1-1-2013

Large Protein Folding and Dynamics Studied by Advanced Hydrogen Exchange Methods

Benjamin Thomas Walters
University of Pennsylvania, ben@btwalters.com

Large Protein Folding and Dynamics Studied by Advanced Hydrogen Exchange Methods

Abstract
Protein folding studies over the past 50 years have been largely focused on small proteins (< 200 residues) leading to a dearth of information about large protein folding. Regardless of protein size, research has generally lacked the structural tools with necessary temporal resolution to provide mechanistic insight into the process. This goal requires incisive information on transient kinetic intermediate conformations that describe the folding pathway. In this work special challenges that hinder large protein folding studies are addressed, and advancements to both HX NMR and HX MS experiments are described that provide unparalleled temporal resolution of structure formation than has been previously possible. These various advanced hydrogen exchange methods are used to study folding behaviors of the large, 370-residue, two-domain maltose binding protein from E. coli and provide a description of its folding pathway in structural detail. This work sheds light on two basic unresolved problems regarding the mechanisms of protein folding, the first being the enigmatic nature of the initial folding collapse event seen in many proteins, and the second concerning the nature of the folding pathway. We find that from an initially heterogeneous hydrophobic collapse, an obligatory intermediate emerges with a 7-second time constant followed by an apparent sequential pathway to the native state. These results add the largest protein studied at structural resolution to-date to the list of proteins known to fold through obligatory, native-like intermediates in distinct pathways and this work highlights strategies that may be employed to interrogate other large systems in future work.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Biochemistry & Molecular Biophysics

First Advisor
S. Walter Englander

Second Advisor
Feng Gai

Keywords
Collapse, Denatured State Ensemble DSE, HDX MS, Hydrogen Exchange, Maltose Binding Protein, Protein Folding

Subject Categories
Analytical Chemistry | Biochemistry | Biophysics

LARGE PROTEIN FOLDING AND DYNAMICS STUDIED BY ADVANCED HYDROGEN EXCHANGE METHODS

Benjamin Thomas Walters

A DISSERTATION

in

Biochemistry and Molecular Biophysics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2013

Supervisor of Dissertation

_____

S. Walter Englander, Ph.D.
Jacob Gershon-Cohen Professor of Medical Science
Professor of Biochemistry and Biophysics

Graduate Group Chairperson

_____

Kathryn M. Ferguson, Ph.D.
Associate Professor of Physiology

Dissertation Committee:

Feng Gai, Ph.D., Professor of Chemistry (Chair)
Kim Sharp, Ph.D., Associate Professor of Biochemistry and Molecular Biophysics
Ben E. Black, Ph.D., Associate Professor of Biochemistry and Molecular Biophysics
Jeffery G. Saven, Ph.D., Associate Professor of Chemistry
Bohdana M. Discher, Ph.D., Research Assistant Professor of Biochemistry and Biophysics
Michael B. Goshe, Ph.D., Associate Professor of Biochemistry, North Carolina State University

LARGE PROTEIN FOLDING AND DYNAMICS STUDIED BY ADVANCED HYDROGEN EXCHANGE METHODS

Dedication

To my wife, Lindsey, who has shown me the meaning of the word dedication in her patience and support of me over the past six years.

ACKNOWLEDGMENT

My undergraduate biochemistry professor, Michael B. Goshe, inspired me to pursue this Ph.D. If he had not pulled me aside as a senior at NCSU and suggested that I work in his lab, I would not be here today. He has tremendously influenced my life.

This work would not have been a success without my mentors Walter Englander and Leland Mayne. They taught me how to be a scientist, how to improve my writing for scientific audiences, how to be a skeptic, how to study protein folding, about the wonderful world of HX, and, most importantly, that no data is better than bad data.

Second to them are Josh Wand and John Gledhill. As a first-year graduate student I rotated with Josh Wand's group. His graduate student, John Gledhill, inspired me to learn how to computer program, and patiently answered all of my questions. This skill proved to be one of the most valuable in my time at Penn. I also appreciate the Wand wet lab team who provided me with unpurified MBP for the studies described in this work.

My thesis committee has been an integral part of advising me throughout my time at the University of Pennsylvania. Permanent members include Feng Gai (chair), Kim Sharp, Ben Black, and Jeffrey G. Saven. Michael Ghoshe and Bohdana Discher kindly agreed to externally and internally review this work, respectively. I am grateful for their help and guidance over the past years.

My lab mates: John Skinner, Alec Ricutti, Zhong-Yuan Kan, Wenbing Hu, and Palaniappan Chetty made the experience a pleasure. I enjoyed many helpful discussions and learned many things from them.

The members of Josh Wand's lab who taught me NMR, namely Kathy Valentine and John Gledhill. Additionally, Sabrina Bédard, Nathaniel Nucci, Vonni Moorman, Jacob Dogan, Joseph Kielic, helped when I needed assistance and we had many useful discussions. The Wand wet lab provided needed protein when I had run out, this was wonderful and very nice of them.

We had a fruitful collaboration with the Sosnick lab at the University of Chicago. Their criticisms and assistance collecting SAXS data helped solidify the reality of our findings in Chapter 5. James Henshaw helped me process and understand the SAXS data.

My family has also provided a wonderful support network to help me through the harder times and celebrate the good times. I want to thank my wife, Lindsey Walters; my mom and dad, Diane and Edward Walters; my brother, Kevin Walters; my uncle Robert Gray; and my grandmothers, "Mema" (Marie Gray) and "Granny" (Joanne Walters). I deeply appreciate my wonderful mother in particular for helping me proofread this text.

This work was supported by National Institutes of Health research grants RO1 GM031847 (to S.W.E) and a structural biology pre-doctoral training grant GM08275 (to B.T.W.), and a National Science Foundation research grant MCB1020649 (to S.W.E.). Use of the Advanced Photon Source, an Office of Science User Facility operated for the U.S. Department of Energy (DOE) Office of Science by Argonne National Laboratory, was supported by the U.S. DOE under Contract No. DE-AC02-06CH11357. This project was supported by grants from the National Center for Research Resources (2P41RR008630-18) and the National Institute of General Medical Sciences (9 P41 GM103622-18) from the National Institutes of Health.

ABSTRACT

LARGE PROTEIN FOLDING AND DYNAMICS STUDIED BY ADVANCED HYDROGEN EXCHANGE METHODS

Benjamin Thomas Walters

S. Walter Englander

Protein folding studies over the past 50 years have been largely focused on small proteins (< 200 residues) leading to a dearth of information about large protein folding. Regardless of protein size, research has generally lacked the structural tools with necessary temporal resolution to provide mechanistic insight into the process. This goal requires incisive information on transient kinetic intermediate conformations that describe the folding pathway. In this work special challenges that hinder large protein folding studies are addressed, and advancements to both HX NMR and HX MS experiments are described that provide unparalleled temporal resolution of structure formation than has been previously possible. These various advanced hydrogen exchange methods are used to study folding behaviors of the large, 370-residue, two-domain maltose binding protein from *E. coli* and provide a description of its folding pathway in structural detail. This work sheds light on two basic unresolved problems regarding the mechanisms of protein folding, the first being the enigmatic nature of the initial folding collapse event seen in many proteins, and the second concerning the nature of the folding pathway. We find that from an initially heterogeneous hydrophobic collapse, an obligatory intermediate emerges with a 7-second time constant followed by an apparent sequential pathway to the native state. These results add the largest protein studied at structural resolution to-date to the list of proteins known to fold through obligatory, native-like intermediates in distinct pathways and this work highlights strategies that may be employed to interrogate other large systems in future work.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

AMORE-HX	-	a multidimensional optimization of radially enhanced NMR-based hydrogen exchange
ANS	-	Dye that binds hydrophobic patches and fluoresces upon binding. (8-Anilino-1-naphthalenesulfonic acid)
BMRB	-	Biological and Magnetic Resonance Database
CID	-	Collision induced dissociation
CSD	-	Charge state distribution
DDA	-	Data dependent acquisition
DFT	-	Discrete Fourier transform
D-MBP	-	Fully Deuterated (backbone amides) MBP
ECD	-	Electron capture dissociation
ESI	-	Electro-spray ionization
ETD	-	Electron transfer dissociation
FD	-	Fully deuterated (sample)
FFT	-	Fast Fourier transform
FRET	-	Fluorescence resonance energy transfer
FT	-	Fourier tranform
H-MBP	-	Fully protiated MBP
HNCO	-	3-atom correlated NMR, magetization starts on the amide H, then to Amide N then (-1) Carbonyl carbon and back to the amide H for detection.
HSQC	-	Hetero-nuclear single quantum coherence
H-T	-	Hydrogen-tritium
HX	-	Hydrogen exchange
IUP	-	Independent unrelated pathways
KHX	-	Kinetic pulse-labeling hydrogen exchange
MBP	-	Maltose binding protein, from *E. coli*
MG	-	Molten Globule
MS	-	Mass spectrometry
NHX	-	Native state hydrogen exchange
NMR	-	Nuclear magnetic resonance spectroscopy

PPOE	-	Predetermined pathways & optional errors
R_g	-	Radius of gyration
RLS	-	Rate limiting step
RMS	-	Root mean squared
RnaseH	-	ribonuclease H
SAXS	-	Small angle x-ray scattering
T_1	-	Longitudinal relaxation lifetime
T_2	-	Transverse relaxation lifetime

LIST OF TABLES

LIST OF ILLUSTRATIONS

PREFACE

In the fall of 2008, I joined the Englander lab and embarked on a quest to understand the folding pathway of MBP (maltose binding protein). We were all intrigued that folding took so long. This meant all sorts of little annoyances that are not a concern with faster folders, such as the fact that it may take a full day of 5-10 minute intervals to collect enough replicates to get good data for some trivial optical experiment such as generating the folding arm of a chevron plot. Things like lamp stability, diffusion artifacts in instruments, many hours in hot, loud laboratory closets, tightly packed with big machines causing one to lose focus and make simple mistakes etc… These things cause one to have to repeat that somewhat irritating day many times before getting it right.

Everything is slower and most things more difficult with a slow folding protein; but, I contend, the lessons learned from these efforts are more beneficial to basic science. It is beneficial firstly because we understand much less about large, slow protein folding – small fast folding protein studies have been popular for decades. It is interesting from a biological perspective because we are seeing folding diseases become more and more prevalent as our life expectancies are increasing and we realize that most of our proteome is comprised of large and, very probably, slow folding proteins. Understanding the conformational dynamics of larger proteins becomes medically relevant with this in mind, and the dearth of information on large protein folding is a result primarily of not having the proper technology to study these molecules. Thus, this work is important because the drive to understand the folding pathway of MBP stimulated the development of new methods and analytical techniques, described herein, that may generally be used to interrogate large protein folding and dynamics in ways that have not been possible in the past.

In the first chapter, a general introduction to the long-standing protein folding problem is given along with a discussion of particular aspects of the problem addressed by this work. This is followed by an introduction to maltose binding protein (MBP) including an overview of background knowledge and justification for using MBP to study protein folding in a large molecule. The chapter closes with an overview of this work along with a set of specifically defined questions addressed herein.

The second chapter is concerned with all things hydrogen exchange. Here I hope to provide the reader with background information that is useful for appreciating what follows in this dissertation. First, I present HX theory, and then describe the various experiments one might employ to utilize hydrogen exchange for studying protein folding and structural dynamics.

This dissertation covers work that lead to three first author publications (1-3), and three supporting author publications (4-6). The bulk of my work is contained in Chapters 3, 4, and 5, which are framed largely around these manuscripts. Method development is contained in Chapter 3 (HX NMR) and Chapter 4 (HX MS); followed by the results of these methods as applied to the study of MBP folding in Chapter 5. These three chapters are previewed below. The dissertation closes with Chapter 6 on general conclusions and potential future directions for studies on MBP.

Chapter 3:

As a first year graduate student, I had begun working with MBP and wanted to measure its hydrogen exchange rates. I rotated with a renowned NMR expert, Joshua Wand, before joining the Englander lab and was interested in the capabilities of NMR; using NMR to measure MBP HX seemed the natural choice. Ultimately, NMR was dropped in favor of mass spectrometry but I did some constructive work along the way which resulted in a co-first author paper (reference (1)) where we developed a 3D HX experiment aimed at providing site resolved real time exchange measurements for large proteins. I additionally learned some useful information from NHX experiments on MBP and created tools that specifically addressed difficulties of doing HX NMR on large proteins.

This chapter follows my work with HX NMR chronologically. In the introduction (section 3.1), I discuss the benefits of studying HX by NMR and generally the challenges facing HX NMR experiments on large proteins along with introducing the reader to my solutions for each that will be fully developed in subsequent sections. In section 3.2, HSQC-NMR based HX strategies are discussed in detail along with a presentation of pilot MBP HX data. In section 3.3, my contributions to the 3D NMR HX experiment (the content of publication (1)) are summarized before closing with a few general sentiments regarding my work and the lessons learned by NMR HX on MBP in section 3.4.

Chapter 4:

As my graduate career progressed, it became clear that I would need to go beyond HX NMR to understand the folding pathway of MBP. Mass spectrometry (MS) offered a solution to challenges faced with NMR; I spent the majority of my time as a graduate student working with this instrument. This work was particularly fruitful resulting in a first author publication (2). My contributions to the design of modern HX MS methodology and analysis of HX MS data further led to three supporting author publications (4-6).

Following a brief history of the HX MS experiment (Section 4.1) to place my work in the appropriate context, the second and third sections discuss solutions developed in our laboratory to combat the two major challenges in HX MS experiments. Section 4.2 addresses the problem of low sequence coverage using content from a supporting author publication (4) to explain the modern fragmentation-separation HX MS experiment and highlight how our procedure ameliorates the issue of low coverage. This is followed in section 4.3 by the content of a first-author publication (reference (2)) where the back-exchange problem is discussed and minimized. Section 4.4 departs from published work and is devoted to my data analysis strategy. Here I describe my approach to extract useful information from mass spectrometry data; in doing so, I am able to highlight a major advantage for HX MS as opposed to HX NMR. This work was not published alone but it has proven essential to all of our folding HX experiments in the lab. It has been highlighted in a recent second author publication that uncovered the folding pathway of ribonuclease H (5) and in my third first-author publication on the folding pathway of MBP (Chapter 5).

My contributions to the HX MS experiment cover all aspects of the technique, both experimental design and data analysis – all efforts serve my purpose of understanding the folding pathway of MBP. Though I spent more than half of my time focused on methodology, every detail of this work contributed to the success of the experiment and results in Chapter 5.

Chapter 5

Chapter 5 is the climax of this work where everything comes together for MBP, both literally (we let it fold) and figuratively (we learn how it folds). This data was published in

reference (3). This was a team effort, Walter, Leland, and I spent perhaps hundreds of hours looking at data, testing ideas, tussling over interpretations, and nearly always becoming puzzled and/or intrigued over the various complexities associated with the kinetic pulse labeling HX MS experiment and/or some aspect of the protein folding problem. All of the method development in the previous two chapters is put to work and we are able to contribute substantially towards the scientific community's knowledge of large protein folding events.

In response to our manuscript (reference (3)) Robert Baldwin, a leader in the field, stated, "Considering the second problem [the nature of protein folding pathways] first, the authors have obtained a clearcut result that will have a major influence on thinking about the protein folding problem."

Chapter 1 - Protein Folding & Maltose Binding Protein

1.1 The Protein Folding Problem

Understanding how a polypeptide acquires native structure represents one of the oldest unsolved problems in molecular biology; we are interested in understanding how proteins are able to fold on a biologically relevant time scale. Clearly, proteins do not fold by exhaustively searching their manifold of available conformations; in formative work, Cyrus Levinthal pointed out that such a random search process would require eons to complete (7) and this led researchers to search for folding pathways.

Biochemical pathways are almost universally characterized by isolating intermediates and characterizing their nature; however, this is most difficult in protein folding studies. Most proteins fold in less than a second and their ephemeral intermediate states are nearly impossible to isolate (8). Research has largely focused on whether folding is 2-state/multi-state etc… and on the kinetic features of the process, but these facts do not provide structural insight into folding mechanisms.

Two leading theories that take surprisingly disparate views of the folding process have emerged. Based largely on theory and simulation, one view is that proteins fold by way of independent unrelated pathways (IUP model) and that intermediates are not productive, they slow down the folding process. Alternatively, based largely from experimentation, the other view is that proteins fold by way of predetermined pathways with optional errors (9-11) (PPOE model) and that intermediates, when observed, are the result of misfolding. This view holds that proteins fold upon a conformational scaffold whereby the native structure is progressively built by the addition of foldons – intermediates speed up folding and the sequential nature of their progression sketches out a macroscopic folding pathway. To settle this dispute, structural information on protein folding processes will be required.

Hydrogen exchange (HX) experiments afford the ability to explore the folding process with sufficient depth to make structural conclusions. In this work, we are able to study the folding of maltose binding protein (MBP) at structural resolution. We show that these results support pathway-directed folding and that ideas originating from proponents of the IUP

perspective are useful in interpreting aspects of the data presented herein. In the following discussion, questions relevant to protein folding research that are addressed by our study of MBP are introduced. Specific issues regarding early events in protein folding and the two aforementioned theories, PPOE and IUP, are explored.

1.1.1 The Unfolded State and Protein Collapse

Many studies (e.g. (12-15)) have demonstrated some change in signal (CD, fluorescence, etc…) that occurs completely within instrumental dead times during refolding experiments[1] and these observations are often referred to as burst events or burst-phases. Determining the causative element for burst events has proven to be most difficult because measurements generally lack structural information. Burst signals are, on occasion, interpreted as a reduction in the radius of gyration (R_g) and thus taken to represent chain collapse events (e.g. (16-31)) which are akin to the concept of a coil-globule transition.

It is commonly thought that one of the earliest events in protein folding involves a collapse of the unfolded chain into a more compact configuration; however, it is clear that a random chain collapse, as has been suggested (32), is not ubiquitous in all proteins studied (33, 34). Broad disagreement also exists as to whether random collapse without the formation of structure occurs. To understand the nature of polypeptide collapse, the physics of unfolded conformations and the mechanism of chemical denaturation deserve attention.

What does it mean to say a molecule is “unfolded” by chemical denaturants? Tanford’s classical experiments (35) demonstrated that many proteins exhibit hydrodynamic properties of random coils in high concentrations of chemical denaturant. The term “unfolded” usually implies conditions where random-coil polymer behavior is observed (36) and backbone □,ψ dihedral angles are uncorrelated with one another. *In vitro* protein folding experiments usually employ denaturant concentrations just slightly beyond that required to observe a cooperative transition (37). However, many lines of evidence indicate the unfolded ensemble in some real proteins continues to swell with increasing concentrations of denaturant.

[1] typically < 10 ms, although burst signals have been observed with dead-times approaching microseconds.

Why do solutions with high concentrations of chemical denaturants promote unfolded states of protein molecules? Both in simulation (38) and experiment (39) guanidinium ions are dehydrated along their planar face suggesting that they participate in hydrophobicity-driven stacking interactions. Guanidinium has also been shown to decrease the hydrophobic effect in a very simple (and therefore more reliable) simulation involving only water, guanidinium, and two hydrophobic plates (40). Thus, it appears that the strength of the hydrophobic effect and the concentration of chemical denaturants are inversely related. By solvating hydrophobic residues, chemical denaturants such as urea and guanidinium selectively promote random coil conformations (41).

Theoretically, if a polypeptide has the proper fraction of hydrophobic residues, as proposed by Ken Dill in 1985, a collapsed conformation devoid of specific contacts may be energetically favorable (42), the hydrophobic effect drives this event. More compact conformations should exhibit some internal structure (43), albeit, not necessarily native structure and the extent of structural induction is expected to increase with compactness. Severe steric constraints facilitate conformational selection. This will lead to the formation of intrachain contacts.

The notion that aspecific polypeptide collapse could occur as a result of changing the solvation conditions (42, 44-46) has led many to conclude that most proteins do collapse randomly before regular structure forms and often this causes a burst-phase; however, finding definitive evidence of this is quite difficult. In experiments where single molecule fluorescence studies involving fluorescence resonance energy transfer (FRET) indicate a collapse transition that precedes structure formation, small angle x-ray scattering (SAXS) experiments fail to corroborate these findings (33, 34, 47, 48).

The question remains, do some proteins collapse in absence of structure formation? If so, what effect, if any, does this have on subsequent folding? Do burst-phases indicate regular structure formation like one would expect in a kinetic folding intermediate, or might they reflect a compact unfolded state? Recent reviews of the evidence for and against non-specific polypeptide collapse (48, 49) and a commentary about early burial in protein folding (50) are useful sources where interested readers may further explore the subject.

Perhaps the question of whether structure forms during the collapse is poorly phrased; "structure" has a variety of definitions. A better question might be, "do some proteins collapse in absence of *stable* structure formation?" The issue here is that often "structure formation" is taken to mean, for example, a change in ANS[2] binding or CD_{222} signals. Certainly, these signals are affected by the presence of structural elements but they are ensemble-averaged measurements and allow substantial leeway in interpretation. ANS binding may reflect the formation of a hydrophobic surface, or it could simply indicate random small patches of hydrophobes (51, 52). Changes in CD signals may reflect the presence of helices, but could also be the result of random aromatic sidechain burial (53-56). Proponents of random collapse often discredit changes in signals when convenient, but then cite observations in these signals, such as multi-exponential behavior, to make strong statements about the folding problem.

Jennings and Wright studied the burst-phase in apomyoglobin using hydrogen exchange methods (57) and found that the burst-phase product exhibited substantial protection from exchange in the A, G, and H helices – this led authors to conclude concurrent collapse and structure formation (6.1 ms dead time). More recently, with a reduced dead time (300 μs), Uzawa et al. (58) concluded apomyoglobin collapsed randomly; however, 50% of the change in CD_{222} signal, observed during folding, occurred simultaneously. This was made consistent with random collapse by suggesting that the helix formation, observed earlier and taken to represent concurrent structure formation, was in fact explained by multi-exponential kinetics in the CD_{222} signal – one phase had a lifetime of 5 ms and authors concluded this must represent the formation of helices A, G, and H.

MBP has a burst-phase that appears to persist for hundreds of milliseconds before subsequent changes in fluorescence and circular dichroism signals are observed. This provides an opportunity to explore the burst-phase product in detail. We employ SAXS measurements to determine the radius of gyration following the burst and compute a rough envelope reconstruction to get an estimate for what it might look like. We also assess any structure that might form during the burst, and stability thereof, using HX.

[2] 8-anilino-1-naphthalene sulfonate

We are able to verify that the MBP burst is the result of a nonspecific collapse event and likely a representation of the unfolded state under permissive folding conditions. A few very short regions of chain do appear to develop structure quickly. We cannot distinguish between induced structure resulting from the collapse, and the rapid association of structural elements causing the collapse. One thing is certain, we find no indication these structural elements are related to foldons or the kinetic folding pathway. We also present evidence of low levels of heterogeneous structure throughout the molecule as a product of the burst collapse. This suggests that each molecule likely has formed very many intrachain H-bonds that reorganize rapidly. Perhaps large proteins fold slowly because their conformational search speed is reduced in the collapsed state.

1.1.2 Kinetic Protein Folding Models

Over the many years that researchers have been interested in folding phenomena, two models that seek to describe how a protein is able to find its native conformation within a biologically relevant time scale are most prevalent today.

Predetermined Pathways with Optional Errors (PPOE Model)

The classical model of protein folding by pathways has persisted through the years and the PPOE model (59) formally describes the emergence of macroscopic folding pathways from the basic physical principle of cooperativity and an interaction principle known as sequential stabilization (9, 60, 61). In this model, cooperative elements of structure appear in the order of their relative stabilities, the basic cooperative unit is called a foldon and these are determined by experiment. The interaction principle suggests when a two foldons associate, they are mutually stabilized. By the combination of these two principles, pathways emerge. In the absence of folding errors, folding will be a two-state process with the rate-limiting step being consolidation of the first foldon. Subsequent steps are faster than preceding steps as conformational freedom reduces with each foldon addition. Subsequent foldons coalesce onto the growing native conformational scaffold in the order of their relative stabilities. Pathway branching may occur (62) when subsequent foldons are energetically similar, but this does not change the classical paradigm whereby proteins fold by sequential processes nor does it violate the PPOE hypothesis.

Complicated multi-exponential kinetic signatures arise from optional misfolding errors. Misfolding errors represent non-native intrachain contacts that must be broken for folding to proceed; this can slow a subset of the folding population and would appear indistinguishable from independent folding pathways by optical measurements. Structural characterizations of intermediate conformations are required to distinguish between these two possibilities (59).

The observation of an obligatory intermediate would be strongly suggestive of a folding mechanism where all molecules are channeled through the same conformation *en route* to the native state. *Unrelated* pathways should not have a *common* intermediate, by definition. However, proving that an intermediate is obligatory requires strong evidence that all molecules visit the intermediate. In a favorable case for cytochrome C, an intermediate can be made to accumulate to roughly 85% (63).

Experimental support for the PPOE model may be found largely in HX studies of the folding process, perhaps because HX experiments represent one of the only ways to obtain broad level information regarding time-dependent structure formation. Foldons generally are hidden in kinetic folding experiments; they are not observable for 2-state folding proteins, nonetheless they are intermediates. Evidence for foldons comes primarily from equilibrium hydrogen exchange experiments (see the review (11)) but also from sulfhydryl labeling experiments (64, 65) and theoretical studies (66-68). Recently, we were able to demonstrate that the folding pathway of ribonuclease H, elucidated in structural detail by HX experiments, folds by a PPOE mechanism (5), evidence was not found suggesting any misfolding in this case.

Independent and Unrelated Pathways (IUP/Funnel Model)

Many researchers turned to simple statistical mechanics models and computer simulations in an attempt to circumvent difficulties in experimental protein folding in the early 1990's (e.g. (69)). These efforts were undertaken to explain complex folding kinetics measured by optical spectroscopy and led to a model of the folding process meant to assist interpretation. This model, sometimes referred to as the "New View" (70), de-emphasizes folding pathways and the importance of specific intermediate structures (71-82). It was popularized and described by Dill and Chan in 1997 metaphorically, "…folding is seen as more like the trickle of water down mountainsides of complex shapes, and less like the flow through a single gulley (82)."

The IUP model suggests kinetic partitioning (32) gives rise to ensembles of molecules that fold on characteristically different time scales, leading to the idea of *fast* and *slow* folding tracks. Multi-exponential kinetic signatures observed from spectroscopic experiments are suggested to represent, in some cases, these different tracks. In this way, it is possible that burst-phase signals could represent a fraction of molecules folding to the native state on a so-called fast track. In our state-sensitive HX experiments with MBP, if there are multiple tracks to the native state, they will be seen directly.

Neither the absence of a fast folding track, the observation of foldons, nor lack of evidence for multiple independent pathways would disprove the IUP model. Proponents of IUP would argue that the model does not preclude pathway-directed folding, *macroscopically*. We explore whether this model is useful for understanding MBP data.

The Importance of Kinetic Models

Conclusions reached in many folding experiments often depend on the question being asked; the thoroughly studied hen egg white lysozyme (HEWL) folding intermediate serves as an example. Comprehensive experiments and analysis of kinetic folding/unfolding data and denaturant dependencies for HEWL led authors to conclude (83-87), in every case, that folding proceeds by independent unrelated pathways. In 2007, Englander and Krishna (59), using the same data, were able to show that the model of pre-determined pathways with optional errors fit the same data with comparable or lower χ^2 scores and fewer fitting parameters. PPOE and IUP are directly opposed from one another in spirit; yet, by law, both fit the spectroscopic data nicely.

Without additional information, the parsimonious conclusion is that HEWL folds by a PPOE mechanism – this example demonstrates both that optical spectroscopy is insufficient for the question being considered, and that kinetic models of protein folding do tend to guide how we imagine the process. Spectroscopic studies do assist understanding of folding processes. One can define a rough estimate of the time scale associated with the global folding event and can explore whether a given perturbation slows or enhances the global folding rate. The fact remains, however, that structural folding information is required to properly distinguish IUP from PPOE mechanisms.

We chose to rely on hydrogen exchange for decisions regarding IUP or PPOE. HX will enable us to measure multiple independent pathways if they exist. Likewise, if intermediates are present, HX will provide necessary structural resolution for their characterization. Deciding between the models will likely require demonstrating a preponderance of evidence for one model that is inconsistent with the other. Information is needed regarding broad-level folding studies of a large diversity of different protein molecules (sequence composition, size). MBP is more than 100 residues larger than the largest protein with a folding mechanism characterized at structural resolution, to our knowledge.

1.2 Large Protein Folding with MBP

Kinetics and thermodynamics of large multi-domain protein folding processes are poorly understood (88-90); however, roughly 40-65% of prokaryotic proteins and 65-80% of eukaryotic proteins contain more than a single domain, as demonstrated by analysis of different genomes (91-95). Most protein folding reactions, *in vivo*, involve much larger systems than those typically studied in biophysics.

There have been a small number of folding studies involving large proteins (see reviews by Jaenicke (96) and Clarke (97)); however, none attain the site-specific conformational resolution requisite for true insight into folding processes. Defining the folding pathway for a large multi-domain protein, like MBP, will facilitate insights into how the basic driving forces for folding scale with protein size.

1.2.1 Background

MBP Structure

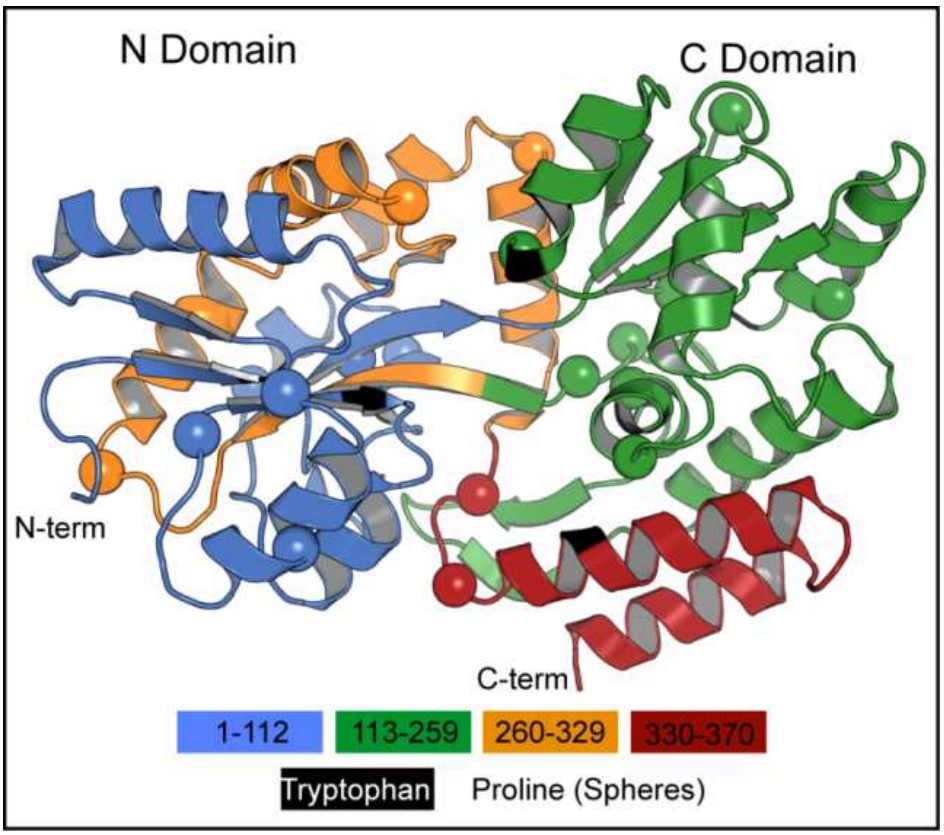

Figure 1.1: The complex topology of the two domain MBP. The chain crisscrosses between domains three times, naturally dividing into four sections on this basis and represented with colors as indicated above. In both cases, domains form from two sequentially discontinuous sections of the chain. There are 21 trans-proline residues (c_α spheres) and tryptophan residues are colored black. PDB ID: 1OMP (98)

Complex topology of the 41 kDa, 370-residue maltose binding protein is shown in Figure 1.1 where the two domains formed by discontinuous regions of the primary sequence are highlighted by color. The primary structure is composed of 21 prolines, 8 tryptophans (good for fluorescence spectroscopy) and 176 hydrophobic[3] residues (47.6%). It would be predicted to have a rather compact unfolded state based on work mentioned earlier by Ken Dill (42). MBP displays an α/β fold architecture with 164 residues forming 19 helices (44%) and 74 residues comprising 22 beta strands (20%) which associate to form 3 sheets in the native protein (predicted by DSSP (99)).

[3] Hydrophobic residues were defined as A,I,L,V,F,W,Y,P

General Interest in MBP

Much of our interest in MBP began after finding a large discrepancy between the predicted folding rate, based on Plaxco and Baker's work (100), and our own spectroscopic folding studies. They found that folding rates for small, single-domain proteins seemed to be well correlated with their relative contact order, defined as the average sequence separation between contacting residues compared to the chain length. Using Baker's contact order calculator (101), MBP (PDB ID: 1OMP) has a relative contact order of 9.8% and is predicted to fold with a rate of 10,000 s^{-1}. The contact order prediction is incorrect by a factor of 10^5!

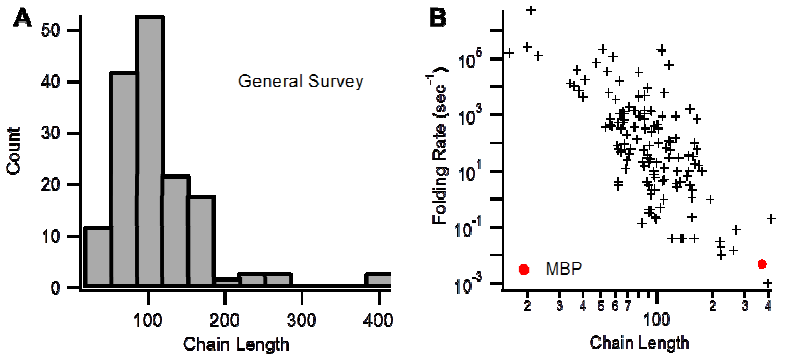

Figure 1.2: Proteins studied by biophysical methods. This information was taken from a literature survey. **(A)** Shows a bias in the size of proteins studied by biophysical methods. **(B)** An illustration of the relationship between protein size and folding, including our measurements for MBP. *References - data shown in A & B were taken mainly from (102) and updated with rates from (14, 15, 103, 104) and included non-redundant data taken from (105).*

MBP folds over a time scale of minutes with several optically resolvable kinetic rates (106-108) separated by more than three orders of magnitude and exhibits reversible 2-state unfolding (109). In later work which sought to understand why multi-state folders deviated from the contact order prediction (Ivankov et al. (102), Table 1) published a fairly comprehensive list of proteins that had been studied by optical spectroscopy reporting information such as protein size and the relaxation rate[4] measured during refolding experiments. From plotting this information[5] (Figure 1.2), I noticed that the cumulative knowledge of the protein folding process was largely based on small proteins containing ~100 amino acids, verifying sentiments offered in

[4] If multi-exponential kinetics were observed, the slowest rate with appreciable amplitude, not due to proline mis-isomerization, was used.

[5] Panel A includes information published in 2003, mainly. It was updated with additional proteins in 2013 from work by the same authors but this did not change the trends – this information motivated us to study MBP in 2008. See the figure legend for complete referencing.

many biophysics texts (88-90), that we know very little about large protein folding. Obviously, there is some relationship between chain length and folding complexity, perhaps a deeper understanding of large protein folding processes could suggest reasons for this trend.

Native-like on-pathway intermediates have been observed in a number of cases (110-124). There have actually been a number of proteins characterized at structural resolution; the overwhelming evidence in those cases (5, 9, 125, 126) point towards a PPOE-type model for folding. There are studies involving large protein folding (14, 15, 104, 127, 128); but none provides structural insight into the folding mechanism. The absence of information drove us in the direction of studying large protein folding.

MBP Burst-phase & Compaction

In seminal MBP folding work, Lynn Randall's group observed rapid (within instrument dead time) formation of what they termed a pre-rate-limiting step intermediate (now called the burst-phase), accounting for ~20-30% of the total change in tryptophan fluorescence upon folding (106). This burst-phase structure exists at what appears to be steady state for the first few seconds of folding; perhaps due to a dramatic reduction in subsequent rates, it is hypothesized that the burst species may either fold to the native state, or, polymerize in a process of "reversible aggregation (109)".

FRET data suggests the burst-phase to have roughly 70% of the native state compaction (129) and therefore representative of a collapse event. Though where FRET indicates collapse, often SAXS demonstrates an expanded ensemble causing many researchers to doubt that proteins collapse without forming substantial native-like structure (41).

Does MBP collapse and bury tryptophan residues from the solvent leading to the burst-phase fluorescence signal? Does it form any regular structure during the burst-phase? Is the burst-phase simply a new unfolded state, reflecting "poor" solvation once diluted out of chemical denaturant? If it is stabilized by spurious intrachain contacts, what effect might this have on subsequent folding? Answers to these questions (Chapter 5, p. 81) require experiments that can provide structural information, such as HX.

1.2.2 Exploring Kinetic Folding Models with MBP

Perhaps MBP folds via IUP mechanisms. An ultra-fast track to the native state could explain why 20-30% of the fluorescence signal is recovered in the burst and help support an IUP model. A rate such as the one predicted by MBP's contact order (10,000 sec^{-1}) would be unresolvable using the instruments employed thus far, perhaps this estimate represents an ultra-fast track to native. The multi-exponential relaxation behavior observed in spectroscopic refolding experiments could be reconciled by independent folding tracks. Such explanations could be easily evaluated with an experiment that provided reliable state and structural information, such as HX.

Perhaps MBP folds via PPOE mechanisms. The presence of sequential folding, such as the observation of native-like structure building upon itself with respect to time, would be suggestive of sequential stabilization. Accumulation of a stable obligatory intermediate would indicate the presence of a single macroscopic pathway and strongly support the PPOE model; this would reconcile the slow folding observed in MBP. These questions also will require some state and structurally sensitive measurement, such as HX.

Finally, one must consider the possibility that ideas from both mechanisms, PPOE and IUP, could be useful for describing the folding of MBP. If MBP collapses to a random ensemble of conformations, it may be that very different energetic barriers separate different molecules from assuming a conformation that is permissible for downstream pathway-directed folding, as has been suggested recently (130). Such a mechanism might be better represented metaphorically by a funnel that empties into a pathway. In any case, it will be necessary to measure the temporal acquisition of structure in a state sensitive manner as the many molecules transition from the unfolded to the native state.

1.3 Dissertation Overview

Large proteins are harder to study at structural resolution for many reasons described by others (97) and in chapters that follow. This is the likely explanation for the dearth of structural information on large protein folding. No folding pathway or mechanism has been structurally characterized for chain lengths greater than ~250 residues to our knowledge. Following a description of HX theory and general experimental overview (Chapter 2, page 15),

we describe technological advances in HX by NMR (Chapter 3, page 25) and MS (Chapter 4, page 49) developed, in part, to assist our work on large protein folding. Equipped with the necessary tools, we examine and present the folding of MBP (Chapter 5, page 81) in structural detail. We close with a discussion of future directions for MBP research and larger protein folding studies in general (Chapter 6, page 106).

1.4 Impact and Specific Questions Addressed In This Dissertation

Using HX NMR, HX MS, optical fluorescence (tryptophan, ANS), circular dichroism, and SAXS, we characterize the folding pathway, the burst-phase, questions regarding collapsed and unfolded states, and assess the application of IUP and PPOE models to these results. We specifically explore the following questions:

- Does MBP collapse?
    - Is regular structure present?
        - What does it look like?
- What causes the optical burst in MBP?
    - Is the burst a rapid re-equilibration of the denatured ensemble?
    - Does the burst represent a classical intermediate?
    - Does the burst represent a fast folding track to N?
- Are there multiple pathways from U → N?
- Are there folding intermediates?
    - What do they look like?
    - Are they obligatory?
    - What might they suggest about energetic barriers?
- Is there evidence of sequential folding/unfolding behavior?
    - Do we see foldons?
- Can we generalize about large protein folding?
    - Why does MBP fold slowly?
    - Do we see evidence of similar features in other large proteins?

The combination of many technologies and advancements described in this thesis allowed us to study and provide answers, directly in most cases, to these questions in chapter 5.

We hope the technologies developed for this work and described herein will enable others to overcome the various issues faced in large proteins and contribute, as we have, to further our structural understanding of large protein folding processes.

Chapter 2 - HX Theory & Experiment

Labile protons, such as those bound to nitrogen and oxygen atoms, continually exchange with solvent protons. The exchange reaction is catalyzed by acids and bases, most frequently H+ or OH-. As such, the main factors contributing to the measured first order reaction rate are temperature and pH; however, protons participating in hydrogen bonding interactions are sequestered from catalyst and are not free to exchange. This unique sensitivity to hydrogen bonding has led to the growing popularity of hydrogen exchange labeling experiments in structural biology and biophysics. Most useful for these purposes, and the focus of this thesis, are backbone amide protons, one for each amino acid aside from proline. In addition, the sidechains Asp, Glu, Asn, Gln, Ser, Thr, Tyr, Lys, Arg, Trp, and His each have one labile proton; however, their rates of exchange are much faster than the amide proton and are less useful for structural studies.

2.1 Amide Hydrogen Exchange Basic Principles

2.1.1 The Chemical Basis

Proton transfer reactions may be written as:

$$AH + Cat \underset{k_{-1}}{\overset{k_1}{\leftrightarrow}} AH \cdots Cat \underset{k_{-2}}{\overset{k_2}{\leftrightarrow}} A \cdots HCat \underset{k_{-3}}{\overset{k_3}{\leftrightarrow}} A + HCat,$$

and an overall rate constant for $AH + Cat \xrightarrow{k_{\text{int}}^{cat}} A + HCat$ may be conveniently approximated (131) as:

$$k_{\text{int}}^{cat} = k_1 \frac{10^{\Delta pK_a}}{\left(10^{\Delta pK_a} + 1\right)}, \qquad \text{Eq. 2.1}$$
$$\Delta pK_a = pK_a(HA) - pK_a(Cat).$$

The maximum rate being given by the diffusion limited rate constant, k_1, predicted long-ago by Debye to be roughly 10^{10} M^{-1} sec^{-1}. By considering the effects of neighboring functional groups to individual amide pK_a, values one may determine the pH- and structure-dependent first-order rate constant, k_{ch}, the "chemical rate" for each exchangeable site on the protein backbone,

$$k_{ch} = k_{\text{int}}^{\text{OH}}(\lambda,\rho)\left[\text{OH}^-\right] + k_{\text{int}}^{\text{H}_2\text{O}}(\lambda,\rho)\left[\text{H}_2\text{O}\right] + k_{\text{int}}^{\text{H}_3\text{O}}(\lambda,\rho)\left[\text{H}_3\text{O}^+\right]. \qquad \text{Eq. 2.2}$$

Aside from an expected dependence on temperature and solvent isotopes (not shown), the second-order "intrinsic" rate constants in Eq. 2.2, $k_{\text{int}}^{catalyst}(\lambda,\rho)$, depend on inductive and steric effects from the local chemical environment (nearest sideschains, represented by (λ,ρ)) and to a smaller degree, the ionic strength of the solution. HX chemistry is now fairly well understood – the intrinsic exchange rate constants are available for all 19 relevant amide protons under any combination (λ,ρ) that might be present in protein structures (132) and a spreadsheet is available to make these calculations with ease at www.hx2.med.upenn.edu.

