_{u}{}^{BP}{}_{C}$, were each assigned to be half of the *m*-value for the equilibrium unfolding reaction, 1.04 kcal mol^{-1} M^{-1} for NT-NtrC and 0.93 kcal mol^{-1} M^{-1} for Spo0F. The procedure is described in more detail in a previous paper.[37]

A total of 32 FL kinetic traces for NT-NtrC and 27 for Spo0F obtained under a variety of unfolding and refolding conditions were then fit to the two models, and the parameters were optimized using the Levenberg Marquardt algorithm. The microscopic rates, kinetic *m*-values, native and unfolded signals and the Z values (relative signal contribution from each species normalized to the difference between the signals for the native and unfolded species) were modeled globally. The kinetic m-values were

Figure 3.7 – Simulated three-state model with a fast folding intermediate. The thick black lines in panel a represent the microscopic relaxation times of the U → N and N → U refolding and unfolding reactions in the absence of an intermediate. In the presence of an intermediate, I, that rapidly equilibrates with the unfolded state, U, the relaxation time of the slower I ⇆ N reaction (crosses in panel a) is sensitive to the properties of the faster reaction through the relative populations of I and U (determined by the ΔG and m-value of the U ⇆ I equilibrium). This effect is independent of the rate of the fast refolding reaction, (blue and red circles in panel a and blue and red lines in panel b) as long as the difference in relaxation times is > 100 fold and the effect is independent of the folding mechanism (on- or off- pathway intermediate).

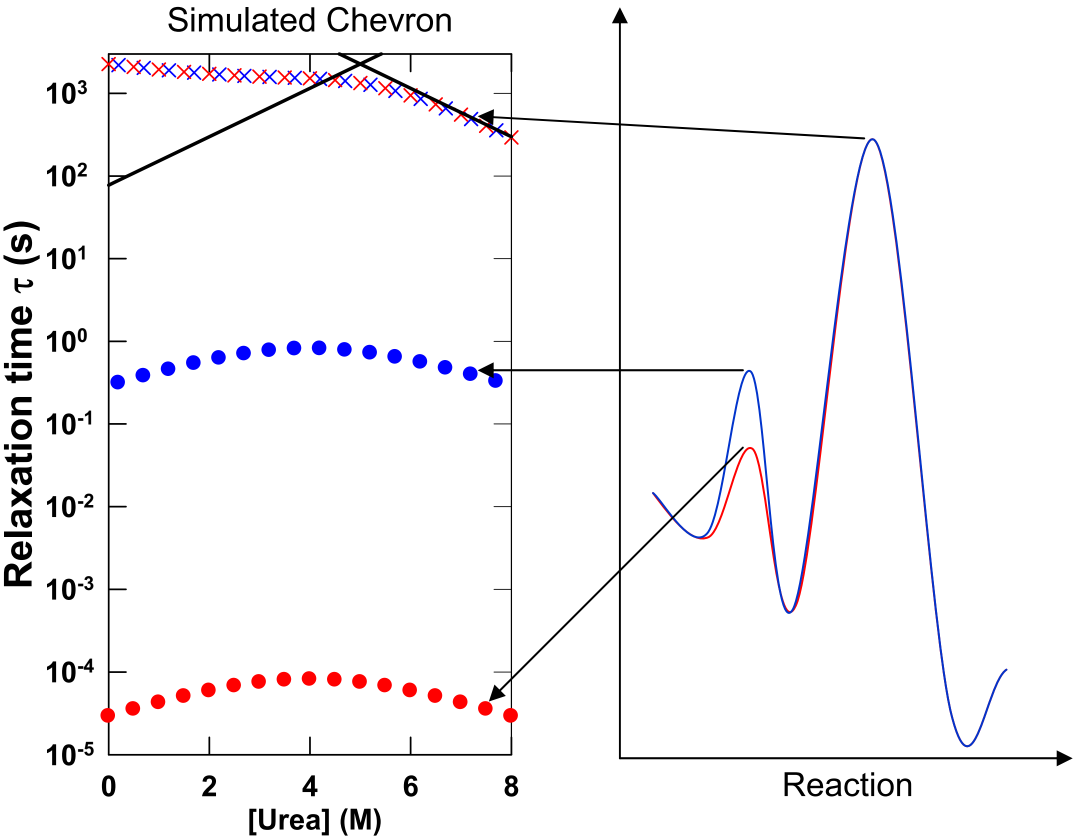

Simulated Chevron
Relaxation time τ (s)
10^3
10^2
10^1
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
0
2
4
6
8
[Urea] (M)
Reaction

Figure 3.7

constrained such that refolding m-values are ≥ 0 and unfolding m-values are ≤ 0. The signal offsets were allowed to vary for each kinetic trace. The protein concentration for each kinetic trace was allowed to vary, albeit with strongly constraints (within 3% of the measured value) to account for possible errors in the measurement of the protein concentration (Fig. 3.8).

The quality of the fits to on- and off-pathway models was assessed by comparison of their reduced chi-square values and by visual comparisons of the predicted chevrons and the amplitudes for the globally-minimized parameters for both proteins. The equilibrium populations of the intermediates for both NT-NtrC (Fig. 3.9a and 3.9b) and Spo0F (Fig. 3.9c and 3.9d) in both models are sufficiently low at all urea concentrations (< 6%) as to remain undetectable, and the predicted equilibrium denaturation profiles are consistent with the experimentally observed two-state behavior (Fig. 3.3a and 3.3d). Although either model provides credible fits of the kinetic traces, the reduced chi-square value for the off-pathway model is 10% lower than that obtained from the fit for the on-pathway model for both NT-NtrC and Spo0F (number of degrees of freedom ~ 3,000, p-values < 0.01).

Representative refolding and unfolding traces for NT-NtrC and Spo0F along with their fits using the parameters from the global analysis of the off-pathway model are shown in Figure 3.10. The microscopic rate constants and their urea-dependences for the on-pathway model, Model 1, are provided in Table 3.2 and Figure 3.5c for NT-NtrC and Figure 3.5e for Spo0F and for the off-pathway model, Model 2, in Table 3.3 and Figure 3.5d for NT-NtrC and Figure 3.5f for Spo0F. The large errors for the rates related with

Figure 3.8 – Global analysis. Normalized correction factors applied to the protein concentration for the various unfolding and refolding traces employed in the global analysis. The protein concentration for each trace was allowed to vary during the global analysis to correct for small random errors in measurement of protein concentrations. A strong constraint was applied, so that the final concentration does not drift more than 3% from the expected concentration.

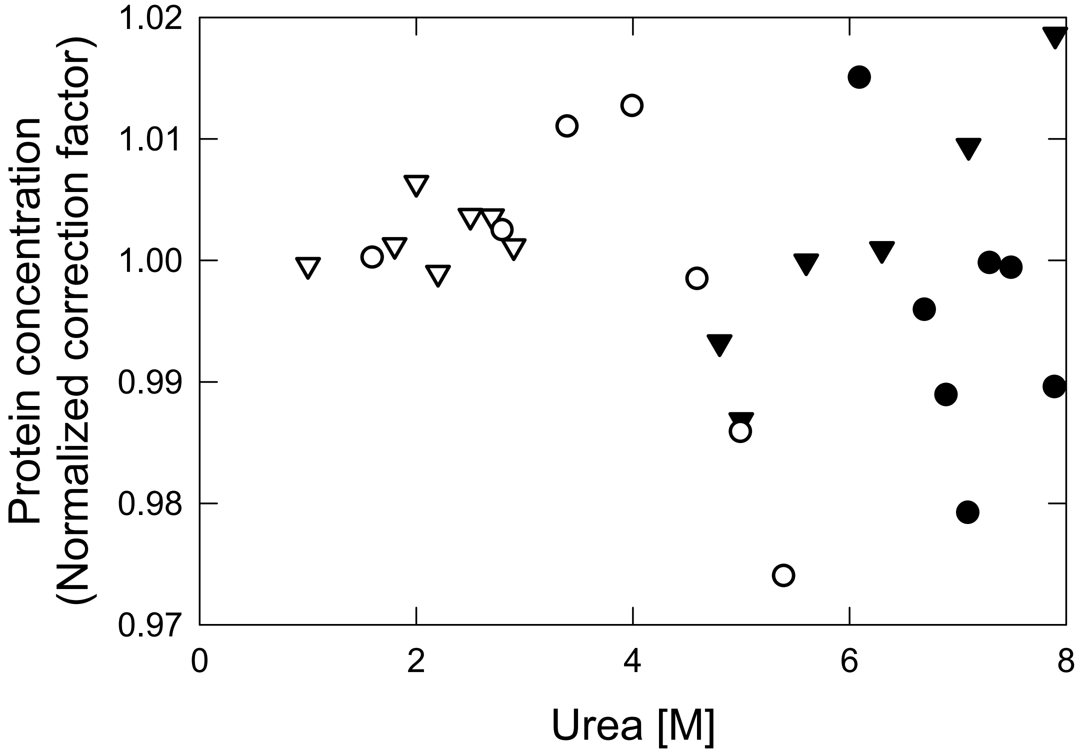

1.02
1.01
1.00
0.99
0.98
0.97
Protein concentration
(Normalized correction factor)
0
2
4
6
8
Urea [M]

Figure 3.8

Figure 3.9 – Equilibrium populations of NT-NtrC and Spo0F predicted by global analysis. (a) The equilibrium population of species predicted for NT-NtrC by the on-pathway model. The native states are represented by blue lines, the unfolded states by red lines and the intermediate states by magenta lines. The dotted lines represent the species in the cis state, while the broken lines represent those in the trans state. **(b)** The equilibrium population of species predicted for NT-NtrC by the off-pathway model. **(c)** The equilibrium population of species predicted by the on-pathway model for Spo0F. **(d)** The equilibrium population of species predicted by the off-pathway model for Spo0F. The legend in (b), (c) and (d) is the same as in (a).

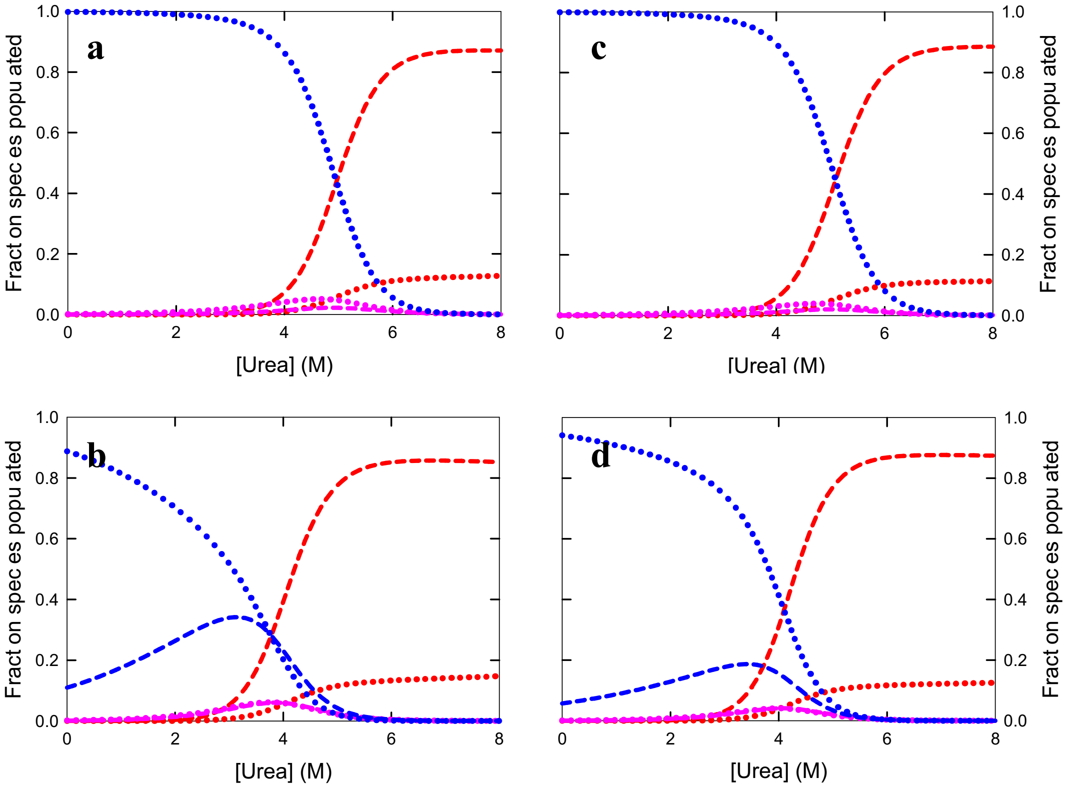

a
c
b
d
Fraction species populated
[Urea] (M)
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8

Figure 3.9

Figure 3.10 – Global analysis of (a) NT-NtrC and (b) Spo0F – the off-pathway model. Representative kinetic traces from the refolding experiments (open circles), unfolding experiments (closed circles), and their fits to the respective off-pathway model obtained from the global analyses (refolding: solid lines and unfolding: dashed lines) are shown. The buffer conditions are described in the caption for Fig. 2 & Fig. 3, and the final urea concentration for each trace is shown.

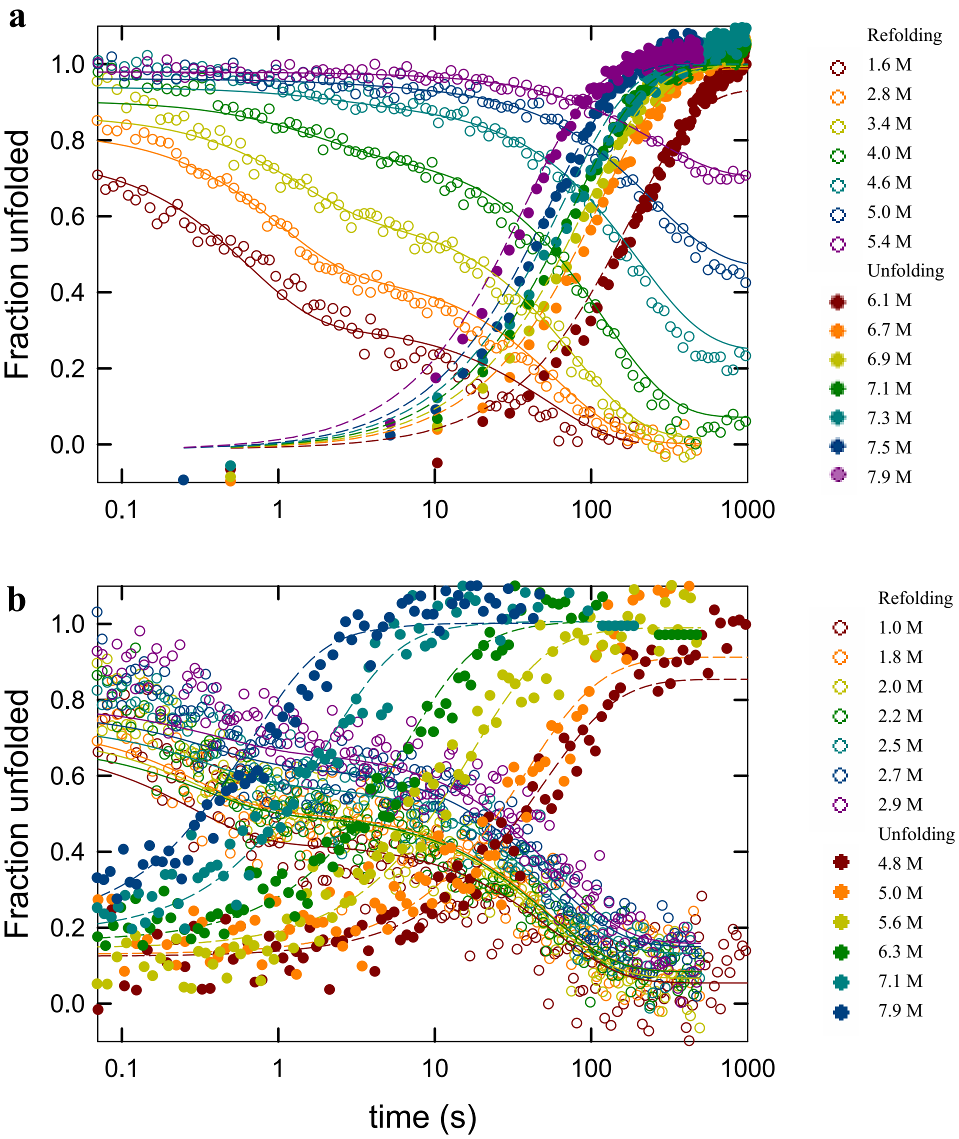

a
Fraction unfolded
1.0
0.8
0.6
0.4
0.2
0.0
0.1
1
10
100
1000
Refolding
○ 1.6 M
○ 2.8 M
○ 3.4 M
○ 4.0 M
○ 4.6 M
○ 5.0 M
○ 5.4 M
Unfolding
● 6.1 M
● 6.7 M
● 6.9 M
● 7.1 M
● 7.3 M
● 7.5 M
● 7.9 M
b
Fraction unfolded
1.0
0.8
0.6
0.4
0.2
0.0
0.1
1
10
100
1000
time (s)
Refolding
○ 1.0 M
○ 1.8 M
○ 2.0 M
○ 2.2 M
○ 2.5 M
○ 2.7 M
○ 2.9 M
Unfolding
● 4.8 M
● 5.0 M
● 5.6 M
● 6.3 M
● 7.1 M
● 7.9 M

Figure 3.10

Table 3.2 – Microscopic rate constants and their associated urea dependences determined by a global fit of kinetic and equilibrium folding data to the on-pathway model (Model 1).

Data for NT-NtrC were obtained from a global analysis of 22 refolding and 10 unfolding kinetic traces and for Spo0F from a global analysis of 10 refolding and 17 unfolding kinetic traces. The errors reported are standard errors from the global fits of the data obtained by standard propagation methods.

[a] The lower limits of the rate constants for the burst-phase reactions are reported. The *m*-values for the burst-phase species are equally distributed to the forward and reverse reactions.

[b] The uncertainty in the rates associated with the I^{BP}_C and N_T species reflect the small/negligible contributions of these species to the fitted kinetic traces.

[c] The N_T species is not detected (ND) in NT-NtrC.

Microscopic step	NT-NtrC		Spo0F	
	k (s^{-1})	m-value	k (s^{-1})	m-value
$U_C \rightarrow U_T$	$9.62\times10^{-2} \pm 5.51\times10^{-3}$	0	$1.99 \pm 1.49\times10$	0
$U_T \rightarrow U_C$	$1.07\times10^{-2} \pm 1.92\times10^{-3}$	0	$2.29\times10^{-1} \pm 1.90$	0
$U_C \rightarrow I^{BP}{}_{C}{}^{a}$	$> 2.38\times10^{4}$	~ 0.51	> 13000	~ 0.52
$I^{BP}{}_{C} \rightarrow U_C{}^{a}$	> 5.14	~ -0.52	> 11	~ -0.51
$U_T \rightarrow I^{BP}{}_{T}{}^{a}$	$> 2.63\times10^{3}$	~ 0.40	> 3500	~ 0.47
$I^{BP}{}_{T} \rightarrow U_T{}^{a}$	$> 6.41\times10$	~ -0.40	> 45	~ -0.46
$I^{BP}{}_{C} \rightarrow I^{BP}{}_{T}{}^{b}$	$8.47\times10^{-2} \pm 1.55\times10^{-2}$	$-0.13 \pm 3.90\times10^{-2}$	1.63 ± 3.90	$-0.03 \pm 6.94\times10^{-1}$
$I^{BP}{}_{T} \rightarrow I^{BP}{}_{C}{}^{b}$	$1.07 \pm 2.76\times10^{-1}$	$0.08 \pm 7.62\times10^{-2}$	2.62 ± 5.16	$0.04 \pm 8.93\times10^{-1}$
$I^{BP}{}_{C} \rightarrow N_C$	$2.72\times10^{-2} \pm 2.09\times10^{-3}$	$0.04 \pm 2.03\times10^{-2}$	$3.91\times10^{-2} \pm 9.42\times10^{-2}$	$0.02 \pm 5.08\times10^{-1}$
$N_C \rightarrow I^{BP}{}_{C}$	$3.83\times10^{-5} \pm 1.65\times10^{-6}$	$-0.48 \pm 2.36\times10^{-3}$	$7.72\times10^{-5} \pm 5.81\times10^{-5}$	$-0.72 \pm 6.97\times10^{-2}$
$I^{BP}{}_{T} \rightarrow N_T{}^{bc}$	ND	ND	$2.57\times10^{-2} \pm 5.57\times10^{-2}$	$1.28\times10^{-7} \pm 7.83\times10^{-1}$
$N_T \rightarrow I^{BP}{}_{T}{}^{bc}$	ND	ND	$2.57\times10^{-4} \pm 3.47\times10^{-4}$	$-0.48 \pm 3.24\times10^{-1}$
$N_C \rightarrow N_T{}^{bc}$	ND	ND	$1.79\times10^{-3} \pm 2.55\times10^{-2}$	-0.14 ± 1.41
$N_T \rightarrow N_C{}^{bc}$	ND	ND	$1.46\times10^{-2} \pm 2.07\times10^{-1}$	0.19 ± 1.43

Table 3.2

Table 3.3 – Microscopic rate constants and their associated urea dependences determined by a global fit of kinetic and equilibrium folding data to the off-pathway model (Model 2).

Data for NT-NtrC were obtained from a global analysis of 22 refolding and 10 unfolding kinetic traces and for Spo0F from a global analysis of 10 refolding and 17 unfolding kinetic traces. The errors reported are standard errors from the global fits of the data obtained by standard propagation methods.

[a] The lower limits of the rate constants for the burst-phase reactions are reported. The *m*-values for the burst-phase species are equally distributed to the forward and reverse reactions.

[b] The uncertainty in the rates associated with the $I^{BP}{}_{C}$ and N_T species reflect the small/negligible contributions of these species to the fitted kinetic traces.

[c] The N_T species is not detected (ND) in NT-NtrC.

Microscopic	NT-NtrC		Spo0F	
step	k (s^{-1})	m-value (kcal mol^{-1} M^{-1})	k (s^{-1})	m-value (kcal mol^{-1} M^{-1})
$U_C \rightarrow U_T$	$5.34\times10^{-2} \pm 8.10\times10^{-3}$	0	1.15 ± 8.65	0
$U_T \rightarrow U_C$	$5.95\times10^{-3} \pm 1.02\times10^{-3}$	0	$1.26\times10^{-1} \pm 1.03$	0
$U_C \rightarrow I^{BP}{}_{C}$[a]	$> 1.95\times10^{4}$	~ 0.52	$> 3.17\times10^{4}$	~ 0.52
$I^{BP}{}_{C} \rightarrow U_C$[a]	> 4.22	~ -0.51	$> 2.69\times10^{1}$	~ -0.52
$U_T \rightarrow I^{BP}{}_{T}$[a]	$> 1.65\times10^{3}$	~ 0.40	$> 4.23\times10^{3}$	~ 0.47
$I^{BP}{}_{T} \rightarrow U_T$[a]	> 40.9	~ -0.39	$> 5.25\times10^{1}$	~ -0.48
$I^{BP}{}_{C} \rightarrow I^{BP}{}_{T}$[b]	$1.52\times10^{-1} \pm 1.16\times10^{-2}$	$-0.16 \pm 1.47\times10^{-2}$	1.42 ± 5.52	-0.04 ± 1.09
$I^{BP}{}_{T} \rightarrow I^{BP}{}_{C}$[b]	$1.95 \pm 2.79\times10^{-1}$	$0.07 \pm 3.89\times10^{-2}$	2.27 ± 9.78	0.03 ± 1.46
$U_C \rightarrow N_C$	93.4 ± 1.83	$0.99 \pm 1.54\times10^{-3}$	$2.68\times10^{1} \pm 8.82$	$0.89 \pm 3.87\times10^{-2}$
$N_C \rightarrow U_C$	$2.84\times10^{-5} \pm 9.94\times10^{-7}$	$-0.51 \pm 1.60\times10^{-3}$	$2.14\times10^{-5} \pm 9.88\times10^{-6}$	$-0.84 \pm 3.91\times10^{-2}$
$U_T \rightarrow N_T$[bc]	ND	ND	$1.24\times10^{-1} \pm 4.68\times10^{-1}$	$0.69 \pm 7.08\times10^{-1}$
$N_T \rightarrow U_T$[bc]	ND	ND	$1.48\times10^{-5} \pm 3.86\times10^{-5}$	$-0.75 \pm 4.75\times10^{-1}$
$N_C \rightarrow N_T$[bc]	ND	ND	$2.02\times10^{-3} \pm 2.08\times10^{-2}$	-0.03 ± 1.01
$N_T \rightarrow N_C$[bc]	ND	ND	$3.33\times10^{-2} \pm 3.47\times10^{-1}$	0.24 ± 1.07

Table 3.3

the N_T species in Spo0F and the $I^{BP}{}_C$ species in both proteins reflect the small/negligible contributions of these species to fitted kinetic traces.

Although the folding of the intermediate is too fast to be directly detected by stopped-flow methods, the parameters derived for the on- and off-pathway models, in combination with the reduced chi-square statistic, enable one to choose the more likely model. Specifically, for the on-pathway model, the refolding m-value for the $I^{BP} \rightarrow N$ reaction was constrained to be greater than or equal to zero as expected for a progressive folding reaction. If this constraint is relaxed, the m-value for the $I^{BP} \rightarrow N$ reaction becomes less than zero and, effectively, Model 1 reverts to Model 2, with an off-pathway intermediate (Fig. 3.11). Thus, the optimal fitting of the data is achieved with the off-pathway intermediates.

A triangular model, wherein the native states N_C and N_T have direct access to both the corresponding unfolded states and the corresponding intermediate states, was also tested (data not shown). For NT-NtrC, the model reverts back to a five-state off-pathway model (Model 2, with the N_T state being inaccessible). However, the fit with Spo0F is equally good as the off-pathway model. Model 2 is favored for Spo0F because it accounts for the observed responses as well as the triangular model, but uses fewer parameters.

The lower chi-square values for the off-pathway models provide support for off-pathway intermediates but are not conclusive in eliminating the on-pathway model. Additional support for the off-pathway model in NT-NtrC and Spo0F is provided by the inspection of urea dependence of the microscopic rate constants. In the on-pathway

Figure 3.11 – Three-state models for the on- and off-pathway mechanisms. A negative m-value for the refolding leg of the I → N reaction implies that the m-value for the U ⇆ I equilibrium (difference between thick magenta solid and dashed lines in panel a) is larger than the differences in the m-values for the I → N transition (difference between thick black solid and dashed lines in panel a). In other words, the transition state is less affected by urea than the intermediate and, the transition state is therefore less structured. This situation suggests that the unfolding of the intermediate is required to attain the productive transition state. In effect, this is equivalent to the off-pathway mechanism (panel b).

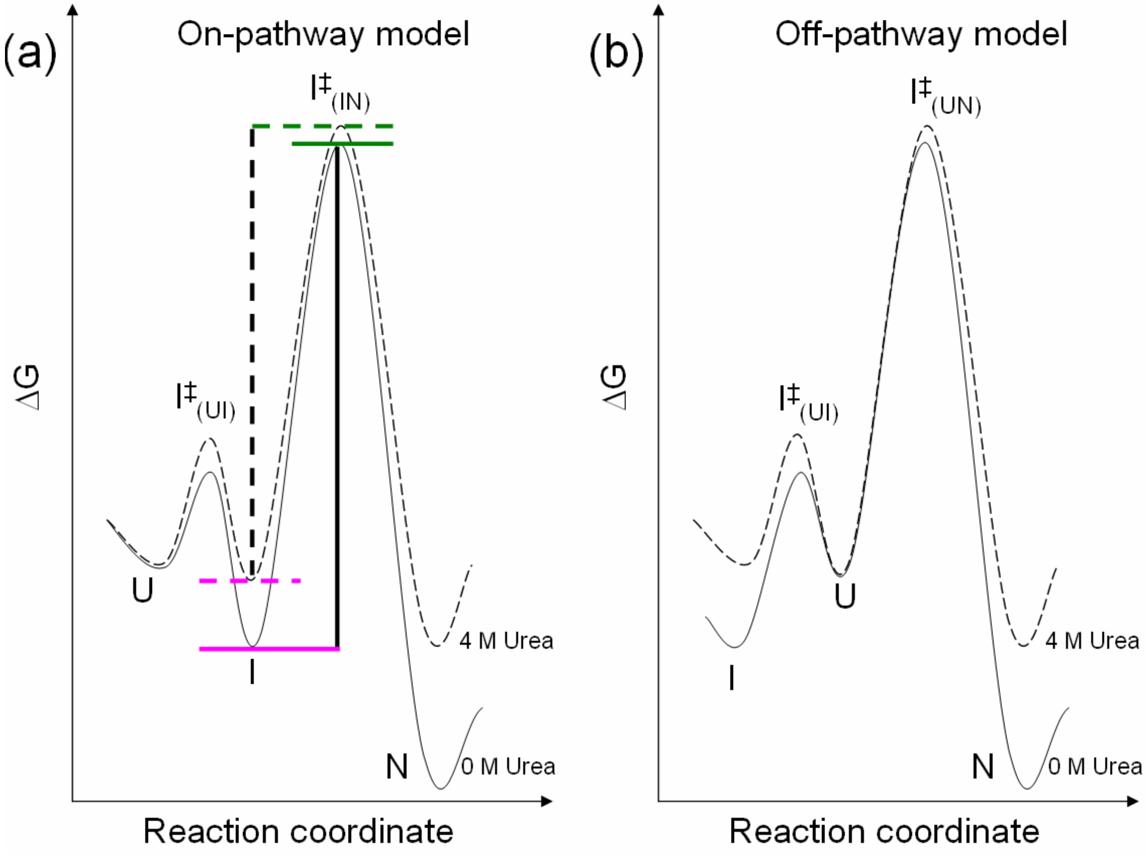

(a)
On-pathway model
ΔG
I‡(IN)
I‡(UI)
U
I
N
4 M Urea
0 M Urea
Reaction coordinate
(b)
Off-pathway model
ΔG
I‡(UN)
I‡(UI)
U
I
N
4 M Urea
0 M Urea
Reaction coordinate

Figure 3.11

model for NT-NtrC (Fig. 3.5c), the nearly urea independent refolding rate constant of the $I^{BP}{}_{C}$ to N_C reaction implies that the productive TSE does not bury a significant amount of surface area relative to that observed in the intermediate. The same prediction is made by the on-pathway model for the refolding of Spo0F (Fig. 3.5e). While the direct refolding of compact non-native intermediates to the native state via a TSE that requires internal repacking has been observed under extreme conditions of high salt[144] or high denaturant,[35] the folding kinetics of many proteins typically reveal that the TSE is more similar to the native state in terms of surface area buried.[98] In accordance with this expectation, the typical urea dependencies for the $U_C \rightarrow N_C$ refolding reactions in the off-pathway models for both NT-NtrC and Spo0F demonstrate the burial of a significant fraction of exposed surface area in the productive TSE of both proteins (Fig. 3.5d and 3.5f). Thus, by both statistical and folding behavior criteria, the global analyses of the experimental data favor, but do not definitively prove an off-pathway folding mechanism for NT-NtrC and Spo0F.

Simulation Analysis

Thermodynamic as well as kinetic Gō-model simulations of folding were carried out for NT-NtrC and Spo0F using methodology previously developed for the homologous CheY folding reaction.[62] To aid comparison of the influence of sequence variation on the folding of the topologically-equivalent proteins a flavored variant of the traditional Gō-model was employed in which heterogeneity of the native contact energies is added to incorporate sequence effects (see Materials and Methods).

The sequence of events was mapped by examining the dependence of the free energy on several structural properties. For the equilibrium thermodynamic calculations, multi-canonical umbrella sampling was used to ensure the entire accessible landscape was sampled, including unfolded, native and high-energy intermediate species. Although more exact methods for characterizing the folding TSE are available,[145,146] the goal of the present work was to elucidate the folding mechanism by determining the most probable order of events in the formation of structure. Multi-canonical equilibrium simulations have proven to be useful in the study of complex folding behavior when multiple reaction coordinates are necessary to describe the essential features of folding mechanisms.[147 149] To lend support to the mechanism defined by the equilibrium simulations, 100 independent, unbiased kinetic folding simulations were also performed starting from a random coil unfolded structure under conditions promoting the native state. The relative sequence of folding events observed in the ensemble kinetic simulations was in good agreement with the most probable pathway revealed from thermodynamic landscape calculations.

Thermodynamic simulations

The free energy landscape was characterized at the folding transition temperature so that both the native and unfolded basins could be clearly defined. The fraction of native contacts formed, denoted Q, is a useful progress variable for monitoring the formation of secondary and tertiary structure. The free energy was computed as a function of the fraction of native contacts formed in the N- and C-subdomains (Fig. 3.1a) at the transition temperature (Fig. 3.12). The resulting Gō-simulation landscapes for NT-

Figure 3.12 – Gō-model results for the thermodynamic characterization of the N-terminally nucleated folding landscapes for **(a)** NT-NtrC **(b)** and Spo0F. The free energy, G, is shown as a function of the fraction of native contacts formed within the N-subdomain ($Q_{\text{N subdomain}}$) and the fraction of native contacts formed within the C-subdomain ($Q_{\text{C subdomain}}$). Off-pathway frustration is evident for Spo0F for which the prematurely structured C-subdomain must unfold in order for the N-subdomain to fold and drive the progression to the native state. Contours are drawn every kcal mol ^{1}; values exceeding 10 kcal mol ^{1} and regions not sampled are shown in yellow. The free energy is computed at the folding transition temperature such that the folded and unfolded states are equally populated.

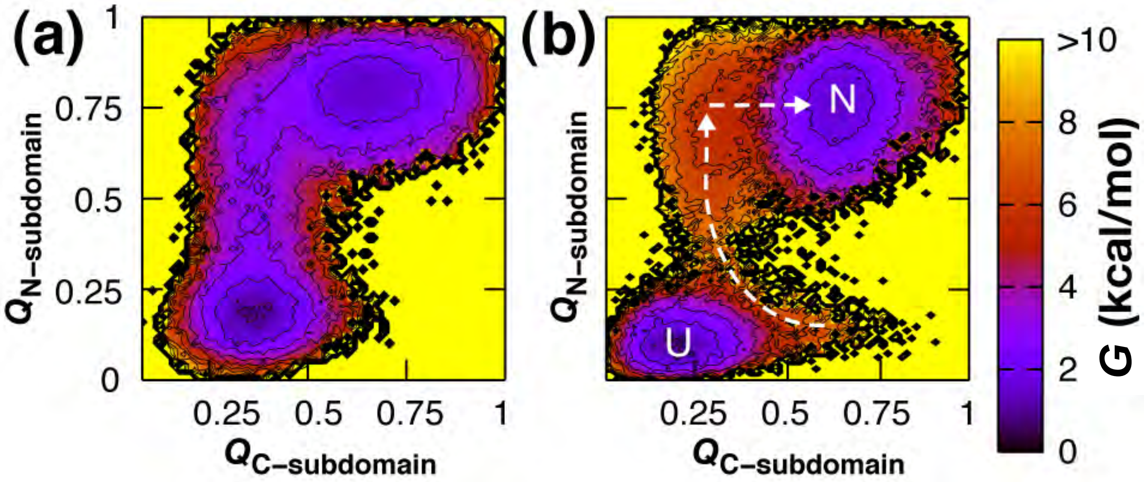

(a)
(b)
$Q_{\text{N-subdomain}}$
$Q_{\text{C-subdomain}}$
0
0.25
0.5
0.75
1
N
U
G (kcal/mol)
>10
8
6
4
2
0

Figure 3.12

NtrC and Spo0F are similar to CheY[127] in that the N-subdomain is partially structured in the productive folding transition state whereas the C-subdomain is not. The C-subdomain does not access its folded basin until the N-subdomain has folded, and the C-subdomain relies on contacts at the subdomain-interface for its stability. This behavior suggests that the N-subdomain serves as the folding nucleus for the C-subdomain. Additionally, the C-subdomain for NT-NtrC and Spo0F exhibited dynamic instability as evidenced by the large width of their native basins where both structured and largely unstructured states are sampled.

The instability of the C-subdomain can be attributed to a lower density of vdW contacts. The Gō-model assigned an average of 1.2/1.1 (NT-NtrC) and 1.6/1.3 (Spo0F) native contacts per residue in the N-/C-subdomains, respectively. The lower average number of contacts for NT-NtrC compared to Spo0F reflects the limitations of the solution NMR structure for NT-NtrC compared to the crystal structures for Spo0F and CheY.[150,151] The smaller number of native contact potentials to promote folding for NT-NtrC results in a lower energy barrier between the unfolded and folded basins than in the case of Spo0F (Fig. 3.12) or CheY.[62] Recent simulations of protein G also demonstrated the dependence of the folding barrier on the experimental structure from which the Gō-model is derived.[151]

Insight into the sequence of folding events for NT-NtrC and Spo0F can be gained by examining the formation of three topologically equivalent $\beta\alpha\beta$ modules in the flavodoxin fold, $\beta1\alpha1\beta2$, $\beta3\alpha3\beta4$ and $\beta4\alpha4\beta5$. The dependence of the free energy on the fraction of contacts formed in each of these three modules reveals the relative order of their formation. In NT-NtrC, elongation of the central β-sheet proceeds from the N-terminus, as has been observed for CheY[62] The most probable folding pathway involves the progressive formation of the β1α1β2 module, the β3α3β4 module and the β4α4β5 module (Fig. 3.13). In Spo0F, folding spreads in both directions from the central β3α3β4 module, first towards the β1α1β2 module and then towards the β4α4β5 module (Fig. 3.14). This finding suggests that the folding nucleus initially identified by Lopez-Hernandez *et al.* as β1α1β2α2β3 in CheY[127] be extended in the case of Spo0F to include helix α3 and strand β4.

Topological frustration in Spo0F

The α2β3α3β4 region in CheY was previously shown to cause significant frustration in folding simulations by partially forming prior to the folding of the N-subdomain.[62] This topologically-frustrated structure was observed to unfold before productive folding in the N-subdomain could occur and proceed to the native state. Topological frustration was not observed in the C-subdomain of NT-NtrC; premature structure in β4 and β5, and the α3 and α4 helices do not preclude productive folding in the N-subdomain of NT-NtrC (Fig. 3.12a, 3.15a and 3.15b). In Spo0F, the β3α3β4 module is no longer a site of frustration as the region serves to initiate productive folding. Frustration in the thermodynamic landscape is evident, however, within its C-subdomain, α3β4α4β5 (Fig. 3.12b, 3.15c and 3.15d).

