Title
Facilitation of protein 3-D structure determination using enhanced peptide amide deuterium exchange mass spectrometry (DXMS)

Permalink
https://escholarship.org/uc/item/05h1b0jv

Author
Pantazatos, Dennis Peter

Publication Date
2006

Peer reviewed|Thesis/dissertation
UNIVERSITY OF CALIFORNIA, SAN DIEGO

FACILITATION OF PROTEIN 3-D STRUCTURE DETERMINATION USING ENHANCED PEPTIDE AMIDE DEUTERIUM EXCHANGE MASS SPECTROMETRY (DXMS)

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in

Biomedical Sciences

by

Dennis Peter Pantazatos

Committee in charge:

Professor Virgil L. Woods Jr., Chair
Professor Philip Bourne
Professor Jeff Esko
Professor Mortin Printz
Professor Francisco Villareal

2006
Copyright

Dennis Peter Pantazatos, 2006

All rights reserved.
The dissertation of Dennis Peter Pantazatos is approved, and is acceptable in quality and form for publication on microfilm:

__________________________________________________________________

__________________________________________________________________

__________________________________________________________________

__________________________________________________________________

__________________________________________________________________

__________________________________________________________________

Chair

University of California, San Diego

2006
DEDICATION

This dissertation is dedicated in loving memory of my father

whose early inspiration led me along my life's path, my brother,

and for my loving mother whose strength and courage have taught me how to endure.
EPIGRAPH

"It's not the accomplishments in life that help you grow

but the experience of the journey."

Dennis Peter Pantazatos
TABLE OF CONTENTS

DEDICATION........................................................................................................................................IV
EPIGRAPH.............................................................................................................................................. V
TABLE OF CONTENTS.......................................................................................................................... VI
LIST OF FIGURES AND TABLES .........................................................................................................XI
ACKNOWLEDGEMENTS..................................................................................................................... XV
CURRICULUM VITA............................................................................................................................... XVII
ABSTRACT OF THE DISSERTATION................................................................................................. XIX
CHAPTER 1...........................................................................................................................................1

HYDROGEN/DEUTERIUM EXCHANGE MASS SPECTROMETRY FOR
INVESTIGATING PROTEIN-LIGAND INTERACTIONS

1.1 ABSTRACT........................................................................................................................................1
1.2 INTRODUCTION.............................................................................................................................. 2
1.3 STANDARD APPROACHES FOR HIGH-RESOLUTION PROTEIN
STRUCTURE ANALYSIS...................................................................................................................... 3
1.4 THEORY OF HYDROGEN EXCHANGE............................................................................................ 4
1.5 AMIDE HYDROGEN/DEUTERIUM EXCHANGE STUDIES.............................................................. 5
1.6 EX1 AND EX2 KINETICS FOR HYDROGEN/DEUTERIUM EXCHANGE ............................... 8
1.7 HYDROGEN/DEUTERIUM EXCHANGE MASS SPECTROMETRY ......................................... 9
1.8 INSTRUMENTATION AND DATA ANALYSIS............................................................................ 10
CHAPTER 5

ON THE USE OF DXMS TO PRODUCE MORE CRYSTALLIZABLE PROTEINS – STRUCTURES OF THE THERMOTOGA PROTEINS TM0160 AND TM1171

5.1 ABSTRACT

5.2 INTRODUCTION

5.3 METHODS

5.4 RESULTS

5.5 DISCUSSION

5.6 ACKNOWLEDGEMENTS

5.7 REFERENCES

CHAPTER 6

METHODS FOR THE DETERMINATION OF PROTEIN THREE-DIMENSIONAL STRUCTURE EMPLOYING HYDROGEN EXCHANGE ANALYSIS TO REFINE COMPUTATIONAL STRUCTURE PREDICTION

6.1 ABSTRACT

6.2 INTRODUCTION

6.3 RESULTS

6.4 ACKNOWLEDGMENTS

6.5 REFERENCES

CHAPTER 7

SUMMARY AND FUTURE DIRECTIONS
7.1 REFERENCES................................................................................................................. 248
LIST OF FIGURES AND TABLES

Figure 1-1: Schematic representation of hydrogen/deuterium exchange mass spectrometry procedure.................................................................20

Figure 1-2: Flow chart depicting automated protein processing and data analysis.....21

Figure 1-3: Refinement of mass spectrometry data for peptide identification ..........22

Figure 1-4(A): Thrombin-Lepirudin Complex.......................................................23

Figure 1-4(B): Thrombin-Lepirudin Complex.......................................................24-25

Figure 1-5: Ribbon diagram of thrombin-hirudin complex.................................26

Figure 2-1: Spectrin Fragmentation map: Comparison of the fragments generated by pepsin and pepsin plus FPXIII is shown........................................49

Figure 2-2: Deuterium accumulation plots for pepsin generated peptides of spectrin. 50

Figure 2-3: Composite spectrin fragmentation map, manual rate map, and High Resolution exchange rate map.........................................................51

Figure 2-4 Overlay of the COREX-calculated and experimentally-determined protection factor profiles deduced for α-spectrin R1617.................................52

Figure 2-5: Mapping of exchange rate classes on to the 3-D crystal structure of the alpha spectrin R1617 construct:..............................................53

Figure 2-6: the definition of the AU for the first 15 amino acid segment of R1617....71

Figure 2-7: HR-DXMS deconvolution of Spectrin simulated data: .........................72
Figure 2-8: Validation of HR-DXMS by producing simulated DXMS deuterated fragment datasets based on published NMR-determined experimental hydrogen exchange rate data from horse cytochrome c.................................................................73

Figure 3-1: Determination of dissociation constant (Kd) and binding stoichiometry for doxycycline-matrilysin complex by equilibrium dialysis........................................95

Figure 3-2: Doxycycline inhibits degradation of native collagen by matrilysin.........96

Figure 3-3: Amide hydrogen/deuterium exchange maps for matrilysin................97

Figure 3-4: Time-dependent deuterium incorporation of matrilysin-doxycycline complex.........................................................................................................................99

Figure 3-5A: Deuterium incorporation for peptides at the amino and carboxyl termini of Matrilysin in relation to the three-dimensional structure.............................100

Figure 3-5B: Deuterium incorporation for peptides at the zinc-binding regions of matrilysin (residues 156-175, Zn\textsubscript{struct}; residues 210-230, Zn\textsubscript{cat}).................................101

Figure 3-5C: Deuterium incorporation for peptides in matrilysin at the putative doxycycline-binding regions (residues 145-153; residues 193-204)..........................102

Figure 3-6: Probing conformation changes in matrilysin induced by doxycycline...103

Figure 3-7: Schematic depiction of matrilysin inhibition by doxycycline..............104

Figure 4-1: 10-second deuteration results for 21 proteins analyzed by DXMS........131

Figure 4-2A: 10-second amide hydrogen/ deuterium exchange map for TM0449....132

Figure 4-3: The on-exchange map of TM0505 ......................................................133

Figure 4-4: The on-exchange maps of TMTM1171 and TM0160 ..........................134

Table 4-1: Description of T. maritima proteins studied, as classified by crystallization history ...................................................................................................................135-136
Figure 4-5: Summary of DXMS analysis of 21 T. Maritima proteins. ..........140-143
Table 4-2: Summary of Data Collection and Refinement Statistics for TM0160 and
       TM1171 .......................................................................................................................... 172
Figure 5-1A: DXMS time exchange data and sequence alignments for TM0160 and
       TM1171. .......................................................................................................................... 174
Figure 5-1B: Sequence Alignment of TM1060 and its closest homologues............ 175
Figure 5-1C: Structural alignment of TM1171 and its two closest structural
       homologues ...................................................................................................................... 177
Figure 5-2A,C: Structure of TM0160...............................................................179-180
Figure 5-2B,D: Structure of TM0160...............................................................181-182
Figure 5-3: Putative active site region for TM0160 ........................................... 183-184
Figure 5-4: Overall Structure of TM1171. .................................................................185-186
Figure 5-5: Comparison of TM1171 with E. coli transcription regulator. R .......... 187
Figure 5-6: Superposition of TM1171 cNTP domain with its counterpart in E. coli.
       (PDB code 1RUN [PDB]). .......................................................................................... 188
Figure 6-1A: Deuterium on exchange: ................................................................. 222
Figure 6-1B: Deuterium localization and quantification:................................. 223
Figure 6-2: an example of validation studies of HR-DXMS de-convolution HR-DXMS
       deconvolution of Cytochrome-c .................................................................224
Figure 6-3: high-density fragmentation DXMS data on a 221 aa two-tandem- repeat
       construct of chicken brain α-spectrin (16th-17th repeats).........................225
Figure 6-4A: Structure predictions for CASP4 target protein #102...............226
Figure 6-4B: Structure predictions for CASP4 target protein #102 .................227
Figure 6-5: analysis of the relationship between variance in protein 3-D structure prediction accuracy and variance in COREX- calculated exchange-rate fingerprints for prediction vs. true structures ................................................................. 228

Figure 6-6: Ten-second amide hydrogen/deuterium exchange map for one of the proteins studied, TM1079. .................................................................................................................. 229

Figure 6-7: Deuteration results for all of the 21 proteins that were analyzed using DXMS for surface exposed residues. ................................................................. 230

Figure 7-1: Table listing ongoing structural genomics initiatives ......................... 247
ACKNOWLEDGEMENTS

First, I would like to thank my family for their tremendous support and love, Virgil Woods for creating an environment of flexibility and independence as a graduate student. I thank my entire thesis committee and the BMS faculty and staff for their continued support throughout my graduate years. Finally I would like to thank Angelo Gaitos and Angela Pantazatos for their assistance in the formatting of this dissertation. And many thanks to Iva Smith for her review and editing of the dissertation. Also many thanks to all the members of the Woods lab: Jack Kim, who ran many of the samples, Chris Gessner, who was invaluable in managing the informatics, data storage and assisting in the data analysis and Krissi Hewitt who also assisted me in data analysis. Further I would like to thank Phil Bourne for his encouraging support and development of my ideas, and James Torrance who aided in the DXMS-COREX simulation studies reported in Chapter VI. I would also like to acknowledge Vince Hilser and his lab (UTMB) for assisting me on the use of the COREX algorithm and David Baker and lab (University of Washington) for their assistance and advice on using Rosetta.

1990  BS Biomedical Engineering, Northwestern University Evanston, IL
1996  MS Human Genetics, University of Michigan Ann Arbor, MI
2006  Ph.D., Biomedical Sciences University of California San Diego, CA

PUBLICATIONS


Abstracts:


Three dimensional structure determination and analysis of proteins is necessary for the understanding of how proteins participate in human disease, and are critical for the effective design of therapeutics for clinically important targets. Current efforts for determining protein structures are centered on novel high-throughput (HT) approaches. These include high throughput (HT) crystallization efforts and global structure prediction efforts monitored through the Critical Assessment of Structure Prediction
(CASP) experiments where progress has been incremental at best. Protein structure analysis of conformational changes and protein-protein interactions can be monitored by biophysical methods which include fluorescence spectroscopy, differential scanning calorimetry, circular dichroism and ultra centrifugation. These methods provide adequate low resolution information on global changes in secondary and tertiary structure but are limited in providing detailed information on protein structure, protein conformational changes and protein-protein interactions. Therefore, there is a great need for improvements in the speed and ease of determining and analyzing protein structures and protein dynamics. Hydrogen/Deuterium (H/D) exchange rates are highly dependent on protein structure and amide hydrogen solvent accessibility. Exchange rates can report structure stability at the individual amino acid scale and provide important information on the secondary and tertiary structure.

The dissertation is arranged as follows:

**Chapter 1** is an introduction to Hydrogen/Deuterium exchange mass spectrometry and also reports my studies on the thrombin-Lepirudin complex.

**Chapter 2** is in preparation for submission and reports the application of DXMS for characterizing the molecular dynamics of spectrin. It also presents the development and validation studies for a computational method for generating amide exchange rate maps from DXMS data, a critical component of the structure determination method described in Chapters six and seven.

**Chapter 3** reports the application of DXMS for structural analysis of drug-protein interactions.
Chapter 4 reports methods for using DXMS to improve the crystallizability of protein constructs for 3D structure determination by x ray crystallography.

Chapter 5 reports the detailed 3-D structures of the first two proteins that were successfully studied with the DXMS-guided construct design method.


Chapter 7 summarizes my conclusions from the foregoing studies and outlines future directions of these studies.
CHAPTER 1

HYDROGEN/DEUTERIUM EXCHANGE MASS SPECTROMETRY FOR INVESTIGATING PROTEIN-LIGAND INTERACTIONS

1.1 ABSTRACT

Amide hydrogen/deuterium exchange mass spectrometry is rapidly becoming a powerful method for high-resolution analyses of protein dynamics, structure, and function. Hydrogen/deuterium exchange approaches can provide information that greatly augments and refines information derived from high-resolution structural studies, and can provide detailed information on native protein structure when structural information is unavailable. Application of this method for rapid analyses of protein-ligand complexes could prove useful for studies of important disease-related protein complexes. The following review covers fundamentals of hydrogen/deuterium exchange and its applications to the study of protein-ligand complexes. In addition, hydrogen/deuterium exchange mass spectrometry studies on a protein-inhibitor complex are presented.
1.2 INTRODUCTION

Changes in protein tertiary structure, protein dynamics (i.e. movement), and association state are critical for the proper function of many protein systems. Anomalous alterations in these properties can affect normal protein function, and in some cases compromise cell viability. Detailed structural investigations of these protein systems can therefore aid in understanding the causal relationships between protein structure and normal or abnormal physiological function.

In this era of proteomics and molecular medicine, mass spectrometry has become a powerful platform for profiling proteins in disease. High accuracy measurements of protein mass (via mass-to-charge ratio), rapid turnover of experiments, instrument automation, and ready access to protein sequence databases have made mass spectrometry one of the technologies of choice for the systematic study of proteins. More recent applications of mass spectrometry to characterize protein structural changes have shown great promise. One such approach utilizes isotopic labeling of proteins via amide hydrogen/deuterium exchange followed by proteolytic fragmentation of labeled protein and analysis by mass spectrometry. Hydrogen/deuterium exchange mass spectrometry has been used to track structural changes in proteins involved in processes such as viral infection\textsuperscript{1,2}, blood coagulation\textsuperscript{3-5} and kinase-mediated signal transduction\textsuperscript{6-11}. Specific application of hydrogen/deuterium exchange to study key disease-related proteins could aid in understanding the mechanisms by which these proteins function during disease progression.
The use of amide hydrogen/deuterium exchange to study protein structure, dynamics, protein-protein interfaces, and small-molecule ligand-binding sites is the subject of this review. What follows is an overview of the theory of amide hydrogen exchange, experiment methodology, and potential applications of hydrogen/deuterium exchange in drug development and proteomics research. In addition, model studies to look at protein-ligand interactions with the serine protease thrombin and a potent thrombin inhibitor, lepirudin, are presented.

1.3 STANDARD APPROACHES FOR HIGH-RESOLUTION PROTEIN STRUCTURE ANALYSIS

Nuclear magnetic resonance spectroscopy (NMR) and X-ray crystallography remain the standards to which all other protein structural methods must be compared. NMR spectroscopy provides detailed information on structural, thermodynamic, and kinetic properties of proteins. NMR is particularly well suited for molecular characterization of three-dimensional protein structure and protein dynamics in physiological-like solutions. Advancements in NMR have made structural analyses of proteins in the 30 - 40 kDa range more routine \(^{12}\). Larger proteins, however, are not routine. Molecular mass (> ~40 kDa) therefore represents the major limitation of NMR. X-ray crystallographic methods also provide high-resolution structural information of proteins. Although the structures are static, many different physiologically-relevant states of a protein may be resolved using the appropriate conditions. One major limitation of protein crystallography is the crystallization process itself; some proteins are
simply not amenable to crystallization. Both methods require milligram amounts of concentrated protein (mM) and the throughput of data analysis is slow. By contrast, mass spectrometry methods require relatively little protein (micrograms at most), mass measurements are exact, and information is obtained almost instantaneously. Highly detailed information of micro-scale solvent exposure for small groups of backbone amides can provide a detailed macro-scale view of protein conformation. Structural studies using mass spectrometry coupled with hydrogen/deuterium exchange can be carried out in a number of physiologically-relevant contexts including those that mediate ligand binding, self-association, and conformational switching. Advancements in other techniques such as Raman spectroscopy also hold promise for use in high-resolution high-throughput protein structure and dynamics studies.

1.4 THEORY OF HYDROGEN EXCHANGE

Exchange of protons between a protein and the surrounding aqueous solvent occurs as a spontaneous chemical process. The intrinsic rate of exchange for a particular proton depends on several factors including, but not limited to, the degree of solvent exposure, local inductive effects caused by adjacent amino acids, temperature, pH, and the concentration of the exchange catalyst (–OH, H₃O⁺, acidic or basic electrolytes). Proteins contain a variety of exchangeable protons. Fast exchanging protons can be found on a number of functional groups located on protein side chains (–OH, -SH, NH₂, -COOH) that exchange too rapidly to be measurable by isotope exchange methods. Protons bound directly to carbon have high covalent character such that hydrogen
exchange is unlikely to occur without a catalyst. Protons found on backbone amide groups of proteins exchange hydrogen with water at rates ranging from milliseconds to many years \(^{15}\); these are measurable by isotope exchange. The variation in exchange rates reflects the diversity of local environments for individual amide hydrogens. Solvent-exposed amide hydrogens will readily exchange protons with water, while those excluded from solvent are less likely to exchange protons. In a folded protein, amide hydrogens on the protein surface or within unstructured regions exchange within several seconds, while those buried within the hydrophobic core or those involved in hydrogen bonding will not exchange unless changes in structure expose them to solvent and hydrogen bonding is perturbed, respectively \(^{16}\). Thus, the propensity of hydrogen to exchange provides information on the conformational properties of a folded protein.