The fortunate aspect of hydrogen exchange for the protein biophysicist as we will see is the particular ΔpK_a between relevant catalysts (OH^-, H_3O^+) and the amide proton. Most other hydrogen exchange reactions are either too fast or too slow for convenient operational use. The chemical rate constants for HX between aqueous catalysts and protein amide protons are adjustable over many orders of magnitude – at low pH, exchange lifetimes are hundreds of minutes; at high pH, lifetimes are microseconds. This provides tremendous flexibility for experimental design; conditions may be selected to be most suitable for the system being studied. A useful rule-of-thumb for design purposes: at pH 7.0 and 273K, the amide proton lifetime is roughly 1 second and increases by a factor of 10 per pH unit and similarly per 20K increase in temperature. Primary structure effects, including full reasoning for the neglect of non-amide proton transfer reactions herein may be found in references (131-134).

2.1.2 The Structural Basis of Hydrogen Exchange

In the early fifties, Professor Linderstrøm-Lang began investigating hydrogen-deuterium exchange in small peptides and proteins (135) and in attempt to describe slowly exchanging protons, proposed that protein conformational states influence exchange competence – there are conformations where exchange does not occur. In subsequent years (136), it became clear that H-bonding interactions were responsible for slowing exchange rates.

Hydrogen exchange provides time-resolved information on structural dynamics because H-bonding interactions must first break or open before exchange can proceed. H-bonding

interactions protect the proton from catalyst and prevent the exchange reaction. This is conveyed in the following diagram:

$$\text{NH(closed)} \underset{k_{cl}}{\overset{k_{op}}{\rightleftarrows}} \text{NH(open)} \xrightarrow{k_{ch}} \text{exchanged} \qquad \text{Eq. 2.3}$$

Thus, HX chemistry (k_{ch}) provides an upper limit for the measured exchange rate. The measured exchange rate, k_{ex}, is slowed by the presence of protein structure, permitting access to H-bond opening (k_{op}) and closing (k_{cl}) rates. For a thorough treatment of the mathematics, see the seminal work of Aase Hvidt (137, 138). My purpose here is to provide a functional explanation of HX phenomenon as it relates to protein structure. A full review of these concepts along with a history of HX experiments may be found in reference (139) and more recently in Englander et al. (11). The structural basis of HX slowing may be described by the following:

$$k_{ex} = \frac{k_{op} k_{ch}}{k_{op} + k_{cl} + k_{ch}} \qquad \text{Eq. 2.4}$$

A plethora of alternative explanations for observed slowing has been explored in the experimental literature along with testing on staphylococcal nuclease NMR data in a recent study (140) – while other factors may potentially slow HX reactions, the dominant cause for HX slowing, by many orders of magnitude, is protection by H-bonded structure.

2.1.3 Limiting Conditions

Two limiting cases emerge when one considers structure effects (141-145). The EX2 limit occurs when two conditions are satisfied. First, any exposed proton must have a greater likelihood of protecting than exchanging ($k_{cl} >> k_{ch}$) and second, the protected state must be favored over the exposed state ($k_{op} << k_{cl}$). If these conditions are met, proton exchange competes with proton H-bonding. By rewriting Eq. 2.4 to express EX2 conditions, the equilibrium constant (K_{op}) for H-bonded structure and its free energy (ΔG_{HX}) may be determined from the measured exchange rate (k_{ex}) in the following way:

$$K_{op} = \frac{k_{op}}{k_{cl}}, \; k_{ex}^{\text{EX2}} = \frac{k_{op} k_{ch}}{k_{cl}} = K_{op} k_{ch} \qquad \text{Eq. 2.5}$$

$$\Delta G_{HX} = -RT\ln K_{op} = -RT\ln\left(\frac{k_{ex}^{EX2}}{k_{ch}}\right) \quad \text{Eq. 2.6}$$

The second limiting condition occurs when the chemical exchange rate is much faster than H-bond formation ($k_{ch} >> k_{cl}$) and there is no longer a competition for exchange. In this condition, known as EX1, every opening event leads to exchange, every solvent exposed proton exchanges. Here, the measured exchange rate and the opening rate of the H-bonded structure are numerically equivalent:

$$k_{ex}^{EX1} = k_{op} \quad \text{Eq. 2.7}$$

2.2 HX Labeling In Practice

Hydrogen exchange is a labeling method. Most labeling techniques involve covalent modification of protein structure, often introducing new functional groups and can have undesirable side effects; hydrogen exchange involves the simplest modification possible, the addition of a single neutron. The technique is minimally invasive. Furthermore, exchange experiments may be done in both directions, H-to-D and D-to-H to verify that the label has no meaningful influence on the system. Additionally, because each amino acid is independently labeled, HX measurements principally provide structural information for each residue in the polypeptide. Other labeling techniques are typically restricted in resolution because labeling generally involves only a subset of residues in the protein and many different labeling strategies must be combined to provide a global picture. These features set HX apart from all other labeling techniques.

2.2.1 Native State Hydrogen Exchange (NHX-type)

"A protein cannot be said to have 'a' secondary structure but exists mainly as a group of structures not too different from one another in free energy… the molecule must be conceived as trying out every possible structure each in accordance with its Boltzmann factor." This statement was written in 1959 by K.U. Linderstrom-Lang and John Schellman (136) working together on protein hydrogen exchange and long before the exact types of motion that determine HX behavior were known.

The NHX namesake arises because of collecting these experiments under conditions where all molecules have structurally equilibrated and often where the native state is favored. In such conditions, the ensemble average conformation is native; however, all molecules are continuously exploring higher energy, exchange competent forms and the frequency of these deviations are reflective, under EX2 conditions, of the free energies of local structure. Structures with lower stability exchange faster than those with higher stability.

NHX experiments have been successfully utilized to study unfolding reactions of proteins, notably, the method was used to structurally characterize independently unfolding subunits of structure in a protein that optically unfolds in a two state manner (9). NHX experiments are also used to characterize structural changes that result from ligand binding. In one example for a lipid binding protein (146), regions where labeling rates change in the presence of lipid provide information on the induction of structure by binding. In Chapter 3, NHX experiments on MBP analyzed by NMR are discussed briefly.

2.2.2 Pulse-Labeling Hydrogen Exchange (KHX-type)

Kinetic pulse-labeling experiments (KHX) provide time-resolved insight into temporally fleeting events that are not accessible via the NHX method. In contrast to NHX where the experiment is conducted under equilibrium conditions, KHX experiments are generally synchronized-start experiments and involve interrogating a system as it relaxes to equilibrium. KHX has traditionally been used for studying protein folding reactions; but has also been used for more exotic purposes such as studying conformational changes occurring during the catalytic cycle of chymotrypsin (147).

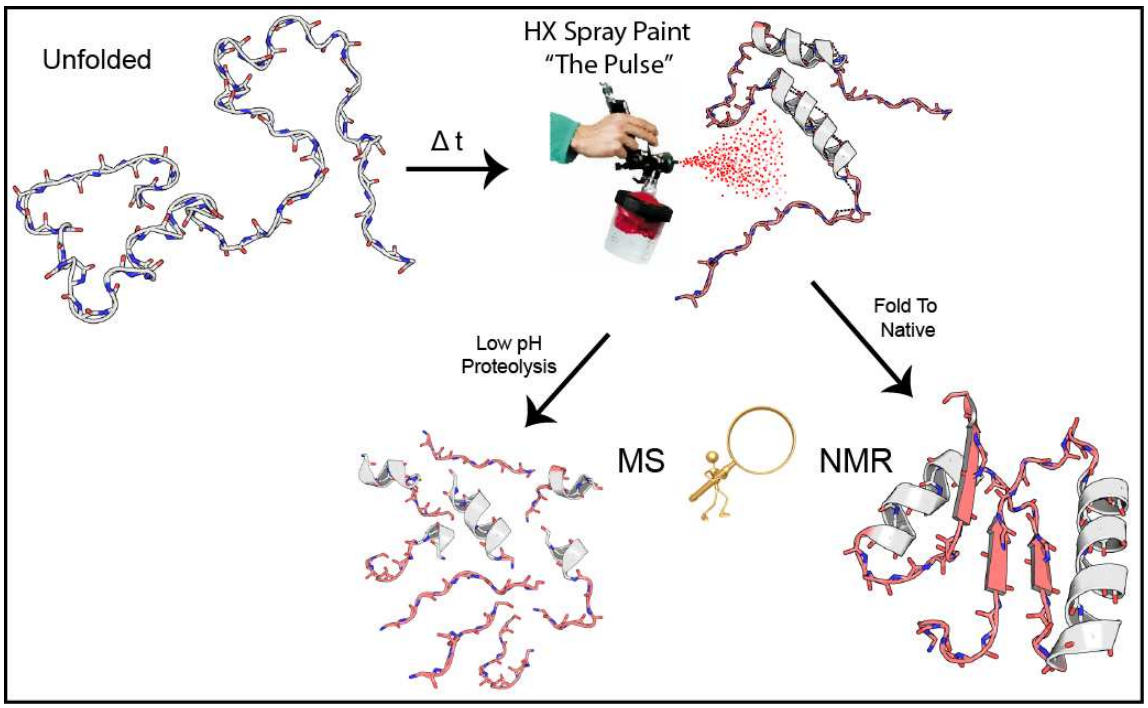

Figure 2.1: The kinetic pulse labeling HX experiment. The protein is fully unlabeled in the unfolded state before diluting into permissive refolding conditions. At a variable time during folding, the HX pulse is applied. This is analogous to spray-paint that will not stick to hydrogen bonded residues, such as those in the helical conformation above. Following the pulse, for MS analysis, the sample is digested by proteases and the fragments separated, this is described in Chapter 4, page 51. For NMR analysis, the sample is allowed to continue folding until the native structure is acquired. Those regions with no paint were folded at the time of the pulse.

KHX experiments, in the context of protein folding investigations, are typically performed with a rapid mixing device to facilitate fine control over labeling times. Starting from a chemically unfolded state, folding is initiated by dilution of denaturant to permissible folding conditions where exchange is neutralized. Folding is allowed to proceed for a variable amount of time before rapidly exchanging solvents, usually by rapid dilution, into a brief, high pH condition where exchange is accelerated. This is quite analogous to the application of spray-paint only to those regions of structure that are not H-bonded, as is illustrated in Figure 2.1. This process results in a labeling pattern which reports on the structure that formed during the folding phase, before application of the labeling pulse.

Under favorable conditions where both structural opening and closing reactions are negligible over the period of the pulse, KHX is a binary experiment and easily interpreted. Amide protons that develop H-bonded structure during the folding phase, such as in an intermediate, are protected from exchange whereas sites that remain unfolded will label to completion during

the pulse. Structural information may then be directly inferred from the time dependence of H-bonding. Regions of the sequence involved in intermediate structures will display protection from labeling on a similar time scale.

If structural opening reactions occur on the time scale of the pulse, residues that would have been completely protected during the pulse may lose some label and introduce a second order effect on the measurement. In these situations, changing the pulse length, pH, or temperature can each provide information on the equilibrium constant for H-bonding at each protected site and allow one to draw inferences regarding structural heterogeneity. In Chapter 5, I characterize an obligatory intermediate in the folding pathway of MBP using the binary experiment and discover heterogeneous structural features of the MBP burst-collapse event using variable pulse lengths at a fixed folding time.

2.3. State Sensitivity in HX MS

HX MS provides an opportunity to independently monitor the partition of molecules between definable non-degenerate conformational states. State sensitivity is essential for discerning between multiple pathway models such as IUP and sequential pathway models such as PPOE. There are an exceedingly small number of ways to achieve state-sensitive measurements. To our knowledge, HX MS is the only direct approach. By using an example from the MBP pulse labeling HX MS dataset (presented in Chapter 5), this exceptionally rare capability is demonstrated here.

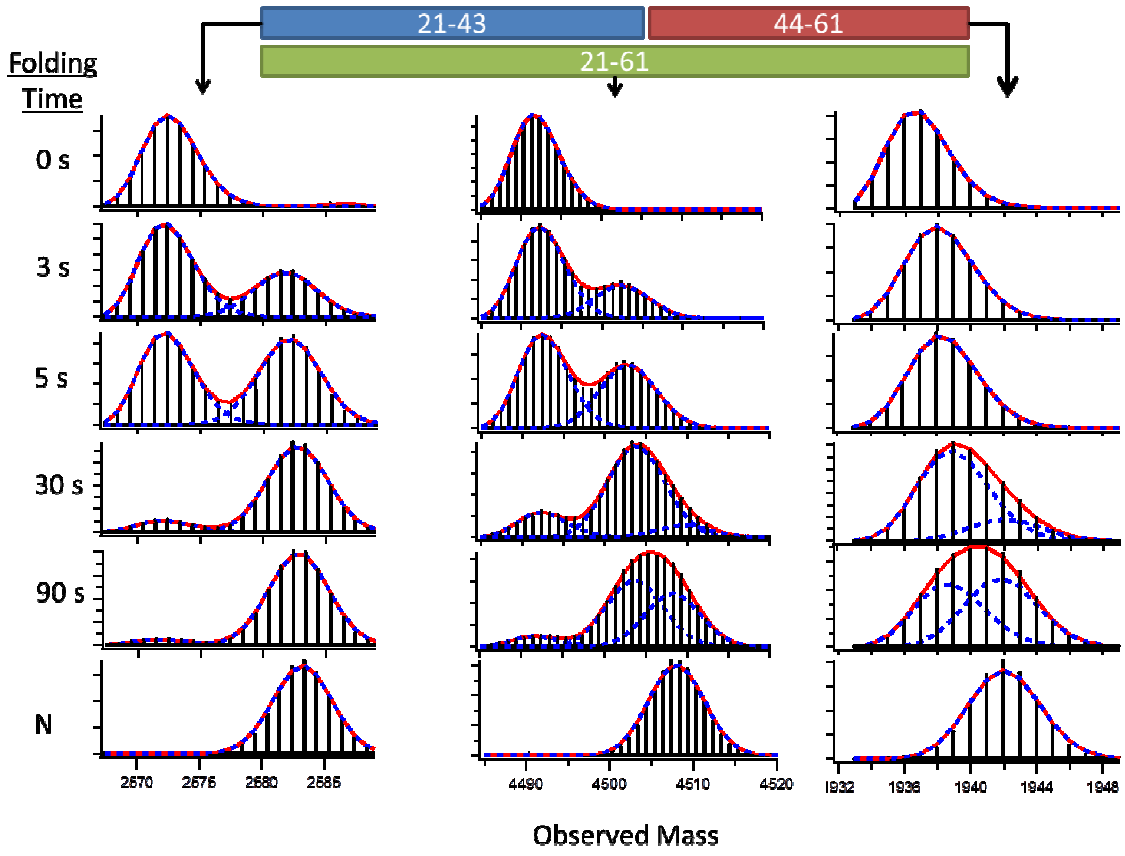

Figure 2.2: State sensitivity in HX MS experiments. Spectra are shown at different refolding times for three MBP peptides. Three states appear due to the presence of an obligatory intermediate (7 second lifetime) that contains peptide 21-43 and not peptide 44-61. A peptide covering both is also shown to demonstrate that we are not limited to two states with this approach. If states are degenerate with respect to the number of H-bonds, they will appear indistinguishable; states are also hard to resolve if they are not sufficiently separated by mass.

In KHX experiments, we interpret protection from exchange as representative of H-bonded structure that formed prior to application of the labeling pulse (or spray-paint in Figure 2.1). Because mass measurements are not ensemble-averaged in the classical sense, using the HDpop program, described on page 73, for analysis of mass distributions provides unbiased state sensitivity; this is illustrated in Figure 2.2 where the blue dashed lines demarcate individual populations whose sum is traced in red. The area contained within each blue dashed line represents the relative population fraction of molecules in the particular HX state, defined by the average number of deuterium retained.

The KHX experiment, a focus of this dissertation, tends to have a binary nature with respect to stable H-bonded structure formed before the labeling pulse. For example, if an

initially deuterated 20 residue peptide has 10 protecting H-bonds due to structure that formed in 50% of the molecules prior to pulsing the sample with H_2O, we would expect to see two populations – one with the mass of the unfolded control and the other shifted by roughly 10 Da after back exchange correction. Each population would have an equivalent area reflecting the 50/50 distribution of molecules. Figure 2.2 shows MBP peptide 21-43 who behaves in this manner. This peptide stably forms a tentative helix that protects 12 deuterons with a folding lifetime of 7 seconds; thus, we observe a heavy population who transitions from the unfolded, low mass, distribution to the +12 mass distribution with an appropriate time signature. Similarly, peptide 44-61 also gives a binary result although the transition occurs on a different time scale. The power of HX MS state sensitivity is most clearly demonstrated in peptide 21-61 which clearly shows three populations who transition between one-another on timescales reflective of the two different behaviors in the smaller peptides in Figure 2.2.

Generally, each population in a KHX experiment has a constant mass with respect to folding time; but, if observed, the time dependence of mass changes for a given population may provide useful information. For example, the light population centroid in peptide 44-61 appears to shift by approximately +3 Da between the 0 s and 30 s time points after correcting for back exchange. Upon quantitation of the time dependence, we observe this mass shift to coincide with the bulk population transition shown in peptide 21-43. This can occur when two populations are not resolved from one another – in this case, the centroid of the cumulant[6] population will move, just as observed in peptide 44-61. We interpret such observations to represent, in most cases, two populations who are unresolved. In this particular example, the mass shift observed in 44-61 represents a β-strand with three H-bonds whose concerted formation with the helix described above and proximity of the residues in the native structure implies both structures form together with a characteristic lifetime of seven seconds. This is explored in detail in Chapter 5**.**

A binary KHX result, such as the one just described, requires that the opening rate lifetime of protecting structure exceed the duration of the pulse by at least a factor of three;

[6] Cumulant – I have defined this to describe two populations who are unresolved from one another. As the fraction of molecules in one unresolved population transition to the other unresolved population, the centroid mass of the cumulant envelope will shift from being dominated by the first to being dominated by the second unresolved mass centroid.

otherwise, preformed structure will open in a significant fraction of the molecules during the pulse and subsequently exchange. In this case, there will appear to be fewer molecules in the protected population than actually existed when the labeling pulse was applied. Modulating the pulse duration allows us to test for such behavior during the MBP collapse (p. 91).

2.4 Conclusions

The amide hydrogen exchange experiment is most useful for studying the dynamics of protein structure because of the similarity in pK between HX catalysts and the backbone amide group of polypeptides. Because structure leads to H-bonding and H-bonding in turn sequesters any amide proton from exchange, by measuring the difference between the well-calibrated free exchange rate (k_{ch}) and that observed in experiment (k_{ex}), one may draw conclusions about the nature of protecting structure.

Due to the sensitivity of the exchange rate to pH, under equilibrium conditions, the HX measurement can provide information on the free energy of protecting structure and on its opening rate. When operating in kinetic mode, HX may also be used to study the temporal acquisition of structure during protein folding reactions. Beyond these provisions, HX coupled to MS provides unbridled state-sensitivity, a feature that we capitalize on throughout this work. Much more detail about the HX MS experimental methodology and data processing algorithms are described in Chapter 4. For completion, my work involving NMR and primarily the NHX experiment are also included and described in Chapter 3.

Chapter 3 - Nuclear Magnetic Resonance & HX NMR

3.1 Introduction

Fundamentally, HX experiments are interested in measuring either the rate of proton exchange (NHX type experiments) or the proton:deuteron ratio (KHX type experiments) at each amide on the protein back-bone. The measurement of HX data by NMR capitalizes on the fact that ^{2}H is NMR silent. Therefore intensity of any given amide ^{1}H signal will decrease linearly as molecules exchange ^{1}H with deuterated solvent and are replaced by ^{2}H. NMR is the only method capable of directly providing site-resolved hydrogen exchange measurements, as such, the value of an HX NMR experiment relies acutely on ability to resolve each amide and precisely measure the change in signal intensity with respect to time. HX NMR on larger protein systems faces three problems that directly interfere with the ability to make reliable and comprehensive HX measurements.

3.1.1 Molecular Tumbling Time

The challenge of NMR, in general, for larger proteins lies in the effect of molecular tumbling time on the line width of the signal. Generally, as protein size increases, so does molecular tumbling time. Longer tumbling times broaden the NMR signal due to their effect on transverse relaxation (see (148) for a thorough explanation of this effect). Increased tumbling time effectively reduces the precision of the HX measurement and magnifies issues of peak overlap.

MBP sits on the edge of this issue, the tumbling time is slow (~19 ns (149)) but still amenable for NMR at 37°C. The issue of slow tumbling time is mentioned here in an effort to present a complete picture of the challenges facing NMR for large proteins. Were it not for the faster tumbling time at 37°C without a loss in stability for MBP, this would have been an issue. Josh Wand's group has pioneered a strategy involving reverse micelle encapsulation to address the slow tumbling problem (150); this strategy has been explored for the purposes of HX NMR with limited success.

3.1.2 Spectral Crowding in Large Proteins

The number of NMR signals increases linearly with the number of amino acids for HX experiments. Even for proteins with a high degree of spectral dispersion, peak overlap or spectral crowding (for two-dimensional experiments) becomes an issue for proteins greater than 100 amino acids (148) and the problem worsens as the number of signals (residues) increases; MBP has 370 residues. As long as a signal is resolved[7], the HX rate may be measured; however, the number of resolved signals will decrease as the total number of signals increase and therefore the amount of useful information recorded during the HX NMR by HSQC is diminished. For HX NMR experiments on large proteins where spectral crowding is an issue, a method to determine which peaks are resolvable in 2D and then match observed peaks with their appropriate assignment is needed.

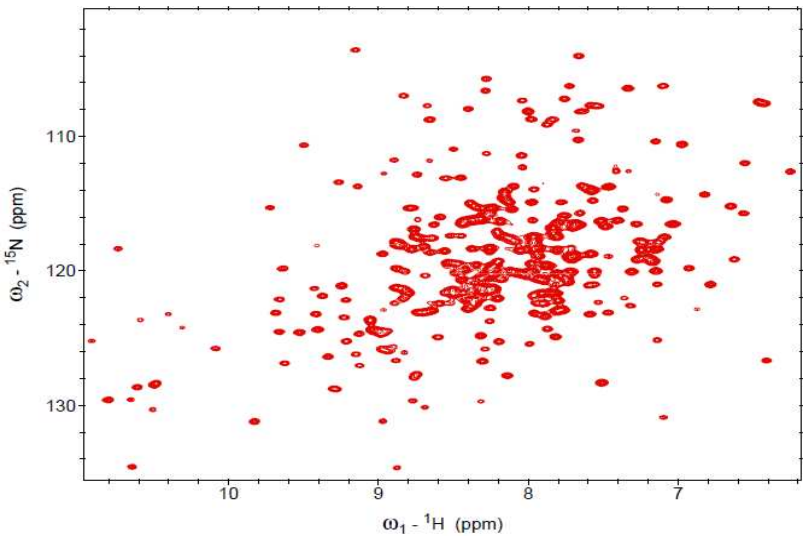

Figure 3.1: ^{1}H-MBP HSQC spectrum collected at pH 9.6. Using the peak picking routine in Sparky, this spectrum was determined to have 253 peaks, 334 peaks are expected. This difference is a result of spectral crowding.

Using the 370-residue MBP, my strategies to allay the effects of spectral crowding are discussed. We first had to develop an algorithm to determine which peaks in the observed spectrum could be matched unambiguously to known chemical shifts for MBP deposited in the Biological Magnetic Resonance Data Bank (BMRB). The resulting algorithm not only deals with

[7] A peak (m_1) is resolved if the following condition is met for all 'x' nearby peaks: $m_1 - m_x > \sigma_1 + \sigma_x$, m =midpoint, σ = Gaussian width. This is written for the 1D case, see later text for expansion into multiple dimensions.

spectral crowding, but also automates peak matching for any NMR experiment (any dimensionality) and is available upon request (ben@btwalters.com). Using this algorithm (described in section 3.2.1) we determined that 172 of the 253 peaks in Figure 3.1 were fully resolved and unambiguously assigned; thus one could reliably measure 51% of the available amide protons in MBP. As a result, I was able to determine an upper bound on the free energy of the native state for MBP.

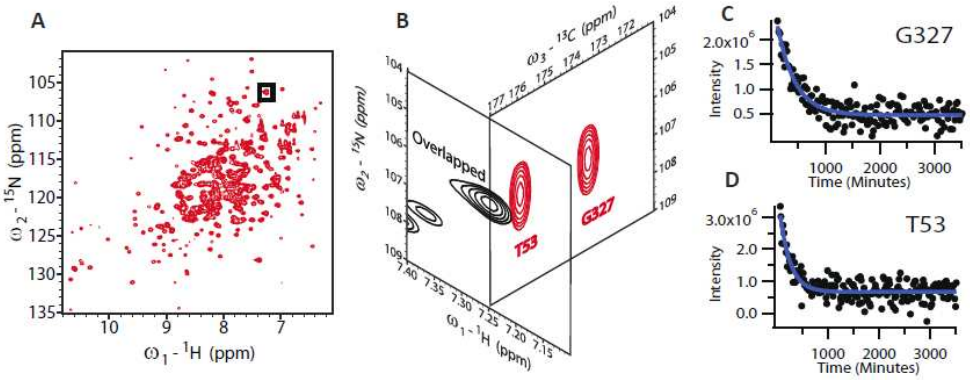

Figure 3.2: The advantage of HX 3D-NMR. **(A)** A typical HSQC spectrum for MBP with a high degree of spectral crowding and paying attention to the peak highlighted by the black box, **(B)** these two peaks are only resolved by the addition of a second dimension that required ~16 hours of acquisition time. **(C)** Using the AMORE-HX method, both peaks in panel B are resolved with the appropriate time resolution for measuring the rate of exchange as is shown by the solid blue line drawn through the data points.

To avoid the spectral crowding issue altogether, we created a novel three-dimensional experiment called AMORE-HX (1). Peaks that are unresolved due to crowding in the HSQC (Figure 3.1 & Figure 3.2A) may be resolved by the inclusion of a third dimension (Figure 3.2B) and their exchange rates for amides overlapped in 2D may be measured (Figure 3.2C-D). The main challenge here was time resolution. A single HSQC spectrum may be collected in ~40 minutes whereas the HNCO typically requires a full day to collect with equivalent signal-to-noise. This time resolution makes HX measurements nearly impossible. In the AMORE-HX experiment, we present a cadre of strategies that focus on increasing the time resolution of the ^{1}H-^{15}N-^{13}C backbone HNCO correlation experiment for the purposes of HX NMR. Figure 3.2 collectively demonstrates the capability of our solution, termed affectionately, AMORE-HX. Where only 51% of the available chemical shifts were resolved in the HSQC, in section 3.3, I describe the AMORE-HX experiment and show how 92% of the deposited chemical shifts for

MBP are resolvable using this new experiment. This work resulted in a first-author publication, see reference (1).

3.1.3 Spectral Density is Sensitive to The Presence of Deuterium

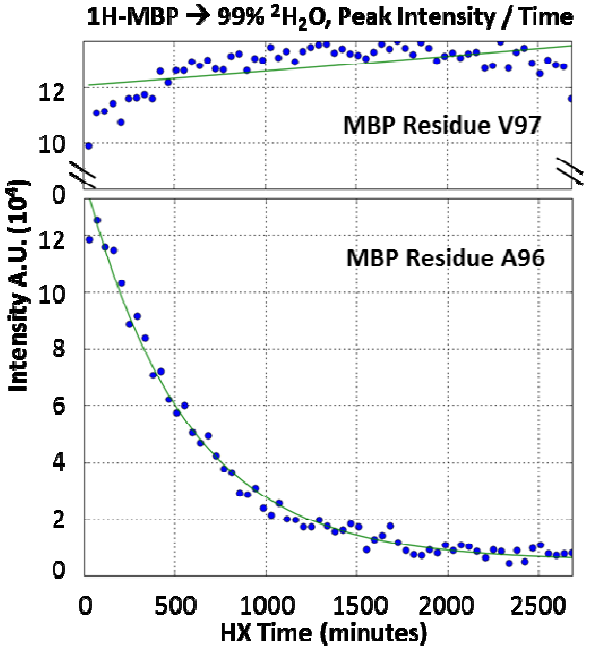

Figure 3.3: Second order issues in HX NMR. The top trace shows what would appear to be a signal growing more intense with time (this should be impossible) before starting to decay later. Upon inspection of the neighbor (lower trace) we find a decaying signal, as expected, with roughly the same lifetime as the growing trace above. The spectral density changes as more protons are replaced by deuterons and spectral density influences transverse relaxation rates. This is the cause for the growing signal observed in the top trace which results from the exchange shown in the lower trace.

The final issue regards an effect on both longitudinal (T_1) and transverse relaxation (T_2) that manifests through the continuously changing spectral density during the HX experiment. The rate of transverse relaxation for an amide ^1H signal increases as neighboring protons are replaced by deuterium. This narrows the linewidth of the ^1H signal and has the effect of increasing the measured amplitude as shown in Figure 3.3 – the amplitude of residue 97 (top) increases while its neighbor, residue 96 (bottom) decays. As both sites in Figure 3.3 are initially fully protonated, the signal should only decay upon dilution into ^2H_2O. Detrimental effects on the HX measurement because of this narrowing may be avoided by measuring the change in cross-peak volume or area as opposed to amplitude with respect to time; however, accurate estimations of peak areas or volumes may be complicated by spectral crowding. Ironically, this

effect of line narrowing that results from a reduced spectral density has been exploited in structural NMR studies by selective deuteration to overcome line broadening from longer tumbling times in large proteins (149, 151, 152), whereas in HX NMR this introduces a second order error in the HX measurement and is undesirable.

The same changing spectral density has the opposite effect on longitudinal relaxation rates (spin-lattice relaxation) which decrease during the HX experiment. A recycle or inter-scan delay is set in between pulses to allow the system to relax and realign with the magnetic field. If the relaxation time of a given spin increases during an experiment beyond the fixed inter-scan delay it will introduce non-equilibrium effects from over-pulsing. As more and more neighbors become deuterated during an HX experiment, the longitudinal relaxation time increases and so too does the degree of over-pulsing. As more and more deuterium is exchanged in, the remaining protons relax more slowly. To the HX measurement, exchange and over-pulsing have equivalent effects – both reduce the number of NMR-active nuclei from one scan to the next, if not corrected, this artificially causes an overestimation of the HX rate. The magnitude of this artifact (30-35% difference in the HNCO) can be quite large as is discussed in section 3.3.2.

Recent developments in pulse sequence design have brought about the use of selective excitation pulses termed SOFAST (153-158) and BEST (159, 160) which may largely solve this problem when used in HX NMR experiments. Selective excitation pulses only excite a subset of the nuclei of a particular type and results in more pathways existing for spin-lattice relaxation. In turn, inter-scan delays in non-HX applications can be reduced by a factor of ~50 as these pulses greatly accelerate spin-lattice relaxation rates. Speeding up the NMR experiment was the motivation for development of these selective pulses. With a faster intrinsic spin-lattice relaxation rate, incorporating selective excitation into the traditional experiment should eliminate the over-pulsing problem altogether. We explored the potential for using this strategy while developing the AMORE-HX experiment and show that selective excitation is able to completely ameliorate the issue arising from over-pulsing.

3.2 HX NMR using the HSQC on MBP

In recent years, the ^{15}N-HSQC has become the standard method to measure NHX experiments. The first step entails acquisition of the fully protonated spectrum before rapidly

replacing the buffer with D_2O and then collecting sequential spectra as peaks decay due to exchanging NMR active hydrogen with NMR inactive deuterium. Though resolving the maximum number of cross-peaks is challenging for larger protein systems, misinformation resulting from incorrectly assigned peaks is arguably much more of an issue. Avoiding misinformation was a key motivation in our efforts to develop the AMORE-HX experiment that is discussed in later (page 38). In this section, I will present my unique solution for avoiding misinformation resulting from the spectral crowding problem along with an example of HX NMR using the HSQC on MBP.

3.2.1 An Algorithm to Handle Spectral Crowding

The 2D HSQC experiment may not be able to resolve the rates of exchange for every site of a large protein; however, there are resolved peaks that may be monitored. To do this, one must distinguish resolved from overlapped and unresolvable peaks. The task of accurately mapping known assignments onto a highly crowded spectrum presents unique challenges. The number of expected peaks (334 for MBP from the BMRB, entry 4986, (161)) outnumbers observed peaks found in the experimental spectra because in the observed spectrum, a single degenerate peak results from the presence of multiple overlapped peaks. Peaks that result from overlapped and unresolved peaks are difficult to manually detect and complicate matching the observed signals with known assignments.

The parity mismatch between the number of expected signals and the number of observed signals requires, at the first level, that we determine which peaks, defined by chemical shifts deposited in the BMRB, are expected to be resolved at the field strength and experimental conditions being employed, ultimately this helps avoid misinformation. Once this has been done, the next challenge is to assign the observed peaks to one or more peaks in the resolved BMRB spectrum in a consistent and unbiased way. Ultimately, we must determine which of the observed peaks are unambiguously assignable to only one entry in the BMRB as these peaks are followed during the NHX experiment. We also need to determine which of the observed peaks are not matched directly with one BMRB entry, these come in two flavors. First, there are those observed peaks that are assigned to an overlapped multi-peak in the BMRB. There are also observed peaks that may not be overlapped in reality but that match to more than one entry in the BMRB. These peaks are flagged because they often are incorrectly assigned and ambiguous. This algorithm is explained in the following discussion.

Defining peaks that are expected to be unresolved in the BMRB at the field strength and solution conditions employed for the HX experiment involves first collecting a reference HSQC spectrum such as is shown in Figure 3.1. Most of the available NMR visualization packages provide automated peak picking routines, the resolution determination algorithm described here is written in python to interface with the Sparky NMR program. Sparky provides an automated peak picking routine to determine peak positions and their respective line widths at half height. This information along with resonance assignments from the BMRB are used as inputs to the algorithm here.

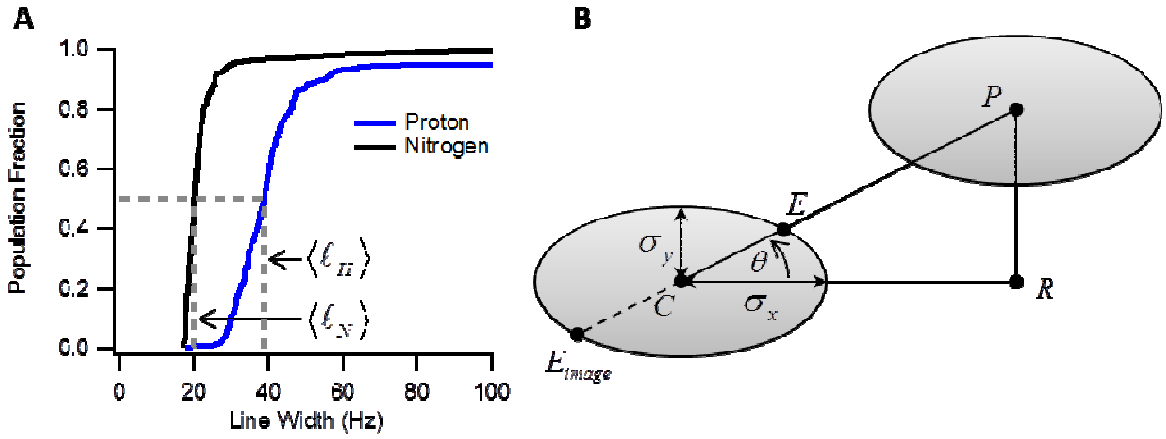

Figure 3.4: Resolution determination. **(A)** The line widths from an observed spectrum are plotted as cumulative distributions to determine the expected line width of an observed peak in each experimental dimension. **(B)** A diagram to accompany Eq. 3.2-Eq. 3.5.

Using the accepted notion that two Gaussian distributions are resolved if they are separated by more than the sum of their individual standard deviations to measure peak resolution, the algorithm determines the expected Gaussian standard deviations of the observed peaks along each experimental dimension using information from the line widths at half maximum in the observed spectrum. By plotting line-width information acquired from a reference HSQC (Figure 3.1 for MBP) as a cumulative distribution we can clearly see in Figure 3.4A, that these distributions are narrow and asymmetric. The skew in these distributions occurs because of peak overlap. To avoid bias in the arithmetic mean of non-normal distributions, the median of the distribution is taken as the expected line width $\langle \ell \rangle$ in each dimension. Line widths are related to the Gaussian standard deviation by

$$\sigma = \frac{\langle \ell \rangle}{2\sqrt{2\ln 2}}. \qquad \text{Eq. 3.1}$$

The algorithm iterates through the BMRB peaks and for each, performs a binary comparison with all other BMRB peaks to determine whether the peak should be resolved from the others in a 2D HSQC given the operational conditions of the experiment. Figure 3.4B shows a sketch of the quantities used in this process to accompany Eq. 3.2-Eq. 3.5 below.

For each pair of peaks in the BMRB, we are given two points, C and P, representing the mid-points of the peaks. The point R is then determined,

$$R = \left(P_x, C_y\right), \qquad \text{Eq. 3.2}$$

allowing us to define the angle between the two peaks,

$$\theta = \cos^{-1}\left(\frac{\overrightarrow{CP}\square\overrightarrow{CR}}{\left|\overrightarrow{CP}\right|\left|\overrightarrow{CR}\right|}\right). \qquad \text{Eq. 3.3}$$

We then use θ to determine a directional Gaussian standard deviation by defining the point E[8],

$$E = \left(\sigma_x \cos(\theta) + C_x, \sigma_y \sin(\theta) + C_y\right), \qquad \text{Eq. 3.4}$$

and then classify the peak C as resolved if the inequality,

$$\left|\overrightarrow{CP}\right| > 2\left|\overrightarrow{CE}\right|, \qquad \text{Eq. 3.5}$$

holds for all possible overlapping peaks. If the peak is not resolved, we create a multi-peak to represent the overlapped peak and write down the identities of each peak contained within the multi-peak, and compute a new midpoint by determining the multi-peak's center of mass.

[8] Depending on the direction of $\overrightarrow{CP}$, we may be defining location of the image of point E (the image of E is shown by the dashed line in figure 4B, however in the case shown, the actual point is determined). Due to symmetry and the purpose of the algorithm, whether we have defined E or its image is of no consequence.

To briefly summarize, we take the median spectral linewidth at half peak height (Figure 3.4A) in each dimension and convert it to a Gaussian standard deviation, $\langle \ell \rangle_x , \langle \ell \rangle_y \rightarrow \sigma_x , \sigma_y$ (Eq. 3.1). For each pair of peaks, we are able to determine a directional standard deviation, $\left| \overrightarrow{CE} \right|$ using Eq. 3.2-Eq. 3.4. Finally, we consider the peaks resolved if separated by greater than the sum of their standard deviations (Eq. 3.5). BMRB peaks that pass the test are considered resolvable using the instrument and conditions present during acquisition of the reference spectrum. BMRB peaks who fail the test are combined into a single overlapped multi-peak, the center of mass of each degenerate peak cluster defines a new peak position for each multi-peak.

With this catalogue of peaks who are resolvable, the next task is to assign these peaks with the observed peaks in our reference spectrum.

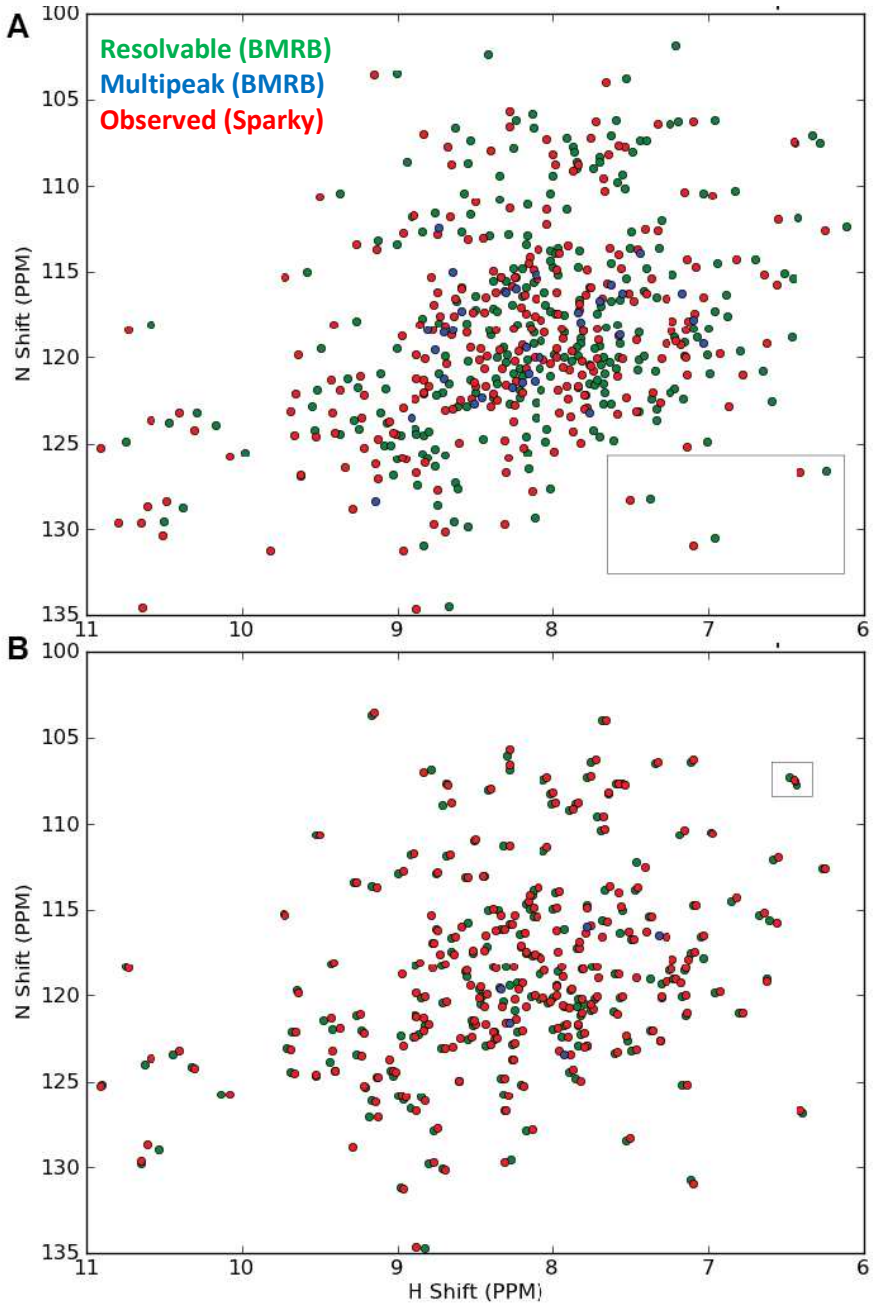

Figure 3.5: Peak assignment algorithm. **(A)** Before running the peak assignment algorithm, a non-uniform offset, highlighted, in the grey rectangle, is observed between peaks found in Figure 3.1 and those deposited in the BMRB (Entry 4986). There are 334 H-N assignments deposited in the BMRB (green), 264 of which are resolvable, 71 are unresolvable and result in 32 overlapped peaks, shown in blue. **(B)** Out of 253 observed peaks in panel A (red), 199 were matched to one or more peaks from the BMRB. Unambiguous assignments were found for 172 peaks, 20 peaks were found to match 2 entries in the BMRB (one example is shown in the gray rectangle), and 7 peaks were paired with 3 or more potential assignments.