The dissolution of prematurely formed native contacts, termed backtracking, ascribed to topological frustration has been previously reported in the literature.[77,152 154] Computational and experimental studies of TIM barrel proteins by Finke and colleagues

Figure 3.13 – Sequential assembly of three topologically equivalent triad segments within NT-NtrC. The free energy is shown as a function of the fraction of native contacts formed in the entire protein and within the regions spanning (**a**) strands 1 and 2 and helix 1, (**b**) strands 3 and 4 and helix 3, and (**c**) strands 4 and 5 and helix 4. Folding is seen to proceed in the order $\beta_1\alpha_1\beta_2$: $\beta_3\alpha_3\beta_4$: $\beta_4\alpha_4\beta_5$.

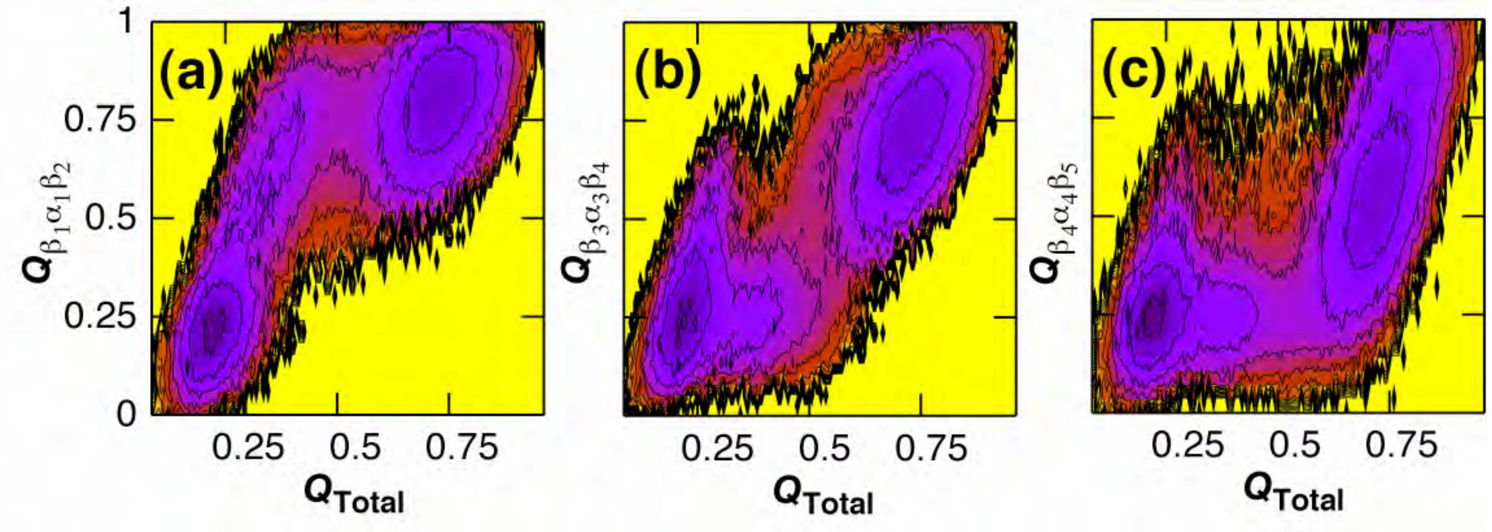
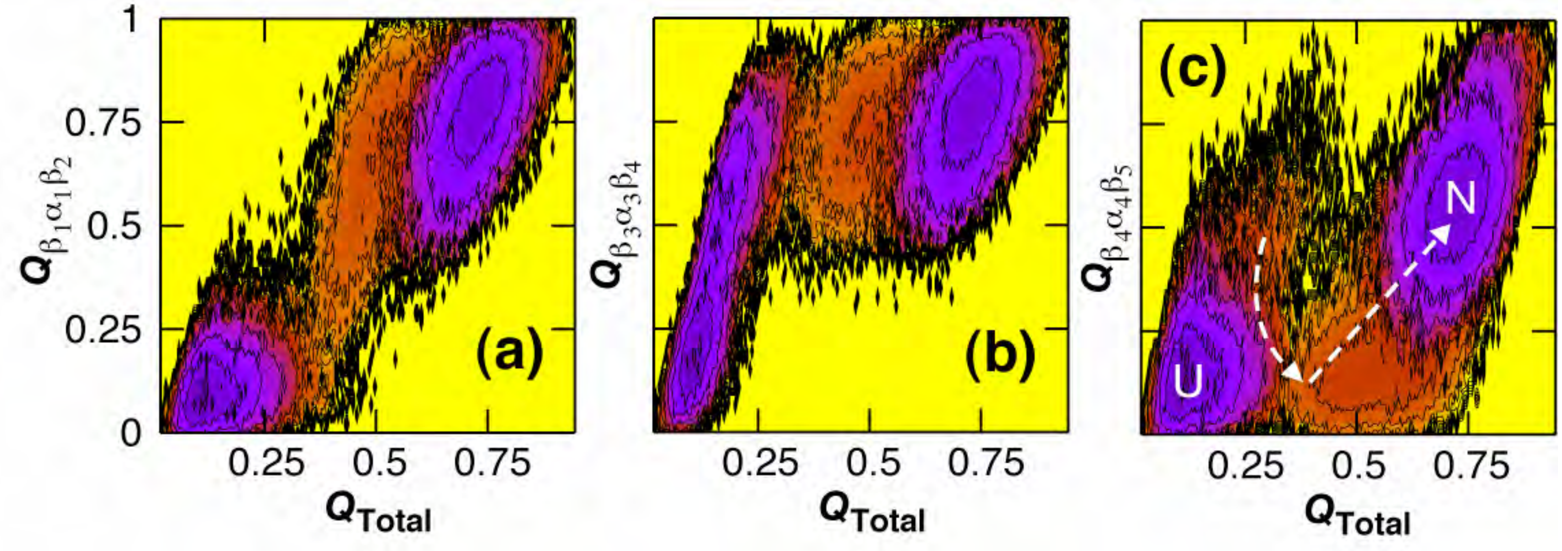

(a)
(b)
(c)
$Q_{\beta_1\alpha_1\beta_2}$
$Q_{\beta_3\alpha_3\beta_4}$
$Q_{\beta_4\alpha_4\beta_5}$
0
0.25
0.5
0.75
1
0.25
0.5
0.75
Q_{Total}

Figure 3.13

Figure 3.14 – Sequential assembly of three topologically equivalent triad segments within Spo0F. The free energy is shown as a function of the fraction of native contacts formed in the entire protein and within (**a**) the regions spanning strands 1 and 2 and helix 1, (**b**) strands 3 and 4 and helix 3, and (**c**) strands 4 and 5 and helix 4. Significant frustration is seen for the helix 4 triad, as evidenced by the high energy barrier bisecting the upper half of the pathway from *U* to *N* at Q_{Total} 0.4 (c). Overall, folding is seen to proceed in the order $\beta_3\alpha_3\beta_4 : \beta_1\alpha_1\beta_2 : \beta_4\alpha_4\beta_5$.



$Q_{\beta_1\alpha_1\beta_2}$
1
0.75
0.5
0.25
0
0.25
0.5
0.75
Q_{Total}
(a)
$Q_{\beta_3\alpha_3\beta_4}$
0.25
0.5
0.75
Q_{Total}
(b)
$Q_{\beta_4\alpha_4\beta_5}$
(c)
N
U
0.25
0.5
0.75
Q_{Total}

Figure 3.14

Figure 3.15 – Frustration in the C-subdomain of NT-NtrC and Spo0F. The free energy is shown as a function of the fraction of native contacts formed within the N-subdomain and between β4 and β5 **(a, c)**, and within the N-subdomain and between α3 and α4 **(b, d)**. Premature structure in the C-subdomain of NT-NtrC **(a, b)** does not preclude N-subdomain folding. For Spo0F **(c, d)**, N-subdomain folding is seen to accompany an initial unfolding of C-subdomain contacts, as evidenced by the high energy barrier bisecting the path to *N* at $Q_{\text{N subdomain}}$ 0.4. The energy scale is described in the caption to Figure 4.

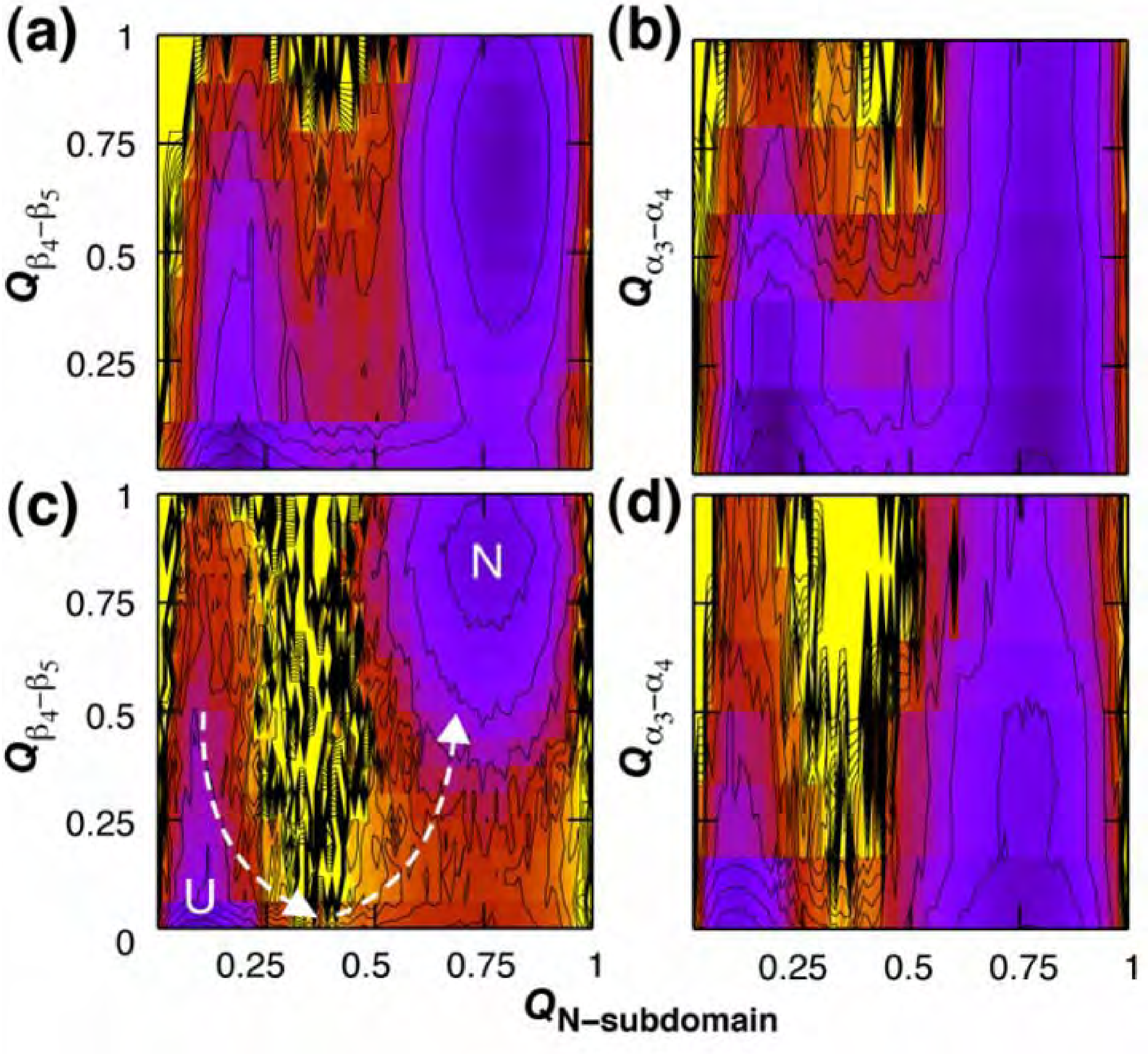

(a)
(b)
(c)
(d)
$Q_{\beta_4-\beta_5}$
$Q_{\alpha_3-\alpha_4}$
$Q_{\beta_4-\beta_5}$
$Q_{\alpha_3-\alpha_4}$
1
0.75
0.5
0.25
0
N
U
0.25
0.5
0.75
1
$Q_{N-subdomain}$

Figure 3.15

have implicated backtracking as a general mechanism for assisting the protein in reaching the native state.[36,155] Premature structure formation followed by backtracking is a likely scenario in the maturation of tertiary structure in multicomponent proteins, in which subdomains must compete for structural contacts.

Kinetic simulations

Ensemble kinetic simulation data are in accord with the thermodynamic results. The fraction of contacts formed at each time point was computed for the 100 kinetic folding simulations and ensemble-averaged. Unfolding of prematurely-folded structure in the C-subdomain is evident in the overall ensemble-averaged kinetic time course for Spo0F but not for NT-NtrC (Fig. 3.16). The negative slopes at Q_{Total} 0.4 indicate local disruption of contacts between β4 and β5 and between α3 and α4 in Spo0F while N-terminal folding is still in progress.

Discussion

As hypothesized for the $(\beta\alpha)_5$ motif, both chain topology and amino acid sequence modulate its folding properties. The equilibrium unfolding mechanisms of NT-NtrC and Spo0F are best described by a two-state model, demonstrating that only the unfolded and native states of these proteins are measurably populated at equilibrium. However, a kinetic analysis reveals a more complicated picture of the folding free energy landscapes for NT-NtrC and Spo0F, which are in many ways similar to the landscape observed for CheY.[37]

Figure 3.16 – Influence of frustration on kinetics. The fraction of native contacts formed at each time point was computed for 100 independent kinetic folding simulations and ensemble-averaged. **(a)** The mean fraction of C-subdomain contacts formed is shown as a function of the fraction of native contacts formed in the entire protein for NT-NtrC, solid line, and Spo0F, broken line. **(b)** The mean fraction of contacts formed in different regions of a protein is shown as a function of the fraction of native contacts formed in the entire protein. For Spo0F, contacts between β4 and β5 are in red and those between α3 and α4 in green. The corresponding regions for NT-NtrC are in blue and pink respectively. The large negative slopes at Q_{Total} 0.4 indicate local unfolding, or backtracking, of C-subdomain contacts in Spo0F.

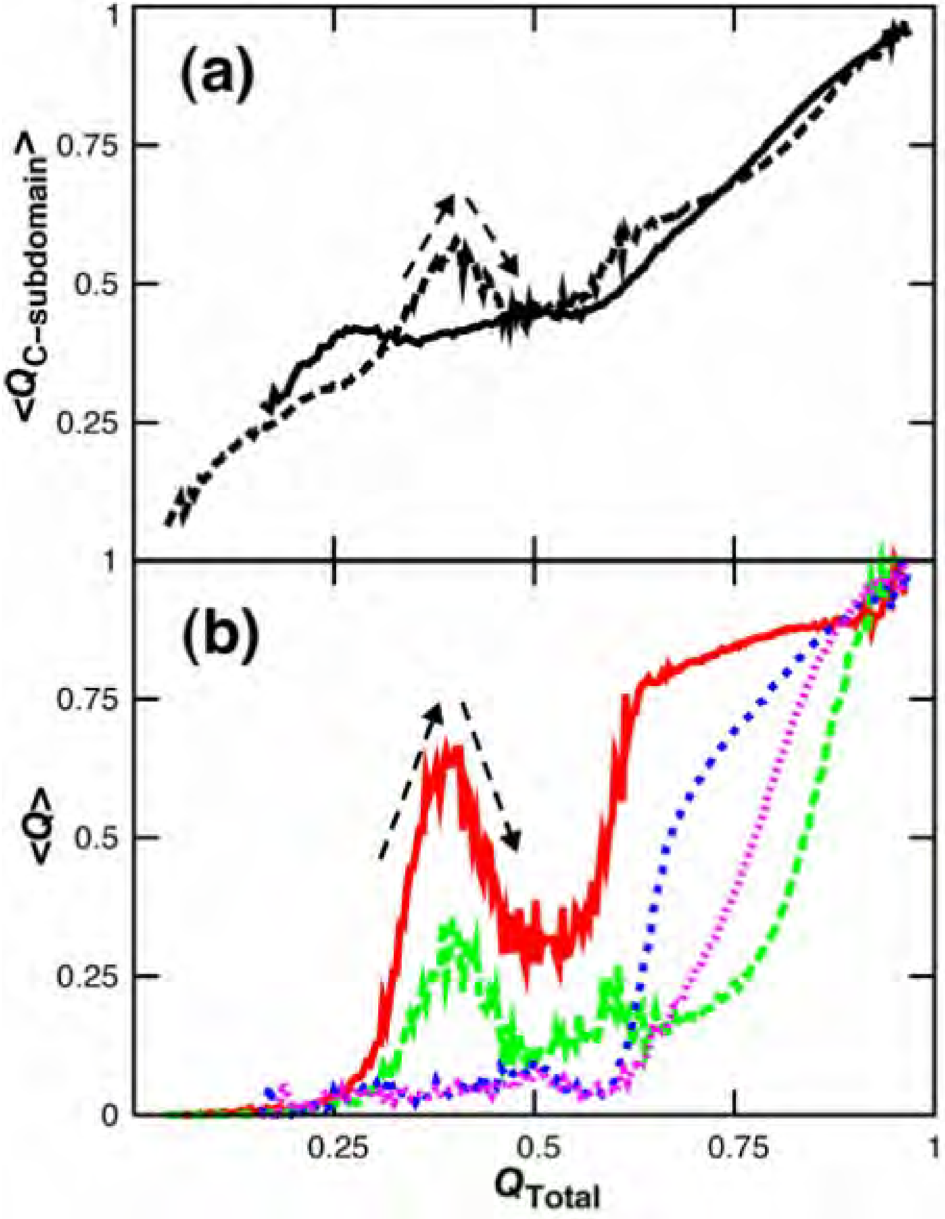

(a)
<$Q_{C\text{-subdomain}}$>
1
0.75
0.5
0.25
(b)
<Q>
1
0.75
0.5
0.25
0
0.25
0.5
0.75
1
Q_{Total}

Figure 3.16

Experimental analysis

Global analysis of equilibrium unfolding transitions, the stabilities of the burst-phase intermediates and the kinetic FL traces derived from a series of unfolding and refolding reactions under a variety of denaturant concentrations for NT-NtrC and Spo0F were best described by parallel channel models with off-pathway intermediates. The previous conclusion that a similar mechanism is operative for CheY,[37] underscores the role of topology in defining the basic features of the folding energy landscape of these three βα-repeat proteins.

However, significant differences between the populations of kinetic species and the rate-limiting reactions were observed during the refolding of the three proteins. Under strongly refolding conditions for NT-NtrC (Fig. 3.17a), both the major and minor unfolded populations, U_T and U_C, rapidly collapse to the corresponding burst-phase, off-pathway intermediates, $I^{BP}{}_T$ and $I^{BP}{}_C$, respectively (Fig. 3.3a and Table 3.1). The subsequent fast refolding reaction corresponds to the isomerization of the prolyl peptide bond accompanying the conversion of $I^{BP}{}_T$ to $I^{BP}{}_C$ (Fig. 3.5d and Table 3.3). The slow refolding reaction gives rise to the acquisition of native structure and is dependent on at least partial unfolding of the $I^{BP}{}_C$ species, the rate-limiting step in the conversion of U_C to N_C. However, at higher urea concentrations where the intermediate is destabilized, the direct refolding of U_C to N_C becomes rate-limiting as evidenced by the roll-over of the slow refolding phase (Fig. 3.3b). The acceleration of the slow phase by cyclophilin can be explained by its dependence on the flow of material from the preceding isomerization reaction. Unlike in CheY,[37] (Fig. 3.17c), the slow equilibration to an alternate native

Figure 3.17 – Folding Mechanism of CheY-like proteins. Mechanism for **(a-c)** refolding of NT-NtrC, Spo0F and CheY, respectively, under strongly refolding conditions and **(d-f)** unfolding under strongly unfolding conditions predicted by the off-pathway model. The progress of the reaction is shown as thick arrows, while the reactions not accessible under the respective conditions are represented by thin gray arrows. The rate-limiting reactions are shown as broken and dotted lines, and the minor channels are shown as broken lines. **(a)** Refolding of NT-NtrC. The dominant unfolded state with the K104-P105 bond in the trans isomer, U_T, collapses within the burst-phase of stopped-flow instrumentation (~ 5 ms) to an off-pathway intermediate, $I^{BP}{}_T$. Isomerization of the prolyl bond gives rise to the fast refolding phase followed by the slow phase corresponding to at least partial unfolding of the intermediate to access the productive TSE between U_C and N_C. A small contribution to the burst-phase reaction from the minor unfolded population, U_C is also shown. **(b)** Refolding of Spo0F. The progression of events is identical to that of NT-NtrC, with the exception that the native state slowly isomerizes to an alternate native state, N_T, which is populated to ~ 5% at equilibrium. **(c)** Refolding of CheY. The isomerization reaction of the $I^{BP}{}_T$ intermediate in CheY is significantly slower than that observed in the other two proteins. This reaction gives rise to the only observable refolding phase that masks all subsequent reactions. A small fraction of the intermediate can also fold to the N_T state, which is populated to ~ 15% at equilibrium. **(d)** Unfolding of NT-NtrC. Under strongly unfolding conditions, the native state unfolds globally by a single unfolding phase. The acquisition of the equilibrium population of the U_T state is optically silent. **(e)** Unfolding of Spo0F is similar to that of NT-NtrC. An additional small amplitude unfolding phase is explained by the independent unfolding of a small population of the N_T state, similar to that observed during the unfolding of CheY **(f)**.

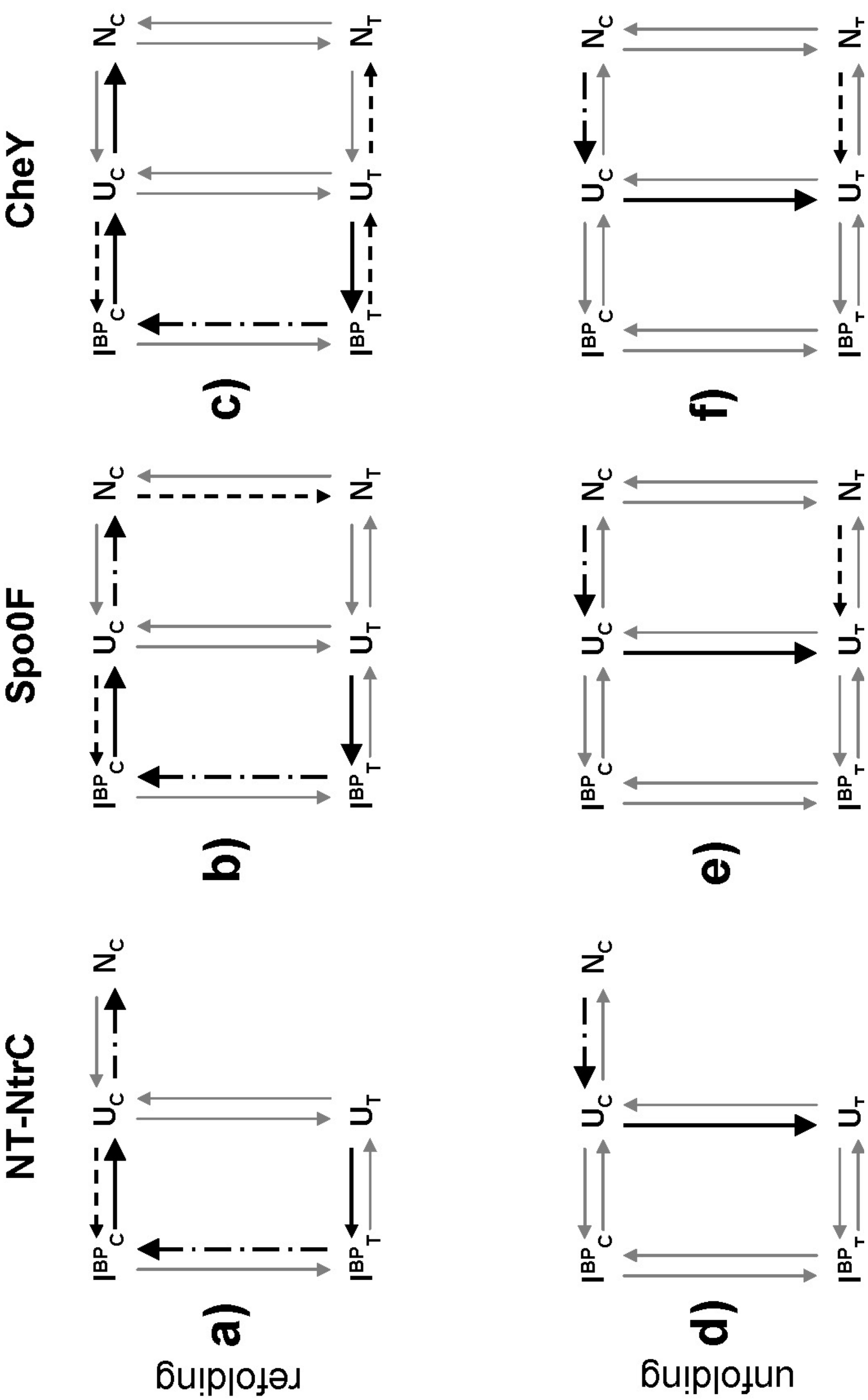

NT-NtrC
Spo0F
CheY
a)
refolding
b)
c)
d)
unfolding
e)
f)
$I^{BP}{}_{C}$
U_C
N_C
$I^{BP}{}_{T}$
U_T
N_T

Figure 3.17

state N_T during refolding is not observed in NT-NtrC. The refolding reaction in Spo0F, under strongly refolding conditions (Fig. 3.17b), is similar to that observed in NT-NtrC, with the exception that the native state, N_C, slowly isomerizes to an alternate conformation, N_T, that is also seen in CheY.[37] CheY differs from both NT-NtrC and Spo0F in that the slow refolding phase corresponds to the prolyl isomerization reaction and the fast refolding reaction to the refolding of U_C to N_C (Fig. 3.17c).[37]

Under strongly unfolding conditions, N_C, the only measurably populated state in NT-NtrC (Fig. 3.17d) and the dominant native state in Spo0F (Fig. 3.17e) and CheY (Fig. 3.17f),[37] rapidly unfolds to U_C. The subsequent slow equilibration to the dominant unfolded state, U_T, is spectrally silent in all three proteins. A small amplitude unfolding phase corresponding to the unfolding of the minor N_T population is also observed in Spo0F (Figs. 3.3f and 3.17e) and CheY (Fig. 3.17f).

The progressive increase in the stability of the N_C state for CheY, 5.4 kcal mol ^{1}, to Spo0F, 6.0 kcal mol ^{1}, to NT-NtrC, 7.5 kcal mol ^{1}, provides a rationale for the inverse correlation with the fractional population of the N_T state, 20%, < 10% and 0%. The higher stability of the N_C state also appears to be reflected in the higher stability of the $I^{BP}{}_C$ state in NT-NtrC and Spo0F relative to the same species in CheY. While the enhanced stability accelerates the isomerization reaction ~ 100-fold in these two proteins,[156,157] (2.10 s ^{1} in NT-NtrC, and 3.69 s ^{1} in Spo0F) (Table 3.3), relative to CheY, (0.08 s ^{1}),[37] it impedes access to the TSE that is primarily structured in the N-subdomain and distal to the site of prolyl bond isomerization. Access to the TSE for the $I^{BP}{}_C$ to N_C reaction ($k_{(U_C \rightarrow N_C)} \times k_{(I^{BP}{}_C \rightarrow U_C)} \div k_{(U_C \rightarrow I^{BP}{}_C)}$) (Table 3.3) is slowest for NT-NtrC

(2.02×10^{2} s^{1}), intermediate for Spo0F (2.27×10^{2} s^{1}) and fastest for CheY (6.40×10^{1} s^{1}).[37] This anti-Hammond behavior,[158] of inverse correlation of refolding rate with stability (Table 3.1) of the native state possibly relates to the packing densities of the folding nuclei (see below).

Gō-model simulations

The simulations of NT-NtrC and Spo0F reveal the development of a productive folding nucleus in their N-subdomains. However the sequence of events and the extent of topological frustration differ from that seen in CheY.[62] In NT-NtrC the reaction proceeds from N-subdomain to the C-subdomain, while in Spo0F topological frustration in the C-subdomain, precedes the appearance of the folding nucleus at the interface of the N- and C-subdomains. In CheY, C-subdomain frustration is readily apparent by both experiment[37] and simulation.[62] However, the site of nucleation is distinctly restricted to the N-subdomain for the subsequent productive folding reaction.

As noted above, CheY[138] and NT-NtrC[136] contain an alanine-rich cavity between α4 and β4β5, which is flexible in the inactive state before undergoing rearrangement upon phosphorylation. Sequence alignment and structural visualization of the three proteins reveals that the same region in Spo0F is filled in with several bulkier side-chains (Fig. 3.1c, 3.1d and 3.1e). As a consequence of these sequence variations, CheY, NT-NtrC and Spo0F have 0.83, 0.88 and 1.06 native contacts per residue in this cavity. Comparing the overall amino acid abundances for the three proteins, they have rather similar numbers of each of the twenty amino acids with the exception that Spo0F has half as many alanines (7 in Spo0F vs. 16 in CheY and 15 in NT-NtrC) and twice as many isoleucines (15 in Spo0F vs. 6 in CheY and 8 in NT-NtrC) as NT-NtrC and CheY (Table 3.4). Several of the alanine replacements by bulkier side-chains occur in the cavity between α4 and β4β5 (Fig. 3.1d and 3.1e).

Topological frustration in Spo0F can be understood in terms of a high density of native contacts in β4α4β5. This property drives the early development of structure in this region as evidenced by the high $Q_{\mathrm{C\ subdomain}}$ contacts, > 0.6, while the $Q_{\mathrm{N\ subdomain}}$ contacts are low, < 0.4 (Fig. 3.12b and 3.15c). The absence of C-subdomain contacts when $Q_{\mathrm{N\ subdomain}}$ increases to 0.4 (Fig. 3.15c) suggests that disruption of preformed structure in the C-subdomain of Spo0F is required to access the productive folding TSE. Indeed, the disruption of C-subdomain contacts is seen to coincide with the appearance of N-subdomain folding in kinetic simulations. The N-subdomain, with its greater density of native contacts, eventually out-competes the C-subdomain and induces its local unfolding. The shift in the location of folding initiation in Spo0F to β3α3β4 can also be understood in terms of the packing density. Spo0F contains approximately 82% more contacts between helices α2 and α3 and 39% more contacts between strands β3 and β4 than CheY and NT-NtrC. The lower density of contacts in the β1α1β2 module is consistent with the flexibility of α1 observed in NMR relaxation measurements of the unphosphorylated state of Spo0F.[140] Finally, a kinetic trap in the C-subdomain of NT-NtrC was not observed in the simulation results, possibly due to the lower packing density of the C-subdomain of NT-NtrC when compared to the packing densities of the corresponding regions in CheY and Spo0F. Thus, the density of native contacts in the

Table 3.4 – Amino acid occurrences for NT-NtrC, CheY and Spo0F.

Protein	Ala	Arg	Asn	Asp	Cys	Glu	Gln	Gly	His	Ile
NT-NtrC	15	6	2	11	1	6	6	7	3	8
CheY	16	4	8	8	0	11	2	10	0	6
Spo0F	7	5	6	11	0	10	5	7	1	15

Protein	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	Total
NT-NtrC	14	4	5	3	6	7	5	2	3	10	124
CheY	15	11	6	6	3	4	5	1	2	10	128
Spo0F	14	11	7	4	4	2	4	0	4	7	124

Table 3.4

C-subdomain appears to be a good predictor of early misfolding reactions in the $(\beta\alpha)_5$ motif.

Kinetic traps in βα-repeat proteins

βα-repeat proteins belonging to the TIM barrel family, α subunit of tryptophan synthase (αTS) from *E. coli*,[59] indole-3-glycerol phosphate synthase (IGPS) *Sulfolobus solfataricus* (sIGPS)[58] and a hypothetical protein IOLI from *Bacillus subtilis*,[74] and the flavodoxin fold proteins, *E. coli* CheY,[37] apo-flavodoxin from *Anabaena sp.*[35] and from *Azotobacter vinelandii*,[159] experience early kinetic traps during folding. The local-in-sequence/local-in-space structure characteristic of these motifs provides ready access to intermediates that cannot directly access the native conformation.

Sequence local clusters of branched aliphatic side-chain (BASiC) residues, isoleucine, leucine and valine (ILV), have previously been implicated in the formation of off-pathway intermediates in CheY[37] (Fig. 3.18a) and in TIM barrel proteins.[58,59,74] The correlation between the location, size and connectivity of clusters of ILV side-chains and the predicted location of prematurely-formed structure in both $(\beta\alpha)_5$ and $(\beta\alpha)_8$[37,59,62] proteins is consistent with an important role for ILV clusters in the appearance of kinetic traps. These branched aliphatic side-chains, along with alanine and glycine, are the only side-chains that do not spontaneously transfer from the vapor phase to water.[103] Therefore, these clusters are especially resistant to fluctuations that would allow the penetration of water into the cluster or to the underlying peptide linkages that support the hydrogen-bonding networks in β-strands and α-helices. The synergy between the tertiary

Figure 3.18 – Cluster analysis of CheY-like proteins. Clusters of branched aliphatic side-chain residues in **(a)** NT-NtrC (1DC7.pdb[122]), **(b)** Spo0F (1SRR.pdb[123]) and **(c)** CheY (3CHY.pdb[90]). Cartoon representation of the NMR solution structure of NT-NtrC and the crystal structures of Spo0F and CheY are shown. α-helices are colored cyan, β-strands are magenta and loops are in light orange. ILV residues that bury greater than 10 Å^2 by contacting other ILV residues are highlighted, and the VDW surfaces of the heavy atoms of these residues are shown as spheres. Two major clusters of ILV residues are observed in all three proteins, one on either side of the central β-sheet. The cluster on the side facing helices α2, α3 and α4 is designated Cluster 1 and is colored blue, while the cluster on the side facing helices α1 and α5 is designated Cluster 2 and is colored red. An additional group of four ILV residues (light blue) that appears contiguous with Cluster 1 is observed in Spo0F and is considered to be a part of Cluster 1. Cluster 1 in all three proteins comprises residues that are closer in sequence that those in Cluster 2.

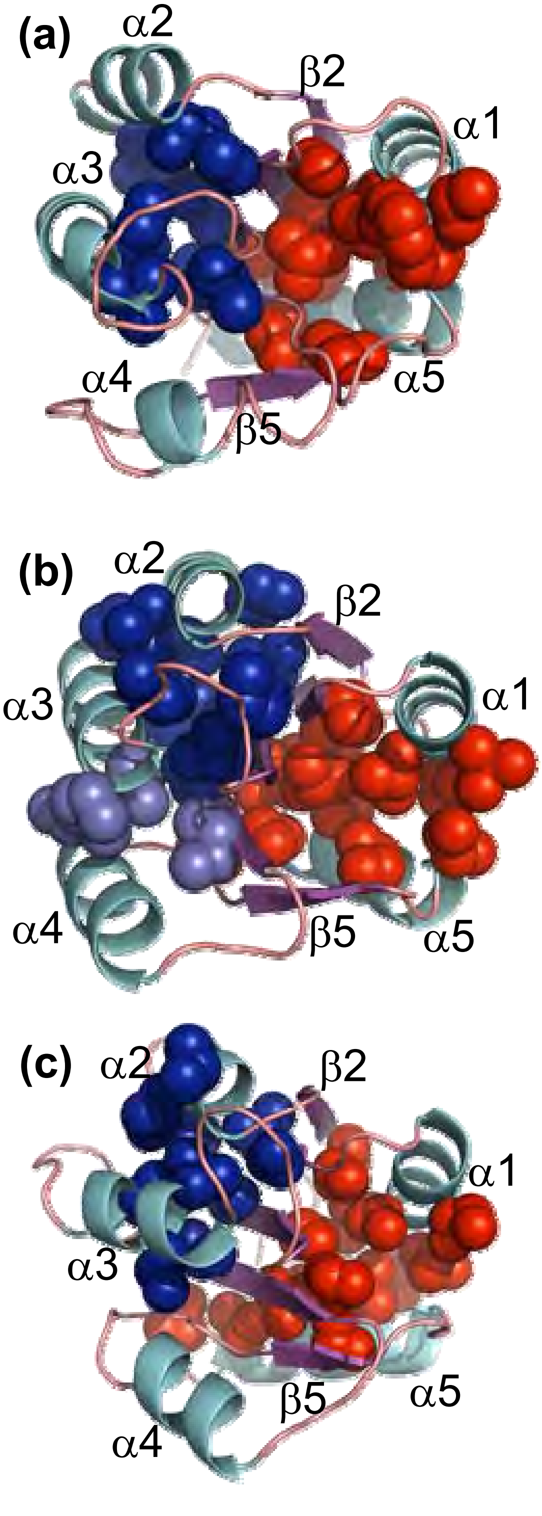


Figure 3.18

and secondary structures, mediated by the exclusion of water from the peptide linkages,[104] would provide a molecular explanation for the two-state cooperativity observed for many proteins. Locally-connected ILV clusters of sufficient size appear to be able to drive off-pathway unproductive folding reactions while both local and non-local clusters could serve to stabilize the native conformations of their resident proteins. Because NT-NtrC and Spo0F also have ILV clusters fusing both of the helical layers to the intervening β-sheet, it was of interest to explore their relationship to the folding mechanisms of both proteins.

For NT-NtrC, a large cluster of 16 ILV side-chains from β1, β3, β4, β5, α1 and α5 and a total buried surface area (BSA) of 1219 Å^2 is observed on one face of the β-sheet (Fig. 3.18b). A smaller cluster of eight side-chains from β1, β3, β4, α2 and α3, is also observed on the opposing face of the β-sheet. While the latter cluster only buries a total of 407 Å^2 and does not reach the cut-off previously chosen to define a stable cluster, 10 side-chains and a BSA of 500 Å^2,[37] it was retained in the analysis for comparison with the corresponding cluster in CheY.[37] Retaining the nomenclature adopted for CheY, the cluster on the α2, α3 and α4 side of the β-sheet in NT-NtrC is designated as Cluster 1 and the cluster on the α1 and α5 face of the β-sheet is designated as Cluster 2.