1.5 AMIDE HYDROGEN/DEUTERIUM EXCHANGE STUDIES

Isotopic exchange of hydrogen has been used to study peptides and proteins since its conception by Kaj Linderstrom-Lang in the 1950’s \(^{17-19}\). In the 1960’s, Walter Englander demonstrated that hydrogen exchange with tritium could be used to characterize different “kinetic classes” of exchangeable hydrogens on ribonuclease, and thus gained insight into the unique structural elements of the protein \(^{20}\). Since that time, hydrogen/deuterium exchange has been coupled with spectroscopic methods such as NMR \(^{21}\), resonance Raman \(^{22}\), and mass spectrometry \(^{23}\) to study protein structure. Weakly acidic peptide amide hydrogens (Peptide-H; H = Hydrogen) that are exposed to solvent and not hydrogen bonded readily undergo chemical exchange with deuterated water (D-OH; D = deuterium), as shown below (Eq. 1).
In an unfolded polar polypeptide, deuterium exchange rates are influenced by the flanking amino acid sequence and solvent conditions, and rates can vary from 10 to 1000 msec\textsuperscript{24,25}. However, in a more complex system, such as a large folded protein with multiple structural domains, the degree of amide hydrogen/deuterium exchange can vary drastically. To complicate matters even further, protein motion (thermally induced protein “breathing” and localized unfolding) can alter structure to briefly expose regions of the protein that were previously inaccessible to solvent. Hydrogen/deuterium exchange in folded proteins can be described by the model shown below (Eq. 2),

\[
\text{Protein-H (closed)} \rightleftharpoons \text{Protein-H (open)} \rightleftharpoons \text{Protein-D (open)} \rightleftharpoons \text{Protein-D (closed)} \quad \text{(2)}
\]

where protein in the closed state contains an inaccessible amide hydrogen and protein in the open state contains an accessible exchange-competent amide hydrogen. For this equation, \(k_1\) is the rate constant for opening and \(k_{-1}\) for closing, and \(k_{\text{int}}\) is the intrinsic rate constant for hydrogen/deuterium exchange in the open state, which is dependent on pH, temperature, and primary amino acid sequence\textsuperscript{26,27}.

Stable proteins in the native state have a higher propensity for the closed form than for the open form, and as such, \(k_{-1} >> k_1\). The observed hydrogen/deuterium exchange rate (\(k_{\text{ex}}\)) can therefore be represented as

\[
k_{\text{ex}} = \frac{k_1 k_{\text{int}}}{(k_{-1} + k_{\text{int}})} \quad \text{(3)}
\]
Moreover, since \( k_1 \gg k_{\text{int}} \), the observed hydrogen/deuterium exchange rate \( k_{\text{ex}} \), can be reduced to

\[
k_{\text{ex}} = K_1 k_{\text{int}}
\]

where \( K_1 \) is the equilibrium constant for the opening reaction in Equation 2. In other words, the observed hydrogen/deuterium exchange rate \( k_{\text{ex}} \) of a unique proton is determined by its intrinsic exchange rate \( k_{\text{int}} \) multiplied by the equilibrium constant for the opening reaction \( K_1 \).

Under conditions that obey EX2 kinetics\(^{26}\) (described below), the Gibbs free energy change for the opening reaction (\( \Delta G^\circ_{\text{open}} \)) can be calculated using the equilibrium constant \( K_1 \), as shown in Equation 5

\[
\Delta G^\circ_{\text{open}} = -RT \ln(K_1)
\]

where \( R \) is the gas constant and \( T \) is the absolute temperature. It is important to reiterate that measurements of \( \Delta G^\circ_{\text{open}} \) here are only valid under EX2 conditions.
1.6 EX1 AND EX2 KINETICS FOR HYDROGEN/DEUTERIUM EXCHANGE

Hydrogen/deuterium exchange of stably folded proteins (where $k_{-1} >> k_1$) can display two distinct types of exchange kinetics, monomolecular exchange, EX1, and bimolecular exchange, EX2. When the closed form of the protein predominates, that is $k_{-1} >> k_1$, and the closing rate ($k_{-1}$) is much slower than the intrinsic rate of hydrogen/deuterium exchange ($k_{\text{int}}$), the observed exchange kinetics are termed EX1. Under EX1 conditions, the hydrogen/deuterium exchange rate, $k_{\text{ex}}$, equals the opening rate $k_1$. Regions of proteins exhibiting EX1 kinetics can exchange all amide hydrogens during one unfolding event. Thus, regions of proteins undergoing slow folding and refolding may display amide EX1 exchange kinetics after a short duration of deuterium exposure. In the case where the closed form of the protein predominates, that is $k_{-1} >> k_1$, and the closing rate ($k_{-1}$) is much faster than the intrinsic rate of hydrogen/deuterium exchange ($k_{\text{int}}$), the observed exchange kinetics are termed EX2. This means that under EX2 exchange conditions, the opening-closing process may occur many times before one proton is exchanged. Hydrogen/deuterium exchange in folded proteins usually displays EX2 kinetics. However, it is likely that both mechanisms operate simultaneously in a protein having regions that are in the native state and those undergoing slow local unfolding and refolding. A more extensive review of hydrogen exchange kinetics is described by Clarke and Itzhaki.

---

$\text{8}$
1.7 HYDROGEN/DEUTERIUM EXCHANGE MASS SPECTROMETRY

A general procedure for hydrogen/deuterium exchange analysis is shown in Figure 1. The experiment can be divided into four parts: 1) deuterium on-exchange; 2) denaturation and fragmentation; 3) mass spectrometry; and 4) peptide identification and mapping. On-exchange of deuterium is performed under native conditions in D2O buffer. The effects of ligand binding on structure are probed during this step. Ligand binding is predicted to mask specific regions from deuterium exchange by steric hindrance at the binding site, and/or by changes in structure that limit solvent exposure (Fig. 1). The exchange reaction is stopped by simultaneous lowering of temperature (~0ºC) and by addition of a “quench” solution that denatures the protein and reduces the pH (or pD) to 2-3. This minimizes back exchange of deuterium with hydrogen: amide hydrogen exchange is slowest between pH 2-3 for an unfolded polypeptide.29 The denatured protein is proteolyzed under conditions of low pH and temperature. Fragments typically ranging from 10-20 amino acids31 are separated by reverse-phase chromatography to minimize mass overlap. Successive protease treatments can be used to generate smaller fragments for mass spectrometric analysis31. Identification of proteolyzed peptide fragments is carried by mass spectrometry and compared to the primary amino acid sequence of the protein. With a properly calibrated instrument, error in mass measurements for a peptide is between 0.02-0.2 kDa.13 Absolute levels of deuterium incorporation are determined for each peptide fragment by mass spectrometry. Rapidly exchanging sites will show a greater shift in molecular weight due to deuterium
incorporation. Parallel experiments in the absence of deuterium provide a reference for comparisons of deuteration level. With reliable identification of peptides and assessments of deuterium exchange for all fragments, a high-resolution map can be pieced together depicting regional levels of deuterium incorporation (Fig. 1).

1.8 INSTRUMENTATION AND DATA ANALYSIS

For our structure analyses, automated processing of deuterated protein begins with placement of samples into a cryogenic autosampler (Spectraphysics AS3000 Autosampler) under external computer control. Frozen samples (-45°C) are lifted from their wells with a robotic arm, rapidly melted to slightly above 0°C, and loaded onto an injector loop for protein fragmentation and peptide separation by liquid chromatography (Fig. 2). Protein fragmentation typically involves proteolytic processing through two protease columns arranged in series (1: pepsin protease; 2: fungal protease; Fig. 2). Fragmented-protein effluent from the protease columns is fractionated by reverse-phase separation to prevent peak overlap during mass spectrometry. A T-flow configuration divides effluent from reverse-phase chromatography for direct tandem mass spectrometric analyses using an electrospray Micromass Q-TOF mass spectrometer and a Thermo Finnigan LCQ electrospray ion trap mass spectrometer. For identification of potentially ambiguous fragments having similar molecular weights, daughter ion scanning (i.e., MS/MS) on the LCQ spectrometer is employed. Data analyses and preliminary peptide identifications using data acquired on both spectrometers are performed with the SEQUEST software program (Thermo Finnigan Inc.) Tentative
identifications are tested with custom-designed deuterium exchange data reduction software developed in collaboration with Sierra Analytics LLC (Modesto CA). This software searches mass spectral data for scans of each of peptide, selects scans with optimal signal-to-noise, averages the selected scans, calculates centroids of isotopic envelopes, screens for peptide misidentification by comparing calculated and known centroids, then facilitates visual review of each averaged isotopic envelope assessing "quality" (yield, signal/noise, resolution, peptide identity and calculated centroids). An example of data processing with this software is shown in Figure 3. Correct first-round assignment of peptide charge state by SEQUEST software is shown in Figure 3A. In some cases, the initial analysis by SEQUEST results in an incorrect assignment of charge state and consequently requires reanalysis (Fig. 3B). A second round of data analysis using deuterium exchange data reduction software allows correct identification of peptides (Fig. 3C), thus ensuring that the pool of peptides used for structural analyses is of high quality.

1.9 THROMBIN-LEPIRUDIN COMPLEX: MODEL STUDIES USING HYDROGEN/DEUTERIUM EXCHANGE MASS SPECTROMETRY

The serine protease thrombin plays an essential role in blood coagulation by proteolytic activation of several blood-clotting proteins and by activation of circulating platelets. Thrombin has therefore garnered much attention as a therapeutic target for anti-coagulation drug therapy. The small protein hirudin (65 amino acids) is the most potent natural inhibitor of thrombin, and as such, an effective anticoagulant. Hydrogen/deuterium exchange has been employed to characterize the binding
interactions between thrombin and lepirudin, a hirudin derivative displaying 98% sequence homology with the structurally characterized hirudin\textsuperscript{35-37}. Data obtained by deuterium exchange not only confirm known regions of contact determined from a previous crystal structure of the thrombin-hirudin complex\textsuperscript{36}, but also demonstrate additional regions of conformational change upon binding of lepirudin that could not be detected by crystallography.

### 1.10 MAPPING LEPIRUDIN BINDING SITES ON THROMBIN

Amide hydrogen/deuterium exchange was used to map the binding regions of the thrombin inhibitor lepirudin. Using the procedures outlined in Figures 1-1 and 1-2, approximately 94% of the thrombin sequence was covered with multiple overlapping peptides for precise localization of deuterium incorporation, as shown in Fig. 1-4A. Proteolytic processing resulted in two small gaps in sequence coverage at the amino terminus (stretch of twelve amino acids) and at a single glycosylation site (within stretch of sixteen amino acids) of thrombin. Two peptide fragmentation maps for thrombin are shown in Figure 1-4A. The top fragmentation map corresponds to the unliganded apo form of thrombin, and the bottom map corresponds to the lepirduin-bound form. Placement of deuterium on specific amides was done as described in Chapter 4, deuterium labeling was manually assigned to residue positions within the protein by first optimizing consensus in deuterium content of overlapping peptide probes, followed by further clustering of labeled amides together in the center of unresolved regions, so that a consensus map was generated. Five major regions of hydrogen/deuterium exchange are
apparent in the peptide maps (denoted by circles), each having at least six consecutive deuterated amides. The green-circled regions represent the anion-binding exosite (circle 2) and the catalytic active site (circles 4 and 5) and the blue-circled regions unstructured loops (circles 1 and 3)\textsuperscript{36}. For unliganded thrombin, amino acid regions circled in green and blue show extensive deuterium incorporation. By contrast, lepirudin binding to thrombin hinders deuterium exchange at the exosite (circle 2) and at the active site regions (circles 4 and 5; Fig.1-4A, bottom map). Amides within these protected regions that do not exchange deuterium are boxed in the lower map. Protected amides in the green-circled regions of the bottom map that do not exchange deuterium are consistent with known hirudin-thrombin contact regions identified in the crystal structure of the complex\textsuperscript{36}. Contact residues from the hirudin-thrombin crystal structure are shown as gray boxes above the fragmentation maps. Two additional regions of fast deuterium exchange apparent in the top fragmentation map (blue circles 1 and 3) remain highly deuterated after lepirudin binding. Lack of change in deuteration level indicates that these regions are not in contact with lepirudin after complex formation.

1.11 ALLOSTERIC CHANGES IN THROMBIN STRUCTURE INDUCED BY LEPIRUDIN BINDING

Deuterium exchange rates were plotted for all peptides (386 generated for thrombin) and analyzed for regions of contact (discussed above) and allosteric changes. Relative rates of amide exchange for selected regions of lepirudin-bound thrombin are shown in Figure 1-(4B). The top graph shows the percentage of deuterium exchange for
a select peptide (residues 109-123) within the thrombin exosite, which resides in a known binding region for hirudin\textsuperscript{36}. Deuterium exchange within the hirudin-contact region decreased 50% upon lepirudin binding at the fastest on-exchange time point (10sec). Later time points revealed constant magnitudes of deuterium exchange decreases (50% decrease relative to apo thrombin), indicating that the structure in this region is stably hindered from deuterium exchange over time. This type of exchange profile is associated with ligand binding\textsuperscript{6,7,10,38}. For some peptides, there is little difference between apo and ligand-bound deuterium-exchange rates, which suggest that structure at those locations is essentially unaffected by ligand binding. Deuterium exchange rates for a structurally unaffected peptide (residues 146-151) are shown (Fig. 1-4B, middle graph). The bottom graph in Figure 1-4B shows deuterium exchange profiles for a peptide (residues 210-226) undergoing allosteric changes in structure. This peptide exhibits slowed deuterium-exchange when lepirudin is bound to thrombin. Residues 210-226 correspond to a loop region within the heavy chain of the structure exclusive of the hirudin contact regions\textsuperscript{36}. Slowed amide exchange within this region is indicative of decreased dynamics of the loop region, rather than site-specific ligand binding. Levels of deuterium incorporation for this peptide are closely matched at the earliest time point (10 sec.) for apo and ligand-bound thrombin but steadily diverge reaching a difference of approximately 40% in deuterium incorporation at the latest time point (Fig 1-4B, bottom graph). These data suggest that lepirudin binding triggers allosteric changes in structure that reduce the dynamics of the loop containing residues 210-226, which is located outside of the ligand-binding site. Such allosteric changes can alter the opening-closing equilibrium that modulates hydrogen exchange with solvent. Similar deuterium-exchange profiles have
been observed for cyclic AMP-dependent protein kinase A undergoing ligand-induced conformational changes\textsuperscript{6,7}. The ribbon diagrams in Figure 5 show comparisons of structural information from X-ray crystallography (hirudin-thrombin) and deuterium exchange mass spectrometry (lepirudin-thrombin) mapped onto the hirudin-thrombin structure. Regions of hirudin contact determined by X-ray crystallography at the catalytic active site and exosite (Fig. 1-5A, colored pink) show extensive overlap with those obtained by deuterium exchange methods for the lepirudin-thrombin complex (Fig. 1-5B, colored red), indicating that hirudin and lepirudin interact with thrombin at identical contact sites. This observation coincides with NMR structural data on lepirudin that revealed structural homology between the two hirudin forms\textsuperscript{37}. In addition, regions showing variable levels of amide hydrogen exchange outside of the ligand-contact regions are shown (Fig. 5B), and are attributed to allosteric effects mediated by lepirudin.

\textbf{1.12 RESOLVING CONFORMATIONAL CHANGES AND LIGAND BINDING}

Changes in hydrogen/deuterium exchange rates can take place following ligand binding by steric hindrance and by ligand-induced allosteric changes in structure\textsuperscript{39-41}. Reliable correlations between local decreases in hydrogen/deuterium exchange and ligand binding are reasonable when structural information of ligand binding sites is available\textsuperscript{11,42}. However, interpretation of hydrogen/deuterium exchange data is more complex in the absence of three-dimensional structures. One strategy for discriminating between regions of protein-ligand interactions from those due to allosteric changes in structure using hydrogen exchange mass spectrometry is described.
Under conditions that promote ligand binding, contact regions on the receptor protein are predicted to be accessible to ligand to facilitate complex formation. Random deuterium labeling of all solvent exposed amides will therefore label regions involved in ligand binding. In the absence of ligand, protein is subjected to deuterium exchange with deuterated solvent. Ligand-binding residues located on the surface of the protein that are in contact with deuterated solvent will undergo rapid hydrogen/deuterium exchange (within milliseconds). Labeled protein is then complexed with ligand at ligand concentrations that shift the equilibrium to ~100% complex. Ligand-bound protein is diluted in non-isotopic solvent for deuterium off-exchange under conditions that prevent protein-ligand dissociation (i.e. in presence of ligand in undeuterated buffer). Bound ligand will trap labeled deuterium within the contact interface, while solvent exposed deuterium will off-exchange with solvent. Therefore, proteolyzed fragments that show extensive incorporation of deuterium are predicted to be at the ligand-binding interface. For assessment of conformational changes, peptides that show variable rates of deuterium exchange (faster or slower relative to unliganded control) represent regions were structural changes can be inferred. Studies on human α-thrombin complexed with thrombomodulin (a complex of unknown structure) by Mandell and coworkers using hydrogen/deuterium exchange and MALDI mass spectrometry have proven the effectiveness of this approach. Residues involved in the thrombin-thrombomodulin interface were identified and distinguished from regions of conformational changes due to allosteric effects. In the same report, regions of a kinase inhibitor and ATP-binding sites on protein kinase A were also determined by deuterium exchange methods. However, it must be noted that data interpretation can be ambiguous when the exchange
rates for ligand binding resemble those due to conformational switching. In these situations, complementary biophysical approaches should be implemented.