Before the matching of resolvable BMRB chemical shifts with the observed peaks, one must determine expected offsets and the variance in offsets between the reference spectrum and the BMRB chemical shifts entries. This is easily visualized by plotting the sets of points together as has been done in Figure 3.5A, observed or reference peaks are red, resolvable BMRB peaks are green, and BMRB overlapped multi-peak centers of mass are blue. By inspection of Figure 3.5A, and in particular those peaks highlighted by the gray rectangle, one notices a variable offset between the peaks deposited in the BMRB with those found in the reference spectrum. The algorithm presents this information to the user in an interactive interface where peaks that are thought to be paired are selected, such as those in the gray rectangle. Once five or more pairs have been selected, the algorithm determines the mean offset and variance in each dimension to facilitate the matching process that follows.

To begin matching, the algorithm first shifts the observed spectrum by the mean offsets determined previously and then enters a series of N, usually 1000, iterations to match the two sets of points. For each point in the observed spectrum, at each iteration, all points in the BMRB set that are in the neighborhood of each observed peak are compared with one another. Using similar mathematics as in the first part of the algorithm (Eq. 3.2 and Eq. 3.3), the angle between an observed peak *C* and each potential candidate peak *P* is computed, θ_{CP}, along with the distance between the two peaks, $\left|\overrightarrow{CP}\right|$. For a given iteration, n_i, an acceptable matching radius between the two peaks is defined by

$$r(\theta_{CP}, n_i \mid N, \nu_x, \nu_y) = 3\frac{n_i}{N}\sqrt{\left(\sqrt{\nu_x}\cos(\theta_{CP})\right)^2 + \left(\sqrt{\nu_y}\sin(\theta_{CP})\right)^2}, \qquad \text{Eq. 3.6}$$

and an assignment is made between two peaks, A and B, if the distance between them is less than the matching radius, $\left|\overrightarrow{AB}\right| \le r(\theta_{AB}, n_i \mid N, \nu_x, \nu_y)$. Here, the variances to the offsets between the observed peaks and BMRB peaks are represented by ν_x & ν_y and the total number of iterations given by N. In the final iteration, the matching radius expands to three standard deviations from the mean offset in all directions. After evaluating all potential matches in a given iteration, unambiguous matches are removed. If an observed peak matches more than

one peak in the BMRB in a single iteration, the assignment is ambiguous and all matching points are removed from the spectrum.

Peaks that were matched to resolved peaks and multi-peaks from the BMRB are shown in Figure 3.5B. Out of 253 peaks in the observed spectrum, unambiguous assignments were found for 172 peaks; 27 peaks matched either a multi-peak or more than one peak in the BMRB and 54 observed peaks were not matched. Thus, with MBP in the 2D HSQC, only 51% of the assigned resonances can be followed with sufficient resolution. This problem will grow worse with protein size – however, the algorithm described here should allow one to proceed collecting the maximum amount of unambiguous information possible.

In cases where spectral crowding is not an issue, the algorithm allows one to match observed spectra to reference spectra in an automated fashion even when there is a non-uniform offset between the two. Previously, this was done manually and this may introduce a source of bias. Manual peak matching may often present ambiguous cases, which lead to incorrect assignment because of human inconsistencies. This algorithm provides a consistent means to match assignments quickly. It is also not limited to 2D experiments. With little effort, the algorithm may be modified to n-dimensional spaces. This formalism was not included because the equations would be redundant and the modifications are directly obvious.

3.2.2 pD 9.6 HX NMR Experiment on MBP.

Though there were many challenges that needed to be solved, using my peak matching algorithm (section 3.2.1), we were able to unambiguously assign 172 cross peaks distributed throughout the protein and proceeded with a NHX style experiment on MBP at pD 9.6. All work on MBP had been done at pH 7.5, but we needed to use higher pH to study the folding pathway (Chapter 5) and I wanted to verify that the structural dynamics at this higher pH were equivalent to pH 7.5. Qualitatively, we knew MBP had maintained native structure because the dispersion pattern at pH 9 matched pH 7.5; however, this would be true even if there were stability changes.

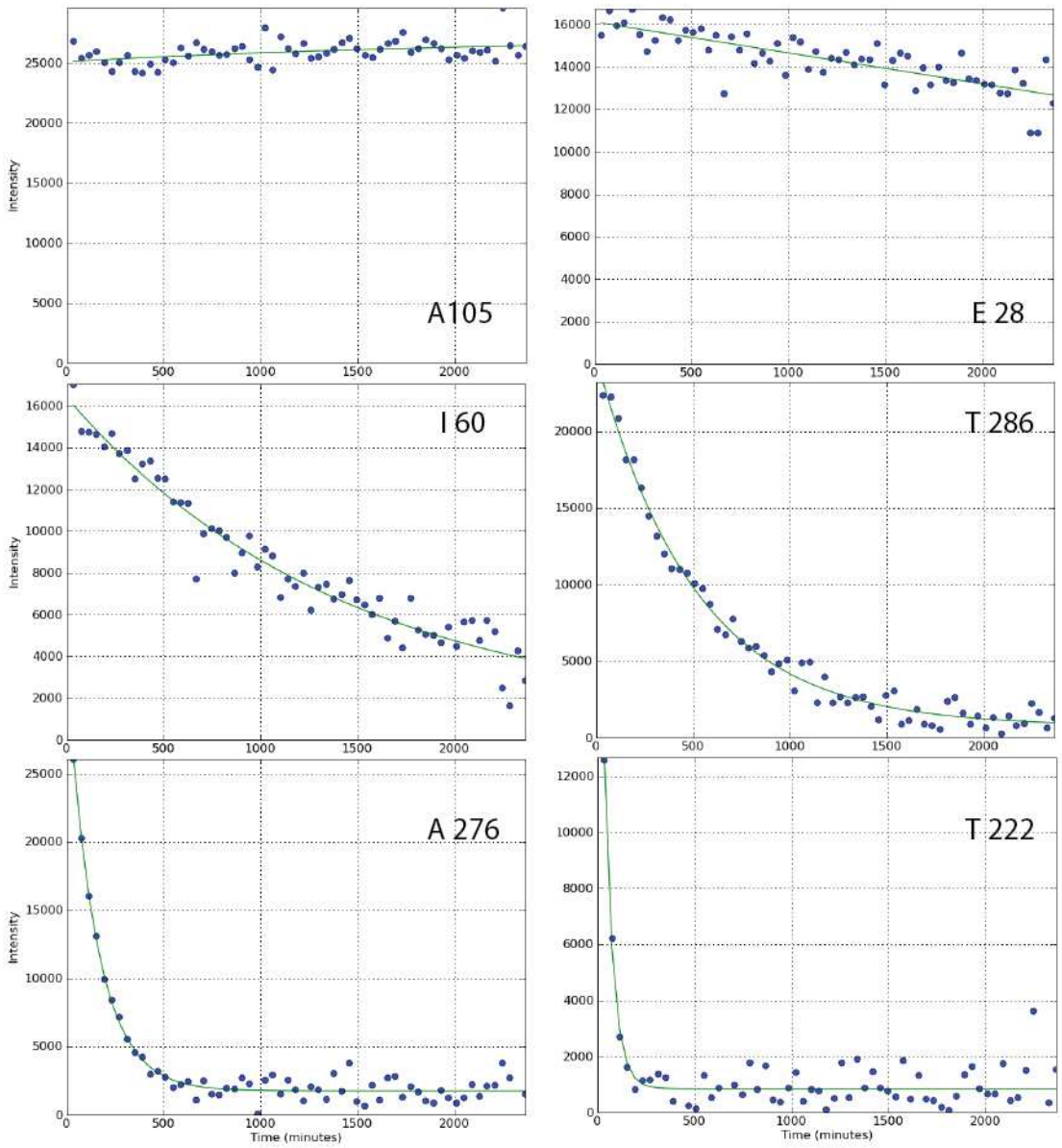

Figure 3.6: Six representative HX profiles observed in a pD 9.6 experiment. Many peaks do not exchange, these traces are represented in the top two panels. Of the 172 peaks with assignments, only 76 were followed through the experiment, many exchange too fast to measure. Blue dots are intensity values for consecutive HSQC's, the green traces show single exponential fits to the exchange profiles. The residue assignment is indicated in each panel.

At pH 7.5, the stability of MBP had been measured by multiple groups using a standard optical denaturant melt experiment (37, 162-165) and was found to be between 10-14 kcals/mol, depending on the denaturant used. Using Eq. 2.6 (page 17), I determined that at pD 9.6 and 310K, on average, an exchangeable site with 12.5 kcal/mol stability should have an HX lifetime of roughly 24 hours. With this in mind, I collected a series of standard HSQC

experiments, over a period of 24 hours with each HSQC requiring 23 minutes. Any site with a stability of less than 9-9.5 kcal/mol will completely exchange before the first time point.

Six representative decay traces are shown in Figure 3.6 from this experiment. At the first time point, only 76 of the 172 unambiguously assigned peaks remained. By the end of the 24-hour period, 21 peaks appeared to have not exchanged a single time, such as A105 in Figure 3.6. This implies a global stability in excess of 15 kcal/mol for the most stable sites. These measurements exceeded the estimated stability from optical experiments. This is not surprising, HX stability measurements often are larger than those gleaned from optical denaturant melts due to a flaw in the analysis of denaturant melts for non-two-state systems (166). This experiment was completed in 2009 – recently a group published data on MBP from HXNMR at pH 7.5 (167). All 15 cross peaks who did not appreciably exchange at pH 7.5 in their study were all among the 21 non-exchanging peaks in my data. Taken together, we conclude that the stability at pH 9 is similar to pH 7.5.

3.2.3 Concluding Remarks

This experiment at pD 9.6 did provide useful facts. In Chapter 5, we measure the stability of MBP by optical denaturation and find the stability to be much lower than this experiment indicated; reasons for this discrepancy are given in Chapter 5. We ultimately pursued kinetic folding experiments at pH/pD 9.0. This experiment confirmed that the native state of MBP at elevated pH was similar in stability to that of pH 7.5. We reasoned from this that information gleaned from folding at pH 9.0 would be relevant to the majority of MBP folding studies conducted at pH 7.5.

3.3 AMORE-HX: a multidimensional optimization of radial enhanced NMR-sampled hydrogen exchange

This section is the result of a co-first-author (1) publication, which resulted from a collaboration with John M. Gledhill Jr. of the laboratory of Joshua Wand.

3.3.1 Introduction

In theory, increasing to a three-dimensional experiment could overcome resolution issues which grow with protein size in two-dimensional experiments such as the HSQC. Figure

3.2 exemplifies the additional resolution from the third dimension. Peaks corresponding to residues T53 and G327 from MBP are coincident in the HSQC (Figure 3.2A) and are resolved by the HNCO (Figure 3.2B). The traditional Cartesian sampled HNCO may generally resolve the spectral crowding issue in 2D NMR because of its third dimension; but, in requiring > 12 hours per spectrum, this experiment simply lacks the necessary time resolution required for hydrogen exchange measurements.

Sparse sampling techniques such as radial sampling (168, 169) with attendant processing schemes (170-175) combined with optimization of the longitudinal relaxation properties (153, 159, 176, 177) can be employed to reduce the time required per HNCO spectra and increase the time resolution for HX purposes. Transverse relaxation optimization reduces acquisition time by decreasing the delay between transient scans. Typically, the inter-scan delay is the longest delay during a pulse sequence, thus substantial time savings is achieved by reducing this delay. Alternatively, sparse sampling achieves a decrease in acquisition time by reducing the number of increments collected in indirect dimensions. Sparsely sampled data cannot be processed using traditional approaches and the resulting frequency domain spectra require special treatment for extraction of HX measurements. Our efforts here focus primarily on sparse sampling and show that radial sampling time can be further reduced with transverse relaxation optimization.

This 3D HX NMR experiment, termed a multidimensional optimization of radially enhanced NMR-based hydrogen exchange in proteins or AMORE-HX, exploits various features of longitudinal relaxation optimization and radial sampling to increase the time resolution, sensitivity, and accuracy of large protein HX data using the radial-HNCO experiment. The advantage of this method is seen directly in Figure 3.2C-D where two peaks which were overlapped in the two dimensional experiment are resolved and their exchange rates measured using the AMORE-HX experimental design and HX processing scheme described here.

3.3.2 Description of AMORE-HX

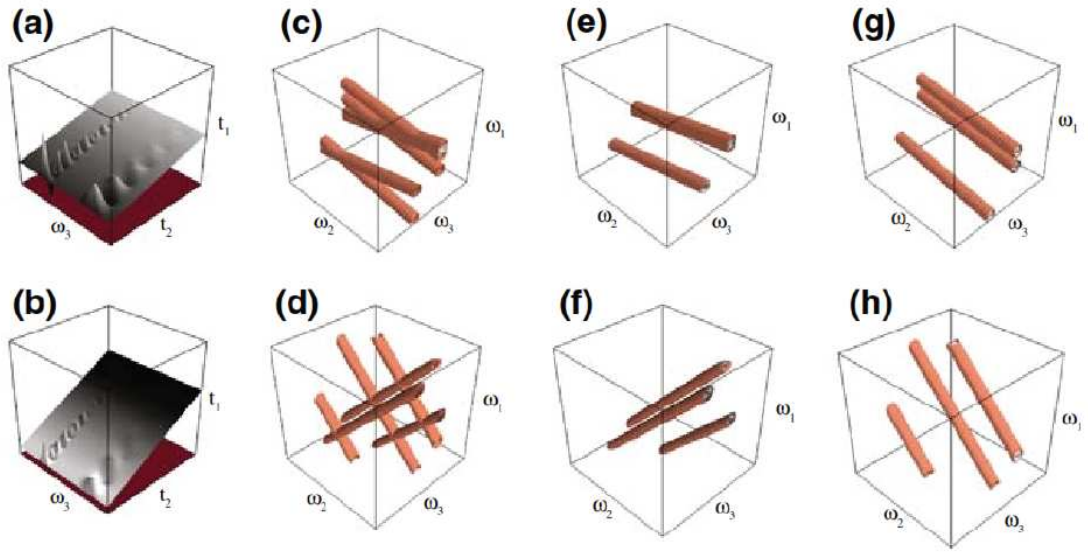

Figure 3.7: Graphic representation of the ridge artifacts from radial sampling. Radial interferrograms are simulated for three fictitious peaks after FFT of the direct ^{1}H dimension (ω_3) for 10° **(A)** and 45° **(B)** sampling angles. Ridge artifacts **(C,D)**, resulting from the simultaneous 2D-FT of the co-evolved dimensions, t_1 and t_2, are shown as cylinders in 3D frequency space; the separable positive **(E,F)** and negative components **(G,H)** are shown for angles 10° and 45°, respectively. Note the convention: $\omega_i = \mathrm{F}\,(\mathrm{t}_i)$, F is the Fourier transform operator. *Figure reprinted from (1).*

The radial HNCO (178) is accomplished by sampling the directly detected dimension (t_3, ω_3) normally and linking the indirect dimensions by $t_1 = \tau\cos(\alpha)$, $t_2 = \tau\sin(\alpha)$ and linearly sampling the time period, τ. The result of this linkage is shown Figure 3.7A-B for sampling angles $\alpha = 10^{\circ}$ (A) and $\alpha = 45^{\circ}$ (B). When processed by a true two-dimensional Fourier transformation (170, 171, 175), linkage of the indirect dimensions results in a fundamental artifact manifested as a ridge of intensity extending through the peak positions at the sampling angle $\pm\ 90^{\circ}$, this is illustrated in Figure 3.7C-D. The positive and negative components of each sampling angle may be separated (Figure 3.7E-H) using a sum and difference of matching and non-matching Fourier transforms (174).

A geometric method, described previously (179), may be used to determine which peaks are resolved by any given sampling angle provided the chemical shifts and line widths of all peaks are known. If two authentic peaks fall on the same ridge, the intensity will be the sum of the two (notice how there are three components in Figure 3.7G whereas coincident ridges sum

leaving only two components in Figure 3.7E). Consequentially, not all chemical shifts will be resolved for any given sampling angle. The selection of sampling angles is perhaps one of the most important aspects of AMORE-HX. Each angle requires a fixed amount of acquisition time and one must collect multiple angles to resolve all peaks.

Selection of Sampling Angles

It was recently shown how one could determine which peaks are resolved by any given sampling angle (179). This work also provided a best-first-sorting algorithm to select a set of sampling angles sufficient to resolve all peaks in the Cartesian HNCO. The best sampling angle by best-first-sorting is the angle that resolves the largest number of peaks. The best angle is selected and all peaks resolved by the best angle are removed from the lists of resolved peaks for all other angles. This task proceeds iteratively, each time selecting the best angle, until all peaks are resolved by at least one angle. This may result in more angles than are necessary to resolve all peaks and thus reduces the experimental time resolution.

To illustrate the weakness of the best-first-sorting approach, suppose we have a fictitious spectrum with six peaks, *1-6*, and three potential sampling angles, *A, B,* & *C*. The peaks resolved by each angle are computed and stored as lists, one list for each angle. Suppose angle *A* resolves peaks *1-4*; angle *B* resolves peaks *2, 3,* and *5*; and angle *C* resolves *1, 4,* and *6*. The best first sorting algorithm first selects angle *A* because it resolves four peaks; however, by selecting angle *A* first, both of the remaining angles are needed to resolve all peaks. Had we instead chosen either angle *B or C* first, we would have then been able to resolve all peaks by angles *B and C*. The severity of this problem scales with the number of peaks and therefore with protein size. This problem with best-first-sorting motivated a new angle selection algorithm described below.

The optimal set of angles is defined as the minimal set of angles that together will resolve all peaks. To find this set we use a combination of Boolean logic and linear algebra and exploit the fact that some peaks are resolved by a large number of angles while other peaks are resolved by relatively few. Once the peaks resolved by each potential sampling angle have been catalogued, the information is tabulated as a $M \times N$ Boolean array, $\mathbf{A}$, where N is the total number of cross peaks and M is the total number of angles to choose from. For any given

angle/peak combination, $\mathbf{a}_{n,m}$ is set to true (1) if peak n is resolved by angle m and set to false (0) otherwise.

To generate a measure of resolution for each peak we define a count matrix $\mathbf{C}$ as the inner product of a unity row matrix of length M and the Boolean resolution data M x N matrix $\mathbf{A}$

$$\mathbf{C} = \left[1_1, k, 1_M\right] \cdot \mathbf{A} . \qquad \text{Eq. 3.7}$$

The vector $\mathbf{C}$ contains an inventory of the number of times a peak is resolved for each radial sampling angle. We define a weight function to facilitate choosing angles based on uniqueness,

$$\omega(\mathbf{C}) = 1 - \frac{\mathbf{C}}{\max(\mathbf{C})} . \qquad \text{Eq. 3.8}$$

The largest element of $\mathbf{C}$ is given by $\max\left(\mathbf{C}\right)$. The elements of the weight function $\omega\left(\mathbf{C}\right)$ approach unity when a peak is only resolved by a few angles – this weighting quantitates which peaks are difficult to resolve and which angles resolve these peaks. The final step is to compute the measure function by taking the inner product between the Boolean matrix $\mathbf{A}$ and the transposed weight function, Eq. 3.8,

$$\mu\left(\mathbf{C}^T\right) = \mathbf{A} \cdot \omega(\mathbf{C})^T . \qquad \text{Eq. 3.9}$$

Each element in $\mu\left(\mathbf{C}^T\right)$ corresponds to the sum of the now weighted terms in $\mathbf{A}$, the angle identified by $\max\left(\mu\left(\mathbf{C}\right)\right)$ is selected. Subsequent angles are selected after removing all peaks resolved by the selected angle from $\mathbf{A}$ and the process is repeated.

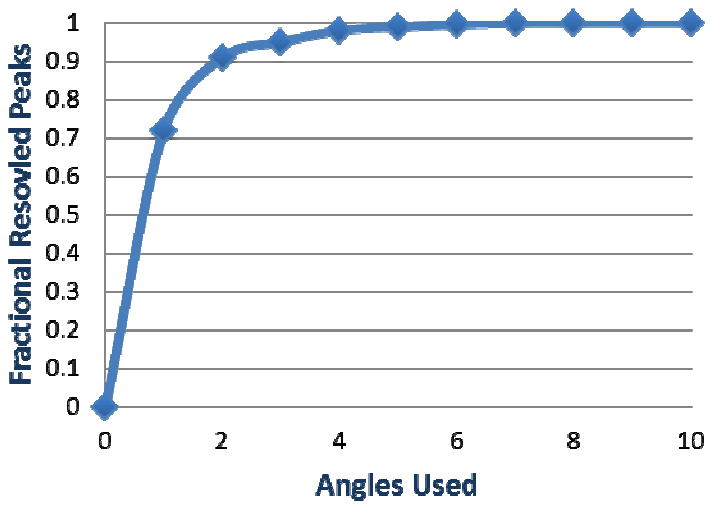

Figure 3.8: Performance of the minimal angle selection algorithm. Fraction of peaks resolved by including an additional *best* angle using the angle selection algorithm described herein.

In a pilot experiment, Figure 3.8 shows the results of the selection algorithm described here. Out of 290 identified peaks in the Cartesian HNCO, 284 peaks are resolved using only four angles. For comparison, the top four angles using the best-first-sorting algorithm resolve 37 fewer peaks. Additional angles give diminishing returns; a balance between peak and time resolution may be necessary. In order to resolve all 290 peaks, seven angles would be required. We settled on the collection of four angles by the AMORE-HX method for MBP as this requires approximately 43 minutes of acquisition time and this is comparable to the time required for a single 2D HSQC.

AMORE-HX Processing Scheme

Individual HNCO experiments, one for each sampling angle, are collected sequentially from the first angle selected to the last. A single time point is composed of one spectrum for each angle, or four radial experiments. As such, there is a substantially more data collected in an AMORE-HX experiment when compared to the traditional 2D HX NMR method.

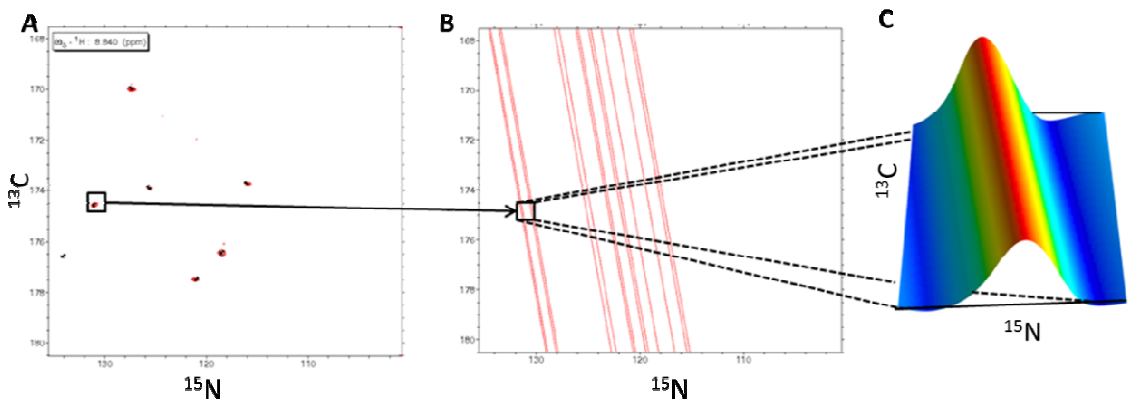

Figure 3.9: AMORE-HX sub-windows. **(A)** The Cartesian sampled spectrum and **(B)** corresponding negative ridge component spectrum for $\alpha = 81^{\circ}$. Frequency pairs are transformed using the 2D-FT to provide a **(C)** sub-window containing the peak intensity information.

Typically, one would process the entire frequency range defined by the sampling increment; however, this is unnecessary in AMORE-HX. In the context of large protein hydrogen exchange experiments one is only interested in determining the rate of exchange, k_{ex}, for each amide proton by accurately measuring the change in peak intensity as a function of time. In contrast to FFT, the direct 2-D FT can process a limited frequency range without artifact. As the chemical shifts of all peaks are known, only the limited region (sub-window, Figure 3.9C) of the indirect plane corresponding to the center of the peak needs to be processed. Figure 3.9 illustrates the difference between Cartesian (A) and radial sampling (B) .

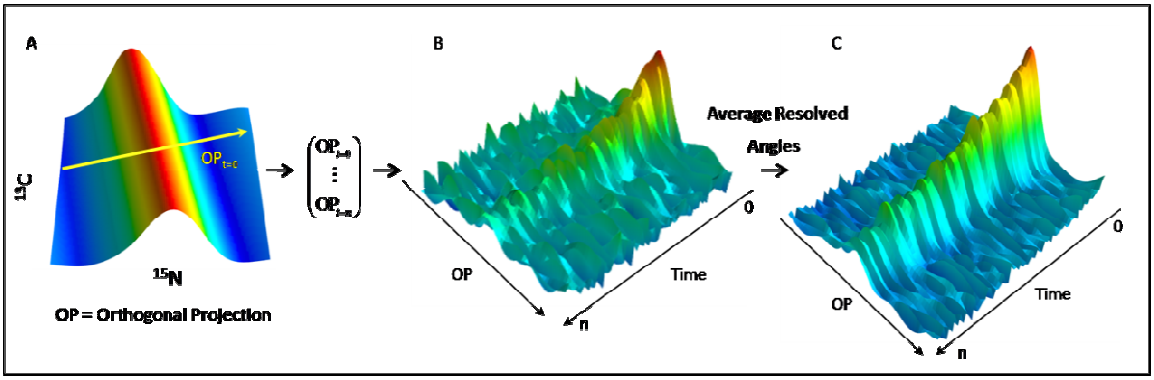

Figure 3.10: AMORE-HX Orthogonal vectors and the average advantage. **(A)** an orthogonal projection is taken for each sub-window By interpolating across the intensity matrix orthogonal to the sample angle (-81°) a slice orthogonal to the ridge is extracted. **(B)** By stacking orthogonal projections from each time point, the decay in intensity of the peak caused by hydrogen exchange is shown for -81°. **(C)** Stacked plots for each sampling angle that resolve any particular peak may be averaged for increased precision in determining the HX rate, this is the average of 4 angles.

Ultimately, sub-windows are not required; they provide a means to inspect the data. Orthogonal vectors (projections in Figure 3.10) may be computed directly using the 2D FT by supplying the corresponding paired frequency components. The resulting one-dimensional vector provides a cross-section of the peak. This vector is stacked with respect to time and an example of this is shown for peak D197 in MBP in Figure 3.10. By fitting a single exponential to the maximum intensity of each orthogonal vector or its area with respect to time, the exchange rate is obtained. If the decay rate is sufficiently slow, multiple angles that resolve the peak may be averaged. Averaging angles will reduce the time resolution; however, when appropriate, the advantage of averaging can be seen clearly in the observable S/N increase between the single angle in Figure 3.10B and the average result in Figure 3.10C. Notably, the S/N is much better in the HSQC than in the HNCO and averaging may be necessary in some cases.

3.3.3 Over-pulsing effect is minimized using shaped excitation pulses

The NMR measurement is sensitive to the effects on the relaxation properties of any given amide hydrogen by exchange of a neighboring amide proton for deuterium. This effect manifests in both the longitudinal (T_1) and transverse (T_2) relaxation rates; the latter of which may be overcome under opportune circumstances by measuring cross peak integrals as discussed earlier. The effect on amide proton T_1 may not be easily overcome; it results from non-equilibrium effects introduced by not allowing angular momentum vectors to relax and re-align with the bulk magnetic field in between pulses.

Inversion recovery measurements with the HSQC and HNCO using a pre-deuterated $^{15}N^{13}C$ MBP sample at HX equilibrium in either 10% or 90% D_2O were employed to assess the magnitude. Effective T_1 rates were approximated by fitting a single exponential to signal recovery curves (signal intensity versus recycle delay). The mean percent difference[9] in T_1 lifetimes for HSQC measurements was -7 ± 6% indicating a small but significant effect. With $^{13}C^{15}N$ MBP sampled by the Cartesian HNCO we found a much larger deviation of -35 ± 40%.

Since the majority of time during an NMR experiment is spent between pulses, NMR practitioners generally try to reduce this inter-scan delay as much as possible; as a result, inter-

[9] $\frac{T_1(10\%D_2O) - T_1(90\%D_2O)}{T_1(10\%D_2O)} \cdot 100$

scan delays and T_1 lifetimes are generally quite similar because inter scan delay duration is usually minimized using fully protonated samples. Differences as small as -7% observed in the HSQC experiment will introduce a second order error in HX measurements and artificially increase apparent exchange rates non-uniformly throughout the protein. If these effects are present and ignored, HX by NMR is more detrimental to scientific advancement than it is useful. *No data is better than bad data.*

We found selective excitation pulses used in the BEST approach (153-155, 159) to be particularly useful at mitigating this effect. The same measurements, using the BEST-HNCO (Cartesian-HNCO with selective excitation pulses), gave a negligible mean fractional deviation of +0.06 ± 0.24 %. By selectively exciting only the protons of interest, T1 relaxation rates are much faster; therefore, modest slowing of any individual rate has no effect because the inter-scan delay greatly exceeds relaxation lifetimes.

3.3.4 Summary of AMORE-HX

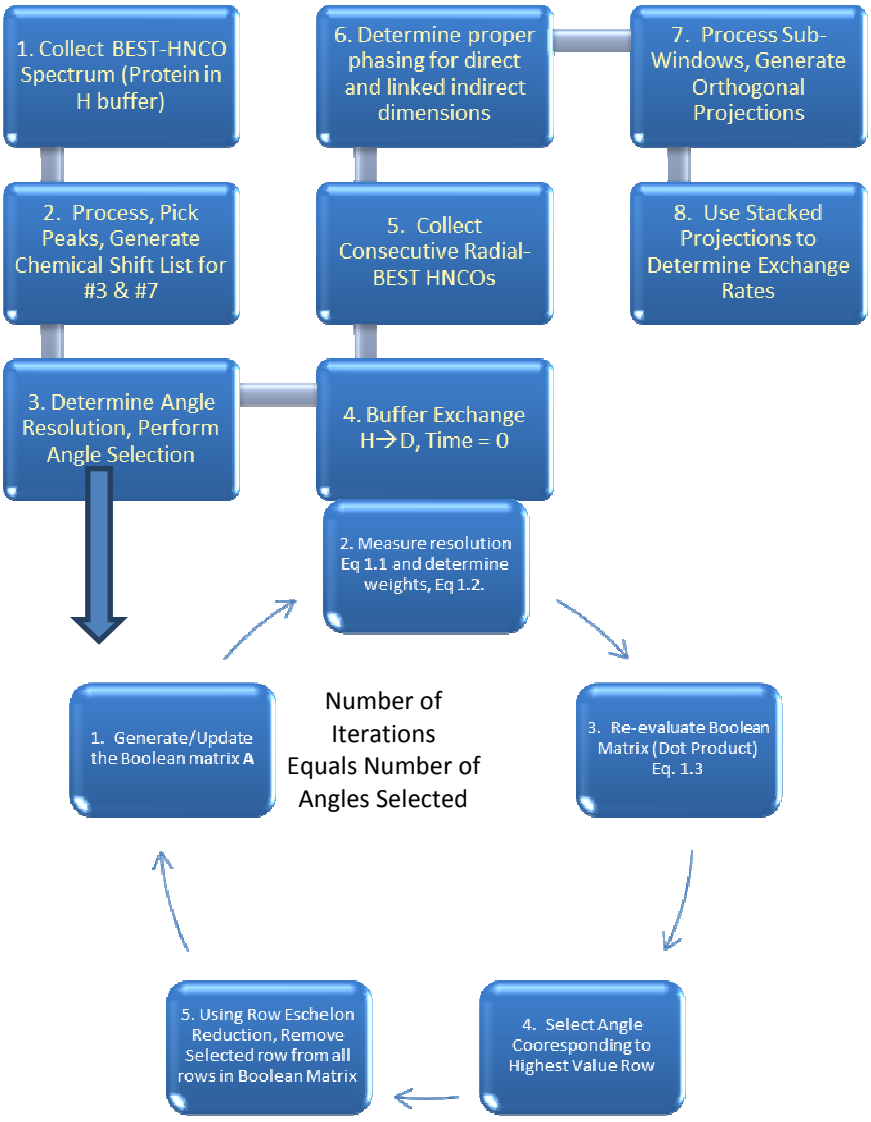

Figure 3.11: A work-flow diagram for the AMORE-HX experiment.

The overall workflow for AMORE-HX is shown in Figure 3.. The first step here is similar to the standard HX NMR HSQC procedure, a reference spectrum in protic buffer (HNCO) allows determination of the coordinates for each cross peak and downstream matching of assignments to observed cross-peaks. Once the reference has been collected, one must select the

appropriate angles for the experiment. All other differences between the AMORE-HX protocol and the traditional HSQC based methods are kept hidden by way of a computer program that I wrote to automate this process and described earlier. The product of the two methods is the same, decay traces that allow determination of the exchange rate. This was done on purpose with the hopes of making this complicated experiment accessible to researchers who have not traditionally focused on NMR.

My computer program for the AMORE-HX experiment called RidgeGlider interfaced with a pythonic version of Al-NMR (180), written by Dr. John Gledhill Jr. for phasing and transforming the sparsly-sampled data (2D FT) into the frequency domain. Without Al-NMR, this project would not have been possible.

3.4 Concluding Remarks

HX experiments traditionally measured by NMR are becoming increasingly rare. Researchers seem to be favoring mass spectrometry for HX measurements as assessed by the volume of publications. This is not surprising when one considers the various issues in HX NMR discussed in section 3.1, page 25. HX MS experiments tend to be easier to conduct, require far less protein, and are much cheaper as the protein does not need to be isotopically labeled for analysis. These reasons and others are explored in detail in Chapter 4.

Chapter 4 - Hydrogen Exchange by Mass Spectrometry

4.1 A Historical Context for HX MS Experiments

The first method proposed generally for HX experiments involved the use of a tritium tracer (181) and provided exchange information on a whole molecule level. Pioneering work by Walter Englander (182-185) and Rosa and Richards (186) were able to provide sub-global structural information using the method of fragmentation-separation. For the first time, the authors showed that HX could provide structurally resolved information on conformational dynamics and their method remains popular in HX MS experiments to this day. In fragmentation-separation experiments, following HX labeling, the exchange reaction is quenched[10] by reduction of temperature and pH (5+ orders of magnitude) allowing time to prepare the sample for analysis without extensive loss of the experimental label, known as back exchange. The quenched sample is proteolyzed by the addition of acid proteases (pepsin, generally) and fragments separated using LC. In its original form, the extent of exchange was monitored by scintillation counting of LC fractions and fragment identities obtained from amino acid hydrolysate analysis. This process was very time consuming, difficult to perform, and generally suffered from low sequence coverage.

HX NMR was developing (187) and was first shown to provide exchange information on each amino acid in the primary sequence of cytochrome C by Josh Wand in 1986 (188). For more than a decade, NMR became the preferred method for HX measurements due to this site-resolution capability.

Despite its popularity, HX NMR has serious problems that are exacerbated by larger protein systems. Protein solubility remains a challenge because of the high concentration of protein required. One also needs to resolve and assign each amide hydrogen in the spectrum; a task that proves to be exceptionally difficult in larger molecules (see page 30 for discussion). Perhaps the biggest issue for HX NMR lies in its inherent protein size limitation. Slow molecular

[10] It is common to refer to the low pH, low temperature condition as quenched. The reaction is still occurring albeit orders of magnitude slower than during HX labeling, roughly 1% per minute for a solvent exposed amide.

tumbling times for larger proteins relegate most biologically relevant molecules inaccessible to NMR measurement. Interest in HX MS stemmed from a need to circumvent these hurdles.

The first HX MS experiments were introduced in the early 90s (189, 190) using a methodology very similar to the earlier H-T fragmentation-separation experiments (182). Since then, HX MS experiments have become dominant in HX literature. Unlike NMR, the MS method has no size limitations making HX experiments possible for larger molecules; however, sequence resolution remains an issue. The first HX MS demonstration (190) found nine fragments in total reporting on less than 50% of the protein's primary sequence. Localization of exchange information, paramount to achieving structural resolution, in HX MS is limited to the size of each proteolytic fragment; this turned out to be a persistent issue.

The same problems of low sequence coverage and loss of information due to back-exchange that hindered the first HX MS experiments have remained burdensome for nearly two decades. Among other things, these problems are solved in the work described below.

4.2. The HX MS Experiment

4.2.1 Introduction

Hydrogen exchange measurements in large biologically important protein systems that are inaccessible to NMR may be routinely measured by the HX MS fragmentation-separation[11] method. Unlike the typical HX NMR experiment where real-time spectra are acquired while the sample is exchanging, measurements by MS involve a staged approach whereby labeling and measurement are decoupled in time. As in the earlier H-T exchange method, proteins are labeled using either NHX- or KHX-type experiments (Chapter 2, page 18), samples are taken in time before being partially quenched by a reduction of pH and temperature. This dramatically slows the exchange reaction facilitating time to prepare the sample for MS measurement. The quenched sample is proteolytically digested (fragmentation) and run through HPLC separation. The LC eluent is ionized by electro-spray and injected into the mass spectrometer to measure the number of incorporated deuterium in each peptide fragment.

[11] HX MS experiments described in this dissertation are always conducted using the method of fragmentation-separation unless otherwise stated.

One major problem in current HX MS technology is sequence coverage, peptide fragments often wind over sizeable[12] regions of the protein sequence and generally represent less than the whole sequence. Where the NMR measurement provides direct site-resolution, MS resolution is limited to the size of each fragment. Changes in the HX rate of a whole peptide cannot distinguish precisely which residues within the peptide have changed. When a change is detected, one cannot determine whether it involves large changes to a few residues or small changes to all residues. This hinders interpretation.

Collision induced dissociation (CID) of peptide fragments in-flight is a popular way to determine the sequence of a peptide and was initially explored as a possible strategy to help localize the label and achieve site resolution. Unfortunately, extreme bond vibrational energies caused by CID result in a redistribution or randomization of the label within each peptide, this is known as scrambling (191). Non-ergodic methods such as electron transfer or electron capture dissociation have been shown to avoid the scrambling problem (192-196) and have been able to provide site-resolution in a limited number of cases.

Alternatively, through comparison of many overlapping peptides, one may be able to infer where in the peptide the label has gone, this is equivalent to site-resolution; the challenge becomes one of identifying and accurately measuring as many peptides as possible. In a group publication (4) where I was a supporting author, we describe a home-built system capable of producing, on average, 10 unique peptides per residue in the native protein for HX MS experiments. This section summarizes those endeavors.

4.2.2 Measuring HX Labeling by LC ESI MS

A labeled and quenched sample must be prepared for mass measurement; broadly, this involves proteolytic digestion to produce peptide fragments, washing of the fragments to remove molecules that would interfere with mass measurement, and separation of the fragments by HPLC. During preparation, back-exchange occurs and degrades the labeling pattern; it is necessary to move quickly through these steps. For this purpose, we developed a cold-flow online system to facilitate digestion, washing, and peptide separation all within a

[12] The average peptide length for HX MS MBP experiments is 12 residues.

closed apparatus that is temperature controlled at or slightly below 0° C. The online system and instrumentation employed are described below.

The On-line System

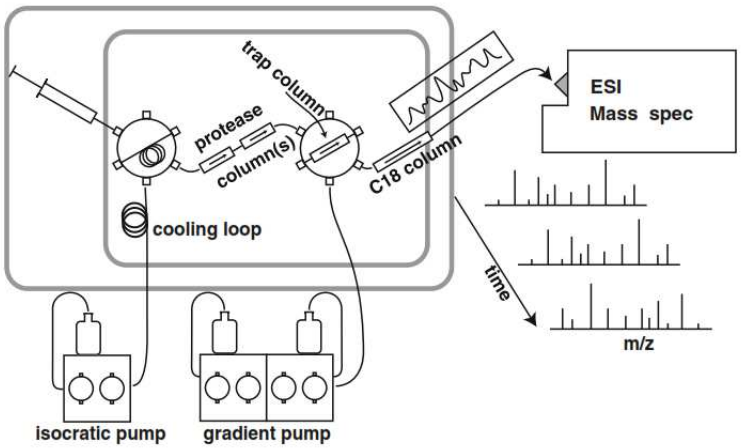

Figure 4.1: On-line HX MS analysis system *(reprinted from reference (4)).* The entire flow system is contained within a Peltier cooled chamber (21x15x25 cm; from an automobile accessory supplier) that maintains 0±1 °C, monitored with a thermocouple thermometer. An internal fan circulates air across the Peltier element and through the chamber. Liquid flow is precooled in a large loop positioned in the airflow. Switching valves are mounted on a ridged board backed with foam insulation with valve handles outside easily accessible for manual operation. The wetted parts of the valves, the injection loop, and columns are contained in the cold chamber. Liquid lines and the thermometer leads are threaded through small holes in the insulation and mounting board. A short length of outflow tubing from the C_{18} analytical column to the MS electrospray source is packed with ice-filled plastic bags.

Illustrated in Figure 4.1, a Peltier cooled chamber contains all components of the online flow system. The precise details of injection volumes and flow rates depend on factors that may differ depending on the protein and type of HX experiment being conducted. For KHX experiments with MBP, the injection loop is loaded with 300 µl of 166 nM MBP in the acid quenched and partially unfolded (1M GdmCl) condition. Flipping of the first valve diverts pre-cooled washing buffer (buffer compositions are described in section 4.3), flowing at 100 µl/minute into the injection loop and passes the sample through protease columns for digestion. Peptide fragments are trapped as they exit the protease column by flowing through a small C_8 column (1 x 5 mm, 5 µm beads). After three minutes, the flow rate from the isocratic pump is doubled and the sample is washed for two minutes to prepare for HPLC separation. Five minutes after injecting the sample, the second switch is flipped, a low volume HPLC pump flows through

the trap and then onto a C_{18} analytical column (0.3 x 50mm, 3 µm beads) for a rough peptide separation. Peptides are directed into the instrument as they elute from the column. To minimize eluent overlap and achieve constant numbers of peptides per unit time, we employ a non-linear 10-15 minute water/acetonitrile gradient (8 µl/min, 10-50% AcCN), gradient shaping is described in Appendix B.

A cleaning gradient is applied between each run consisting of 2-3 sequential 0-100-0% AcCN up-downs, 3 minutes each, to maintain low column backpressure and elute very large peptides that remain bound to the analytical column. To avoid the problem of peptide carry-over (197) the protease column is washed using two injections of 1M GdmCl at low pH while the trap and analytical columns are being washed. Blank injections containing no protein sample are run periodically to check for carry-over. In total, each run requires 20 minutes for HX MS preparation and measurement followed by a 10-15 minute cleaning cycle and a final 5 minutes for pressure re-equilibration before the next run.

Proteolysis

Acid proteases are used to digest the protein samples. Though this may be done in free-solution, passage through an immobilized protease column has proven to be more effective (198, 199). Often, protease digestion may be improved by protein unfolding by the addition of chemical denaturants (2 M Urea, 1 M GdmCl, or 0.5 M GdmSCN) and TCEP to reduce disulphide bonds (200), if needed. Commercial protease columns are available; however, compelling evidence suggests that these columns may promote back-exchange (198). We prefer using 2mm x 20mm guard columns packed with POROS AL beads to which are ligated either pepsin or fungal protease XIII following manufacturer instructions. The protease is gel-filtered prior to ligation removing any contaminating amines that may be present in buffers used to commercially purify the enzyme as they will compete for binding sites on the POROS AL. In literature, it is common to perform the coupling reaction in the cold for an unknown reason; however, the manufacturer instructions clearly state that the coupling reaction must be done at room temperature. Sodium sulphate is used during coupling for salting out the protein, this increases the density of ligated molecules on the surface of the bead – ligating in the cold is inefficient because of reduced Na_2SO_4 solubility.