In Spo0F (Fig. 3.18c), Cluster 1 is comprised of 12 ILV side-chains from β1, β3 and β4 and helices α2 and α3, and it buries 655 Å^2. Cluster 2 is comprised of 13 ILV side-chains from β1, β3 and β4 and helices α1 and α5, and it buries 731 Å^2. These two clusters resemble CheY in their size and the elements of secondary structure involved. However, an additional group of four residues that is adjacent to Cluster 1 is also seen in

Spo0F (Fig. 3.18c). While none of the individual residue-residue contacts between members of this smaller group and those in Cluster 1 bury more than the 10 Å^2, previously defined for participation in a cluster,[37] the total surface area buried between the two clusters is greater than 50 Å^2. Moreover, the continuity of secondary structure elements between Cluster 1 (β1, α2, β3, α3 and β4) and the small cluster represented by these four residues (α3, β4 and α4) suggests that both should be considered in Cluster 1 (Fig. 3.18c).

In assessing the roles of these clusters in the folding reactions of NT-NtrC and Spo0F, it has proven to be useful to include information on the sequence disposition of the side-chains in the clusters. While the BSA provides a measure of the hydrophobicity and the vdW energy contributions to stability, the connectivity reflects the conformational entropy penalty required to form the mutual ILV contacts. The absolute contact order (ACO) algorithm, developed previously for CheY,[37] is based upon earlier algorithms devised by Baker and his colleagues,[40,41] and is similar in spirit to other approaches,[45] and provides a useful metric for comparing the clusters in NT-NtrC, Spo0F and CheY (Table 3.5). The striking difference in the connectivity of Clusters 1 and Clusters 2 in all three proteins suggests that these two clusters may play different roles in the folding of these proteins. The low ACO values for Cluster 1 in all three proteins (Table 3.5) may drive the early formation of non-productive intermediates. The premature formation of these clusters may preclude direct folding to the native state because the side-chains in Cluster 2, on the opposing face, require the formation of non-local interactions. Although the unfavorable kinetic competition with Cluster 1 formation

Table 3.5 – Cluster analysis of CheY-like proteins

Clusters represent networks of Isoleucine, Leucine and Valine side-chains that are in contact with each other and bury a surface area of 10 Å^2 or more per contact.

[a] The number of residues that bury more than 50 Å^2 of their side-chain atoms by contacts within the cluster are shown in parentheses.

[b] The average surface area of each residue buried within the cluster is shown in parentheses.

[c] The average surface area buried by each contact between two residues is shown in parentheses.

Protein	NT-NtrC		Spo0F		CheY[16]	
Cluster	Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 1	Cluster 2
Secondary structure elements	β1, β3, β4, α2 & α3	β1, β3, β4, β5, α1 & α5	β1, β2, β3, β4, α2, α3 & α4	β1, β3, β4, β5, α1 & α5	β1, β3, β4, α2 & α3	β1, β2, β3, β4, β5, α1 & α5
# of residues (> 50 Å^2)[a]	8 (4)	15 (11)	16 (9)	13 (8)	10 (7)	15 (8)
BSA Total (BSA/residue)[b]	407 (50.8)	1050 (70.0)	870 (54.4)	731 (56.3)	608 (60.8)	838 (55.9)
ACO (BSA/contact)[c]	23.33 (18.3)	45.26 (27.2)	23.88 (34.8)	47.56 (29.0)	21.64 (27.6)	36.09 (26.2)

Table 3.5

may preclude access to the productive TSE, the larger sizes of Cluster 2 may ultimately drive the folding reaction to the native conformation.

While the sources of frustration are readily apparent in coarse-grained simulations of folding for CheY and Spo0F, the simulations do not report this behavior for NT-NtrC. It is interesting to speculate that the smaller size of Cluster 1 in NT-NtrC does not provide a sufficient contribution to the simplified potential function to be apparent in the Gō-model simulations.

Topology vs. sequence and folding energy landscapes

As has been observed previously for other motifs, [120,121,160 162] both topology and sequence contribute to defining the folding free energy landscapes for the three CheY-like proteins examined in the present study. Similar two-state thermodynamic behavior and off-pathway folding intermediates have also been observed for a pair of proteins with the closely-related flavodoxin fold.[35,159] The misfolding reaction may be an inherent property of the βα-repeat motif that is a defining feature of the native conformation and, therefore, cannot be eliminated through evolution.

Similarities in the folding free energy landscape of structurally homologous proteins is expected to arise from constraints of their shared topology, which in turn may be defined by a set of conserved, structurally important residues. This core set of residues may serve not only to maintain the native state topology, but also to direct the rapid folding of the polypeptide chain to the native state. Analysis of the β-sandwich motif[163] revealed a strong preference for VLIF residues at the interface between a quartet of interlocking β-strands contributed by both β-sheet layers. Mutational analysis of one member of this family, azurin, demonstrated that a sub-set of the equivalent side-chains also are involved in the TSE leading to the native state.[164] A subsequent mutational analysis of another β-sandwich protein, a fibronectin domain, showed that the position of the folding nucleus can vary slightly, depending on variations in the sequence.[161] In both examples, clusters of ILV residues serve to define the β-sandwich motif. A statistical analysis of several different motifs, including CheY,[165] also found preferred conservation of ILV side-chains in their folding nuclei. In this case, the conservation was at the level of the group of branched beta residues, not at the level of the individual amino acids. The conservation of the group of side-chains is quite logical, given the propensity of the ILV side-chains to form clusters of a significant size that can stabilize the underlying hydrogen bond network by the preferential exclusion of water.

These and many other results show that a great deal of variation can be tolerated in sequence space without altering the topology. The sequence variations enable the development of entirely novel functional properties, a case in point being the plethora of reactions catalyzed by the TIM barrel, $(\beta\alpha)_8$, motif.[166] The present study shows, however, that sequence-induced variations on topology-defined folding landscapes can result in substantial redistributions of the flow of protein through partially-folded states during the folding reaction or along folding trajectories. Thus, the variations in sequence that support functional divergence can also modulate folding mechanisms that are primarily defined by the topology.

Materials and methods

Protein expression and purification

The expression plasmid pJES820 with the gene encoding NT-NtrC was obtained from Dr. David Wemmer at UC Berkeley, and the plasmid pET20 with the gene encoding Spo0F was obtained from Dr. James A. Hoch at the Scripps Research Institute. The DNA sequence was confirmed at the UC Davis sequencing facility. The *Escherichia coli* strain BL21 Codonplus®(DE3)RIL was used for expression of NT-NtrC and BL21(DE3)PlysS® was used for expression of Spo0F. Both 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 and 4 ºC. The refolded protein was concentrated, applied to a Q Sepharose column and eluted using a salt gradient from 0 to 400 mM NaCl for NT-NtrC and 0 to 200 mM NaCl for Spo0F. Further purification was done using a Sephadex® G-75 gel filtration column in 10 mM potassium phosphate at pH 7.0. The identity and purity (> 98%) was confirmed using nano-spray mass-spectrometry at the Proteomics Facility at the University of Massachusetts Medical School. An extinction coefficient of 14060 M ^{1}cm ^{1} at 280 nm and 7000 M ^{1}cm ^{1} at 275 nm was used for NT-NtrC[167] and Spo0F,[168] respectively, to determine the protein concentration.

Stability analysis

Samples of 10 μM NT-NtrC in 10 mM potassium phosphate at pH 7.0 were equilibrated overnight in 0 M to 8 M urea at concentration increments of 0.2 M urea. The far-UV CD spectra of each sample at 25 °C, using a 1cm cuvette in a Peltier-style thermostatted sample compartment, were recorded on a JASCO model J810 CD spectrophotometer. The CD spectra were recorded between 215 nm and 260 nm, with a band width of 2.5 nm, and a step size of 0.5 nm, integrated for 1 s and averaged over three traces. The measurements were repeated twice, and the reversibility of the reaction was confirmed by coincidence of the equilibrium transition curve obtained by starting from the unfolded state in 8 M urea. The steady-state FL emission spectra of 8 μM NT-NtrC under similar conditions to the CD equilibrium titration were recorded between 295 nm and 500 nm at a 1 nm interval, after excitation at 290 nm using a T-format Horiba Fluorolog fluorimeter. After correcting the spectra for contributions from the buffer, the transition curves at 222 nm for CD and 315 nm for FL emission were plotted as a function of urea concentration and fitted to a two-state model, N ⇆ U, where N is the native form of the protein and U is its denatured form. The free energy change associated with unfolding in the absence of denaturant was determined by assuming a linear dependence of the apparent free-energy change on the denaturant concentration.[112,113]

$$\Delta G^{\circ}_{[\mathrm{Urea}]} = \Delta G^{\circ}_{H_2O} - m[\mathrm{Urea}] = -RT\ln\left(K_{\mathrm{eq[Urea]}}\right) \qquad \textbf{(1)}$$

where ΔG°(H_2O) is the standard unfolding free energy change in the absence of urea, ΔG°[Urea] is the standard unfolding free energy change at any urea concentration, [Urea], and *m* is its dependence on the concentration of urea.[112,113] A nonlinear regression analysis module of the software Savuka[34] was used to fit the data to this model. Fitting to a three-state model did not improve the fit significantly. The two-state fit was confirmed by globally fitting the FL and CD data across all wavelengths using singular value decomposition (SVD) vectors (for description, see Ionescu *et al*.[114] and Gualfetti *et al*.[115] and references therein). Only two significant vectors were observed.

Similar unfolding and refolding equilibrium titrations between 0 M and 8 M urea were performed for 5 μM samples of Spo0F using CD and 10 μM samples using steady state FL emission after excitation at 280 nm. The data were fitted using the method described above.

Nuclear magnetic resonance

NMR experiments were performed using a Varian Unity INOVA 600 operating at a ^{1}H frequency of 599.7 MHz and equipped with an inverse triple-resonance cryogenically cooled probe and preamplifier. The sample for NT-NtrC, 2.3 mM solutions of the Ac-P103-K104-P105-F106-D107-NH_2 pentapeptide, and Spo0F, 3.2 mM solutions of the Ac-A103-K104-P105-F106-D107-NH_2 pentapeptide, were obtained from New England Peptides Inc., dissolved in 10 mM potassium phosphate buffer (pH 7.0) containing 10% D_2O. Spectra were obtained using a simple pulse-acquire sequence with excitation sculpting for solvent suppression.[116] A total of 16,000 complex data points were recorded over a spectra width of 8 kHz. All data were apodized with 0.5-Hz exponential line broadening prior to Fourier transformation. All data were processed using the NMRpipe software package.[117]

Kinetics

Fluorescence

The change in FL emission associated with refolding or unfolding of NT-NtrC was monitored using an Applied Photophysics SX 17MV instrument (dead time 2 ms). The excitation wavelength was 290 nm, and emission was monitored using a 320 cut-off filter. The relaxation times and the associated amplitudes were calculated by fitting the kinetic data to the equation

$$A(t) = A(\infty) + \sum_{i=1}^{n} A_i \exp\left(-t / \tau_i\right) \quad \textbf{(2)}$$

where $A(\infty)$ is the observed signal at infinite time, $A(t)$ is the observed signal at time t, A_i is the signal and τ_i is the relaxation time associated with phase (i) and n is the number of exponentials. The kinetic data were fit to a series of exponentials using an in-house non-linear least squares fitting program, Savuka.[34] The logarithm of the relaxation time was plotted as a function of final denaturant concentration in the form of a chevron analysis.[96]

Refolding: NT-NtrC was equilibrated overnight in 7.4 M urea and 10 mM potassium phosphate at pH 7.0 and at 25 °C and refolded by rapid mixing into refolding buffer with varying final concentrations of denaturant (1 M urea to 4 M urea), and 10 μM final protein concentration. The same experiments were performed with 10 μM Spo0F equilibrated in 6 M urea and 10 mM potassium phosphate at pH 7.0 and at 25 °C

Unfolding: NT-NtrC was unfolded by rapid mixing into high concentration of urea buffered with 10 mM potassium phosphate at pH 7.0 and 25 °C, to final urea concentrations ranging from 2.5 M to 8.0 M, and 10 μM protein. The same experiments were performed with 10 μM Spo0F in 10 mM potassium phosphate at pH 7.0 and at 25 °C.

Circular dichroism

Refolding was also monitored by the far-UV CD ellipticity at 222 nm using an AVIV model 202 stopped-flow CD spectrophotometer (dead time 5 ms) and a JASCO model J810 CD spectrophotometer (manual mixing dead time ~10 s). The conditions for the experiments were as described above, and the data were fitted by the same method used for FL emission.

Stability of burst-phase intermediate: The refolding kinetics upon 10-fold dilution of 100 μM protein unfolded with 6 M urea into refolding buffer were monitored by CD at 222 nm, buffer-corrected and extrapolated to 0 s to determine the signal associated with the burst-phase intermediate. The amplitude of the signal was then plotted against the final denaturant concentration. The sigmoidal unfolding curve was then fitted to a two-state model, $I^{BP} \leftrightarrows U$, where I^{BP} is the burst-phase intermediate and U is the denatured form. The free energy change associated with the unfolding of the intermediate in 0 M urea and its dependence on urea concentration was determined by the method described above.

Global analysis

Both the CD and the FL kinetic traces from the refolding, unfolding and double jump experiments were fit globally to several different models. The Levenberg-Marquart method[97] was then used to obtain the best fit to the kinetic data. Details of the methods used are described previously.[34,37]

Analysis of hydrophobic clusters

The contact surface area between atoms was calculated using the CSU software developed by Sobolev *et al.*[119] The method for analysis of hydrophobic ILV clusters has been described previously.[37] The application of the Absolute Contact Order[41] (ACO) algorithm to the ILV clusters is also described in earlier work.[37]

Coarse-grained simulations

Gō-model simulations were performed with NT-NtrC and Spo0F using the coarse-grained model developed by Karanicolas and Brooks,[47] previously developed for the study of CheY folding.[62] Briefly, the protein backbone is represented as a string of beads connected by virtual bonds. Each bead represents a single amino acid and is located at the α-carbon position. Bond lengths are kept fixed, bond angles are subject to a harmonic restraint, and dihedral angles are subject to potentials representing sequence-dependent flexibility and conformational preferences in Ramachandran space. Nonbonded interactions are represented using a Gō-model in which only residues that are in contact in the native state (taken to be structures 1DC7.pdb[122] and 1SRR.pdb[123] for NT-NtrC and Spo0F, respectively) interact favorably. Backbone hydrogen bonds and side-chain pairs with non-hydrogen atoms separated by less than 4.5 Å interact via a pairwise 6-10-12 potential that consists of an energy well and a small desolvation barrier. To incorporate sequence effects, the interaction energies of side-chain native contacts are scaled according to their abundance in the Protein Data Bank as reported by Miyazawa and Jernigan.[169] Residues not in contact in the native state interact via a repulsive volume exclusion term. A complete description of the model potential and its parameters can be found in Karanicolas and Brooks.[47]

Molecular dynamics simulations were performed in Cartesian space using CHARMM[170] within the *gorex.pl* module of the MMTSB Tool Set.[171] Langevin dynamics with a 1.36 ps ^{1} friction coefficient was used to maintain thermal equilibrium, and the time step was set at 22 fs. For kinetic folding simulations, 100 independent runs were each performed for 2×10^8 dynamics steps at a temperature highly favoring the native state, namely at 0.87 T_f, where T_f is the folding transition temperature defined by the maximum in the heat capacity curve, $C_v(T)$. Note that absolute timescales cannot be obtained due to the coarse-grained nature of the Gō-model and the lack of explicit solvent molecules. Unfolded starting structures for the folding runs were generated by equilibration at 1.5 T_f for 10^7 dynamics steps starting from randomly assigned initial velocities. Conformational snapshots were recorded every 10^5 dynamics steps. The fraction of native contacts formed, Q, was used to monitor 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.

To characterize the entire accessible free energy landscape, a two-dimensional extension of replica-exchange molecular dynamics[172] was performed. Each replica was assigned one of four temperatures (0.87, 0.97, 1.08 or 1.20 T_f) and one of seven harmonic biasing restraints on the radius of gyration, R_g, for a total of 28 replicas. To ensure overlap between the R_g distributions harmonic potentials were used with minima at 1.0, 1.1. 1.2, 1.3, 1.5, 1.7 and 2.0 R_g^0, where R_g^0 is the radius of gyration of the native state, with force constants 0.5, 5.0, 5.0, 5.0, 4.0, 0.8 and 0.5 kcal/mol-Å^2, respectively. Stronger restraints were required at intermediate radii to sample the high energy transition region between the unfolded and native states. Conformational exchanges between temperature windows and restraints were attempted every 40,000 dynamics steps, and the snapshots were recorded. The exchange frequency remained between ~10% and 40% throughout the 6×10^8-step simulation. Finally, conformations were combined from all 28 replicas for a total of 4.2×10^5 structures, and the multidimensional weighted histogram analysis method[173,174] was used to obtain the unbiased free energy at T_f projected along various progress coordinates. The above procedure was carried out in its entirety for both NT-NtrC and Spo0F.

Acknowledgments

Simulations were performed using the modeling package available through the NIH resource (RR12255) Multiscale Modeling Tools for Structural Biology (http://mmtsb.org). The authors would like to thank Dr. David Wemmer and James A. Hoch for generously providing the plasmids with the genes encoding NT-NtrC and Spo0F, respectively. R.D. Hills Jr. thanks the La Jolla Interfaces in Science Training Program for financial support, the Center for Theoretical Biological Physics (www.ctbp.ucsd.edu) for providing a stimulating intellectual environment and D.A. Case for providing laboratory space. S.V. Kathuria thanks Osman Bilsel for guidance in the global analysis, Can Kayatekin and J.A. Zitzewitz for valuable discussions and their help with preparation of the manuscript. This work was supported by NIH grant GM48807 to C.L. Brooks III and NSF grants MCB0327504 & MCB0721312 to C.R. Matthews.

Chapter IV – Clusters of Isoleucine, Leucine and Valine Side Chains Define Cores of Stability in Globular Proteins: Sequence Determinants of Structure, Stability and Folding

This chapter is being prepared as a manuscript for publication in the Journal of Molecular Biology as *Kathuria SV, Matthews CR. "Clusters of Isoleucine, Leucine and Valine Side Chains Define Cores of Stability in Globular Proteins: Sequence Determinants of Structure, Stability and Folding."*

The work presented in the following chapter was a collaborative effort. The BASiC hypothesis was conceived by Dr. C. Robert Matthews. I developed and analyzed the databases. Dr. C Robert Matthews and I wrote the manuscript.

Abstract

Measurements of protection against exchange of main chain amide hydrogens with solvent and mutational analyses of the kinetic and thermodynamic folding properties of globular proteins have provided remarkable insights into the structures of rare high-energy states and of transition state ensembles (TSEs) that guide folding reactions. Lacking, however, has been a unified theory that rationalizes these high-energy states and the TSEs in terms of the structures and sequences of their resident proteins. The **B**ranched **A**liphatic **Si**de **C**hain (BASiC) Hypothesis was recently developed to explain observed patterns of protection for main chain amide hydrogens against exchange with solvent in a pair of TIM barrel proteins. This hypothesis supposes that the side chains of isoleucine, leucine and valine (ILV) residues often form hydrophobic clusters that very effectively impede the penetration of water to their underlying hydrogen bond networks. The linkage between the secondary and tertiary structures enables these ILV clusters to serve as cores of stability in high-energy states. Very good correlations between the locations of ILV clusters and both strong protection against exchange and the positions of folding nuclei for a variety of proteins reported in the literature support the generality of the BASiC hypothesis. The results also illustrate the necessity to elaborate this simple hypothesis to account for the roles of adjacent hydrocarbon moieties in defining stability cores of partially-folded states along folding reaction coordinates.

Introduction

It has been 50 years since Walter Kauzmann published his seminal review of the protein denaturation reaction.[50] His analysis of the various factors involved in stabilizing the native conformation of proteins anticipated an important role for the hydrophobic effect. Based upon solubility data for nonpolar analogs of side chains and the propensity of detergents and organic solvents to denature proteins, he reasoned that aliphatic and aromatic side chains would be preferentially sequestered in the interior of proteins. This view was subsequently verified when the first crystal structure of a protein, myoglobin, appeared in the same time frame.[175] Although a lively debate has ensued about the relative importance of the hydrophobic effect and hydrogen bonding to stability,[104,176,177] it is well accepted that buried and tightly-packed nonpolar side chains are crucial to the stabilization of the native, functional forms of globular proteins.

The contribution of individual nonpolar side chains to the stability of the native conformations of numerous proteins have been examined using mutational analysis.[178 180] Replacements almost invariably decrease the stability of the native conformation, reflecting the global connectivity of the numerous non-covalent interactions that result in highly-cooperative unfolding reactions. The molecular underpinnings of these observations reflect both the relative differences in the transfer free energies of the side chains from nonpolar solvents to water[103] and the loss in the vdW interactions when large nonpolar side chains are replaced by smaller counterparts, typically alanine. At the point that the polypeptide chain reaches the native conformation, essentially all of the buried side chains contribute to its stability.

Where the buried side chains do distinguish themselves is in the formation of transition state ensembles (TSE) and other high energy states that exist in equilibrium with the native state. Structural information on these rare states has been inferred by several different methods, including mutational analyses of kinetic folding reactions (phi analysis[53] and psi analysis[181]) and by measurements of the protection of main chain amide hydrogens against exchange with solvent.[182] HX followed by NMR is a powerful tool to study the role of individual amino acid residues in these high energy states.[183] In favorable cases, one or more partially-folded states, whose size decreases with increasing free energy, have been observed.[58,184] The most resistant amide hydrogens to exchange often correspond to clusters of side chains formed by adjacent elements of secondary structure in the native conformation.

Although the implied structures of these transient and marginally-populated states have been mapped for a large number of proteins,[185] there has not yet been a unifying hypothesis that rationalizes these structures in terms of their constituent side chains. Computer simulations, both coarse-grained Gō-models[47,62,186] and high resolution molecular dynamics simulations,[187,188] have had some success in probing the structures and energetics of these high energy states for small proteins or domains of <100 amino acids. As computing power increases and, perhaps, as force fields become more accurate, it might ultimately be possible to predict the structures and energetics of partially-folded states in larger proteins and protein complexes from their primary structures. Until that time, however, it seems reasonable to suppose that the fundamental principles underpinning protein folding reactions can be further elucidated by experimental methods.

Folding Reactions of TIM barrel Proteins

Recent results from mutational and/or hydrogen exchange experiments carried out on a pair of structurally conserved $(\beta\alpha)_8$, TIM barrel proteins, the alpha subunit of tryptophan synthase (αTS) and indole-3-glycerolphosphate synthase (IGPS), provided significant insights into the role of sequence in determining the cores of stability. The individual replacement of 10 leucines, isoleucines or valines in a large cluster of 31 of these aliphatic side chains, near the N-terminus of αTS and between the β-barrel and the α-helical shell, eliminated an off-pathway sub-millisecond folding intermediate.[59] Mutations in other, smaller ILV clusters slightly destabilized but preserved this partially-folded state. Hydrogen exchange NMR studies revealed preferential and strong protection in the same region identified as crucial to the off-pathway intermediate by the mutational analysis (Fig. 4.1a).[189] Hydrogen exchange mass spectrometry analysis of a similar early off-pathway species in IGPS[36] and the subsequent on-pathway equilibrium intermediate[58] showed strong protection in peptic peptides whose ILV side chains form a pair of adjacent hydrophobic clusters (Fig. 4.1b) in a different region of the structure than that for αTS. Because the HX protection is not uniform throughout the barrel and is not associated with a specific set of $\beta\alpha$-repeat units in these two proteins, it is clear that the protection does not simply reflect the $(\beta\alpha)_8$ topology. Although the ILV composition of β-strands in $(\beta\alpha)_8$ barrels, ~ 40%,[190] is significantly more than in other regions of globular proteins, ~ 20%, the protection against exchange occurred preferentially in

Figure 4.1 – HX patterns in TIM barrels. a) The measured hydrogen-exchange free energies, 6.5 to 11.7 kcal mol^{-1}, in 0 M urea, mapped on the crystal structure of αTS.[189] The figure is generated using PyMOL from the Protein Data Bank file 1BKS for αTS. The residues colored in purple offer the strongest protection, followed by blue, green, yellow and red. The ILV cluster is shown in spheres. b) Two-dimensional contact map of the ILV cluster of sIGPS (PDB Id: 2C3Z). The circles represent contacts made between any two residues represented on the x- and y-axes. The size of the circle is proportional to the amount of surface area buried by the contacting residues. The two clusters are color coded red and black circles. The blue box encompasses the most strongly protected region and represents structured regions in the on-pathway I_a intermediate, and the red box encompasses regions that are at least moderately protected and represents structured regions in the on-pathway I_b intermediate. The purple box represents the most dense region of sequence local contacts in the protein, which is predicted to be the driving force for the off-pathway intermediate.[58]

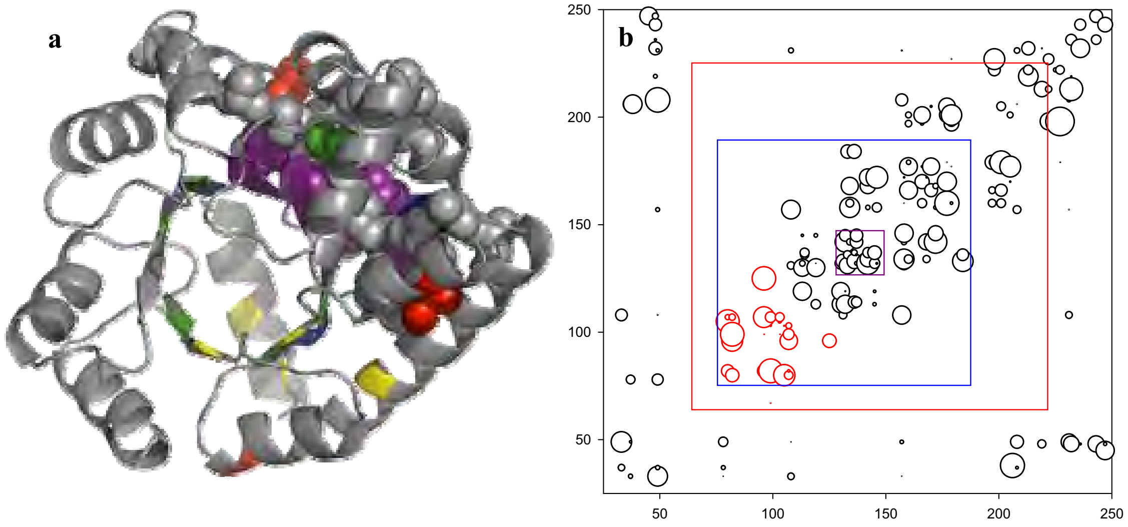


Figure 4.1

segments corresponding to large clusters of ILV side chains (>10 residues), not to those in smaller clusters dispersed throughout the structures.

The Hydrophobic Effect

Hydrophobicities of amino acids are typically determined by measuring the partitioning of the amino acids between a nonpolar solvent and water. The nonpolar solvents have included cyclohexane and 1-octanol, among others,[191] and, depending upon the non-aqueous phase, the relative hydrophobicities differ to some extent from scale to scale (Fig. 4.2). In addition to this inherent variability, the difficulty in applying any single scale is its inability to accurately reflect the complex and heterogeneous interior of a protein. Isoleucine, leucine and valine are among the top-ranked hydrophobic residues on all of the scales,[103,192] but ILVs are generally not considered to make the strongest contribution to the hydrophobic effect stabilizing proteins. Rather, the aromatics, phenylalanine, tryptophan and tyrosine, are commonly thought to be the most hydrophobic amino acids. On the basis of the empirical hydrophobicities of the individual side chains, therefore, it was not obvious why large clusters of ILV side chains might be responsible for stabilizing the off-pathway intermediates in αTS and sIGPS or for protection against HX in the stable on-pathway intermediates for αTS and IGPS.

Insight into a possible explanation for this behavior was obtained from the results of all-atom molecular dynamic simulations in explicit solvent. Chandler and his colleagues[106] have long studied dewetting transitions in model systems and had speculated that similar behavior might be observed during the docking of hydrophobic

Figure 4.2 – Hydrophobicity scales a) energetic penalties for solvation of different residues from a) Octanol to water and b) Cyclohexane to water c) Kyte and Doolittle,[193] used a combination of amino acid distribution in protein interiors and the energetic penalties calculated for solvation of residues from the gas phase by d) Wolfenden 1981[103].

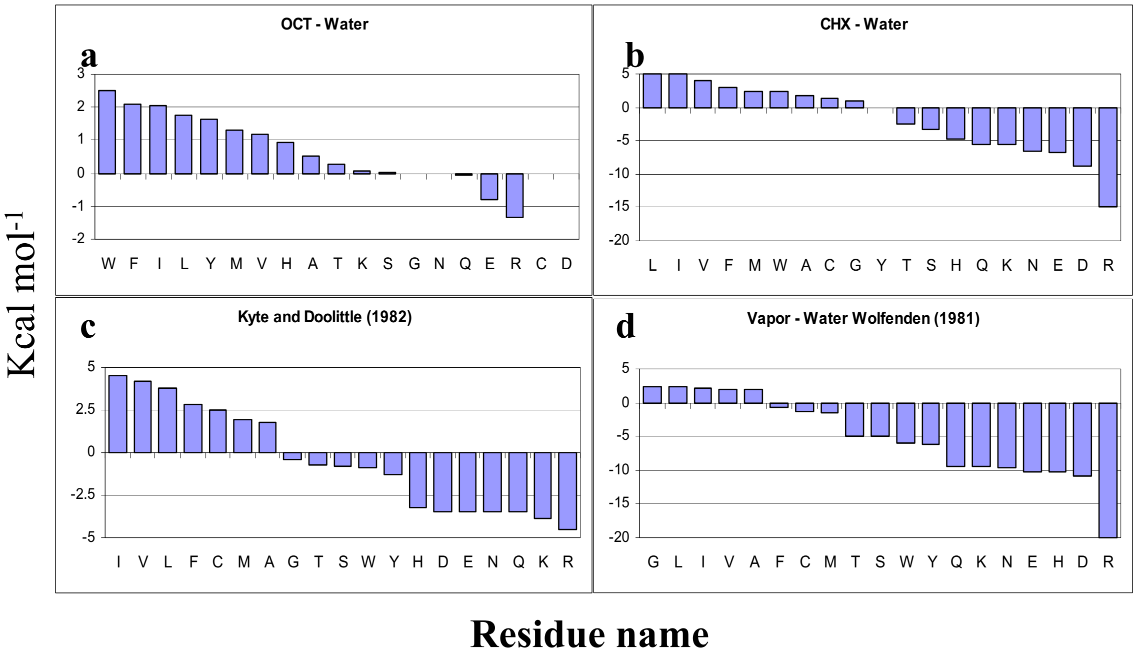

OCT - Water
a
3
2
1
0
-1
-2
W F I L Y M V H A T K S G N Q E R C D
CHX - Water
b
5
0
-5
-10
-15
-20
L I V F M W A C G Y T S H Q K N E D R
c
Kyte and Doolittle (1982)
5
2.5
0
-2.5
-5
I V L F C M A G T S W Y H D E N Q K R
d
Vapor - Water Wolfenden (1981)
5
0
-5
-10
-15
-20
G L I V A F C M T S W Y Q K N E H D R
Kcal mol-1
Residue name

Figure 4.2

surfaces in protein folding and/or association reactions. An initial simulation of the docking of the hydrophobic interfaces for the two domains in a dioxygenase, BPHC from *Pseudomonas sp.*, in the Berne lab only showed a small decrease in the density of the intervening solvent prior to docking.[107] Another simulation on tetrameric melittin,[194] however, showed a dramatic decrease in the water density between the two pairs of chains (the four α-helix complex was initially separated into two-chain pairs) as they approached each other. The dewetting transition preceded the subsequent collapse of the chains into the complex. Inspection of the amino acid compositions for the domain-domain interface in the dioxygenase and the four-helix interface for melittin showed a significant difference. While the dioxygenase interface was a heterogeneous mixture of nonpolar side chains, the melittin interface was almost entirely composed of ILV side chains: 4 leucines, 3 isoleucines, 2 valines and 1 tryptophan per chain. Individual replacements of the isoleucines with smaller aliphatic side chains, alanine, glycine and valine, showed that dewetting was retained for replacements at two of the isoleucines but not the third. Thus, robust dewetting in the melittin tetramer requires a substantial aliphatic surface. A subsequent molecular dynamics analysis of dewetting transitions in other multimeric protein complexes showed a preference in dewetting for those whose interface is dominated by ILV residues.[195]

The BASiC Hypothesis

The close association of large clusters of ILV side chains with the cores of stability in partially-folded states of TIM barrel proteins[58,59] has led to the proposal of the

Branched Aliphatic amino acid Side Chain (BASiC) Hypothesis. In its original formulation,[59] this hypothesis supposed that clusters of 12 or more branched aliphatic side chains, whose self-contact density exceeded 2 contacts per side chain, are primarily responsible for the stabilization of structure in high energy states that are in equilibrium with the native TIM barrel state. A subsequent survey of 55 non-redundant TIM barrel proteins suggested that a minimum of 10 ILV side chains with at least 500 Å^2 of mutual contact surface area is required for a stable cluster (See below).[37]

The molecular basis of this hypothesis resides in the uniquely unfavorable interactions of saturated hydrocarbons with water.[176] Radzicka and Wolfenden[103] measured the partitioning of side chain analogs of the 20 naturally-occurring amino acids between the vapor phase and water. Only alanine, glycine, isoleucine, leucine and valine have a positive free energy of transfer from the vapor phase to water. All other side chain analogs, including those for phenylalanine, tyrosine and tryptophan, spontaneously dissolve in water. The iso-butyl, n-butyl and prolyl groups for isoleucine, leucine and valine, however, would be more effective than the methyl group of alanine or the hydrogen of glycine at excluding water from the main chain amide group and have a greater capacity for stabilizing vdW interactions. The unique properties of ILV residues are evident in their preferential sequestration in the interior of globular proteins[192,196] and, for isoleucine and valine, in the unusually potent protection against exchange with solvent for their associated amide hydrogens in simple peptides.[197]

Interestingly, a further consequence of burying any residue in the interior of proteins is the ~10% decrease in the volume of its peptide moiety.[104] The absence of water removes a source of competition for the N-H donor and C O acceptor and decreases the dielectric constant of the surrounding medium. As a result, the strength of the main chain-main chain hydrogen bond is increased. A cluster of ILV side chains in TIM barrel proteins would be expected to strengthen a network of main chain-main chain hydrogen bonds that define the β-barrel and the surrounding shell of α-helices. The decrease in the volume of the peptide group could, in turn, force a closer packing of the aliphatic side chains, enhancing their vdW interactions. The more tightly-packed ILV cluster would even more effectively exclude water from the underlying hydrogen bond network. The resulting synergy between the tertiary and secondary structures, based upon the exclusion of water from the polypeptide backbone, could be a significant factor in defining the two-state cooperativity that is the hallmark of globular proteins. Other nonpolar side chains, e.g., phenylalanine, tyrosine or tryptophan, or the hydrocarbon portions of polar side chains, e.g., those linking the charged groups in lysine or arginine to the Cα carbon, could be associated with these clusters. However, these alternative side chains would play a supportive role in stabilizing these hydrophobic clusters and may serve as an amphipathic interface between the ILV clusters and the solvent in partially-folded states. Ultimately, water is excluded from almost all buried non-polar side chains, including the aromatics, methionine and cysteine.

Definition of clusters

Previous work on αTS suggested that the residues involved in the formation of an off-pathway intermediate and most protected during native state HX were clustered together in the protein. The definition of this cluster was based on presumed favorable interactions between BASiC residues within vdW distances of each other. A distance of 4.2 Å was used to define favorable vdW interactions. A cluster of ILV residues with a minimum of 12 interconnected residues and involved in least two such favorable interactions was considered to be sufficient to guide the folding reaction. One of the limitations of this method is that the distance cutoff criterion does not recognize the orientation of interaction between the two residues. This parameter can significantly alter the nature of the interaction between the residues.

Subsequently, a more rigorous analysis of the clusters of ILV residues was performed on a non-redundant set of 55 TIM barrel proteins (data not shown). To distinguish between strong and weak interactions, the surface area buried between two residues was calculated using the CSU software. The interaction network between the ILV residues was analyzed as a function of the amount of surface area buried (SAB) between residues. As the SAB cutoff is increased, the weakest links disappear rapidly. Only a stable set of interactions remain that form the clusters which are relatively unaffected by minor changes in the cutoff. As the cutoff is increased further, the clusters begin to fragment and eventually disappear altogether. The expected result is represented in Figure 4.3a and a few representative results from the 55 TIM barrel database is shown in Figure 4.3b. A cutoff of 10 Å^2 appears to represent a stable interaction and is used for further evaluations.

A recent study on the formation of amyloid nuclei,[198] suggests that a minimum of 10 interconnected hydrophobic residues are required to form a stable core that can

Figure 4.3 – Buried surface area cut-offs for cluster contacts. Conceptual (a) and observed (b) results from a subset of 55 TIM barrel proteins, for different surface area cutoff criteria in determining inter-residue contact strength. Contiguous placement of ILV residues in the core of proteins is represented as hydrophobic clusters. As the criteria for the strength of interactions between residues, measured by the extent of buried surface area, is increased, the clusters are expected to fragment. This is represented by the small increase in the number of clusters when the criterion is increased from 0 to 2 Å^2. Beyond this a further increase in the stringency of the criteria does not alter the cluster composition dramatically, as is represented by the constant number of clusters. As the criteria become more stringent, the clusters fragments into numerous smaller clusters, which eventually disintegrate at very high cutoff values. The representative traces are labeled with the PDB ID of the corresponding protein.