**1.13 HYDROGEN/DEUTERIUM EXCHANGE TO CHARACTERIZE DRUG-PROTEIN INTERACTIONS**

Structural information on drug-binding sites, drug-induced allosteric changes, and alterations in protein dynamics can provide key insights into the structural and chemical nature of the drug-protein interaction. Efforts are already underway to implement proteome-scale crystallography methods for use in high-throughput structural analyses. Creation of large-scale structural libraries (i.e., structural informatics) from these efforts could provide active-site geometries for rational drug design. Recently, Pantazatos and co-workers demonstrated that hydrogen/deuterium exchange mass spectrometry can be used as a reliable high-throughput platform to profile disordered regions of proteins: high-resolution information on sequence and structure dynamics for twenty-one proteins were amassed and analyzed within two weeks. Data from these studies were used for post hoc refinements of protein constructs for improved protein crystallography; two protein constructs were successfully crystallized following refinement. A similar approach can be employed in targeted drug development where small-molecule or protein-based drugs can be structurally modified to improve their ability to bind within the active site(s) of target proteins, and/or induce inactive-state protein conformations. Relationships between native structure and physiological function, and the inhibitory actions of specific drug candidates can be characterized.
Such data could be used to infer molecular modes of drug inhibition. In addition, examination of structural changes induced by drug binding among protein isoforms could lead to elucidation of unique drug-binding structural motifs.

1.14 CONCLUSION

The union of classic hydrogen/deuterium exchange methods with modern mass spectrometry has resulted in a powerful platform for high-resolution studies of protein structure, ligand binding, and protein dynamics in solution. As an independent technology for structural analyses, the method is impressive in its ability to generate unambiguous information on microenvironment for individual amides, or small groups of amides. Mapping of intrinsic exchange rates of individual amides can be viewed as a “fingerprint” that relates to local environment, and ultimately structure. In addition, other protein systems inherently difficult to analyze by NMR and crystallography, such as membrane proteins, can be analyzed by hydrogen/deuterium mass spectrometry. Deuterium exchange studies on the transmembrane fragment of the M2 protein of Influenza A reconstituted into lipid vesicles revealed that the weak hydrogen exchange of backbone amides in the transmembrane domain can be influenced by protein conformation and dynamics, and the properties of the surrounding lipids. A particularly novel application of hydrogen/deuterium exchange involves in situ analyses of protein structure in the cell. Such studies offer a unique opportunity to view true native protein conformation in a proper physiological setting. Proteins derived from cells cultured in deuterated medium and purified from lysates (without loss of deuterium)
could be subjected to hydrogen/deuterium exchange analyses to assess structure. In situ studies could be used to examine proteins under normal and pathological conditions for any number of diseases. Hydrogen/deuterium exchange approaches can also be used as an adjunct to other structural methods, where novel insights can be obtained on unique regions of proteins, such as unstructured loops of large proteins, that are inherently difficult to examine by crystallography (described above). The prospects for even better data accuracy and faster computational data analyses are on the horizon as instrumentation and related technologies move forward. Efforts to couple amide hydrogen/deuterium exchange mass spectrometry methods to computational protocols for structure-activity relation analyses are already underway. Hydrogen/deuterium exchange mass spectrometry should prove useful in the larger context of proteomics and structural genomics for high-throughput structure determination.

1.15 ACKNOWLEDGEMENTS

Chapter 1 is a reprint of the material as it appears in *Assay Drug Dev Technol*, 2004, Garcia, R.A., D.P. Pantazatos, and F.J. Villarreal. I was co-equal first author, and all the studies of the thrombin-hirudin binding interaction were designed, performed, and interpreted solely by me. The co-authors assisted in the writing of the manuscript. I’d like to thank Virgil Woods Jr. and Shirley Reynolds for their critical reading of the manuscript and Chris Gessner for assistance with archived data retrieval. I’d also like to thank Angela Pantazatos for assistance with fragmentation mapping.
Figure 1-1: Schematic representation of hydrogen/deuterium exchange mass spectrometry procedure.
Figure 1-2: Flow chart depicting automated protein processing and data analysis
(Figure taken from Garcia, R.A., D.P. Pantazatos et al. 2004)
Figure 1-3: Refinement of mass spectrometry data for peptide identification. (A) On-screen panel of deuterium exchange reduction software showing accurate prediction of charge state after first-round analysis by SEQUEST. The red curve is the raw mass spectral data and the blue curve is the theoretical fit. (B) Poor fit to data after first-round SEQUEST analysis. (C) Correct prediction of charge state following post hoc analysis with deuterium exchange data reduction software. (Figure taken from Garcia, R.A., D.P. Pantazatos et al. 2004)
**Figure 1-4(A): Thrombin-Lepirudin Complex.** Peptide fragmentation map of thrombin in the absence (apo) and presence of lepirudin (+Lepirudin) with regions corresponding to light and heavy chains shown. Proteins were subjected to 10 seconds of deuterium exchange followed by proteolysis and mass spectrometry. Peptides are color coded reflecting amide deuterium incorporation per amino acid, where deuterated amides are colored red and undeuterated amides are blue. Gray vertical lines indicate boundaries of fast-exchanging amides. Amides protected from deuterium exchange by lepirudin binding are boxed in lower map. Gray-colored boxes above each fragmentation map denote hirudin-thrombin contact regions determined by crystallography.
**Figure 1-4(B): Thrombin-Lepirudin Complex.** Relative rates of amide exchange for selected regions of thrombin after lepirudin binding. Top panel shows rates of deuterium exchange at peptide (residues 109-123) within the thrombin exosite, which is a known binding region for hirudin. Deuterium exchange within this hirudin-contact region decreases 50% upon lepirudin binding. Middle panel shows lack of difference in deuterium-exchange profiles for apo and lepirudin-bound thrombin at a non-contact site (residues 146-151). Bottom panel shows deuterium exchange of peptide (210-226) found in a region where allosteric changes are implicated. Exchange rates for unliganded (red line) and lepirudin-bound thrombin (blue line) are shown. (Figure taken from Garcia, R.A., D.P. Pantazatos et al. 2004)
Peptide 109-123

Peptide 146-151

Peptide 210-226

lepirudin contacts within exosite

non-contact site

lepirudin-induced allosteric changes

apo

+ lepirudin
Figure 1-5: Ribbon diagram of thrombin-hirudin complex. (A) X-ray crystal structure of the hirudin-thrombin complex showing regions of contact (colored pink). Arrows denote hirudin-thrombin catalytic active site and exosite contacts. Thrombin light chain is colored black and hirudin/lepirudin green. (B) Lepirudin-thrombin model based on crystal structure. Contacts for lepirudin-thrombin determined by hydrogen/deuterium exchange mass spectrometry (colored red). Contacts overlap with those determined by X-ray crystallography. Sites of allosteric change suggested by hydrogen/deuterium exchange are colored orange (high decrease in dynamics) and blue (moderate decrease in dynamics) based on the degree of slowing in deuterium exchange. (Figure taken from Garcia, R.A., D.P. Pantazatos et al. 2004)
1.16 REFERENCES


10. Anand GS, Hughes CA, Jones JM, Taylor SS, Komives EA. Amide H/2H exchange reveals communication between the cAMP and catalytic subunit-


CHAPTER 2

HIGH RESOLUTION DEFINITION OF THE STABILITY LANDSCAPE OF CHICKEN BRAIN $\alpha$-SPECTRIN (R1617) WITH ENHANCED HYDROGEN/DEUTERIUM EXCHANGE MASS SPECTROMETRY (DXMS): IMPLICATIONS FOR THE BASIS OF SPECTRIN ELASTICITY.

2.1 ABSTRACT

Spectrin is a cytoskeletal protein that functions as an elastic molecule with a distinctive compliance behavior, where tension remains within a relatively narrow range despite considerable lengthening. To better understand the molecular basis of this behavior, we have used enhanced methods of peptide amide hydrogen-deuterium exchange-mass spectrometry to determine the exchange-rate profile of a two repeat construct of chicken brain $\alpha$-spectrin with unprecedented resolution and comprehensiveness. The experiment revealed the relative thermodynamic stability or “energetic landscape” of the spectrin molecule at the individual amino acid level.

Remarkably, each of the six long $\alpha$-helices present in the construct was not a uniformly stable structure, but demonstrated gradients in hydrogen exchange rates, with amides in the middle 1/4 to 1/3 portions of each helix having slow exchange rates, progressively increasing to more than 1000 times faster rates towards the ends of the
helices. Comparable $\alpha$-helix stability gradients were not present in five other proteins studied.

These findings support and extend previous models of $\alpha$-spectrin behavior that propose conformational rearrangements involving helix-loop transitions, with migration of the short loop regions back and forth along the $\alpha$-spectrin helices, and indicate that the stability gradients provide the mechanism by which the molecule stores mechanical energy when stretched. These studies also demonstrate the impressive ability of these advanced hydrogen exchange techniques to deduce protein dynamics comprehensively and at high resolution.

Abbreviations: DXMS, high throughput, high resolution amide hydrogen-deuterium exchange mass spectrometry; HR-DXMS, High Resolution, residue-specific determination of amide hydrogen exchange rates from DXMS data; Gu HCl, guanidinium hydrochloride; TCEP, tris 2-carboxyethyl phosphine; TBS, Tris buffered saline.
2.2 INTRODUCTION

The cytoskeleton of blood cells includes many components necessary for maintaining membrane structural integrity and allowing the cells to withstand the stresses of traversing the circulatory system [2, 3]. The major component of this skeletal network is the elastic protein $\alpha$-spectrin, which consists of an $\alpha$-monomer of 20 tandem repeats and a $\beta$-monomer of 16 repeats. X-ray crystal structures of constructs composed of two such tandem repeats of the $\alpha$-subunit reveal that each is composed of three well-formed, long antiparallel $\alpha$-helices connected by short turns or loops, forming a “Z”-shaped three-helix bundle [1], with the tandem repeats connected by a short $\alpha$-helical “linker” region.

While spectrin plays a critical role in the reversible deformation of the membrane, the molecular basis of this behavior, particularly its dynamic aspects, are unclear. Investigation of cloned repeats using chemical and thermal denaturation, as well as atomic force microscopy (AFM), have yielded important advances in the understanding of the physical and biomechanical properties of unfolding and refolding of repeating units, as well as the function of $\alpha$-spectrin as a whole. Force-extension curves from AFM studies have demonstrated that spectrin is a highly non-linear spring with substantial relatively small peak unfolding forces (20-50pN) per repeating unit. Other studies have also indicated that these repeats unfold independently and/or in tandem [4, 5] with the presence of intermediates [6]. In AFM experiments $\alpha$-spectrin demonstrates a distinctive
“sawtoothed” compliance behavior, where tension rises only gradually, and remains within a relatively narrow range despite considerable lengthening of the molecule.

Several models, based on crystallization and/or AFM studies, have been proposed to account for α-spectrin’s elasticity, including tension-induced bending of the linker regions, tension-induced unwrapping or melting of the ends of α-helices into elongated loops; and catastrophic unfolding of triple helical bundles, in which the sawtoothed compliance observed is attributed to multiple tandem bundles sequentially popping open with increased tension.

A fourth mechanism has been proposed in which there is an end-to-end lengthening of the triple helical bundles, resulting from helix-loop transitions along the α-helices, accomplished by relatively little change in the total amount of helix in each bundle [1]. Observations presented herein support and extend this model.

Evaluation of these and other proposed mechanisms for α-spectrin elasticity would be facilitated by a detailed characterization of the thermodynamic stability or “energetic landscape” of the α-spectrin molecule. To this end, we have probed its structural stability at the individual amino acid scale employing enhanced methods of peptide amide hydrogen-deuterium exchange lc-mass spectrometry, which we term DXMS. Peptide amide hydrogens are not permanently attached to a protein, but continuously and reversibly interchange with hydrogen present in water. The chemical mechanisms of the exchange reactions are understood, and several well-defined factors can profoundly alter exchange rates [7-10]. One of the factors that determine the rate of exchange is the extent to which particular exchangeable hydrogen is exposed (accessible) to water. The exchange reaction proceeds efficiently only when particular peptide amide hydrogen is
fully exposed to solvent. Peptide amide hydrogens that are freely accessible to water exchange at their maximal possible rate with an average half-life of exchange of approximately one second at 0 °C and pH 7.0. [11, 12]. The precise rate of exchange of a particular fully-solvated amide can vary more than thirty-fold from this average rate, depending upon the identity of the two amino acids flanking the amide bond. Exact exchange rates expected for fully solvent-exposed amide hydrogens can be reliably calculated from knowledge of the temperature, pH and primary amino acid sequence involved [11, 12].

In a structured protein, most peptide amide hydrogens exchange more slowly (up to $10^9$-fold slower) than the maximal, fully solvated exchange rate, as they are not efficiently exposed to solvent water. Protein structure is not static, but best considered as an ensemble of transiently unfolded states: the native state ensemble [13-15]. Amide hydrogen exchange occurs only when a particular transient unfolding event fully exposes an amide to solvent. The ratio of exchange rates for a particular amide hydrogen in the folded vs. random coil states is referred to as the exchange protection factor and directly reflects the free energy change in the atomic environment of that particular hydrogen between unstructured and structured states of the protein. In this sense, amide hydrogens can be treated as atomic-scale sensors of highly localized free energy change throughout a protein and the magnitude of free energy change reported from each of a protein's amides in a folded vs. unfolded state is precisely equal to $-RT \ln \text{(protection factor)}$ [16]. In effect, each peptide amide's exchange rate in a folded protein (when measured) directly and precisely reports the protein's thermodynamic stability at the individual amino acid scale [10, 16].
Deuterium exchange methodologies, coupled with Liquid Chromatography Mass Spectrometry (LCMS), presently provide the most effective approach to perform hydrogen exchange studies of proteins larger than 30 kDa in size [17-28]. Building upon the pioneering work of Walter Englander and David Smith [17, 29, 30], we have developed and implemented a number of improvements to their methodologies and experimental equipment that have significantly improved throughput, comprehensiveness, and resolution. We term these collective improvements enhanced Deuterium Exchange-Mass Spectrometry (DXMS) (19-37).

In the present study, we demonstrate that DXMS can be used to obtain sufficient information on the exchange behavior of a two repeat construct (R1617) of chicken brain α-spectrin (16th-17th repeats) to allow construction of a peptide amide hydrogen exchange rate map at near single-amide resolution, from which the thermodynamic stability or “energetic landscape” of the molecule could be assessed at the individual residue level. Our results demonstrate that the long α-helices within the tandem repeats exhibit marked gradients in stability. If the “loop-migration” model is operative in α-spectrin, then these gradients provide the mechanism by which mechanical energy is stored in the stretched α-spectrin molecule.
2.3 METHODS

Preparation of R1617 of chicken brain α-spectrin. The two repeat fragment composed of repeat 16 and repeat 17 from chicken brain α-spectrin was cloned from the cDNA for chicken brain α-spectrin expressed in E. coli, BL21 DE3, and purified therefrom as previously described [1]. The sequence of the spectrin fragment was ascertained by DNA sequencing, by mass spectrometry of the purified protein, and by identification of the amino acid sequence from the crystal structure. The protein was a monomer in solution by analytical ultracentrifugation to sedimentation equilibrium.

Establishment of protein fragmentation maps. Thirty microliters of stock “exchange quench” solutions (0.8% formic acid, 0M/.8M/1.6M/3.2M/6.4M GuHCl) was added to 20ul of sample (final concentration 0.5% formic acid, 0M/.05/1.0/2M/4M M GuHCl) containing 10-15ug of protein in TBS, transferred to autosampler vials, and then frozen on dry ice within one minute after addition of quench solution. Vials with frozen samples were stored at –80 deg C until transferred to the dry ice-containing sample basin of the cryogenic autosampler module of the DXMS apparatus. Samples were individually melted at 0 deg C, then injected (45 ul) and pumped through protease columns (0.05% TFA, 250ul/min, 16 seconds exposure to protease). Proteolysis was catalyzed by immobilized pepsin (66 μl column bed volume, coupled to 20AL support from PerSeptive Biosystems at 30 mg/ml) or similarly immobilized Aspergillus satoi Fungal Protease XIII (20mg/ml, 66ul bed volume column). Protease-generated fragments were collected onto a C18 HPLC column, eluted by a linear acetonitrile gradient (5 to 45 % B
in 30 minutes; 50 μl/min; solvent A, 0.05% TFA; solvent B, 80% acetonitrile, 20% water, 0.01% TFA), and effluent directed to the mass spectrometer with data acquisition in either MS1 profile mode or data-dependent MS2 mode. Mass spectrometric analyses used a Thermo Finnigan LCQ electrospray ion trap type mass spectrometer operated with capillary temperature at 200 °C or an electrospray Micromass Q-ToF mass spectrometer, as previously described [40-46]. The Sequest software program (Thermo Finnigan Inc) was used to identify the likely sequence of the parent peptide ions. Tentative identifications were tested with specialized DXMS data reduction software developed in collaboration with Sierra Analytics, LLC, Modesto, CA. This software searches MS1 data for scans containing each of the peptides, selects scans with optimal signal-to-noise, averages the selected scans, calculates centroids of isotopic envelopes, screens for peptide misidentification by comparing calculated and known centroids, then facilitates visual review of each averaged isotopic envelope allowing an assessment of "quality" (yield, signal/noise, resolution), and confirmation or correction of peptide identity and calculated centroid [40-46].

On-exchange deuteration of proteins. After establishment of fragmentation maps amide hydrogen exchange-deuterated samples of R1617 were prepared and processed exactly as above, except that 5μl of each protein stock solution was diluted with 15μl of Deuterium Oxide (D2O), containing 5mM Tris, 150mM NaCl, pH (read) 7.0, and incubated at 22 degrees C for 3,10,30,100,300,10³,3x10³,10⁴,2.5x10⁵, 3.4x10⁵ seconds, at which time samples were supplemented with 30μl of a quench solution (0.8% formic acid, 0.8M GuHCl ) at 0 degrees C, and samples immediately frozen at –80 degrees C.
until further processed as above. Data on the deuterated sample sets was acquired in a single automated 8-hour run, and subsequent data reduction performed on the DXMS data reduction software as previously described. Corrections for loss of deuterium-label by individual fragments during DXMS analysis (after “quench”) were made through measurement of loss of deuterium from reference α-spectrin R1617 samples that had been equilibrium-exchange-deuterated under denaturing conditions, as previously described [40-46]. High Resolution, residue-specific determination of amide hydrogen exchange rates from DXMS data (HR-DXMS) was performed as described in Supplemental Material.