Using the previously stated flow rates, transit time on the protease column ranges from 3 to 35 seconds. When necessary, increasing or decreasing the flow rate during digestion can be useful to bias peptide size distributions towards smaller or larger fragments, respectively. In an effort to discover the limits of our system, we have used pepsin and fungal protease XIII alone and in tandem for three separate injections. In section 4.3 and later in Chapter 5, only pepsin is employed as this condition yields an abundance of overlapping peptides sufficient for the level of sequence resolution needed in each case. This is the cause for differences in the reported number of peptides in different sections of this dissertation.

Instrumentation

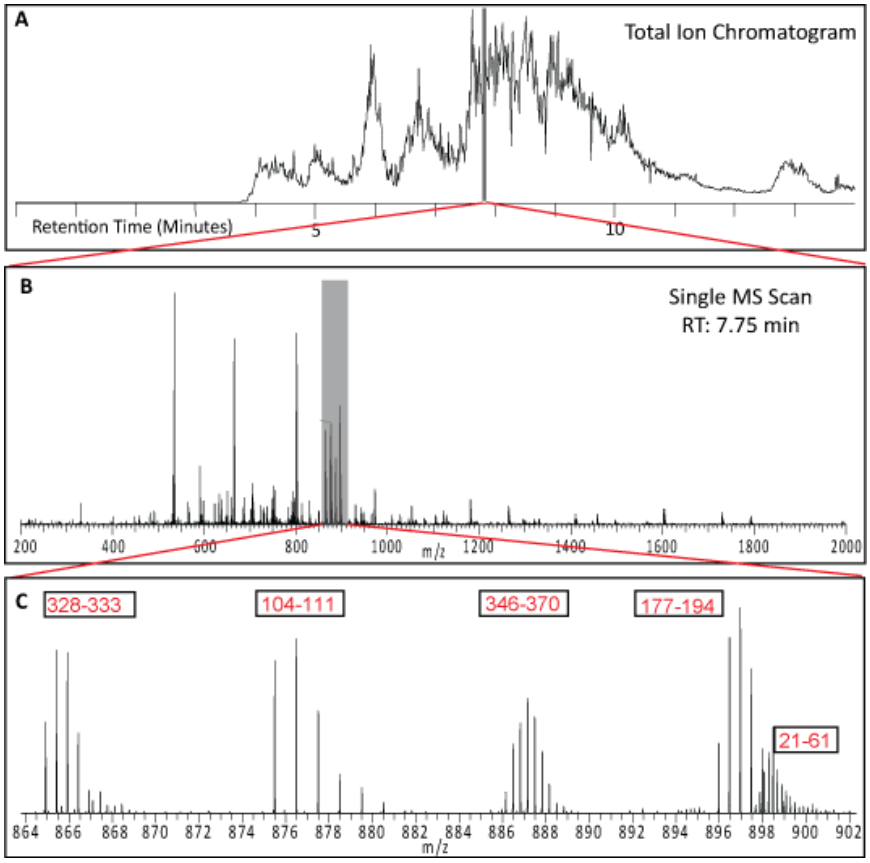

Figure 4.2: Illustration of the many peptides available and the requirement for high resolution instrumentation. **(A)** The retention time chromatogram for a 15 minute linear 10-40% AcCN gradient on peptic fragments from MBP. **(B)** A MS scan taken at 7.75 minutes in panel A. **(C)** Five co-eluting peptides in MBP observed within a small mass range from panel B (indicated by the gray shading in B).

High-resolution MS instruments are necessary to obtain large numbers of peptides in the experiment. Figure 4.2 illustrates this point using data taken on MBP. Figure 4.2A shows the

full retention time chromatogram where the sum intensity of the ions being injected is plotted on the y-axis. Inspection of a single scan (Figure 4.2B) taken at 7.75 minutes retention time demonstrates the presence of many peptides eluting simultaneously. Frequently, peptide mass distributions overlap with one another and may reduce the number of useful peptides for HX. The density of peptides is highlighted in Figure 4.2C by zooming in on the shaded region in Figure 4.2B. Here, five peptides are shown to occupy only 2% of the total mass-to-charge axis per scan (200-2000 m/z). Were it not for the high resolution (60,000 at 1 s/scan) of our ThermoScientific LTQ Orbitrap XL, we would have no hope of resolving 177-194 from 21-61 and would be unable to use the data shown for either peptide as they contaminate each other.

4.2.3 A Method to Obtain Many Overlapping Peptides for HX MS Experiments

Peptides are not fragmented by CID in the HX MS experiment. To identify deuterated peptides, a peptide pool database is first constructed containing the retention times, monoisotopic masses, and sequences for all peptides that may be present in the labeling experiment. Building the database consists of running data-dependent MS/MS on fully protonated samples and identifying precursor ions from daughter CID fragmentation spectra with the SEQUEST algorithm (201). The identifications are evaluated using an independently calibrated quality score to reduce the false discovery rate to less than 0.1%. Our in-house program, ExMS, described elsewhere (202), is then used to further validate these peptides under conditions that mimic HX MS experiments with partially deuterated samples.

Peptide Identification & The Peptide Pool

To construct the peptide pool, unlabeled samples are prepared for mass analysis using the online system (page 52) and conditions to be employed in downstream labeling experiments. We use tandem MS in data-dependent acquisition mode (DDA) for mass analysis to generate fragment spectra that are subsequently used for peptide identification. In DDA mode, relative intensities of eluting ions are determined by full MS (parent) scans using the orbitrap. In each parent scan, the masses of the four most abundant ions that are not on a dynamic exclusion list are sent to the LTQ for CID fragmentation and subsequent measurement. Ions selected for fragmentation are added to the dynamic exclusion list, each entry expires after 30 seconds.

In practice, we use three DDA MS/MS runs for each protease condition to build the peptide pool. Peptides identified by the SEQUEST algorithm with acceptable P_{pep} scores (described below) for each condition are added to a static exclusion list for use in subsequent MS/MS runs. The static exclusion list is operationally equivalent to the dynamic exclusion list except entries on the static list are only excluded for ± 1 minute to their respective retention times. This procedure ensures inclusion of lower abundance peptides in the peptide pool. One could continue beyond three runs but the effort produces diminishing returns.

We use SEQUEST (ThermoScientific Bioworks 3.3.1) for searching MS/MS results against a database containing the protein of interest and all possible contaminants. Briefly, the database contains the *E. coli* proteome (including MBP), entries for all proteins studied in this lab, a variety of human and dog[13] keratins, and the sequences of both pepsin and fungal protease XIII. We use a search tolerance of 4 ppm for parent ions and 0.1u for fragments and calibrate our instrument daily using positive-ion CalMix (ThermoScientific).

P_{pep} Calibration and Use

P_{pep} is a proprietary statistical goodness-of-fit parameter provided by the Bioworks software for each peptide identified by SEQUEST. Though it was designed apparently for a different purpose, we found it useful in eliminating false positive identifications from our MS/MS runs. To calibrate the P_{pep} score for our purpose, we constructed a decoy database by sequence reversal of the real database; this adds decoy entries and allows us to check for false positive identifications. MS/MS data were searched against both the real and decoy databases (203), any fragment identified in the decoy database is a false positive. Figure 4.3 shows the P_{pep} distribution of four proteins studied in the lab, each protein was searched separately and the results merged, blue data represents searching against the real database, red data are decoy results.

[13] Montana, a golden retriever, is a member of the lab and frequently stops by for a pat on the head while we work. Including dog keratins seemed logical although we have not identified dog or human keratins in our samples.

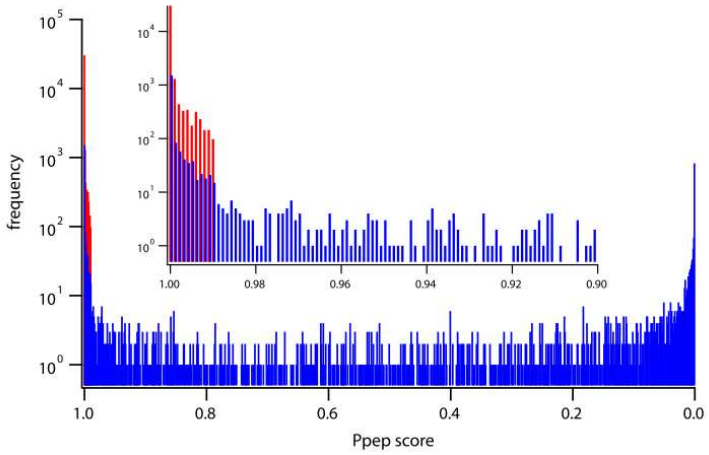

Figure 4.3: Distributions of Bioworks P_{pep} Scores *(reprinted from reference (4))*. Known false identifications (red) represent SEQUEST hits for MS/MS data run against a large database with reversed sequences. Data in blue represents hits on the experimental protein for MS/MS data run against the large database with all proteins in their forward sequences. Results for the four proteins studied are merged to provide a large statistical sampling. The inset focuses on the poorest scoring matches. These results show that the cutoff P_{pep} value necessary to reduce false identifications to <0.1% is 0.990.

By inspection of the data shown in Figure 4.3, it was determined that elimination of any peptide match with a P_{pep} score > 0.990 would reduce the number of false positive identifications to < 1/1000. Though P_{pep} is not a statistical p-value, it is similar; larger scores indicate lower confidence in the identification. We adopted a P_{pep} threshold of < 0.990 for our work. Elimination of all hits at or above this value reduced the number of identifications by 40%.

A secondary requirement to confirm the usefulness of these peptides for HX MS experiments was that they be found in 50% deuterated MS spectra by our ExMS program (for a thorough description of this program, see reference (202)). For each protease condition, we ran three injections through the online system that had been equilibrated in 50% D_2O. The ExMS program searches deuterated MS spectra using peptide pool information described earlier. Upon identification of a peptide, ExMS provides a selected ion chromatogram where sequential scans deemed acceptable are summed and the data for each peptide is exported for further analysis, or, in this case, manual verification. For each deuterated run, identified peptides were manually inspected – correctly identified peptides were added to our peptide inventory. The intersection of peptides for all three runs was taken as a measure of the useful identifications.

The ExMS program found nearly 100% of the respective peptide pools for all unlabeled samples, but a smaller fraction of peptide pool entries were identified in deuterated samples. The 50% deuterated condition is most challenging because isotope envelope widths and spectral overlap are maximized in this state. Searching this data provided a mechanism for us test and optimize user defined parameters in ExMS. Once optimized, ExMS was able to find and internally confirm 50 to 75% of the peptide pools in each deuterated run. Many of the remaining peptides could be manually verified upon inspection; however, ~20% of the pool was eliminated through this process.

Peptide Identifications and Further Validation

Useful peptides are those that met all selection criteria, they all have P_{pep} scores less than 0.990 and were found in all three 50% deuterated replicates for each protease condition. Collectively, the MBP peptide pool has 443 unique peptides and over 70% are found in more than one charge state.

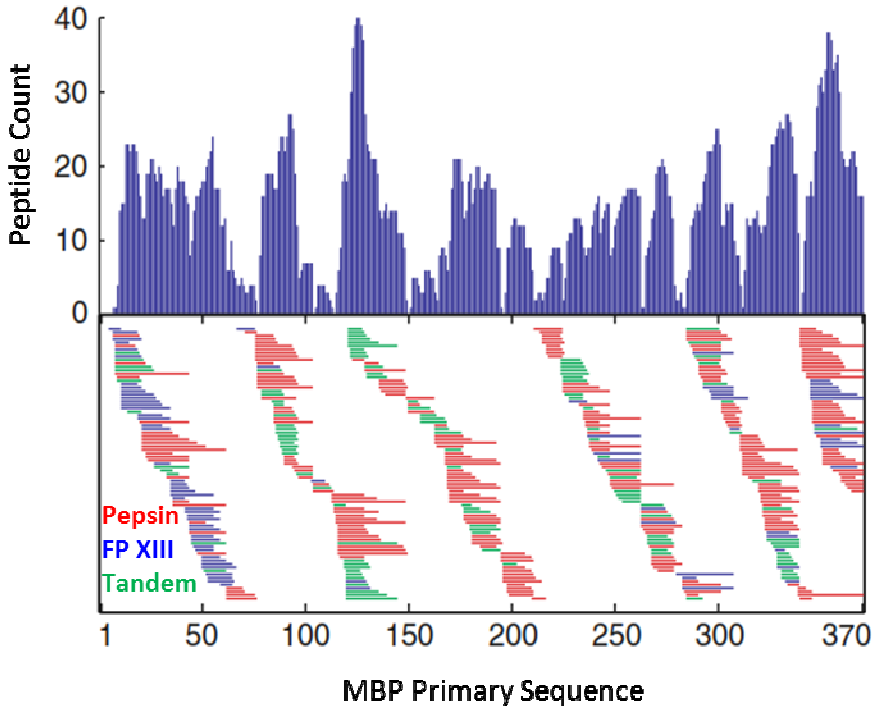

Figure 4.4: The Peptide Inventory for MBP. *(reprinted from reference (4))*. In the top panel, a histogram shows the number of peptides reporting on each of the 370 residues in MBP. The lower panel shows unique peptides found in each protease condition and the regions of sequence covered. Red peptides are found in the pepsin digest, blue peptides are found in the fungal protease XIII digest, and green are those peptides found in the tandem condition.

The inventory of useful peptides for MBP is shown in Figure 4.4. Notice that many of the peptides share a common N-terminus. This speaks to the validity of our method, these peptide groups occur at high probability cut sites. Upon checking for peptides produced by prohibited cut sites (204) (no Arg, His, Lys, or Pro at the P1 site) no violations were found for MBP. Over all four proteins used in the group analysis, only one violation was found at an Arg-Phe site.

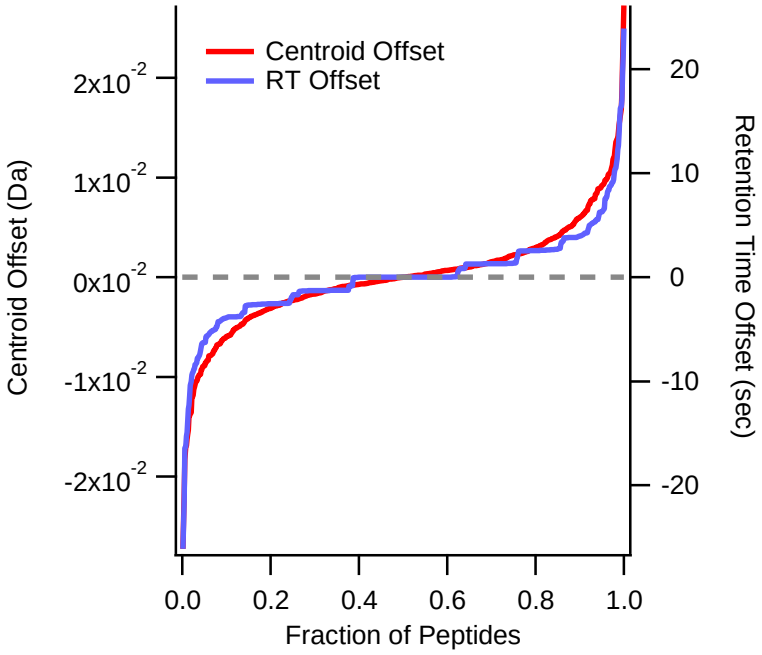

Figure 4.5: Comparing Retention Times and Centroids Using Redundant Ions. 651 Ions are plotted representing 225 *useful* unique MBP peptides, generated by pepsin proteolysis. The two y-axes plot the difference in retention times and centroid values for each ion from the average values of all ions reporting on the peptide. The apparent steps in the RT offset distribution (blue) result from a quantized time axis due to the instrument providing 1 scan per second.

Peptides with the same sequence but different charge states should have similar retention times and mass centroids. We find remarkable agreement in both features as shown in Figure 4.5. The information in this figure was produced as follows: for each unique peptide identified, if there were more than a single charge state/ion, the differences of centroid mass and retention time were taken from the average of each variable across all charge states of the peptide and then collected to produce the cumulative distributions shown in Figure 4.5. On average, retention times for redundant charge states were within 5 seconds of one another. Likewise, deuterated centroids were all within 0.01 Daltons. This result would be highly unlikely if our peptide pool was fraught with incorrect identifications. These observations further verify

that our method accurately identifies a very large number of overlapping peptide fragments for use in HX MS experiments.

4.2.4 Conclusions

There are many reasons to strive for as many overlapping peptides as possible in HX MS experiments. Many overlapping peptides are necessary for accurate site-resolution efforts (6). Overlapping peptides also provide an internal consistency check – agreement in overlapping peptides increases ones confidence. In comparative analyses or in epitope mapping studies, complete sequence coverage is paramount; however, in most cases this has not been achieved. Typical HX MS studies in the literature are plagued by far fewer peptides than we readily find.

Our analysis shows that many more peptides may be available than what is commonly reported in the literature. Here (4), we have shown how to verify peptide identifications and lay down a procedural map that obtains on average 10 overlapping peptides per residue – that this coverage redundancy was true for all proteins tested speaks to the generality of our method. Important determinants are a reliable on-line cold flow system, high resolution and sensitivity in the mass spectrometer, and multiple on-line protease conditions. In a publication that accompanied the one described here, ExMS was shown to outperform other available programs designed for HX MS in terms of finding deuterated peptides, this program is described elsewhere (202). Without the combination of the methods described here and the ExMS program's ability to find deuterated peptides, our coverage would have certainly been far less impressive. The strategy presented here should be broadly applicable to all HX MS experiments in other laboratories where high resolving powers are available.

4.3 Minimizing the Back Exchange Problem

This section focuses on the content of my manuscript published in the Journal of the American Society for Mass Spectrometry in 2012 titled "Minimizing Back Exchange in the Hydrogen Exchange-Mass Spectrometry Experiment", the full citation is included in the bibliography (2).

4.3.1 Introduction

Hydrogen exchange investigations of larger and biologically more interesting protein systems can be achieved by a proteolytic fragmentation method (182) followed by mass spectrometry analysis (190, 205-207). In this method, protein samples taken from an H-D exchange experiment are proteolytically fragmented and separated in preparation for MS analysis to determine the quantity and position of carried D-label at a fragment-resolved level. A problem is that some D-label is variably lost during sample preparation due to back exchange in the H_2O solutions used. The different residues in any given peptide fragment unavoidably lose D-label at different rates (132) (refer to Chapter 2), and this residue-level variability cannot be reconstructed and corrected for when one has only fragment-level data. The problem can only be minimized by reducing the level of back exchange.

Because back exchange quickly degrades HX MS analysis, it continues to receive a great deal of attention (198, 208-217). The typical level of D-label recovery reported in the fragment separation literature is about 70% (30% back exchange). Higher reported values generally depend on results for only one or a few peptides. However, we find that different peptide fragments experience a wide range of back exchange values. Among other implications, any computational correction for back exchange will be flawed.

One popular method to correct for back exchange involves spiking a fully deuterated reference peptide into each sample when the experimental labeling phase is quenched. This approach leads to serious error because of the wide range of back exchange values observed in peptide fragments. We correct each peptide using the level of back exchange observed for that particular peptide in a fully deuterated (FD) control experiment. Here, an FD sample is quenched and subjected to the same preparation and mass measurement conditions used in labeling experiments. By doing this, one is able to determine an expected level of back exchange for each experimental peptide. We find that run-to-run back exchange variability is roughly 2%. Correction factors obtained for individual peptides are certainly more reliable than a uniform correction factor determined by a reference peptide; however, the correction remains fundamentally flawed since different amide sites will be labeled in experimental and control situations and each site will lose label to back exchange at different rates. Thus, MS measurement accuracy of deuterium incorporation during the labeling phase of an HX

experiment depends heavily on the minimization of back exchange while preparing the sample for analysis.

We systematically studied the conditions that determine back exchange including pH, ionic strength, ion transfer tube temperature, the interaction of peptides with reverse phase columns, and the time consumed at each stage of sample preparation. The optimization of these variables reduces back exchange by a factor of two to three.

4.3.2 Results and Discussion

In the typical HX MS fragment separation analysis, an experimental protein is exposed to H-D exchange for a period of time. Each amide hydrogen exchanges at its own rate determined by solvent conditions, its intrinsic chemical rate, and protecting structure, modified by the experimental variable being studied. To measure the extent of D-labeling, protein samples are taken and prepared for MS analysis by quenching into a minimum HX rate condition (low pH and temperature). The protein unfolds but HX is greatly slowed, allowing a short time for sample preparation without excessive loss of D-label. In the present experiments, the protein was proteolytically fragmented (immobilized pepsin column), the peptide fragments were caught on a trap column, washed and buffer exchanged, roughly separated by fast reverse phase chromatography, and then injected by ESI into the spectrometer to determine the mass of each fragment and thus the amount of carried D. These sample preparation steps were performed in an online flow system described on page 52.

To study the effect of various preparatory conditions on back exchange, we used maltose binding protein (MBP, 370 residues) that had been fully deuterated by exchange in D_2O. We measured the recovery of D-label for each of many MBP peptide fragments after passage through the entire analysis. Recovery calculations are described in Appendix A. Although our methods found 225 MBP fragments (pepsin proteolysis alone), we used for each experimental series only the peptides that were observed in all experiments in order to ensure unbiased comparisons. Identification and analysis of these many peptides used SEQUEST (ThermoScientific Bioworks 3.3.1) and the ExMS program (202).

pH and Ionic Strength

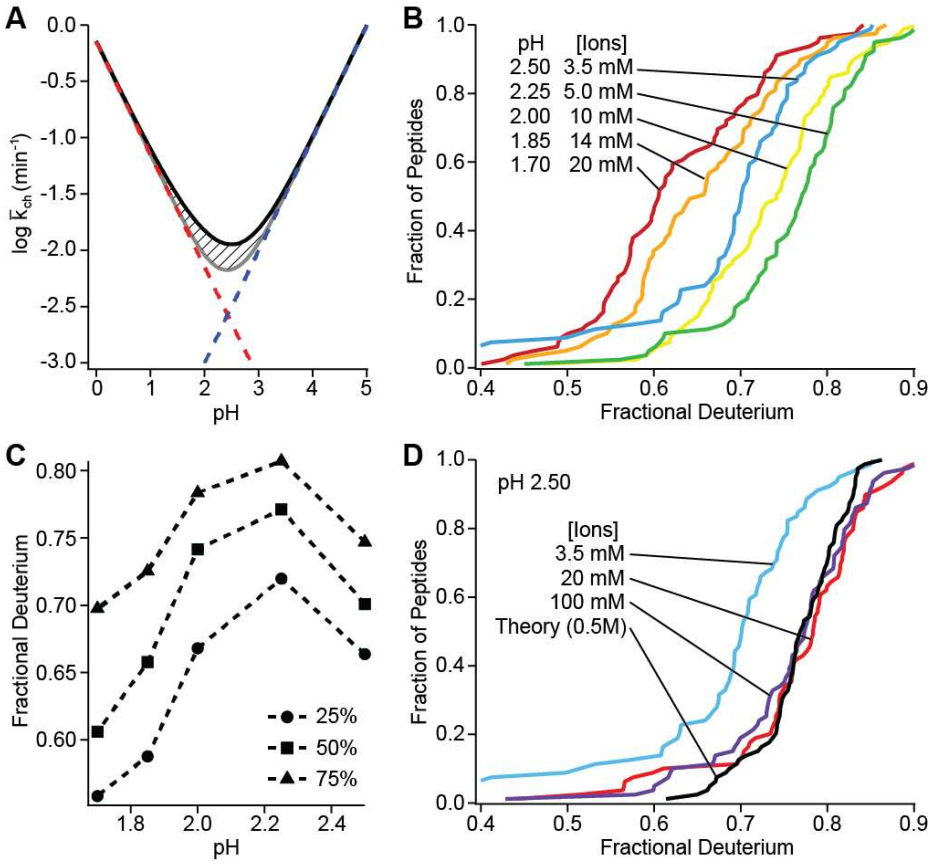

Figure 4.6: Dependence of HX rates on pH and ionic strength at 0 °C. *(reprinted from reference(2))*. **(A)** Expected pH dependence for the hypothetical peptide GGVALISTDENQRHKCMFTW. The rate at any pH is the sum of the H_3O^+ ion catalyzed reaction (red) and the OH^- ion catalyzed reaction (blue), each of which varies by 10-fold per pH unit. The additional pH-independent contribution due to water catalysis is shown with hatch marks. The averaged fragment-level HX rate constant shown is taken as the geometric mean (log averaged) of the 20 amides, each of which exchange with somewhat different rate constants. **(B)** Cumulative population distribution pH series. **(C)** Slices taken across B at given population percentiles. **(D)** Effect of ionic strength at pH 2.5.

Figure 4.6A shows the expected dependence of HX rate on pH for a hypothetical peptide with all amino acids, calculated from standard reference values (132, 134). The theoretical minimum HX rate, computed by taking the log-average rate for a hypothetical peptide containing all 19 amino acids and two N-terminal alanines is expected to be reached at pH 2.5. Accordingly, the quench and running buffers in fragment-separation experiments have always been prepared near this condition (182). To test this expectation we performed a series of D-recovery experiments over a range of experimental pH values (Figure 4.6B). Single peptide values are often used as a back exchange reference in the literature. In fact, different peptides display a wide range of D-label recoveries. This can be expected since amide HX rate varies with amino acid type and nearest neighbors (132, 134). Unexpectedly however, the minimum rate

with significantly reduced back exchange was reached at pH 2.25 (Figure 4.6C). Upon further inspection of the literature, we found in Z. Zhang's 1995 Purdue dissertation (218) an example where pH 2.3 was promoted as the ideal quench pH; however, no explanation for the observed difference between measurement and theory was presented.

Testing showed that the shift in the pH of minimum rate depends on ionic strength. When ionic strength is 20 mM or higher, HX rate is a minimum at pH 2.5 and matches expected values (Figure 4.6D). The earlier HX rate calibrations (132, 134), which underlie theoretical rate computations, were done in high salt (0.5 M KCl) purposely to shield against extraneous charge effects. However, MS analysis requires electrospray solutions with low salt where, we find, the pH of minimum HX rate is significantly shifted. The amide group acts like it has a small net positive partial charge which, at low ionic shielding, favors the OH^- -catalyzed reaction and disfavors H_3O^+, shifting the pH-rate curve to the left.

The present results show how these different requirements for minimizing back exchange rate can be satisfied. Experimental HX samples normally contain significant salt and after quench due to the presence of GdmCl (0.5 to 2 M) added to promote protein unfolding and improve digestion in the proteolysis step. Therefore, in this first stage of sample preparation we use quench buffer with 1 M GdmCl at pH 2.5. The sample is then caught on a trap column, washed, eluted with an acetonitrile gradient through the LC step, and injected online into the mass spectrometer. These latter steps should use low salt, desirable for ESI MS, and the lower pH. We use wash and elution buffers with 0.1% formic acid adjusted to pH 2.25 with TFA. Solution pH values were measured and adjusted in the pertinent solutions at room temperature and then used at 0° C.

Desolvation Temperature

Details of the ion source depend on instrument design. In our spectrometer, after nebulization at the electrospray needle, droplets of solution are pulled by a pressure differential through a heated capillary (~200 °C), which speeds solvent evaporation and ionization. As exchange rates in solution depend sharply on temperature (~3-fold per 10 °C) (132, 134), sample heating in the capillary might greatly promote back-exchange.

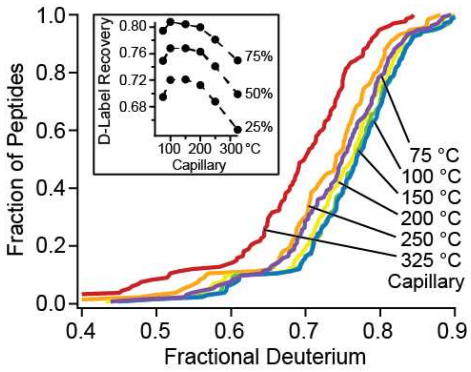

Figure 4.7: The dependence of D-label recovery on transfer tube temperature *(reprinted from reference(2))*.

We measured back exchange as a function of capillary temperature. Cumulative recovery distributions are shown in Figure 4.7. The results show a broad maximum in D-recovery when capillary temperature is set between 100 and 200 °C, with declining recovery at higher and lower temperature. We did not observe a difference in recovery between charge states of the same peptide as reported before (216), apparently due to instrumental differences. Results for given peptides with different charge state agreed in these and our other experiments to < 0.1 D.

Interestingly, the 75 °C data shows distinctly reduced recovery. Less efficient evaporation at 75 °C could lead to increased time at temperature above 0 °C in the liquid phase before solvent evaporation, leading to increased back-exchange. Given these results, we adopted a capillary temperature setting of 100 °C.

Time on the LC Column

To study the HX behavior of peptides bound to the C_{18} media of reverse phase columns, we compared HX rates of column-bound peptides with rates expected from earlier calibrations in free solution. Fully deuterated MBP samples were placed into quench conditions in H_2O, injected into the online flow system**,** digested, and washed onto the trap column (5 min elapsed time). Peptides were held on the column for an additional experimental delay time between 0 and 45 min, then eluted from the trap column, through the analytical LC column, and into the mass spectrometer (3 to 18 minutes additional time).

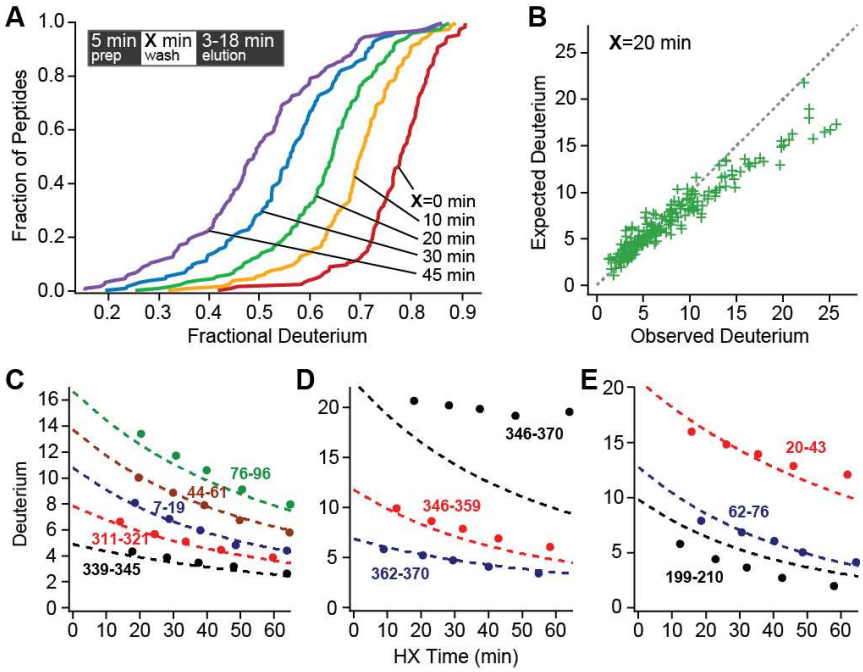

Figure 4.8: Exchange on the column *(reprinted from reference(2))*. **(A)** Observed D-recovery for all peptides across the delay time series. The inset places the variable delay time during sample preparation. **(B)** Recovery for the various peptides after a 20 minute delay on the trap column. **(C–E)** Observed (data points) and theoretical (dashed lines) D-label recovery. **(C)** Some peptides with normal recovery. **(D)** A peptide with large slowing on the column due to structure formation and two component peptides with normal recovery. **(E)** Some histidine-containing peptides with accelerated early loss.

Figure 4.8 shows cumulative recovery distributions across the time series, and compares these results with the expected time-dependent loss of D-label in free solution (Appendix A describes how the expected recovery was determined during preparatory steps where the pH and temperature are not constant). We assumed that D-label on side chains (132) and the N-terminal amino group is lost too rapidly to measure (expected rate > 10 s^{-1}), and similarly for the amide on the second residue. The accelerated rate for the second residue is due to the absence of an amide group on the prior residue (10-fold in rate), and it is promoted by another 10-fold by the fixed positive charge on the neighboring N-terminal amino group, especially at the low salt concentration used here. This effect is contained in the older literature on HX of peptide models (see Table 1 in Molday *et al*. (133)) and has been directly measured more recently (219).

A comparison between observed and expected D-recovery from the 20 minute delay experiment is shown for the whole peptide population in Figure 4.8B and for a number of

individual peptides across the delay series in Figure 4.8C-E. During the sample preparation time including proteolysis and column interaction, most of the peptides exchange as expected. A few are much slower. Large retardation with HX slowing up to 20-fold while bound to the column matrix was seen for 11 overlapping peptides between the C-terminal residues 340 to 370. Figure 4.8D shows one of these and two shorter component peptides which exchange as expected. Interestingly, in native MBP this segment adopts a helix-turn-helix motif and docks with a hydrophobic interface on the C-domain. Evidently, this peptide and some subfragments are induced to form mildly stable H-bonded structure, perhaps aided by hydrophobic interaction with the hydrocarbon chains of the reverse phase column. The slowing factor decreases systematically as either (helix) segment is cut back. Similar but more modest slowing, up to 4-fold, was seen for sets of peptides derived from several other protein segments (116-149, 169-194, 283-301, 312-330), apparently due to tentative helix formation.

A recent publication (220) that following the manuscript described here affirmed the idea that helix formation may be promoted by binding to the LC column. In this work, peptides were designed to form amphiphilic helical surfaces. The authors found that these peptides exhibited reduced exchange rates when bound on an LC column than when free in solution. This work appears to confirm our hypothesis that binding to the LC column may induce helix formation and provide additional protection from back exchange.

Some peptides show a small but noticeable negative offset between the expected and observed number of D atoms at the earliest time point, indicating additional D-loss. This included all of the 15 peptides that contain one of the three MBP histidine residues, suggesting some (acid) catalysis of nearby residues by the imidazolium side chain. However, we have not seen indications of this phenomenon with histidine-containing peptides in some other proteins.

Sample Preparation Time

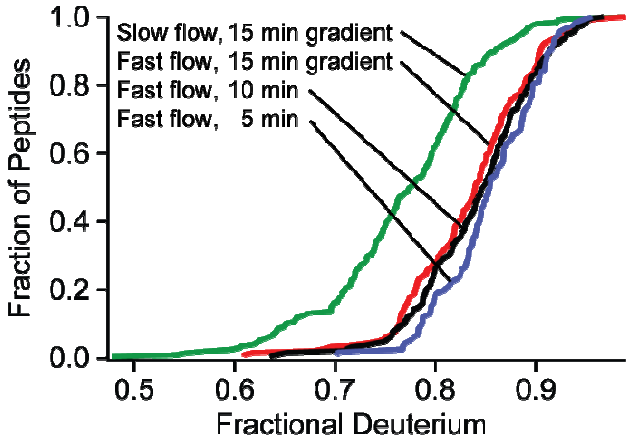

Figure 4.9: The effect of reducing LC gradient length (time) on peptide recoveries *(reprinted from reference(2))*. Sharper elution gradients provide little reduction in back exchange and sacrifice peptide fragment yield. Minimizing preparation time by increasing flow rates (termed the fast flow condition, see text) increased recovery levels from the green to the colored distributions.

Most previous attempts to minimize back exchange focus on minimizing the time that samples spend on the reverse phase column, with modest improvement. We find that time reduction accomplished by shortening the acetonitrile elution gradient (15, 10 or 5 minutes) produces surprisingly small gains (Figure 4.9). The reason appears to be that early eluting peptides experience almost no time reduction while later eluting peptides tend to have a slower intrinsic HX rate (132) (more large apolar side chains (132), more time in higher acetonitrile) so that increased exposure time has less than the expected effect on back exchange. Consider the following, whereas the amides of polar residues lose label at the rate of 1% to 2% per minute, the large apolar residues do so ~4 times more slowly (132). Experimental evidence for this view comes from the fact that we find no correlation between the level of D-recovery and column elution time. In these experiments, total back exchange time varied between 10 and 20 minutes.

Elution Profile Length	5 Minutes	10 Minutes	15 Minutes
Recovery Average	0.85	0.84	0.83
Peptide Count (total)	273	306	296
Peptide Count (unique)	**116**	**190**	**183**

Table 4.1: The effect of reducing LC gradient length on recovery and peptide coverage. The information in this table was taken from our "fast flow" condition, described in the text.

In fact, the reduction of column retention time proved counter-productive. In order to obtain ultimate HX resolution at the amino acid level, it will be necessary to obtain a large

number of sequentially overlapping peptides and multiple residue coverage. Chromatographic crowding became a problem in the 5-minute gradient resulting in 40% fewer useful peptides (see Table 4.1). Gradient shaping (Appendix B) used to equalize peptide density through the chromatogram reduces but does not overcome this problem.

We more broadly reduced the time required to navigate the free volume in our flow system by increasing overall system flow rates. An increase in flow rates, to "fast flow" conditions (300 µl/min during digestion, 450 µl/min for buffer exchange, 10 µl/min during peptide elution) reduced overall sample preparation time by 4.3 minutes. These flow rates maintained pressures below 2000 psi as recommended for POROS media (protease column). The increased D-recovery illustrated in Figure 4.9 is consistent with the expected back exchange loss rate of about 1 to 2% per minute on average of carried D-label at the pH minimum and 0 °C (132).

Other Considerations

Are these results for maltose binding protein typical for proteins in general? Each test shown here used ~90 peptides, and they vary over a wide range in size, amino acid content, hydrophobicity, etc. It seems unlikely that sets of peptides from other proteins will behave differently. In agreement, we have now used our previous sample processing conditions and the improved conditions described here in ongoing experiments with other proteins (cytochrome c, staphylococcal nuclease, ribonuclease H, apolipoprotein A-I, Hsp104). The gain in D-recovery was comparable in all cases.

When is back exchange important? For HX MS experiments in which one attempts to define epitopic or ligand binding sites, one may be satisfied with crude peptide-level changes. These are less dependent on back exchange. Back exchange becomes most important when reaching for the amino acid level of resolution that has made the HX NMR experiment so powerful for protein studies. Recent progress using ECD and ETD to strive for site resolution minimizes the back exchange problem. In this case, the whole protein can be injected directly into the mass spectrometer, avoiding the fragment separation analysis. However, a major advantage of the fragment separation analysis is the ability to study much larger and biologically more important proteins than HX NMR can accomplish. This goal probably exceeds the capability of direct ECD/ETD methods. In order to study large proteins by these methods, it

seems likely that the fragment separation approach will be required as an initial step, resurrecting the back exchange problem.

The ability of the HX MS method to achieve high structural resolution depends on obtaining high quality data for many overlapping peptide fragments. We previously described methods for obtaining (4) and efficiently analyzing (202) hundreds of useful protein fragments with data accuracy to ~0.1 D. The present work shows that attention to the various factors that determine back exchange can increase D-recovery into the range 75 to 95%, as summarized in Figure 4.9. These capabilities taken together give the investigator freedom to choose among different options. For example, if the effort to reach single amino acid resolution requires exceptionally low back exchange, the sacrifice of the lower half of the peptide population shown in Figure 4.9 would still retain a very large number of peptides with high data quality, i.e. with D-recovery in the range of 90 ± 5%.

4.3.3 Conclusions

A systematic study of the factors that influence back exchange in the typical HX MS setup reveals a number of surprises and shows how the back exchange problem can be minimized. We find that different peptides exhibit a range of back exchange levels. Among other implications, this situation negates the use of any one or a small number of peptides as a reference marker for the degree of back exchange or its correction by computation. This must be done peptide by peptide and even then is imperfect since different amide sites will be detected in experimental and reference situations.

Results show that there is no single best back exchange condition; it varies with ionic strength. The first stage of sample preparation, involving proteolysis and sample trapping, is best performed at pH 2.5 and 0^{o}C in high ionic strength, often with substantial GdmCl. The trapped peptides should then be washed and passed through the analytical HPLC column in pH 2.25 solution at low ionic strength. The common approach of trying to limit chromatographic time (reduced column size; shorter elution gradient) yields limited gains and is potentially counter-productive in respect to the yield of useful peptides. Sample exposure time can be more simply minimized by using high flow rates to rapidly clear system free volume. Putting aside the previously unexpected ionic strength effect, the loss of D-label through the sample preparation time proceeds closely as predicted from previous amide HX rate calibrations (132,

134) although peptide-column interaction can have some unexpected effects such as structure formation. The combination of previously described methods for producing (4) and analyzing (202) many peptide fragments together with the ability to largely negate deleterious back exchange moves toward the goal of obtaining ultimate amino acid structural resolution for HX MS analysis.

4.4 Extracting Information from HX MS Peptides

4.4.1 Introduction

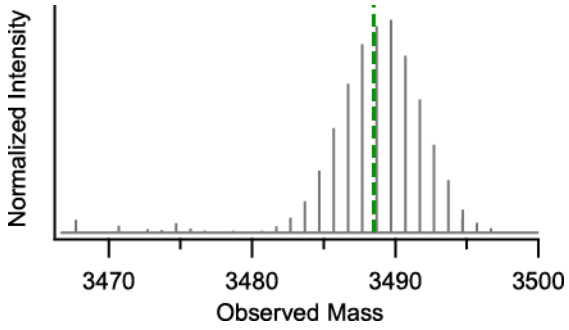

Figure 4.10: The fully deuterated MBP peptide 163-195 with centroid indicated by the green dashed line.

The principle information of interest in an HXMS experiment is the average deuteration level of each peptide. Typically, each peptide contains a single mass distribution and its average deuteration level is easily obtained by computation of the peptide's mass centroid, such as shown in Figure 4.10. However, during kinetic biochemical processes such as protein folding, molecular ensembles may be distributed over many independent conformational states. When interrogated by HXMS pulse labeling, each state typically has a different degree of protection from HD exchange. This results in multiple mass distributions present within a single peptide mass envelope and renders the centroid computation irrelevant.

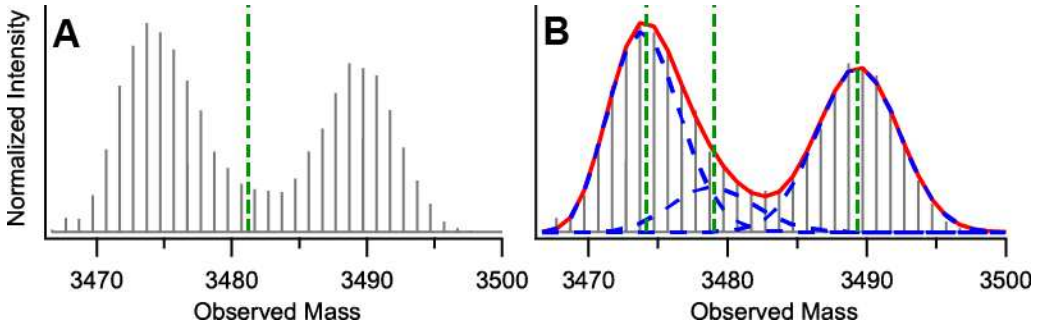

Figure 4.11: Multimodal Distributions. **(A)** demonstrates the failure of a single centroid computation (green dashed line) to extract useful information from a multimodal mass spectrum. **(B)** The results from fitting this data with HDpop. The red line is the sum of the individual population fits shown in blue. There are three populations, each with a different centroid. The green dashed lines demonstrate the centroids of each population, as computed by HDpop.