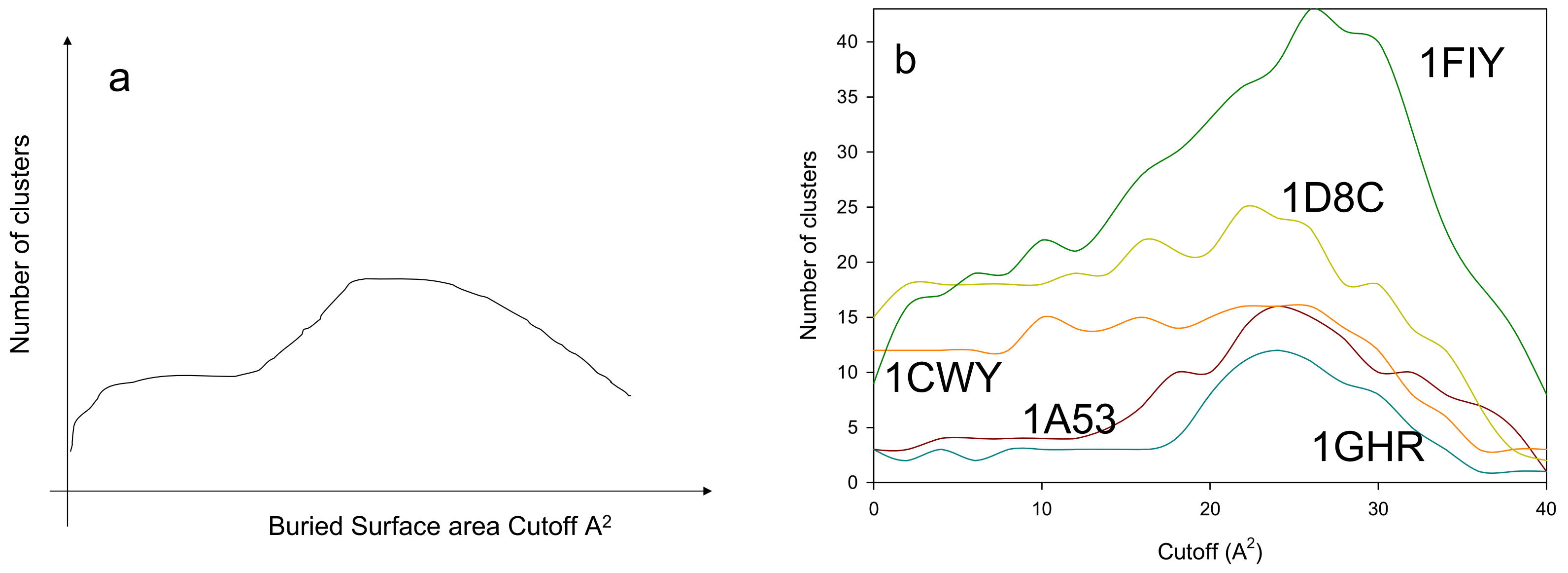

a
Number of clusters
Buried Surface area Cutoff A^2
b
Number of clusters
0
5
10
15
20
25
30
35
40
1FIY
1D8C
1CWY
1A53
1GHR
0
10
20
30
40
Cutoff (A^2)

Figure 4.3

Figure 4.4 – Correlation between number of residues in a cluster and the amount of surface area buried by the cluster. 500 Å^2 corresponds to the surface area buried by a cluster of 10 ILV residues (red arrow).

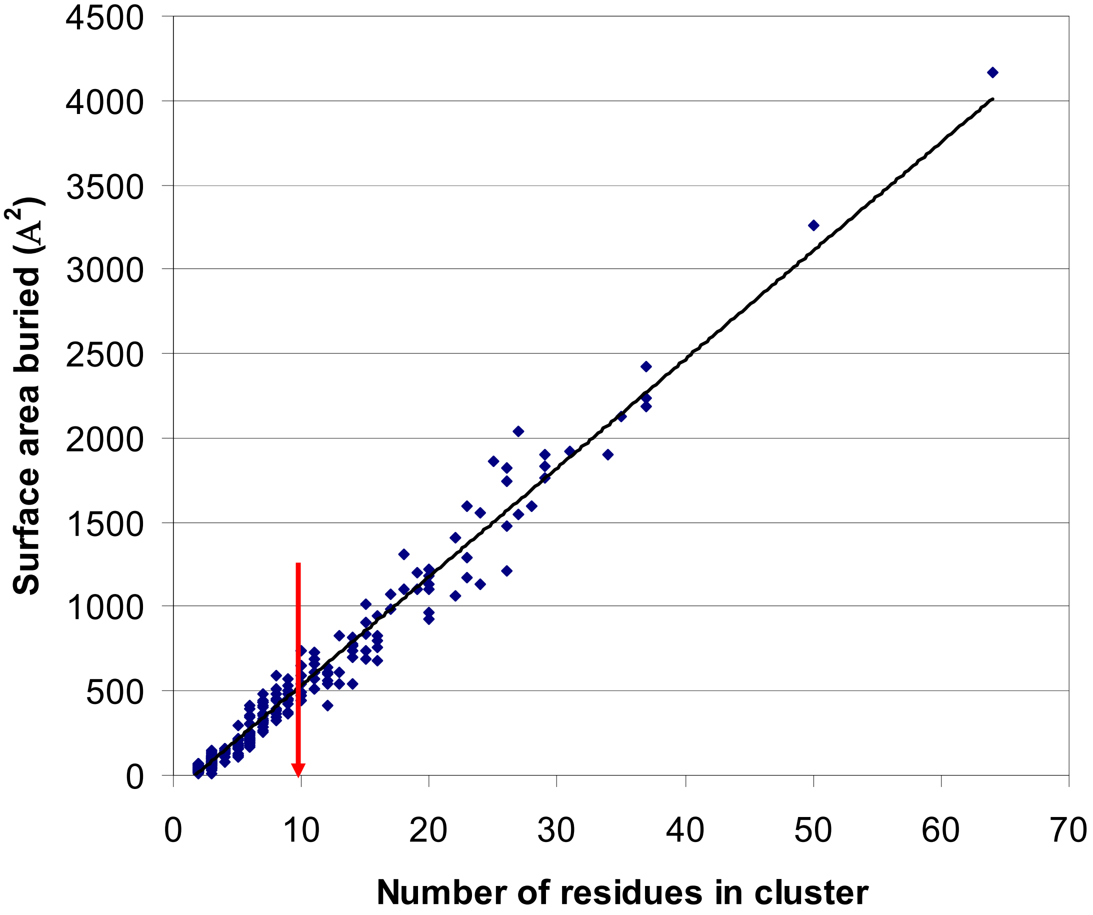

4500
4000
3500
3000
2500
2000
1500
1000
500
0
0
10
20
30
40
50
60
70
Surface area buried (A2)
Number of residues in cluster

Figure 4.4

propagate the aggregation reaction. The total amount of surface area buried by a cluster is directly proportional to the number of residues involved in it (Fig. 4.4). The SAB by a cluster of 10 residues, ~ 500 Å^2, corresponds to the burial of 6 iso-butyl groups,[199,200] that is the minimum core for hydrophobic collapse in globular proteins. These two criteria, i) inter-residue SAB > 10 Å^2, and ii) 10 interconnected residues with 500 Å^2 buried, are used henceforth to define clusters.

Test for the validity of the BASiC Hypothesis

The wealth of information available on the native state HX of several proteins from different topological groups makes it possible to test the validity of the BASiC hypothesis. The presumption is that amide hydrogens that are close to hydrophobic clusters and are not on the surface in the native structure of their resident proteins should be protected from exchange. A database of thirty three proteins with residue specific (NMR) native state HX information was developed from the literature to test this hypothesis (Table 4.1). Residues were classified into four broad categories: 1) those that exchange faster than the dead time of the experiment, i.e., unprotected, 2) those residues for which the exchange rate can be measured but does not correspond to a large scale unfolding events, i.e., weakly protected, 3) those that exchange at rates comparable with the global unfolding of the protein, i.e., strongly protected and 4) those with intermediate exchange rates (subglobal), i.e., medium protection. Five of the proteins, representing different folds and comprising both α-helices and β-strands were selected as a test set to define the subset of residue types that best correspond to the protected core of the protein.

Table 4.1 – List of proteins with residue specific native state HX data. The PDB ID of the proteins used in the test set are in bold font. The number in the last column corresponds to the protein number in figure 4.9.

	Protein Name	Reference	PDB	Protein number for ILV cluster analysis
1	Hen egg-white lysozyme	Pederson JMB 1991[201]	193L	1
2	Human α-lactalbumin	Schulman JMB 1995[202]	1A4V	2
3	Apo-cytochrome b 562	Chu Biochemistry 2002[203]	1APC	
4	Protein A - B domain	Bai Prot Sci 1997[204]	1BDD	
5	α-subunit of tryptophan synthase	Vadrevu JMB 2008[189]	**1BKS**	3
6	Barstar	Bhuyan Proteins 1998[27]	1BTA	4
7	Dynein Light Chain Dimer	Mohan J Biomol NMR 2009[205]	1F3C	5
8	Apo-leghemoglobin	Nishimura JMB 2008[206]	1FSL	6
9	Entamoeba histolytica Calcium binding protein	Mukherjee Biochemistry 2007[207]	1JFK	
10	Human acidic fibroblast growth factor	Chi JBC 2002[208]	1JQZ	7
11	Lys N	Alexandrescu JMB 1999[209]	1KRS	8
12	Apo-myoglobin	Hughson 1990 Science[210]	1MBC	9
13	Cold shock protein A	Rodriguez Biochem 2002[211]	1MJC	
14	Tendamistat	Qiwen Biochem 1987[55]	1OK0	
15	Kinase inducible transactivation domain	Schanda JMB 2008[212]	1SB0	10
16	Staphylococcal nuclease	Bedard JMB 2008[213]	**1SNP**	11

Table 4.1

Table 4.1 (cont.)

	Protein Name	Reference	PDB	Protein number for ILV cluster analysis
17	Chicken src SH3 domain	Grantcharova Biochem 1997[214]	1SRL	
18	Ubiquitin	Sidhu thesis personal communication[215]	**1UBQ**	12
19	Apo-flavodoxin	Yves JM Bollen PNAS 2006[159]	**1YOB**	13
20	Human carboxy anhydrase I	Kjellsson 2003 Biochemistry[216]	2CAB	14
21	Equine lysozyme	Ludmilla JMB 1997[217]	2EQL	
22	Protein G B1	Orban Biochem 1995[218]	2GB1	
23	Outer surface protein A	Yan JMB 2002[219]	2I5V	15
24	Single chain fragment variable antibody	Freund FEBS 1997[220]	2MCP	16
25	Turkey ovomucoid third domain	Arrington JMB 1999[221]	2OVO	
26	Protein L	Yi Prot Sci 1996[222]	2PTL	
27	RibonucleaseH	Chamberlain Nature 1996[223]	2RN2	17
28	Thioredoxin	Bhutani Prot Sci 2003[224]	2TRX	18
29	Chemotaxis protein Y	Lacroix JMB 1997[225]	**3CHY**	19
30	Ribonuclease A	Mayo Science 1993[226]	3DH5	
31	Bovine pancreatic trypsin inhibitor	Kim Biochemistry 1993[56]	5PTI	
32	Ribonuclease T1	Mullins Prot Sci 1997[227]	9RNT	

Table 4.1 (cont.)

The composition of residues closest to the protected core of the protein is expected to be different from that of the unprotected dynamic exterior of the protein. Figure 4.5, shows the composition of residues within 4 Å of the amide groups that are protected. The probability of occurrence of AFILMYV residues within vdW distances of strongly protected amides is higher than their normal distribution in protein sequences (Fig. 4.5), while the probability of occurrence of CW residues in the vicinity of weakly protected amides is higher than their respective occurrence in protein sequences. To a lesser extent, FIMW residues appear to be more numerous than random within 4 Å of amides with medium protection. When normalized for their occurrence in protein sequences, the composition of alanines is higher close to the strongly protected amides. However, this preferential distribution of alanines rapidly returns to normal as the distance from strongly protected amides is increased (Fig. 4.6), possibly due to a combination of their small size and their frequent occurrence in proteins. The rapid return to normal distribution for F and Y residues, suggests that only a few contacts with strongly protected residues contributed to the higher than normal distribution in the immediate vicinity of strongly protected amides. The rise in their numbers at further distances and the correlation of FW residues with the medium protected amides may be indicative of a peripheral role for the aromatic residues. For ILMV residues the preferential distribution is seen even out to 10 Å away from the strongly protected amides and to a smaller extent from those with medium protection.

These observations suggest that there is a specific pattern of residues packing

Figure 4.5 – Residue composition in a shell of 4 Å around the backbone amide groups of residues protected to different extents. The residue composition of the test set of five proteins is also shown. AILMFYV residues appear to be concentrated around strongly protected residues. CW residues are concentrated around residues that are weakly protected.

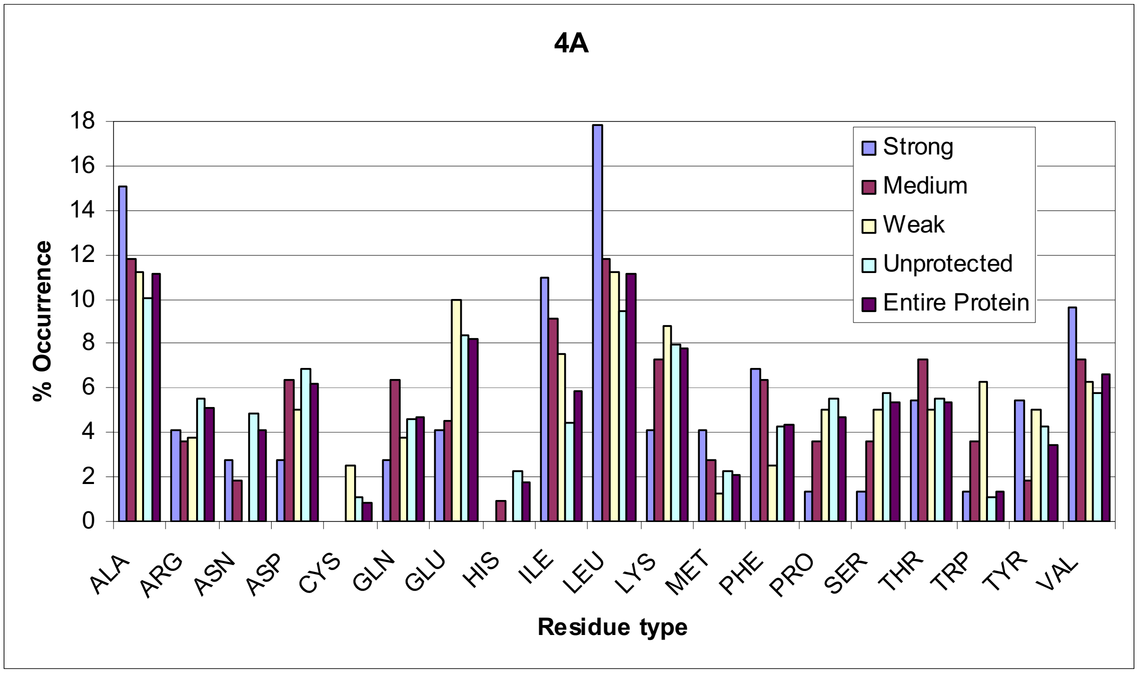

4A
18
16
14
12
10
8
6
4
2
0
% Occurrence
Strong
Medium
Weak
Unprotected
Entire Protein
ALA
ARG
ASN
ASP
CYS
GLN
GLU
HIS
ILE
LEU
LYS
MET
PHE
PRO
SER
THR
TRP
TYR
VAL
Residue type

Figure 4.5

Figure 4.6 – Residue composition in a shell around residues protected to different extents, as a function of an increasing shell size. The residue composition is normalized for occurrence in the test set of five proteins. Distribution of ILMV residues is correlated with the strongly protected amides. FY residues have a binomial distribution. Ala residues show a small signal for preferential distribution in the immediate vicinity of strongly protected residues. CW residues appear to be correlated with weakly protected residues.

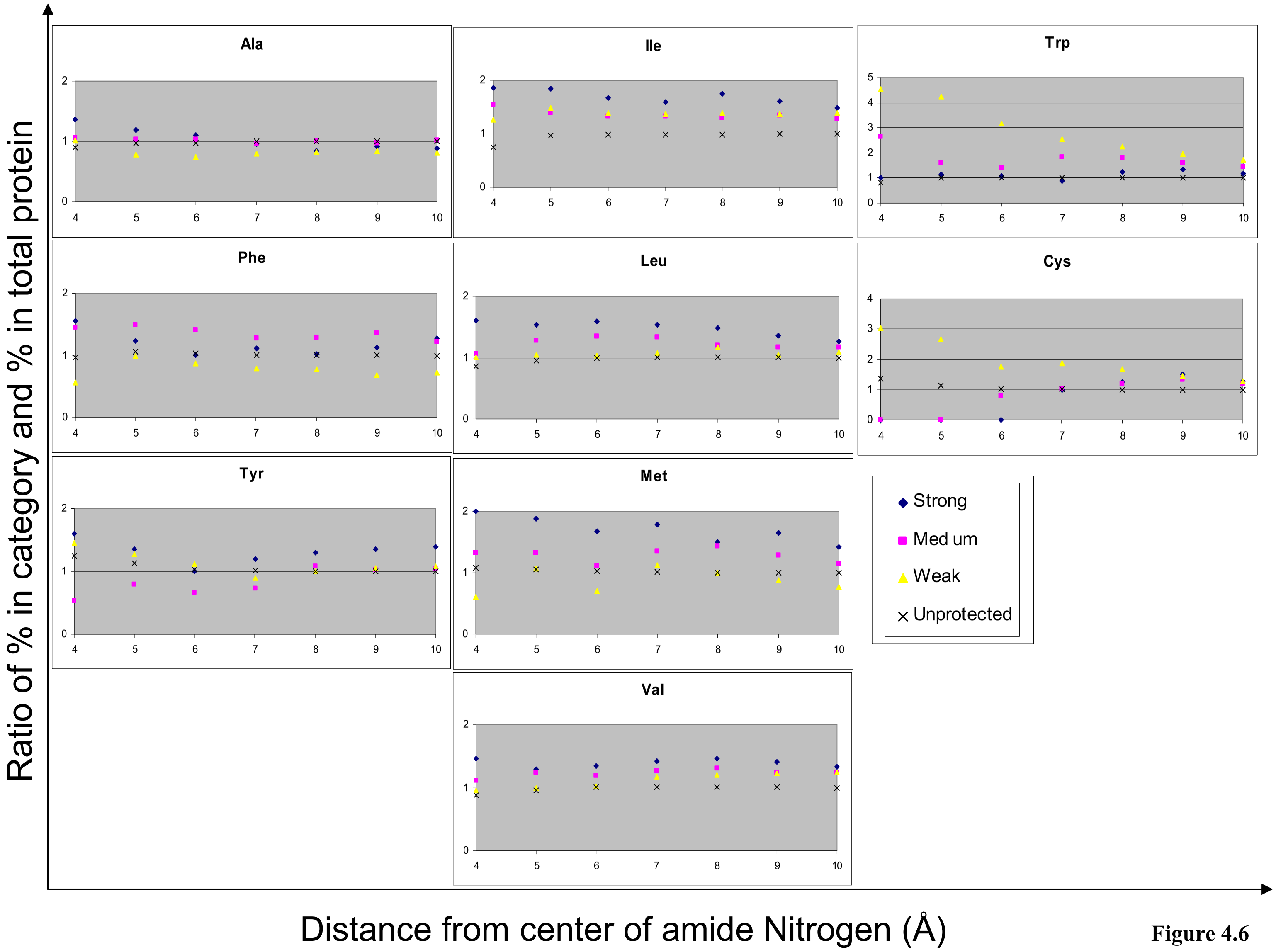

Ala
Ile
Trp
Phe
Leu
Cys
Tyr
Met
Val
0
1
2
3
4
5
4
5
6
7
8
9
10
◆ Strong
■ Med um
▲ Weak
× Unprotected
Ratio of % in category and % in total protein
Distance from center of amide Nitrogen (Å)

Figure 4.6

Figure 4.6 (cont). – Residue composition in a shell around residues protected to different extents, as a function of an increasing shell size. The distribution patterns of DEHKNPQRST residues.

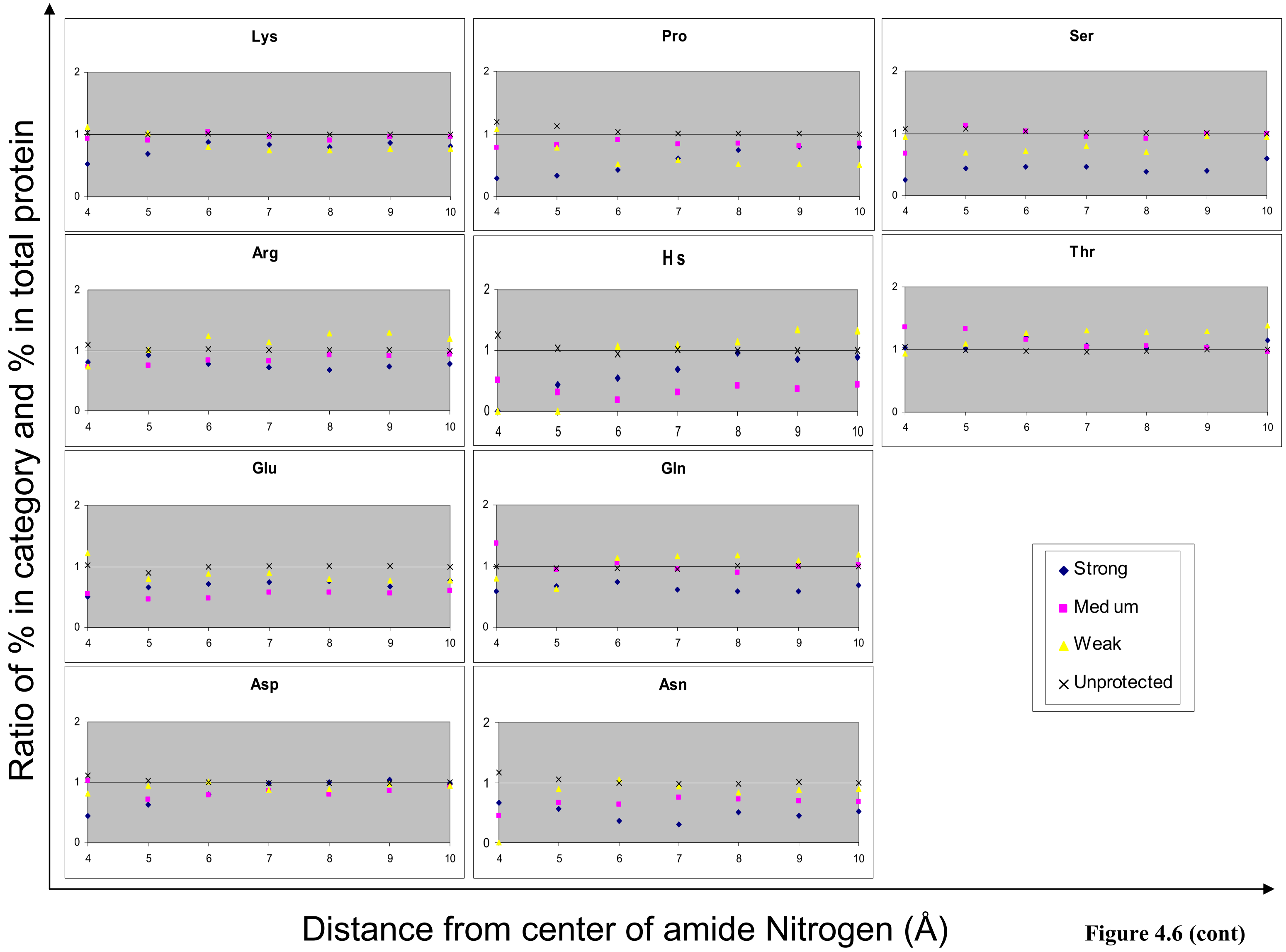


Figure 4.6 (cont)

around the protected core of the protein. The innermost regions that are in direct contact with the protected amide hydrogen atoms can be comprised of any of the hydrophobic residues, AFILMYV. The next layer of residues preferentially comprises ILMV residues. The Ala residues that are not in direct contact with the protected amides are distributed evenly throughout the protein, whereas the remaining FY and also the W residues form a peripheral layer around the ILMV core. It follows that the majority of ILMVs should be localized in the protected core region, in close contact with strongly protected amides and forming an interconnected network of contacts.

The percentage of these residues in the vicinity of protected residues is shown in Figure 4.7. Only ~40% of the ILMV residues in the protein sequence pack against strongly protected amides (within 6 Å), while ~50% pack against those amides that have medium protection from solvent exchange. While these numbers are higher than those observed for other residue types (data not shown), they imply that the remainder of ILMV residues are randomly distributed. This observation is further supported by the random distribution of ILMV residues around unprotected residues (Fig. 4.6). One explanation for this distribution pattern is that only 40 50% ILMV residues take the shape of large hydrophobic clusters in the protected core region, while the remainder form isolated small hydrophobic patches that may not offer sufficient protection to the neighboring amides from exchange with the solvent.

Although the methionine residues have a high probability of occurring in the vicinity of protected residues, their small numbers are unlikely to be the defining feature of the clusters. And, among the aromatics, only phenylalanines are sufficiently numerous

Figure 4.7 – BASiC residues and protected amides. Percentage of any given residue type in the vicinity of residues protected to different extents (strong, medium, weak and unprotected) as a function of distance from the center of the amide nitrogen.

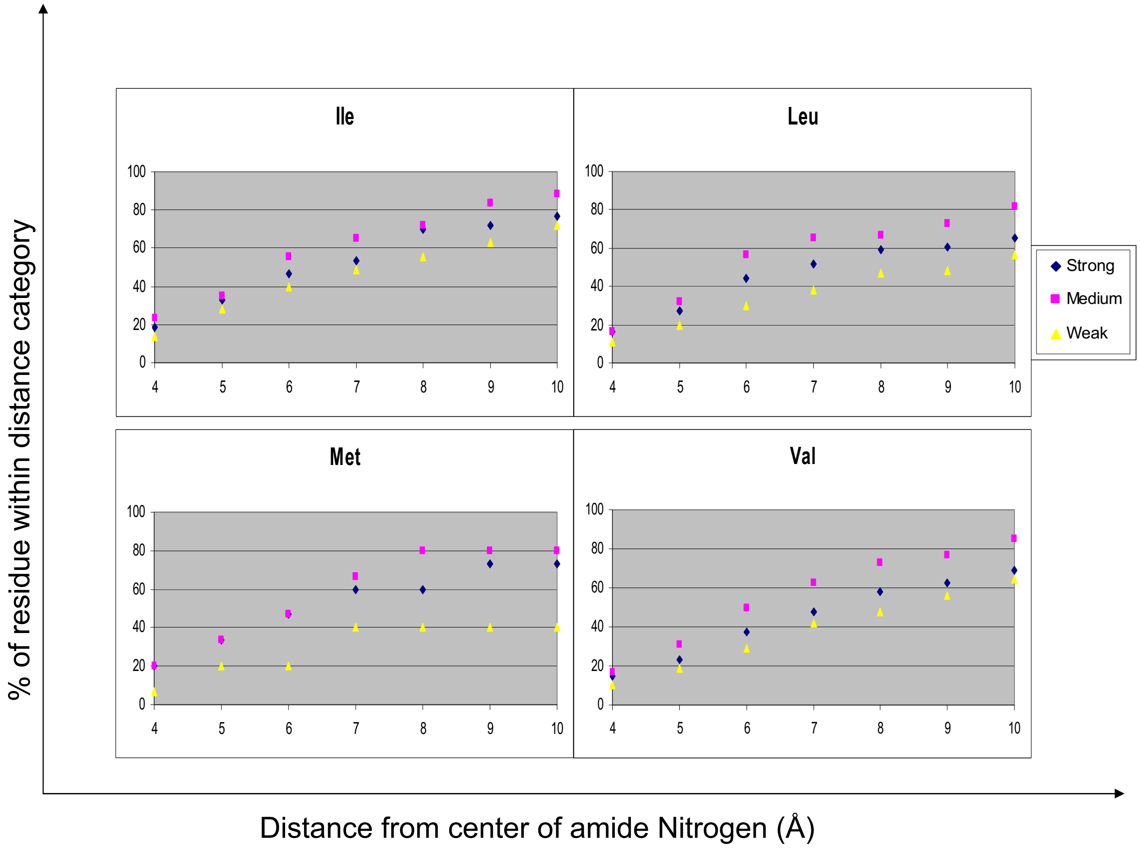

Ile
Leu
Met
Val
100
80
60
40
20
0
4
5
6
7
8
9
10
◆ Strong
■ Medium
▲ Weak
% of residue within distance category
Distance from center of amide Nitrogen (Å)

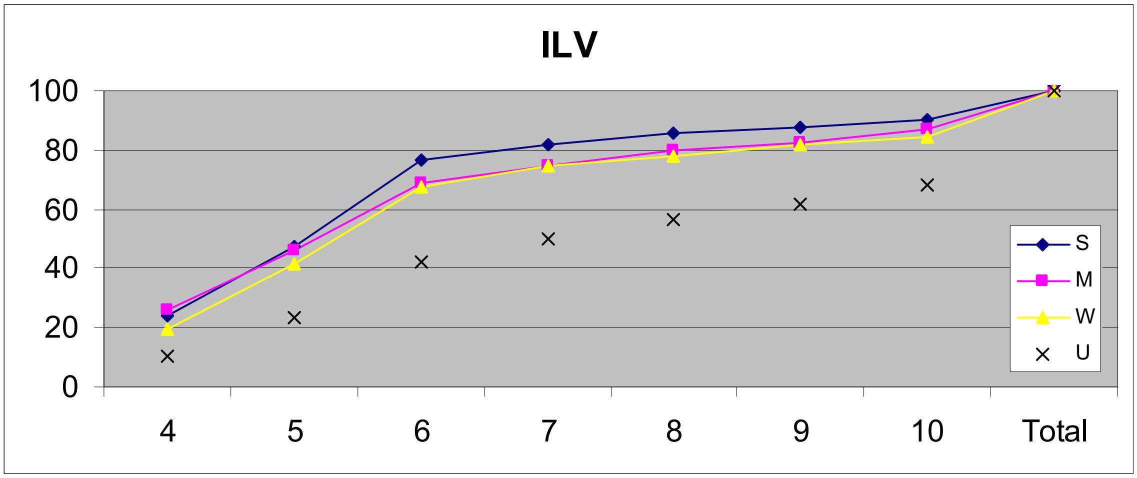
Figure 4.7

to alter the cluster definition. However, the occurrence of the aromatics in the periphery of the clusters is apparent from their correlation with the medium protected residues instead of the strongly protected ones. Only the ILV residues were used as components of the hydrophobic clusters in subsequent analyses. Altering the composition of the clusters to include methionines did not alter the results (data not shown).

From the dataset of 32 proteins with residue specific HX information, only 19 contain clusters of ILV residues that meet the requirements defined above. Further analyses were carried out on this set of proteins. The percentage of residues within an increasing distance cutoff from the ILV clusters is shown in Figure 4.8. As expected, the majority, ~70%, of the protected residues (strong, medium and weak) are within 6 Å of ILV clusters. However, ~50%, of the residues within 6 Å of ILV clusters belong to the unprotected group (overprediction). Only a slight improvement in the predictive power of this method can be made by excluding surface exposed amides (surface area exposed to solvent predicted by CSU < 2 Å^2) (data not shown). Two measures of a successful prediction are, 1) the positive identification of experimental data, and 2) the percentage of false positives. These two measures, termed sensitivity and specificity,[228] respectively, are represented for the 19 proteins in Figure 4.9. On average, this method of predicting cores of protein stability based on the BASiC hypothesis has a high sensitivity score (70 %), but a lower than desired level of specificity (~ 50%). The examples discussed later in this chapter demonstrate the reasons for the limitation of this method.

Several other methods including COREX,[229] FIRST,[230] GNM,[228] and other empirical potential functions[231] have attempted to predict the protection patterns in

Figure 4.8 – Hydrophobic clusters and protection patterns. Number of amide nitrogens in the vicinity of ILV side chains that are involved in hydrophobic clusters, as a function of distance from the center of the closest side chain atom.



Distance from the center of amide nitrogen

Figure 4.8

Figure 4.9 – Prediction of protection patterns. a) Sensitivity and **b)** specificity, two measures of the accuracy of prediction of protection patterns. Sensitivity represents the number of true positives as a percentage of residues experimentally determined to be protected (overlap between prediction and experiment) / (experiment). Specificity represents the number of true positives as a percentage of the group predicted to be protected (overlap between prediction and experiment) / (predicted). The values for each protein (for protein names refer to table 4.1) are represented as diamonds (sensitivity) and squares (specificity). The broken line represents the value obtained for random distribution.

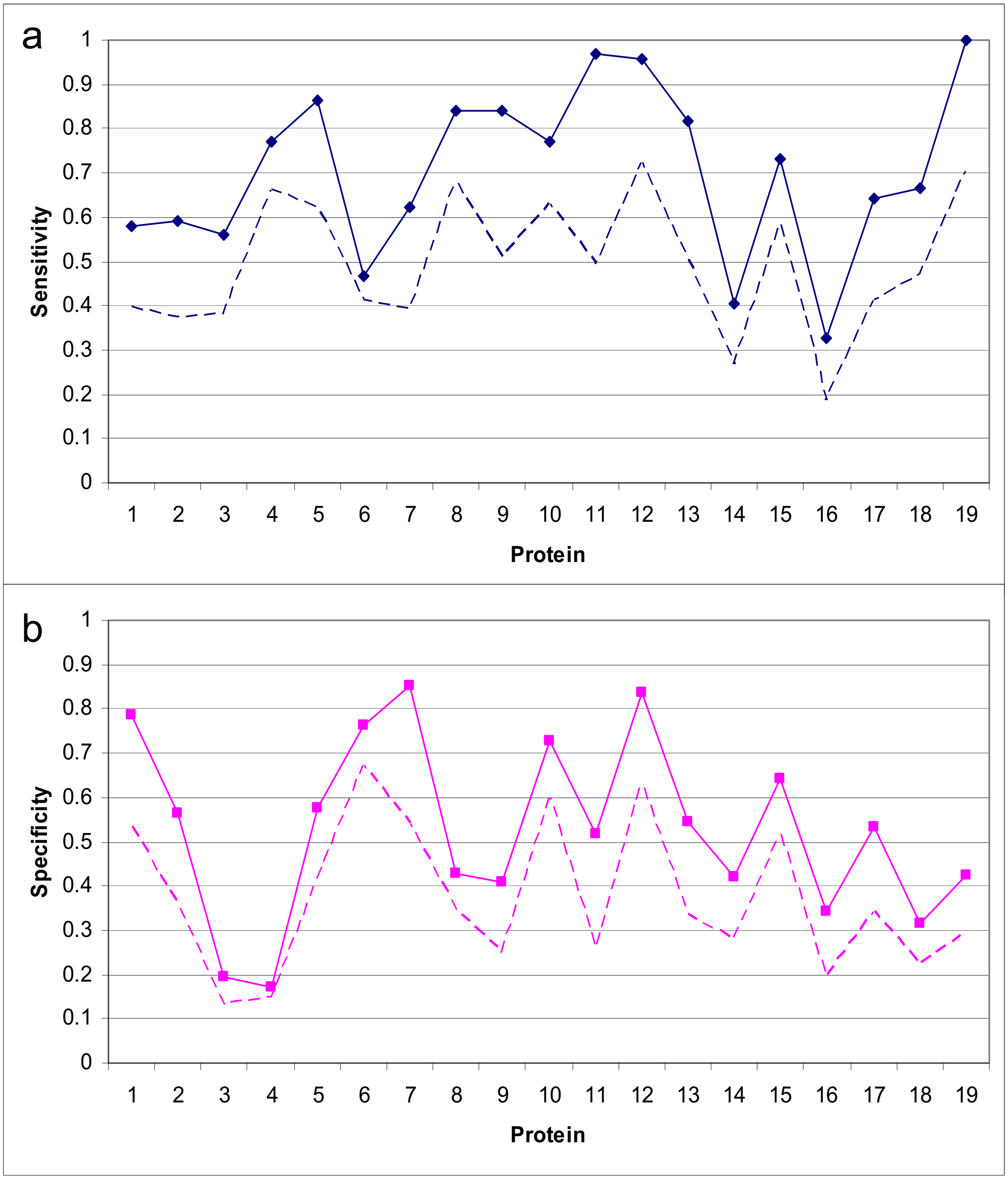

a
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Sensitivity
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
Protein
b
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Specificity
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
Protein

Figure 4.9

proteins with comparable success rates. The COREX method analyzes the effects of local fluctuations on the global structure of the protein to determine regional stabilities. The FIRST (Floppy Inclusions and Rigid Substructure Topography) method, defines constraints on the molecule, based on the interaction network of the native state. The strength of each H-bond is tested by determining the rigidity or flexibility of an all atom protein model in the presence or absence of that bond. The GNM method replaces all interactions in the protein with elastic springs. The collective motions of the elastic network determine the overall flexibility of the protein. All of these approaches are modeled on inter-residue interactions in the native state and are sensitive to packing densities in the protein and are hence possibly sensitive to the location of ILV clusters. However, they do not attempt to explicitly distinguish between different types of interactions. The BASiC hypothesis provides the means to incorporate sequence information based on physical and chemical properties of hydrophobic side chains that form the underlying determinants of stability.

ILV Clusters and Cores of Stability in Globular Proteins

In the following examples, a rationale is presented for how native-like clusters of ILV residues stabilize thermodynamic states that are in equilibrium with the native state that can be accessed by hydrogen exchange. It is, however, important to note that clusters observed in the native state can be rearranged or alternately packed in higher energy states, and thus the predictive power of this method is best suited for the exchange that is occurring through the native state or native-like intermediates and to a lesser extent for global unfolding events.

Staphylococcal nuclease

The rare high energy states in staphylococcal nuclease (SNase), an α + β protein, have recently been examined by native-state HX-NMR.[213] Under conditions that greatly favor the native conformation and preserve partially-folded states of marginal stability, i.e., in the absence of or in low concentrations of chemical denaturants, the patterns of protection of main chain amide hydrogens against exchange with solvent hydrogens can provide insights into the secondary structures of these very rare conformers. Application of this method to SNase demonstrated significant protection in the core of the β-sheet region and an α-helix that docks on the β-sheet. Measurable protection was also observed in segments of two other helices that dock on the β-sheet. Examination of the x-ray structure shows that a cluster of 13 ILV side chains contribute to the hydrophobic core (Figure 4.10). Eight of nine ILV's have amide hydrogens with protection factors whose ΔG^{o}_{HX} values exceed 8 kcal mol ^{1}; the protection factors of the remaining four amide hydrogens were not reported. Although 20 other side chains also display strong protection against exchange, the ILV cluster links all five β-strands and three α-helices and, as such, is a central element in the core structure.

Ribonuclease H

Ribonuclease H (RNase H) provides another example where an ILV cluster serves as a subset of side chains associated with the HX-NMR-determined stability core. Marqusee and her colleagues[223] have found that main chain amide hydrogens in the A

Figure 4.10 – Cartoon representation of the crystal structure of staphylococcal nuclease (1SNP). The amide nitrogens of protected residues are shown as colored spheres, with blue corresponding to those that are most protected (global unfolding).[213] Those in green form the next level of protection, followed by yellow and finally red. The grey spheres represent the hydrophobic cluster made up of ILV residues. The proximity of the highly protected residues with the ILV core is apparent.