Equipment configuration. The equipment configuration consisted of electrically-actuated high pressure switching valves (Rheodyne), connected to two position actuators from Tar Designs Inc., Pittsburgh, as described previously [40-46]. A highly modified Spectraphysics AS3000 autosampler, partially under external PC control, employed a robotic arm to lift the desired frozen sample from the sample well, then automatically and rapidly melted and injected the sample under precise temperature control [40-42, 44-46]. The autosampler basin was further thermally insulated and all but 20 vial positions were filled with powdered dry ice sufficient to keep samples colder than -45° C for 18 hours. Four HPLC pumps (Shimadzu LC-10AD) were operated by a Shimadzu SCL-10A pump controller. One produced forward flow over the pepsin column, another backflushed the protease column after sample digestion (0.05 % aqueous TFA), and two delivered solvents to a downstream HPLC column for gradient elution (A: 0.05 % aqueous TFA; B: 80% acetonitrile, 20% water, 0.01% TFA; 1 x 50 mm C18 Vydac # 218MS5105, pH 2.3).
Valves, tubing, columns and autosampler were contained within a refrigerator at 2.8° C, with pepsin and HPLC columns immersed in melting ice. The timing and sequence of operation of the DXMS apparatus fluidics were controlled by a personal computer running an in-house written LabView-based program, interfaced to solid-state relays (digital input/output boards, National Instruments), controlling pumps, valve actuators, and MS data acquisition [40-46].

COREX Calculations of $\alpha$-spectrin R1617 amide exchange rate protection factors.

Fyrestar, operating the COREX algorithm, was installed on the Blue Horizon supercomputer at the San Diego Supercomputer Center, and run in a sparse Monte Carlo mode against the structural coordinates of the $\alpha$-spectrin R1617 construct employing a COREX window size of 8 and a sampling of 1000 states per partition. This resulted in a sampling of 8000 states of the total 900 million possible. Hydrogen exchange rates and protection factors were calculated from the stability profile as previously described [13-15, 33].
2.4 RESULTS

Increased production of overlapping peptides of α-Spectrin construct R1617. The ability to localize and quantify detailed hydrogen exchange behavior with DXMS is largely determined by the degree to which a densely overlapping set of peptides can be proteolytically generated from the deuterated protein prior to LCMS. Prior to deuterium on-exchange analysis, digestion of exchange-quenched, undeuterated R1617 was performed on samples in 0, 0.5, 1.0, 2.0, and 4M GuHCl for different lengths of time to determine optimum conditions for maximally overlapping fragmentation. At 0.5M GuHCl and 250ul/min flow rate over the pepsin column (66ul bed volume), 114 reproducibly obtained and well-behaved fragments were produced. A second, higher resolution fragmentation map was also obtained by employing these conditions, but with the addition of an *Aspergillus satoi* Fungal Protease XIII (FP XIII) column (66ul bed volume) after the pepsin column, resulting in the generation of an additional 86 peptides. A comparison of the fragments generated by pepsin and pepsin plus FPXIII is shown in Figure 2-1. Pepsin-generated fragments are indicated in blue, and novel fragments produced by pepsin plus fungal protease indicated in orange. A total of 200 fragments were obtained with the combination of the pepsin and fungal protease columns. Such extensive fragmentation and redundancy in the overlapping of peptides was essential to successful calculations of reliable exchange rates for each residue in the spectrin construct.
Once the optimal quench-compatible fragmentation conditions were established, the R1617 construct was incubated in 150mM NaCl, 5mM Tris, pH (read) 7.0 containing 75% mole-fraction deuterated water at 22 degrees C. for times varying from 3 seconds to $3.4 \times 10^5$ seconds, and then aliquots exchange-quenched by supplementation to 0.5% formic acid, 0.5M GuHCl at 0 degrees C, followed by immediate cooling to and storage at –80 degrees C. Quenched, deuterated samples were then enzymatically fragmented, and subjected to LCMS under continued quench conditions as described in Methods. The deuterium content of each of the 200 peptides that had been generated from each sample was then calculated from the LCMS data, for all on-exchange times, employing specialized data reduction software and corrections for back-exchange (loss of deuterium from peptides after institution of “quench”) as previously described (19-37).

Construction of a low-resolution exchange rate map for spectrin R1617. Plots of deuterium accumulation for each peptide vs. on-exchange time were constructed from data obtained by analysis of 114 pepsin-only generated peptides, as shown in Figure 2-2 for three representative peptides. The time axis was arbitrarily divided into three regions, (fast, medium, and slow-exchanging; Figure 2) and the number of amides on each peptide scored that acquired deuterium in the fast, medium and slow rate classes. The latter class was grouped with the very slow class unmeasured in the limited on-exchange times ($<10^5$ sec) used in this experiment; Figure 2). A map of rate-class vs. construct sequence (Figure2-3B) was then assembled from this information, employing a strategy in which the (generally smaller) peptides containing one rate class were first placed in amino acid sequence register, followed by placement of peptides with two, and then three, rate
classes, in a manner that required that placements of the three rate classes of amides in each peptide conform with the preceding placements. The resulting “α-Spectrin Consensus Rate Map” is indicated by the arrow.

This map demonstrated features that might reasonably be anticipated from the protein's structure: the short loops between the long α-helices were uniformly fast-exchanging (Figure 3B, short horizontal black bars below the consensus map) while substantial regions within the α- helices were much more slowly exchanging (the locations of the long α-helices are shown as light-blue bars below Figure 3C and in Figure 2-5). The “Consensus Map was mapped on to the 3-D crystal structure of R1617 and revealed excellent correlation between exchange rate class and secondary structure (Figure 2-5). The map further indicated that substantial gradients in exchange rates were present within each α-helix, with slowly-exchanging regions of helix gradually giving way to fast exchanging-regions. The invariant tryptophans involved in stabilizing the triple helical bundle (47) are located in the most stable regions of the helices which are known to be located in the middle. Remarkably, these gradients in rates were not restricted to the ends of the helices, but occurred across most of their lengths. For example, helix A’ was fast-exchanging at its N-terminal and C–terminal third, with its slowest exchanging region in the middle of the helix flanked by short medium exchanging regions. Interestingly, lower stability was mapped on the linker region connecting the two repeating units.
Construction of a high-resolution exchange rate map for spectrin R1617. To study these gradients in spectrin α-helix exchange rates at higher resolution, we developed a computational method for the deconvolution of aggregate, time-dependent peptide deuteration data, to specific exchange rates for individual amide hydrogens within the native protein. This method, which we term “High Resolution, residue-specific determination of amide hydrogen exchange rates from DXMS data” (HR-DXMS), employs an algorithm centered on use of a two-phase numerical technique, linear programming (LP) for an initial rate estimation followed by a nonlinear least squares fit (NLS). Essential to success of the method is the derivation and incorporation of residue-specific corrections for deuterium loss during “back-exchange” in contrast to the use of “peptide-average” loss corrections usually employed in hydrogen exchange data analysis. The method can also make use of additional hydrogen exchange data, obtained by systematically varying the duration of the usually deleterious “back-exchange”, to allow resolution of individual amide rates within protein regions not sufficiently resolved by enzymatic fragmentation alone. A detailed description of the method, the validation studies that have been performed, and examples of the use of the enabling software are presented in Supplemental Material.

Figure 2-3C shows the results of application of HR-DXMS to the data from the 200 deuterated α-spectrin R1617 construct fragments obtained by the combined action of pepsin plus FP XIII. Results are expressed as the ΔG_{exchange} (the difference in Gibbs free energy of exchange) between the folded and unfolded form of the protein, according to equation 7.
\[ \Delta G_{\text{exchange},i} = -RT \ln \left( \frac{k_{\text{ex},i}}{k_{\text{int},i}} \right) \]  

(7)

where \( k_{\text{ex},i} \) and \( k_{\text{int},i} \) are the experimental and intrinsic (random coil) exchange rates at amide i as determined from the intrinsic rates of random coil model peptides [31, 32].

To facilitate comparison with the low resolution consensus rate map in Figure 2-3B, Figure 2-3C is divided by two horizontal dashed lines that are placed at \( \Delta G_{\text{exchange}} \) values corresponding to the arbitrary rate divisions imposed in the generation of the low resolution rate map. There is considerable agreement between the results of the two methods, and the computational approach resulted in a more finely detailed and less subjective description of the exchange rate distribution within the \( \alpha \)-helices, clearly demonstrating the extensive exchange rate gradients that traverse the helices. The A’ and A” helices have gradients with a stable central region that decreases in stability towards each end, while the B’ and B” helices demonstrate more monotonic gradients with stable C-termini that gradually become less stable at the N-terminus. The tandem-repeat linker region, which is seen to be an \( \alpha \)-helix in the crystal structure, has a distinctly lower stability than the amides of the helices that immediately adjoin it, helix C’ and A”. This pattern of stability is illustrated in Figure 2-5 on the 3D crystal structure of R1617.

**Calculation of the hydrogen exchange rate map of \( \alpha \)-spectrin R1617 from its crystallographically determined structure.** We then compared our experimentally determined exchange rate map for \( \alpha \)-spectrin R1617 with purely computational estimates of hydrogen exchange rates that can be obtained with use of the COREX algorithm.
COREX (implemented in the Fyrestar software of Redstorm Scientific, Houston TX) is a computational tool that utilizes the high-resolution structure of a protein as a template to generate a large ensemble of incrementally different conformational states. COREX represents proteins as ensembles of conformations, rather than as discrete structures, and has been shown to predict amide hydrogen exchange rates with remarkable accuracy and precision when tested against available NMR-derived experimental data, suggesting that the calculated ensemble captures the general features of the actual ensemble, and thus provides a realistic physical description of proteins[13-15, 33]. COREX was run in a sparse Monte Carlo mode against the structural coordinates of the α-spectrin R1617 structure [1] and hydrogen exchange rate protection factors were calculated as described in Methods.

Figure 2-4 overlays the COREX-calculated (red line) and experimentally-determined (black line) protection factor maps deduced for α-spectrin R1617 by DXMS analysis. There is significant agreement in the overall pattern of stability between the experimental and computationally derived protection factor profiles, both confirming the α-spectrin α-helix stability gradients and cross-validating the ability of HR-DXMS to derive protection factor maps that substantially match those obtained computationally by COREX analysis of known three dimensional structures.

Substantial gradients in helix stability in α-spectrin R1617. Contrary to expectation, gradients in exchange rates, and corresponding gradients in helical stability, extended across most of the length of each of the six α-helices in the two-repeat α-spectrin construct. It had been anticipated that a few amides near the end of each helix (turn
residues) might exchange faster than the bulk of the helix. Instead, a gradient of stability increased, and subsequently decreased from one end to the other end of each 20 to 35-residue helix. To evaluate the significance of these stability gradients, we surveyed the hydrogen exchange rates of $\alpha$-helices within five other proteins for which experimental measurements were available: horse cytochrome c, BPTI, SNASE, HEWL, and equine lysozyme [34-37]. Typically, the exchange rates of the first 4 amides of the N-termini of helices in these proteins could not be determined by NMR, indicating that exchange occurred too rapidly to be experimentally determined at the shortest time point experimentally accessible (generally 2 minutes). This is expected since the first 4 amino acids at the N-terminus of a typical $\alpha$-helix (known as the N-cap residues) do not usually have robust cis-hydrogen bond acceptors. Amino acids interior to the N-cap residues in these proteins had typical free energies of hydration between 6-8 kcal/mol, values that were found only in the linker and most stable central portions of the helices in $\alpha$-spectrin R1617. The N-terminus of the B" helix in R17 showed values well below 6 kcal/mol fully 15 residues into the helix.

These values indicate that substantial portions of the $\alpha$-spectrin R1617 helices are much less stable under solution conditions than the comparable regions of the $\alpha$-helices in the five comparison proteins. Although the conformational helix-loop transitions in the crystal structure are located in the BC loop of R17, we cannot exclude that there may be structural differences in crystalline versus solution phases of R1617. Nevertheless, the asymmetric pattern of stability at the ends of the helices provides further support for potential conformational rearrangements in these regions.
Figure 2-1: Spectrin Fragmentation map: Comparison of the fragments generated by pepsin and pepsin plus FPXIII is shown. Pepsin-generated fragments are indicated in blue, and novel fragments produced by pepsin plus fungal protease indicated in orange. A total of 200 fragments were obtained with the combination of the pepsin and fungal protease columns.
Figure 2-2: Deuterium accumulation plots for pepsin generated peptides of spectrin.
Plots of deuterium accumulation for each peptide vs. on-exchange time were constructed
from data obtained by analysis of 114 pepsin-only generated peptides (Figure 2-1) for
three representative peptides spanning the indicated amino acids (141-151, 114-117, and
8-15). The time axis was arbitrarily divided into three regions, (fast, medium, and slow-
exchanging) and the number of deuteron incorporated on each peptide is indicated in
t parenthesis. The latter class was grouped with the very slow class unmeasured in the
limited on-exchange times (<105 sec) used in this experiment.
Figure 2-3: Composite spectrin fragmentation map, manual rate map, and High Resolution exchange rate map. (A) Pepsin generated fragmentation map resulted in 114 fragments spanning 100% of the sequence. (B) Manual assignment of rate-class to construct sequence “α-Spectrin Consensus Rate Map” is indicated by the arrow. Black horizontal bars indicate loop regions and correlate with locations of fast exchanging amides. (C) Application of HR-DXMS to the data from the 200 deuterated α-spectrin R1617 construct fragments obtained by the combined action of pepsin plus FP XII. Locations of the invariant W in both constructs are indicated as red stars.
Figure 2-4 Overlay of the COREX-calculated and experimentally-determined protection factor profiles deduced for α-spectrin R1617. The HR-DXMS algorithm was used to calculate the individual amide exchange rates (black line) from DXMS generated raw data of the 200 overlapping fragments produced by pepsin and FPXIII digestion. (Figure 2-1). Theoretical exchange rate profile (red line) was generated using COREX from the R1617 spectrin construct. Exchange rates are plotted as the log of the protection factor (PF).
Figure 2-5: Mapping of exchange rate classes on to the 3-D crystal structure of the alpha spectrin R1617 construct: The crystal structure of R1617 reveals excellent correlation with secondary structure with. The map indicates substantial gradients in exchange rates are present within each α-helix, with slowly-exchanging regions of helix gradually giving way to fast exchanging-regions. Surprisingly, there is lower stability in the alpha helical linker region. Invariant tryptophans involved in stabilizing the helical bundle (47) are shown located in the slowest exchanging and more stable parts of the helices. (PDB code 1CUN [PDB])
2.5 DISCUSSION

In this study, high-resolution protein stability profiles were derived for the prototypic α-spectrin two tandem-repeat R1617 by a novel experimental approach (HR-DXMS) and by use of the well-validated COREX algorithm operating on the 3-D structure of α-spectrin R1617. The two independently-derived profiles were highly concordant and demonstrated marked, unanticipated gradients in the stabilities of the several long α-helices in the construct. The discovery of these gradients has important implications for proposed mechanisms of α-spectrin elastic behavior.

Spectrin elongation is mediated by tension-induced catastrophic unfolding. Atomic force microscopy measurements of α-spectrin constructs reveal that repeats abruptly unfold at forces of 20 – 50pN [4]. When several repeats are in tandem, the tension-length relationship exhibits a distinctive “sawtoothed” behavior in which tension gradually rises with increasing length until an abrupt drop in tension occurs, returning almost to baseline, followed immediately by repeated tension rise and collapse with continued elongation. The result is that the α-spectrin molecule can be elongated up to multiples of its resting length, with tension constrained to a constant, relatively narrow, range. The abrupt drops in tension have been attributed to catastrophic unfolding of individual repeats, and there is considerable evidence to support this model.

However, the mechanism responsible for the short-range rise in tension with each “sawtooth” is less clear. The force required to “snap open” the repeats is well above that
typically found to be exerted on the \(\alpha\)-spectrin molecule in simulations of membrane deformation, which are more in the range of 5-10pN\[5, 6\]. Taken together, these observations indicate that the mechanisms that account for the rise in tension with each “sawtooth” are central to understanding \(\alpha\)-spectrin’s elasticity. Studies have indicated that repeats may undergo more subtle conformational changes that mediate elasticity in this 5-10pN regime before catastrophic unfolding of the same or tandem repeats occurs with higher tension \[5, 6\].

**\(\alpha\)-Spectrin elastic behavior requires efficient storage of mechanical energy.** Models for \(\alpha\)-spectrin elastic behavior should explain how mechanical energy is stored by tension-induced conformational change so as to allow efficient, low hysteresis recoil when tension is released. Models have been proposed in which tension induces gradual unwinding or melting of the ends of \(\alpha\)-helical regions into elongated, relatively disordered loops\[6, 38\]. In both catastrophic unfolding and helix-melting models it is unclear how mechanical energy could be stored without undue hysteresis: the forces that mediate the non-covalent binding interactions within the structures of proteins operate over short distances and once the distances are exceeded, the forces in large part disappear. Once tension is released, entropic forces may allow reassembly of the unwrapped helical regions, but with resulting large losses in the mechanical energy that unwrapped them.
α-Spectrin elasticity may be mediated by energy-storing loop migration. Crystal structures have been determined for two-repeat constructs consisting of the same sequence as α-spectrin R1617, but with small variations in the particular N- and C-terminal residues chosen to begin and end the construct: i.e. having differing “phases”. These phase-differing constructs exhibited discrete differences in their structures when crystallized [1]. These differences indicated that α-spectrin could reduce its end-to-end distance by helix to loop and loop to helix transitions without change in the overall amount of sequence in loop or helix conformation [1]. This model is further supported by studies of the crystallized 16th repeat of Drosophila α-spectrin which showed a conformational rearrangement of a loop region into a helix [39].