A folding time point for the same peptide as in Figure 4.10 but before native equilibrium is shown in Figure 4.11. This time point was chosen to illustrate the issue that arises from multiple populations. The raw centroid is drawn using a dashed green line in Figure 4.11A highlighting that the computation is meaningless in multi-modal mass envelopes – it reports information about the whole envelope and remains oblivious to the presence of multiple populations.

The challenge for practitioners collecting such data is to determine the number of mass distributions or populations present in a mass envelope and then further to characterize the fraction of molecules in each population and the average deuteration level of each one. In this section, a program referred to as HDpop is introduced and described to satisfy this challenge. For the peptide in Figure 4.11, HDpop determines that three populations are present; the HDpop results are presented in panel B. Each population is identified by the dashed blue lines and the sum of all populations traces the peptide envelope in red. Theoretical centroids, one for each population found by HDpop, are shown in green dashed lines.

A number of groups have addressed this challenge by fitting peptide mass distributions using a Gaussian (202, 221-224) or binomial function (225) written to include the possibility of up to two populations. These methods are often capable of providing an accurate estimate of the deuteration level and mole fraction of each distribution or state; however, none provides any mechanism to determine the statistical significance of incorporating an additional population into the model and all are limited to two populations. A statistical mechanism for

determining how many populations are present is necessary to avoid bias in interpretation, and the limited expansion to two populations may be inadequate as is shown by the three-population envelope in Figure 4.11.

4.4.2 The HDpop Program

The method introduced here and implemented in HDpop offers two important advantages over what is currently available. First, this method incorporates the F-test (a parametric statistical hypothesis test) to determine the merit of incorporating an additional population to the fit; a valid *p*-value accompanies the decision to expand the number of populations. This establishes a statistical population detection mechanism whose false positive rate is adjustable to a user-desired confidence interval (the $p-value$). Second, the number of degrees of freedom in the dataset determines the maximum number of populations possible. This circumvents the arbitrary limitation to two populations. A description of this method in the context of our programmatic implementation, HDpop, is described below accompanied by an example of the application of HDpop to define the folding pathway of ribonuclease H.

Preparing the Data for Analysis

* EUCLID typeface indicates an ordered list of scalar variables.

* $\mathrm{Y}[i]$ '*i*' indexes the list Y, square brackets in this context refer to list indices. Numbering begins with i=0.

The first step of HDpop is to compute the natural abundance mass spectrum[14] for the peptide. Intensities[15] of similar isotopologues (ie. +1 ^{13}C or +1 ^{15}N) are combined by addition resulting in one intensity for each integer offset from the monoisotopic peak. A full discussion regarding nominalization of the mass axis, the monoisotopic peak, and the details for computing the natural abundance distribution are presented in Appendix C.

[14] Natural Abundance Mass Spectrum refers to the mass distribution resulting from isotopes of C, N, O, S.
[15] Model intensities refer to the probability of observing a particular mass in the mass distribution.

After computing the natural abundance distribution, peak intensities are stored in a list and the list is culled to reflect instrument sensitivity. Starting from the last entry and progressing towards the first, peaks are removed until the first peak is encountered with sufficient intensity to be detected; this detection threshold is defined by the user. This list $\mathrm{P}_{natural} = y_0, y_1, \ldots, y_k$ contains the natural abundance distribution. Nominal mass information is encoded by list indices. The monoisotopic peak takes the subscript '0', +1 isotopologues take the subscript '1' and so on.

HDpop then computes the degrees of freedom for the measured data which is used later to define the maximum number of fit populations. The variable k represents one less than the number of theoretically detectable peaks in the natural abundance distribution, $\mathrm{P}_{natural}$; thus, k defines the degrees of freedom in a non-deuterated version of the mass envelope and accounts for the experimental signal-to-noise by way of the user-defined sensitivity described earlier. Experimental degrees of freedom for a deuterated peptide are

$$dF_{data} = k + s, \qquad \text{Eq. 4.1}$$

where the variable s represents the number of exchangeable sites on the peptide. The number of exchangeable sites is determined by the number of non-proline residues that exist on the peptide not counting the first two residues, as they are known to back-exchange completely (132, 134). If the peptide was fully deuterated, the natural abundance distribution would be nominally shifted by s, giving $+s$ potentially detectable peaks.

Our peptide identification program (ExMS,(202)) determines one peak intensity per integer offset from the monoisotopic mass and reports a 42-entry list, Y_{ExMS}, indexed in the same way as $\mathrm{P}_{natural}$ for every peptide identified. During import into HDpop, a normalization constant is determined and the data is read as a normalized list of peak intensities:

$$\eta = \sum_{i=0}^{\mathrm{Y}_{ExMS}} \mathrm{Y}[i], \qquad \text{Eq. 4.2}$$

$$\mathrm{P}_{data} = \mathrm{Y}_{ExMS} \frac{1}{\eta}. \qquad \text{Eq. 4.3}$$

N-population Binomial Fitting

HDpop employs a non-linear least squares optimization (the Levenburg-Marquardt algorithm (226), implemented in python (227)) to minimize the error function[16]:

$$\mathrm{E}_{fit} = \mathrm{P}_{data} - \left(\mathrm{P}_{fit} * \mathrm{P}_{natural}\right) \qquad \text{Eq. 4.4}$$

Here, the convolution operator is represented by $*$, various HDpop implementations of this operator are described in Appendix C.

To evaluate the error function, an s+1 element list, P_{fit}, is constructed by the N-population binomial function which contains all floating parameters:

$$B\left(x:N,\mathrm{A},\mathrm{PR}\right) = \sum_{i=0}^{N-1} \mathrm{A}\left[i\right]\left(\frac{s!}{x!(s-x)!}\right)(\mathrm{PR}\left[i\right]^{x})(1-\mathrm{PR}\left[i\right])^{s-x}, \qquad \text{Eq. 4.5}$$

$$\mathrm{P}_{fit}[x] = B\left(x;N,\mathrm{A},\mathrm{PR}\right) \mid 0 \le x \le s\,. \qquad \text{Eq. 4.6}$$

Fit parameters A and PR contain N amplitude and probability entries, one for each population. The fraction of molecules in the i^{th} population, $f_i = \mathrm{A}\left[i\right]$, and the expected number of incorporated deuterons, $d_i = \mathrm{PR}\left[i\right] \square s$, are obtained directly from the fit parameters. The total number of floating degrees of freedom in the fit model is

$$dF_{fit}(N) = 2N - 1\,. \qquad \text{Eq. 4.7}$$

One is subtracted because regardless of the size of N, one amplitude is determined *a priori* due to the normalization condition: $\sum \mathrm{A} = 1$. Before construction of P_{fit} (Eq. 4.6) and subsequent evaluation of the error function (Eq. 4.4), an arbitrary amplitude of 1 is appended to A and then the amplitudes are normalized.

[16] Before evaluation of the error function, all lists must contain $dF_{data} + 1$ elements and this is achieved by either list truncation or zero-padding. The 42-element list from ExMS, P_{data}, is truncated by removing trailing zeros. The k+1 element and s+1 element lists, $\mathrm{P}_{natural}$ and P_{fit} (described in the text) are zero-padded.

Population Detection

HDpop detects the number of statistically justified populations at user-determined confidence interval by hypothesis testing via the F-test. The F-test may be used to choose between two nested[17] models referred to as reduced (N populations) and expanded (N+1 populations), note that $dF_{reduced} < dF_{expanded}$ [18] must be true. As a result, the expanded model will always be able to fit the data at least as well as the reduced model. The F-test determines if the expanded model fits the data *significantly* better.

HDpop takes the null hypothesis to be that the expanded model (one more population than the reduced model) does not fit the data *significantly* better than the reduced model. To test the hypothesis, first the residual sum of squares scalar is computed from the error function (Eq. 4.4) for each model and then used to compute the F-statistic as follows:

$$RSS_{fit} = \sum_{i=0}^{dF_{data}} \mathrm{E}_{fit}\left[i\right]^2, \qquad \text{Eq. 4.8}$$

$$F_{statistic} = \frac{\left(\frac{RSS_{reduced} - RSS_{expanded}}{dF_{expanded} - dF_{reduced}}\right)}{\left(\frac{RSS_{expanded}}{dF_{data} - dF_{expanded}}\right)}. \qquad \text{Eq. 4.9}$$

The probability of falsely rejecting the null hypothesis (accepting the additional population when we should not have done so) is given by the $p-value$, obtained by integrating the parameterized F-distribution, from the F statistic to zero:

$$p-value = 1 - \int_0^{F-statistic} \frac{1}{\int_0^1 t^{\frac{d_1}{2}-1}\left(1-t\right)^{\frac{d_2}{2}-1} dt} \left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}} x^{\frac{d_1}{2}-1} \left(1+\frac{d_1}{d_2}x\right)^{-\frac{d_1+d_2}{2}} dx. \qquad \text{Eq. 4.10}$$

[17] The term *nested* means that by setting certain coefficients in the expanded model to zero, the nested and expanded models are equivalent.

[18] $dF_{reduced} = dF_{fit}(N)$, $dF_{expanded} = dF_{fit}(N+1)$

The F-distribution has two degrees of freedom defined by $d_1 = dF_{expanded} - dF_{reduced}$ and $d_2 = dF_{data} - dF_{expanded}$, estimation of the integral is performed using python and hyper trigonometric functions. If the $p-value$ is less than or equal to the user supplied critical value (1 – C.I.), the reduced model is rejected.

For any given peptide, If $dF_{data} \leq dF_{fit}(2)$ is true, the F-test does not run and HDpop defaults to a single population fit – in this rare case, a p-value is not computed. Typically, HDpop begins by fitting a single population (N=1, reduced model) and a double population binomial (N=2, expanded model) to the data before performing an F-test to determine whether the double population binomial gives a statistically significant improvement in the residual sum of squares with respect to the changes in available degrees of freedom. Then the program enters an F-test loop where the previous expanded model becomes the reduced model, and the new expanded model increases by one population. The new expanded model is fit and the result F-tested. The program continues iterating over the F-test loop until stopping criteria are met.

There are two stopping criteria in HDpop. The obvious conclusion is when the F-test fails to reject the null hypothesis and accepts the reduced model. The loop also closes if the maximum number of populations have been tested, determined either by the user or by the program. If determined by HDpop, the maximum number of populations has been tested when the inequality $dF_{data} \leq dF_{fit}(N+1)$ is true.

The program reports these p-values in a spreadsheet that is presented to the user following analysis.

4.4.3 A Sequential Folding Pathway Defined with HDpop

HDpop was recently highlighted in a HXMS pulse-labeling study (5) whereby the mechanism and complete folding trajectory of RNaseH was fully determined in structural detail. Previous studies (111) found that RNaseH folds by way of a rapid unresolved burst-phase, complete by 15 ms of folding, followed by a slower phase leading to the native state. These studies were not able to attain definitive structural resolution that was achieved here (5) by using HX MS pulse labeling followed by HDpop analysis.

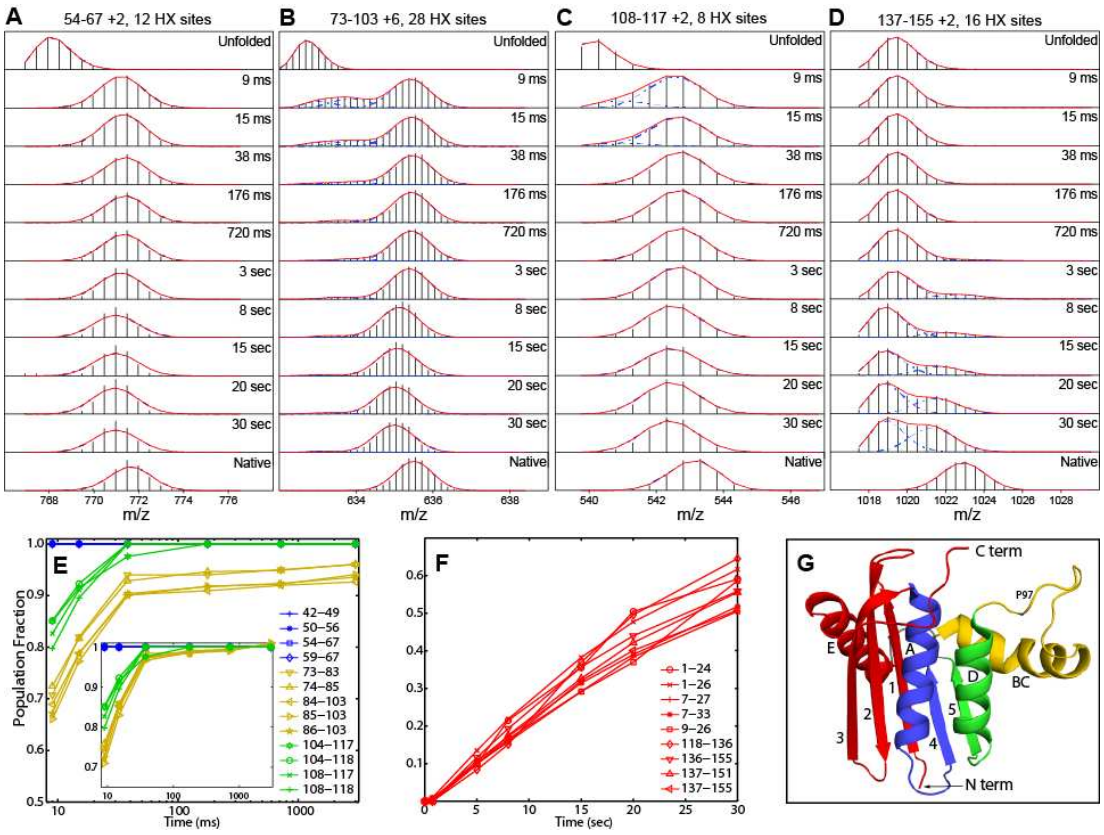

Figure 4.12: HX MS data for RNase Folding *(reprinted from reference (5)).* **(A-D)** Four peptides represent the four general categories of folding behaviors observed in the experiment. **(E)** The fraction of molecules that are heavy (indicative of native structure) showing three sequential foldons plotted as a function of folding time. The inset renormalizes the yellow curves to demonstrate that they are indeed trailing the green curves in time. **(F)** The slow folding peptides plotted in the same way as panel E. **(G)** The structure of RNase H colored to represent the order of folding as defined by the folding rates shown in panels E and F.

In this study, a population of fully deuterated and unfolded RNaseH molecules was diluted into refolding conditions where exchange rates are slow. After a variable amount of folding time, the sample was subjected to conditions that rapidly favor exchange of unprotected amides (the pulse). During the 10 ms pulse, the mean HX lifetime of unprotected backbone amides was roughly 0.4 ms giving unprotected sites ~25 lifetimes to exchange. Any deuterium remaining after the pulse may then be attributed to the formation of H-bonded or tertiary structure during the variable folding phase of the experiment.

During the progression of folding, any given fragment will convert from a lighter unfolded state to a heavier state, and the heaviest observed state is taken to represent the native conformation. Before a given segment has fully reached the native state, molecules will

be found in both states and peptide mass envelopes are expected to be multi-modal. The sub stoichiometric ratios of each mass distribution reflect the fraction of molecules in the particular state and provide a measure of the folding rate for the particular segment. Examples of multi-modal mass distributions fit by HDpop can be seen in Figure 4.12A-D representing the various folding rate categories observed in the experiment.

HDpop was integral to the analysis of this data in two distinct ways. First, the program was employed to resolve the overlapped peptide mass distributions and provide a measure of the fraction of molecules in each state. By following the fraction of molecules in the heavy population with respect to time, a sequential nature to the folding pathway was discovered as shown in Figure 4.12E-F. The population detection feature combined with HDpop's accurate determination of the fraction of molecules in each population directly led to the disambiguation of fast folding behavior into three sequential events occupying regions of the molecule colored blue, green, and yellow in Figure 4.12G. Without a mechanism to detect and resolve heavily overlapped populations, such as shown in Figure 4.12B-C, the folding rates of these three distinct regions would have been blurred. Though blue, green, and yellow peptides would have all been observed to fold faster than the slowest folding red peptides. The differences within the fast group and resulting structural implications could not have been defined with such clarity without HDpop. Second, HDpop was used indirectly as a filter for a program called HDsite that was used to determine the deuteration levels of individual sites utilizing information from overlapping peptides. The HDsite algorithm requires strictly uni-modal mass envelopes for reasons that are described elsewhere (6) and briefly mentioned in Appendix C, page 124. Only peptides having a single population, as determined by HDpop at the 95% confidence interval, were included in the HDsite analysis. Importantly, the site-resolved information from HDsite verified that the protection observed in the pulse labeling experiment corresponded to native secondary structure elements. This capability indirectly depended on HDpop.

4.4.4 Conclusions

HX MS provides unparalleled insights into folding pathways as a product of its state sensitivity; however, without the proper analytical tools, much of this information is ignored. A

proper analytical tool for this purpose must take an unbiased approach in fitting the data as the number of populations found potentially changes the interpretation. The program HDpop utilizes the statistical F-test to make an unbiased decision regarding the number of populations utilizing a user adjustable confidence interval. Because the maximum number of populations is defined by degrees of freedom in the dataset, the analysis is not limited to two populations as in other available programs. HDpop should be useful for all HX MS data in other laboratories.

HDpop has proven essential to our work in the Englander laboratory. We check every dataset with HDpop, regardless of whether we expect multi-population behavior. The analytical method has been employed to unravel the folding pathway of RNaseH in exquisite detail and also the folding pathway for MBP (discussed in Chapter 5).

4.5 Concluding Remarks

In the early days of HX experiments, NMR was preferred because of the ability to measure deuterium occupancy at the site-resolved level. NMR methods have some serious negatives that are overcome by the advent of high-resolution mass spectrometry and the strategies presented in this chapter.

Specifically, with respect to studying the folding pathway of MBP, NMR could not have been employed for kinetic pulse labeling HX measurements discussed in Chapter 5. This is because of the slow folding of MBP and the NMR requirement that the molecules be properly folded for pairing chemical shifts with their residue identities. The protein will not fold at the low pH condition necessary for slowing hydrogen exchange rates to minimize back exchange; thus, at elevated pH required for folding, all of the label would be lost. The fragmentation-separation experiment by mass spectrometry has no such requirement. Directly after pulsing the sample, we drop the pH to quench exchange and immediately digest the protein into peptides for MS measurement. This allows us to preserve the experimental labeling pattern.

In this chapter, we have shown methods that advance the HX MS experiment. The needs to obtain many overlapping peptides, minimize back exchange, and properly analyze experimental results were hurdles for HX MS in the past that are hindrances no longer. With these methods, we are now equipped to study the folding of MBP at structural resolution in the next chapter.

Chapter 5 – Studies on MBP Folding

The work presented in this chapter was submitted to PNAS in October, 2013 (condensed, less text) and appeared in the PNAS Early Edition on November 4th, 2013. Citation information is given in reference (3).

5.1 Introduction

The protein folding problem is fundamental for understanding *in vitro* protein biophysics and *in vivo* biological proteostasis. Yet, 50 years after Anfinsen's seminal demonstration that an unfolded protein can refold spontaneously when placed under native conditions, major questions concerning the folding process are still ambiguous (48, 228-230). Important questions relate to the condition of the unfolded state, its degree of compaction, the reality and character of residual structure before folding begins and its possible role in guiding the folding process (49, 130, 231, 232). Analogous questions relate to the folding pathway more broadly. Do proteins fold through many alternative independent pathways, an IUP-type mechanism, as earlier theoretical investigations have suggested, (71-73) or do they fold through predetermined intermediates in a distinct pathway (9), a PPOE-type mechanism, as a growing list of experimental observations indicate (5, 11)?

To answer these questions it will be necessary to define experimentally the residual structure that exists in the unfolded state and the intermediate forms that proteins move through on their way to the native state. This goal is beyond the reach of the usual high resolution crystallographic and NMR structural methods. The great majority of experimental folding studies have therefore relied on low-resolution optical methods that can follow folding in real time but rarely provide the kind of structural information necessary to resolve the basic mechanistic questions. Recent work has demonstrated an advanced hydrogen exchange - pulse labeling - mass spectrometry technology (HX MS) that is able to detect and characterize local structure even when it is only transiently present during the course of kinetic folding (5, 114). The method provides a snapshot of main chain amide sites that are protected against HX labeling by H-bonds that are present at the time of the labeling pulse. The measurements can

determine the position, stability, and dynamic behavior of native and non-native H-bonded structure, independently of whether it persists or dissipates in subsequent folding. In recent work the method was able to describe the structure and time-dependent formation of three sequential native-like folding intermediates in the 155 residue ribonuclease H protein (5).

Protein folding studies, whether theoretical or experimental**,** have been limited to relatively small proteins, with few exceptions. However, biological proteomes and the considerations they raise are dominated by larger proteins. Here we extend an advanced HX MS technology, introduced in Chapter 4, to the two-domain, 370 residue, maltose binding protein (MBP). MBP is synthesized in the *E. coli* cytoplasm and transported to the periplasm (233) where it serves as a soluble receptor for the high affinity capture and transport of maltose and maltodextrins (234). The protein folds *in vivo* after deletion of a signal sequence; we study here the mature protein with the signal sequence deleted.

When unfolded MBP is placed into native conditions, it rapidly adopts a heterogeneously collapsed dynamic state, which can lead to aggregation *in vitro* and inclusion body formation *in vivo* when the concentration is >1 μM (109). Folding to the native state occurs much more slowly even in the absence of aggregation, moving through the formation of an obligatory intermediate substructure (~7 sec) and thence to the native state (~100 sec). The HX MS experiment provides incisive information on the nature of the initially collapsed state, the slow formation and identity of the on-pathway native-like intermediate, and the even slower emergence of native structure.

5.2 Results & Discussion

5.2.1 Optical Measurements

Because folding rates could be sensitive to pH and we needed to pulse label during the HX experiment discussed later at pH 9.0; we uniformly conduct optical studies using a pH of 9.0 and use a deuterated protein in D_2O solvent.

Equilibrium Stability and Unfolding at pH 9.0

The global stability and thermodynamic characterization of wild-type MBP has been previously assessed by chemical denaturation over a range of temperatures (Urea and GdmCl)

and calorimetric analysis (162, 164, 165) at pH 7.1. In GdmCl, Sheshadri et al. (162) measured a stability of 12-14.5 kcal/mol with a GdmCl $C_{½}$ of 1.03 M ($C_{½}$ is the [GdmCl] where $k_f=k_u$). In all of these studies, the pre- and post-transition baselines had been removed. Baseline behavior is useful for interpretation. We also wanted to assess the stability at pD 9.0 since published data was for lower pH values.

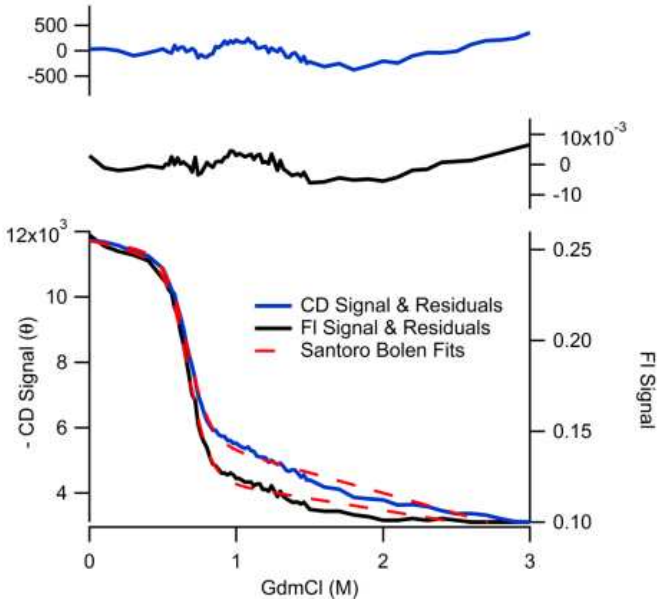

Figure 5.1: Equilibrium melt of MBP at pH 9.0 and 20 °C. CD absorption was measured at 222nm, fluorescence measurements were acquired with 280nm excitation and a 309 nm long-pass cutoff filter between the sample and photomultiplier for measurement.

In our hands, denaturation at pD 9.0 follows a distorted sigmoidal curve with non-linear baselines, shown in Figure 5.1. The melting curve can be extrapolated to a global stability of 5.12 ± 0.33 and 5.38 ± 0.32 kcal/mol for CD_{222} and fluorescence, respectively. This extrapolation assumes unfolding and refolding are two-state processes; we know that MBP likely violates this assumption. The usual 6-parameter Santoro-Bolen equation (37) will produce a relatively good fit to any reasonably sigmoidal curve, regardless of whether the system is two state (166).

Baseline curvature such as we observe in the post-transition region of Figure 5.1, may also cause the Santoro-Bolen equation to produce misleading results. Using only the data between 0.2-1.2 M GdmCl, the Santoro-Bolen stability estimate increases to 6.5 kcal/mol (see Figure 6.3, page 110). Additionally, the baseline curvature may indicate a second transition.

We turn to NHX experiments for reliable estimates of global stability, as has been done before (140). When assessed by HX NMR in 2D, (page 36), we find 21 amides (NMR cross peaks) that do not appreciably exchange (< 5%) in 24 hours at pD 9.6, nor do they exchange in 2 months at pD 7.5 (data not shown). Thus, we can suggest a lower limit for the global stability to be ~15.0 kcal/mol, consistent with the estimates provided by calorimetric studies. Native stability was not the focus of this work; we are interested in the denaturant melt because it allows us to draw inferences about the burst behavior as discussed later.

Kinetic Refolding at pH 9.0

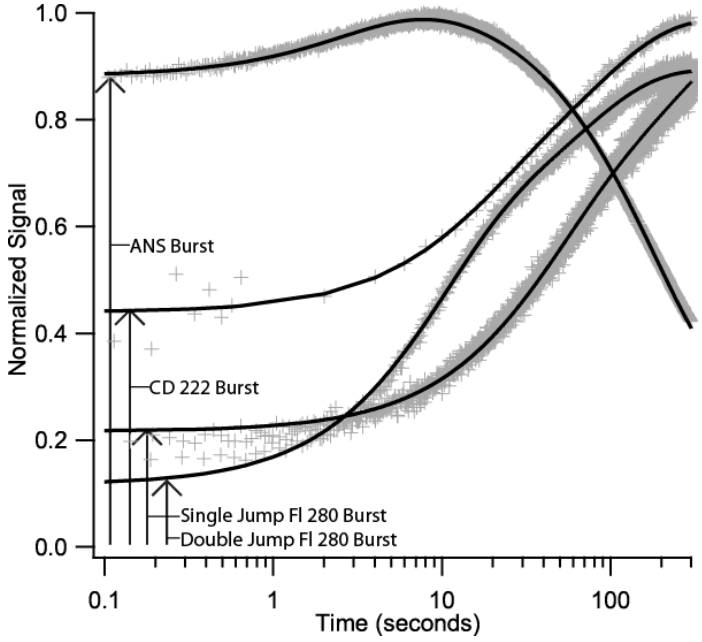

Figure 5.2: MBP folding and burst amplitude assessed by a variety of optical probes.

D-MBP (10 µM) was unfolded in 2 M GdmCl and then diluted 1:9 into pD 9 buffer to initiate refolding, subsequent kinetics were observed by a variety of optical probes shown in Figure 5.2 – solid lines drawn through the data represent double exponential fits to the data. With the exception of the ANS binding data, all measurements shown in Figure 5.2 were normalized to U and N endpoints[19]. Due to the growth and decay phases in the ANS data and

[19] We observed a low level of photo-bleaching to different extents in the tryptophan fluorescence traces (excited with 280nm photons) necessitating a point-wise normalization scheme, $Norm(y(t)) = y(t) - \hat{y}_U(t) \Big/ \hat{y}_N(t) - \hat{y}_U(t)$; for tryptophan fluorescence, N and U baselines were fit by exponentials,

because both endpoints (U=0.44, N=0.53) gave lower signals than the observed data, the signal intensity directly from the photomultiplier for ANS data is multiplied by a constant such that the maximum is coincident with 1.0 on the y-axis. Within the dead time[20] of spectroscopic observations, MBP exhibits a fast initial burst-phase increase in tryptophan fluorescence, ANS binding, and the formation of ~40% of its native circular dichroism signal (CD_{222}) corresponding to as much as ~20% helical content. The burst-phase signal persists for approximately 500 ms, as if it has reached pseudo equilibrium before subsequent changes in signal are observed.

		Single Jump		Double-Jump	ANS
		CD	Fl	Fl	*U = 0.44, N=0.53*
	Burst Amplitude	44%	20%	13%	*0.46 (+)*
λ_1	*Amplitude*	21%	18%	44%	*0.08 (+)*
	Lifetime	15.1 ± 2.1 sec	16.0 ± 0.5 sec	9.3 ± 0.1 sec	3.6 ± 0.01 sec
λ_2	*Amplitude*	34%	62%	43%	*0.47 (-)*
	Lifetime	80.5 ± 6.3 sec	92.8 ± 0.8 sec	74.1 ± 0.7 sec	159.5 ± 0.5 sec

Table 5.1: Fit parameters for kinetic optical data shown in Figure 5.2. *Italicized values represent fits to non-normalized values.* For ANS data, the signs + and – represent growth and decay, respectively; the N signal for ANS fluorescence is equivalent to the fit value for y_0 and the U signal was determined by averaging the unfolded baseline.

We fit the optical data to double exponentials (solid lines in Figure 5.2), $y(t) = y_0 - A_1 e^{-\left(\frac{t}{\tau_1}\right)} - A_2 e^{-\left(\frac{t}{\tau_2}\right)}$, and determined the burst-phase amplitude by $y_0 - A_1 - A_2$; the fit parameters along with errors (95% confidence band) are given in Table 5.1. These results are similar to those reported in the literature (107) for single jump experiments; no double jump refolding data has been published.

A double jump experiment allows one to test for proline mis-isomerization effects on folding kinetics. As molecules are only unfolded for a few seconds in the double jump, proline residues do not have time to re-isomerize; thus, the effect of proline mis-isomerization on relaxation constants may be eliminated. We find a small difference in relaxation properties

variables with hats represent values predicted from this fit. For CD and ANS data, variables with hats represent the average value of the N and U baselines. This was done to avoid propagating error.

[20] 23 ms for CD, 3.2 ms for Fl double jump, 8.3 ms for Fl single jump and ANS. Early time points shown in Figure 5.2 were symmetrically averaged (0-200ms) such that the first point was 100 ms to increase signal-to-noise. We found that the burst-phase was unresolved at our shortest dead-time of 3.2 ms.

between single and double jump experiments, apparently due to proline mis-isomerization; but, the differences are not substantial, the protein still folds slowly.

Fast Initial Compaction & The Burst-phase

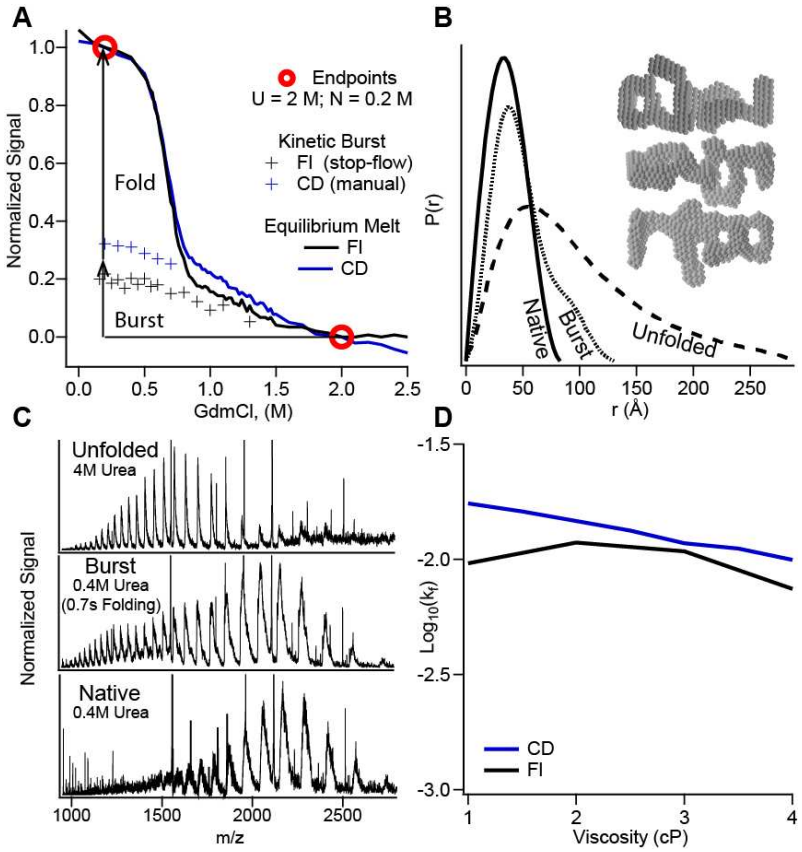

Figure 5.3: Exploring the burst-phase in MBP. **(A)** Fitted (burst) amplitudes from a kinetic refolding experiments (pH = 9, 20 °C, Figure 5.2) with normalized melt data for comparison (same data as Figure 5.1). **(B)** Experiments collected by SAXS measurements and (inset) envelope reconstructions for the burst species. **(C)** Charge-state distribution experiments (see text for description). **(D)** Viscosity dependence, of folding assessed using glycerol to increase viscosity and reported in units of centipoise. k_f was determined using a single stretched exponential[21] over two rounds of fitting. First by floating the stretching parameter (β) and rate for each condition and then refitting the rates with β fixed to the average value from the first round of fits (β = 0.63 as shown here).

As discussed in Chapter 1, unfolded states may gradually expand in response to increasing concentrations of chemical denaturants. Extrapolation of the unfolded baseline measured in thermodynamic melting experiments has been used to test for this, if the burst amplitude is equivalent to the baseline extrapolation, one concludes that the burst-phase signal change does not represent a barrier-crossing event. In the denaturant melt (Figure 5.1) there is

[21] $y(t) = \left\{ y_0 - Ae^{-(k_f t)^{\beta}} \mid 0 < \beta \leq 1 \right\}$

no clear demarcation between the cooperative transition and the unfolded baseline therefore it is difficult to decide on what data to use for the baseline. As a qualitative test, in Figure 5.3A, we normalize the melt data (blue and black lines for CD and Fl data) to kinetic endpoints (N=1, U=0; red circles) and compare it to burst amplitudes (plus symbols) observed in a refolding denaturant series. The kinetic burst amplitudes seem to follow the same trend as the post-transition equilibrium measurements and the post-transition shoulder that appears in both CD and Fl melt data may represent a cooperative transition. Perhaps the burst represents some cooperative folding event. Structural information is required to determine the cause for burst behavior observed in MBP.

We collected small angle x-ray scattering data to determine the R_g (radius of gyration) of the mature burst species (Figure 5.3B). In these experiments, 11-second exposures were collected under continuous flow conditions with an experimental dead time of 0.7 seconds. Native and Unfolded profiles were collected using standard SAX equilibrium protocols. In 2M GdmCl, unfolded MBP has an R_g 73.3±0.5 Å, similar to the value expected for a random coil of appropriate length, 69 Å, computed by $R_g=2N^{0.6}$ (36). The native structure has an R_g of 22.3±0.7 Å in 0.2M GdmCl, this is practically identical to the expected value of 23 Å calculated from the crystal structure (PDB ID: 1OMP). After 0.7±0.1 seconds of refolding, the measured R_g has dropped to 36.4±1.1 Å representing 75% of the total ΔR_g between native and unfolded states. The P(r) curves clearly demonstrate the near-native compaction occurring early during refolding experiments. Three representative envelope reconstructions generated from the P(r) distribution are shown in the Figure 5.3B inset. Unlike the image brought to mind by term *molten globule*, this species in MBP appears to be polyglobular, characterized by multiple small clusters likely stabilized by hydrophobic side-chain contacts.

Figure 5.3C shows the charge state distribution (CSD) produced by injecting MBP by electrospray ionization (ESI) into a mass spectrometer (LTQ orbitrap XL) within ~50 ms of initiating folding. The spectrum is shifted from the high charge state pattern characteristic of unfolded protein toward the much lower charge state distribution of the native protein, consistent with a significant compaction and reduction in surface exposure to solvent (235, 236). A small population fraction with CSD like that of the unfolded protein is also seen but it can be noted that the populations measured in this way are greatly biased toward exaggerating the

more unfolded component (235). This indicates that the collapsed conformation(s) observed in SAXS experiments at 0.7 seconds of folding time are perhaps present in the initial burst-phase.

Others (232) have suggested that collapsed states should have relatively high internal friction because of the random intrachain interactions that must form to stabilize collapsed ensembles. If this were the case, one might expect the folding rate to be relatively insensitive to solvent viscosity because chain diffusion rates would depend on first breaking spurious contacts formed in any given collapsed state. CD and fluorescence data show virtually no dependence of folding rates on solvent viscosity (Figure 5.3D). This observation indicates that polypeptide reconfiguration during the conformational searching that ultimately organizes the native structure is not limited by diffusional searching of the polypeptide chain through free solvent but rather by the difficulty of conformational reorganization within and between condensed polyglobular regions. Significant differences between our findings and timescales suggested for restricted chain diffusion in smaller proteins commonly attributed to so-called internal friction (237, 238) should be appreciated. The present time scale is nine orders of magnitude slower than is observed in small molecules, in part due to the degree of chain collapse, but also because the folding event measured here requires a specific nearly simultaneous multi-point interaction rather than a general two-point interaction as for example in a FRET experiment.

Conclusions from Optical Experiments

In summary, many measurements agree that, upon dilution from unfolding denaturant, MBP experiences a fast molecular collapse into an ensemble of compact polyglobular forms and then folds slowly in a way that is limited by the difficulty of chain reconfiguration therein. The optical burst signal is most likely the result of transitioning from an expanded random coil state to the collapsed state upon dilution from high denaturant concentration. However, these widely used methods only monitor whole molecule behavior. They provide little detailed information about structure in the compact state or the folding mechanism that produces the native state.

5.2.2 Kinetic Pulse Labeling

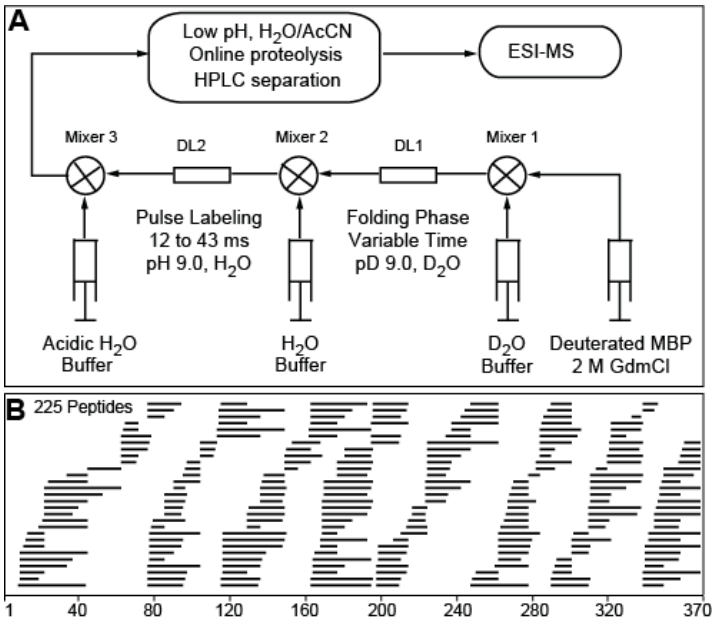

Figure 5.4: KHX experiment diagram and MBP peptides used in this work (*figure taken from reference (3)*). **(A)** Experimental work-flow for pulse labeling experiments. The box labeled "Low pH, H_2O/AcCN…" represents the on-line system shown in Figure 4.1 on page 52. **(B)** 225 unique peptides in the peptide pool for pepsin-proteolyzed MBP.

To achieve a structural picture of the folding process, we employed a quench-flow HX pulse labeling experiment, diagramed in Figure 5.4A, measured by mass spectrometry. During kinetic refolding, D-MBP, unfolded in 2 M D_6-GdmCl, is diluted 1:9 into a pD 9.0 D_2O solvent initiating refolding at 0.2 M D_6-GdmCl. After variable refolding time, the labeling pulse is applied by further diluting the sample 1:4 with a pH 9.0 H_2O buffer for a duration of 12 to 43 ms at 20 °C. The average exchange lifetime of an unprotected amide is ~1 ms in these conditions. Amides in pre-existing H-bonds when the pulse is applied tend to retain the deuterium label. Immediately after the pulse, the labeling reaction is slowed by pH reduction and cold temperature (section 4.3, page 60), the sample digested and washed, the fragments separated by HPLC, and their mass distributions measured by MS (section 4.2, page 50).

We find 225 unique peptides from pepsin proteolysis (often with multiple charge states) that each monitor the folding behavior of the protein segment from which it was derived. These peptides are shown with respect to their location in the MBP primary sequence in Figure 5.4B.

The peptides are identified as described in Chapter 4 and the extent of labeling determined by the HDpop program (section 4.4.2, page 73). We only used 116 peptides of the highest quality regarding signal-to-noise and with a larger separation in mass between populations to make inferences about MBP folding in this work; within this set, every peptide had at least one overlapping peptide to provide consistency checks.

In the discussion below, it will be useful to recall earlier discussions regarding HX MS state sensitivity introduced in Chapter 2 on page 21. For HX MS pulse labeling experiments, we collect an unfolded control, pulsed in 2M GdmCl, and this identifies the lightest population observed for any given peptide. We also collect a native control by pulsing a fully deuterated sample that has been refolded on the bench for one hour prior to pulse application – this mass distribution reflects how the peptide will appear if properly folded. All data is corrected for back exchange as indicated in Chapter 4. The relative areas of each sub-population in a given mass spectrum reflect the partition of molecules between observed states (recall Figure 2.2 on page 22), identified by HDpop at the 99% confidence interval, and the centroid masses of each, after back-exchange correction, indicate the number of protected sites.

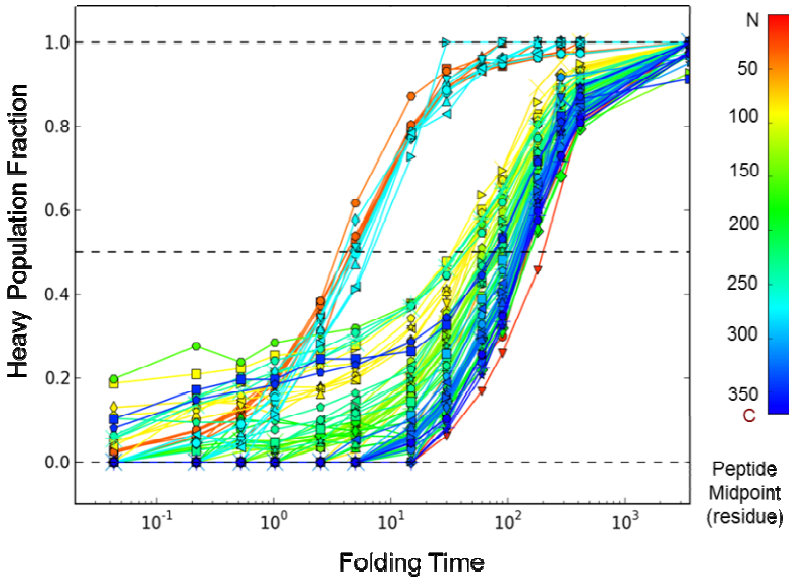

Figure 5.5: The heavy population amplitude vs. folding time for 116 peptides.