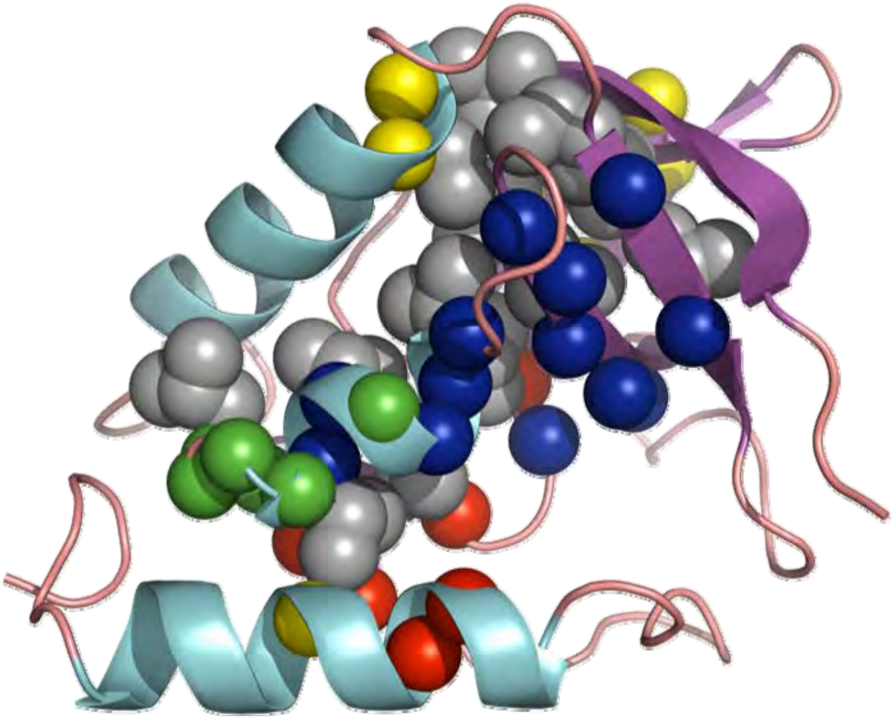


Figure 4.10

Figure 4.11 – Cartoon representation of the crystal structure of ribonuclease H (2RN2). The amide nitrogens of protected residues are shown as colored spheres, with blue corresponding to those that are most protected (global unfolding).[223] Those in green form the next level of protection, followed by red and finally yellow. The grey spheres represent the hydrophobic cluster made up of ILV residues. The proximity of the highly protected residues with the ILV core is apparent.

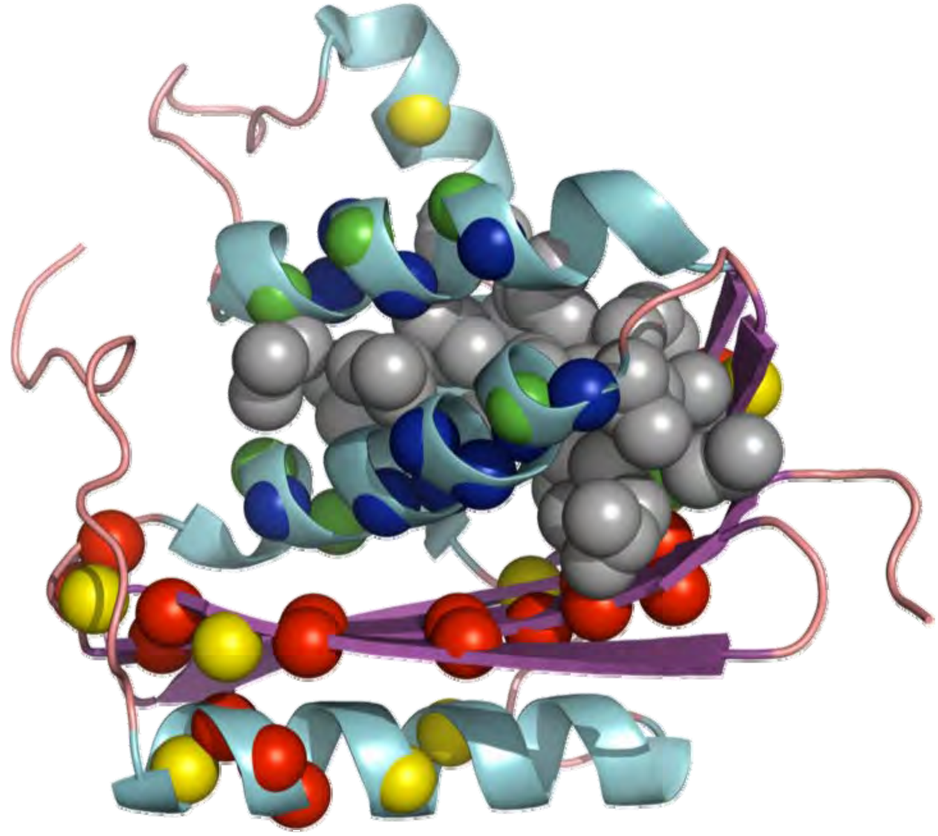


Figure 4.11

and D helices are most strongly protected against exchange. These helices dock on each other, are rich in ILV residues, and are central to the structure of this protein. ILV cluster analysis shows that 16 ILVs form a large buried cluster in RNase H and that these side chains are derived from all 4 α-helices and 4 of the 5 β-strands (Fig. 4.11). Comparison with the HX protection map reveals that the amide hydrogens in all 16 ILVs are protected to some degree and 8 exchange through a global unfolding mechanism. The very strong protection for the A and D helices may reflect the large number of intra-cluster ILV contacts for their resident side chains and the effective exclusion of solvent.

ILV Clusters and Folding Reactions

Mutational analysis has been extensively employed to identify the residues that participate in high energy transition state ensembles (TSE), saddle points on folding free energy surfaces that limit access to transient intermediates and the native conformation during folding.[53,96,232] The side chains whose replacement perturbs the refolding rate constants highlight the minimal structural requirements for the productive flow of protein to the native conformation. Very often the residues that contribute to this minimal structure are a subset of the highly protected amides in native state exchange measurements.[183] The relation between ILV clusters and the structures of high energy intermediates is consistent with a previous statistical analysis of folding nuclei in globular proteins. Mirny and Shakhnovich[165] found that the side chains involved in folding nuclei are often significantly more conserved during evolution than those elsewhere in the structure. Further, the segregation of ILV side chains from all others provided the most significant measure of conservation. In the context of the BASiC Hypothesis, this segregation reflects the participation of these branched beta side chains in clusters that are especially resistant to the penetration of water to the amide linkage. A comparison of the phi-values of several proteins with the location of their ILV clusters, defined by the BASiC hypothesis, would provide valuable insights into the role of hydrophobic clusters in driving protein folding reactions.

Conclusion

The BASiC Hypothesis is founded on the unfavorable hydration of the branched aliphatic amino acid side chains, isoleucine, leucine and valine. Although the relevance of their vapor phase-to-water transfer free energies to protein folding and stability is not obvious for globular proteins in aqueous environments, the limited survey presented in this communication supports this view. Hydrogen exchange data clearly show that other side chains also play crucial roles in cores of stability for thermodynamic states and for TSEs. However, these stability cores invariably contain clusters of ILV side chains that would provide a platform to recruit aliphatic segments of other side chains or sulfur-containing side chains or aromatic side chains whose polarizable character would mediate interactions with water in partially-folded states. Although the BASiC Hypothesis underestimates the structures of TSEs and intermediates, recognition of the essential roles of ILV clusters in folding and stability provides a starting point for more comprehensive analyses.

The simplicity of the BASiC Hypothesis enables tests of its validity for explaining a variety of biological phenomena. For examples, an argument can be made that ILV clusters are crucial in defining a variety of structural motifs including leucine zipper coiled coils, ankyrin repeat proteins, leucine-rich repeat proteins and left-handed β-helix proteins. Protein-protein interactions may, in some cases, be mediated by ILV clusters. Examples include calmodulin/peptide complexes, SNARE complexes responsible for protein trafficking, dimerization domains of Hsp 90[233] and IRF5[234] and chaperones/client interactions observed for GroEL and DnaK. ILV clusters may also be involved in the formation of amyloids thought to be responsible for human pathologies[235,236] and infectious prions.[237] All of these potential applications would serve to reemphasize the role of the amino acid sequence in folding and stability, as originally envisioned by Anfinsen and his colleagues.

Acknowledgement

The authors would like to thank Mr. Noah Cohen for his help with data mining, Dr. Jill A. Zitzewitz for valuable discussions and editing the manuscript, Dr. Osman Bilsel and Dr. Can B. Kayatekin for helpful discussions. Dr. Ramakrishna Vadrevu, Dr. Ying Wu and Dr. Zhenyu Gu, for their insights in the TIM barrel data analysis. A Special thanks to Dr. David Lapointe for developing and hosting the web based version of the cluster algorithm. This study was funded by the NSF (Grant MCB-0721312 to CRM) and by the NIH (Grant GM23303 to CRM).

Chapter V – βα-Hairpin Clamps Brace βαβ Modules and can Make Substantive Contributions to the Stability of TIM Barrel Proteins

This chapter has been published previously as *Yang X, Kathuria SV, Vadrevu R, Matthews CR. "Beta alpha-hairpin clamps brace beta alpha beta modules and can make substantive contributions to the stability of TIM barrel proteins." PLoS One. 2009 Sep 29;4(9):e7179.*

The work presented in the following chapter was a collaborative effort. Dr. Xiaoyan Yang, Dr. Ramakrishna Vadrevu and Dr. C Robert Matthews conceived and designed the experiments. Dr. Xiaoyan Yang performed the experiments. Dr. Xiaoyan Yang, analyzed the experimental data. I developed and analyzed the database. Dr. Xiaoyan Yang, Dr. Ramakrishna Vadrevu, Dr. C Robert Matthews and I wrote the manuscript.

Abstract

Non-local hydrogen bonding interactions between main chain amide hydrogen atoms and polar side chain acceptors that bracket consecutive βα or αβ elements of secondary structure in αTS from *E. coli*, a TIM barrel protein, have previously been found to contribute 4-6 kcal mol ^{1} to the stability of the native conformation. Experimental analysis of similar βα-hairpin clamps in a homologous pair of TIM barrel proteins of low sequence identity, IGPS from *S. solfataricus* and *E. coli*, reveals that this dramatic enhancement of stability is not unique to αTS. A survey of 71 TIM barrel proteins demonstrates a 4-fold symmetry for the placement of βα-hairpin clamps, bracing the fundamental βαβ building block and defining its register in the $(\beta\alpha)_8$ motif. The preferred sequences and locations of βα-hairpin clamps will enhance structure prediction algorithms and provide a strategy for engineering stability in TIM barrel proteins.

Introduction

The $(\beta\alpha)_8$, TIM barrel is one of the most common folds in biology, supporting a myriad of catalytic functions essential to life[166]. Experimental[238] and bioinformatics[166,239,240] analyses of TIM barrel proteins have led to the conclusion that a pair of adjacent parallel β-strands and the intervening anti-parallel α-helix, i.e., the βαβ module, serve as the minimal unit of stability. Gene duplication of this elemental βαβ building block into higher-order structures has been suggested to result in several common βα-repeat structures, including the TIM barrel, Rossman, flavodoxin and leucine-rich folds[239]. The interactions stabilizing this super-secondary structure include: (1) main chain-main chain (MC-MC) hydrogen bonds (H-bonds) between the β-strands, (2) intra-helical MC-MC H-bonds, (3) hydrophobic interactions between the side chains (SC) protruding from the β-strands and the α-helix, (4) side chain-side chain (SC-SC) H-bonds and salt bridges, (5) dipole-dipole interactions between the α-helix and the pair of β-strands on which it docks[241] and (6) main chain-side chain (MC-SC) H-bonds[143,242,243]. The surprising role of a subset of non-local MC-SC H-bond interactions in structure and stability is the subject of this communication.

A majority of MC-SC interactions in proteins are local in sequence, usually within six residues[242,243], and are often involved in capping either the N- or the C-termini of α-helices[244,245]. Mutational analysis has shown that these non-covalent interactions usually contribute modestly, typically in the range of 1-2 kcal mol ^{1}, to the stability of their resident proteins[246 249]. In contrast, the removal of three non-local MC-SC H-bond interactions each reduce the stability of the alpha subunit of tryptophan synthase (αTS), a TIM barrel protein, by 4-6 kcal mol ^{1}, and disrupt the complete formation of the TIM barrel motif[143]. These three interactions in αTS, between MC amide H-bond donors and SC H-bond acceptors, connect the N-terminus of one element of secondary structure, either β-strand or α-helix, to the C-terminus of the subsequent element of structure, either α-helix or β-strand, respectively. These non-local MC-SC interactions were designated as βα-hairpin clamps and αβ-hairpin clamps, respectively[143]. The significant contribution to structure and stability by three such clamps in αTS[143] raises the possibility that potent βα- and αβ-hairpin clamps may be an important general feature of TIM barrel proteins.

A two-pronged approach was taken to probe the significance of βα-hairpin clamps in TIM barrel proteins. First, mutational analysis of two representative TIM barrel proteins, indole-3-glycerol phosphate synthase (IGPS) from *S. solfataricus* (sIGPS) and *E. coli* (eIGPS), shows that a subset of their βα-hairpin clamps make significant contributions to protein stability. Second, a survey of 71 TIM barrel proteins[250] explored the frequency, location and sequence preferences of all βα-hairpin clamps. The observed preferences for location and sequence for the βα-hairpin clamps and their contribution to the structure and stability of TIM barrel proteins suggest that the recognition of these interactions can enhance protein structure prediction algorithms and provide a strategy for engineering stability in TIM barrel proteins.

Results

Experimental analysis of βα-hairpin clamp interactions in two TIM barrel proteins

The generality of the potent hairpin clamps found in αTS was tested by mutational analysis of βα-hairpin clamps in two homologous TIM barrel proteins with low sequence identity (<30%) to αTS and to each other. sIGPS (Fig. 5.1a) and eIGPS (Fig. 5.1b), each contain three βα-hairpin clamps (Figs. 5.1c and 5.1d), some of which are conserved in location with those in αTS and others between sIGPS and eIGPS. Figure 5.1 displays the distances between the donor and acceptor atoms of the βα-hairpin

Figure 5.1 – Ribbon diagrams of IGPS - sIGPS (**a**) and eIGPS (**b**) highlighting the βα-hairpin clamps. Panels (**c**) and (**d**) display the intervening elements of secondary structures between the residues forming the clamps for sIGPS: sIGPS-β2α2-S104$_{\mathrm{NH}}$→$_{\mathrm{O}\varepsilon 1}$E74; sIGPS-β3α3-I107$_{\mathrm{NH}}$→$_{\mathrm{O}\delta 1}$D128; and sIGPS-β7α7-K207$_{\mathrm{NH}}$→$_{\mathrm{O}\delta 2}$N228 and for eIGPS: eIGPS-β1α1-F50$_{\mathrm{NH}}$→$_{\mathrm{O}\gamma}$S82; eIGPS-β3α3-I111$_{\mathrm{NH}}$→$_{\mathrm{O}\delta 2}$D132; and eIGPS-β7α7-V211$_{\mathrm{NH}}$→$_{\mathrm{O}\delta 1}$N231. The SCs involved in the clamp interactions are highlighted with the H-bond donor and acceptor atoms shown in blue and red, respectively. The distances between the donor and acceptor atoms are indicated. The solvent exposed surface areas of the H-bond donor and acceptor atoms is shown in parenthesis. The H-bonds and their corresponding distances were determined by using the program HBPLUS[251]. The structures were generated using PyMOL *v* 0.99[252], and the PDB codes are 2C3Z for sIGPS[253] and 1PII for eIGPS[254].

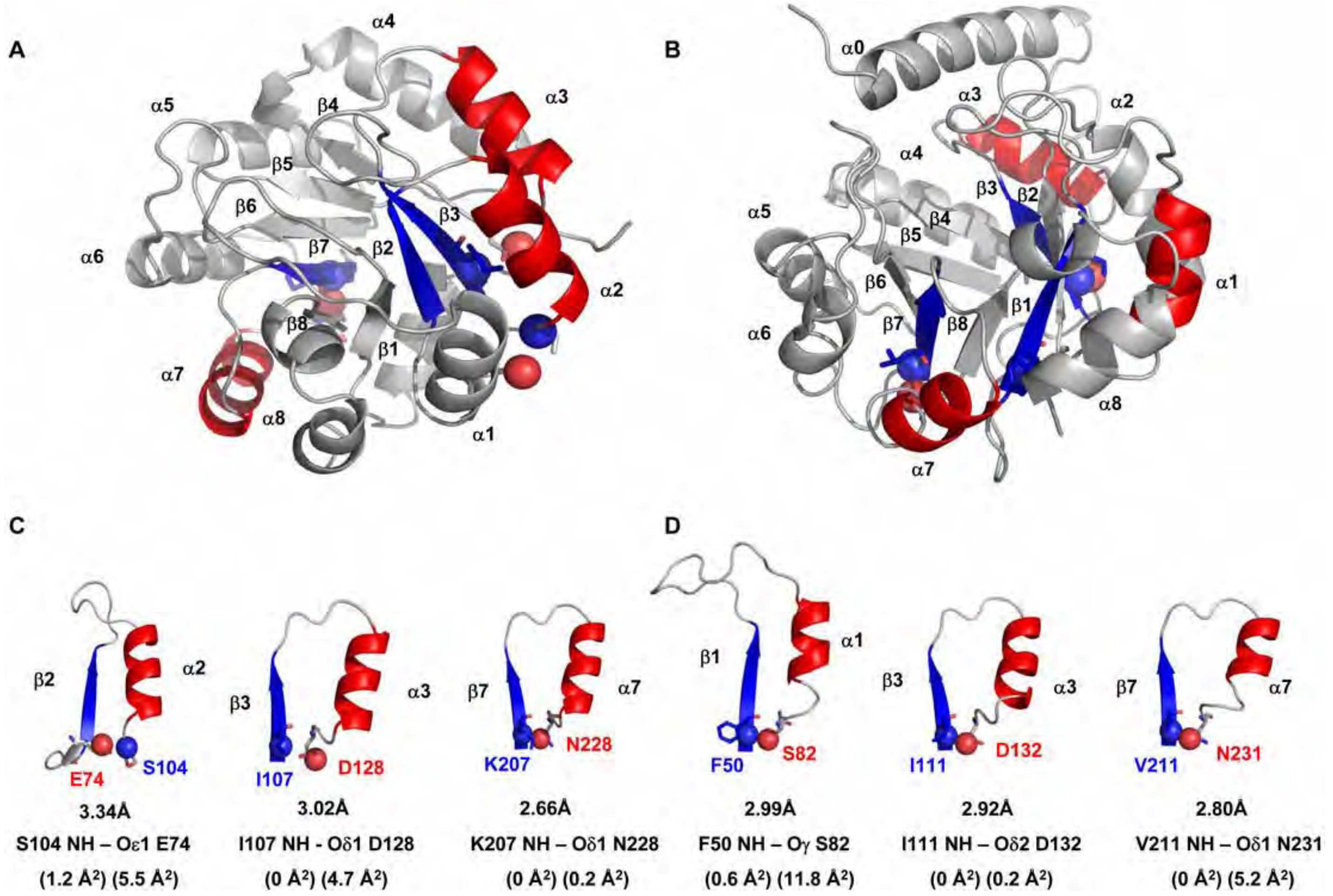

A
α4
α3
α5
β4
β5
β6
β3
α6
β7
β2
α2
β8
β1
α7
α8
α1
B
α0
α3
α2
α4
β3
β2
α5
β4
β5
α1
β6
β1
α6
β7
β8
α8
α7
C
β2
α2
E74
S104
3.34Å
S104 NH – Oε1 E74
(1.2 Å²) (5.5 Å²)
β3
α3
I107
D128
3.02Å
I107 NH - Oδ1 D128
(0 Å²) (4.7 Å²)
β7
α7
K207
N228
2.66Å
K207 NH – Oδ1 N228
(0 Å²) (0.2 Å²)
D
β1
α1
F50
S82
2.99Å
F50 NH – Oγ S82
(0.6 Å²) (11.8 Å²)
β3
α3
I111
D132
2.92Å
I111 NH – Oδ2 D132
(0 Å²) (0.2 Å²)
β7
α7
V211
N231
2.80Å
V211 NH – Oδ1 N231
(0 Å²) (5.2 Å²)

Figure 5.1

clamps interactions observed in sIGPS (Fig. 5.1c) and eIGPS (Fig. 5.1d). The solvent-exposed surface area of the H-bond acceptor atoms ranges from 0.2 Å^2 (0%) to 11.8 Å^2 (~25%), while the MC H-bond donor amide is typically completely excluded from solvent (Figs. 5.1c and 5.1d). The β1α1 clamp is observed in αTS and eIGPS (αTS-β1α1-F19$_{NH}$→$_{O\delta2}$D46, eIGPS-β1α1-F50$_{NH}$→$_{O\gamma}$S82), the β2α2 clamp only appears in sIGPS (sIGPS-β2α2-S104$_{NH}$→$_{O\varepsilon1}$E74), the β3α3 clamp is observed in all three proteins (αTS-β3α3-I97$_{NH}$→$_{O\delta2}$D124, sIGPS-β3α3-I107$_{NH}$→$_{O\delta1}$D128 and eIGPS-β3α3-I111$_{NH}$→$_{O\delta2}$D132), and the β7α7 clamp is observed in sIGPS and eIGPS (sIGPS-β7α7-K207$_{NH}$→$_{O\delta2}$N228 and eIGPS-β7α7-V211$_{NH}$→$_{O\delta1}$N231).

Perturbation of the secondary and tertiary structure by clamp deletion in sIGPS and eIGPS

The contribution of each βα-clamp interaction to the structure of the TIM barrel proteins, sIGPS and eIGPS, was assessed by replacing the H-bond acceptor SC with alanine and monitoring the effects on the secondary and tertiary structure by far-UV and near-UV circular dichroism (CD) spectroscopy. The far-UV CD spectra for the wild-type (WT) and clamp-deletion variants of sIGPS (sIGPS-WT, sIGPS-Δβ2α2-E74A, sIGPS-Δβ3α3-D128A and sIGPS-Δβ7α7-N228A) and eIGPS (eIGPS-WT, eIGPS-Δβ1α1-S82A, eIGPS-Δβ3α3-D132A and eIGPS-Δβ7α7-N231A) are shown in Figures 5.2a and 5.2b, and the near-UV CD spectra are shown in Figures 5.2c and 5.2d. Relatively small changes in the far-UV and near-UV CD spectra are observed for sIGPS-Δβ3α3-D128A, eIGPS-Δβ1α1-S82A and eIGPS-Δβ3α3-D132A compared to

Figure 5.2 – Ellipticity of wild-type and clamp-deletion variants of sIGPS and eIGPS. Far-UV (**a, b**) and near-UV (**c, d**) CD spectra of sIGPS (**a**) and (**c**): sIGPS-WT (——), sIGPS-Δβ2α2-E74A (•), sIGPS-Δβ3α3-D128A (•••), and sIGPS-Δβ7α7-N228A ( ); and eIGPS (**b**) and (**d**): eIGPS-WT (——), eIGPS-Δβ1α1-S82A (•), eIGPS-Δβ3α3-D132A (•••), and eIGPS-Δβ7α7-N231A ( ). Buffer conditions: 10 mM potassium phosphate, 0.2 mM K_2EDTA, 1 mM βME, pH 7.8 for sIPGS and pH 7.0 for eIGPS at 25 °C.

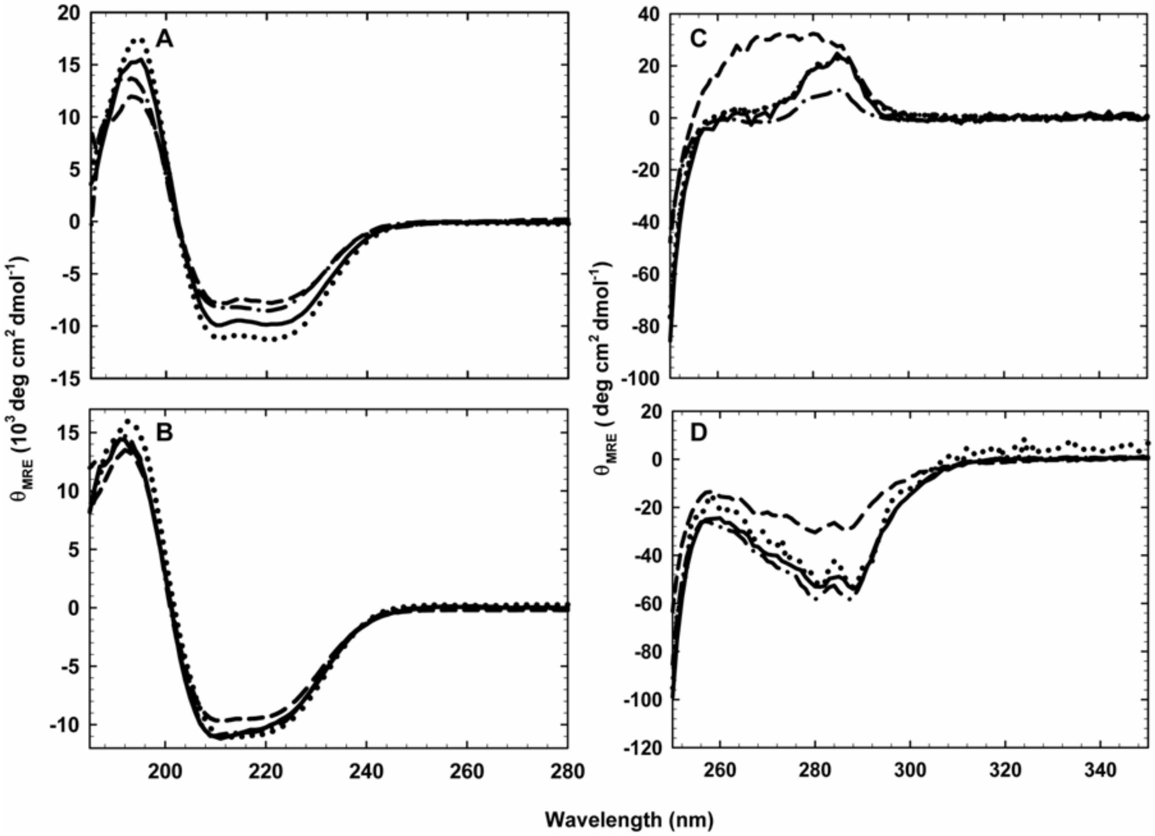

A
B
C
D
20
15
10
5
0
-5
-10
-15
40
-20
-40
-60
-80
-100
-120
200
220
240
260
280
300
320
340
θ_{MRE} (10^3 deg cm^2 dmol^{-1})
θ_{MRE} (deg cm^2 dmol^{-1})
Wavelength (nm)

Figure 5.2

their respective WT sequences. However, the significant changes in the near-UV CD spectra for the sIGPS-Δβ2α2-E74A, sIGPS-Δβ7α7-N228A and eIGPS-Δβ7α7-N231A variants imply that the deletion of the β7α7 clamps in both proteins and the β2α2 clamp in sIGPS result in altered aromatic side chain packing.

Perturbation of stability by clamp deletion in sIGPS and eIGPS

The effect of βα-hairpin clamp deletion on the stability of sIGPS and eIGPS was determined by urea denaturation. As for αTS[255], both sIGPS[75] and eIGPS[256] unfold via a highly populated intermediate, and their unfolding titration curves are well described by a three-state model, N ⇆ I ⇆ U. With the exception of eIGPS-Δβ7α7-N231A, the urea-induced unfolding transition for each of the remaining five clamp-deletion variants is also well-described by this three-state model (Fig. 5.3a and 5.3b). Because a distinct transition between the native state (N) and the intermediate state (I) is not observed during the urea induced denaturation of eIGPS-Δβ7α7-N231A (Fig. 5.3b), kinetic unfolding experiments were performed to verify the existence of I and measure the free energy difference between N and I[143].

The presence of I in eIGPS-Δβ7α7-N231A is verified by the observation of a slow kinetic unfolding phase, whose relaxation times decrease with increasing final denaturant concentration[96], when eIGPS is subjected to an unfolding jump from 0 to 3 M urea where I is expected to be highly populated. Because the amplitude of the unfolding phase is proportional to the population of N from which the reaction initiates, the decrease in the amplitude as a function of increasing initial urea concentrations (Fig. 5.4) can be fit to a two-state model, N ⇆ I, to extract the stability, ΔG°_{NI}, and the urea

Figure 5.3 – Stability perturbation of sIGPS and eIGPS by clamp deletion. Panels (**a**) and (**b**) display urea denaturation equilibrium unfolding curves of WT and clamp-deletion variants of IGPS, the lines represent fits of the data for each variant to a 3-state equilibrium folding model as described in the text. (**a**) sIGPS: sIGPS-WT (●,——), sIGPS-Δβ2α2-E74A (▲, •), sIGPS-Δβ3α3-D128A (◆,•••), and sIGPS-Δβ7α7-N228A (■,). (**b**) eIGPS: eIGPS-WT (●,——), eIGPS-Δβ1α1-S82A (▲, •), eIGPS-Δβ3α3-D132A (◆,•••), and eIGPS-Δβ7α7-N231A (■,). Panels (**c**) and (**d**) are bar graphs representing the free energy differences for the N to I step in unfolding, ΔG^{o}_{NI}, (black bars) and the I to U step, ΔG^{o}_{IU} (gray bars) for WT and the clamp-deletion variants of sIGPS (**c**) and eIGPS (**d**). The urea denaturation equilibrium unfolding curve of sIGPS-WT (**a**) and the corresponding folding free energy changes (**c**) are adapted from Forsyth *et al.*[75]

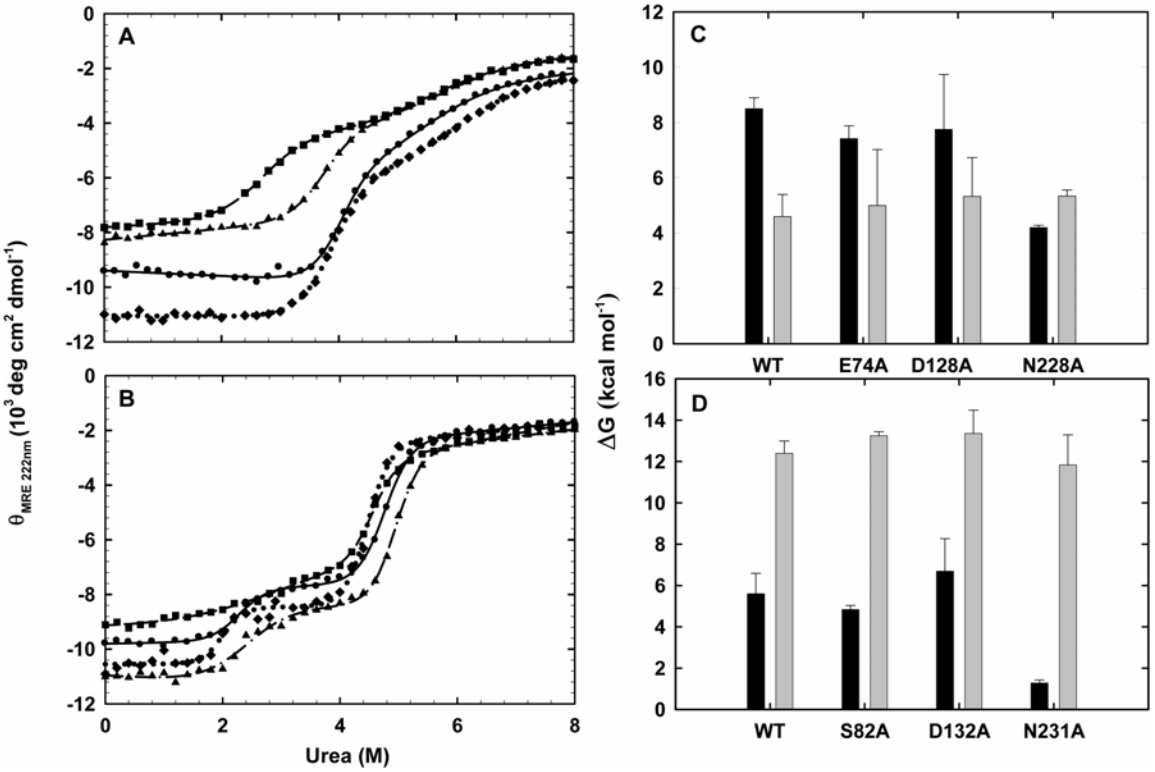

A
B
C
D
$\theta_{MRE\ 222nm}$ (10^3 deg cm^2 dmol^{-1})
Urea (M)
ΔG (kcal mol^{-1})
0
-2
-4
-6
-8
-10
-12
0
2
4
6
8
12
10
8
6
4
2
0
16
14
12
10
8
6
4
2
0
WT
E74A
D128A
N228A
WT
S82A
D132A
N231A

Figure 5.3

Figure 5.4 – Stability of eIGPS-Δβ7α7-N231A. The dependence of the amplitude for the unfolding phase for (**a**) eIGPS-WT (●,——) and (**b**) eIGPS-Δβ7α7-N231A (■,) on the initial urea concentration; the final urea concentration in all cases was 3 M urea. The lines represent the fit of the data to a two-state model with ΔG° 5.29 ± 1.71 kcal mol ^{1} and *m* 2.34 ± 0.74 kcal mol ^{1} M ^{1} for eIGPS-WT and ΔG° 1.28 ± 0.15 kcal mol ^{1} and *m* 0.89 ± 0.11 kcal mol ^{1} M ^{1} for eIGPS-Δβ7α7-N231A.

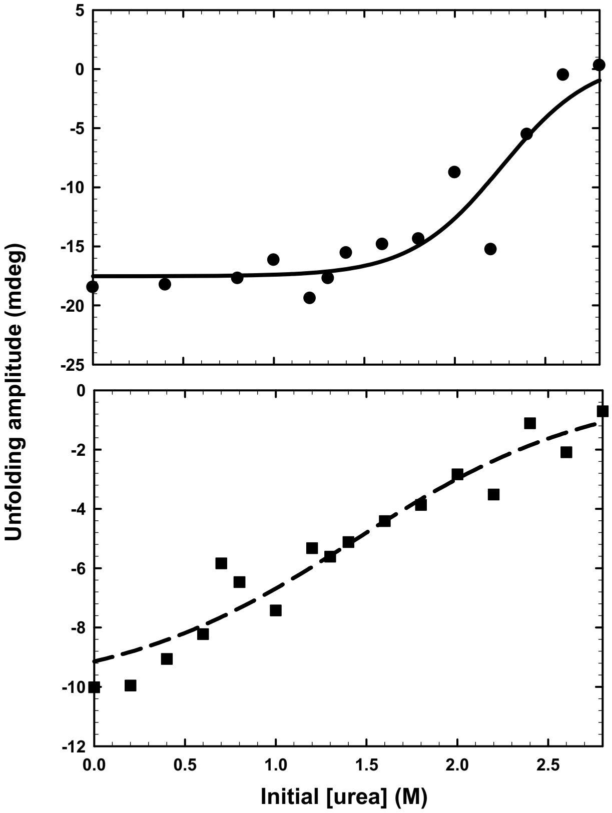

Unfolding amplitude (mdeg)
5
0
-5
-10
-15
-20
-25
0
-2
-4
-6
-8
-10
-12
0.0
0.5
1.0
1.5
2.0
2.5
Initial [urea] (M)

Figure 5.4

dependence of the stability, m_{NI}. These parameters are used to fit the CD unfolding transition for eIGPS-Δβ7α7-N231A (Fig. 5.3b) and to extract $\Delta G^\circ_{\mathrm{IU}}$ and m_{IU} (Materials and Methods).

The stabilities of N and I for the clamp-deletion variants and the WT parent sequences are illustrated graphically in Figures 5.3c and 5.3d for sIGPS and eIGPS, respectively. The free energy differences between N and I, $\Delta G^\circ_{\mathrm{NI}}$, and between I and the unfolded state, U, $\Delta G^\circ_{\mathrm{IU}}$, as well as the *m*-values, are tabulated in Table 5.1. The deletion of the β2α2 clamp in sIGPS, sIGPS-Δβ2α2-E74A, only reduces the stability of N by 1.08 kcal mol ^{1}, and the deletion of the β3α3 clamp, sIGPS-Δβ3α3-D128A, has no significant effect on its stability. By striking contrast, the elimination of the β7α7 clamp, sIGPS-Δβ7α7-N228A, reduces the stability of N by 4.30 kcal mol ^{1}. Consistent with the absence of these clamps in I for all of these variants, the free energy differences between I and U for the clamp-deletion variants are comparable to the corresponding value for sIGPS-WT (Fig. 5.3c and Table 5.1). Similar results are obtained for eIGPS. Only eIGPS-Δβ7α7-N231A decreases the stability of N significantly, ΔΔG°   4.32 kcal mol ^{1}. eIGPS-Δβ1α1-S82A and eIGPS-Δβ3α3-D132A have no significant effect on the stability of N, and none of the clamp-deletion variants perturb the stability of I relative to U (Fig. 5.3d and Table 5.1). Thus, while the elimination of either the β1α1, β2α2 or β3α3 clamps has only marginal effects on sIGPS and eIGPS, the β7α7 clamps in both proteins contribute significantly to both the structure and the stability of the native states for their resident TIM barrel protein.