This model is particularly relevant for elastic processes, as it features low hysteresis due to ready reversibility of conformational rearrangements, which involve changes in short-range interactions only within the repeats and no net change in the fraction of loop and helix structure. If the loop migration model is operative during stretching of α-spectrin, the loops could migrate into the progressively more stable regions of the α-helices, with reformation of less stable helix behind them, storing mechanical energy. Hence, DXMS-demonstrated gradients of stability in helical regions are crucial for the storage of mechanical energy by this mechanism. This energy can be recovered when the molecule relaxes, and the loop migrates back into less stable regions of the helices, allowing reformation of the stable helical regions.
**α-Spectrin elasticity may be mediated through linker-region flexibility.** The linker between the 16\(^{th}\) and 17\(^{th}\) repeating units of R1617 and the linker-flanking sequence in each unit (helix C’ and A’’) are present in the crystal structures as a single, very long uninterrupted α-helix [1]. The crystallization study also revealed different degrees of bending of the linker region among the 5 structures determined, which suggested that in solution, the linker region might be significantly more flexible than other α-helical regions of the molecule. Our data support this inference, as we found a significant decrease in the ΔG\(_{\text{exchange}}\) (higher exchange rate) of 2-4 kcal/mol in the linker region when compared to the more stable helical regions flanking the linker. Hence, the linker, although α-helical and interacting with residues of the flanking repeats, may be intrinsically less stable in solution. It is unclear whether this decrease in ΔG\(_{\text{exchange}}\) is due to enthalpic or entropic contributions at the linker region. Presumably the flanking helices, C’ and A’’, are enthalpically more stable due to the hydrophobic packing of the triple helical bundle and are less vulnerable to the dynamic processes governing helical-coil transitions in single helices.

Whether exchanges in the linker region occur via local or global unfolding can be addressed by comparing ΔG\(_{\text{exchange}}\) of the linker region with that of the amides in the flanking helices. The 2-4 kcal/mol lower ΔG\(_{\text{exchange}}\) of the linker region than that of the amides in flanking helices indicates that predominately local unfolding is occurring in the former. Thus the helical linkers are unusually dynamic in solution, and likely to be significantly more flexible than other parts of the molecule. Although the presence of an α-helix joining adjacent repeats may appear to be a "stiff" linkage, our results support the
inference drawn from comparative analysis of differently “phased” structures, that under solution conditions, these linkers are dynamic structures that can provide significant configurational entropy for \(\alpha\)-spectrin chains.

**High resolution DXMS.** In the course of this study, we have developed novel methods (“HR-DXMS”) by which high quality amide hydrogen/deuterium mass spectrometry data can be computationally resolved into near single-amide resolution hydrogen exchange rate profiles for an entire protein construct. The method was validated by demonstrations of its ability to accurately deconvolute realistically-simulated raw DXMS fragmentation data, employing fragmentation densities, ranges of on-exchange times and data precision routinely obtained with DXMS analysis. It was further validated by demonstration of its ability to derive an exchange rate profile for \(\alpha\)-spectrin R1617 which substantially matched that produced by analysis of the crystal structure of \(\alpha\)-spectrin R1617 with the COREX algorithm. With this capability, HR-DXMS now rivals the resolution of NMR-based methods for amide hydrogen exchange rate measurement, with the substantial advantages of being able to measure even the fastest exchanging amides, and to do this with substantially larger proteins, and less material than is required for NMR approaches. The principal requirement for application of this method is that reproducible exchange data on a sufficiently large number of peptides with extensively overlapping sequences be obtained in order to assign a reliable exchange rate to each residue. This is now achievable through application of the enhanced data acquisition methods (“DXMS”) we have developed and recently reported [40-46].
There is a further, unique advantage of the HR-DXMS approach: The functional deuteration step of the analysis can be performed under conditions where the dynamic properties of the study protein can be directly manipulated, with experimental design limited only by the ingenuity of the investigator. For example, with the present \( \alpha \)-spectrin study as a foundation, one can test the “loop migration” hypothesis, and other models for \( \alpha \)-spectrin elasticity by performing comparative HR-DXMS studies with the deuteration step performed while the \( \alpha \)-spectrin molecule is being progressively elongated, for example by shear stress fields generated by fluid flow or stirring. After induction of quench, exactly the same method of analysis that was employed in the present study would allow rigorous assessment of the several proposed mechanisms for \( \alpha \)-spectrin elasticity.

The COREX algorithm was developed with the goal of representing the ensemble thermodynamic behavior of proteins in a computationally accessible manner. It scales well when implemented in a (massively) parallel manner, as opposed to typical molecular dynamics calculations. The amide hydrogen exchange-rate calculating ability of COREX was originally developed to allow validation of the stability profiles it generated by comparison with NMR- derived exchange rate measurements. We anticipate that the rate-calculating ability of COREX will play an important role in the manner in which HR-DXMS-derived protein stability profiles and exchange rate maps are interpreted and exploited. The close agreement between HR-DXMS and COREX-derived exchange rate profiles for \( \alpha \)-spectrin R1617 has heightened this expectation. There are, however, minor portions of the COREX-derived profile that deviate from the experimental profile: the linker region and a portion of the C-terminal region of the molecule are shown to be more
stable by COREX analysis than by HR-DXMS analysis. These differences may result from an inadequate sampling of states in the Monte Carlo mode we employed or from real differences between the calculations from crystal structure and DXMS data.

2.6 ACKNOWLEDGEMENTS

Chapter 2, in full, will be submitted for publication in the *Journal of Molecular Biology* with coauthors Chris Gessner, Jack S. Kim, Krissi Hewett, Vincent J. Hilser, Steven T. Whitten, Ruby I. MacDonald and Virgil L. Woods, Jr. My contribution to the work included the design, data reduction, and interpretation of experiments. I was responsible for running COREX on supercomputing platforms. I was principally responsible for constructing the algorithm and software for High Resolution, residue-specific determination of amide hydrogen exchange rates from DXMS data: HR-DXMS described in Supplemental Material. This software allows construction of high-resolution protein rate-maps ("rate fingerprints") needed in the work described in Chapter 6. I thank Walter Englander, David Wemmer, Pat Jennings, David Baker, Russell Doolittle, Philip Bourne, Halbert White and Karin Maxson for their support and guidance. I’d also like to thank Pavel Pevzner, Neil Jones, and Bryant Forsgren for their expert collaboration in the design and development of the precursor of the HR-DXMS rate map calculating algorithm and software as well as Redstorm Scientific for their generous provision of computational and technical resources. This work was supported by NIH grants CA099835 (VLW), GM57692 (RIM) and by grants from the University of California BioStar and LSI programs, grants S97-90, S99-44, L98-30 (VLW), with the matching
corporate sponsor for these grants being ExSAR Corporation, Monmouth Junction, NJ.

VLW has an equity interest in ExSAR Corporation.
2.7 SUPPLEMENTAL MATERIAL

Algorithm and Software for High Resolution, residue-specific determination of amide hydrogen exchange rates from DXMS data (HR-DXMS):

There does not currently exist an algorithm that will reliably find a globally minimum value for an arbitrary non-linear function. A common numerical difficulty in non-linear optimization is the discovery of local optima which exhibit many of the properties of globally minimal points, but are not in fact globally minimal. All currently known algorithms for non-linear optimization are susceptible to the problem of incorrectly terminating at a local minimum; however, in the case of smooth and continuous objective functions, one can often ameliorate this problem by initializing the numerical optimization with a solution that is likely to be near the global minimum.

The following algorithm centers on use of a two-phase numerical technique, linear programming (LP) followed by nonlinear least squares (NLS). The computational problem is to determine the mass gain ("shifts") for each smallest segment of the protein's sequence (here termed “atomic unit” or AU) that is resolved by differences between each DXMS-generated overlapping fragments' sequence, at each time point measured in the experiment. We apply linear regression of the AU shifts to determine a rate for each AU. The rates from the linear regression analysis of the shifts are then fed into the nonlinear least-squares technique as initial rates.
Linear Programming Method: Linear programming (LP) is a technique that optimizes an objective function subject to certain predefined linear constraints. Given a protein sequence $P$ where $P_i$ denotes the $i$-th character of $P$, a fragment $f_{i,j}$ is simply a substring of $P$: $P_i, P_{i+1} \ldots P_j$ (Figure below). Fragments are generated by the protein digestion phase of the DXMS experiment, and are generally fixed for a given data analysis problem. A position $k$ in the protein is covered by any fragment $f_{i,j}$ when $i \leq k \leq j$. An AU is the largest consecutive substring whose positions is covered by the same set of fragments, and can not overlap (Figure 1A). The concatenation of all AU generated by the fragments in an experiment will cover each position of $P$ only once if, and only if, $P$ is the union of all the fragments. Therefore the set of AU is determined entirely by the set of fragments generated by protein digestion. With the $\alpha$-spectrin R1617 fragmentation map of 114 peptides there are 65 AU spanning 100% of the protein sequence with sizes ranging from 1 to 13 amides in length. By calculating the AU from the fragmentation pattern and knowing the mass shift of each AU at each time point we can calculate the mass shift, $\delta_{au}(t)$, for each AU as well as the corresponding error in our estimation. Note that if our fragmentation pattern produced 212 AU that would represent single amide coverage over the entire 212 amino acid sequence of $\alpha$-spectrin R1617. Therefore we let $A$ be the set of AU determined by $F$ (the set of fragments). For each fragment $f$, where $f$ is a subset of $F$, there exists a set of AU whose positions are covered by $f$. For each AU, $A(i)$, we define a variable $s_{i,t}$ that represents the mass shift of AU $i$ at time $t$. For each fragment $f$, we define a variable $E_{f,t}$ which represents the experimental error in the mass shift measurement for fragment $f$ at time $t$. The computational problem is to determine the mass gain ("shifts"), $s_{i,t}$, for each AU, $A(i)$, at each time point.
measured in the experiment. Figure 2-6A illustrates the definition of the AU for the first 15 amino acid segment of R1617. Atomic units (A1, A2... A8) are defined by the set of fragments (f1, f2... f12) and each fragment shift is the additive contribution of the calculated shifts for each AU, Figure 2-6B. The fragment shift at each time point s(n,t) can be defined as the sum of the shifts of the AU in each fragment at time. We can define an optimal solution for determining the mass shift of all atomic units as the sum of the error of each fragment (f) at time=t E(f,t). We apply linear regression to minimize the global error of all 114 fragments E(114,t) to globally determine a rate for all atomic units. The rates so calculated represent average rates of exchange of all amide hydrogens within the AU and provide us with good initial starting rates to seed into our non-linear least squares fit.

**Non-Linear Least Squares Fit:** We can approximate the exchange process in a protein of N amino acids as N independent chemical reactions where each obeys first-order reaction kinetics. In particular, if amino acid i has rate constant k_{ex,i} then the amount of deuterium D_i(t), at time t at position i is simply

\[ D_i(t) = 1 - e^{-k_{ex,i}t} \]  

(1)

The rate constant k_{ex,i} is a function of pD, temperature, protein sequence, and protein conformation. For a fragment f composed of n amides the amount of deuterium incorporated is
where $D_{F(\alpha),t}$ is the total amount of deuterium on fragment $f$ starting at amino acid residue $m$ through amino acid residue $n$ at time $t$, and $k_{ex,i}$ is the exchange rate constant of amide $i$, where $m \leq i \leq n$. For the nonlinear least-squares technique the computational problem is to find rate constants that minimize the squared difference between the theoretical deuteration level of all measured fragments (equation 2) and the fragments’ observed levels. The objective function is aimed to minimize the global error (GE) and includes a form of equation 2 for all fragments ($p$) at all time points ($z$) and attempts a global fit over all parameters according to a simplified equation 3 for our spectrin analysis of 114 fragments.

$$
\sum_{f=1}^{p} \left[ \sum_{t=1}^{z} \left( D_{obs,F(\alpha),t} - D_{F(\alpha),t} \right)^2 \right] = GE
$$

The parameters in this optimization ($k_{ex,i}$) are exactly the quantities of interest. Equation 3 can be used if the back exchange (loss of deuterium from protein/peptides after the institution of “exchange-quench” conditions) of the peptides is corrected by using the standard peptide average exchange method. For a more rigorous correction of back exchange we can correct for the loss of deuterium on each amide independently by modifying equation 3 to include published off exchange rates of model peptides [31, 32].

In a completely unstructured polypeptide chain, all peptide amide hydrogens are freely accessible to water and exchange at their maximal possible rate, with a half-life of
exchange of approximately one second at 0 °C and pH 7.0. Exact exchange rates for particular amide hydrogens in fully unstructured sequence can be reliably calculated from knowledge of the temperature, pH and primary amino acid sequence involved [31, 32]. The precise rate of exchange of a particular amide in random coil can vary more than thirty-fold from the average rates for all amides in a peptide under such conditions, with the precise rate depending upon the identity of the two amino acids flanking the particular amide bond, and whether or not the amide is at the c- or n-terminus of the peptide [31, 32]. The N-terminal amide in a peptide generally exchanges 20 times faster than the average rate for the other amides in most peptides, a phenomenon that is important to take into account in data reduction calculations. Because DXMS analysis fragments and denatures peptides, we can model the off-exchange of amide deuterium from the fragments as a random coil and represent it as

$$D_{\text{off},i}(T) = e^{-k_{\text{int},i}(T_q)}$$

(4)

where $D_{\text{off},i}(T)$ is the fraction of deuterium left on a deuterated amide given an off-exchange time of $T$, and $k_{\text{int}}$ is the intrinsic exchange rate of the amide under quench conditions (pH 2.3, 273K) calculated from known rates of model peptides from Bai and Englander [31, 32]. $T$ represents the time upon quench to the time the fragment is analyzed in the mass spectrometer and is the sum of the fragment's retention time and the system lag time (SLT), the time between induction of exchange quench and sample loading onto the C18 column (2-5min). Although the retention times for each fragment...
are readily determined, the SLT can be better approximated to a value which results in
the least amount of error in the overall fit, as described below.

Equation 2 for the total deuterium on a fragment can be readily modified to
incorporate amide specific back exchange rate for every amide on that fragment by
substitution of equation 4 to produce equation 5.

\[
D_{\text{corr},F(f),t} = \sum_{i=m}^{n} D_{\text{off},i}(T)(1-e^{-k_{\text{ex}},i,t})
\]

(5)

Now \(D_{\text{corr},F(f),t}\) is the corrected total amount of deuterium for fragment \(f\) at time \(t\) taking
into account amide specific back-exchange rates which are dependent on the fragments
retention time in the system under quench conditions. Substituting \(D_{\text{corr},F(f),t}\) into
equation 3 allows us to refit our exchange rates with the corrections automatically taken
into account (equation 6.)

\[
\sum_{f=1}^{p} \left[ \sum_{t=1}^{z} \left( D_{\text{obs},F(f),t} - D_{\text{corr},F(f),t} \right)^2 \right] = \text{GE}
\]

(6)

**Validation studies of HR-DXMS.** The success of the method relies on the extent
of overlapping fragmentation, the number of sampled on–exchange time points, and the
number of post-quench off exchange time points sampled. Simulation studies were
performed to determine the overall performance of this approach when we used values
for these parameters that were readily achievable in the present study.
α-Spectrin R1617 construct simulations. Studies were performed to determine how accurately the HR-DXMS method could deconvolute input DXMS data, with a fragmentation intensity and number of on-exchange time points similar to those employed with the α-spectrin R1617 construct in the present study (Figure 2-7). We generated an arbitrary “true” exchange rate map for a hypothetical “Hyp R1617 protein” that was approximately based on the rates calculated for α-spectrin R1617 by COREX analysis). Given these hypothetical rates, we generated predicted deuteration levels for each of the same peptide fragment sequences (200 fragments) collected in the actual DXMS experiment with α-spectrin, with and without incorporation of a normally-distributed random variable to simulate experimental and instrumental error. We then fed the HR-DXMS algorithm this simulated deuterated fragmentation data, and compared the resulting deconvoluted amide-specific determinations (expressed as the free energy of exchange; Figure 2-7A,B lavender line) to the free energy of “true” exchange profile of the “Hyp R1617 protein” used to generate the data (Figure 2-7A,B blue line). There was an excellent agreement between the two profiles when using simulated fragmentation data without error. The algorithm produced a normalized global error of .0044. When we introduced a 20% error in the peptide deuteration levels prior to deconvolution the normalized error increased to 19.4 indicating the ability of this method to accurately calculate exchange rates. Since there has been no published data on the values of the global free energy of unfolding of R1617 to exceed 12kcal/mol, we incorporated a further constraint on the algorithm in the form of an energetic cap of 12kcal/mol. This represents the fact that the algorithm cannot produce an exchange rate resulting in a free energy of unfolding greater than 12kcal/mol. In our experience, peptide deuteration levels are
typically measured with a precision of 5-10% in DXMS, due in large part to the reproducibility resulting from the extensive automation we employ. Regions where individual amides considerably diverged can be observed by the lack of overlap between the blue and lavender colored lines (Figure 2-7A,B).