By following the relative area of the heavy (native) population during a folding experiment, one is able to estimate the folding lifetime of each peptide. The entire time series,

pulsed for 43 ms at each time point, is shown in Figure 5.5 and peptides are colored by the midpoint of exchangeable sites to show their positions in the protein sequence. This data is analyzed first by breaking the peptides into categories based on the time dependence of their protection pattern.

Early HX Protection

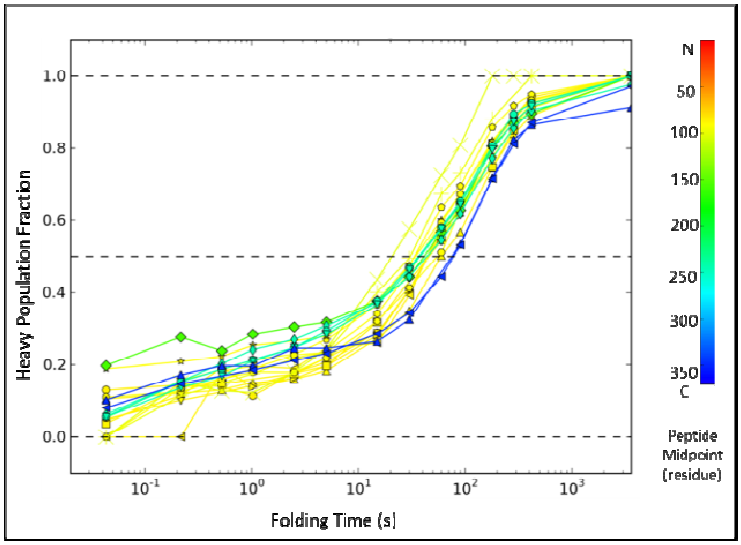

Figure 5.6: Peptides that show early protection (13 peptides with high signal-to-noise are shown)

	Location	D's
Region 1	90-97	7
	104-111	3
Region 2	149-162	3
Region 3	235-247	6
	251-262	4
Region 4	340-347	6 or 7
	Total	**30**

Table 5.2: Regions of the molecule identified by peptides in our KHX experiment that show 20-30% protection early during refolding experiments.

Incomplete protection was observed in the thirteen peptides shown in Figure 5.6, these peptides segregate into four distinct regions of the protein. By comparing these peptides with others, we are able to define sequence boundaries, given in Table 5.2, for this behavior. Only 20-30% of the molecules were found in the protected heavy populations for these peptides; however, two regions (1 and 4) stood out because the numbers of protected deuterium were

equivalent to that seen in their respective native controls, the heavy populations in regions 2 and 3 were less protected than native.

Incomplete protection could not be the result of a fraction of the molecules reaching the native state on a fast folding track because we did not similar behavior in the other peptides. There are other possibilities. The fractional behavior could be the result of 20-30% of the molecules in these regions being protected, while the others are not. Alternatively, it could mean that 100% of the molecules were protected during early folding in these 4 regions but 70-80% of those molecules opened and exchanged during the 43 ms pulse. In most other peptides the light population mass did appear to grow slightly heavier in the first second of folding – this would be expected if spurious low-level protection resulted from an aspecific increase in chain density. We could not decide between these explanations without measuring structural opening rates.

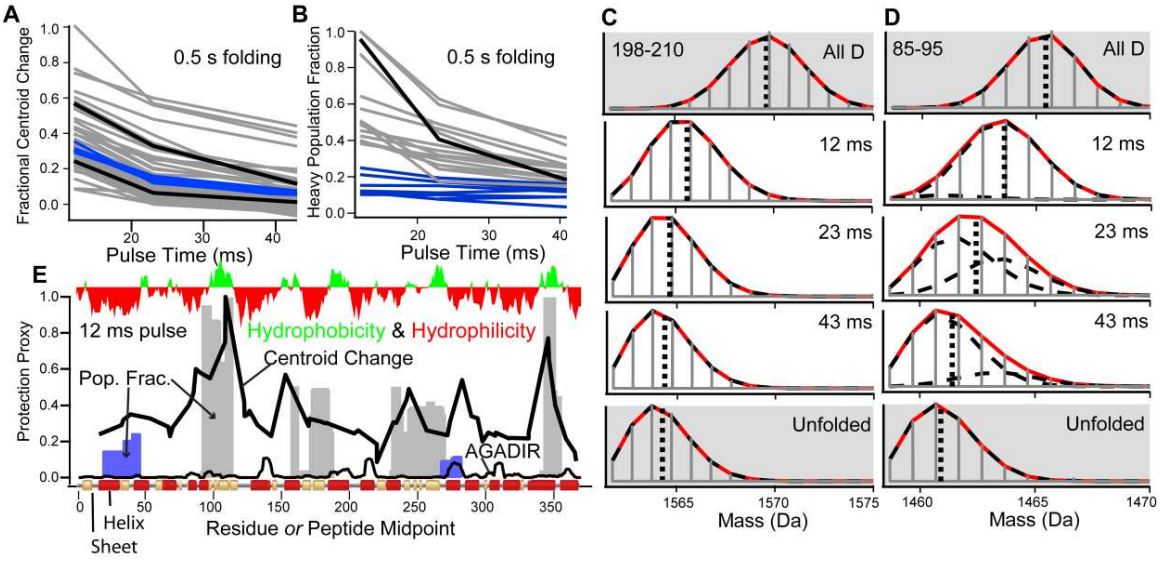

Figure 5.7: Modulation of the pulse time at a fixed folding time *(reprinted from reference(3))*. **(A)** All peptides, including those with and without two populations are plotted in terms of their fractional change in centroid mass with respect to the unfolded and fully deuterated endpoints. **(B)** Peptides that display bimodal spectra are plotted as the population fraction protected versus the pulse length (strength; blue traces represent peptides that form the 7 s intermediate (already 7% folded). **(C,D)** Non-bimodal and bimodal data, identified by black lines in panels A & B, are shown to illustrate respective spectral responses to pulse modulation. Centroids are indicated with vertical dashed lines and HDpop envelope fits in red with sub-population(s) shown in black dashed lines. **(E)** A visual presentation of the 12 ms pulse length data. Peptides with two populations are shown as bars with heights representative of the fraction of molecules heavy (ie. panel B & D) and fractional centroid values (ie. panels A, C-D) are given for all

peptides with a solid black line, each drawn through peptide midpoints. AGADIR helical propensities, and Kyte-Doolittle hydrophobicity[22] (15 – residue sliding window) are also shown.

To explore these possibilities, we modulated the pulse duration between 12 and 43 ms and held the folding time fixed at 0.5 seconds; the results are shown in Figure 5.7. For all peptides, regardless of whether they have two populations, the mass centroid taken for each pulse length was scaled to the mass centroids of experimental endpoints[23] that are unaffected by pulse length (0s and All D). Fractional centroid values are shown with respect to pulse time in Figure 5.7A. Representative unimodal and bimodal mass spectra are shown in Figure 5.7C-D – the centroids of both example peptides, shown in black in Figure 5.7A, are represented by the vertical dashed lines for each pulse time in panels C and D. The majority of one-population peptides exhibited low-level protection with roughly equivalent responses to reduction in pulse time; however, this was not universal – the thirteen peptides from Figure 5.6 with two populations at 0.5 seconds of folding deviated from this trend.

For all peptides with clearly two populations at 0.5 s, we are able to plot, in Figure 5.7B, the change in heavy and protected populations with respect to pulse time. Blue lines represent peptides which go on to form the 7 second intermediate, these areas of MBP are already 7% folded at 0.5 seconds of folding and are discussed later. From the four regions in Table 5.2 that show 20-30% of the molecules in a heavy population in the 43 ms pulse, the same two regions (1 and 4) that stood out earlier appear to be fully protected in the 12 ms pulse. Many overlapping peptides confirm this observation.

What can be said about these observations collectively? The fact that we see two discrete populations does indicate regions of MBP exchange by some cooperative unfolding mechanism and the similar responses to pulse time reduction in regions 1 and 4 could be suggestive that both are members of the same cooperative structural unit. These two segments of primary structure do not interact in the native protein; thus, we do not expect this to correspond to native-like structure but cannot eliminate the possibility that this represents an early, non-native, low stability intermediate present in almost every molecule. The other regions

[22] Computed using the setting "Hphob. / Kyte & Doolittle" on http://web.expasy.org/protscale/ with a window size of 15 residues.

[23] Centroids in these endpoints do not respond to changes in pulse length, thus we were able to use the information for scaling purposes.

of the protein that exhibit low levels of protection appear to be engaged in heterogeneous H-bonding with low stability because of severe steric restraints in the collapsed state and the energetic penalty of non-paired potential H-bonds[24]. This can be demonstrated by analyzing the centroid changes in unimodal peptides with respect to a uniform structural protection factor, such as one expects for cooperative protection.

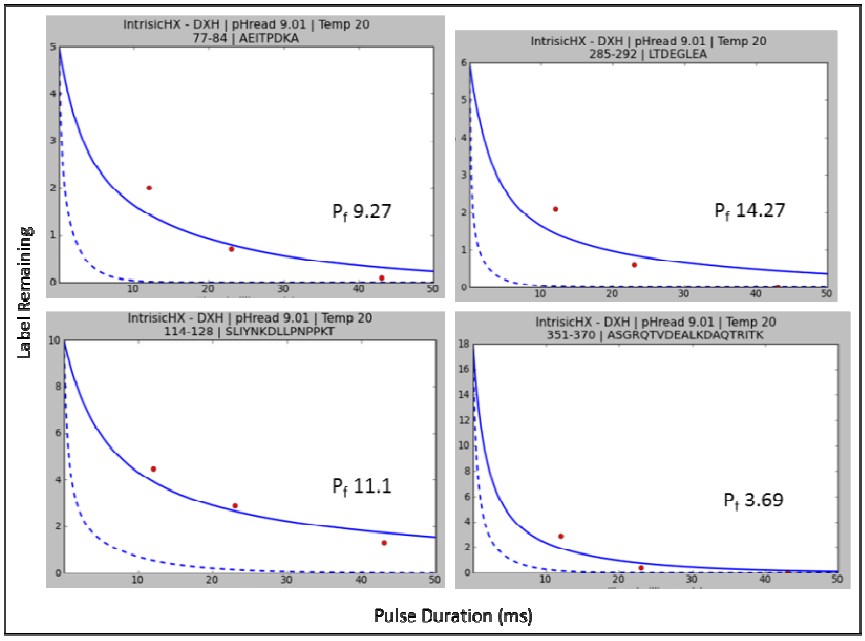

Figure 5.8: Broad level protection at 0.5 s folding time. The abscissa represents pulse time at the fixed folding time of 0.5s, the ordinate contains the amount of deuterium present after correction for D_2O in the labeling pulse and for back-exchange during preparation. In each, the residues and sequence of the peptide are written in the header, the dashed line shows the expected amount of deuterium if the peptide was solvent exposed, the solid blue line represents the protection factor that results from structural slowing. Fractional protection is indicated by text in each plot.

Further analysis of the centroid changes in peptides with only one discernible population highlights the heterogeneity in protection (Figure 5.8), where we see a non-uniform 10-fold slowing[25]. The deviation between solid blue curves, computed for the uniform protection factor, and the actual data points shown in Figure 5.8, indicates a large spread in protection factors within each peptide. We interpret low-level non-uniform protection to

[24] This does not imply that the collapsed state is dehydrated; but, the local concentration of water is likely reduced compared to more expanded conformations and perhaps promotes random intrachain H-bonds.

[25] Slowing factors (P_f) are computed as follows: $Label(t) = Sites - \sum_{i}^{Sites} e^{-\frac{k_{ch,i}t}{P_f}}$

represent rapid reconfiguration of random H-bonding patterns in the polyglobular state, behavior that is reminiscent of what might be expected in the Vijay Pande kinetic hub model of protein folding (239). The two regions of the molecule with apparent correlated protection are not expected in the kinetic hub model, this behavior suggests classical pathway-directed folding. We are able to hypothesize that that chain reconfiguration times within the randomly collapsed high-energy polyglobule are likely on the millisecond time scale (ie opening reaction lifetimes are ~10 ms) and that opening reactions are generally followed by H-bond reformation times faster than ~1 millisecond.

Burial has been shown to drive H-bonding (240); earlier work by Ken Dill demonstrated that collapsed but unfolded chains will create an environment that favors helix formation (42); both regions that near 100% protection in the 12 ms pulse are capable of forming amphipathic helices. It is likely that the energetic need to satisfy H-bonding requirements of the polar backbone amide group accounts for the large CD_{222} burst shown in Figure 5.2; however, the regions showing substantial protection do not correlate with AGADIR predictions for helical propensity (241-243). Perhaps the four regions in MBP that show early protection and especially those that appear to be completely protected at the shortest pulse lengths have a higher probability of helix formation in the collapsed polyglobule. Figure 5.7E shows that these regions with nearly 100% protection are geometrically correlated with two of the four regions of high hydrophobicity in MBP (see footnote 22, page 93); it is possible that the high hydrophobicity of the chain in these sections results in a higher probability for burial of these segments during collapse.

Consistent with the idea that a compact relatively unstructured polyglobule should promote random H-bonding interactions throughout the molecule, we observe only one population in the majority of peptides and their mass centroids progressively grow heavier with a reduction in pulse duration after 0.5 s of folding time. This data indicates heterogeneous low-level protection throughout the molecule and is highlighted by the black line in Figure 5.7E representing the fractional centroid value for all peptides at the 12 ms pulse.

These results indicate widespread, diverse, low-level HX protection among the various sites within each protein molecule and for any given segment distributed throughout the

protein population. This behavior is consistent with the development of heterogeneous non-native molecular collapse in the initial denaturant dilution step. Apparently, structural collapse is too fast to allow native-like structure formation initially, and produces a condition like that in the equilibrium melting experiment after the main melting transition (Figure 5.1 & Figure 5.3A). Distinct, stable structure forms on a much slower time scale, suggestively because the randomly collapsed milieu interferes with subsequent conformational searching.

The 7 s Obligatory Intermediate

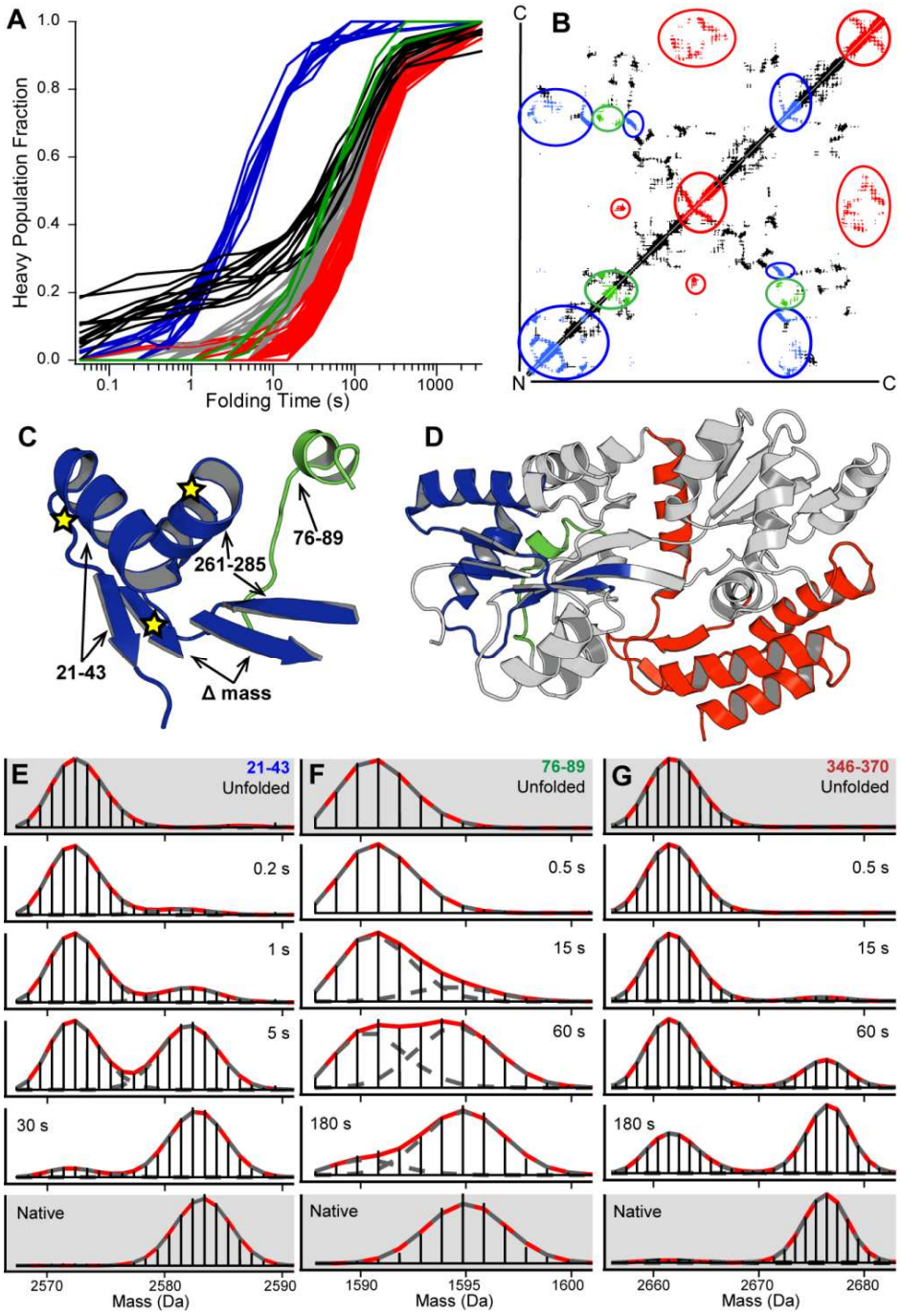

Figure 5.9: Intermediate and Slow Folding Behavior for MBP (*figure taken from reference (3)*). **(A)** Peptides originally shown in Figure 5.5 are now recolored to represent our interpretation of this data. Blue traces are peptides from the 7 s intermediate and black traces represent early folding peptides shown in Figure 5.6. Others that fold in the very slow grouping are graded from early (green), through gray, to late (red). **(B)** Contact map colored to match panel A. **(C, D)** Placement of the blue, green, and red segments in native MBP, stars indicate the sites of slow folding mutants, see text. **(E-G)** HX mass spectra showing the time-dependent folding of color-marked segments in panel A.

From the dynamic polyglobular collapsed state, a stable, obligatory intermediate emerges with a lifetime of 7 seconds. Fifteen peptides with high signal-to-noise are shown in blue in Figure 5.9A – their positions in the native state are colored blue in panels B-D (with exception of the two strands in panel C labeled Δ mass). The MS data for one of the fifteen fragments is shown in Figure 5.9E, comparing the heavy mass distribution with the native mass distribution leads one to conclude that the number of protected deuterons roughly matches the number of deuterons protected in the native structure. We interpret this concerted transition to represent the formation of an obligatory intermediate comprising the core sheet and two flanking helices of the N-domain. We observe a small mass shift in the heavy population of this peptide (+3 deuterons) occurring after the 7 second behavior. This likely reflects that the opening rate of the protecting structure is reduced as additional structure adds onto the 7 s intermediate and increases its stability. The pulse time series performed to understand the earlier folding behavior could verify this conjecture if performed for later folding times.

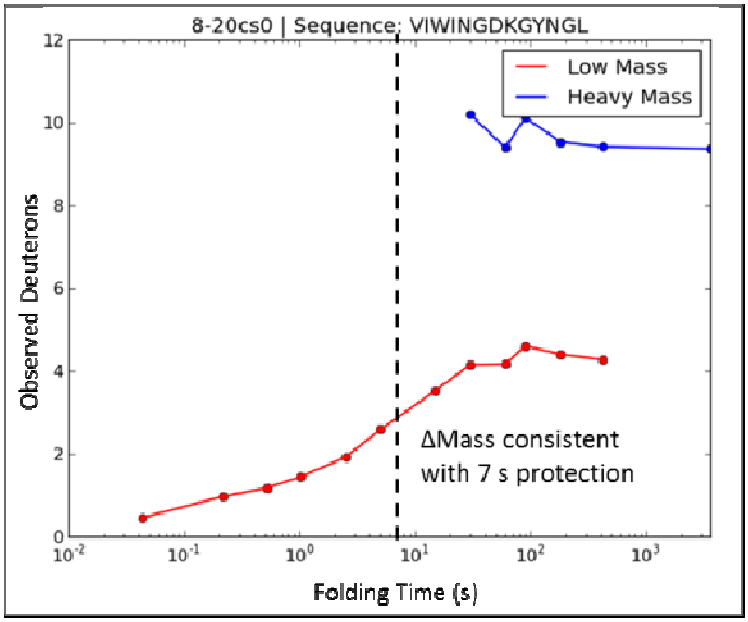

Figure 5.10: Mass shifts provide structural information. A *cumulant distribution (see text)* formed by two unresolved populations exhibits a characteristic mass shift as the molecules transition from one to the other unresolved population. A fragment that reports on residues 8-20 exhibits a mass shift that coincides with the 7 s time constant of the blue traces in Figure 5.9A. This information, along with a similar observation in overlapping fragments for the two strands highlighted as Δ mass in Figure 5.9C implies that these two strands form in the 7 s intermediate.

In Figure 5.9C two strands are labeled "Δ mass" to reflect that their inclusion in the intermediate was determined on the basis of mass shifts, not population transitions; these two

strands can be observed to protect by paying attention to their unfolded populations centroid masses. As first mentioned in footnote 6 on page 23, if two or more populations are not sufficiently separated, HDpop will default to a single cumulant population whose time dependent change in mass reflects the fractions of molecules (f) in each unresolved subpopulation. The mass of a cumulant population is described by $\langle Mass \rangle_{cumulant} = f_1 \langle Mass \rangle_1 + f_2 \langle Mass \rangle_2 \,/\, f_1 + f_2$. In general, mass shifts in this dataset are restricted to the heavy population and reflect stability increases during folding (i.e. Figure 5.9E, heavy population); however, for the two small β strands in the 7 s intermediate, we observe a very rare light population mass shift, shown in Figure 5.10, for peptide 8-20. Inferring population transitions from mass shifts are less convincing than if we had measured those population transitions directly; but, is hard to imagine how the other strands in Figure 5.9C, with resolved population transitions, could occur without these two other strands also forming simultaneously. All evidence is consistent with this interpretation.

The slow native transition

The remaining peptides transition in a two-state although not accurately exponential way to a highly protected form, on a 60-120 s time scale (Figure 5.9A, red, grey, green). This time scale matches the spectroscopic folding kinetics in Figure 5.2 and Table 5.1, which tracks native state formation. The spread of folding rates among the late folding segments is clearly broader than is seen for the 7-second class (Figure 5.9). Earlier and later folding peptides are colored. The bimodal transition behavior of peptides drawn from the earlier and later regions are shown in Figure 5.9F&G. Folding halftimes for the differently colored regions are ~55 s for the earlier (green) peptides and ~100 s for the later (red) ones.

The spread suggests that different protein segments fold with somewhat different time signatures, but the large number of peptides with apparent half-times between 60-120 seconds are too close to resolve into clearly separate groupings. It is interesting that peptides that occupy the earliest part of the spread (green) are adjacent to the 7 s intermediate structure in the native protein, as would be expected from the sequential stabilization mechanism (59). Similarly, the slowest folding peptides shown in red are remote and even in the other C-terminal domain of native MBP.

5.3 Impact of This Work

This work used standard optical techniques and a developing HX pulse labeling method to study the folding of the large, 2-domain, 370-residue maltose binding protein. Upon mixing into folding conditions, the unfolded MBP polypeptide quickly condenses. The results characterize the condensed state and display the subsequent formation of an obligatory on-pathway intermediate and the even slower folding to the native state.

5.3.1 Insight into Earlier MBP Folding Work

N-domain mutants (V8E, Y283D) located within the 7 s intermediate (stars show their locations in Figure 5.9C) have been previously found to slow folding and increase aggregation propensity (106). In one case, the malE31 double loop mutant (G32/I33 to D32/P33) reduced the earliest resolvable Fl_{280} lifetime to 110 s and almost all of the protein, upon over-expression, aggregated into inclusion bodies. This mutant has little effect on the unfolding rate (107). In the same work, a similar loop double mutation made in the C-domain has little effect on the earliest resolvable F_{280} lifetime, does slow the later lifetime, and does not result in increased aggregation propensity. The C-domain mutant increases the unfolding rate. One way to interpret these observations would be to suggest that unfolding of the C-domain is rate limiting for unfolding studies, and that formation of the 7 s N-domain obligatory intermediate may protect the protein from aggregation. When the aggregation sensitive collapsed state is present for an extended amount of time, by these N-domain mutations, the concentration threshold for aggregation is reduced compared to that of the wild-type protein; therefore, increased aggregation propensity is observed.

Inquiries into the increased aggregation propensity resulting from these N-domain mutants began many years ago (106). These mutants were combined (244) in GroE chaperone studies. Where the wild-type protein does not interact favorably with GroE, the slow and aggregation prone N-domain mutants experience folding rate enhancement by interaction with the GroE chaperonin system (129). Later it was shown that this rate enhancement is due to the prevention of aggregation (245). It is clear that N-domain mutants located within the 7 s intermediate disrupt folding; thus, we are able to suggest that the conformation that is susceptible to aggregation is the aspecific polyglobule.

A rate-limiting step (RLS) for MBP folding has been hypothesized to be the formation of N-domain structure (106), this hypothesis has been repeated in many other manuscripts so much that it has become something of a fact. More recently (246), authors suggested that the first step of MBP folding involves a concerted core assembly in both domains. Although these were reasonable conclusions, in both cases, the experiments were wholly incapable of addressing the question; our work is able to address these statements directly and find both to be incorrect.

Because we are able to define its structure, the 7 s obligatory intermediate suggests that the core of both the N and C domains are not formed in a concerted manner; the N-domain core forms first. We reconcile the observation that N-domain mutations (in the 7 s intermediate) slow both phases of folding whereas C-domain mutants only affect the slower of the two relaxation constants (107) by being able to show that the N-domain intermediate is obligatory – slowing formation of this intermediate will slow everything subsequent to this step. We are also able to reject that formation of the N-domain core is rate limiting because the N-domain core forms in 7 s. Folding of the C-domain is much slower suggesting that the RLS for folding lies in the formation of C-domain structure. A preponderance of evidence suggests that all molecules in transit to the native state must form the N-domain intermediate that we have identified at 7 seconds.

5.3.2 Protein condensation

A quantity of work on relatively small proteins has focused on the character of the unfolded state and its possible role in guiding subsequent folding (49, 76, 130, 232, 247, 248). Is the denatured state ensemble compact under native conditions at the start of the folding process? Is significant pre-folding structure present? If so, does it help to guide or hinder the folding process? These questions are significant for understanding protein folding and other biophysical properties such as the quality of water as a polypeptide solvent, the energy balance between the favorable drive to occlude hydrophobic surfaces and the unfavorable entropy of chain collapse, and the accompanying requirement to satisfy the H-bonding propensity of polar groups that become buried in the collapse step.

The methodology used here provides some answers, although large proteins like MBP with many more hydrophobic interaction possibilities are likely to bias more toward the collapsed condition in initial folding (249, 250). When unfolded MBP is mixed into folding conditions, the polypeptide chain rapidly condenses to a polyglobular form. Low level HX protection in the loosely compacted chain indicates that structure is dynamic, relatively unstable, and heterogeneous. Hydrophobic interactions that drive condensation bring together sites that allow ANS-to-protein binding and can similarly promote protein-to-protein aggregation. This includes binding of the exposed hydrophobic sites of apparently condensed but still unfolded proteins to the apical hydrophobic sites of GroEL. Large proteins like MBP are the common client substrates of the GroE chaperonin system and other chaperonins (251) which function to bind and shield them from aggregation during their vulnerable slow-folding time period (252). The present characterization of the MBP pre-folded state supports this view.

Over time the entire MBP refolding population self-organizes, forming a distinctly structured, native-like intermediate, and then proceeds to assemble the native state. It appears that the compacted condition contributes to the slow folding observed, in part because it constrains chain reconfigurational searching. Results show that the weakly H-bonded interactions detected in the earliest collapse do not contribute to formation of the early intermediate and probably not to later folding steps.

5.3.3 Nature of the Protein Folding Pathway

Figure 5.9E, F & G show isotopic envelopes for peptides that help to define the 7 s intermediate (blue), the following step (green), and a slowest (red) peptides. These spectra directly show that 100% of the population adopts each of these structures. Given that the pulse time was 43 ms (pH 9, 20^{o}C, where intrinsic HX lifetime is ~1 ms) and the D-occupancy is maintained at the native level, the protection of these structures against pulse labeling is > 100, corresponding to > 3 kcal/mol of stability assuming EX2 behavior, and/or greater than 150 ms unfolding lifetimes. In comparison, the D-occupancy observed for peptides protected in the burst-phase decreases even for shorter pulse lengths indicating protection stability < 1 kcal and unfolding lifetimes on the order of 12 msec.

In the early history of the protein folding field, it was assumed that proteins fold through discrete intermediates in discrete pathways, like other biochemical pathways. Early theoretical efforts to study protein folding mechanisms led to a view of the pre-folded protein that seems reminiscent of the initial condensed state studied here. It was inferred that proteins then fold through multiple pathways (71, 73-82). In spite of a dearth of experimental verification, this view is still current and experimental as well as theoretical folding results are often phrased in this language. Some spectroscopy-based experiments have been taken to suggest a small number of folding pathways (85, 86, 253-255), but it has been shown that this kind of data cannot distinguish alternative parallel pathways from a given pathway with alternative misfolding barriers (59, 256).

A quantity of more recent experimental work using hydrogen exchange and associated methods has found much more organized folding behavior. Many proteins have been found to form at least one specifically structured native-like on-pathway intermediate (110-116, 120-124, 257-259), and even more impressively an organized folding sequence that progressively assembles the native protein (5, 9, 48, 116, 126). The present work used an advanced mass spectrometry analysis to extend HX pulse labeling folding studies to a larger protein with multi-state folding where multiple pathways, if they exist, should be more evident. The results clearly define the formation of an initial obligatory native-like intermediate and the subsequent formation of adjacent structure apparently in a stepwise sequential stabilization way. This work adds to the growing list of experimental demonstrations that proteins tend to fold through distinct intermediates in distinct pathways. Recent progress in physically-based molecular dynamics simulations now finds similar, repeatable folding pathways for a number of small proteins (67, 260).

5.4 Methods

5.4.1 Protein Purification

The version of *E. coli* apo-MBP (Protein Data Bank (PDB) ID: 1OMP) used herein is the wild-type mature protein without its 26-residue leader sequence. Expression and purification have been described previously (106).

5.4.2 Optical Experiments

Equilibrium melting experiments used 0.8 μM [MBP] in a 20 μM borate buffer at pD 9.0. Following a change in denaturant, samples were allowed to equilibrate for 20 minutes prior to a 20-second signal acquisition. Kinetic refolding experiments were collected using the same conditions for refolding as described below for pulse labeling. For measurement of ANS binding, 126 μM of ANS (8-Anilino-1-naphthalenesulfonic acid) was added to refolding buffers, 380 nm photons were used for excitation and emission recorded using a long wavelength pass filter (CWI Mellties Griot) with 50% transmission at 450 nm. Double jump experiments were performed by first diluting 10 μM deuterated protein into 3 M D_6-GdmCl at pD 9 for 3 seconds before a 10-fold dilution into refolding conditions (0.5 μM protein, 0.3 M D_6-GdmCl, pD 9). SAXS experiments were collected using the same refolding conditions employed for pulse labeling, dilution was performed in a home-built t-mixer before passing through the BioCAT (APS, Argonne National Labs) under continuous flow conditions measured for an 11-second exposure to increase signal-to-noise following 0.7 seconds of dead time. These samples were measured by our collaborator Dr. Tobin Sosnick and his graduate student James Henshaw in November of 2012.

5.4.3 KHX MS Experiments

Unfolded MBP, initially fully deuterated at exchangeable hydrogen sites in D_2O, was diluted into folding conditions (pD 9, D_2O, 0.8 μM protein, 0.2 M D_6-GdmCl, 20^oC), allowed to fold for some predetermined time, and then probed by a brief pulse of D to H labeling (usually 43 ms) to obtain a snapshot of the structure that had been formed to that point. Initial dilution into D_2O instead of the usual H_2O was used to avoid back exchange loss of D-label during the lengthy (many seconds) pre-pulse period. For pulse labeling, the refolding protein was diluted by 5-fold into H_2O buffer at pH 9. D-label on amide sites not yet protected by H-bonding exchanges to H during the brief pulse (average unprotected HX time constant ~1-2 msec). The labeling pulse was terminated by dilution into low pH (pH 2.5, 1.2 M GdmCl, ~0^oC) as suggested by recent work (2). For the 0 s time point (unfolded control), the pre-pulse folding phase was eliminated and 2M GdmCl was added to the pulse buffer. For the native control, the refolding dilution was performed manually and allowed to fold for one hour before being pulsed and quenched as described above. To estimate back-exchange, a fully deuterated sample was

collected as described in previous work (2). Immediately following the low pH dilution, the samples were injected into an online flow system (4) where the protein was cleaved into many peptide fragments in an immobilized pepsin column, the fragments were caught in a trap column and washed, and then separated by HPLC (shaped H_2O/AcCN gradients, see Appendix B, were employed (2)) and injected by electrospray ionization (ESI) into the mass spectrometer. The resulting mass spectra were analyzed by the ExMS program (202) to identify the many peptides and our in-house program, HDpop, to measure their remaining D-label. Spectra for multiple charge states of the same peptide were added to increase S/N. The reproducibility of measured bound D per peptide in replicate experiments was in the range ± 0.01 D/peptide. Comparison of data for multiple overlapping peptides allows many internal consistency checks. For structural analysis, we used subsets of peptides with high abundance and signal/noise (116 peptides of the 225 shown in Figure 5.4B).

Chapter 6 –Concluding Remarks & Future Directions

6.1 Summary & Sentiment

Understanding the kinetic accessibility of the native conformation has been a goal of protein folding science since the very beginning. The main problem has been a lack of necessary tools to study the process. We, as protein folding researchers, have learned a great deal from small protein studies; though, large proteins, such as MBP, dominate our proteome and we know very little about their folding dynamics. Experiments that have been sufficient for smaller proteins are not particularly well suited for larger species. In this dissertation, the folding behaviors of the largest protein studied at structural resolution to date are described using an advanced hydrogen exchange mass spectrometry experiment created for this purpose. The experiment provides incisive structural information on the process; here we will take a moment to review the key findings for MBP, their significance, and then address potential future directions for both MBP and large protein folding research in general.

The argument has been raised that because the unfolded state is characterized by random coil conformations, at the onset of folding some of these random conformations should be better suited to fold than others, there should be a spread of folding rates and perhaps a plethora of independent and unrelated routes leading to the energetic minimum of the native state. This explanation must soothe the mind because it has become immensely popular. After all, it is reasonable to ask, how could stochastic chain diffusion give rise to an apparently predetermined pathway?

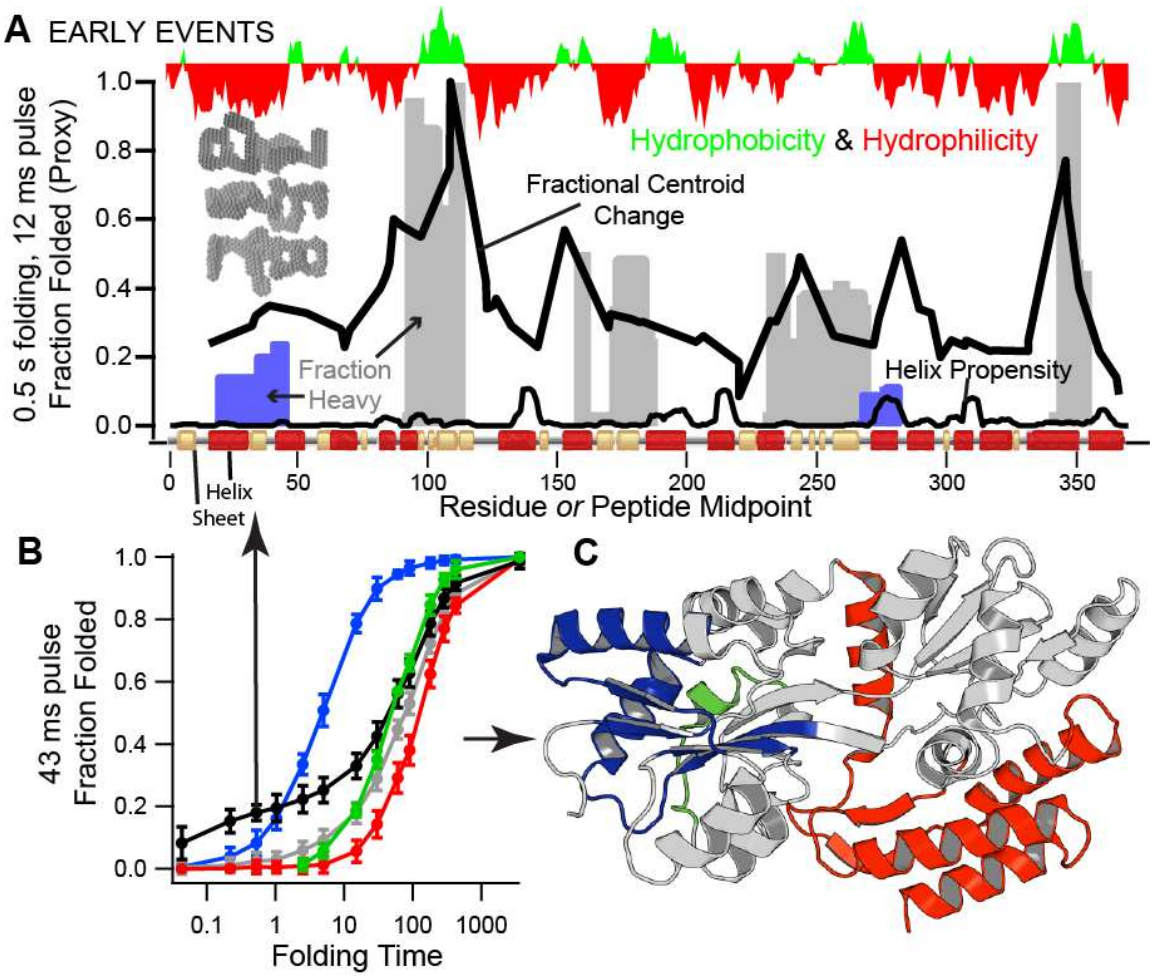

Figure 6.1: Summary of the early and late folding events in MBP. **(A)** Almost identical to Figure 5.7E, expanded for ease of inspection and with the envelopes generated by SAXS measurements shown in the inset. This is the result of pulsing for only 12 ms at 0.5 seconds of folding. The heavy population fraction is shown in bars for those peptides with two populations. The fractional mass increment interpolated between unfolded and fully deuterated controls is shown for every peptide in the thick black line. Secondary structures in the native state and sequence helical propensities and hydrophobicity are shown for reference. More information can be found in Chapter 5. **(B)** Groups of peptides shown individually in Figure 5.9 have here been averaged to give a better representation of each group. **(C)** The same as Figure 5.9D, reproduced for convenience.

It appears that the aspecific collapse, summarized in Figure 6.1A represents a plethora of chain configurations, with a nearly native radius of gyration, which leads to the optical burst signal. This state represents the denatured state ensemble (DSE) under permissible folding conditions. About the presence of structure in DSEs, we do observe low levels of protection throughout the molecule as shown by the centroid changes and find two particular regions with exceptionally high levels of early protection. We find no indication that these two elements are native-like; however, they are the primary candidates for the gain in CD_{222} signal observed

during the burst and may represent cooperative folding. This potential cooperativity is explored and shown below in Figure 6.3.

If there were multiple pathways to the native state, one expects such behavior to materialize from the early heterogeneity observed in MBP; however from the initial multi-pathway-like and heterogeneous collapse behavior that we measure in MBP, an obligatory intermediate and folding pathway emerges (described in Chapter 5, and recapitulated in Figure 6.1). We also demonstrate definitively that no fast folding tracks are operative in MBP.

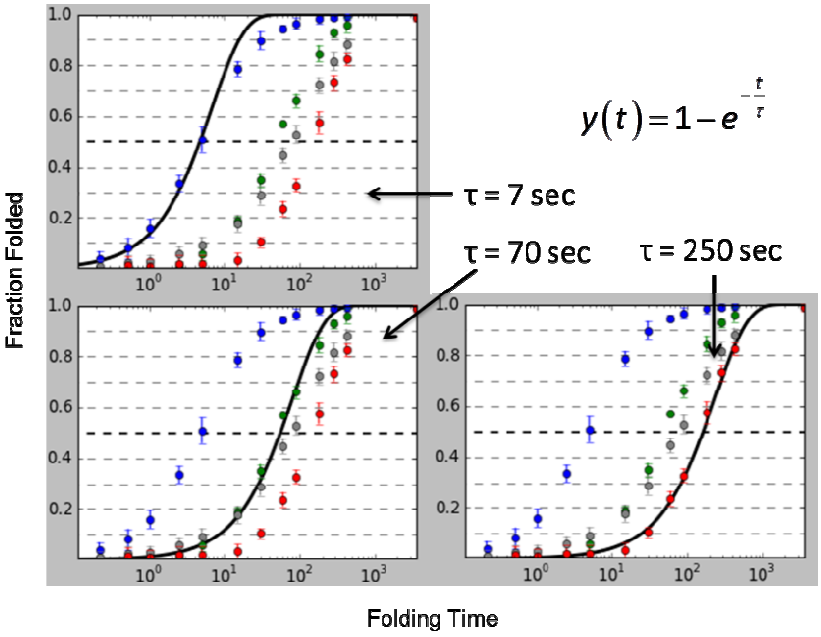

Figure 6.2: Testing the observed data for single exponential behavior. If we observe simple exponential behavior, it is interesting and relevant as discussed in the text below.

The slowest folding elements in Figure 6.1B (red) show no evidence of protection until after most molecules have formed the 7 s intermediate (blue). We do not necessarily expect the time dependence of non-two-state folding to be well characterized by a simple exponential; however, the slowest folding elements unexpectedly fit a single exponential almost perfectly (see Figure 6.2). This suggests that the final event is rate-limiting for all molecules and could be taken alone as evidence that folding is pathway directed. It is bona fide proof that there are no fast pathways to the native state. Not only do we see an obligatory transition at 7 seconds but we also measure a single step transition to native. The slower folding behaviors are bookended

by features expected for pathway-directed folding. The energetic descent to the native conformation in MBP fits the predetermined pathway model.

The emergence of a predetermined pathway from stochastic diffusive chain kinematics is not surprising when one remembers two well-established physical principles. The first being the principle of cooperativity and the second being an interaction principle observed throughout biology known as sequential stabilization. Cooperative structural units have a tendency to explore particular regions of conformational space more frequently due to the energetic benefits of those configurations relative to the others that are available. We call these basic cooperative units foldons. All foldons are randomly appearing and disappearing during the early stages of folding, albeit some foldons appear at a higher frequency than others. When the most stable foldon collides with the second most stable foldon, they mutually stabilize one another through the free energy of association. Together, the foldons persist for much longer than either taken in isolation. This increases the probability that the first two foldons are folded properly when the next foldon transiently forms and so on. The third and later foldons only experience the additional stability from binding to the conformational scaffold containing all earlier foldons; else pathway branching may occur. A *deterministic* sequence of *macroscopic* events emerges as less stable foldons are sequentially stabilized by their addition to a growing conformational scaffold; their addition further stabilizes the scaffold.