Table 5.1 – Thermodynamic parameters for the urea-induced unfolding of sIGPS, eIGPS, αTS and eight βα-hairpin clamp-deletion variants[a].

a. Buffer conditions: 10 mM potassium phosphate, 0.2 mM K_2EDTA, 1 mM βME, pH 7.8 for sIPGS and pH 7.0 for eIGPS at 25 °C.

b. Perturbation in stability for the N to I reaction, calculated by $\Delta\Delta G^\circ_{NI} = \Delta G^\circ_{NI}$ (H_2O, variant) - ΔG°_{NI} (H_2O, WT).

c. Values are from Forsyth *et al.*[75]

d. Determined by fitting the urea dependence of the amplitude of the unfolding kinetic phase to a two-state model.

e. Determined by fitting the equilibrium unfolding data to a three-state model with parameters for the N to I transition fixed to the values determined as described in footnote c.

f. Values are from Yang *et al.*[143]

	Donor and acceptor pairs	Donor and acceptor distance (Å)	variants	$\Delta G°_{NI}$ (H_2O) (kcal mol^{-1})	-m_{NI} (kcal mol^{-1} M^{-1})	$\Delta G°_{IU}$(H_2O) (kcal mol^{-1})	-m_{IU} (kcal mol^{-1} M^{-1})	$\Delta\Delta G°_{NI}$ (kcal mol^{-1})^{b}
sIGPS			WT^{c}	8.50±0.40	2.10±0.10	4.60±0.80	0.86±0.13	-
	S104 –E74	3.3	E74A	7.42±0.46	1.97±0.13	5.00±2.02	0.86±0.36	-1.08±0.61
	I107-D128	3.0	D128A	7.75±1.99	1.99±0.12	5.33±1.40	0.89±0.23	-0.75±2.03
	K207-N228	2.7	N228A	4.20±0.08	1.56±0.03	5.34±0.22	0.97±0.04	-4.30±0.41
eIGPS			WT	5.60±0.99	2.46±0.42	12.39±0.60	2.60±0.13	-
	F50-S82	3.0	S82A	4.84±0.19	2.05±0.08	13.24±0.20	2.68±0.04	-0.76±1.01
	I111-D132	2.9	D132A	6.69±1.57	3.34±0.77	13.36±1.12	2.99±0.26	1.09±1.86
	V211-N231	2.8	N231A	1.28±0.15^{d}	0.89±0.11^{c}	11.84±1.45^{e}	2.62±0.31^{e}	-4.32±1.00
αTS^{f}			WT	7.19±0.58	2.85±0.24	3.04±0.85	0.81±0.17	-
	F19 –D46	2.8	D46A	1.98±0.45	0.78±0.17	4.97±1.96	1.07±0.39	-5.21±0.73
	I97-D124	2.6	D124A	2.53±0.40	1.12±0.19	3.81±0.64	0.79±0.16	-4.66±0.70

Table 5.1

Survey of βα-hairpin clamps in the TIM barrel proteins

The observation that βα-hairpin clamps can have a significant effect on structure and stability in three TIM barrel proteins motivated a survey of the prevalence of such non-local MC-SC H-bonds in the TIM barrel fold. This analysis was carried out for a structural database of 71 TIM barrel domains, previously reported as a non-redundant representation of the TIM barrel fold[250]. H-bonds between main chain amide hydrogens and polar side chains ($\mathrm{MC}_{NH} \rightarrow \mathrm{SC}$) that serve as βα-hairpin clamps in the TIM barrel domains were identified (Materials and Methods) for a direct comparison with experimental results.

In the 71 TIM barrel proteins examined, there are 131 $\mathrm{MC}_{NH} \rightarrow \mathrm{SC}$ βα-hairpin clamps. As can be seen in Table 5.2, there is a very significant preference, > 42% of the clamps (χ^2_{Yates} 592.49, n 131, d 3, p-value 4.26 × 10^{128}), for aspartic acid SCs forming H-bonds with the MC amide hydrogen of isoleucine, leucine and valine residues. Inspection of the location of the donor and acceptor residues in the β-strands reveals that every βα-hairpin clamp secures the N-terminus of one β-strand to the loop preceding the subsequent β-strand in the barrel.

The locations of the entire group of 131 $\mathrm{MC}_{NH} \rightarrow \mathrm{SC}$ βα-hairpin clamps are displayed in Figure 5.5a, with each clamp interaction represented as a bridge across two adjacent β-strands. A very strong preference (77%) is seen for β1α1, β3α3, β5α5 and β7α7 clamps, where the SC acceptor is C-terminal to the MC H-bond donor. With the exception of 13 β8α8 clamps, the paucity of β2α2, β4α4 and β6α6 clamps is distinct

Table 5.2 – Sequence preferences for MC$_{NH}$→SC βα-hairpin clamps in 71 TIM barrel proteins.

Numbers in parenthesis are the values expected from the distribution of MC$_{NH}$→SC H-bonds in 71 TIM barrel proteins that do not form βα-hairpin clamps and have at least 15 residues between the donor and acceptor residues.

MC_{NH} →

SC↓	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	SER	THR	TRP	TYR	VAL
ASN	1 (1)	1 (1)	(1)	1 (1)	1 (0)	(1)	(1)	2 (1)	(0)	4 (2)	2 (1)	2 (1)	(0)	(1)	(1)	(1)	(0)	1 (1)	2 (1)
ASP	2 (2)	2 (2)	(2)	(2)	3 (1)	(1)	(1)	(5)	1 (2)	**18** (1)	**14** (1)	2 (1)	2 (1)	5 (2)	(5)	2 (3)	(1)	1 (2)	**23** (3)
GLN	1 (1)	(0)	(0)	1 (1)	(0)	(0)	(0)	(1)	(0)	1 (1)	1 (1)	(1)	(0)	1 (0)	(0)	2 (1)	(0)	(1)	1 (1)
GLU	(3)	1 (1)	(2)	(2)	(0)	2 (0)	(1)	(4)	(0)	2 (1)	1 (3)	1 (1)	1 (1)	1 (1)	2 (5)	(2)	(0)	(1)	1 (1)
HIS	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	1 (0)	1 (0)	(0)	1 (0)	(0)	(0)	(0)	(0)	(0)	(0)
SER	(1)	(1)	(0)	(0)	(0)	(0)	(1)	2 (3)	(1)	1 (0)	2 (1)	(0)	(0)	(1)	(1)	(1)	(0)	(1)	2 (1)
THR	(0)	(0)	(1)	1 (0)	(0)	(0)	(0)	1 (2)	(0)	1 (0)	(0)	(0)	(0)	1 (0)	(1)	1 (0)	(0)	(0)	2 (0)
TYR	(1)	(0)	(0)	(1)	(0)	(1)	(0)	(1)	(0)	(1)	1 (1)	(0)	(0)	(0)	(1)	(0)	(0)	(0)	(1)

Table 5.2

Figure 5.5 – Positional preference of βα-hairpin clamps in 71 TIM barrel proteins. (**a**) The TIM barrel architecture is represented by a cross-sectional view of the 8 β-strands, represented as rectangles and the strand number is indicated. The number of MC$_{NH}$ → SC βα-hairpin clamp interactions connecting adjacent β-strands with SC H-bond acceptors C-terminal to the MC$_{NH}$ donors (——), and with SC H-bond acceptors N-terminal to the MC$_{NH}$ donors (   ) are indicated. The number of βα-hairpin clamps with (I/L/V) MC → SC (D) is represented in parenthesis. (**b**) The positional preference of (I/L/V) MC → SC (D) relative to the β-strands. The MC donor prefers either the first or second position of the β-strand and the SC acceptor prefers to be in the loop immediately preceding the subsequent β-strand. The number of times each pair of interactions occurs in the 55 I/L/V MC → SC D sub-set is indicated.

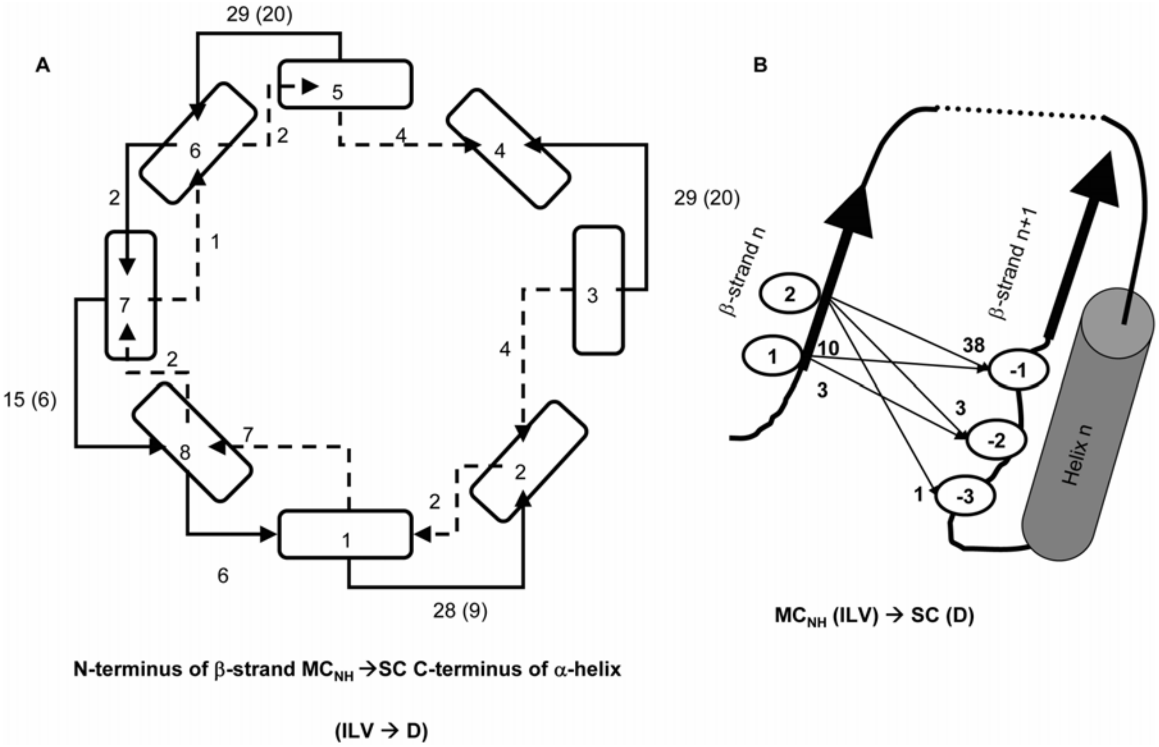

A
29 (20)
5
6
2
4
4
29 (20)
2
1
7
3
4
2
15 (6)
8
7
2
2
1
6
28 (9)
N-terminus of β-strand MC$_{NH}$ →SC C-terminus of α-helix
(ILV → D)
B
β-strand n
β-strand n+1
2
1
10
3
38
-1
3
-2
1
-3
Helix n
MC$_{NH}$ (ILV) → SC (D)

Figure 5.5

from their odd β-strand counterparts. The relatively large number of clamps for the β8β1 interface may reflect the necessity for securing the N- and C-terminal β-strands. Far fewer βα-hairpin clamps, in which the SC acceptor is N-terminal to the MC H-bond donor, are observed. Highlighting the significance of this distribution pattern, the 55 Ile, Leu and Val (I/L/V) MC → SC Asp (D) sub-group of βα-hairpin clamps always have their MC H-bond donor I/L/V located in the odd-numbered stands, β1, β3, β5 or β7, and their SC acceptor, D, is always located before the succeeding even-numbered β-strands, β2, β4, β6 and β8. There is also a strong preference for the I/L/V residue to occupy the 2nd position in the odd-numbered β-strand and for the D residue to occupy the position immediately preceding the even-numbered β-strand (Fig. 5.5b). This positional preference braces two consecutive and adjacent β-strands, along with the intervening helix, and reinforces the β-strand register required for the canonical TIM barrel architecture[101].

Discussion

Experimental analysis of βα-hairpin clamps between MC H-bond donors and SC H-bond acceptors in three TIM barrel proteins, αTS[143], sIGPS and eIGPS, has shown that a subset of these non-covalent interactions make substantive contributions to stability. Comparisons of the potency of the βα-hairpin clamps in these three proteins shows no correlation between the contributions of these clamps to stability and either the location of the clamps in the structure, their contributing residues or their relative exposure (0 -

25 %) to the solvent. The observation of potent clamps formed by the neutral N228 in sIGPS and N231 in eIGPS, the β7α7 clamps, also shows that the formal negative charge on the aspartic acid H-bond acceptors in the remaining two potent βα-hairpin clamps is not determinative of the strength of the clamp interaction. An examination of the crystal structures of the three proteins, however, suggests that the length of the H-bond in each structure differentiates between the clamps that make major or minor contributions to stability (Table 5.1). Although the nominal resolutions of the crystal structures of these proteins, 2.0 to 2.8 Å[253,254,257], dictate that the correlation between H-bond length and the clamp contribution to protein stability be viewed as tentative, it appears βα-hairpin clamps whose H-bonds are less than 2.8 Å in length are those, which when replaced with alanine, reduce the stability of the native state by 4-6 kcal mol ^{1}. The apparent correlation provides a logical and testable hypothesis for future experiments on βα-hairpin clamps in other TIM barrel proteins.

The assay for the contribution of the βα-hairpin clamps to the stability of three TIM barrel proteins involves the replacement of the polar side chain H-bond acceptors, asparagine, aspartic acid, glutamic acid and serine, with alanine. The absence of the H-bond acceptor moiety is accompanied by the introduction of a potential void for these buried side chains, reflecting the absence of chemical mass as the side chain is truncated to the β-carbon. The perturbations in the secondary and/or tertiary structures induced by the mutations (Fig. 5.2c and 5.2d) show that the loss of the clamp is propagated to numerous other non-covalent interactions via the global cooperativity of the native conformation.

The absence of the βα-hairpin clamps in the I states of all three TIM barrel proteins demonstrates that the potent effects of these clamps only appear as the N state appears[59]. Kinetic folding studies on αTS revealed further that each clamp is crucial for accessing the transition state ensemble required to reach the properly-folded structure[143]. Although the local connectivity of the βαβ modules might have been expected to enable the clamp to have a role in the early stages of the folding reaction, the primary role of the potent set of clamps is to drive the final stage of the reaction to completion and fully develop global cooperativity.

The 4-fold symmetry of the preferred βα-hairpin clamps is mirrored, not only in the symmetry of the βαβ modules, but also in the packing of the side chains in the interior of the β-barrel. A residue oriented towards the inside of the β-barrel from an odd-numbered β-strand is at the same level as corresponding residues from the three remaining odd-numbered β-strands. The next layer is comprised of the four side chains from the even-numbered β-strands (Fig. 5.6); the third, and usually final layer, is comprised again of side chains from the odd-numbered β-strands[101]. The layering of side chains inside the barrel has its origin in the tilt of the β-strands (35°) with respect to the central axis of the β-barrel[258]. The resulting S   8 shear[101,258] provides a favorable orientation for the H-bonding network between adjacent parallel β-strands and provides opportunities for MC-SC βα-hairpin clamp interactions. Together, these non-covalent interactions and others stabilize the $(\beta\alpha)_8$, TIM barrel fold (Fig. 5.6). The observation of similarly placed non-local MC-SC interactions in a limited survey of flavodoxin fold

Figure 5.6 – Architectural principles of the TIM barrel fold. The strand number of the 8 β-strands of the TIM barrel architecture (↗) is indicated at the C-terminus of each β-strand. To convey the closed barrel architecture, β8 is shown adjacent to β1 (↗) as well as adjacent to β7. The position of each residue on the β-strands, with SC pointing into the β-barrel (⬬) and SC pointing towards the α-helices (⬭), is indicated. The one letter code for the most common amino acids (>15%) in the loop preceding the β-strand in the 71 TIM barrel proteins database is shown. The H-bond network for the β-barrel (⇢), the βα-hairpin clamp interactions between the second residue of an odd-numbered β-strand and the side chain of the residue immediately preceding the subsequent even-numbered β-strand (→), and the MC-MC interactions between the same two residues (→) are indicated

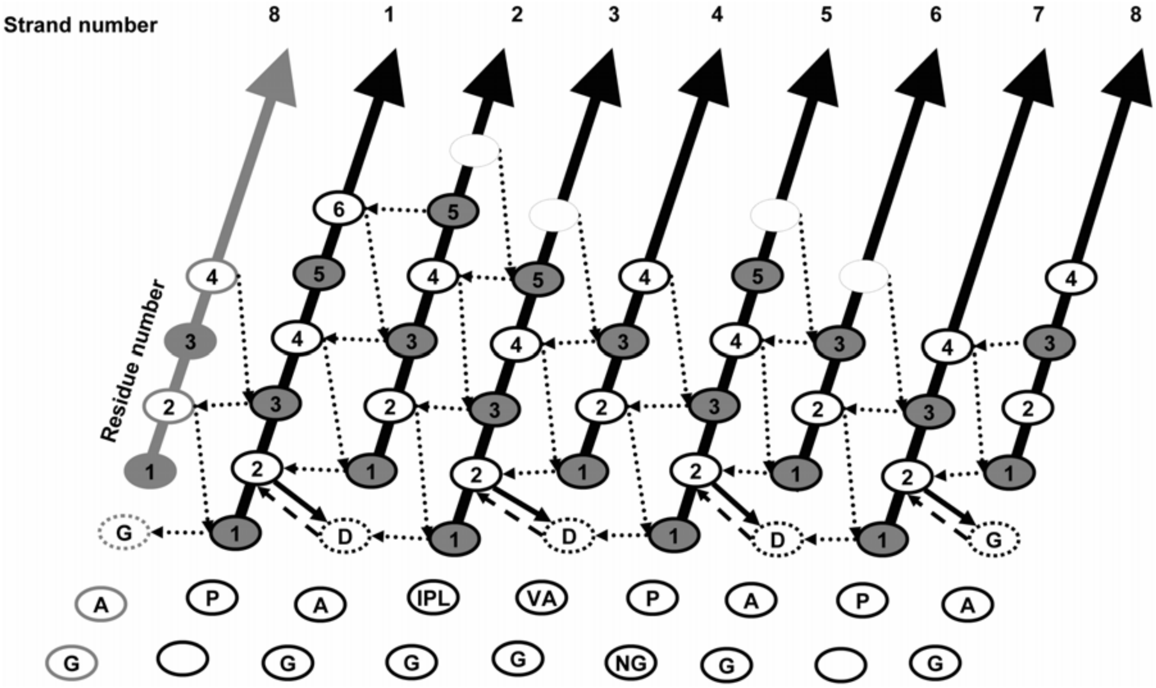

Strand number
8
1
2
3
4
5
6
7
8
Residue number
G
D
D
D
G
A
P
A
IPL
VA
P
A
P
A
G
G
G
G
NG
G
G

Figure 5.6

proteins (data not shown) suggests that βα-hairpin clamps are a common structural feature of βα-repeat proteins.

The chemical origin for the asymmetry between odd- and even-numbered β-strands is apparent from an inspection of the residue preference (>15%) at positions preceding the N-terminus of each β-strand (Fig. 5.6). The conserved proline just before odd-numbered β-strands provides a kink in the backbone that marks the beginning of a β-strand[241]. The preferred sequence pattern of the tight turn connecting the α-helix and the subsequent even-numbered β-strand (Fig. 5.6), GAD, has been reported previously[166]. The positive φ angle allowed by glycine and the hydrophobic nature of alanine immediately following the α-helix enables a Schellman motif for the C-terminal capping of the helix[245] and a tight turn to the next β-strand. The aspartic acid just prior to the beginning of even-numbered β-strands forms the βα-hairpin clamp and braces the βαβ module. This N-terminal cap for the odd-numbered β-strand is very often complemented by a MC-MC H-bond, with the amide group of the aspartic acid acting as the donor to the MC carbonyl oxygen of the partner residue. While other SC acceptors are observed (Table 5.2), the length of the aspartic acid side chain appears to be optimal for the reinforcement of the MC-SC H-bond with the MC-MC H-bond, providing a plausible explanation for its higher frequency in βα-hairpin clamps.

The preference for I/L/V residues at the MC_{NH} H-bond donor position may reflect, in part, the > 40% occurrence of these residues in parallel β-strands of TIM barrel proteins[101]. Further, along with alanine and glycine, I/L/V are the only amino acids that do not partition favorably from the vapor phase to water[103]. As such, these large aliphatic side chains are especially effective at excluding water from MC-SC H-bonds in the βα-hairpin clamps. The exclusion of water, that is apparent from the limited access to solvent for the H-bond donor and acceptor atoms of potent clamp interactions in αTS, sIGPS and eIGPS (Fig. 5.1c and 5.1d), is expected to strengthen these H-bonds and make them more resistant to exchange with solvent, as observed previously for αTS[58,59,189]. This presumption is supported by the conclusions of Gao *et al.*[259], who recently reported that the strength of a MC-MC H-bond is inversely related to the polarity of its local environment. Valine more effectively screened an underlying β-sheet MC-MC H-bond from solvent than alanine in a Pin WW domain, increasing the strength of the H-bond by up to 1.2 kcal mol ^{1}.

The occurrence of the βαβ motif in a large number of protein families[101,239] suggests that the N-terminal capping of β-strands by βα-hairpin clamps, akin to the analogous N-capping of α-helices[244,245], may be a useful property for the refinement of protein fold prediction and for engineering stability in βα-repeat proteins. βα-repeat proteins are readily recognized from their sequences and the predicted alternating patterns of α-helices and β-strands[260]. The refinement of the 3D structures predicted from knowledge-based potentials[261], threading[262] and homology modeling[263] of these protein sequences, could be enhanced by screening for βα-hairpin clamps between the MC amide hydrogens at favored positions near the N-terminus of a β-strand and H-bond acceptor SC in the loop before the subsequent β-strand (~25 residues apart in sequence).

These clamps would establish the register of the pair of β-strands, and, with the very short loop linking the intervening α-helix to the second β-strand, it might be possible to establish the register of the α-helix on the β-strand pair in the βαβ module. Although TIM barrel proteins typically contain only a few βα-hairpin clamps, defining the spatial relationships of the components of a subset of βαβ modules might increase the probability of predicting the packing of adjacent βα-repeats in the structures. The effect of accurately predicting the structure of one βαβ module might, therefore, propagate throughout the TIM barrel protein.

The TIM barrel architecture provides a scaffold that is capable of a very diverse set of enzymatic functions[166], and this property has enabled TIM barrel enzymes to be re-engineered in order to accommodate alternative substrates[264 267] and even to catalyze non-biological reactions[268]. Because the active sites of TIM barrel enzymes are invariably comprised of the loops protruding from the C-termini of the β-strands, engineering βα-hairpin clamps at the N-termini of the β-strands offer a unique opportunity to enhance the stability of TIM barrel proteins without compromising function.

Materials and Methods

Clamp-deletion variants

The plasmid encoding a truncated version of sIGPS, in which the non-canonical additional α-helix (α00) at the N-terminus was deleted to eliminate aggregation during folding, pTNI4[75], was obtained from Dr. K. Kirschner (University of Basel, Switzerland). The plasmid coding for eIGPS, pJB122[269], was obtained from Dr. J. M. Blackburn (University of the Western Cape, South Africa). The eIGPS, with an additional Ala residue after the start codon and a C-terminal FLAG peptide sequence (GSDYKDDDDK), is fully folded and catalytically active[269]. Oligonucleotides for mutagenesis were purchased from Eurofins MWG Operon (Huntsville, AL), and the Quickchange^{TM} site-directed mutagenesis kit was obtained from Stratagene (La Jolla, CA). The site-directed mutations were confirmed by DNA sequence analysis (Genewiz Inc, NJ).

Protein expression and purification

The sIGPS protein and its variants were expressed in BL21/DE3 cells and purified as described previously[75]. The expression and purification of eIGPS and its variants followed the same protocol, with the exception that the procedures were conducted at pH 7.0. The purity (>95%) was demonstrated by the appearance of a single band Coomassie blue stained PAGE and confirmed using electrospray mass spectrometry at the Proteomics Facility at the University of Massachusetts Medical School (Worcester, MA).

Circular dichroism

Far- and near-UV CD spectroscopy was employed to monitor the secondary and the tertiary structure near aromatic side chains, respectively. Spectra were obtained on a Jasco Model J-810 spectropolarimeter equipped with a thermoelectric cell holder. Far-UV CD data were collected from 280 nm to 185 nm at a scan rate of 50 nm/min and at 1 nm intervals using a 0.1 cm pathlength cell, a bandwidth of 2.5 nm, with an averaging time of 8 s. Three replicate spectra were collected and averaged. The protein concentration was 5 μM. Near-UV CD data were collected from 350 nm to 250 nm at 5 nm/min using a 0.5 cm path length cell, and the protein concentration was 50-150 μM. The temperature was maintained at 25 ºC with a computer-controlled Peltier system.

Thermodynamic measurements

The stability of the IGPS clamp-deletion variants was measured by urea denaturation as described previously[75] in a buffer containing 10 mM potassium phosphate, pH 7.8 for sIGPS and pH 7.0 for eIGPS, 0.2 mM K2EDTA, and 1 mM βME. A Hamilton 540B automatic titrator was used to prepare the samples containing 0 to 8 M urea at concentration increments of 0.2 M urea to enhance the precision of the measurements. The samples were incubated overnight at 25 ºC to ensure equilibration.

Kinetic experiments

CD manual-mixing kinetic experiments were performed on a Jasco Model J-810 spectropolarimeter equipped with a thermoelectric cell holder using a 1 cm pathlength cell, a bandwidth of 2.5 nm, and an averaging time of 1 s. The dead-time of the experiments was 3 s, and the instrument response time was about 5 s. The change in ellipticity as a function of time was monitored at 222 nm. Kinetic unfolding experiments to determine the stability of the N state of the clamp-deletion variant were performed by jumping from different initial urea concentration (0-2.8 M) to a final concentration 3 M urea. Protein samples were first equilibrated in the initial urea concentration overnight and then jumped to 3 M urea in buffered solutions by a 1:10 dilution. The final protein concentration was 5 μM.

Data analysis

Equilibrium CD data at 222 nm were fit to a three-state model, N $\leftrightarrows$ I $\leftrightarrows$ U, as described previously[34]. All thermodynamic folding data were fit using Savuka version 5.2, an in-house, non-linear, least-squares program[34].

Survey of TIM barrel proteins

A database of 71 TIM barrel proteins has been previously developed (http://www.cbrc.jp/~gromiha/tim/proteinlist.html[250]) from the SCOP[85] and HOMSTRAD[270] databases, with a pair-wise sequence homology of < 25%. The highest resolution structure for each domain was chosen from the Protein Data Bank[271]. The secondary structure was calculated using the DSSP program[272] and the H-bond interaction parameters were calculated using default settings of the HBPLUS program[251].

Definitions of βα-hairpin clamp interactions

For each protein, the 8 canonical β-strands and α-helices in the context of the TIM barrel architecture were identified and labeled accordingly. H-bonding partners identified using the HBPLUS program[251], were subjected to the following filters: 1) the H-bonds must be between a MC amide donor and a SC acceptor, 2) the amino acid chain length between the donor and acceptor must be ≥ 15 residues thereby eliminating shorter-range helix-capping interactions[244,245] and 3) the chain must include exactly one β-strand and one α-helix identified in the context of the TIM barrel architecture. For the case of the β8α8-hairpin clamps, MC_{NH} → SC H-bonds between the residues prior to β1 and β8 were included. The H-bonds that passed each stage of the filtering process were exported to a PyM0L[252] script in color-coded fashion for manual confirmation.

Statistical significance of residue preference for βα-hairpin clamps

The frequency of MC_{NH} → SC H-bonds in the 71 TIM barrel proteins, where the donor and acceptor residues were at least 15 amino acids apart and were not involved in βα-hairpin clamp interactions, was determined. This frequency was used to calculate the expected frequency of H-bonding between any two types of residues and compared to that observed in βα-hairpin clamps. Four categories, Ile MC → SC D, Leu MC → SC D, Val MC → SC D, Other MC → SC Other, were used to determine the χ^2 distribution probabilities, with Yates correction[273], of observed βα-hairpin clamps.

Acknowledgements

The authors would like to thank Dr. K. Kirschner (University of Basel, Switzerland) and Dr. J. M. Blackburn (University of the Western Cape, South Africa) for sharing the plasmids encoding the sIGPS and eIGPS proteins, respectively. We also thank Dr. Jill A. Zitzewitz, Dr. Osman Bilsel, Dr. Zhenyu Gu, and Can Kayatekin for helpful discussions and assistance in the preparation of this manuscript.

Chapter VI – Discussion

Summary

As demonstrated in Chapters II and III, three members of the CheY-like family of proteins, CheY, NT-NtrC and Spo0F, all appear to follow an apparent two-state like equilibrium profile. In all three cases, a large burst-phase ellipticity, implying significant secondary structure, is observed during refolding. The refolding kinetics are further complicated by the presence of a conserved *cis*-proline in the native structure. Global analyses of the equilibrium and kinetic data in all three proteins suggest that the intermediate formed is a kinetically-trapped species that must at least partially unfold before the native state can be accessed. Observation of fast folding kinetics in the microsecond timescale would provide a test of the conclusion that the kinetic intermediate is indeed off the productive folding pathway. Low resolution structural studies of this intermediate by CD, FRET and SAXS will also provide insights into the driving forces for its formation.

Gō-simulations performed by the Brooks group have suggested structurally different sources for the topological frustrations in the CheY-like family of proteins. The striking correspondence of the location and extent of sequence local clusters of isoleucine, leucine and valine residues with the topologically-frustrated regions predicted by Gō-simulations suggest a unique role for these hydrophobic clusters in directing the early misfolding of CheY-like proteins. Another cluster of ILV residues, comprised predominantly of non-local interactions, is thought to be essential for the productive folding to the native state. Non-native packing of residues belonging to this latter cluster resulting from the early formation of the sequence-local cluster possibly adds to the stability of the off-pathway intermediates. Mutations of residues involved in the two clusters are expected to have differential effects in the formation of the off-pathway intermediate and the native state. The working hypothesis is that a mutation in the sequence local cluster will affect the formation of the intermediate and will perturb the stability of the native state, whereas a mutation in the non-local cluster will destabilize the native state, but only marginally affect the intermediate.

It has been demonstrated in a pair of TIM barrel proteins, αTS[59] and sIGPS,[36] that not only do non-native interactions between residues of ILV clusters direct the early misfolding reaction; the same core is also responsible for protection of the underlying amide hydrogens from exchange against solvent. However, the location and extent of the core in these proteins is not conserved. These observations have led us to investigate the role of clusters of Branched Aliphatic Side Chain (BASiC) residues as the cores of stability in globular proteins (Chapter IV). Greater than 70% of the residues that are shown to be protected in native state hydrogen exchange experiments are within 6 Å of a large cluster of BASiC residues. However, of all the residues in the vicinity of the cluster, only 50% are protected from HX. On closer inspection of the results, we find that the BASiC cluster method can help identify the general region of the protein that would be most protected. Improving the specificity of this technique would require the inclusion of other parameters, such as distance from the surface of the protein, hydrogen bonding partners, crystallographic B-factors, etc. Further analyses of the correlations between BASiC residue clusters and residues with high phi-values will provide insights into their roles in defining the folding pathways and the structures of the transition state ensembles.

Another type of influence a side-chain can have on the stability of a protein is the formation of H-bonds. An interesting category of H-bonds that contribute significantly to the stability of TIM barrel proteins has been reported recently by the Matthews lab.[143] Prompted by the work of Xiaoyan Yang, another graduate student in the Matthews lab, a survey of 71 TIM barrel proteins was conducted. This survey demonstrated that the non-local H-bonds between main chain amides and side chain carboxyl groups, linking the N-terminus of a β-strand to the C-terminus of the subsequent α-helix, appear to be a recurring feature in the $(\beta\alpha)_8$ repeat topology of these proteins (Chapter V). Interestingly, specific patterns in the distribution of these "clamp" interactions suggest an underlying linkage with the architecture of the TIM barrel fold and a possible evolutionary significance with respect to the βαβ precursor module of $(\beta\alpha)_n$-repeat proteins. The underlying role of neighboring hydrophobic clusters in defining the potency of clamp interactions remains to be investigated.

The off-pathway mechanism

It is intriguing that several members of the $(\beta\alpha)_n$ repeat proteins studied form off-pathway intermediates early during their folding reactions.[34 39] The predominant contacts in both of these topologies are due to the packing of sequential elements of secondary structure elements and only a few non-local contacts are formed between the two termini.

In the case of αTS, the cylindrical barrel, comprised of 8 β-strands, has been shown to be formed in an on-pathway intermediate.[189] Thus the subsequent transition state must require the formation of the long-range contacts between the N- and C- termini. Similarly in CheY, the productive transition state intermediate has been shown to comprise at least part of the non-local hydrophobic cluster.[127] It is, therefore, possible that the preponderance of local contacts in both of these topologies permit the independent formation of several locally-connected nuclei, each with a very low probability of propagating to the native structure. On the other hand, the early formation of native-like non-local contacts would sufficiently restrict the conformational space available and rapidly access the productive transition state on route to the native state.

The global modeling presented in Chapters II and III is consistent with off-pathway intermediates for all three proteins and is corroborated by Gō-simulations for CheY and Spo0F. However, conclusive evidence for the off-pathway mechanism for all three proteins has been elusive as it is non-trivial to distinguish between the following three scenarios: i) an off-pathway intermediate, ii) an obligate on-pathway intermediate and iii) a non obligate on-pathway intermediate, where a significant proportion of the molecules can directly access the native state.

Off-pathway intermediate

$$\begin{tikzcd}
& I \arrow[dl, shift left] & \\
U \arrow[ur, shift left] \arrow[rr, shift left] & & N \arrow[ll, shift left]
\end{tikzcd}$$

Obligate on-pathway intermediate

$$\begin{tikzcd}
& I \arrow[dl, shift left] \arrow[dr, shift left] & \\
U \arrow[ur, shift left] & & N \arrow[ul, shift left]
\end{tikzcd}$$

non obligate on-pathway intermediate

$$\begin{tikzcd}
& I \arrow[dl, shift left] \arrow[dr, shift left] & \\
U \arrow[ur, shift left] \arrow[rr, shift left] & & N \arrow[ul, shift left] \arrow[ll, shift left]
\end{tikzcd}$$

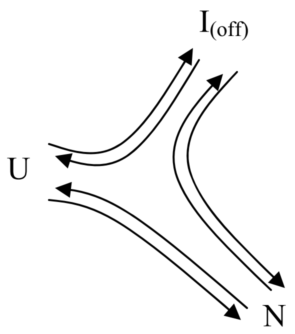
The m-values of the intermediates and the transition states, which are indicators of the relative amounts of surface areas buried, are used in Chapters II and III to demonstrate that the intermediate is more structured than the transition state preceding the native conformation and thus not consistent with the obligate on-pathway model. No significant differences are observed between the fits for the global modeling based on the off-pathway model and the non obligate on-pathway model. Although the rearrangement of the intermediate to form the native state (on-pathway) may be possible, it is not the dominant means for the folding mechanism.

Further, as suggested by the Gō-simulations[38,62] the backtracking reaction requires only the unfolding of specific contacts between the folding sub-domains; and for subsequent access to the productive transition state. Thus only partial unfolding of the off-pathway intermediate is required for folding to the native state. One means of representing this partial unfolding of the off-pathway intermediate is the following

$$\begin{tikzcd}
& I_{(off)} \arrow[d, shift left] & \\
U \arrow[r, shift left] & I_{(on)} \arrow[l, shift left] \arrow[r, shift left] \arrow[u, shift left] & N \arrow[l, shift left]
\end{tikzcd}$$

This model suggests that the equilibrium m-value and the m-value calculated from the chevron for the direct folding path may be different in order to account for the surface area buried by $I_{(on)}$. However, the conditions under which the off-pathway species is not visited, > 2 M Urea, and the protein folds directly to N, would also be unfavorable for $I_{(on)}$. Thus it is not possible to estimate the m-value, the stability or the kinetics for the formation of this species. Moreover, the sequence of events leading to the off-pathway intermediate may not require the formation of the same species.

A more accurate representation of the off-pathway model is shown below



In this model the sequence of events leading to the formation of the off-pathway species and the sequence of events during its partial unfolding on-route to the native state may be different; the unfolding of the off-pathway species, on-route to the native state does not require access to the unfolded state; no addition intermediate is required in the equilibrium folding pathway.

Future directions

As a means of validating the off-pathway folding mechanism in CheY-like proteins, two independent studies are proposed:

Mutagenesis to study the role of ILV clusters in the burst-phase intermediate

Mutations in the two ILV clusters are expected to have differential effects in the stability of the folding intermediate and the native state. Building upon the limited mutagenesis analysis performed by Serrano and his colleagues on CheY, a modified phi-analysis can be used to determine the extent of structure in the intermediate state.

However, due to the limited stability of CheY (5.4 kcal mol ^{1}), large perturbations to the core of the protein may not be feasible. To overcome this limitation, CheY containing the F14N mutation (pWT-CheY), which has been reported to enhance the native state stability of CheY by ~3 kcal mol ^{1},[91] can be used as the host for the amino acid replacements.

Several mutations to the ILV core have been engineered in the background of the stabilizing pWT-CheY (Fig. 6.1). As mentioned earlier, mutations in the sequence-local cluster are predicted to affect the formation and stability of the intermediate and to perturb the stability of the native state. By contrast, a mutation in the non-local cluster will destabilize the native state, but only marginally affect the more loosely-packed region of the intermediate.

Including the fast folding data in the global analysis

One of the limitations of the global analyses presented in Chapters II and III is that the burst-phase reactions of CheY, NT-NtrC and Spo0F are modeled as single events occurring in the microsecond timescale during the refolding of the three proteins. The denaturant dependence of the burst-phase amplitude follows a sigmoid curve and a fit of these data to a two-state model is used to determine the relative apparent stabilities of the intermediates. The extracted m-values are used to estimate of the amounts of surface areas buried by the burst-phase intermediates. With the ability to observe microsecond kinetic refolding data, it may be possible to obtain the refolding rates for the burst-phase intermediates and determine a higher degree of confidence on their m-values. However, even with the inclusion of these data in the global analysis, it may be difficult to

Figure 6.1 – Cartoon representation of the crystal structure of CheY. (3CHY). The sequence local ILV cluster is represented by red spheres, and the long-range cluster is shown as blue spheres. The proposed β-strand mutations in the ILV residues that contribute to the two clusters are highlighted as yellow spheres.