**Horse cytochrome c.** We further extended our test of HR-DXMS by producing simulated DXMS deuterated fragment datasets based on published NMR-determined experimental hydrogen exchange rate data from horse cytochrome c [34]. Residues where the rates of exchange had been too fast to be measurable in the NMR experiments were assigned arbitrary values. Since horse cytochrome c is 104 amino acids in length we used the same fragmentation pattern as that obtained for α-spectrin R1617 for the first 104 amino acid residues. Figure 2-8 shows that the experimentally-determined exchange amide-specific free energy profile of exchange of cytochrome c (blue line) agree closely with the HR-DXMS-deconvoluted rate profile of the simulated data (lavender line). An important requirement for proper behavior of the fitting algorithm is the imposition of upper and lower bounds during the nonlinear least squares fit. Since the slowest exchanging peptides reached 50% deuteration level at $10^5$ sec this corresponds to an average exchange rate on the order of $10^{-6}$/sec. We set our lower boundaries 2 orders of magnitude lower so as to not exclude the possibility that a single amide may show slower rates within a given peptide, with the exception of regions of the protein sequence that had peptides that were maximally deuterated at the 10 sec time point. With these peptides, the lower boundary of exchange was calculated at .92/sec, corresponding to
99.99% deuteration at 10secs. The upper boundaries for the fit were set to the maximum exchange rates of each amide [31, 32].
Figure 2-6: the definition of the AU for the first 15 amino acid segment of R1617. (A) Atomic units (A1, A2… A8) are defined by the set of fragments (f1, f2… f12). (B) Each fragment shift is the additive contribution of the calculated shifts for each AU at time (t). After determining the mass shift for each AU at each time point we apply linear regression of the AU’s shifts to globally determine a rate for each AU by minimizing the global error of the defined optimal solution. The rates so calculated represent average rates of exchange of all amide hydrogens within the AU and provide us with good initial starting rates to seed into our non-linear least squares fit.
Figure 2-7: HR-DXMS deconvolution of Spectrin simulated data: (A) 20% Error (B) 0% Error. Simulated data was generated on the fragmentation map of the 200 peptides (Figure 2-1.) approximately based on the rates calculated for α-spectrin R1617 by COREX analysis (Figure 2-4) to generate a hypothetical “true” rate profile (blue line). There is an excellent overlap between the “true” profile and the deconvoluted profile (lavender line) when no error was present and only marginal degradation with incorporation of 20% error in the simulated data. An cap of 12kcal/mol was imposed as a further constraint.
Figure 2-8: Validation of HR-DXMS by producing simulated DXMS deuterated fragment datasets based on published NMR-determined experimental hydrogen exchange rate data from horse cytochrome c. The fragmentation pattern obtained for $\alpha$-spectrin R1617 for the first 104 amino acid residues was used to generate simulated data. Residues where the rates of exchange had been too fast to be measurable in the NMR experiments were assigned arbitrary values. Experimentally determined exchange profile (represented as free energy of exchange) of cytochrome c (blue line) closely agrees with the HR-DXMS-deconvoluted rate profile of the simulated data (lavender line) of cytochrome c.
2.8 REFERENCES


43. Woods-Jr., V., Methods for identifying hot- spot residues of binding proteins and small compounds that bind to the same., in *U.S. Patent # 6,599,707*. 2003: USA.


47. Pantazatos DP, MacDonald RI. Site-directed mutagenesis of either the highly conserved Trp-22 or themoderately conserved Trp-95 to a large, hydrophobic residue reduces the thermodynamic stability of a spectrin repeating unit. *J Biol Chem.* 1997 Aug 22;**272**(34):21052-9
CHAPTER 3

MOLECULAR INTERACTIONS BETWEEN MATRILYSIN AND THE MATRIX METALLOPROTEINASE INHIBITOR DOXYCYCLINE INVESTIGATED BY DEUTERIUM EXCHANGE MASS SPECTROMETRY

3.1 ABSTRACT

Matrix metalloproteinases (MMPs) play an essential role in normal and pathological extracellular matrix degradation. Deuterium exchange mass spectrometry (DXMS) was used to localize the binding regions of the broad-spectrum MMP inhibitor doxycycline on the active form of matrilysin (residues 95-267), and to assess alterations in structure induced by doxycycline binding. DXMS analyses of inhibitor-bound versus inhibitor-free forms of matrilysin reveal two primary sites of reduced hydrogen/deuterium exchange (residues 145-153; residues 193-204) that flank the structural zinc binding site. Equilibrium dialysis studies of doxycycline-matrilysin binding yielded a $K_d$ of 73 µM with a binding stoichiometry of 2.3 inhibitor molecules per protein, which compares well with DXMS results that show principal reduction in deuterium exchange at two sites. Lesser changes in deuterium exchange evident at the amino and carboxyl termini are attributed to inhibitor-induced structural fluctuations. Tryptophan fluorescence
quenching experiments of matrilysin with potassium iodide suggest changes in conformation induced by doxycycline binding. In the presence of doxycycline, tryptophan quenching is reduced by approximately 17% relative to inhibitor-free matrilysin. Examination of the X-ray crystal structure of matrilysin shows that the doxycycline-binding site at residues 193-204 is positioned within the structural metal center of matrilysin, adjacent to the structural zinc atom and in close proximity to both calcium atoms. These results suggest a mode of matrilysin inhibition by doxycycline that could involve interactions with the structural zinc atom and/or calcium atoms within the structural metal center of the protein.

3.2 INTRODUCTION

Remodeling of the extracellular matrix (ECM) is critical for the maintenance of life. Important physiological processes that depend on efficient remodeling of ECM include organ growth and morphogenesis, tissue repair, embryonic development, and general maintenance of cell and tissue structural support. Tight regulation of these processes is vital for proper physiological function. Aberrant remodeling of ECM can have deleterious effects on tissue structural stability and physiological function, and can lead to pathological tissue remodeling.

Degradation of ECM is catalyzed by ubiquitous enzymes called matrix metalloproteinases (MMPs). The specificity of ECM cleavage by MMPs is broad: MMPs can cleave virtually any ECM component (Egeblad and Werb, 2002). To date, more than twenty-five forms of MMPs have been identified in vertebrates, suggesting
more global roles for MMPs in tissue remodeling (Lindsey, 2004). MMPs are members of the zinc-dependent metallopeptidases (Visse and Nagase, 2003). Members of this family typically contain an auto-inhibitory pro domain, a catalytic zinc-binding domain that operates via a “cysteine switch” mechanism (Wart and Birkedal-Hansen, 1990), and a substrate-binding hemopexin-like domain (Visse and Nagase, 2003). In addition, MMPs contain a second zinc atom and at least two calcium atoms that are not part of the catalytic center and appear to play roles in stabilizing tertiary structure (Lowry et al., 1992; Willenbrock et al., 1995). MMPs are expressed as zymogens (pro-MMPs), and are either secreted from the cell to the extracellular space or are anchored to the cell membrane via a transmembrane domain. The accepted paradigm for in vivo activation of most pro-MMPs involves conversion of extracellular pro-enzyme to the catalytically active form via proteolytic removal of the pro domain (Visse and Nagase, 2003).

Alternate in vivo activation pathways have been proposed including direct proteolytic activation by other MMPs (Nagase, 1997), autolytic activation via sulfhydryl oxidation of the catalytic cysteine residue (Fu et al., 2001), and protease-independent pro-MMP activation via sulfhydryl modification of the catalytic cysteine residue (Okamoto et al., 2001).

Matrilysin (also called MMP-7) is the smallest member of the MMP family and lacks the hemopexin-like domain. Matrilysin has been linked to a number of pathological processes ranging from inflammatory diseases (Matsuno et al., 2003), to heart failure (Boixel et al., 2003), to colorectal, prostate, and breast cancers (Adachi et al., 1999; Davies et al., 2001; Heppner et al., 1996). Of these processes, its purported involvement in cancer is by far the most documented. Studies using a mouse model of
acinar-to-ductal pancreatic metaplasia revealed that deletion of the matrilysin gene inhibited the development of cancerous pancreatic lesions after ductal ligation (Crawford et al., 2002). Similarly, in a mouse model of multiple intestinal neoplasia, matrilysin gene knockout resulted in reduction of tumor multiplicity and a significant decrease in tumor diameter (Wilson et al., 1997). Such studies have brought much attention to matrilysin as a potential therapeutic target for small-molecule inhibitor development.

Doxycycline is the only clinically-approved drug used to inhibit MMP activity and is marketed under the name of Periostat (CollaGenex) for use in the treatment of periodontitis (Peterson, 2004). Doxycycline is a member of the tetracycline family of antibiotics and is known to exert biological effects that are independent of its antimicrobial activity. Doxycycline has been used for the experimental treatment of other pathological conditions that involve MMPs such as aortic aneurysms (Thompson and Baxter, 1999), myocardial infarction (Villarreal et al., 2003), and cancer (Onoda et al., 2004). Although the primary mechanism of action of doxycycline on attenuating ECM degradation is ascribed to its MMP-inhibitory capacities, it remains unclear how doxycycline mediates inhibition.

Given the importance of matrilysin in the pathological progression of the diseases described, we sought to characterize critical sites of inhibitor-protein interactions on matrilysin with the broad-spectrum MMP inhibitor doxycycline. We used an enhanced form of deuterium exchange mass spectrometry (DXMS) (Hamuro et al., 2002b; Hamuro et al., 2003; Woods-Jr, 2002; Woods-Jr and Hamuro, 2001) and equilibrium dialysis to localize putative drug-binding regions on the protein and to correlate them
with the determined number of drug molecules bound per protein. DXMS studies were performed without a priori knowledge of the drug binding sites on the protein, demonstrating the utility of DXMS in identifying and characterizing drug-binding regions on proteins. The studies presented herein suggest a mechanism for doxycycline inhibition of matrilysin activity via drug interactions with the structural zinc and/or calcium atoms.
3.3 MATERIALS AND METHODS

Materials. Matrilysin was obtained from Calbiochem (San Diego, CA). Doxycycline and Sigmacote were purchased from Sigma-Aldrich (St. Louis, MO) and [3H] doxycycline (specific activity: 5 Ci/mmol) was from American Radiolabeled Chemicals, Inc (St. Louis, MO). Formic acid and common reagents were purchased from Fisher Scientific (Fairlawn, NJ). D20 (99.9%) was from Isotec (Miamisburg, OH).

Equilibrium dialysis of matrilysin-doxycycline complex. Equilibrium dialysis experiments were performed under optimized conditions for preservation of enzyme stability. Attempts to perform experiments in buffer containing moderate to high doxycycline (>200 µM) with high calcium (>1mM) concentrations resulted in protein precipitation. Samples of purified matrilysin were diluted to 1 uM and 10 uM in buffer containing 50 mM HEPES, 150 mM NaCl, 50 mM CaCl₂, pH 7.4. Volumes of 100 ul were placed into Slide-A-Lyzer Mini Dialysis units (7 kDa cutoff; Pierce Biotechnology, Rockford, IL) and dialyzed against 1050 µl of buffer in siliconized microfuge tubes with 0, 0.1, 1, 10, 50, 100, 200, 500, 1000, 1200 mM doxycycline for 20-24 hours at room temperature. Samples with drug contained 1 uCi [³H]doxycycline. Minimum time to reach equilibrium (~10 hours) was determined with non-radioactive doxycycline by absorbance at 345 nm. Sample volumes were measured at the end of each experiment. Radioactivity of entire protein sample volumes and equivalent volumes of dialyzate were determined by tritium counting in 5 ml of scintillation liquid.
on a 1900CA Tri-Carb Scintillation Analyzer (Packard Bioscience; Billerica, MA).

Total ligand concentration in the protein compartment ([Dox]_{protein}) was determined from total tritium counts in the chamber (CPM ml^{-1}) and the specific activity of the radiolabeled ligand. Free ligand concentration ([Dox]_{dialyzate}), which is the same in both compartments, was determined from tritium counts measured in the dialyzate sample (CPM ml^{-1}). Concentration of doxycycline bound to matrilysin ([DOX]_{bound}) was the difference between total and free ligand concentration, that is, \[ [\text{DOX}]_{\text{bound}} = [\text{Dox}]_{\text{protein}} - [\text{Dox}]_{\text{dialyzate}}. \]

Data were plotted as fraction of protein containing bound drug, [DOX]_{bound}/[matrilysin], versus log total drug concentration, [DOX]_{total}. Data were analyzed by nonlinear least squares regression with GraphPad Prism (version 3.0) to determine equilibrium dissociation constant (K_d) and binding stoichiometry. Values are reported as mean ± standard deviation.

**IC_{50} Determination.** Inhibition of matrilysin activity by doxycycline was measured by absorbance spectroscopy using a colorimetric peptide substrate, and with purified collagen in a gel-based assay. Colorimetric assays were performed with MMP-thiopeptolide substrate (BIOMOL; Plymouth Meeting, MA). Assays contained 0.6 µM matrilysin in buffer (50 mM HEPES, 150 mM NaCl, 50 mM CaCl_2, 1mM DTNB, pH 7.4) in the presence of increasing concentrations of doxycycline (0, 0.1, 1, 5, 10, 50, 100, 500, 1000 µM). Protein was incubated with drug for 1 hour at 37°C in the dark. Cleavage reactions were started by addition of substrate to a final concentration of 100 µM MMP-thiopeptolide substrate. Kinetic readings were measured at 430 nm every 90 seconds for 1 hour. Substrate cleavage rates were determined from the linear regions of
kinetic curves. IC$_{50}$ was determined from graphs of MMP activity versus log [doxycycline].

Gel-based inhibition assays were conducted with type I collagen purified from newborn rat skin as described (Miller and Rhodes, 1982). Purified collagen was dispersed in 0.5 M acetic acid and extensively dialyzed against buffer (50 mM HEPES, 150 mM NaCl, pH 7.4) using Slide-A-Lyzer cassettes (10 kDa cutoff, Pierce Biotechnology). Amount of collagen used for cleavage assays was determined empirically by visualization of undigested collagen band intensity in Coomassie-stained SDS-polyacrylamide gels: a single concentration that yielded intense resolvable bands was used. Matrilysin (5 uM) was incubated in buffer (50 mM HEPES, 150 mM NaCl, 45 mM CaCl$_2$ pH 7.4) in the absence and presence of doxycycline (0.1, 1, 10, 100, 500, 1000, 5000 uM) at room temperature for 30 minutes. Collagen was added (∼50 µg) to tubes, incubated at 37ºC for 1 hour, and run on SDS-polyacrylamide gels. Collagen band intensities were quantified by densitometry using Image J (1.31v) and IC$_{50}$ was estimated from graphs.

Deuterium Labeling Experiments. Samples of matrilysin (80 µM in 10 mM HEPES pH 7.4, 100 mM NaCl, 250 mM CaCl$_2$) were complexed with the MMP inhibitor doxycycline (8.3 mM) for 30 minutes at room temperature and chilled on ice for 5 minutes. Proteins were deuterated by diluting samples into deuterium exchange buffer (20 mM MOPS pD 7.4, 50 mM NaCl, 1 mM DTT; 70% D$_2$O) for a final doxycycline concentration of 2.5 mM. Samples were immediately incubated on ice for indicated times (10, 100, 1000, 10000 sec), deuterium exchange was stopped by addition of
quench buffer (30 µl of 0.8% formic acid, 0.8M guanidine-HCl), quickly frozen in dry ice, and stored at -80°C for subsequent analyses by mass spectrometry. Nondeuterated control experiments were processed as above in nondeuterated buffer.

Mass Spectrometry Experiments. Protein samples were processed as described (Hamuro et al., 2003). Briefly, denatured deuterated samples were fragmented with a pepsin 20-AL column (porcine pepsin, Sigma, coupled to 20AL support material, Perceptive Biosystems). Fragmented samples were passed through a coupled C18 column for reverse-phase separation of peptides (1 mm X 50 mm, Vydac) using a linear acetonitrile gradient of 5% - 45% over 10 minutes at a flow rate of 50 µl per minute. For precise temperature control, valves, tubing, columns, and auto-sampler were refrigerated at 2.8°C with columns immersed in a melting ice bath. Effluent from the C18 column was analyzed on both a Thermo Finnigan LCQ electrospray mass spectrometer (Woburn, MA) and a Micromass Q-TOF mass spectrometer. Data were collected for each run and saved for subsequent data analyses.

DXMS Data Analysis. DXMS analysis was used to confirm sequence identity and track time-dependent deuterium buildup for all proteolyzed peptides. Briefly, the SEQUEST program (Thermo Finnigan; Woburn, MA) was used to identify the likely amino acid sequence of peptide fragments and analyzed as described (Pantazatos et al., 2004; Woods-Jr, 2002). DXMS analysis was performed using SEQUEST output MS1 and MS2 data, i.e., the parent and daughter tandem mass spectrometric data. SEQUEST identifications were confirmed with DXMS by comparing the predicted isotopic
envelopes for the proteolyzed peptides generated from SEQUEST against MS1 data acquired from the Q-TOF mass spectrometer. DXMS data reduction was used to track deuterium buildup for each peptide fragment over time (10, 100, 1000, 10000 sec). Selected peptides passing this automated quality-control step were manually checked for correct fit and mass identification (Pantazatos et al., 2004). Peptide maps were generated from DXMS data showing localized areas of deuterium buildup on each peptide fragment over time. Areas where deuterium exchange was localized were colored red, and areas without deuterium were colored blue. Specific locations of deuterium within matrilysin were assigned by forming a consensus map of deuterium exchange within regions of overlapping peptides, as described (Pantazatos et al., 2004).

**Fluorescence Quenching of Matrilysin.** Fluorescence measurements were performed using an AlphaScan fluorometer (Photon Technology International; South Brunswick, NJ) at room temperature with an excitation wavelength of 285 nm and slit widths of 8 nm. Tryptophan emission spectra for each sample were measured from 300-400 nm. Spectra were taken in 1 nm increments with 1 second integration per increment. Protein samples of 5 μM were prepared in buffer (50 mM HEPES, 100 mM KCl, 50 μM CaCl₂, pH 7.4) with either 0 or 400 μM doxycycline and incubated in the dark for 30 minutes at room temperature. Competitive absorbance by doxycycline precluded use of higher drug concentrations. Titrations were conducted with freshly prepared potassium iodide (5M stock solution) supplemented with 0.1 mM sodium thiosulfate to prevent oxidation of iodide. Quenching of tryptophan emission was performed by successive additions of potassium iodide (50 - 500 mM) to the cuvette. The sample was gently mixed,
incubated for 3 minutes after each addition of potassium iodide, and the spectra were recorded.