The founding father of folding research, Christian B. Anfinsen, introduced everyone to the longstanding view that the kinetic accessibility of the native conformation is determined by the sequence of amino acids alone. This specificity of the primary sequence in proteins serves as the key difference between the physics of protein folding and that of simple polymers and spin-glasses[26]. This can be shown simply by the fact that random sequences of amino acids do not fold into specific structures.

Experimental evidence has shown that at least one specifically structured native-like on pathway intermediate exists for many proteins (110-116, 119-124, 257, 258). Organized folding sequences have also been demonstrated experimentally (5, 9, 48, 116, 126) and more recently

[26] Models of multiple independent and unrelated pathways were born out of simple polymer physics & theory.

in simulation for small proteins (67, 260). *Definitive* evidence for innumerable independent folding pathways, as implied by the IUP model, have not been observed in experiments that have the capacity to directly measure this type of behavior such as the one presented here.

6.2 Future Directions

6.2.1 Low pH Molten Globule & The Kinetic Polyglobule

Prajapati et al. conducted a study of molten globule states in periplasmic binding proteins, one of which was MBP (261). They found that at low pH, MBP adopts a molten globule (MG) conformation and binds the hydrophobic dye, ANS, just as we observe early during the folding of MBP (shown in Figure 5.2, page 84). In a separate study, it was shown that unfolding of the MG in MBP was characterized by significant ΔC_p and ΔH (162).

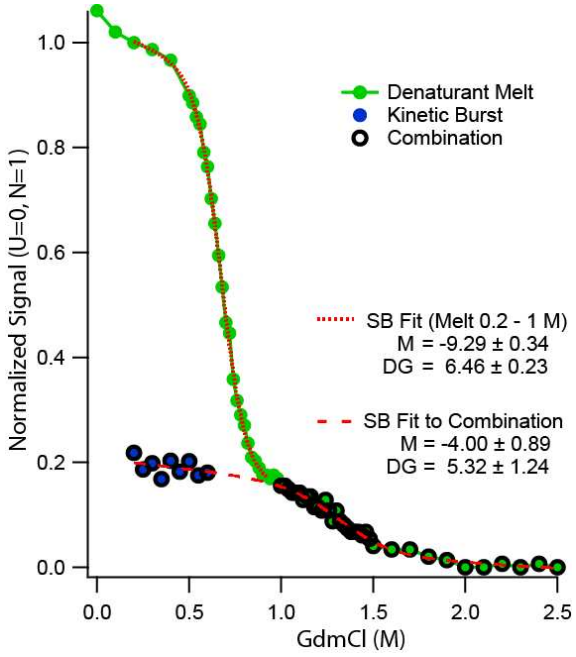

Figure 6.3: Re-analyzing the denaturant melt of MBP as measured by optical fluorescence. The data shown here is also presented in Figure 5.3A page 86. The Santoro-Bolen fit to the entire GdmCl range along with the residuals are shown in Figure 5.1, page 83. This is merely to play with the melt data and ponder whether there is any significance in the energetic values estimated by this approach. Certainly by summing the free energies (DG, sum = 11.78 kcal/mol), one arrives closer to the stability estimate provided by HX NMR measurements in Chapter 3 of ~ 14-15 kcal/mol for the most stable H-bonds.

Could it be possible that the post-transition baseline we observe in a denaturant melt of MBP (shown in Figure 5.3A) is actually a second transition that describes unfolding of an MG?

Perhaps our kinetic polyglobule and the equilibrium MG described here at low pH are similar. To first test this hypothesis, using the melt data for MBP at pH 9.0, normalized to kinetic endpoints (0.2 M GdmCl → N, 2 M GdmCl → U) so that we could fit sub-regions of the melt separately, we see in Figure 6.3 two transitions. The transition connected to our kinetic burst data is similar to that observed in these other studies of the low pH MBP molten globule (162, 261). Additionally, by summing the free energies, the optical melt allows one to more closely approach the roughly 14-15 kcal/mol stability measured by HX NMR in Chapter 3. Playing around with data in this manner is recreational, and not meant to suggest a deep physical relationship between our polyglobule and the low pH MG; however, it does tickle the mind a little.

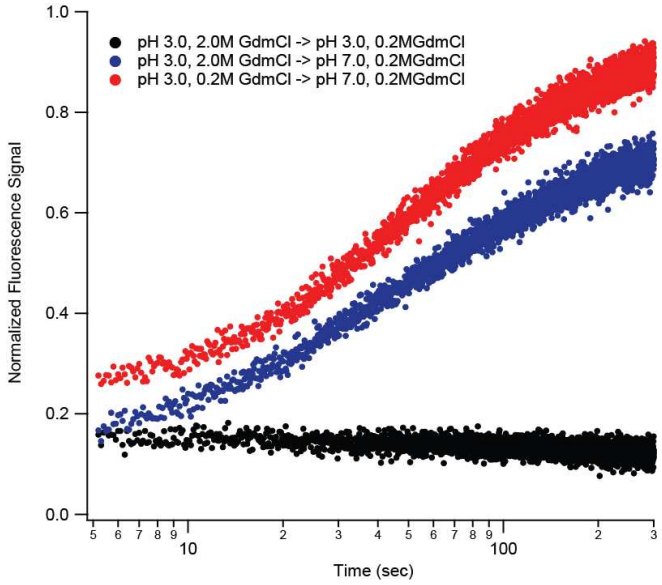

Figure 6.4: Optical similarities between the kinetic burst and low pH MG. The normalized signal acquired by starting from an unfolded state equilibrated at pH 7.0 (not shown) is nearly coincident with the blue data, at pH 9.0 the later signal fills in slightly quicker but the early behavior also resembles the blue data (see Figure 5.2, page 84 for comparison).

By optical measurements shown in Figure 6.4, the difference in the amplitude of the kinetic burst phase and the difference between the chemically denatured state and the low pH molten globule are similar; perhaps both species have a similar change in solvent accessible surface area. One assumes the polyglobule state gives rise to the burst-phase we observe in optical kinetics, but the transient nature of this state presents some difficulties that may be overcome if conditions were found that promoted this species at equilibrium. Perhaps the low pH condition favors a polyglobular state similar to the one described in this work.

In data not shown in this dissertation, we performed a native hydrogen exchange experiment where we passingly attempted to define the structure of the low pH molten globule. The behavior did resemble the same sort of low-level protection observed in our pulse modulation experiment (see Figure 5.8, page 94), but we did not see specific areas of increased protection in the low pH MG. Perhaps this is because the low pH MG exchanges largely by way of local fluctuations (262), or experiences electrostatic slowing of chemical exchange rates from the collapsed state (263) and therefore gives a spread of unremarkable slowing factors.

A denaturant series, similar to what has been done in the past for cytochrome C (9) might shed light onto the structure of the low pH MG and facilitate comparisons with what we observe in our kinetic polyglobule. Additionally, pulsing the low pH MG for a few milliseconds, similar to our pulse time modulation strategy may reveal the same regions with high protection observed in the kinetic polyglobule to be present also in the low pH MG.

6.2.2 Pulse power modulation

In our current work, we modulated pulse time at 0.5 seconds of folding and were able to show that while many peptides did not exhibit multiple populations, they were more structured than was implied by focusing only on the time dependence of their heavy population fractions; additionally, the fraction of heavy molecules in the biphasic peptides (black traces in Figure 6.5B) increased as we reduced the pulse strength. Pulse power modulation shows that a few of these peptides were actually 100% protected after 0.5 seconds of folding but their structural opening rates were such that it appeared only 20-30% were protected in a 43 ms pulse. Those peptides that formed the 7 s intermediate did not respond to the change in pulse strength because their protecting structure was sufficiently stable.

Many other peptides deviated from single exponential behavior (such as the grey traces in Figure 1B), albeit not as strongly as the early biphasic peptides. Perhaps by modulating the strength of the pulse, we could uncover similar findings in these later events shown in Figure 6.1B. Changes in stability during folding, such as those seen in a 12 ms pulse (Figure 6.1A) compared to the 43 ms pulse (Figure 6.1B) are suggestive that opening rates decrease as more and more structure adds onto the growing native-like scaffold. Pulse-power modulation at all

folding times may uncover other interesting aspects of the folding pathway that we did not observe in our work.

6.2.3 Double Jump Pulse Labeling HX MS

By eliminating the effects of proline mis-isomerization, some of the slower behavior in MBP might be resolved more clearly. Upon inspection of Figure 6.1B, one notices the formation of the blue intermediate; but, beyond this, separable events are less clearly defined. It is possible that mis-isomerized proline residues lead to ambiguity in the sequence of later events; perhaps the grey and red groups in Figure 6.1B are actually three separate groups; the degree of overlap in the current data prevents us from making this conclusion. We know that proline mis-isomerization does slow down the folding process somewhat (compare single and double jump traces in Figure 5.2, page 84). Figure 5.4 on page 89 shows that all four syringes on our stop flow are used in the single jump experiment. To perform a double jump HX pulse labeling study, we would need an additional syringe and delay line. We would be curious to see whether a double jump HX pulse labeling experiment would resolve additional behaviors in the slower folding regions of MBP.

6.3 Moving Forward

Perhaps the most interesting questions are those that we are now inclined to ask about other large proteins. These questions could be addressed directly using the advanced hydrogen exchange methods outlined in this work. Do other large proteins possess a collapsed unfolded state in permissible refolding conditions? Do large molecules frequently collapse aspecifically? Are all collapsed states sensitive to reversible aggregation? Do most large proteins fold by way of a predetermined pathway? We do not have answers to these questions because there are no other studies of large protein folding with the degree of structural resolution presented here. The technological advancements presented in this dissertation provide a road map for future studies. We hope the technologies developed for this work and described herein will enable others to overcome the various issues faced in large proteins and contribute, as we have, to further our structural understanding of large protein folding processes.

Appendix A - Calculation of Expected D-Recovery

Fractional recovery is represented as

$$R_{obs} = \frac{Mass_{FD} - Mass_{H}}{s}. \qquad \text{Eq. 7.1}$$

where *s* represents the number of exchangeable amides on the peptide, which is the number of residues in the peptide minus the first two and minus the number of prolines beyond the first two residues. The subscripts *FD* and *H* refer to the protein sample being fully deuterated or protonated, respectively. Masses were experimentally determined using the appropriate centroids in the standard way: *Mass* = *z***centroid*, where centroid is in m/z units and *z* represents the charge state.

To compute the expected recovery (or fractional deuteration) of a given peptide in the reported data (or any multi stage HX experiment), the following tuples are defined for each different chemical environment (n= number of different environments):

$$\begin{aligned} \varphi &= \{f_{D1}, f_{D2}, \ldots, f_{Dn}\} \\ \kappa &= \{\{k_1, k_2, \ldots, k_s\}_1, \{k_1, k_2, \ldots, k_s\}_2, \ldots, \{k_1, k_2, \ldots, k_s\}_n\}. \\ \tau &= \{t_1, t_2, \ldots, t_n\} \end{aligned} \qquad \text{Eq. 7.2}$$

Phi contains the fractional solvent deuteration level for each condition. Kappa contains a complete set of chemical exchange rates (132) with 's' entries for each exchange site for each pH/temperature condition. Tau holds time intervals for each pH/temperature condition such as digest/wash time, elution time, *etc.* for each different condition encountered in the experiment. Fractional deuteration at some time, $F_d(t)$, can be represented as follows,

$$F_d(t)=\left\{\begin{array}{lll} \frac{1}{s}\sum_{j=1}^{s}\varphi_1-[\varphi_1-f_{initial}]e^{-k_{1j}t} & if & t\le\tau_1 \\ \frac{1}{s}\sum_{j=1}^{s}\varphi_2-[\varphi_2-F_d(\tau_1)]e^{-k_{ij}(t-\tau_1)} & if & \tau_1<t\le\tau_1+\tau_2 \\ \vdots & \vdots & \vdots \\ \frac{1}{s}\sum_{j=1}^{s}\varphi_n-[\varphi_n-F_d(\sum_{p=1}^{n-1}\tau_p)]e^{-k_{nj}(t-\sum_{p=1}^{n-1}\tau_p)} & if & \sum_{p=1}^{n-1}\tau_p<t\le\sum_{p=1}^{n}\tau_p \end{array}\right\}, \qquad \text{Eq. 7.3}$$

where $f_{initial}$ is the initial fractional deuteration of the peptide present before the experiment begins. The form of Eq. 7.3 describes all HX expectation curves in either direction (exchange-in or -out) so long as the sign conventions are followed, but does not consider the back reaction, $ND+H_{solvent}$ $NH+D_{solvent}$, because free deuteron (or proton if in the reverse direction) levels in solution are negligible during the majority of preparation time.

Appendix B - LC Gradient Shaping

The effort to reach single amino-acid resolution requires high precision mass measurements on many overlapping peptides with minimal back-exchange. To accomplish this we want to efficiently separate peptides using reversed-phase chromatography in the shortest possible time; however, reduction of chromatography time leads to chromatographic crowding which can significantly reduce the number of peptides resolved. The effect of crowding was evidenced by the 40% reduction in unique peptides identified between a 10- and 5-minute gradient (section 4.3.2, page 65). Chromatographic shaping is based on the idea that constant peptide elution density per unit time will be the most efficient separation in terms of minimization of gradient length while maximizing the number of unique peptide overlaps identified.

B.1 Linear and Shaped LC Gradients

Typically, in ESI-MS experiments, peptides are eluted from a reverse phase column using a linear elution gradient which can are described in the following way:

$$\rho(t) = f(ACN(t)), \quad \text{Eq. 8.1}$$

$$AcCN(t) = m_{gradient}t + AcCN_{t=0}. \quad \text{Eq. 8.2}$$

The term $\rho(t)$ represents the population density of unique peptides eluting at a particular time. Using a linear ACN gradient, eluate peptide density may vary significantly over the gradient. Linear gradients may be inefficient depending on the composition of peptides bound to the LC column, there may be H_2O:AcCN compositions where many peptides elute simultaneously interspersed between compositions that yield few peptides. For data dependent tandem MS experiments, such as those used to create the peptide pool (described in Chapter 4, page 51), when many unique peptides are eluting simultaneously, it may be impossible for the instrument to select and fragment each peptide and this will result in fewer identifications. This inefficiency is shown in Figure B.1A for a 10-minute linear ACN gradient (red trace) which spans the same ACN range as our 10-minute shaped condition in the main text. The linear gradient gives 177 unique peptides (blue circles).

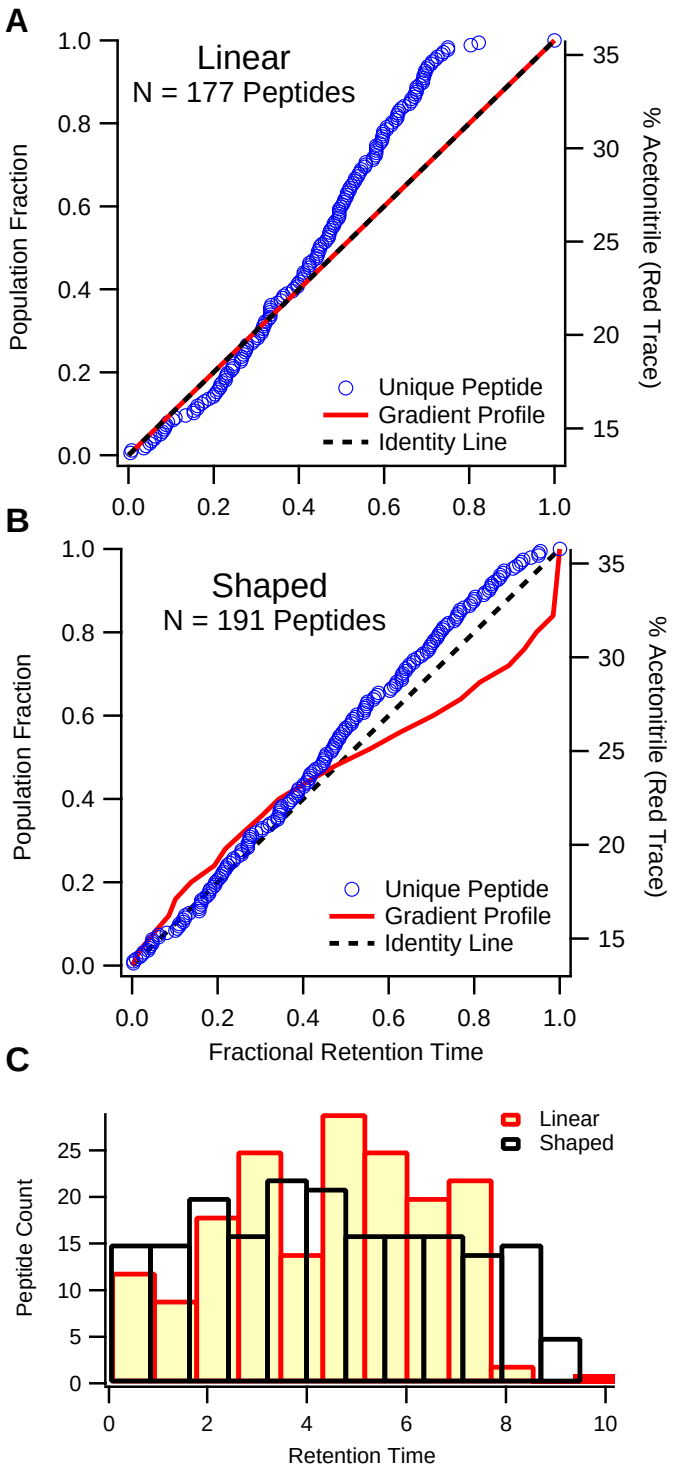

Figure B.1: Chromatographic optimization by shaped gradients.

Inverting Eq. 8.1 will provide a constant elution density per unit time and increase chromatographic efficiency:

$$f^{-1}(\rho(t)) = ACN(t)\,. \qquad \text{Eq. 8.3}$$

This inversion is shown in Figure B.1B where we resolved 191 unique peptides with a 10 minute shaped gradient. Its effect on producing a constant elution density per unit time is contrasted with the linear gradient in Figure B.1C.

B.2 Implementation and Discussion of Gradient Shaping

Our HPLC pump was designed for traditional step-gradients therefore we used a discrete implementation of Eq. 8.3. This involved first collecting a slow reference linear gradient from 0-50% AcCN. Retention times for identified peptides were used to generate 30 evenly spaced time bins where each bin was assigned the value of the number of peptides eluting over that time range (our pump software accepted 30 steps). We then matched retention time with %ACN during elution which allowed us compute the necessary $\Delta\%ACN$ for each bin such that the bins contain equal numbers of peptides.

Chromatographic shaping improved our 10-minute gradient from 177 (linear) to 191 (shaped) uniquely identified peptides using a mass resolving power of 100,000. High resolution instruments with resolving powers, $\frac{mass}{\Delta mass_{50\%\ \text{intensity}}}$, at or above 100,000 somewhat mitigate the benefit of chromatographic shaping as peak capacity increases with resolving power. The difference in number of identified peptides between our 10 minute shaped and linear gradients are somewhat understated due to the high resolving power available with our instrument. However, many laboratories employ QTOF and lower resolution instruments for the acquisition of hydrogen exchange data. We imagine chromatographic shaping will improve the number of peptides identified significantly in those cases.

While we only increased our peptide resolution by 14 unique identifications, Figure B.1C demonstrates that shaping achieves our goal of equal peptide density per unit time during chromatography. As HX MS experiments move to larger and larger protein systems, researchers will not have the flexibility of running longer linear gradients to achieve adequate separation.

Proteins much larger than MBP with potentially thousands of unique peptides should show a larger effect. Though not always necessary, one can easily envisage situations that would improve tremendously by gradient shaping.

Appendix C - Simulating MS Data

C.1 Nominalization of the Mass Axis

Nuclide	Atomic Mass	Nominal Offset	ρ (probability)
^{12}C	12.0000	+ 0	0.9893
^{13}C	13.0034	+1	0.0107
^{14}N	14.0031	+0	0.9964
^{15}N	15.0001	+1	0.0036
^{16}O	15.9949	+0	0.9979
^{18}O	17.9992	+2	0.0021
^{32}S	31.9721	+0	0.9495
^{33}S	32.9715	+1	0.0076
^{34}S	33.9679	+2	0.0429

Table C.1: Isotopic abundances and nominal offsets for atoms relevant to peptide MS.

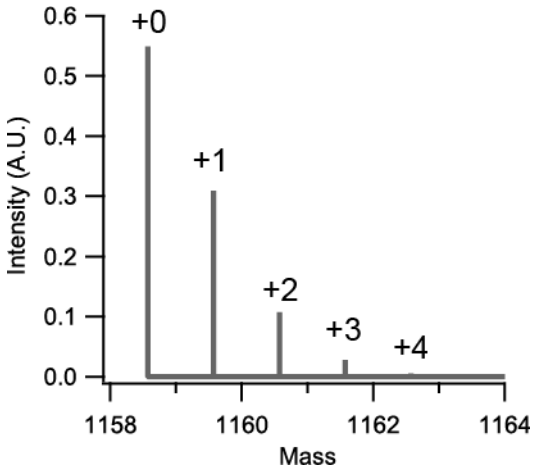

Figure C.1: Peak nominalization for a peptide spectrum. The monoisotopic mass is defined as +0. Each additional peak is defined by an integer offset from the monoisotopic mass as indicated above each peak.

Table C.1 contains the relevant information for stable isotopes encountered in peptide mass spectrometry. The monoisotopic mass of the molecule is the molecular weight computed by the masses of the most abundant stable isotope for each atom (red in Table C.1, +0 in Figure C.1). An example natural abundance distribution computed by HDpop for the MBP peptide SAGINAASPNKE is shown in Figure C.1 and the monoisotopic peak is the lightest peak in the

distribution. Isotopes lighter than the most abundant for atoms observed in peptide mass spectrometry (C, O, N, & S) have negligible natural abundances and are always expected to be far below the limit of detection. Therefore, the monoisotopic mass or formula mass of the molecule composed of the most abundant isotopes for each atom type (red in Table C.1) is always the lightest peak of the peptide mass distribution. To nominalize the mass axis, the monoisotopic peak is indexed '+0' and the absolute mass information for each additional peak may be replaced by an appropriate integer offset from the monoisotopic peak. These nominal offsets, one for each isotopologue[27] group are written above each peak in Figure C.1. All isotopologues of the molecule containing only one ^{13}C or only one ^{15}N contribute to the measured intensity of the +1 nominal mass as shown in Figure C.2 below.

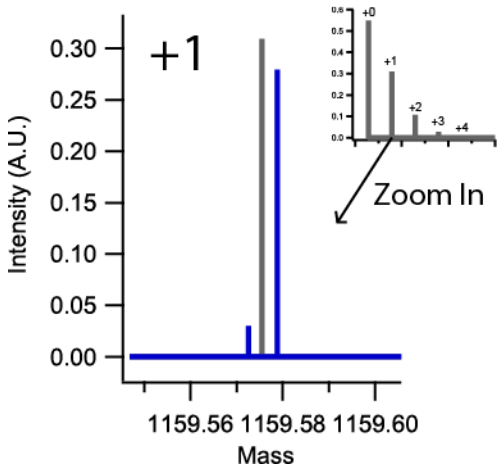

Figure C.2 Unresolved isotopologues (blue) for the +1 peak in Figure C.1. The area of the gray peak is equivalent to the sum of the areas for each isotopologue.

Each peak or isotopologue of the molecule in the mass spectrum results from some combination of the isotopes in Table C.2; however, many isotopologues will overlap and are indistinguishable. Currently, the maximum resolution[28] used for HX experiments in the literature is 100,000 and the typical resolution is 20,000. To resolve the two isotopologues (blue) shown

[27] An isotopologue, as defined in the IUPAC Compendium of Chemical Terminology, is a molecular entity that differs only in isotopic composition (number of isotopic substitutions). For example, the carbon isotopologues for ethane are $^{12}C_2\,^1H_6$, $^{12}C^{13}C\,^1H_6$, $^{13}C_2\,^1H_6$, only stable nuclides are considered.

[28] MS resolution is defined by $mass / \Delta mass_{50\%Intensity}$

for the +1 peak in Figure C.2, one would need greater than 150,000 resolution. At 100K and lower resolutions, isotopologues within a single nominal group blend and are therefore represented by a single peak in the computed natural abundance distribution mentioned in Chapter 4 on page 73(gray, Figure C.2). These single peaks in Figure C.1 are the sum of the probabilities for observing each isotopologue in a given nominal group as shown in Figure C.2.

C.2 HDpop Implementation

Computing the expected mass spectrum of a peptide is equivalent to determining the isotopic abundances of all isotopologues for a given chemical formula. This task can become particularly laborious as the number of atoms increase in the molecule. HDpop has two methods for simulating mass distributions, they are described below.

The method most typically employed is valid in the case that similar mass isotopologues are unresolved, in this case, only the nominal natural abundance distribution needs to be computed. This is the only method used for fitting HX data simply because our peptide identification program, ExMS, does not attempt to resolve isotopologues. ExMS determines the intensity of the peaks at each integer offset from the monoisotopic peak in experimental data and sends those intensities to HDpop for analysis. Therefore, HDpop only needs to determine one peak intensity per nominal offset from the monoisotopic peak for the typical HX MS dataset. The convolution method employed in HDpop, page 73, involves application of the discrete Fourier transform and is similar to the work of Rockwood et al (264). This approach is much less demanding from a computational perspective than the second exact method, which may be employed to determine the exact isotopologue distribution.

This exact polynomial method is not employed for the work presented in Chapter 4, page 73. However, a description is included as there are some interesting implications regarding site resolution of D occupancies that were discovered in the process of development. After developing this method, I found an early paper recommending the use of Diophantine equations (265) for this purpose – this approach is very similar to the one described here, although truncations are made reducing the accuracy of that other method. The exact method is discussed following the DFT method.

C.3 Nominal Natural Abundance Distribution Using the DFT

To compute the natural abundance distribution, HDpop first determines individual distributions for each atom type and then convolves them together using the discrete Fourier transform, using the python *SciPy* function `signal.convolve` (227). For each atom, the natural isotopic distribution can be modeled using a binomial expansion[29],

$$I(x:k,\rho)=\frac{k!}{x!(k-x)!}\times\rho^{x}\times(1-\rho)^{k-x}, \qquad \text{Eq. 8.4}$$

$x=0$ corresponds to the probability of observing the molecule containing only the most abundant isotope of the atom, $x=1$ gives the probability of observing only one heavy isotope of the particular atom and so on. Fixed variables k and ρ represent the number of atoms of a given type in the chemical formula and the probability of observing the heavy isotope, respectively.

A list is constructed using equation Eq. 8.4 for each non-zero nominal mass in table 1, $\mathrm{P}_{atom}=(0:k_{atom},\rho_{atom}),(1:k_{atom},\rho_{atom}),\ldots,(k:k_{atom},\rho_{atom})$. Conveniently, the indices of all atom lists except ^{18}O and ^{34}S directly correspond to the nominal mass shifts associated with each probability in the list. However, as is indicated by the +2 nominal masses for ^{18}O or ^{34}S, these lists must be modified for the previous statement to hold. New lists are created for these atoms by inserting a zero between each entry such that the new list equals twice the cardinality of the original minus one.

The ability to ignore differences between isotopologues with similar atomic masses means that the indices of each P_{atom} list correspond directly to the intensity contribution of each particular atom to the intensity of the peaks in the full spectrum given by the nominal offsets (+0, +1, …) introduced in earlier. Thus, the grids for each atom are uniform with respect to each other and with respect to each consecutive entry; the data does not require resampling in order to apply the DFT convolution. All lists are zero padded to equivalent lengths before convolution.

[29] It is worth noting that sulfur is not properly modeled using the binomial, there are three possibilities and the distribution is trinomial. In practice, modeling the sulfur distribution using binomials does not introduce significant error because of the low number of sulfur atoms in the average peptide, typically zero. Using the binomial for sulfur (there are two binomial sulfur distributions) was chosen for method 1.

The Fourier transform of a convolution is equivalent to the point-wise products of Fourier transforms. Restated, convolution in the time domain is equivalent to multiplication in the frequency domain. Point-wise multiplication involves far fewer computations than algebraic convolution making this approach attractive. Let $*$ denote convolution and $\square$ point-wise multiplication operators. Let F denote the discrete Fourier transform operator, and F^{-1} its inverse. Convolution may then be represented as:

$$\mathrm{A} * \mathrm{B} = \mathrm{F}^{-1}\left\{\mathrm{F}\left\{\mathrm{A}\right\} \square \mathrm{F}\left\{\mathrm{B}\right\}\right\} \qquad \text{Eq. 8.5}$$

Where A and B are arbitrary lists of equal cardinality on a uniform grid. The natural abundance distribution is $\mathrm{P}_{natural} = \mathrm{P}_{^{13}C} * \mathrm{P}_{^{15}N} * \mathrm{P}_{^{18}O} * \mathrm{P}_{^{33}S} * \mathrm{P}_{^{34}S}$. Computationally, this is done in an iterative fashion, first we form a kernel: $\mathrm{P}_{\text{kernal}} = \texttt{scipy.signal.convolve}(\mathrm{P}_{atom1}, \mathrm{P}_{atom2})$, then loop through the remaining atoms, $\mathrm{P}_{\text{kernal}} = \texttt{scipy.signal.convolve}(\mathrm{P}_{\text{kernal}}, \mathrm{P}_{atom})$. When no atoms remain, $\mathrm{P}_{\text{kernal}} = \mathrm{P}_{natural}$.

C.4 Exact Natural Abundance Distribution Using a Polynomial Method

HDpop also has the option to generate the expected natural abundance distribution at infinite resolution using an exact method. This method is computationally unbearable for large molecules. For each atom, a relation must be created which allows regeneration of mass information following polynomial convolution. For example, to cast the carbon distribution appropriately (under the terms set in Table C.1), we define the following relationships:

$$\begin{aligned} \Delta^{x}_{^{12}C} &= x \cdot mass_{^{12}C} \\ \Delta^{x}_{^{13}C} &= x \cdot mass_{^{13}C} \end{aligned} \qquad \text{Eq. 8.6}$$

We then cast the distribution for carbon as a binomial using the natural abundances of carbon isotopes as coefficients and the atom counts as exponents:

$$\left(\rho_{^{12}C}\Delta^{1}_{^{12}C} + \rho_{^{13}C}\Delta^{1}_{^{13}C}\right)^{k_{carbon}}, \qquad \text{Eq. 8.7}$$

k_{carbon} is the number of carbon atoms in the molecule and $\rho_{^{12}C}$ is the probability of observing a single ^{12}C atom; this is reflected the exponent which is set to one. For an atom requiring a trinomial, such as sulfur, the distribution is written the same way, $\left(\rho_{^{32}S}\Delta^{1}_{^{32}S}+\rho_{^{33}S}\Delta^{1}_{^{33}S}+\rho_{^{34}S}\Delta^{1}_{^{34}S}\right)^{k_{sulfur}}$.

The natural abundance distribution is obtained by multiplying all individual atom n-nomials (Eq. 8.6) for a molecule together using standard algebraic distribution. The result is a long string of terms and each represents a particular isotopologue in the mass spectrometer. Following distribution, each isotopologue will have scalar probabilities, $\rho^{exponent}_{atom}$, which refer to intensity. Numerically evaluating scalar probabilities $\rho^{exponent}_{atom}$ to the power indicated by the exponent and then combining all by multiplication gives the intensity. Each term also has a number of $\Delta^{exponent}_{atom}$ parts that behave differently than the scalar probabilities. These exponents after distribution are used to convert each part to a mass (as done for carbon in Eq. 8.6). The following illustrates a few operations to show how the mass terms work:

$$\Delta^{A}_{atom1}\Delta^{B}_{atom2}\Delta^{C}_{atom3}=\left(A\cdot mass_{atom1}\right)+\left(B\cdot mass_{atom2}\right)+\left(C\cdot mass_{atom3}\right)$$

$$\Delta^{A}_{atom1}\Delta^{B}_{atom1}=\Delta^{A+B}_{atom1}$$

$$\Delta^{A}_{atom1}\Delta^{A}_{atom1}=2\Delta^{A}_{atom1}=\Delta^{2A}_{atom1}$$

Just as in polynomial distribution, think of the $\rho^{exponent}_{atom}$ terms as coefficients and the $\Delta^{exponent}_{atom}$ as variables that identify the mass position of the intensity defined by $\rho^{exponent}_{atom}$. Any coefficient with identical variables can be combined by addition or subtraction as appropriate, ie: $\rho^{A}_{atom}\Delta^{B}_{atom1}\Delta^{C}_{atom2}+\rho^{A}_{atom}\Delta^{B}_{atom1}\Delta^{C}_{atom2}=\left[\left(\rho_{atom}\right)^{A}+\left(\rho_{atom}\right)^{A}\right]\Delta^{B}_{atom1}\Delta^{C}_{atom2}$, referring to a peak with intensity $\left(\rho_{atom}\right)^{A}+\left(\rho_{atom}\right)^{A}$ and mass $\left(B\cdot mass_{atom1}\right)+\left(C\cdot mass_{atom2}\right)$. However, unlike the type of variable typically encountered, the following, $\rho^{A}_{atom}\Delta^{B}_{atom1}\Delta^{C}_{atom2}+\rho^{A}_{atom}\Delta^{2B}_{atom1}\Delta^{C}_{atom2}$, actually refers to two different peaks, $\Delta^{B}_{atom1}\Delta^{C}_{atom2}$ and $\Delta^{2B}_{atom1}\Delta^{C}_{atom2}$, these $\rho^{exponent}_{atom}$ parts are not combined as in the previous case.

The method is computationally bearable for the size of molecules typically encountered using the fragmentation-separation HX MS strategy presented in section 4.2, page 50. It is included because the form implies that as instrument resolution increases in the future, the information to determine the exact deuterium distribution (as opposed to the average deuteration of each peptide) could be determined directly from analyzing the intensities of the individual isotopologues. One would be able to write an equation for each isotopologue. Together, these equations form an over determined non-linear system with respect to the unknown site D occupancies. By nonlinear methods, one would be able to solve for site D occupancies directly and with high confidence. Overlapping peptides would be required to match the site occupancies with their corresponding residues in the protein.

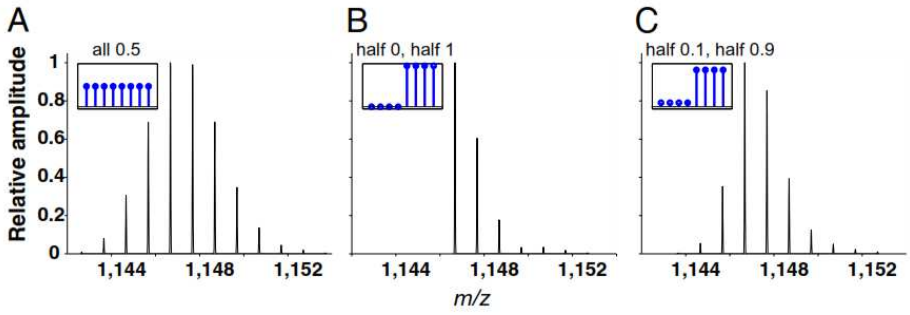

Figure C.3: Another example of non-uniform deuterium distributions (*reprinted from reference (6)*). The centroid value of all three spectra is the same. In each case, the peptide has an average of 5 deuterium. **(A)** Uniform deuteration, all sites have 0.5 deuterium. **(B,C)** Non-uniform deuteration profiles, see the insets.

We have recently shown how site deuterium occupancies may be extracted using unresolved isotopologues so long as a sufficient number of overlapping peptides are available (6). By inspection of the mass spectra in Figure C.3 one notices that the distribution observed reflects the degree of non-uniformity in the deuteration pattern. By fitting each site independently, a non-uniform site D-occupancy distribution may be determined for each peptide. On a per peptide basis, one would need to resolve the individual isotopologues to have an over-determined system; however, by virtue of many overlapping peptides reporting independent measurements on the same site, globally using all overlapping peptides circumvents the issue.

Overlapping information will be required if the site D occupancies that are theoretically determined by the intensities of the individual isotopologues are to be matched to any

particular residue in the protein. By resolving individual isotopologues, the number of peaks observed in any one spectrum dramatically increases. As the distribution of deuterium occupancies influences the all peak intensities in the spectrum, using the polynomial method, we may write equations for each peak. Should instrument resolution achieve resolving powers in excess of one million, the polynomial approach will be able to achieve site resolution in a state sensitive manner. Using the global DFT based analysis at lower resolution is sufficient when the spectra contain only a single population per peptide; however, for 2+ populations, the system will remain under-determined regardless of the number of overlapping peptides and therefore the method unreliable for more than a single population.

BIBLIOGRAPHY

1. Gledhill JM, Jr., Walters BT, & Wand AJ (2009) AMORE-HX: a multidimensional optimization of radial enhanced NMR-sampled hydrogen exchange. *J Biomol NMR* 45(1-2):233-239.
2. Walters BT, Ricciuti A, Mayne L, & Englander SW (2012) Minimizing Back Exchange in the Hydrogen Exchange-Mass Spectrometry Experiment. *J. Am. Soc. Mass Spectrom.* 23(12):2132-2139.
3. Walters BT, Mayne L, Hinshaw JR, Sosnick TR, & Englander SW (2013) Folding of a large protein at high structural resolution. *Proc Natl Acad Sci U S A* x(x):x-xx.
4. Mayne L*, et al.* (2011) Many Overlapping Peptides for Protein Hydrogen Exchange Experiments by the Fragment Separation-Mass Spectrometry Method. *J. Am. Soc. Mass Spectrom.* 22(11):1898-1905.
5. Hu W*, et al.* (2013) Stepwise protein folding at near amino acid resolution by hydrogen exchange and mass spectrometry. *Proc. Natl. Acad. Sci. USA* 110(19):7684-7689.
6. Kan ZY, Walters BT, Mayne L, & Englander SW (2013) Protein hydrogen exchange at residue resolution by proteolytic fragmentation mass spectrometry analysis. *Proc Natl Acad Sci U S A*.
7. Levinthal C (1968) Are there pathways for protein folding? *Journal de Chimie Physique et de Physico-Chimie Biologique* 65:44-45.
8. Englander SW (1993) In pursuit of protein folding. *Science* 262(5135):848-849.
9. Maity H, Maity M, Krishna MM, Mayne L, & Englander SW (2005) Protein folding: the stepwise assembly of foldon units. *Proc Natl Acad Sci U S A* 102(13):4741-4746.
10. Krishna MM, Lin Y, & Englander SW (2004) Protein misfolding: optional barriers, misfolded intermediates, and pathway heterogeneity. *J. Mol. Biol.* 343(4):1095-1109.
11. Englander SW, Mayne L, & Krishna MM (2007) Protein folding and misfolding: mechanism and principles. *Q. Rev. Biophys.* 40(4):287-326.
12. Hills RD, Jr. & Brooks CL, 3rd (2008) Subdomain competition, cooperativity, and topological frustration in the folding of CheY. *J. Mol. Biol.* 382(2):485-495.
13. Kathuria SV, Day IJ, Wallace LA, & Matthews CR (2008) Kinetic traps in the folding of beta alpha-repeat proteins: CheY initially misfolds before accessing the native conformation. *J. Mol. Biol.* 382(2):467-484.
14. Schulenburg C*, et al.* (2009) The folding pathway of onconase is directed by a conserved intermediate. *Biochemistry* 48(35):8449-8457.
15. Agócs G, Szabó Bence T, Köhler G, & Osváth S (2012) Comparing the Folding and Misfolding Energy Landscapes of Phosphoglycerate Kinase. *Biophys. J.* 102(12):2828-2834.
16. Qiu L, Zachariah C, & Hagen SJ (2003) Fast chain contraction during protein folding: "foldability" and collapse dynamics. *Phys Rev Lett* 90(16):168103.
17. Zhou R, Huang X, Margulis CJ, & Berne BJ (2004) Hydrophobic Collapse in Multidomain Protein Folding. *Science* 305(5690):1605-1609.
18. Ratner V, Amir D, Kahana E, & Haas E (2005) Fast collapse but slow formation of secondary structure elements in the refolding transition of E. coli adenylate kinase. *J. Mol. Biol.* 352(3):683-699.
19. Ziv G, Thirumalai D, & Haran G (2009) Collapse transition in proteins. *Phys. Chem. Chem. Phys.* 11(1):83-93.

20. Dasgupta A & Udgaonkar JB (2010) Evidence for initial non-specific polypeptide chain collapse during the refolding of the SH3 domain of PI3 kinase. *J. Mol. Biol.* 403(3):430-445.
21. Di Paolo A, Balbeur D, De Pauw E, Redfield C, & Matagne A (2010) Rapid collapse into a molten globule is followed by simple two-state kinetics in the folding of lysozyme from bacteriophage lambda. *Biochemistry* 49(39):8646-8657.
22. Ballew RM, Sabelko J, & Gruebele M (1996) Direct observation of fast protein folding: the initial collapse of apomyoglobin. *Proc Natl Acad Sci U S A* 93(12):5759-5764.
23. Parker MJ, Sessions RB, Badcoe IG, & Clarke AR (1996) The development of tertiary interactions during the folding of a large protein. *Folding & design* 1(2):145-156.
24. Nath U & Udgaonkar JB (1997) Folding of tryptophan mutants of barstar: evidence for an initial hydrophobic collapse on the folding pathway. *Biochemistry* 36(28):8602-8610.
25. Chen L, Wildegger G, Kiefhaber T, Hodgson KO, & Doniach S (1998) Kinetics of lysozyme refolding: structural characterization of a non-specifically collapsed state using time-resolved X-ray scattering. *J. Mol. Biol.* 276(1):225-237.
26. Segel DJ*, et al.* (1999) Characterization of transient intermediates in lysozyme folding with time-resolved small-angle X-ray scattering. *J. Mol. Biol.* 288(3):489-499.
27. Hagen SJ & Eaton WA (2000) Two-state expansion and collapse of a polypeptide. *J. Mol. Biol.* 301(4):1019-1027.
28. Pollack L*, et al.* (2001) Time resolved collapse of a folding protein observed with small angle x-ray scattering. *Phys Rev Lett* 86(21):4962-4965.
29. Arai M*, et al.* (2002) Fast compaction of alpha-lactalbumin during folding studied by stopped-flow X-ray scattering. *J. Mol. Biol.* 321(1):121-132.
30. Ferguson N & Fersht AR (2003) Early events in protein folding. *Curr Opin Struct Biol* 13(1):75-81.
31. Lapidus LJ*, et al.* (2007) Protein hydrophobic collapse and early folding steps observed in a microfluidic mixer. *Biophys J* 93(1):218-224.
32. Thirumalai D (1995) From Minimal Models to Real Proteins: Time Scales for Protein Folding Kinetics. *J. Phys. I France* 5(11):1457-1467.
33. Jacob J, Krantz B, Dothager RS, Thiyagarajan P, & Sosnick TR (2004) Early collapse is not an obligate step in protein folding. *J. Mol. Biol.* 338(2):369-382.
34. Jacob J, Dothager RS, Thiyagarajan P, & Sosnick TR (2007) Fully reduced ribonuclease A does not expand at high denaturant concentration or temperature. *J. Mol. Biol.* 367(3):609-615.
35. Tanford C (1968) Protein denaturation. *Adv. Protein Chem.* 23:121-282.
36. McCarney ER, Kohn JE, & Plaxco KW (2005) Is there or isn't there? The case for (and against) residual structure in chemically denatured proteins. *Critical Reviews in Biochemistry and Molecular Biology* 40(4):181-189.
37. Santoro MM & Bolen DW (1988) Unfolding free energy changes determined by the linear extrapolation method. 1. Unfolding of phenylmethanesulfonyl alpha-chymotrypsin using different denaturants. *Biochemistry* 27(21):8063-8068.
38. Mason PE*, et al.* (2004) The structure of aqueous guanidinium chloride solutions. *J. Am. Chem. Soc.* 126(37):11462-11470.
39. Mason PE, Neilson GW, Dempsey CE, Barnes AC, & Cruickshank JM (2003) The hydration structure of guanidinium and thiocyanate ions: implications for protein stability in aqueous solution. *Proc Natl Acad Sci U S A* 100(8):4557-4561.