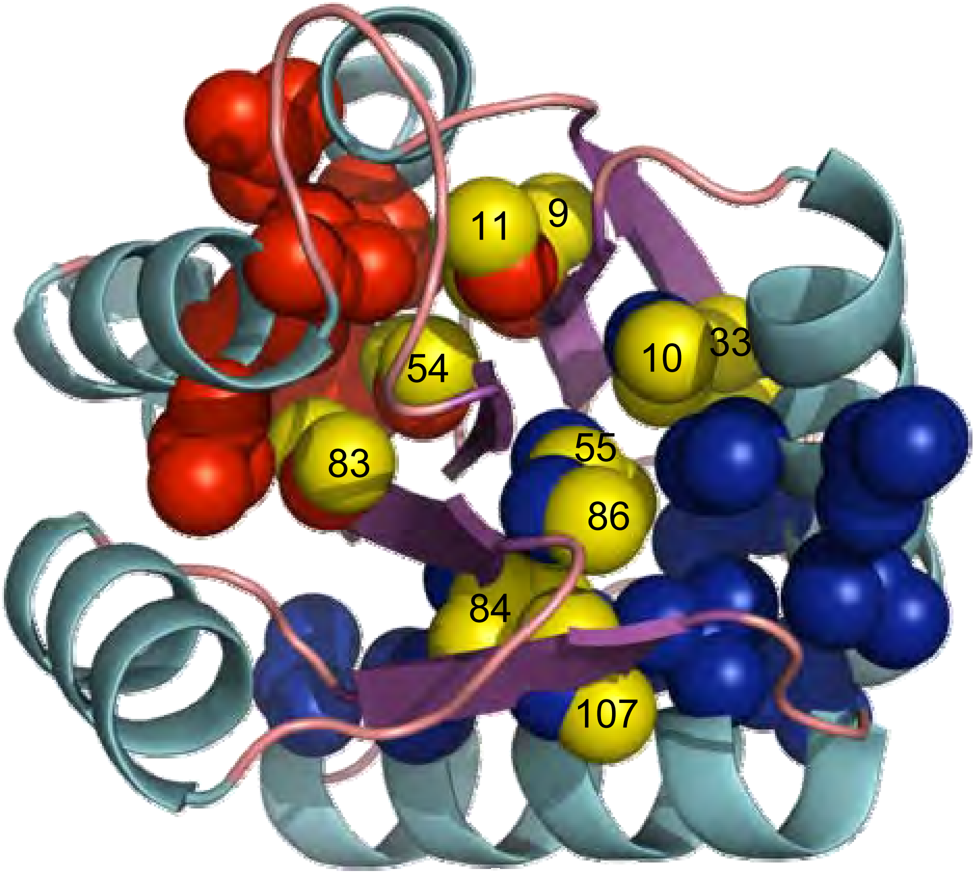


Figure 6.1

completely eliminate the possibility of an on-pathway intermediate that is identical to the TSE.[35,144]

Continuous flow technology with Fluorescence, CD, SAXS, and FRET: Microsecond kinetic data have become accessible by the development of microfluidic mixing devices[79,274] and by T-jump kinetics.[275] The continuous-flow mixer developed in the Matthews lab[274] has been interfaced with a host of spectroscopic instruments to obtain folding kinetics in αTS and *E.coli* dihydrofolate reductase (DHFR) in the 30 μs to 1.2 ms time-range. The kinetics with which low-resolution structural features develop can be accessed by interfacing the mixer with time-resolved fluorescence spectroscopy (tertiary packing interactions of the fluorophores and changes in overall size of the protein), circular dichroism (secondary structure development and changes in tertiary packing interactions of aromatic residues), and small angle x-ray scattering (SAXS) (changes in size and shape of the protein). Changes in distance distribution along specific dimensions can also be measured by the incorporation of FRET probes in the protein. Data obtained by these techniques will provide structural information as a function of refolding time, which can be used to identify the location and extent of structure in the intermediate and will improve the confidence in the folding models for these proteins.

Preliminary results:

NT-NtrC: The equilibrium refolding and unfolding curves of NT-NtrC as measured by fluorescence lifetimes of its two tryptophans can be fit to a two-state model (Fig. 6.2a). The stability and its denaturant dependence are comparable with those

Figure 6.2 – Fast folding kinetics of NT-NtrC. a) The equilibrium unfolding titration of NT-NtrC monitored by tryptophan fluorescence lifetimes. The broken lines represent the unfolded and native baselines. The fit to a two-state model is shown as the solid line. b) Refolding kinetics of 5 μM NT-NtrC in 0.8 M urea and 10 mM potassium phosphate at 25° C and pH 7.0, monitored by tryptophan fluorescence lifetimes. The unfolded and native state values under these conditions are indicated. c) The same refolding kinetics monitored by total intensity of tryptophan fluorescence. d) Chevron analysis of NT-NtrC. The fast folding relaxation times measured by fluorescence intensity are shown as filled circles. The slower millisecond and seconds scale refolding and unfolding kinetics are adapted from Chapter III.

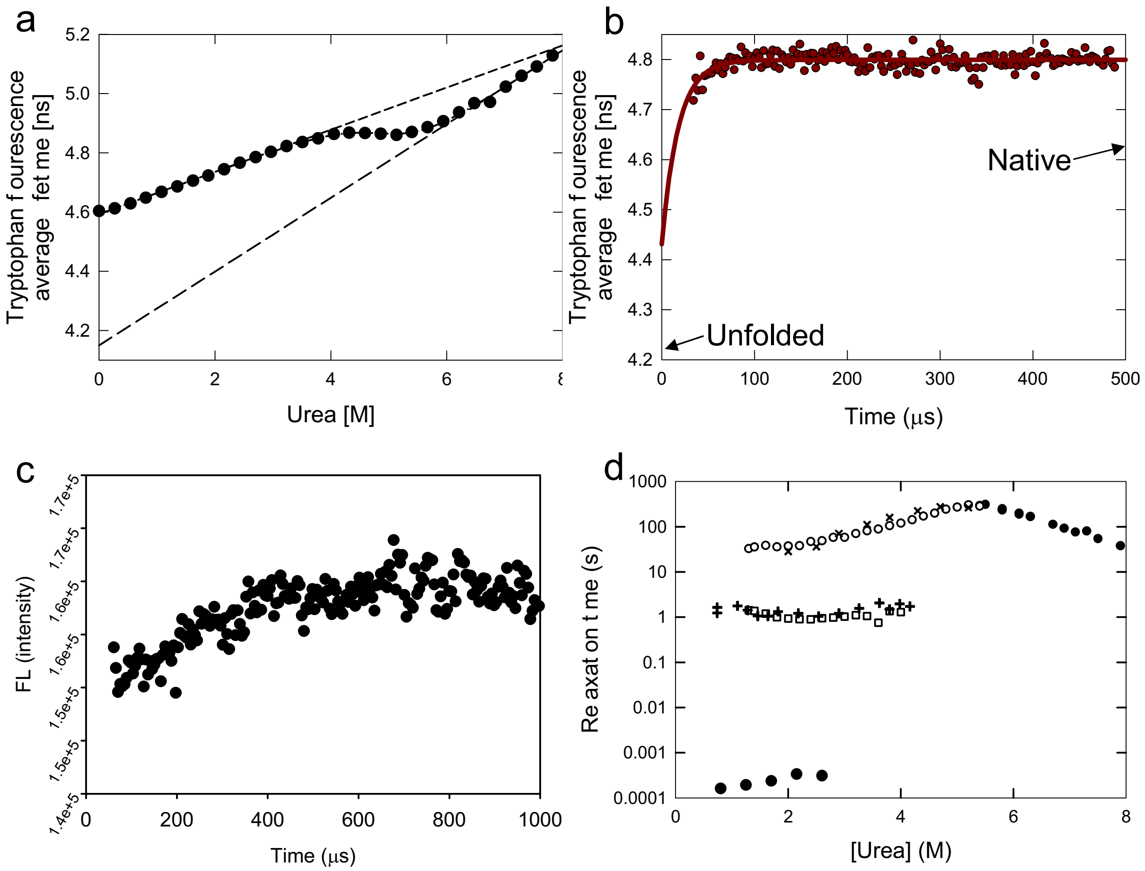

a
Tryptophan f ourescence average fet me [ns]
5.2
5.0
4.8
4.6
4.4
4.2
0
2
4
6
8
Urea [M]
b
Tryptophan f ourescence average fet me [ns]
4.8
4.7
4.6
4.5
4.4
4.3
4.2
Native
Unfolded
0
100
200
300
400
500
Time (μs)
c
FL (intensity)
1.7e+5
1.7e+5
1.6e+5
1.6e+5
1.5e+5
1.5e+5
1.4e+5
0
200
400
600
800
1000
Time (μs)
d
Re axat on t me (s)
1000
100
10
1
0.1
0.01
0.001
0.0001
0
2
4
6
8
[Urea] (M)

Figure 6.2

reported in Chapter III by far-UV CD spectroscopy (ΔG 7.5 kcal mol^{1} m-value 1.5 kcal mol^{1} M^{1}).[38]

The microsecond refolding kinetics measured by fluorescence lifetimes (Fig 6.2b) shows a rapid increase in lifetime that is faster than the dead time of the instrument (30 μs) and could therefore not be fit. This increase in lifetime is consistent with the burial of one or both tryptophan residues (W7 and W17) during refolding. This increase however, results in the non-native packing of at least one of the tryptophan side chains as the resulting fluorescence lifetime is higher than that seen in the native state under similar conditions (Fig. 6.2b). A second kinetic phase, ~200 μs, is observed by total fluorescence intensity measurements (Fig. 6.2c). The non-native packing observed in the first refolding phase is unaffected by this phase. The denaturant dependence of the refolding time constant of the second phase measured by FL intensities is shown in Figure 6.2d.

From these preliminary observations of the microsecond refolding kinetics of NT-NtrC, it is apparent that i) a non-native packing of at least one of the tryptophan residues occurs within 30 μs after refolding is initiated, ii) a subsequent slower (~ 200 μs) refolding phase, leads to the formation of a kinetic intermediate with the tryptophan(s) in the same mispacked conformation, and iii) by visual inspection, the midpoint of transition of the burst-phase intermediate reported in Chapter III, corresponds to the urea concentration at the maximum relaxation time (Fig. 3.3). Thus it will now be possible to improve the global analysis of the folding free-energy of NT-NtrC by providing better constraints on the refolding rate constants and denaturant dependence of the burst-phase intermediate.

Figure 6.3 – pWT-CheY. a) Equilibrium denaturation of pWT-CheY, monitored by SAXS (•). The error bars represent the errors of the fit from each SAXS profile. The fit to a two-state model is represented as a solid line. The radius of gyration measured for the first snap-shot from refolding kinetics (20 s) at different final urea concentrations is shown (○). b) The same data monitored by circular dichroism. c) CD spectrum (black line) of the species populated after 150 μs of refolding (0.8 M urea, 10 mM KPi, pH 7.0) is overlaid with the CD-spectra of native (blue line), unfolded (red line) and I_{BP} species (magenta circles) of WT-CheY (Chapter II).

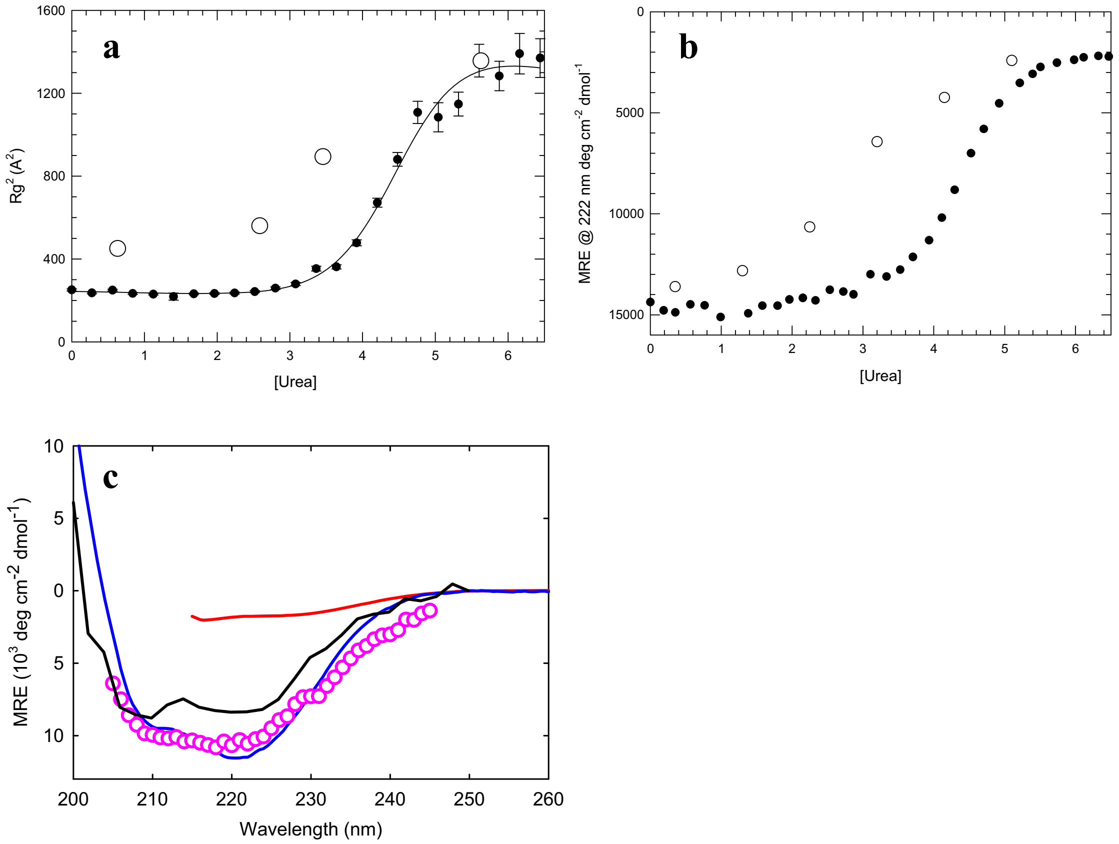

a
Rg^2 (A^2)
1600
1200
800
400
0
0
1
2
3
4
5
6
[Urea]
b
MRE @ 222 nm deg cm^{-2} dmol^{-1}
0
5000
10000
15000
0
1
2
3
4
5
6
[Urea]
c
MRE (10^3 deg cm^{-2} dmol^{-1})
10
5
0
5
10
200
210
220
230
240
250
260
Wavelength (nm)

Figure 6.3

pWT-CheY: Preliminary equilibrium refolding and unfolding studies of pWT-CheY monitored by CD and SAXS suggest a two-state transition (Fig. 6.3a and 6.3b) (ΔG   6.65 ± 0.41 kcal mol^{1}, m-value   1.56 ± 0.1 kcal mol^{1} M^{1}). Manual mixing refolding kinetics reveal a burst-phase intermediate that has a large CD signal, similar to that observed in the WT-CheY (Fig. 6.3a). This intermediate also appears to be very compact as observed by SAXS (Fig. 6.3b). The CD spectrum recorded of the species formed after 150 μs of refolding (Fig. 6.3c) demonstrates that a considerable amount of secondary structure is formed early in the refolding reaction. Whether or not the compaction of the protein is concomitant with the development of secondary structure and the sequence of events involved in the formation of the stopped-flow burst-phase intermediate[276] is yet to be determined.

The BASiC hypothesis

In Chapter IV, the correlation of ILV clusters with high energy intermediates mapped out by native state HX experiments is demonstrated. While the BASiC method is highly sensitive to the location of the protected core of the proteins, it is unable to specifically identify the residues in the region that will be protected. Improvement upon the methodology may be achieved by including simple filters, such as distance from the solvent, hydrogen bonding potential of the residue, etc.

Comparison of the TSEs mapped by phi-analysis and the location of ILV clusters

Another potential application of the BASiC hypothesis is the prediction TSE structures identified by phi-analyses.[53,96,232] CI-2 is a small, 83-residue protein with an α-

Figure 6.4 – Cartoon representation of the crystal structure of CI-2. (2CI2). a) The residues with high phi-values are distributed into two groups; the major core is shown as blue spheres and the minor core as red spheres. b) The side chains involved in the ILV cluster of CI-2 are shown as grey spheres. 8 of the 11 side chains are part of the major core.

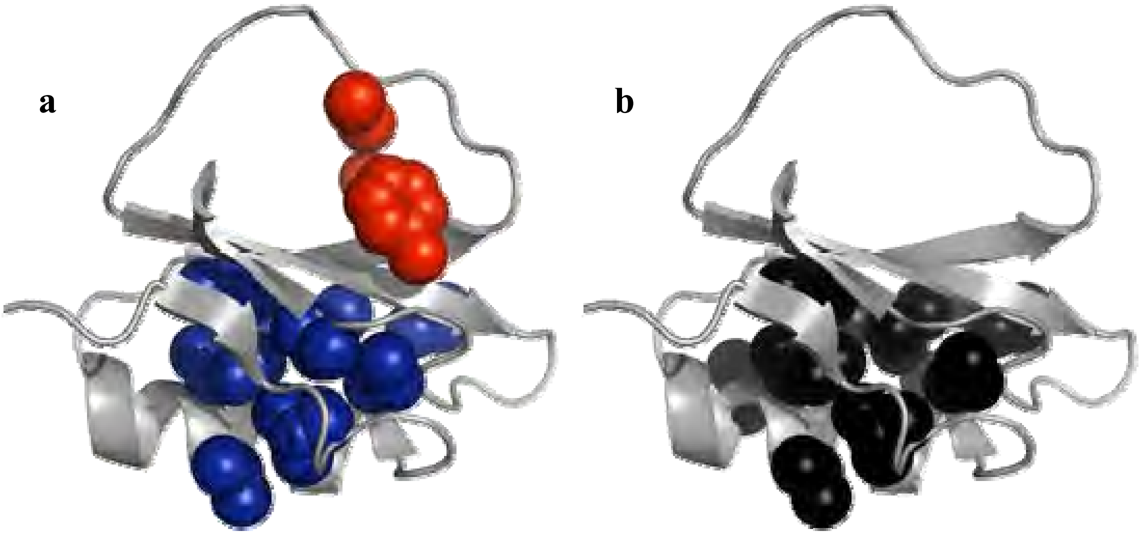


Figure 6.4

helix nestled against a mixed parallel/anti-parallel β-sheet. The active-site loop is located on the opposite face of the β-sheet and links β3 and β4. Mutational analysis showed that the principal stabilizing elements of the TSE are located in the helix and β-strands 3 and 4 on which it docks. ILV analysis of CI-2 reveals an 11 residue cluster (Fig. 6.4) that connects the α-helix (V13, V19, I20 and L21) with β3 (I29 and V31), β4 (V47, L49 and V51) and to a limited extent, β5 (I57). With the exception of I57, the phi values for 7 of the 10 remaining ILVs exceed 0.25, two were not measured and one, V19, was -0.26 (indicating non-native interactions in the TSE). A number of other side chains have non-zero phi values; however, their magnitude tends to decrease with distance from the helix/β3β4 core of structure. Noteworthy among these other side chains are A16, which is buried in the large ILV cluster, and I30, l32 and V38, which form a small cluster on the opposing face of the β-sheet from the helix. The phi value of A16G, 1.06, is the largest observed in CI-2 and indicates a very native-like packing at this position in the α-helix. The phi values for I30 and L32 in β3, 0.31 and 0.19, and V38 in the active-site loop, 0.12, are at the lower end of the spectrum for those in the larger ILV cluster. Perhaps the large ILV cluster in CI-2 can recruit the smaller cluster through the β-sheet, via their common backbone in the β3 segment (I29, I30, V31 and I32). One other small ILV cluster, containing V9 and V60, does not form in the TSE.

Future Directions

A rigorous statistical comparison of the ILV clusters of proteins with the structures of their TSEs, as determined by phi-analyses, must be performed to test the validity of this application. If the structure of the TSE is related to the location of the ILV cluster, it could be hypothesized that a rate-limiting factor in the formation of the native state includes proper packing of the most complex ILV cluster of the protein. The complexity of this cluster could therefore be correlated with the refolding rates of proteins *vis-à-vis* the contact order. As mentioned previously, a single parameter, based on the native state structure, is unlikely to capture the complex problem of protein folding. However, it is tempting to envisage a parameter that will be able to identify the location of TSEs of proteins and predict their folding rates based on physical-chemical phenomena that comprise the essence of the protein folding problem.

Non-local hydrogen bonds

It is intriguing that only a subset of the non-local main chain-side chain H-bonds makes substantial contributions to the stability of the resident proteins. Structurally, a heavy atom H-bond distance of 2.8 Å seems to be sufficient to differentiate between the potent and non-potent H-bonds. However, as evolutionary signatures of the βαβ modules that are the building blocks of TIM barrel proteins, not only should all such bonds be potent, it might be expected that they would appear in the early stages of folding rather than at the last step, defining the native state. In Chapter V, it is supposed that the clamp interactions may have been important stability and structure determinants in the early progenitors of the TIM barrel architecture. However, with the development of the complete $(\beta\alpha)_8$ TIM barrel, other sources of stability, such as large hydrophobic clusters, became more important to the resident proteins. Clamps were probably retained in some proteins for the marginal stability they offer, but may no longer be essential for defining the structure of the protein. In cases where the ILV clusters can only form part of the structure,[143] however, these H-bonds can still play an important role in determining protein stability and structure.

Future directions

An enlarged experimental dataset including a more diverse set of TIM barrel proteins is essential for identifying all of the factors involved in determining the potency of clamp interactions. Further, the role of these evolutionary signatures in other topologies derived from the same βαβ modules, e.g. flavodoxin fold proteins, remains to be determined. Finally, the prospect of using these patterns in structure prediction algorithms and the rational design of proteins will possibly be a step towards the development of a protein folding code.

Broader impact

The role of sequence in determining the folding mechanisms of proteins has been highlighted by several studies,[46,57 59,277] besides the work presented in this thesis. While the field is still a long way from cracking the folding code, specific sequence determinants that can modulate the folding free energy landscape can be readily identified[59] (and Chapters II III IV and V). Identification of stability cores of proteins based on the native structures of proteins can be useful in rational design of proteins. For example, i) the specific packing of ILV residues in the most stable repeat of an ankyrin repeat protein was used to improve the packing in other repeat elements within the protein to create a hyper stable protein,[278] ii) the underlying architectural features of the TIM barrel topology may make it possible to engineer clamp interactions at specific positions in a protein to enhance its stability. Further, as has been discussed earlier, sequence determinants of protein structure, such as the clamp interactions, which have been shown to be evolutionary signatures of the βαβ building block, can be useful in refining the predicted structures of βα-repeat proteins (Chapter V).

From a fundamental perspective, the data from the fast folding kinetics addresses the following issues: i) does chain collapse occur prior to secondary structure development or are the two concomitant? ii) are there specific events in the folding of the intermediate or does it form via multiple independent pathways? iii) is there a conformational bias towards the native state in the unfolded state? These questions, raised earlier in Chapter I, are the key to understanding folding free energy landscapes.

The propensity of large proteins with simple repeat topologies, such as the TIM barrel and CheY-like proteins, to form off-pathway intermediates can have serious consequences for cellular homeostasis. The probability of rapidly populating misfolded structures that are capable of associating with other proteins via non-native interactions can possibly result in aggregation and cause diseases. It may be for this reason that TIM barrel proteins represent the most common motif associated with chaperone proteins in E.coli[279] Sequence and structural features associated with these intermediates, might also provide insights into misfolded, aggregation-prone species in other protein sequences. The recognition of these patterns may enhance our understanding of the molecular determinants of misfolding diseases. [280]

Bibliography

1. Anfinsen, C. B. & Scheraga, H. A. (1975). Experimental and theoretical aspects of protein folding. *Adv Protein Chem* **29**, 205-300.
2. Leventha.C. (1968). Are There Pathways for Protein Folding. *Journal De Chimie Physique Et De Physico-Chimie Biologique* **65**, 44-&.
3. Onuchic, J. N. & Wolynes, P. G. (2004). Theory of protein folding. *Current Opinion In Structural Biology* **14**, 70-75.
4. Gribskov, M., Mclachlan, A. D. & Eisenberg, D. (1987). Profile Analysis - Detection of Distantly Related Proteins. *Proceedings of the National Academy of Sciences of the United States of America* **84**, 4355-4358.
5. Kuhlman, B. & Baker, D. (2000). Native protein sequences are close to optimal for their structures. *Proc Natl Acad Sci U S A* **97**, 10383-8.
6. Rychlewski, L., Zhang, B. H. & Godzik, A. (1998). Fold and function predictions for Mycoplasma genitalium proteins. *Folding & Design* **3**, 229-238.
7. Bell, C. B., Calhoun, J. R., Bobyr, E., Wei, P. P., Hedman, B., Hodgson, K. O., DeGrado, W. F. & Solomon, E. T. (2009). Spectroscopic Definition of the Biferrous and Biferric Sites in de Novo Designed Four-Helix Bundle DFsc Peptides: Implications for O-2 Reactivity of Binuclear Non-Heme Iron Enzymes. *Biochemistry* **48**, 59-73.
8. Allen, B. D. & Mayo, S. L. (2010). An Efficient Algorithm for Multistate Protein Design Based on FASTER. *Journal of Computational Chemistry* **31**, 904-916.
9. Jaroszewski, L., Rychlewski, L., Zhang, B. H. & Godzik, A. (1998). Fold prediction by a hierarchy of sequence, threading, and modeling methods. *Protein Science* **7**, 1431-1440.
10. Chivian, D., Kim, D. E., Malmstrom, L., Schonbrun, J., Rohl, C. A. & Baker, D. (2005). Prediction of CASP6 structures using automated Robetta protocols. *Proteins* **61 Suppl 7**, 157-66.
11. Wu, S., Skolnick, J. & Zhang, Y. (2007). Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Biol* **5**, 17.
12. Kawakami, T., Murakami, H. & Suga, H. (2008). Messenger RNA-Programmed incorporation of multiple N-methyl-amino acids into linear and cyclic peptides. *Chemistry & Biology* **15**, 32-42.
13. Gellman, S. (2009). Structure and Function in Peptidic Foldamers. *Biopolymers* **92**, 293-293.
14. Nauli, S., Kuhlman, B. & Baker, D. (2001). Computer-based redesign of a protein folding pathway. *Nat Struct Biol* **8**, 602-5.
15. Alexander, P. A., He, Y. A., Chen, Y. H., Orban, J. & Bryan, P. N. (2009). A minimal sequence code for switching protein structure and function. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 21149-21154.

16. Zagrovic, B., Snow, C. D., Shirts, M. R. & Pande, V. S. (2002). Simulation of folding of a small alpha-helical protein in atomistic detail using worldwide-distributed computing. *Journal of Molecular Biology* **323**, 927-937.

17. Oldziej, S., Czaplewski, C., Liwo, A., Chinchio, M., Nanias, M., Vila, J. A., Khalili, M., Arnautova, Y. A., Jagielska, A., Makowski, M., Schafroth, H. D., Kazmierkiewicz, R., Ripoll, D. R., Pillardy, J., Saunders, J. A., Kang, Y. K., Gibson, K. D. & Scheraga, H. A. (2005). Physics-based protein-structure prediction using a hierarchical protocol based on the UNRES force field: Assessment in two blind tests. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 7547-7552.

18. Bowie, J. U. & Eisenberg, D. (1994). An Evolutionary Approach to Folding Small Alpha-Helical Proteins That Uses Sequence Information and an Empirical Guiding Fitness Function. *Proceedings of the National Academy of Sciences of the United States of America* **91**, 4436-4440.

19. Bradley, P., Misura, K. M. S. & Baker, D. (2005). Toward high-resolution de novo structure prediction for small proteins. *Science* **309**, 1868-1871.

20. Tomasic, I. B., Metcalf, M. C., Guce, A. I., Clark, N. E. & Garman, S. C. Interconversion of the specificities of human lysosomal enzymes associated with Fabry and Schindler diseases. *J Biol Chem*.

21. Wiseman, R. L., Powers, E. T., Buxbaum, J. N., Kelly, J. W. & Balch, W. E. (2007). An adaptable standard for protein export from the endoplasmic reticulum. *Cell* **131**, 809-21.

22. Johnson, S. M., Wiseman, R. L., Sekijima, Y., Green, N. S., Adamski-Werner, S. L. & Kelly, J. W. (2005). Native state kinetic stabilization as a strategy to ameliorate protein misfolding diseases: a focus on the transthyretin amyloidoses. *Acc Chem Res* **38**, 911-21.

23. Radford, S. E. (2000). Protein folding: progress made and promises ahead. *Trends Biochem Sci* **25**, 611-8.

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

25. Karplus, M. & Weaver, D. L. (1994). Protein folding dynamics: the diffusion-collision model and experimental data. *Protein Sci* **3**, 650-68.

26. Itzhaki, L. S., Otzen, D. E. & Fersht, A. R. (1995). The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J Mol Biol* **254**, 260-88.

27. Bhuyan, A. K. & Udgaonkar, J. B. (1998). Two structural subdomains of barstar detected by rapid mixing NMR measurement of amide hydrogen exchange. *Proteins* **30**, 295-308.

28. Kuwajima, K. & Sugai, S. (1978). Equilibrium and kinetics of the thermal unfolding of alpha-lactalbumin. The relation to its folding mechanism. *Biophys Chem* **8**, 247-54.

29. Fersht, A. R. (1997). Nucleation mechanisms in protein folding. *Curr Opin Struct Biol* **7**, 3-9.

30. Dill, K. A., Ozkan, S. B., Shell, M. S. & Weikl, T. R. (2008). The protein folding problem. *Annu Rev Biophys* **37**, 289-316.

31. Ptitsyn, O. B. (1995). Molten globule and protein folding. *Adv Protein Chem* **47**, 83-229.

32. Dill, K. A. & Chan, H. S. (1997). From Levinthal to pathways to funnels. *Nat Struct Biol* **4**, 10-9.

33. Capaldi, A. P., Kleanthous, C. & Radford, S. E. (2002). Im7 folding mechanism: misfolding on a path to the native state. *Nat Struct Biol* **9**, 209-16.

34. Bilsel, O., Zitzewitz, J. A., Bowers, K. E. & Matthews, C. R. (1999). 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.

35. Fernandez-Recio, J., Genzor, C. G. & Sancho, J. (2001). Apoflavodoxin folding mechanism: An alpha/beta protein with an essentially off-pathway lntermediate. *Biochemistry* **40**, 15234-15245.

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

37. Kathuria, S. V., Day, I. J., Wallace, L. A. & Matthews, C. R. (2008). Kinetic traps in the folding of beta/alpha-repeat proteins: CheY initially misfolds before accessing the native conformation. *Journal Of Molecular Biology* **382**, 467-484.

38. Hills, R. D., Jr., Kathuria, S. V., Wallace, L. A., Day, I. J., Brooks, C. L., 3rd & Matthews, C. R. 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.

39. Bollen, Y. J. M., Sanchez, I. E. & van Mierlo, C. P. M. (2004). Formation of on- and off-pathway intermediates in the folding kinetics of Azotobacter vinelandii apoflavodoxin. *Biochemistry* **43**, 10475-10489.

40. Plaxco, K. W., Simons, K. T. & Baker, D. (1998). Contact order, transition state placement and the refolding rates of single domain proteins. *Journal Of Molecular Biology* **277**, 985-994.

41. Ivankov, D. N., Garbuzynskiy, S. O., Alm, E., Plaxco, K. W., Baker, D. & Finkelstein, A. V. (2003). Contact order revisited: Influence of protein size on the folding rate. *Protein Science* **12**, 2057-2062.

42. Istomin, A. Y., Jacobs, D. J. & Livesay, D. R. (2007). On the role of structural class of a protein with two-state folding kinetics in determining correlations between its size, topology, and folding rate. *Protein Sci* **16**, 2564-9.

43. Gromiha, M. M. & Selvaraj, S. (2001). Comparison between long-range interactions and contact order in determining the folding rate of two-state

proteins: application of long-range order to folding rate prediction. *J Mol Biol* **310**, 27-32.

44. Kamagata, K., Arai, M. & Kuwajima, K. (2004). Unification of the folding mechanisms of non-two-state and two-state proteins. *J Mol Biol* **339**, 951-65.

45. Kamagata, K. & Kuwajima, K. (2006). Surprisingly high correlation between early and late stages in non-two-state protein folding. *J Mol Biol* **357**, 1647-54.

46. Shea, J. E., Onuchic, J. N. & Brooks, C. L., III. (1999). Exploring the origins of topological frustration: Design of a minimally frustrated model of fragment B of protein A. *Proceedings Of The National Academy Of Sciences Of The United States Of America* **96**, 12512-12517.

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

48. Dobson, C. M. (2003). Protein folding and misfolding. *Nature* **426**, 884-90.

49. Bolen, D. W. & Rose, G. D. (2008). Structure and energetics of the hydrogen-bonded backbone in protein folding. *Annu Rev Biochem* **77**, 339-62.

50. Kauzmann, W. (1959). Some factors in the interpretation of protein denaturation. *Adv Protein Chem* **14**, 1-63.

51. Lesser, G. J. & Rose, G. D. (1990). Hydrophobicity of amino acid subgroups in proteins. *Proteins* **8**, 6-13.

52. Tanford, C. (1969). Extension of the theory of linked functions to incorporate the effects of protein hydration. *J Mol Biol* **39**, 539-44.

53. Fersht, A. R., Matouschek, A. & Serrano, L. (1992). The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. *J Mol Biol* **224**, 771-782.

54. Roder, H., Elove, G. A. & Englander, S. W. (1988). Structural Characterization Of Folding Intermediates In Cytochrome-C By H-Exchange Labeling And Proton Nmr. *Nature* **335**, 700-704.

55. Wang, Q. W., Kline, A. D. & Wuthrich, K. (1987). Amide proton exchange in the alpha-amylase polypeptide inhibitor Tendamistat studied by two-dimensional 1H nuclear magnetic resonance. *Biochemistry* **26**, 6488-93.

56. Kim, K. S., Fuchs, J. A. & Woodward, C. K. (1993). Hydrogen exchange identifies native-state motional domains important in protein folding. *Biochemistry* **32**, 9600-8.

57. Scott, K. A., Batey, S., Hooton, K. A. & Clarke, J. (2004). The folding of spectrin domains I: wild-type domains have the same stability but very different kinetic properties. *J Mol Biol* **344**, 195-205.

58. Gu, Z., Zitzewitz, J. A. & Matthews, C. R. (2007). 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.

59. Wu, Y., Vadrevu, R., Kathuria, S., Yang, X. & Matthews, C. R. (2007). 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.

60. Fersht, A. R. & Sato, S. (2004). Phi-Value analysis and the nature of protein-folding transition states. *Proceedings Of The National Academy Of Sciences Of The United States Of America* **101**, 7976-7981.
61. Hocker, B., Beismann-Driemeyer, S., Hettwer, S., Lustig, A. & Sterner, R. (2001). Dissection of a (betaalpha)8-barrel enzyme into two folded halves. *Nat Struct Biol* **8**, 32-6.
62. Hills, R. D., Jr. & Brooks, C. L., III. (2008). Subdomain competition, cooperativity, and topological frustration in the folding of CheY. *Journal Of Molecular Biology* **382**, 485-495.
63. Baldwin, R. L. (1996). On-pathway versus off-pathway folding intermediates. *Fold Des* **1**, R1-8.
64. Lindorff-Larsen, K., Vendruscolo, M., Paci, E. & Dobson, C. M. (2004). Transition states for protein folding have native topologies despite high structural variability. *Nat Struct Mol Biol* **11**, 443-9.
65. Neri, D., Billeter, M., Wider, G. & Wuthrich, K. (1992). NMR determination of residual structure in a urea-denatured protein, the 434-repressor. *Science* **257**, 1559-63.
66. Mohana-Borges, R., Goto, N. K., Kroon, G. J., Dyson, H. J. & Wright, P. E. (2004). Structural characterization of unfolded states of apomyoglobin using residual dipolar couplings. *J Mol Biol* **340**, 1131-42.
67. Smith, L. J., Fiebig, K. M., Schwalbe, H. & Dobson, C. M. (1996). The concept of a random coil. Residual structure in peptides and denatured proteins. *Fold Des* **1**, R95-106.
68. Dyer, R. B., Maness, S. J., Franzen, S., Fesinmeyer, R. M., Olsen, K. A. & Andersen, N. H. (2005). Hairpin folding dynamics: the cold-denatured state is predisposed for rapid refolding. *Biochemistry* **44**, 10406-15.
69. de Alba, E., Jimenez, M. A., Rico, M. & Nieto, J. L. (1996). Conformational investigation of designed short linear peptides able to fold into beta-hairpin structures in aqueous solution. *Fold Des* **1**, 133-44.
70. Peng, Z. Y. & Wu, L. C. (2000). Autonomous protein folding units. *Adv Protein Chem* **53**, 1-47.
71. Brooks, C. L., 3rd. (2002). Protein and peptide folding explored with molecular simulations. *Acc Chem Res* **35**, 447-54.
72. Bai, Y. (1999). Kinetic evidence for an on-pathway intermediate in the folding of cytochrome c. *Proc Natl Acad Sci U S A* **96**, 477-80.
73. Nishimura, C., Dyson, H. J. & Wright, P. E. (2006). Identification of native and non-native structure in kinetic folding intermediates of apomyoglobin. *J Mol Biol* **355**, 139-56.
74. Forsyth, W. R., Bilsel, O., Gu, Z. & Matthews, C. R. (2007). Topology and sequence in the folding of a TIM barrel protein: global analysis highlights partitioning between transient off-pathway and stable on-pathway folding

intermediates in the complex folding mechanism of a (betaalpha)8 barrel of unknown function from B. subtilis. *J Mol Biol* **372**, 236-53.

75. Forsyth, W. R. & Matthews, C. R. (2002). Folding mechanism of indole-3-glycerol phosphate synthase from Sulfolobus solfataricus: a test of the conservation of folding mechanisms hypothesis in (beta(alpha))(8) barrels. *J Mol Biol* **320**, 1119-33.
76. Munoz, V., Lopez, E. M., Jager, M. & Serrano, L. (1994). Kinetic characterization of the chemotactic protein from Escherichia coli, CheY. Kinetic analysis of the inverse hydrophobic effect. *Biochemistry* **33**, 5858-66.
77. Clementi, C., Nymeyer, H. & Onuchic, J. N. (2000). Topological and energetic factors: What determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. *Journal Of Molecular Biology* **298**, 937-953.
78. Lopez-Hernandez, E., Cronet, P., Serrano, L. & Munoz, V. (1997). Folding kinetics of Che Y mutants with enhanced native alpha-helix propensities. *Journal Of Molecular Biology* **266**, 610-620.
79. Roder, H., Maki, K. & Cheng, H. (2006). Early events in protein folding explored by rapid mixing methods. *Chemical Reviews* **106**, 1836-1861.
80. Hocker, B., Schmidt, S. & Sterner, R. (2002). A common evolutionary origin of two elementary enzyme folds. *FEBS Lett* **510**, 133-5.
81. Stock, A. M., Robinson, V. L. & Goudreau, P. N. (2000). Two-component signal transduction. *Annu Rev Biochem* **69**, 183-215.
82. Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., Studholme, D. J., Yeats, C. & Eddy, S. R. (2004). The Pfam protein families database. *Nucleic Acids Res* **32**, D138-41.
83. Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J., Chothia, C. & Murzin, A. G. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res* **32**, D226-9.
84. Lo Conte, L., Brenner, S. E., Hubbard, T. J., Chothia, C. & Murzin, A. G. (2002). SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res* **30**, 264-7.
85. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* **247**, 536-40.
86. Orengo, C. A. & Thornton, J. M. (2005). Protein families and their evolution-a structural perspective. *Annu Rev Biochem* **74**, 867-900.
87. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). CATH--a hierarchic classification of protein domain structures. *Structure* **5**, 1093-108.
88. Filimonov, V. V., Prieto, J., Martinez, J. C., Bruix, M., Mateo, P. L. & Serrano, L. (1993). Thermodynamic analysis of the chemotactic protein from Escherichia coli, CheY. *Biochemistry* **32**, 12906-21.