**Analysis of Fluorescence Quenching Data.** Iodide quenching data were analyzed according to the Stern-Volmer and modified Stern-Volmer equations (Lakowicz, 1983). Under conditions where static and collisional quenching do not occur simultaneously, the Stern-Volmer relationship is represented by

\[ \frac{7F_0}{F} = 1 + K_{SV} [Q] \]

where \( K_{SV} \) is the collisional Stern-Volmer quenching constant, \( F_0 \) and \( F \) are the fluorescence intensities in the absence and presence of quencher, respectively, and \([Q]\) is the quencher concentration. A plot of \( \frac{F_0}{F} \) as a function of \([Q]\) yields a linear plot for homogeneous fluorescence emitters, all equally accessible to quencher, whose slope is equal to \( K_{SV} \). A heterogeneous population of fluorophores indicates that only a fraction of fluorophores is accessible to quencher and results in downward curvature of the Stern-Volmer plot (Lakowicz, 1983). For such cases, the modified Stern-Volmer equation is used,

\[ \frac{F_0}{(F_0-F)} = \frac{1}{[Q] f_a K_Q} + \frac{1}{f_a} \]

where \( f_a \) is the fraction of the fluorescence accessible to quencher, and \( K_Q \) is the quenching constant. A plot of \( \frac{F_0}{(F_0-F)} \) as a function of \( 1/[Q] \) yields a linear plot with a slope equal to \( \frac{1}{f_a K_Q} \) and a y-intercept of \( \frac{1}{f_a} \).
3.4 RESULTS

**Doxycycline-Matrilysin Dissociation Constant and Binding Stoichiometry** - The dissociation constant ($K_d$) and binding stoichiometry of the doxycycline-active matrilysin (residues 95-267; hereafter referred to as “matrilysin”) complex was determined by equilibrium dialysis. Minimum time required for doxycycline to equilibrate between buffer and protein compartments was 10 hours: subsequent experiments with protein were performed for 24 hours. A representative binding experiment is shown in Figure 3-1. The binding isotherm shows that matrilysin achieves saturation with doxycycline at millimolar concentrations of drug. Data plotted as saturation fraction versus free doxycycline concentration were analyzed by nonlinear least-squares fitting and revealed a $K_d$ of 73 ± 8 µM and a binding stoichiometry of 2.3 ± 0.2. To control for Donnan effects, parallel equilibrium dialysis experiments were carried out at higher ionic strengths (Suter and Rosenbusch, 1977) and showed no effects on equilibrium (data not shown).

**Inhibition of Matrilysin Activity by Doxycycline** - To address conflicting data in the literature regarding the capacity of doxycycline to inhibit matrilysin (Kivela-Rajamaki et al., 2003; Peterson, 2004), doxycycline-mediated inhibition of matrilysin was demonstrated with a commonly used matrix metalloproteinase-specific peptide substrate (colorimetric substrate) and with native type I collagen. Cleavage of colorimetric peptide substrate in the presence of increasing concentrations of doxycycline yielded an IC$_{50}$ value of 28 ± 5 µM. Detailed analyses of IC$_{50}$ data for this
substrate are provided (see Supplementary Materials). Additional experiments to assess inhibition by doxycycline with a more physiologically-relevant substrate were conducted with native collagen purified from rat. In these experiments, the degree of preservation of collagen is indicative of doxycycline inhibition. The gel in Figure 3-2 shows two bands at molecular weights of approximately 138 and 128 kDa that correspond to uncleaved alpha-chains of collagen. Complete \textit{in vitro} cleavage of collagen occurs in the presence of matrilysin (5 µM) and decreases in the presence of increasing doxycycline (Fig. 2A). Densitometric analyses of collagen band intensities revealed an IC$_{50}$ of 90 ± 31 µM for inhibition of collagen degradation (Fig. 3-2B).

**DXMS Studies of Doxycycline-Matrilysin Complex** - Deuterium exchange mass spectrometry (DXMS) was used to map the binding regions of doxycycline on matrilysin, and to assess local changes in tertiary structure induced by drug binding. Deuterium on-exchange was performed in deuterated buffer under conditions that favored > 97% doxycycline-matrilysin complex formation based on the calculated K$_d$ reported above. Figure 3-3A shows fragmentation maps for matrilysin and matrilysin-doxycycline complex taken at the earliest deuterium on-exchange time point (10 sec). Protein fragmentation and subsequent filtering of fragments for quality (described by Hamuro and coworkers (Hamuro et al., 2003)) resulted in 114 overlapping peptide fragments that covered the entire length of the protein. Discrete changes in deuterium exchange are evident at various locations of the protein. The most prominent changes in deuterium exchange that occur after drug binding are evident at two sites (residues 145-153 and residues 193-204, magenta circles) that flank the structural zinc-binding
region (from primary amino acid sequence). In the absence of drug, extensive
deuterium exchange occurs within the consensus regions depicted in the fragmentation
map (magenta circles, upper map). Upon drug binding, the degree of deuterium
exchange decreases dramatically at these two sites (magenta circles, lower panel).
These results are in good agreement with the determined binding stoichiometry of
approximately 2.3 doxycycline molecules per matrilysin monomer, indicating that these
sites may represent sites of drug binding. Lesser changes in deuterium exchange are
also apparent at the amino and carboxyl termini, the former being largely composed of
an unstructured loop region (Browner et al., 1995). By contrast, no major changes in
deuterium exchange are evident at the catalytic zinc region (residues 210-230, green
circle) or the structural zinc region (residues 156-175, blue circle), indicating that drug
binding does not affect deuterium exchange at these sites. Graphical representation of
deuterium incorporation throughout the length of protein is shown in Figure 3-3B. As
expected, longer durations of deuterium on-exchange (100-10000 sec) give rise to
increased deuterium incorporation throughout the protein, including locations that were
shown to be exchange resistant at 10 seconds, most notably the zinc binding sites
(Figure 3-4 (A-C).

Locations of key peptides showing enhanced or diminished deuterium exchange
carried by doxycycline binding were mapped onto the three-dimensional structure of
human matrilysin (Browner et al., 1995). Figure 3-5A shows regions of deuterium-
exchange flux at the amino (residues 95-110) and carboxyl (250-257) termini in the
presence and absence of doxycycline, and the corresponding rates of deuterium
exchange. The amino-terminal region, which is largely unstructured, reveals
diminished deuterium-exchange after drug binding (Figure 3-5A, upper graph). By contrast, residues 250-257 comprising a short α-helix (helix C, (Browner et al., 1995)) at the carboxyl terminus show increased deuterium exchange when bound to doxycycline (Figure 3-5A, lower graph), which is indicative of alterations in structure or dynamics (Pantazatos et al., 2004) (Hamuro et al., 2002a) (Hamuro et al., 2003). Residues 156-177 comprise a large part of the structural metal center for matrilysin, where several important metal-coordination ligands for the structural zinc and calcium atoms are located (Browner et al., 1995). Approximately 73% of this amino-acid stretch is unstructured loop. The patterns of deuterium exchange observed for the loop region are similar with inhibitor-free or inhibitor-bound matrilysin, albeit slightly less in the presence of drug (Figure 3-5B upper graph), indicating that this region is generally unaffected by doxycycline binding. Residues 210-230 contain three critical histidine residues that coordinate with the catalytic zinc (His218, His222, His228). Doxycycline binding does not show major effects on deuterium exchange from 10 to 100 seconds, relative to unbound matrilysin. However, increased deuterium exchange is evident at later time points (1000 and 10000 sec, Figure 3-4B lower graph). Residues 145-153 and 193-204 are positioned within regions that show extensive reductions in deuterium exchange based on fragmentation maps (Figure 3-3A). Residues 145-153 are located within β-strand 2 (Browner et al., 1995) distal to the metal binding face of the protein (Figure 3-3), whereas residues 193-204 make up β-strand 5, in close proximity to the structural zinc and calcium atoms (Figure 3-5C). In the presence of doxycycline, these residues show large decreases in deuterium exchange at 10 seconds (≈ 60% and 40%,
respectively, Figure 3-5C). Residues 193-204 exhibit resistance to deuterium exchange in the drug-bound state, which persists to the last time point at 10,000 seconds.

**Tryptophan Fluorescence Quenching of Doxycycline-Matrilysin Complex** - To explore changes in matrilysin structure induced by doxycycline binding, steady-state tryptophan fluorescence quenching experiments were performed using potassium iodide (KI). In these experiments, changes in tryptophan quenching reflect alterations in solvent accessibility of tryptophan to quencher, which implicate changes in protein conformation (Lakowicz, 1983). Changes in fluorescence intensity were monitored upon titration of matrilysin with KI. Tryptophan emission spectra obtained for inhibitor-free and inhibitor-bound matrilysin are shown in Figure 6. From the fluorescence profiles, it is evident that the peak wavelength at 325 nm for matrilysin and matrilysin-doxycycline complex does not shift to higher or lower wavelengths during the course of KI titration, indicating that KI does not disturb tertiary structure. The Stern-Volmer plots for KI quenching are shown in Figure 3-6C. In the absence of doxycycline, KI quenching of tryptophan fluorescence shows a linear trend with a Stern-Volmer constant \( K_{SV} \) equal to 6.7 M\(^{-1}\), which reflects homogeneity in solvent accessibility of the tryptophan residues. By contrast, exposure of matrilysin to doxycycline results in a curved plot with a downward trend indicating differential accessibility of quencher to tryptophan residues. In order to conduct quantitative comparisons of apo and doxycycline-bound matrilysin, modified Stern-Volmer analyses of fluorescence data were performed. The modified Stern-Volmer plots of apo and
doxycycline-bound matrilysin both display linear trends, as shown in Figure 6D. Y-intercepts of 1.0 and 1.2 were determined for apo and doxycycline-bound matrilysin, respectively. The reciprocal of the y intercept represents the fraction of tryptophan fluorescence accessible to quencher ($f_a$) (Lakowicz, 1983). In the case of doxycycline-bound matrilysin, the $f_a$ is 0.83 indicating that only 83% of the initial fluorescence is accessible for quenching. In addition, the quenching constant ($K_Q$) calculated for doxycycline-bound matrilysin 6.7 M$^{-1}$ is smaller than apo 7.3 M$^{-1}$, indicating that doxycycline decreases the rate of tryptophan quenching by iodide. To ensure that iodide quenching of fluorescence was not caused by alterations in ionic strength of the solution, identical titrations were also performed using NaCl. No major decreases in fluorescence intensity due to quenching by NaCl were observed (Fig. 3-5C).
Figure 3-1: Determination of dissociation constant (Kd) and binding stoichiometry for doxycycline-matrilysin complex by equilibrium dialysis. Representative graph of equilibrium binding of matrilysin-doxycycline complex is shown. Analyses of binding data by nonlinear least squares revealed a Kd of 73 ± 8 µM and a binding stoichiometry of approximately 2.3 doxycycline molecules bound per protein monomer. Experiments were performed at least in triplicate. (Figure taken from Garcia, R.A., D.P. Pantazatos et. al. 2005)
Figure 3-2: Doxycycline inhibits degradation of native collagen by matrilysin. Type I collagen purified from rat was subjected to proteolysis by matrilysin. (A) Collagen degradation by matrilysin was analyzed SDS-PAGE. The gel shows preservation of collagen bands as a function of increasing doxycycline concentration. Undigested control (denoted “C”) is shown at left. (B) The corresponding graph shows values for collagen band intensities as a percent of undigested control with respect to doxycycline concentration. Estimated IC50 for inhibition of collagen degradation was 90 µM. (Figure taken from Garcia, R.A., D.P. Pantazatos et al. 2005)
Figure 3-3: Amide hydrogen/deuterium exchange maps for matrilysin.

(A) Upper fragmentation map represents inhibitor-free matrilysin and bottom map represents doxycycline-matrilysin complex. Horizontal blue bars represent fragmented matrilysin peptides generated by proteolysis. Individual peptides were color coded red to reflect the number of deuterons that were incorporated onto each peptide. Deuterium incorporation sites were determined by optimizing consensus regions of deuterium content based on patterns of deuterium exchange of overlapping fragments. Deuterons outside of consensus regions were clustered together in the center of unresolved peptide fragments (Pantazatos et al., 2004). As denoted by the magenta circles, primary sites of reduced deuterium exchange are readily detected at sites flanking the structural zinc-binding region (blue circle). Green circle represents the catalytic zinc site. Additional regions showing lesser reductions in deuterium exchange are evident at the amino and carboxyl termini. Sequence number of residues is shown at top. (B) Bar graph showing percent difference in deuterium incorporation between inhibitor-free and inhibitor-bound matrilysin at fastest exchange time point (10 sec) with single representative peptides spanning length of protein. Positive values reflect reduced deuterium incorporation and negative values reflect increased incorporation relative to inhibitor-free matrilysin. Two regions implicated in doxycycline binding (D1 and D2) show the greatest reduction in deuterium exchange. Residue numbers for matrilysin are given in the x-axis. (Figure taken from Garcia, R.A., D.P. Pantazatos et al. 2005)
Figure 3-4: Time-dependent deuterium incorporation of matrilysin-doxycycline complex. Progression of deuterium incorporation at 100, 1000, and 10000 seconds. Regions of zinc binding (Znstruct, blue; Zncat, green) and regions implicated in doxycycline binding (D1, D2, magenta) are shown (Figure taken from Garcia, R.A., D.P. Pantazatos et al. 2005).
Figure 3-5A: Deuterium incorporation for peptides at the amino and carboxyl termini of Matrilysin in relation to the three-dimensional structure. Representative percent deuterium incorporation curves are shown for peptides from 10-10000 seconds on-exchange. Rate of amide deuterium incorporation is dependent upon solvent accessibility and local environment of the amide. Thus, slower accumulation graphs represent regions of protection or increased stability. Three-dimensional coordinates for matrilysin (1MMQ) were downloaded from the Protein Data Bank and displayed using Swiss-PDBViewer 3.7 (www.expasy.org/spdb/). Ribbon models show residues from 95-259. Zinc atoms are represented as orange spheres and calcium atoms are gray spheres. (Figure taken from Garcia, R.A., D.P. Pantazatos et. al. 2005).
Figure 3-5B: Deuterium incorporation for peptides at the zinc-binding regions of matrilysin (residues 156-175, $\text{Zn}_{\text{struct}}$; residues 210-230, $\text{Zn}_{\text{cat}}$).
Representative percent deuterium incorporation curves are shown for peptides from 10-10000 seconds on-exchange. Rate of amide deuterium incorporation is dependent upon solvent accessibility and local environment of the amide. Thus, slower accumulation graphs represent regions of protection or increased stability. Three-dimensional coordinates for matrilysin (1MMQ) were downloaded from the Protein Data Bank and displayed using Swiss-PDBViewer 3.7 (www.expasy.org/spdb/). Zinc atoms are represented as orange spheres and calcium atoms are gray spheres. (Figure taken from Garcia, R.A., D.P. Pantazatos et. al. 2005).
Figure 3-5C: Deuterium incorporation for peptides in matrilysin at the putative doxycycline-binding regions (residues 145-153; residues 193-204).
Representative percent deuterium incorporation curves are shown for peptides from 10-10,000 seconds on-exchange. Rate of amide deuterium incorporation is dependent upon solvent accessibility and local environment of the amide. Thus, slower accumulation graphs represent regions of protection or increased stability. Three-dimensional coordinates for matrilysin (1MMQ) were downloaded from the Protein Data Bank and displayed using Swiss-PDBViewer 3.7 (www.expasy.org/spdb/). Zinc atoms are represented as orange spheres and calcium atoms are gray spheres. (Figure taken from Garcia, R.A., D.P. Pantazatos et. al. 2005).
Figure 3-6: Probing conformation changes in matrilysin induced by doxycycline. Potassium iodide fluorescence quenching of matrilysin in the absence (A) and presence (B) of 400 µM doxycycline. Competitive absorbance of doxycycline precluded use of higher inhibitor concentrations. C, Stern-Volmer plot with matrilysin and DOX-bound matrilysin. For inhibitor-free matrilysin (blue triangles), the plot of Fo/F as a function of [quencher] yields a linear plot indicating that all tryptophan residues are equally accessible to quencher. In the presence of DOX (red squares), the curve shows a downward trend indicating differential accessibility of tryptophan residues to quencher. Changes in tryptophan fluorescence were not caused by increases in electrolyte concentration, as indicated by NaCl titrations (black triangles). D, modified Stern-Volmer analyses indicate that tryptophan exposure of matrilysin decreases by approximately 17% in the presence of doxycycline. (Figure taken from Garcia, R.A., D.P. Pantazatos ett. al. 2005).
Figure 3-7: Schematic depiction of matrilysin inhibition by doxycycline. Inhibition of matrilysin activity is proposed to occur via doxycycline binding proximal to the structural metal center and chelation of the structural zinc and/or calcium cations. Loss of these critical metals may lead to instability of tertiary structure. (Figure taken from Garcia, R.A., D.P. Pantazatos et. al. 2005).
3.5 DISCUSSION

The data reported here demonstrate that doxycycline binding to matrilysin induces marked changes in deuterium exchange and tryptophan fluorescence quenching. Primary reduction in deuterium exchange occurs at residues 145-153 and residues 193-204. The latter amino acid region is adjacent to the structural zinc atom and two calcium atoms, i.e., the structural metal center of matrilysin. Quenching of tryptophan fluorescence decreases in the presence of doxycycline, suggesting that matrilysin undergoes conformational changes triggered by drug binding that decrease tryptophan accessibility to quencher. The propensity of doxycycline to bind adjacent to the structural zinc and calcium atoms is of particular interest given the ability of tetracycline to chelate calcium and zinc cations (Schneider, 2001).

Doxycycline-Binding Regions. At concentrations favoring inhibitor-protein complex formation (>97%), DXMS analyses of the doxycycline-matrilysin complex reveal two prominent regions of reduced deuterium exchange. These sites are considered protected by doxycycline binding within these regions. This assertion is drawn from several pieces of information.

- Fragmentation mapping of matrilysin in the presence of drug shows consensus in reduced deuterium exchange for overlapping peptides at residues 145-153 and residues 193-204 (Figure 3-3A).
- Deuterium exchange rates for peptides at these locations show high degrees of protection (Figure 3-5C). At 10 seconds, exchange at
residues 145-153 shows a marked decrease (60%) upon doxycycline binding. Similarly, residues 193-204 also show high degrees of protection (40%). In addition, protection is conserved at residues 193-204 throughout the course of deuterium exchange, suggesting that drug binding may be more stable here than at the former location.

- Lastly, reduced deuterium exchange at two distinct sites is in good agreement with the determined binding stoichiometry of approximately 2.3 doxycycline molecules bound per protein.