40. England JL, Pande VS, & Haran G (2008) Chemical denaturants inhibit the onset of dewetting. *J. Am. Chem. Soc.* 130(36):11854-11855.
41. England JL & Haran G (2011) Role of solvation effects in protein denaturation: from thermodynamics to single molecules and back. *Annual review of physical chemistry* 62:257-277.
42. Dill KA (1985) Theory for the folding and stability of globular proteins. *Biochemistry* 24(6):1501-1509.
43. Chan HS & Dill KA (1990) THE EFFECTS OF INTERNAL CONSTRAINTS ON THE CONFIGURATIONS OF CHAIN MOLECULES. *J. Chem. Phys.* 92(5):3118-3135.
44. Dill KA (1990) Dominant forces in protein folding. *Biochemistry* 29(31):7133-7155.
45. Stigter D, Alonso DO, & Dill KA (1991) Protein stability: electrostatics and compact denatured states. *Proc Natl Acad Sci U S A* 88(10):4176-4180.
46. Dill KA & Shortle D (1991) Denatured States of Proteins. *Annu. Rev. Biochem.* 60(1):795-825.
47. Plaxco KW, Millett IS, Segel DJ, Doniach S, & Baker D (1999) Chain collapse can occur concomitantly with the rate-limiting step in protein folding. *Nat. Struct. Biol.* 6(6):554-556.
48. Sosnick TR & Barrick D (2011) The folding of single domain proteins--have we reached a consensus? *Current opinion in structural biology* 21(1):12-24.
49. Haran G (2012) How, when and why proteins collapse: the relation to folding. *Current opinion in structural biology* 22(1):14-20.
50. Sosnick TR & Baxa MC (2013) Revealing what gets buried first in protein folding. *Proc Natl Acad Sci U S A*.
51. Stryer L (1965) The interaction of a naphthalene dye with apomyoglobin and apohemoglobin. A fluorescent probe of non-polar binding sites. *J. Mol. Biol.* 13(2):482-495.
52. Goto Y, Azuma T, & Hamaguchi K (1979) Refolding of the immunoglobulin light chain. *Journal of biochemistry* 85(6):1427-1438.
53. Woody RW (1978) Aromatic side-chain contributions to the far ultraviolet circular dichroism of peptides and proteins. *Biopolymers* 17(6):1451-1467.
54. Kuwajima K, Garvey EP, Finn BE, Matthews CR, & Sugai S (1991) Transient intermediates in the folding of dihydrofolate reductase as detected by far-ultraviolet circular dichroism spectroscopy. *Biochemistry* 30(31):7693-7703.
55. Vuilleumier S, Sancho J, Loewenthal R, & Fersht AR (1993) Circular dichroism studies of barnase and its mutants: characterization of the contribution of aromatic side chains. *Biochemistry* 32(39):10303-10313.
56. Woody RW (1994) Contributions of tryptophan side chains to the far-ultraviolet circular dichroism of proteins. *Eur Biophys J* 23(4):253-262.
57. Jennings PA & Wright PE (1993) Formation of a molten globule intermediate early in the kinetic folding pathway of apomyoglobin. *Science* 262(5135):892-896.
58. Uzawa T*, et al.* (2004) Collapse and search dynamics of apomyoglobin folding revealed by submillisecond observations of alpha-helical content and compactness. *Proc Natl Acad Sci U S A* 101(5):1171-1176.
59. Krishna MM & Englander SW (2007) A unified mechanism for protein folding: predetermined pathways with optional errors. *Protein Sci.* 16(3):449-464.

60. Maity H, Maity M, & Englander SW (2004) How cytochrome c folds, and why: submolecular foldon units and their stepwise sequential stabilization. *J. Mol. Biol.* 343(1):223-233.
61. Krishna MM, Maity H, Rumbley JN, Lin Y, & Englander SW (2006) Order of steps in the cytochrome C folding pathway: evidence for a sequential stabilization mechanism. *J. Mol. Biol.* 359(5):1410-1419.
62. Krishna MM, Maity H, Rumbley JN, & Englander SW (2007) Branching in the sequential folding pathway of cytochrome c. *Protein Sci.* 16(9):1946-1956.
63. Krishna MM, Lin Y, Mayne L, & Englander SW (2003) Intimate view of a kinetic protein folding intermediate: residue-resolved structure, interactions, stability, folding and unfolding rates, homogeneity. *J. Mol. Biol.* 334(3):501-513.
64. Silverman JA & Harbury PB (2002) The equilibrium unfolding pathway of a (beta/alpha)8 barrel. *J. Mol. Biol.* 324(5):1031-1040.
65. Stratton MM, Cutler TA, Ha JH, & Loh SN (2010) Probing local structural fluctuations in myoglobin by size-dependent thiol-disulfide exchange. *Protein Sci.* 19(8):1587-1594.
66. Lindorff-Larsen K, Trbovic N, Maragakis P, Piana S, & Shaw DE (2012) Structure and dynamics of an unfolded protein examined by molecular dynamics simulation. *J. Am. Chem. Soc.* 134(8):3787-3791.
67. Piana S, Lindorff-Larsen K, & Shaw DE (2012) Protein folding kinetics and thermodynamics from atomistic simulation. *Proc Natl Acad Sci U S A*.
68. Dror RO, Dirks RM, Grossman JP, Xu H, & Shaw DE (2012) Biomolecular simulation: a computational microscope for molecular biology. *Annu. Rev. Biophys.* 41:429-452.
69. Bryngelson JD & Wolynes PG (1987) Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. USA* 84(21):7524-7528.
70. Baldwin RL (1995) The nature of protein folding pathways: the classical versus the new view. *J Biomol NMR* 5(2):103-109.
71. Brooks CL, 3rd, Gruebele M, Onuchic JN, & Wolynes PG (1998) Chemical physics of protein folding. *Proc Natl Acad Sci U S A* 95(19):11037-11038.
72. Onuchic JN & Wolynes PG (2004) Theory of protein folding. *Current opinion in structural biology* 14(1):70-75.
73. Wolynes PG, Onuchic JN, & Thirumalai D (1995) Navigating the folding routes. *Science* 267(5204):1619-1620.
74. Plotkin SS & Onuchic JN (2002) Understanding protein folding with energy landscape theory. Part II: Quantitative aspects. *Q. Rev. Biophys.* 35(3):205-286.
75. Plotkin SS & Onuchic JN (2002) Understanding protein folding with energy landscape theory - Part I: Basic concepts. *Quarterly Reviews of Biophysics* 35(2):111-167.
76. Chahine J, Nymeyer H, Leite VB, Socci ND, & Onuchic JN (2002) Specific and nonspecific collapse in protein folding funnels. *Phys Rev Lett* 88(16):168101.
77. Onuchic JN, Nymeyer H, Garcia AE, Chahine J, & Socci ND (2000) The energy landscape theory of protein folding: insights into folding mechanisms and scenarios. *Adv. Protein Chem.* 53:87-152.
78. Socci ND, Onuchic JN, & Wolynes PG (1998) Protein folding mechanisms and the multidimensional folding funnel. *Proteins* 32(2):136-158.
79. Onuchic JN, Luthey-Schulten Z, & Wolynes PG (1997) Theory of protein folding: the energy landscape perspective. *Annual review of physical chemistry* 48:545-600.

80. Bryngelson JD, Onuchic JN, Socci ND, & Wolynes PG (1995) Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins* 21(3):167-195.
81. Wolynes PG (2005) Energy landscapes and solved protein-folding problems. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences* 363(1827):453-464; discussion 464-457.
82. Dill KA & Chan HS (1997) From Levinthal to pathways to funnels. *Nat. Struct. Biol.* 4(1):10-19.
83. Kiefhaber T (1995) Kinetic traps in lysozyme folding. *Proc Natl Acad Sci U S A* 92(20):9029-9033.
84. Kiefhaber T, Bachmann A, Wildegger G, & Wagner C (1997) Direct measurement of nucleation and growth rates in lysozyme folding. *Biochemistry* 36(17):5108-5112.
85. Wildegger G & Kiefhaber T (1997) Three-state model for lysozyme folding: triangular folding mechanism with an energetically trapped intermediate. *J. Mol. Biol.* 270(2):294-304.
86. Bieri O, Wildegger G, Bachmann A, Wagner C, & Kiefhaber T (1999) A salt-induced kinetic intermediate is on a new parallel pathway of lysozyme folding. *Biochemistry* 38(38):12460-12470.
87. Bieri O & Kiefhaber T (2001) Origin of apparent fast and non-exponential kinetics of lysozyme folding measured in pulsed hydrogen exchange experiments. *J. Mol. Biol.* 310(4):919-935.
88. Creighton T (1992) *Proteins: Structures and Molecular Properties* (W.H. Freeman, New York).
89. Creighton T (1992) *Protein Folding* (W.H. Freeman, New York).
90. Fersht A (1998) *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding* (W. H. Freeman, New York).
91. Apic G, Gough J, & Teichmann SA (2001) Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J. Mol. Biol.* 310(2):311-325.
92. Ekman D, Bjorklund AK, Frey-Skott J, & Elofsson A (2005) Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. *J. Mol. Biol.* 348(1):231-243.
93. Gerstein M (1998) How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Folding & design* 3(6):497-512.
94. Liu J & Rost B (2004) CHOP: parsing proteins into structural domains. *Nucleic acids research* 32(Web Server issue):W569-571.
95. Teichmann SA, Chothia C, & Gerstein M (1999) Advances in structural genomics. *Curr Opin Struct Biol* 9(3):390-399.
96. Jaenicke R (1999) Stability and folding of domain proteins. *Progress in biophysics and molecular biology* 71(2):155-241.
97. Batey S, Nickson AA, & Clarke J (2008) Studying the folding of multidomain proteins. *HFSP journal* 2(6):365-377.
98. Sharff AJ, Rodseth LE, Spurlino JC, & Quiocho FA (1992) Crystallographic evidence of a large ligand-induced hinge-twist motion between the two domains of the maltodextrin binding protein involved in active transport and chemotaxis. *Biochemistry* 31(44):10657-10663.

99. Kabsch W & Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22(12):2577-2637.
100. Plaxco KW, Simons KT, & Baker D (1998) Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* 277(4):985-994.
101. Baker D (Determine A Proteins Contact Order.
102. Ivankov DN*, et al.* (2003) Contact order revisited: Influence of protein size on the folding rate. *Protein Sci.* 12(9):2057-2062.
103. Lorenz T & Reinstein J (2008) The Influence of Proline Isomerization and Off-Pathway Intermediates on the Folding Mechanism of Eukaryotic UMP/CMP Kinase. *J. Mol. Biol.* 381(2):443-455.
104. Stagg L, Samiotakis A, Homouz D, Cheung MS, & Wittung-Stafshede P (2010) Residue-specific analysis of frustration in the folding landscape of repeat beta/alpha protein apoflavodoxin. *J. Mol. Biol.* 396(1):75-89.
105. Garbuzynskiy SO, Ivankov DN, Bogatyreva NS, & Finkelstein AV (2013) Golden triangle for folding rates of globular proteins. *Proc. Natl. Acad. Sci. USA* 110(1):147-150.
106. Chun SY, Strobel S, Bassford P, Jr., & Randall LL (1993) Folding of maltose-binding protein. Evidence for the identity of the rate-determining step in vivo and in vitro. *J. Biol. Chem.* 268(28):20855-20862.
107. Raffy S, Sassoon N, Hofnung M, & Betton JM (1998) Tertiary structure-dependence of misfolding substitutions in loops of the maltose-binding protein. *Protein Sci.* 7(10):2136-2142.
108. Betton JM, Sassoon N, Hofnung M, & Laurent M (1998) Degradation versus aggregation of misfolded maltose-binding protein in the periplasm of Escherichia coli. *J. Biol. Chem.* 273(15):8897-8902.
109. Ganesh C, Zaidi FN, Udgaonkar JB, & Varadarajan R (2001) Reversible formation of on-pathway macroscopic aggregates during the folding of maltose binding protein. *Protein Sci.* 10(8):1635-1644.
110. Roder H, Elove GA, & Englander SW (1988) Structural characterization of folding intermediates in cytochrome c by H-exchange labelling and proton NMR. *Nature* 335(6192):700-704.
111. Chamberlain AK, Handel TM, & Marqusee S (1996) Detection of rare partially folded molecules in equilibrium with the native conformation of RNaseH. *Nat. Struct. Biol.* 3(9):782-787.
112. Raschke TM & Marqusee S (1997) The kinetic folding intermediate of ribonuclease H resembles the acid molten globule and partially unfolded molecules detected under native conditions. *Nature structural biology* 4(4):298-304.
113. Nishimura C, Dyson HJ, & Wright PE (2006) Identification of native and non-native structure in kinetic folding intermediates of apomyoglobin. *J. Mol. Biol.* 355:139-156.
114. Pan J, Han J, Borchers CH, & Konermann L (2010) Characterizing Short-Lived Protein Folding Intermediates by Top-Down Hydrogen Exchange Mass Spectrometry. *Anal. Chem.* 82(20):8591-8597.
115. Fuentes EJ & Wand AJ (1998) Local dynamics and stability of apocytochrome b(562) examined by hydrogen exchange. *Biochemistry* 37(11):3687-3698.
116. Feng HQ, Vu ND, & Bai YW (2005) Detection of a hidden folding intermediate of the third domain of PDZ. *J. Mol. Biol.* 346(1):345-353.

117. Zhou Z, Feng H, & Bai Y (2006) Detection of a hidden folding intermediate in the focal adhesion target domain: Implications for its function and folding. *Proteins: Structure, Function, and Genetics* 65(2):259-265.
118. Kato H, Vu ND, Feng H, Zhou Z, & Bai Y (2007) The folding pathway of T4 lysozyme: an on-pathway hidden folding intermediate. *J. Mol. Biol.* 365(3):881-891.
119. Bai Y (2006) Protein folding pathways studied by pulsed- and native-state hydrogen exchange. *Chem Rev* 106(5):1757-1768.
120. Bollen YJ, Sanchez IE, & van Mierlo CP (2004) Formation of on- and off-pathway intermediates in the folding kinetics of Azotobacter vinelandii apoflavodoxin. *Biochemistry* 43(32):10475-10489.
121. Korzhnev DM, Religa TL, Lundstrom P, Fersht AR, & Kay LE (2007) The folding pathway of an FF domain: Characterization of an on-pathway intermediate state under folding conditions by N-15, C-13(alpha) and C-13-methyl relaxation dispersion and H-1/(2) H-exchange NMR Spectroscopy. *J. Mol. Biol.* 372(2):497-512.
122. Korzhnev DM, Religa TL, Banachewicz W, Fersht AR, & Kay LE (2010) A Transient and Low-Populated Protein-Folding Intermediate at Atomic Resolution. *Science* 329(5997):1312-1316.
123. Gsponer J*, et al.* (2006) Determination of an ensemble of structures representing the intermediate state of the bacterial immunity protein Im7. *Proc. Natl. Acad. Sci. USA* 103(1):99-104.
124. Kulakarni SK*, et al.* (2006) A near-native state on the slow refolding pathway of hen lysozyme. *Protein Sci.* 8:35-44.
125. Feng H, Zhou Z, & Bai Y (2005) A protein folding pathway with multiple folding intermediates at atomic resolution. *Proceedings of the National Academy of Sciences USA* 102(14):5026-5031.
126. Yan S, Kennedy SD, & Koide S (2002) Thermodynamic and kinetic exploration of the energy landscape of *Borrelia burgdorferi* OspA by native-state hydrogen exchange. *J. Mol. Biol.* 323:363-375.
127. Dahiya V & Chaudhuri TK (2013) Functional intermediate in the refolding pathway of a large and multi-domain protein Malate synthase G. *Biochemistry*.
128. Graczer E*, et al.* (2009) Symmetrical refolding of protein domains and subunits: example of the dimeric two-domain 3-isopropylmalate dehydrogenases. *Biochemistry* 48(5):1123-1134.
129. Sharma S*, et al.* (2008) Monitoring protein conformation along the pathway of chaperonin-assisted folding. *Cell* 133(1):142-153.
130. Lapidus LJ (2013) Exploring the top of the protein folding funnel by experiment. *Current opinion in structural biology* 23(1):30-35.
131. Englander SW, Downer NW, & Teitelbaum H (1972) Hydrogen Exchange. *Annu. Rev. Biochem.* 41(1):903-924.
132. Bai Y, Milne JS, Mayne L, & Englander SW (1993) Primary structure effects on peptide group hydrogen exchange. *Proteins* 17(1):75-86.
133. Molday RS, Englander SW, & Kallen RG (1972) Primary structure effects on peptide group hydrogen exchange. *Biochemistry* 11(2):150-158.
134. Connelly GP, Bai Y, Jeng MF, & Englander SW (1993) Isotope effects in peptide group hydrogen exchange. *Proteins* 17(1):87-92.

135. Linderstrom-Lang K (1955) The pH-dependence of the deuterium exchange of insulin. *Biochim Biophys Acta* 18(2):308.
136. Linderstrøm-Lang KU & Schellman JA (1959) *Protein Structure and Enzyme Activity* (Academic Press, New York) 2 Ed.
137. Hvidt A (1964) A Discussion of the Ph Dependence of the Hydrogen-Deuterium Exchange of Proteins. *C R Trav Lab Carlsberg* 34:299-317.
138. Hvidt A & Nielsen SO (1966) Hydrogen exchange in proteins. *Adv. Protein Chem.* 21:287-386.
139. Englander SW & Kallenbach NR (1983) Hydrogen exchange and structural dynamics of proteins and nucleic acids. *Q. Rev. Biophys.* 16(4):521-655.
140. Skinner JJ, Lim WK, Bedard S, Black BE, & Englander SW (2012) Protein dynamics viewed by hydrogen exchange. *Protein Sci.* 21(7):996-1005.
141. Krishna MM, Hoang L, Lin Y, & Englander SW (2004) Hydrogen exchange methods to study protein folding. *Methods* 34(1):51-64.
142. Englander SW (2000) Protein folding intermediates and pathways studied by hydrogen exchange. *Annual review of biophysics and biomolecular structure* 29:213-238.
143. Bai Y, Sosnick TR, Mayne L, & Englander SW (1995) Protein folding intermediates: native-state hydrogen exchange. *Science* 269(5221):192-197.
144. Bai Y, Englander JJ, Mayne L, Milne JS, & Englander SW (1995) Thermodynamic parameters from hydrogen exchange measurements. *Methods Enzymol* 259:344-356.
145. Bai Y, Milne JS, Mayne L, & Englander SW (1994) Protein stability parameters measured by hydrogen exchange. *Proteins* 20(1):4-14.
146. Chetty PS*, et al.* (2009) Helical structure and stability in human apolipoprotein A-I by hydrogen exchange and mass spectrometry. *Proc Natl Acad Sci U S A* 106(45):19005-19010.
147. Liuni P, Jeganathan A, & Wilson DJ (2012) Conformer Selection and Intensified Dynamics During Catalytic Turnover in Chymotrypsin. *Angewandte Chemie International Edition* 51(38):9666-9669.
148. Wüthrich K (1986) *NMR of proteins and nucleic acids* (Wiley, New York) pp xv, 292.
149. Tugarinov V, Hwang PM, & Kay LE (2004) Nuclear magnetic resonance spectroscopy of high-molecular-weight proteins. *Annu. Rev. Biochem.* 73:107-146.
150. Nucci NV*, et al.* (2011) Optimization of NMR spectroscopy of encapsulated proteins dissolved in low viscosity fluids. *J Biomol NMR* 50(4):421-430.
151. Tugarinov V, Kanelis V, & Kay LE (2006) Isotope labeling strategies for the study of high-molecular-weight proteins by solution NMR spectroscopy. *Nature protocols* 1(2):749-754.
152. Tugarinov V, Muhandiram R, Ayed A, & Kay LE (2002) Four-dimensional NMR spectroscopy of a 723-residue protein: chemical shift assignments and secondary structure of malate synthase g. *J. Am. Chem. Soc.* 124(34):10025-10035.
153. Schanda P, Kupce E, & Brutscher B (2005) SOFAST-HMQC experiments for recording two-dimensional heteronuclear correlation spectra of proteins within a few seconds. *J Biomol NMR* 33(4):199-211.
154. Schanda P & Brutscher B (2006) Hadamard frequency-encoded SOFAST-HMQC for ultrafast two-dimensional protein NMR. *J Magn Reson* 178(2):334-339.

155. Schanda P, Forge V, & Brutscher B (2007) Protein folding and unfolding studied at atomic resolution by fast two-dimensional NMR spectroscopy. *Proc Natl Acad Sci U S A* 104(27):11257-11262.
156. Kern T, Schanda P, & Brutscher B (2008) Sensitivity-enhanced IPAP-SOFAST-HMQC for fast-pulsing 2D NMR with reduced radiofrequency load. *J Magn Reson* 190(2):333-338.
157. Gal M, Kern T, Schanda P, Frydman L, & Brutscher B (2009) An improved ultrafast 2D NMR experiment: towards atom-resolved real-time studies of protein kinetics at multi-Hz rates. *J Biomol NMR* 43(1):1-10.
158. Amero C*, et al.* (2009) Fast two-dimensional NMR spectroscopy of high molecular weight protein assemblies. *J. Am. Chem. Soc.* 131(10):3448-3449.
159. Lescop E, Schanda P, & Brutscher B (2007) A set of BEST triple-resonance experiments for time-optimized protein resonance assignment. *J Magn Reson* 187(1):163-169.
160. Lescop E, Schanda P, Rasia R, & Brutscher B (2007) Automated spectral compression for fast multidimensional NMR and increased time resolution in real-time NMR spectroscopy. *J. Am. Chem. Soc.* 129(10):2756-2757.
161. Evenas J*, et al.* (2001) Ligand-induced structural changes to maltodextrin-binding protein as studied by solution NMR spectroscopy. *J. Mol. Biol.* 309(4):961-974.
162. Sheshadri S, Lingaraju GM, & Varadarajan R (1999) Denaturant mediated unfolding of both native and molten globule states of maltose binding protein are accompanied by large deltaCp's. *Protein Sci.* 8(8):1689-1695.
163. Thomson J, Liu Y, Sturtevant JM, & Quiocho FA (1998) A thermodynamic study of the binding of linear and cyclic oligosaccharides to the maltodextrin-binding protein of Escherichia coli. *Biophys. Chem.* 70(2):101-108.
164. Novokhatny V & Ingham K (1997) Thermodynamics of maltose binding protein unfolding. *Protein Sci.* 6(1):141-146.
165. Ganesh C, Shah AN, Swaminathan CP, Surolia A, & Varadarajan R (1997) Thermodynamic characterization of the reversible, two-state unfolding of maltose binding protein, a large two-domain protein. *Biochemistry* 36(16):5020-5028.
166. Mayne L & Englander SW (2000) Two-state vs. multistate protein unfolding studied by optical melting and hydrogen exchange. *Protein Sci.* 9(10):1873-1877.
167. Merstorf C*, et al.* (2012) Mapping the Conformational Stability of Maltose Binding Protein at the Residue Scale Using Nuclear Magnetic Resonance Hydrogen Exchange Experiments. *Biochemistry* 51(44):8919-8930.
168. Kupce E & Freeman R (2005) Fast multidimensional NMR: radial sampling of evolution space. *J Magn Reson* 173(2):317-321.
169. Coggins BE & Zhou P (2008) High resolution 4-D spectroscopy with sparse concentric shell sampling and FFT-CLEAN. *Journal of Biomolecular Nmr* 42(4):225-239.
170. Coggins BE & Zhou P (2006) Polar Fourier transforms of radially sampled NMR data. *J Magn Reson* 182(1):84-95.
171. Kazimierczuk K, Kozminski W, & Zhukov I (2006) Two-dimensional Fourier transform of arbitrarily sampled NMR data sets. *J Magn Reson* 179(2):323-328.
172. Kupce E & Freeman R (2004) Projection-reconstruction technique for speeding up multidimensional NMR spectroscopy. *J. Am. Chem. Soc.* 126(20):6429-6440.
173. Yoon JW, Godsill S, Kupce E, & Freeman R (2006) Deterministic and statistical methods for reconstructing multidimensional NMR spectra. *Magn Reson Chem* 44(3):197-209.

174. Gledhill JM & Wand AJ (2007) Phasing arbitrarily sampled multidimensional NMR data. *J Magn Reson* 187(2):363-370.
175. Marion D (2006) Processing of ND NMR spectra sampled in polar coordinates: a simple Fourier transform instead of a reconstruction. *J Biomol NMR* 36(1):45-54.
176. Schanda P, Van Melckebeke H, & Brutscher B (2006) Speeding up three-dimensional protein NMR experiments to a few minutes. *J. Am. Chem. Soc.* 128(28):9042-9043.
177. Schanda P & Brutscher B (2005) Very fast two-dimensional NMR spectroscopy for real-time investigation of dynamic events in proteins on the time scale of seconds. *J. Am. Chem. Soc.* 127(22):8014-8015.
178. Freeman R & Kupce E (2003) New methods for fast multidimensional NMR. *J Biomol NMR* 27(2):101-113.
179. Gledhill JM, Jr. & Joshua Wand A (2008) Optimized angle selection for radial sampled NMR experiments. *J Magn Reson* 195(2):169-178.
180. Gledhill JM, Jr. & Wand AJ (2012) Al NMR: a novel NMR data processing program optimized for sparse sampling. *J Biomol NMR* 52(1):79-89.
181. Englander SW (1963) A HYDROGEN EXCHANGE METHOD USING TRITIUM AND SEPHADEX ITS APPLICATION TO RIBONUCLEASE. *Biochemistry* 2(4):798-&.
182. Englander JJ, Rogero JR, & Englander SW (1985) Protein hydrogen exchange studied by the fragment separation method. *Anal. Biochem.* 147(1):234-244.
183. Behe MJ & Englander SW (1979) MIXED GELATION THEORY - KINETICS, EQUILIBRIUM AND GEL INCORPORATION IN SICKLE HEMOGLOBIN MIXTURES. *J. Mol. Biol.* 133(1):137-&.
184. Englander SW*, et al.* (1980) INDIVIDUAL BREATHING REACTIONS MEASURED IN HEMOGLOBIN BY HYDROGEN-EXCHANGE METHODS. *Biophys. J.* 32(1):577-589.
185. Englander SW & Kallenbach NR (1983) HYDROGEN-EXCHANGE AND STRUCTURAL DYNAMICS OF PROTEINS AND NUCLEIC-ACIDS. *Quarterly Reviews of Biophysics* 16(4):521-655.
186. Rosa JJ & Richards FM (1979) An experimental procedure for increasing the structural resolution of chemical hydrogen-exchange measurements on proteins: application to ribonuclease S peptide. *J. Mol. Biol.* 133(3):399-416.
187. Wagner G & Wüthrich K (1982) Amide proton exchange and surface conformation of the basic pancreatic trypsin inhibitor in solution : Studies with two-dimensional nuclear magnetic resonance. *J. Mol. Biol.* 160(2):343-361.
188. Wand AJ, Roder H, & Englander SW (1986) Two-dimensional 1H NMR studies of cytochrome c: hydrogen exchange in the N-terminal helix. *Biochemistry* 25(5):1107-1114.
189. Katta V & Chait BT (1991) Conformational changes in proteins probed by hydrogen-exchange electrospray-ionization mass spectrometry. *Rapid Commun. Mass Spectrom.* 5(4):214-217.
190. Zhang Z & Smith DL (1993) Determination of amide hydrogen exchange by mass spectrometry: a new tool for protein structure elucidation. *Protein Sci.* 2(4):522-531.
191. Ferguson PL*, et al.* (2006) Hydrogen/Deuterium Scrambling during Quadrupole Time-of-Flight MS/MS Analysis of a Zinc-Binding Protein Domain. *Anal. Chem.* 79(1):153-160.
192. Rand KD, Zehl M, Jensen ON, & Jorgensen TJD (2009) Protein Hydrogen Exchange Measured at Single-Residue Resolution by Electron Transfer Dissociation Mass Spectrometry. *Anal. Chem.* 81(14):5577-5584.

193. Abzalimov RR, Kaplan DA, Easterling ML, & Kaltashov IA (2009) Protein Conformations Can Be Probed in Top-Down HDX MS Experiments Utilizing Electron Transfer Dissociation of Protein Ions Without Hydrogen Scrambling. *J. Am. Soc. Mass Spectrom.* 20(8):1514-1517.
194. Pan J, Han J, Borchers CH, & Konermann L (2009) Hydrogen/Deuterium Exchange Mass Spectrometry with Top-Down Electron Capture Dissociation for Characterizing Structural Transitions of a 17 kDa Protein. *J. Am. Chem. Soc.* 131(35):12801-12808.
195. Zehl M, Rand KD, Jensen ON, & Jorgensen TJD (2008) Electron Transfer Dissociation Facilitates the Measurement of Deuterium Incorporation into Selectively Labeled Peptides with Single Residue Resolution. *J. Am. Chem. Soc.* 130(51):17453-17459.
196. Rand KD, Pringle SD, Morris M, Engen JR, & Brown JM (2011) ETD in a Traveling Wave Ion Guide at Tuned Z-Spray Ion Source Conditions Allows for Site-Specific Hydrogen/Deuterium Exchange Measurements. *J. Am. Soc. Mass Spectrom.* 22(10):1784-1793.
197. Fang J, Rand KD, Beuning PJ, & Engen JR (2011) False EX1 signatures caused by sample carryover during HX MS analyses. *Int J Mass Spectrom* 302(1-3):19-25.
198. Wu Y, Kaveti S, & Engen JR (2006) Extensive deuterium back-exchange in certain immobilized pepsin columns used for H/D exchange mass spectrometry. *Anal. Chem.* 78(5):1719-1723.
199. Wang L, Pan H, & Smith DL (2002) Hydrogen exchange-mass spectrometry: optimization of digestion conditions. *Molecular & cellular proteomics : MCP* 1(2):132-138.
200. Hamuro Y*, et al.* (2002) Domain organization of D-AKAP2 revealed by enhanced deuterium exchange-mass spectrometry (DXMS). *J. Mol. Biol.* 321(4):703-714.
201. Eng JK, McCormack AL, & Yates Iii JR (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* 5(11):976-989.
202. Kan ZY, Mayne L, Chetty PS, & Englander SW (2011) ExMS: Data Analysis for HX-MS Experiments. *J. Am. Soc. Mass Spectrom.* 22(11):1906-1915.
203. Hather G, Higdon R, Bauman A, von Haller PD, & Kolker E (Estimating false discovery rates for peptide and protein identification using randomized databases. *Proteomics* 10(12):2369-2376.
204. Hamuro Y, Coales SJ, Molnar KS, Tuske SJ, & Morrow JA (2008) Specificity of immobilized porcine pepsin in H/D exchange compatible conditions. *Rapid Commun. Mass Spectrom.* 22(7):1041-1046.
205. Konermann L, Pan J, & Liu YH (2011) Hydrogen exchange mass spectrometry for studying protein structure and dynamics. *Chem. Soc. Rev.* 40(3):1224-1234.
206. Marcsisin SR & Engen JR (2010) Hydrogen exchange mass spectrometry: what is it and what can it tell us? *Anal. Bioanal. Chem.* 397(3):967-972.
207. Tsutsui Y & Wintrode PL (2007) Hydrogen/deuterium, exchange-mass spectrometry: A powerful tool for probing protein structure, dynamics and interactions. *Curr. Med. Chem.* 14(22):2344-2358.
208. Hotchko M, Anand GS, Komives EA, & Ten Eyck LF (2006) Automated extraction of backbone deuteration levels from amide H/2H mass spectrometry experiments. *Protein Sci.* 15(3):583-601.

209. Wu Y, Engen JR, & Hobbins WB (2006) Ultra performance liquid chromatography (UPLC) further improves hydrogen/deuterium exchange mass spectrometry. *J. Am. Soc. Mass Spectrom.* 17(2):163-167.
210. Kipping M & Schierhorn A (2003) Improving hydrogen/deuterium exchange mass spectrometry by reduction of the back-exchange effect. *J. Mass Spectrom.* 38(3):271-276.
211. Emmett MR*, et al.* (2006) Supercritical fluid chromatography reduction of hydrogen/deuterium back exchange in solution-phase hydrogen/deuterium exchange with mass spectrometric analysis. *Anal. Chem.* 78(19):7058-7060.
212. Zhang H-M, Bou-Assaf G, Emmett M, & Marshall A (2009) Fast reversed-phase liquid chromatography to reduce back exchange and increase throughput in H/D exchange monitored by FT-ICR mass spectrometry. *J. Am. Soc. Mass Spectrom.* 20(3):520-524.
213. Keppel TR, Jacques ME, Young RW, Ratzlaff KL, & Weis DD (2011) An Efficient and Inexpensive Refrigerated LC System for H/D Exchange Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* 22(8):1472-1476.
214. Rand KD, Lund FW, Amon S, & Jørgensen TJD (2011) Investigation of amide hydrogen back-exchange in Asp and His repeats measured by hydrogen (1H/2H) exchange mass spectrometry. *Int. J. Mass Spectrom.* 302(1-3):110-115.
215. Valeja S, Emmett M, & Marshall A (2012) Polar Aprotic Modifiers for Chromatographic Separation and Back-Exchange Reduction for Protein Hydrogen/Deuterium Exchange Monitored by Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *J. Am. Soc. Mass Spectrom.*:1-9.
216. Coales SJ, Tomasso JC, & Hamuro Y (2008) Effects of electrospray capillary temperature on amide hydrogen exchange. *Rapid Commun. Mass Spectrom.* 22(9):1367-1371.
217. Zhang Z, Zhang A, & Xiao G (2012) Improved Protein Hydrogen/Deuterium Exchange Mass Spectrometry Platform with Fully Automated Data Processing. *Anal. Chem.* 84(11):4942-4949.
218. Zhang Z (1995) Protein hydrogen exchange determined by mass spectrometry: A new tool for probing protein high-order structure and structural changes. Ph.D. 9601613 (Purdue University, United States -- Indiana).
219. Skinner JJ, Lim WK, Bédard S, Black BE, & Englander SW (2012) Protein hydrogen exchange: Testing current models. *Protein Sci.* 21(7):987-995.
220. Sheff J, Rey M, & Schriemer D (2013) Peptide–Column Interactions and Their Influence on Back Exchange Rates in Hydrogen/Deuterium Exchange-MS. *J. Am. Soc. Mass Spectrom.*:1-10.
221. Kreshuk A*, et al.* (2011) Automated detection and analysis of bimodal isotope peak distributions in H/D exchange mass spectrometry using HeXicon. *Int. J. Mass Spectrom.* 302(1-3):125-131.
222. Weis DD, Engen JR, & Kass IJ (2006) Semi-automated data processing of hydrogen exchange mass spectra using HX-Express. *J. Am. Soc. Mass Spectrom.* 17(12):1700-1703.
223. Abzalimov RR & Kaltashov IA (2006) Extraction of local hydrogen exchange data from HDX CAD MS measurements by deconvolution of isotopic distributions of fragment ions. *J. Am. Soc. Mass Spectrom.* 17(11):1543-1551.
224. Zhang J, Ramachandran P, Kumar R, & Gross ML (2013) H/D Exchange Centroid Monitoring is Insufficient to Show Differences in the Behavior of Protein States. *J. Am. Soc. Mass Spectrom.*

225. Chik JK, Vande Graaf JL, & Schriemer DC (2005) Quantitating the Statistical Distribution of Deuterium Incorporation To Extend the Utility of H/D Exchange MS Data. *Anal. Chem.* 78(1):207-214.
226. Levenburg K (1944) A Method for the Solution of Certain Non-Linear Problems in Least Squares. *Quarterly of Applied Mathematics* 2:164-168.
227. Oliphant TE (2007) Python for Scientific Computing. *Computing in Science & Engineering* 9(3):10-20.
228. Thirumalai D, Liu Z, O'Brien EP, & Reddy G (2012) Protein folding: from theory to practice. *Curr Opin Struct Biol*.
229. Dill KA & MacCallum JL (2012) The Protein-Folding Problem, 50 Years On. *Science* 338(6110):1042-1046.
230. Bowman GR, Voelz VA, & Pande VS (2011) Taming the complexity of protein folding. *Current Opinion in Structural Biology* 21(1):4-11.
231. Meng W, Lyle N, Luan B, Raleigh DP, & Pappu RV (2013) Experiments and simulations show how long-range contacts can form in expanded unfolded proteins with negligible secondary structure. *Proc Natl Acad Sci U S A* 110(6):2123-2128.
232. Soranno A*, et al.* (2012) Quantifying internal friction in unfolded and intrinsically disordered proteins with single-molecule spectroscopy. *Proc Natl Acad Sci U S A* 109(44):17800-17806.
233. Park S, Liu G, Topping TB, Cover WH, & Randall LL (1988) Modulation of folding pathways of exported proteins by the leader sequence. *Science* 239(4843):1033-1035.
234. Szmelcman S, Schwartz M, Silhavy TJ, & Boos W (1976) Maltose transport in Escherichia coli K12. A comparison of transport kinetics in wild-type and lambda-resistant mutants as measured by fluorescence quenching. *European journal of biochemistry / FEBS* 65(1):13-19.
235. Konermann L, Rodriguez AD, & Liu J (2012) On the formation of highly charged gaseous ions from unfolded proteins by electrospray ionization. *Anal. Chem.* 84(15):6798-6804.
236. Hall Z & Robinson CV (2012) Do charge state signatures guarantee protein conformations? *J. Am. Soc. Mass Spectrom.* 23(7):1161-1168.
237. Borgia A*, et al.* (2012) Localizing internal friction along the reaction coordinate of protein folding by combining ensemble and single-molecule fluorescence spectroscopy. *Nature communications* 3:1195.
238. Waldauer SA, Bakajin O, & Lapidus LJ (2010) Extremely slow intramolecular diffusion in unfolded protein L. *Proc. Natl. Acad. Sci. USA*.
239. Bowman GR & Pande VS (2010) Protein folded states are kinetic hubs. *Proc. Natl. Acad. Sci. USA*.
240. Fleming PJ & Rose GD (2005) Do all backbone polar groups in proteins form hydrogen bonds? *Protein Sci.* 14(7):1911-1917.
241. Munoz V & Serrano L (1995) Elucidating the folding problem of helical peptides using empirical parameters. II. Helix macrodipole effects and rational modification of the helical content of natural peptides. *J. Mol. Biol.* 245(3):275-296.
242. Munoz V & Serrano L (1995) Elucidating the folding problem of helical peptides using empirical parameters. III. Temperature and pH dependence. *J. Mol. Biol.* 245(3):297-308.
243. Munoz V & Serrano L (1994) Elucidating the folding problem of helical peptides using empirical parameters. *Nat. Struct. Biol.* 1(6):399-409.

244. Tang Y-C*, et al.* (2006) Structural Features of the GroEL-GroES Nano-Cage Required for Rapid Folding of Encapsulated Protein. *Cell* 125(5):903-914.
245. Tyagi NK, Fenton WA, Deniz AA, & Horwich AL (2011) Double mutant MBP refolds at same rate in free solution as inside the GroEL/GroES chaperonin chamber when aggregation in free solution is prevented. *FEBS letters* 585(12):1969-1972.
246. Bechtluft P*, et al.* (2007) Direct observation of chaperone-induced changes in a protein folding pathway. *Science* 318(5855):1458-1461.
247. Bowler BE (2012) Residual structure in unfolded proteins. *Current opinion in structural biology* 22(1):4-13.
248. Orevi T, Rahamim G, Hazan G, Amir D, & Haas E (2013) The loop hypothesis: contribution of early formed specific non-local interactions to the determination of protein folding pathways. *Biophys Rev* 5(2):85-98.
249. Wu Y, Kondrashkina E, Kayatekin C, Matthews CR, & Bilsel O (2008) Microsecond acquisition of heterogeneous structure in the folding of a TIM barrel protein. *Proc Natl Acad Sci U S A* 105(36):13367-13372.
250. Arai M, Iwakura M, Matthews CR, & Bilsel O (2011) Microsecond Subdomain Folding in Dihydrofolate Reductase. *J. Mol. Biol.* 410(2):329-342.
251. Basha E, O'Neill H, & Vierling E (2012) Small heat shock proteins and alpha-crystallins: dynamic proteins with flexible functions. *Trends in biochemical sciences* 37(3):106-117.
252. Houry WA, Frishman D, Eckerskorn C, Lottspeich F, & Hartl FU (1999) Identification of in vivo substrates of the chaperonin GroEL. *Nature* 402(6758):147-154.
253. Radford SE, Dobson CM, & Evans PA (1992) The folding of hen lysozyme involves partially structured intermediates and multiple pathways. *Nature* 358(6384):302-307.
254. Wu Y & Matthews CR (2002) Parallel channels and rate-limiting steps in complex protein folding reactions: prolyl isomerization and the alpha subunit of Trp synthase, a TIM barrel protein. *J. Mol. Biol.* 323(2):309-325.
255. Kamagata K, Sawano Y, Tanokura M, & Kuwajima K (2003) Multiple parallel-pathway folding of proline-free staphylococcal nuclease. *J. Mol. Biol.* 332(5):1143-1153.
256. Bedard S, Krishna MM, Mayne L, & Englander SW (2008) Protein folding: independent unrelated pathways or predetermined pathway with optional errors. *Proc Natl Acad Sci U S A* 105(20):7182-7187.
257. Zhou Z, Feng HQ, & Bai YW (2006) Detection of a hidden folding intermediate in the focal adhesion target domain: Implications for its function and folding. *Proteins-Structure Function and Bioinformatics* 65(2):259-265.
258. Kato H, Vu ND, Feng HQ, Zhou Z, & Bai YW (2007) The folding pathway of T4 lysozyme: An on-pathway hidden folding intermediate. *J. Mol. Biol.* 365(3):881-891.
259. Bai Y (2006) Energy barriers, cooperativity, and hidden intermediates in the folding of small proteins. *Biochemical and Biophysical Research Communications* 340(3):976-983.
260. Lindorff-Larsen K, Piana S, Dror RO, & Shaw DE (2011) How fast-folding proteins fold. *Science* 334(6055):517-520.
261. Prajapati RS, Indu S, & Varadarajan R (2007) Identification and Thermodynamic Characterization of Molten Globule States of Periplasmic Binding Proteins&#x2020. *Biochemistry* 46(36):10339-10352.
262. Maity H, Lim WK, Rumbley JN, & Englander SW (2003) Protein hydrogen exchange mechanism: local fluctuations. *Protein Sci.* 12(1):153-160.

263. Anderson JS, Hernandez G, & Lemaster DM (Sidechain conformational dependence of hydrogen exchange in model peptides. *Biophys. Chem.* 151(1-2):61-70.
264. Rockwood AL, Van Orden SL, & Smith RD (1995) Rapid Calculation of Isotope Distributions. *Anal. Chem.* 67(15):2699-2704.
265. Hsu CS (1984) Diophantine approach to isotopic abundance calculations. *Anal. Chem.* 56(8):1356-1361.