89. Lopez-Hernandez, E. & Serrano, L. (1995). 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.

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

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

92. Brandts, J. F., Halvorson, H. R. & Brennan, M. (1975). Consideration of the Possibility that the slow step in protein denaturation reactions is due to cis-trans isomerism of proline residues. *Biochemistry* **14**, 4953-63.

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

94. Berova, N., Nakanishi, K. & Woody, R. W. (2000). *Circular Dichroism: Principles and Applications*. 2nd edit, Wiley & Sons Inc, New York.

95. Reimer, U., Scherer, G., Drewello, M., Kruber, S., Schutkowski, M. & Fischer, G. (1998). Side-chain effects on peptidyl-prolyl cis/trans isomerisation. *J Mol Biol* **279**, 449-60.

96. Matthews, C. R. (1987). Effect of point mutations on the folding of globular proteins. *Methods Enzymol* **154**, 498-511.

97. Marquardt, D. W. (1963). An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *Journal of the Society for Industrial and Applied Mathematics* **11**, 431-441.

98. Jackson, S. E. (1998). How do small single-domain proteins fold? *Fold Des* **3**, R81-91.

99. Finkelstein, A. V. & Shakhnovich, E. I. (1989). Theory of cooperative transitions in protein molecules. II. Phase diagram for a protein molecule in solution. *Biopolymers* **28**, 1681-94.

100. Shakhnovich, E. I. & Finkelstein, A. V. (1989). Theory of cooperative transitions in protein molecules. I. Why denaturation of globular protein is a first-order phase transition. *Biopolymers* **28**, 1667-80.

101. Branden, C. & Tooze, J. (1999). *Introduction to Protein Structure*. 2 edit, Garland Science Publishing, New York.

102. Makhatadze, G. I. & Privalov, P. L. (1995). Energetics of protein structure. *Adv Protein Chem* **47**, 307-425.

103. Radzicka, A. & Wolfenden, R. (1988). Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *Biochemistry* **27**, 1664-1670.

104. Schell, D., Tsai, J., Scholtz, J. M. & Pace, C. N. (2006). Hydrogen bonding increases packing density in the protein interior. *Proteins-Structure Function And Bioinformatics* **63**, 278-82.

105. Chothia, C. (1976). The nature of the accessible and buried surfaces in proteins. *J Mol Biol* **105**, 1-12.

106. Liu, P., Huang, X., Zhou, R. & Berne, B. J. (2005). Observation of a dewetting transition in the collapse of the melittin tetramer. *Nature* **437**, 159-62.

107. Zhou, R., Huang, X., Margulis, C. J. & Berne, B. J. (2004). Hydrophobic collapse in multidomain protein folding. *Science* **305**, 1605-9.

108. Hua, L., Huang, X., Zhou, R. & Berne, B. J. (2006). Dynamics of water confined in the interdomain region of a multidomain protein. *J Phys Chem B* **110**, 3704-11.

109. Hubner, I. A., Deeds, E. J. & Shakhnovich, E. I. (2006). Understanding ensemble protein folding at atomic detail. *Proceedings Of The National Academy Of Sciences Of The United States Of America* **103**, 17747-17752.

110. Yang, J. S., Wallin, S. & Shakhnovich, E. I. (2008). Universality and diversity of folding mechanics for three-helix bundle proteins. *Proceedings Of The National Academy Of Sciences Of The United States Of America* **105**, 895-900.

111. Gill, S. C. & von Hippel, P. H. (1989). Calculation of protein extinction coefficients from amino acid sequence data. *Anal Biochem* **182**, 319-26.

112. John, A. S. (1978). Solvent denaturation. *Biopolymers* **17**, 1305-1322.

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

114. Ionescu, R. M., Smith, V. F., O'Neill, J. C., Jr. & Matthews, C. R. (2000). Multistate equilibrium unfolding of Escherichia coli dihydrofolate reductase: thermodynamic and spectroscopic description of the native, intermediate, and unfolded ensembles. *Biochemistry* **39**, 9540-50.

115. Gualfetti, P. J., Bilsel, O. & Matthews, C. R. (1999). The progressive development of structure and stability during the equilibrium folding of the alpha subunit of tryptophan synthase from Escherichia coli. *Protein Sci* **8**, 1623-35.

116. Hwang, T. L. & Shaka, A. J. (1995). Water Suppression That Works. Excitation Sculpting Using Arbitrary Wave-Forms and Pulsed-Field Gradients. *Journal of Magnetic Resonance, Series A* **112**, 275-279.

117. Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J. & Bax, A. (1995). NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* **6**, 277-93.

118. Atkins, P. & de Paulo, J. (2006). *Physical Chemistry*. 8th edit, W.H. Freeman.

119. Sobolev, V., Sorokine, A., Prilusky, J., Abola, E. E. & Edelman, M. (1999). Automated analysis of interatomic contacts in proteins. *Bioinformatics* **15**, 327-32.

120. Wensley, B. G., Gartner, M., Choo, W. X., Batey, S. & Clarke, J. (2009). Different members of a simple three-helix bundle protein family have very different folding rate constants and fold by different mechanisms. *J Mol Biol* **390**, 1074-85.

121. Friel, C. T., Capaldi, A. P. & Radford, S. E. (2003). Structural analysis of the rate-limiting transition states in the folding of Im7 and Im9: similarities and differences in the folding of homologous proteins. *J Mol Biol* **326**, 293-305.

122. Kern, D., Volkman, B. F., Luginbuhl, P., Nohaile, M. J., Kustu, S. & Wemmer, D. E. (1999). Structure of a transiently phosphorylated switch in bacterial signal transduction. *Nature* **402**, 894-898.

123. Madhusudan, Zapf, J., Whiteley, J. M., Hoch, J. A., Xuong, N. H. & Varughese, K. I. (1996). Crystal structure of a phosphatase-resistant mutant of sporulation response regulator Spo0F from Bacillus subtilis. *Structure* **4**, 679-690.

124. Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J. & Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-8.

125. Wallace, L. A. & Robert Matthews, C. (2002). Highly divergent dihydrofolate reductases conserve complex folding mechanisms. *J Mol Biol* **315**, 193-211.

126. Bollen, Y. J. M. & van Mierlo, C. P. M. (2005). Protein topology affects the appearance of intermediates during the folding of proteins with a flavodoxin-like fold. *Biophysical Chemistry* **114**, 181-189.

127. Lopez-Hernandez, E. & Serrano, L. (1996). Structure of the transition state for folding of the 129 aa protein CheY resembles that of a smaller protein, CI-2. *Folding & Design* **1**, 43-55.

128. Formaneck, M. S., Ma, L. & Cui, Q. (2006). Reconciling the "old" and "new" views of protein allostery: A molecular simulation study of chemotaxis Y protein (CheY). *Proteins-Structure Function And Bioinformatics* **63**, 846-867.

129. Lee, S. Y., Cho, H. S., Pelton, J. G., Yan, D. L., Henderson, R. K., King, D. S., Huang, L. S., Kustu, S., Berry, E. A. & Wemmer, D. E. (2001). Crystal structure of an activated response regulator bound to its target. *Nature Structural Biology* **8**, 52-56.

130. Stock, A. M. & Guhaniyogi, J. (2006). A new perspective on response regulator activation. *Journal Of Bacteriology* **188**, 7328-7330.

131. Nelson, E. D. & Grishin, N. V. (2006). Alternate pathways for folding in the flavodoxin fold family revealed by a nucleation-growth model. *Journal Of Molecular Biology* **358**, 646-653.

132. Sola, M., Lopez-Hernandez, E., Cronet, P., Lacroix, E., Serrano, L., Coll, M. & Parraga, A. (2000). Towards understanding a molecular switch mechanism: Thermodynamic and crystallographic studies of the signal transduction protein CheY. *Journal Of Molecular Biology* **303**, 213-225.

133. Zhu, X. Y., Rebello, J., Matsumura, P. & Volz, K. (1997). Crystal structures of CheY mutants Y106W and T871/Y106W - CheY activation correlates with movement of residue 106. *Journal Of Biological Chemistry* **272**, 5000-5006.

134. De Carlo, S., Chen, B. Y., Hoover, T. R., Kondrashkina, E., Nogales, E. & Nixon, B. T. (2006). The structural basis for regulated assembly and function of the transcriptional activator NtrC. *Genes & Development* **20**, 1485-1495.

135. Hu, X. H. & Wang, Y. M. (2006). Molecular dynamic simulations of the N-terminal receiver domain of NtrC reveal intrinsic conformational flexibility in the inactive state. *Journal Of Biomolecular Structure & Dynamics* **23**, 509-517.

136. Volkman, B. F., Lipson, D., Wemmer, D. E. & Kern, D. (2001). Two-state allosteric behavior in a single-domain signaling protein. *Science* **291**, 2429-2433.

137. Fraser, J. S., Clarkson, M. W., Degnan, S. C., Erion, R., Kern, D. & Alber, T. (2009). Hidden alternative structures of proline isomerase essential for catalysis. *Nature* **462**, 669-73.

138. Dyer, C. M. & Dahlquist, F. W. (2006). Switched or not?: the structure of unphosphorylated CheY bound to the N terminus of FliM. *Journal Of Bacteriology* **188**, 7354-7363.

139. Simonovic, M. & Volz, K. (2001). A distinct meta-active conformation in the 1.1-angstrom resolution structure of wild-type apoCheY. *Journal Of Biological Chemistry* **276**, 28637-28640.

140. Gardino, A. K., Volkman, B. F., Cho, H. S., Lee, S. Y., Wemmer, D. E. & Kern, D. (2003). The NMR solution structure of BeF3--activated Spo0F reveals the conformational switch in a phosphorelay system. *Journal Of Molecular Biology* **331**, 245-254.

141. Madhusudan, Zapf, J., Hoch, J. A., Whiteley, J. M., Xuong, N. H. & Varughese, K. I. (1997). A response regulatory protein with the site of phosphorylation blocked by an arginine interaction: Crystal structure of Spo0F from Bacillus subtilis. *Biochemistry* **36**, 12739-12745.

142. Varughese, K. I., Tsigelny, I. & Zhao, H. Y. (2006). The crystal structure of beryllofluoride Spo0F in complex with the phosphotransferase Spo0B represents a phosphotransfer pretransition state. *Journal Of Bacteriology* **188**, 4970-4977.

143. Yang, X., Vadrevu, R., Wu, Y. & Matthews, C. R. (2007). Long-range side-chain-main-chain interactions play crucial roles in stabilizing the (betaalpha)8 barrel motif of the alpha subunit of tryptophan synthase. *Protein Sci* **16**, 1398-409.

144. Otzen, D. E. & Oliveberg, M. (1999). Salt-induced detour through compact regions of the protein folding landscape. *Proc Natl Acad Sci U S A* **96**, 11746-51.

145. Du, R., Pande, V. S., Grosberg, A. Y., Tanaka, T. & Shakhnovich, E. S. (1998). On the transition coordinate for protein folding. *Journal Of Chemical Physics* **108**, 334-350.

146. Snow, C. D., Rhee, Y. M. & Pande, V. S. (2006). Kinetic definition of protein folding transition state ensembles and reaction coordinates. *Biophysical Journal* **91**, 14-24.

147. Juraszek, J. & Bolhuis, P. G. (2008). Rate constant and reaction coordinate of Trp-cage folding in explicit water. *Biophysical Journal* **95**, 4246-4257.

148. Cho, S. S., Levy, Y. & Wolynes, P. G. (2006). P versus Q: Structural reaction coordinates capture protein folding on smooth landscapes. *Proceedings Of The National Academy Of Sciences Of The United States Of America* **103**, 586-591.

149. Karanicolas, J. & Brooks, C. L., III. (2003). Improved Go-like models demonstrate the robustness of protein folding mechanisms towards non-native interactions. *Journal Of Molecular Biology* **334**, 309-325.

150. Rey-Stolle, M. F., Enciso, M. & Rey, A. (2009). Topology-based models and NMR structures in protein folding simulations. *J Comput Chem* **30**, 1212-9.

151. Prieto, L. & Rey, A. (2008). Simulations of the protein folding process using topology-based models depend on the experimental structure. *Journal Of Chemical Physics* **129**, 115101.

152. Chavez, L. L., Gosavi, S., Jennings, P. A. & Onuchic, J. N. (2006). Multiple routes lead to the native state in the energy landscape of the beta-trefoil family. *Proc Natl Acad Sci U S A* **103**, 10254-8.

153. Gosavi, S., Chavez, L. L., Jennings, P. A. & Onuchic, J. N. (2006). Topological frustration and the folding of interleukin-1 beta. *Journal Of Molecular Biology* **357**, 986-996.

154. Gosavi, S., Whitford, P. C., Jennings, P. A. & Onuchic, J. N. (2008). Extracting function from a beta-trefoil folding motif. *Proc Natl Acad Sci U S A* **105**, 10384-9.

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

156. Jakob, R. P. & Schmid, F. X. (2008). Energetic coupling between native-state prolyl isomerization and conformational protein folding. *J Mol Biol* **377**, 1560-75.

157. Jakob, R. P. & Schmid, F. X. (2009). Molecular determinants of a native-state prolyl isomerization. *J Mol Biol* **doi:10.1016/j.jmb.2009.02.021**.

158. Fowler, S. B. & Clarke, J. (2001). Mapping the folding pathway of an immunoglobulin domain: structural detail from Phi value analysis and movement of the transition state. *Structure* **9**, 355-66.

159. Bollen, Y. J. M., Kamphuis, M. B. & van Mierlo, C. P. M. (2006). The folding energy landscape of apoflavodoxin is rugged: Hydrogen exchange reveals nonproductive misfolded intermediates. *Proceedings Of The National Academy Of Sciences Of The United States Of America* **103**, 4095-4100.

160. Olofsson, M., Hansson, S., Hedberg, L., Logan, D. T. & Oliveberg, M. (2007). Folding of S6 structures with divergent amino acid composition: pathway flexibility within partly overlapping foldons. *J Mol Biol* **365**, 237-48.

161. Lappalainen, I., Hurley, M. G. & Clarke, J. (2008). Plasticity within the obligatory folding nucleus of an immunoglobulin-like domain. *J Mol Biol* **375**, 547-59.

162. Lam, A. R., Borreguero, J. M., Ding, F., Dokholyan, N. V., Buldyrev, S. V., Stanley, H. E. & Shakhnovich, E. (2007). Parallel foldng pathways in the SH3 domain protein. *Journal Of Molecular Biology* **373**, 1348-1360.

163. Kister, A. E., Finkelstein, A. V. & Gelfand, I. M. (2002). Common features in structures and sequences of sandwich-like proteins. *Proc Natl Acad Sci U S A* **99**, 14137-41.

164. Wilson, C. J. & Wittung-Stafshede, P. (2005). Snapshots of a dynamic folding nucleus in zinc-substituted Pseudomonas aeruginosa azurin. *Biochemistry* **44**, 10054-62.

165. Mirny, L. & Shakhnovich, E. (2001). Evolutionary conservation of the folding nucleus. *J Mol Biol* **308**, 123-9.

166. Nagano, N., Orengo, C. A. & Thornton, J. M. (2002). One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J Mol Biol* **321**, 741-65.

167. Nohaile, M., Kern, D., Wemmer, D., Stedman, K. & Kustu, S. (1997). Structural and functional analyses of activating amino acid substitutions in the receiver domain of NtrC: evidence for an activating surface. *J Mol Biol* **273**, 299-316.
168. Zapf, J. W., Hoch, J. A. & Whiteley, J. M. (1996). A phosphotransferase activity of the Bacillus subtilis sporulation protein Spo0F that employs phosphoramidate substrates. *Biochemistry* **35**, 2926-33.
169. Miyazawa, S. & Jernigan, R. L. (1996). Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *Journal Of Molecular Biology* **256**, 623-644.
170. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983). CHARMM - A Program For Macromolecular Energy, Minimization, And Dynamics Calculations. *Journal Of Computational Chemistry* **4**, 187-217.
171. Feig, M., Karanicolas, J. & Brooks, C. L., III. (2004). MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. *Journal Of Molecular Graphics & Modelling* **22**, 377-395.
172. Sugita, Y. & Okamoto, Y. (1999). Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters* **314**, 141-151.
173. Kumar, S., Bouzida, D., Swendsen, R. H., Kollman, P. A. & Rosenberg, J. M. (1992). The Weighted Histogram Analysis Method For Free-Energy Calculations On Biomolecules .1. The Method. *Journal Of Computational Chemistry* **13**, 1011-1021.
174. Gallicchio, E., Andrec, M., Felts, A. K. & Levy, R. M. (2005). Temperature weighted histogram analysis method, replica exchange, and transition paths. *Journal Of Physical Chemistry B* **109**, 6722-6731.
175. Kendrew, J. C. (1961). The three-dimensional structure of a protein molecule. *Sci Am* **205**, 96-110.
176. Southall, N. T., Dill, K. A. & Haymet, A. D. J. (2002). A view of the hydrophobic effect. *Journal Of Physical Chemistry B* **106**, 521-533.
177. Pace, C. N. (2001). Polar group burial contributes more to protein stability than nonpolar group burial. *Biochemistry* **40**, 310-3.
178. Shortle, D. (1992). Mutational studies of protein structures and their stabilities. *Q Rev Biophys* **25**, 205-50.
179. Matouschek, A. & Fersht, A. R. (1991). Protein engineering in analysis of protein folding pathways and stability. *Methods Enzymol* **202**, 82-112.
180. Srivastava, A. K. & Sauer, R. T. (2002). Mutational studies of protein stability and folding of the hyperstable MYL Arc repressor variant. *Biophys Chem* **101-102**, 35-42.
181. Sosnick, T. R., Dothager, R. S. & Krantz, B. A. (2004). Differences in the folding transition state of ubiquitin indicated by phi and psi analyses. *Proc Natl Acad Sci U S A* **101**, 17377-82.
182. Bai, Y. & Englander, S. W. (1996). Future directions in folding: the multi-state nature of protein structure. *Proteins-Structure Function And Bioinformatics* **24**, 145-51.

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

184. Rojsajjakul, T., Wintrode, P., Vadrevu, R., Robert Matthews, C. & Smith, D. L. (2004). Multi-state unfolding of the alpha subunit of tryptophan synthase, a TIM barrel protein: insights into the secondary structure of the stable equilibrium intermediates by hydrogen exchange mass spectrometry. *J Mol Biol* **341**, 241-53.

185. Brockwell, D. J. & Radford, S. E. (2007). Intermediates: ubiquitous species on folding energy landscapes? *Current Opinion In Structural Biology* **17**, 30-37.

186. Karanicolas, J. & Brooks, C. L., III. (2004). Integrating folding kinetics and protein function: Biphasic kinetics and dual binding specificity in a WW domain. *Proceedings Of The National Academy Of Sciences Of The United States Of America* **101**, 3432-3437.

187. Schaeffer, R. D., Fersht, A. & Daggett, V. (2008). Combining experiment and simulation in protein folding: closing the gap for small model systems. *Curr Opin Struct Biol* **18**, 4-9.

188. Rhee, Y. M. & Pande, V. S. (2006). On the role of chemical detail in simulating protein folding kinetics. *Chemical Physics* **323**, 66-77.

189. Vadrevu, R., Wu, Y. & Matthews, C. R. (2008). NMR analysis of partially folded states and persistent structure in the alpha subunit of tryptophan synthase: implications for the equilibrium folding mechanism of a 29-kDa TIM barrel protein. *J Mol Biol* **377**, 294-306.

190. Janin, J. & Chothia, C. (1980). Packing of [alpha]-Helices onto [beta]-Pleated sheets and the anatomy of [alpha]/[beta] proteins. *Journal Of Molecular Biology* **143**, 95-128.

191. Biswas, K. M., DeVido, D. R. & Dorsey, J. G. (2003). Evaluation of methods for measuring amino acid hydrophobicities and interactions. *J Chromatogr A* **1000**, 637-55.

192. Rose, G. D. & Wolfenden, R. (1993). Hydrogen bonding, hydrophobicity, packing, and protein folding. *Annu Rev Biophys Biomol Struct* **22**, 381-415.

193. Kyte, J. & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **157**, 105-32.

194. ten Wolde, P. R. & Chandler, D. (2002). Drying-induced hydrophobic polymer collapse. *Proc Natl Acad Sci U S A* **99**, 6539-43.

195. Hua, L., Huang, X. H., Liu, P., Zhou, R. H. & Berne, B. J. (2007). Nanoscale dewetting transition in protein complex folding. *Journal Of Physical Chemistry B* **111**, 9069-9077.

196. Pace, C. N., Trevino, S., Prabhakaran, E. & Scholtz, J. M. (2004). Protein structure, stability and solubility in water and other solvents. *Philos Trans R Soc Lond B Biol Sci* **359**, 1225-34; discussion 1234-5.

197. Bai, Y. & Englander, S. W. (1994). Hydrogen bond strength and beta-sheet propensities: the role of a side chain blocking effect. *Proteins* **18**, 262-6.

198. Hills, R. D., Jr. & Brooks, C. L., III. (2007). Hydrophobic cooperativity as a mechanism for amyloid nucleation. *Journal Of Molecular Biology* **368**, 894-901.

199. Tsai, J., Gerstein, M. & Levitt, M. (1997). Simulating the minimum core for hydrophobic collapse in globular proteins. *Protein Sci* **6**, 2606-16.

200. Raschke, T. M., Tsai, J. & Levitt, M. (2001). Quantification of the hydrophobic interaction by simulations of the aggregation of small hydrophobic solutes in water. *Proc Natl Acad Sci U S A* **98**, 5965-9.

201. Pedersen, T. G., Sigurskjold, B. W., Andersen, K. V., Kjaer, M., Poulsen, F. M., Dobson, C. M. & Redfield, C. (1991). A nuclear magnetic resonance study of the hydrogen-exchange behaviour of lysozyme in crystals and solution. *J Mol Biol* **218**, 413-26.

202. Schulman, B. A., Redfield, C., Peng, Z. Y., Dobson, C. M. & Kim, P. S. (1995). Different subdomains are most protected from hydrogen exchange in the molten globule and native states of human alpha-lactalbumin. *J Mol Biol* **253**, 651-7.

203. Chu, R., Pei, W., Takei, J. & Bai, Y. (2002). Relationship between the native-state hydrogen exchange and folding pathways of a four-helix bundle protein. *Biochemistry* **41**, 7998-8003.

204. Bai, Y., Karimi, A., Dyson, H. J. & Wright, P. E. (1997). Absence of a stable intermediate on the folding pathway of protein A. *Protein Sci* **6**, 1449-57.

205. Mohan, P. M., Chakraborty, S. & Hosur, R. V. (2009). NMR investigations on residue level unfolding thermodynamics in DLC8 dimer by temperature dependent native state hydrogen exchange. *J Biomol NMR* **44**, 1-11.

206. Nishimura, C., Dyson, H. J. & Wright, P. E. (2008). The kinetic and equilibrium molten globule intermediates of apoleghemoglobin differ in structure. *J Mol Biol* **378**, 715-25.

207. Mukherjee, S., Mohan, P. M., Kuchroo, K. & Chary, K. V. (2007). Energetics of the native energy landscape of a two-domain calcium sensor protein: distinct folding features of the two domains. *Biochemistry* **46**, 9911-9.

208. Chi, Y. H., Kumar, T. K., Chiu, I. M. & Yu, C. (2002). Identification of rare partially unfolded states in equilibrium with the native conformation in an all beta-barrel protein. *J Biol Chem* **277**, 34941-8.

209. Alexandrescu, A. T., Jaravine, V. A., Dames, S. A. & Lamour, F. P. (1999). NMR hydrogen exchange of the OB-fold protein LysN as a function of denaturant: the most conserved elements of structure are the most stable to unfolding. *J Mol Biol* **289**, 1041-54.

210. Hughson, F. M., Wright, P. E. & Baldwin, R. L. (1990). Structural characterization of a partly folded apomyoglobin intermediate. *Science* **249**, 1544-8.

211. Rodriguez, H. M., Robertson, A. D. & Gregoret, L. M. (2002). Native state EX2 and EX1 hydrogen exchange of Escherichia coli CspA, a small beta-sheet protein. *Biochemistry* **41**, 2140-8.

212. Schanda, P., Brutscher, B., Konrat, R. & Tollinger, M. (2008). Folding of the KIX domain: characterization of the equilibrium analog of a folding intermediate using 15N/13C relaxation dispersion and fast 1H/2H amide exchange NMR spectroscopy. *J Mol Biol* **380**, 726-41.

213. Bedard, S., Mayne, L. C., Peterson, R. W., Wand, A. J. & Englander, S. W. (2008). The foldon substructure of staphylococcal nuclease. *J Mol Biol* **376**, 1142-54.

214. Grantcharova, V. P. & Baker, D. (1997). Folding dynamics of the src SH3 domain. *Biochemistry* **36**, 15685-92.

215. Sidhu, N. S. (2004). Exploring the conformational manifold of ubiquitin by native state hydrogen exchange, University of Iowa.

216. Kjellsson, A., Sethson, I. & Jonsson, B. H. (2003). Hydrogen exchange in a large 29 kD protein and characterization of molten globule aggregation by NMR. *Biochemistry* **42**, 363-74.

217. Morozova-Roche, L. A., Arico-Muendel, C. C., Haynie, D. T., Emelyanenko, V. I., Van Dael, H. & Dobson, C. M. (1997). Structural characterisation and comparison of the native and A-states of equine lysozyme. *J Mol Biol* **268**, 903-21.

218. Orban, J., Alexander, P., Bryan, P. & Khare, D. (1995). Assessment of stability differences in the protein G B1 and B2 domains from hydrogen-deuterium exchange: comparison with calorimetric data. *Biochemistry* **34**, 15291-300.

219. Yan, S., Kennedy, S. D. & Koide, S. (2002). Thermodynamic and kinetic exploration of the energy landscape of Borrelia burgdorferi OspA by native-state hydrogen exchange. *J Mol Biol* **323**, 363-75.

220. Freund, C., Gehrig, P., Holak, T. A. & Pluckthun, A. (1997). Comparison of the amide proton exchange behavior of the rapidly formed folding intermediate and the native state of an antibody scFv fragment. *FEBS Lett* **407**, 42-6.

221. Arrington, C. B., Teesch, L. M. & Robertson, A. D. (1999). Defining protein ensembles with native-state NH exchange: kinetics of interconversion and cooperative units from combined NMR and MS analysis. *J Mol Biol* **285**, 1265-75.

222. Yi, Q. & Baker, D. (1996). Direct evidence for a two-state protein unfolding transition from hydrogen-deuterium exchange, mass spectrometry, and NMR. *Protein Sci* **5**, 1060-6.

223. Chamberlain, A. K., Handel, T. M. & Marqusee, S. (1996). Detection of rare partially folded molecules in equilibrium with the native conformation of RNaseH. *Nat Struct Biol* **3**, 782-7.

224. Bhutani, N. & Udgaonkar, J. B. (2003). Folding subdomains of thioredoxin characterized by native-state hydrogen exchange. *Protein Sci* **12**, 1719-31.

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

226. Mayo, S. L. & Baldwin, R. L. (1993). Guanidinium chloride induction of partial unfolding in amide proton exchange in RNase A. *Science* **262**, 873-6.

227. Mullins, L. S., Pace, C. N. & Raushel, F. M. (1997). Conformational stability of ribonuclease T1 determined by hydrogen-deuterium exchange. *Protein Sci* **6**, 1387-95.

228. Rader, A. J. & Bahar, I. (2004). Folding core predictions from network models of proteins. *Polymer* **45**, 659-668.

229. Vertrees, J., Barritt, P., Whitten, S. & Hilser, V. J. (2005). COREX/BEST server: a web browser-based program that calculates regional stability variations within protein structures. *Bioinformatics* **21**, 3318-9.

230. Hespenheide, B. M., Rader, A. J., Thorpe, M. F. & Kuhn, L. A. (2002). Identifying protein folding cores from the evolution of flexible regions during unfolding. *Journal of Molecular Graphics & Modelling* **21**, 195-207.

231. Chen, M. Z., Dousis, A. D., Wu, Y. H., Wittung-Stafshede, P. & Ma, J. P. (2009). Predicting protein folding cores by empirical potential functions. *Archives of Biochemistry and Biophysics* **483**, 16-22.

232. Mallam, A. L. & Jackson, S. E. (2008). Use of protein engineering techniques to elucidate protein folding pathways. *Prog Mol Biol Transl Sci* **84**, 57-113.

233. Wayne, N., Lai, Y., Pullen, L. & Bolon, D. N. Modular control of cross-oligomerization: analysis of superstabilized Hsp90 homodimers in vivo. *J Biol Chem* **285**, 234-41.

234. Chen, W., Lam, S. S., Srinath, H., Jiang, Z., Correia, J. J., Schiffer, C. A., Fitzgerald, K. A., Lin, K. & Royer, W. E., Jr. (2008). Insights into interferon regulatory factor activation from the crystal structure of dimeric IRF5. *Nat Struct Mol Biol* **15**, 1213-20.

235. Paravastu, A. K., Leapman, R. D., Yau, W. M. & Tycko, R. (2008). Molecular structural basis for polymorphism in Alzheimer's beta-amyloid fibrils. *Proc Natl Acad Sci U S A* **105**, 18349-54.

236. Nelson, R., Sawaya, M. R., Balbirnie, M., Madsen, A. O., Riekel, C., Grothe, R. & Eisenberg, D. (2005). Structure of the cross-beta spine of amyloid-like fibrils. *Nature* **435**, 773-8.

237. Wasmer, C., Lange, A., Van Melckebeke, H., Siemer, A. B., Riek, R. & Meier, B. H. (2008). Amyloid fibrils of the HET-s(218-289) prion form a beta solenoid with a triangular hydrophobic core. *Science* **319**, 1523-6.

238. Zitzewitz, J. A., Gualfetti, P. J., Perkons, I. A., Wasta, S. A. & Matthews, C. R. (1999). Identifying the structural boundaries of independent folding domains in the α subunit of tryptophan synthase, a β/α barrel protein. *Protein Sci* **8**, 1200-9.

239. Gerstein, M. (1997). A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. *J Mol Biol* **274**, 562-76.

240. Frenkel, Z. M. & Trifonov, E. N. (2005). Closed loops of TIM barrel protein fold. *J Biomol Struct Dyn* **22**, 643-56.

241. FarzadFard, F., Gharaei, N., Pezeshk, H. & Marashi, S.-A. (2008). [beta]-Sheet capping: Signals that initiate and terminate [beta]-sheet formation. *Journal of Structural Biology* **161**, 101-110.

242. Baker, E. N. & Hubbard, R. E. (1984). Hydrogen bonding in globular proteins. *Prog Biophys Mol Biol* **44**, 97-179.

243. Stickle, D. F., Presta, L. G., Dill, K. A. & Rose, G. D. (1992). Hydrogen bonding in globular proteins. *J Mol Biol* **226**, 1143-59.

244. Presta, L. G. & Rose, G. D. (1988). Helix signals in proteins. *Science* **240**, 1632-41.
245. Aurora, R. & Rose, G. D. (1998). Helix capping. *Protein Sci* **7**, 21-38.
246. Horovitz, A., Serrano, L., Avron, B., Bycroft, M. & Fersht, A. R. (1990). Strength and co-operativity of contributions of surface salt bridges to protein stability. *J Mol Biol* **216**, 1031-44.
247. Serrano, L., Kellis, J. T., Jr., Cann, P., Matouschek, A. & Fersht, A. R. (1992). The folding of an enzyme. II. Substructure of barnase and the contribution of different interactions to protein stability. *J Mol Biol* **224**, 783-804.
248. Myers, J. K. & Pace, C. N. (1996). Hydrogen bonding stabilizes globular proteins. *Biophys J* **71**, 2033-9.
249. Ibarra-Molero, B., Zitzewitz, J. A. & Matthews, C. R. (2004). Salt-bridges can stabilize but do not accelerate the folding of the homodimeric coiled-coil peptide GCN4-p1. *J Mol Biol* **336**, 989-96.
250. Gromiha, M. M., Pujadas, G., Magyar, C., Selvaraj, S. & Simon, I. (2004). Locating the stabilizing residues in (alpha/beta)8 barrel proteins based on hydrophobicity, long-range interactions, and sequence conservation. *Proteins* **55**, 316-29.
251. McDonald, I. K. & Thornton, J. M. (1994). Satisfying hydrogen bonding potential in proteins. *J Mol Biol* **238**, 777-93.
252. Delano, W. L. (2002). *The PyMOL Molecular Graphics System*, DeLano Scientific.
253. Schneider, B., Knochel, T., Darimont, B., Hennig, M., Dietrich, S., Babinger, K., Kirschner, K. & Sterner, R. (2005). Role of the N-terminal extension of the (betaalpha)8-barrel enzyme indole-3-glycerol phosphate synthase for its fold, stability, and catalytic activity. *Biochemistry* **44**, 16405-12.
254. Wilmanns, M., Priestle, J. P., Niermann, T. & Jansonius, J. N. (1992). Three-dimensional structure of the bifunctional enzyme phosphoribosylanthranilate isomerase: indoleglycerolphosphate synthase from Escherichia coli refined at 2.0 A resolution. *J Mol Biol* **223**, 477-507.
255. Matthews, C. R. & Crisanti, M. M. (1981). Urea-induced unfolding of the alpha subunit of tryptophan synthase: evidence for a multistate process. *Biochemistry* **20**, 784-92.
256. Sanchez del Pino, M. M. & Fersht, A. R. (1997). Nonsequential unfolding of the alpha/beta barrel protein indole-3-glycerol-phosphate synthase. *Biochemistry* **36**, 5560-5.
257. Hyde, C. C., Ahmed, S. A., Padlan, E. A., Miles, E. W. & Davies, D. R. (1988). Three-dimensional structure of the tryptophan synthase alpha 2 beta 2 multienzyme complex from Salmonella typhimurium. *J Biol Chem* **263**, 17857-71.
258. Wierenga, R. K. (2001). The TIM-barrel fold: a versatile framework for efficient enzymes. *FEBS Lett* **492**, 193-8.

259. Gao, J., Bosco, D. A., Powers, E. T. & Kelly, J. W. (2009). Localized thermodynamic coupling between hydrogen bonding and microenvironment polarity substantially stabilizes proteins. *Nat Struct Mol Biol* **16**, 684-90.

260. Kunin, V., Chan, B., Sitbon, E., Lithwick, G. & Pietrokovski, S. (2001). Consistency analysis of similarity between multiple alignments: prediction of protein function and fold structure from analysis of local sequence motifs. *J Mol Biol* **307**, 939-49.

261. Rohl, C. A., Strauss, C. E., Chivian, D. & Baker, D. (2004). Modeling structurally variable regions in homologous proteins with rosetta. *Proteins* **55**, 656-77.

262. Brylinski, M. & Skolnick, J. (2009). FINDSITE: a threading-based approach to ligand homology modeling. *PLoS Comput Biol* **5**, e1000405.

263. Sali, A. & Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234**, 779-815.

264. Gerlt, J. A. & Babbitt, P. C. (2009). Enzyme (re)design: lessons from natural evolution and computation. *Curr Opin Chem Biol*.

265. Jurgens, C., Strom, A., Wegener, D., Hettwer, S., Wilmanns, M. & Sterner, R. (2000). Directed evolution of a (beta alpha)8-barrel enzyme to catalyze related reactions in two different metabolic pathways. *Proc Natl Acad Sci U S A* **97**, 9925-30.

266. Gerlt, J. A. & Raushel, F. M. (2003). Evolution of function in (beta/alpha)8-barrel enzymes. *Curr Opin Chem Biol* **7**, 252-64.

267. Wieczorek, S. J., Kalivoda, K. A., Clifton, J. G., Ringe, D., Petsko, G. A. & Gerlt, J. A. (1999). Evolution of enzymatic activities in the enolase superfamily: Identification of a "new" general acid catalyst in the active site of D-galactonate dehydratase from Escherichia coli. *Journal of the American Chemical Society* **121**, 4540-4541.

268. Rothlisberger, D., Khersonsky, O., Wollacott, A. M., Jiang, L., DeChancie, J., Betker, J., Gallaher, J. L., Althoff, E. A., Zanghellini, A., Dym, O., Albeck, S., Houk, K. N., Tawfik, D. S. & Baker, D. (2008). Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190-5.

269. Altamirano, M. M., Blackburn, J. M., Aguayo, C. & Fersht, A. R. (2000). Directed evolution of new catalytic activity using the alpha/beta-barrel scaffold. *Nature* **403**, 617-22.

270. Mizuguchi, K., Deane, C. M., Blundell, T. L. & Overington, J. P. (1998). HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci* **7**, 2469-71.

271. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res* **28**, 235-42.

272. Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577-637.

273. Yates, F. (1934). Contingency Tables Involving Small Numbers and the Ï‡^{2} Test. *Supplement to the Journal of the Royal Statistical Society* **1**, 217-235.

274. Bilsel, O., Kayatekin, C., Wallace, L. A. & Matthews, C. R. (2005). A microchannel solution mixer for studying microsecond protein folding reactions. *Review of Scientific Instruments* **76**, -.

275. Snow, C. D., Qiu, L. L., Du, D. G., Gai, F., Hagen, S. J. & Pande, V. S. (2004). Trp zipper folding kinetics by molecular dynamics and temperature-jump spectroscopy. *Proc Natl Acad Sci USA* **101**, 4077-4082.

276. Arai, M., Kondrashkina, E., Kayatekin, C., Matthews, C. R., Iwakura, M. & Bilsel, O. (2007). Microsecond hydrophobic collapse in the folding of Escherichia coli dihydrofolate reductase, an alpha/beta-type protein. *J Mol Biol* **368**, 219-29.

277. Nakamura, T., Makabe, K., Tomoyori, K., Maki, K., Mukaiyama, A. & Kuwajima, K. Different folding pathways taken by highly homologous proteins, goat alpha-lactalbumin and canine milk lysozyme. *J Mol Biol* **396**, 1361-78.

278. Mosavi, L. K., Minor, D. L., Jr. & Peng, Z. Y. (2002). Consensus-derived structural determinants of the ankyrin repeat motif. *Proc Natl Acad Sci U S A* **99**, 16029-34.

279. Kerner, M. J., Naylor, D. J., Ishihama, Y., Maier, T., Chang, H. C., Stines, A. P., Georgopoulos, C., Frishman, D., Hayer-Hartl, M., Mann, M. & Hartl, F. U. (2005). Proteome-wide analysis of chaperonin-dependent protein folding in Escherichia coli. *Cell* **122**, 209-20.

280. Fernandez-Escamilla, A. M., Rousseau, F., Schymkowitz, J. & Serrano, L. (2004). Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat Biotechnol* **22**, 1302-6.