Although our interpretation of these data is not unambiguous, results described here support the assertion that doxycycline binds at these two sites. Comparisons of the amino acid content between these regions show only 16.7% sequence homology, thus ruling out a conserved sequence motif for doxycycline binding. Interestingly, however, both amino acid regions contain a single tryptophan residue (Trp149, Trp198). This finding is significant given that tetracyclines are known to bind proteins via combined hydrophobic and electrostatic interactions (Khan et al., 2002; Kisker et al., 1995). Fluorescence studies on the tetracycline-Tet repressor protein complex showed that tetracycline binding is potentiated in the presence of divalent cations (Takahashi et al., 1986). Moreover, tetracycline was shown to be in direct proximity of a tryptophan residue in the drug-protein complex (Takahashi et al., 1986). These data imply that tryptophan residues alone, or in conjunction with protein-bound divalent metals, may represent viable sites for doxycycline (and tetracycline) interactions and/or docking. Although the studies described herein were performed in the presence of micromolar
calcium, which is in agreement with concentrations known to stabilize fibroblast collagenase (Lowry et al., 1992), we cannot rule out that higher, physiological calcium levels that are present within the extracellular matrix might facilitate further binding to matrilysin.

**Amino-Terminal Peptide Docking Site** - Changes in deuterium exchange were also evident at the amino and carboxyl termini (Figure 3-5A). The carboxyl-terminal helix forms a well-defined binding site for the amino terminal peptide that is located outside of - and adjacent to - the catalytic domain. The carboxyl-terminal helix maintains several inter-region contacts with the amino terminal peptide and is proposed to function as a docking site for the amino-terminal peptide for zymogen activation (Browner et al., 1995). Interestingly, increases in deuterium exchange occur at the carboxyl-terminus (residues 250-257) upon drug binding to matrilysin, indicating that solvent accessibility of this region has increased. These data suggest changes in protein conformation (Pantazatos et al., 2004) (Hamuro et al., 2002a) (Hamuro et al., 2003). By comparison, reduced deuterium exchange of the unstructured amino-terminal loop was observed at 10 seconds (20% decrease, Figure 5A). The observed reduction in deuterium exchange is of lesser magnitude than that observed at the residues 145-153 or 193-204, and is likely due to stabilization of the loop region with vicinal regions of the protein, or possibly to low-affinity interactions with doxycycline present in solution. Stabilization of unstructured loop regions of proteins by ligand binding has been observed by hydrogen/deuterium exchange in several ligand-protein complexes (Hamuro et al., 2003; Yamamoto et al., 2004) (Hamuro et al., 2002a).
**Doxycycline-Induced Conformation Changes** - Data obtained by fluorescence quenching of doxycycline-matrilysin complex suggest changes in structure induced by drug binding. Active matrilysin contains four tryptophan residues (Trp104, Trp136, Trp149, Trp198), three of which fall within amino-acid regions that showed reductions in deuterium exchange when doxycycline was present. Quenching studies demonstrated that drug binding decreases solvent exposure of tryptophan residues (modified Stern-Volmer analyses; Figure 3-6D), implying that local structure vicinal to tryptophan is altered, yielding reduced access to quencher. Trp 104 is present within the amino-terminal loop of matrilysin: the loop region is implicated in conformational rearrangement upon drug binding (discussed above). Remaining tryptophan residues may also become less accessible for quenching via local structure changes. Although our quenching studies infer changes in structure adjacent to tryptophan residues in matrilysin, the exact nature of this conformational change is as yet undetermined.

**Mode of MMP Inhibition by Doxycycline.** The current paradigm for inhibition of MMPs involves chelation of the catalytic zinc atom from the enzyme active site (Peterson, 2004). Structures of matrilysin reported by Browner and coworkers were determined in complex with three unique inhibitors containing hydroxamate, sulfodiimide, or carboxylate as the zinc-coordination group. The polar atoms of all inhibitors interact with the same main-chain atoms at the catalytic active site of matrilysin (Browner et al., 1995). In addition, all inhibitors were shown to form unique bonding groups with the catalytic zinc atom. Hydroxamate exhibited ideal bidentate
bonding with the catalytic zinc atom for stable divalent metal coordination within the active site. By contrast, our observation that doxycycline binds proximal to the structural zinc site suggests that inhibition of the MMP matrilysin may occur without direct interactions with the catalytic zinc atom, but rather via interactions with the structural zinc and/or calcium atoms contained within the structural metal center of the protein. Our results are consistent with enzyme kinetics studies with MMP-8 and MMP-13 (Smith-Jr. et al., 1999), and partial-proteolysis experiments of MMP-8 (Smith-Jr. et al., 1996) that support a mechanism for doxycycline inhibition of MMP activity via chelation of calcium. Chelation of the structural zinc atom by tetracyclines has also been hypothesized to inactivate MMPs via destabilization of tertiary structure (Ryan et al., 1996; Seftor et al., 1998). In this scenario, chelation of the structural zinc and/or critical calcium atoms could result in destabilization of native matrilysin structure and lead to protein denaturation and degradation. Figure 3-7 summarizes a proposed mode of matrilysin inhibition by doxycycline that involves chelation of either zinc or calcium from the structural metal center of matrilysin.

The observed avidity of doxycycline for the structural metal center may play a key role in inhibiting matrilysin activity. Future biophysical studies to localize regions of doxycycline interactions with other classes of MMPs will aid in defining the exact mechanisms by which doxycycline inhibits MMP activity.
3.6 ACKNOWLEDGEMENTS.

Chapter 3 is a reprint of the material as it appears in *Mol Pharmacol*, 2005, Garcia, R.A., D.P. Pantazatos, C.R. Gessner, K.V. Go, V.L. Woods, Jr., and F.J. Villarreal. I was the secondary author, and was responsible for data reduction and analysis, and assisted in manuscript writing. The coauthors either supervised the work, or assisted in data acquisition and manuscript writing. We are grateful to Professor Robert MacDonald (Northwestern University) for use of his fluorometer, Mrs. Krissi Hewitt for assistance with early DXMS data analysis and Neil Tolentino for help with binary maps. We also thank Professors Joseph Adams, and Seth Cohen, and Ms. Shirley Reynolds for their critical reading of the manuscript. This work was supported by a grant from the National Institutes of Health (2 R01 HL43617-10A2) to Francisco Villarreal and CA099835 to Virgil Woods, Jr. Support for Ricardo Garcia was provided by a training grant from the National Institutes of Health, Cardiovascular Physiology and Pharmacology Training Program UC San Diego (2 T32 HL07444-21).
3.7 REFERENCES


CHAPTER 4

RAPID REFINEMENT OF CRYSTALLOGRAPHIC PROTEIN
CONSTRUCT DEFINITION EMPLOYING ENHANCED
HYDROGEN/ DEUTERIUM EXCHANGE MASS
SPECTROMETRY (DXMS).

4.1 ABSTRACT

Crystallographic efforts often fail to produce suitably diffracting protein crystals. Unstructured regions of proteins play an important role in this problem, and considerable advantage can be gained in removing them. We have developed a number of enhancements to amide hydrogen/deuterium exchange-mass spectrometry (DXMS technology) that allow rapid identification of unstructured regions in proteins. To demonstrate the utility of this approach for improving crystallization success, DXMS analysis was attempted on 24 Thermotoga maritima proteins with varying crystallization and diffraction characteristics. Data acquisition and analysis for 21 of these proteins were completed in two weeks and resulted in the localization and prediction of several unstructured regions within the proteins. When compared with those targets of known structure, the DXMS-method correctly localized even small regions of disorder. DXMS analysis was then correlated with the propensity of such targets to crystallize and further utilized to define truncations that improved crystallization. Truncations that were defined
solely on DXMS analysis demonstrated greatly improved crystallization and have been used for structure determination. This approach represents a rapid and generalized method which can be applied to structural genomics or other targets in a high-throughput manner.

4.2 INTRODUCTION

It is widely anticipated that access to high-resolution protein structures will be facilitated by novel high-throughput improvements to conventional crystallographic methods. Proteome-wide crystallography is one avenue being pursued by several groups, including the Joint Center for Structural Genomics (JCSG) (1-3). These efforts have benefited greatly from recent technology enhancements in protein expression and crystallization. Despite these enhancements, production of stable proteins that produce suitable crystals continues to be a serious bottleneck. Many generally well-structured proteins contain disordered regions that may serve as passive linkers between structurally autonomous domains, or become ordered when they interact with binding partners that provide stabilizing atomic contacts (4). Regardless of their function, unstructured regions can inhibit crystallization. Unstructured regions of proteins are also particularly susceptible to contaminating cellular proteases. Removing disordered regions may improve homogeneity. The energetics and kinetics of protein crystallization may be facilitated by selective deletion of unstructured sequences (5). Even those proteins that readily crystallize can suffer from poor diffraction, and it is likely that disorder plays a significant role. Truncated constructs should result in better diffraction and,
consequently, result in higher resolution data more amenable to automated map fitting procedures (6, 7).

In principle, information regarding protein dynamics could be used to design truncations that retain structure and maintain biological function but are otherwise depleted of disordered regions. A number of approaches ranging from stability-dependent protein expression screens to computation of stability from primary structure have been reported (8-10). For structural genomics studies, many targets have unknown folds, which limits the utility of bioinformatic predictions. NMR spectroscopy is one of the most powerful techniques to provide protein dynamics information; however, protein quantity, concentration, experimental time, and size are often limiting factors. Though limited proteolysis coupled to mass spectrometry is a preferred approach, its use is time consuming, frequently requiring that multiple proteolytic reactions be refined for optimal cleavage (6). Interpretation of limited proteolysis results is confounded by the possibility that proteolysis may clip internal loops, leading to destabilization and further proteolytic degradation of what originally was a structured region. Most importantly, there is no facile method to confirm that the truncations designed have retained the stable elements of the full-length protein. These approaches are problematic in structural genomics efforts, where throughput and cost are dominating considerations (11).

For more than 40 years, peptide amide hydrogen-exchange techniques have been employed to study the thermodynamics of protein conformational change and the mechanisms of protein folding (12, 13). More recently, they have proven to be increasingly powerful methods by which protein dynamics, domain structure, regional stability and function can be studied (14, 15). Deuterium exchange methodologies
coupled with Liquid Chromatography Mass Spectrometry (LCMS) presently provide the most effective approach to study exchange rates in proteins (15). Proteolytic and/or collision-induced dissociation (CID) fragmentation methods allow exchange behavior to be mapped to subregions of the protein (15-26). Building upon the pioneering work of Englander and Smith (14, 15, 27), we have developed and implemented a number of improvements which have significantly improved throughput, comprehensiveness, and resolution. We term the method employing these enhancements high-throughput and high-resolution Deuterium Exchange-Mass Spectrometry (DXMS) (28-37).

Peptide amide hydrogens are not permanently attached to proteins, but reversibly interchange with hydrogen present in solvent water. The chemical mechanisms of the exchange reactions are understood, and several well-defined factors can profoundly alter exchange rates (12, 38-40). One of these factors is the extent to which a particular exchangeable hydrogen is exposed (accessible) to water. In a completely unstructured polypeptide sequence, peptide amide hydrogens are always maximally accessible to water and exchange at their maximal rate, which is approximately (within a factor of 30) the same for all amides; their half-life of exchange is in the range of one second at 0°C and pH 7.0 (41, 42). Most amide hydrogens in structured peptides or proteins exchange much more slowly (up to $10^9$-fold reduction), reflecting the fact that exchange occurs only when transient unfolding fluctuations fully expose the amides to solvent water. The exception is the set of very fast exchanging amides in structured regions that have their amides fully solvent-exposed at all times, reflecting their protein-surface disposition. In effect, each amide's exchange rate in a native protein directly and precisely reports solvent accessibility to it, thereby revealing the protein's thermodynamic stability on the
scale of individual amino acids. Measurement of the exchange rates of a protein’s amides can therefore allow direct identification and localization of structured/unstructured regions of the protein; unstructured regions are those where substantial contiguous stretches of primary sequence exhibit the maximal possible exchange rates, indicative of complete and continuous solvation of the amide hydrogens in such segments (12, 13).

With its high-throughput capabilities, DXMS can rapidly localize disorder within crystallographic targets using a minimum of protein sample.

The JCSG is a 5-year pilot project funded by the NIH NIGMS Protein Structure Initiative (PSI; www.nigms.nih.gov) to develop technologies for high throughput protein structure determination. One aspect of these studies is focused on proteins from *Thermotoga maritima* (3). An unbiased set of *T. maritima* targets, 1376 of the 1877 predicted open reading frames, were processed through expression and purification attempts. Of these, 542 proteins were expressed in soluble form and setup for crystallization trials with 434 resulting in preliminary crystal hits. This large dataset provides the basis to select proteins for DXMS analysis based on their propensity to crystallize. To sharply focus this analysis, we have investigated a subset of *T. maritima* proteins selected for their range of known crystallization behavior. In the present study, we have used DXMS to improve crystallographic construct design under high-throughput conditions.
4.3 METHODS

See Supplemental Material for further description of methods.

Protein expression and purification. Twenty-four T. maritima proteins were selected for analysis (see Table 1). These proteins, and the subsequently designed truncated constructs, were freshly prepared for this study as previously described (3). In brief, all targets were expressed in either E. coli DL41 or HK100 from plasmids based on the expression vector pMH1 or pMH4. These vectors encode a 12 amino acid tag containing the first 6 amino acids of thioredoxin and 6 His residues placed at the N-terminus. Expression was induced by the addition of 0.15% arabinose for 3 hours. Bacteria were lysed by sonication, cell debris pelleted, and proteins purified from the soluble fraction by nickel chelate chromatography. Proteins were concentrated to a final volume of 0.75ul with concentrations ranging from 15 to 50 mg/ml in 20mM TrisHCl, pH8.0 with 150 mM NaCl(3).

Establishment of protein fragmentation probe maps. Aliquots of each of the 24 proteins were adjusted to a concentration of 10 mg/ml in Tris-Buffered Saline (5mM Tris, 150mM NaCl, pH 7.0; TBS), and all subsequent steps performed at 0° C on melting ice. In a 4° C cold room, five ul of each solution was further diluted with 15ul of TBS in a microtiter plate employing multichannel pipettors for simultaneous manipulation. Thirty microliters of a stock “exchange quench” solution (0.8% formic acid, 1.6 M GuHCl) was then added to each sample (final concentration 0.5% formic acid, 1.0 M GuHCl), samples transferred
to autosampler vials, and then frozen on dry ice within one minute after addition of quench solution as previously described (28-37). Vials with frozen samples were stored at −80º C until transferred to the dry ice-containing sample basin of the cryogenic autosampler module of a DXMS analysis apparatus designed and operated as previously described (35-37). In brief, samples were melted at 0º C, proteolyzed for 16 seconds by exposure to immobilized pepsin, fragments collected on a c18 HPLC column, with subsequent acetonitrile gradient elution. Column effluent was analyzed on both a Thermo Finnigan LCQ electrospray mass spectrometer and a Micromass Q-Tof mass spectrometer, as previously described (31-37). The Sequest software program (Thermo Finnigan Inc) identified the likely sequence of the parent peptide ions and these tentative identifications were confirmed with specialized DXMS data reduction software as previously described (35-37).

On-exchange deuteration of proteins. After establishment of fragmentation maps for each protein, amide hydrogen exchange-deuterated samples of each of the 24 proteins were prepared and processed exactly as above, except that 5ul of each protein stock solution was diluted with 15 ul of Deuterium Oxide (D2O) containing 5mM Tris, 150mM NaCl, pD (read) 7.0, and incubated for ten seconds at 0º C on melting ice before quench and further processing. Data on the deuterated sample set were acquired in a single automated 30-hour run and subsequent data reduction performed with the DXMS software. Corrections for loss of deuterium label were made as previously described (35-37). The total time elapsed for data acquisition and analysis (both fragmentation maps and deuteration study) was two weeks. A total of 100 ug of each protein was used
to complete the study. For subsequent comparative analysis of the exchange rates of amide hydrogens within truncated protein constructs versus their full-length forms, both proteins were contemporaneously on-exchanged as above, but quenched at varying times (10, 30, 100, 300, 1000, 3000, 10,000, and 30,000 seconds), and further processed as above, employing the fragmentation maps established for the full-length protein.

**Protein crystallization and diffraction data acquisition.** Proteins were crystallized using the vapor diffusion method with 50nl or 250nl protein and 50nl or 250nl mother liquor respective volumes as sitting drops on customized 96 well microtiter plates (Greiner). Each protein was setup using 480 standard crystallization conditions (Wizard I/II, Wizard Cryo I/II [Emerald Biostructures], Core Screen I/II, Cryo I, PEG ion, Quad Grid [Hampton Research]) at 4° and 20° C. Images of each crystal trial were taken at least twice, typically at 7 and 28 days after setup with an Optimag Veeco Oasis 1700 imager. Each image was evaluated using a crystal detection algorithm and scored for the presence of crystals (43). Images at days 7 and 28 were also evaluated manually. Diffraction data were provided by the JCSG from automated data collection at 100K on beamlines of the SSRL Structural Molecular Biology/Macromolecular Crystallography Resource, and the Advanced Light Source beamlines 5.0.2 and 5.0.3 as described previously (3).
4.4 RESULTS

DXMS defines rapidly-exchanging regions of *T. maritima* proteins. In DXMS analysis, fragmentation parameters are initially optimized, including denaturant (GuHCl) concentration, protease type(s), proteolysis duration to maximize the number of peptide fragment probes available for use with the target protein, and then the protein is examined using a broad range of on-exchange times. This approach optimizes our ability to measure the widely ranging exchange rates for most of the peptide amides in the protein (31-37). In the present study, we sought to localize only disordered amides that exchanged very fast in the native protein. Based on prior experience, we employed a single set of fragmentation conditions and on-exchanged samples for a single, brief (10-seconds, 0°C) interval to selectively label only the most rapidly exchanging amides.

Generation of fragmentation maps and acquisition and analysis of deuteration data were completed in two weeks time for 24 samples. Fragmentation maps covering the entire protein sequence were obtained for sixteen proteins, nearly complete coverage for five proteins, and inadequate coverage for three proteins (Table 1). Deuterium on-exchange studies were performed on the 21 proteins that had generated useful fragmentation maps (Figure 1). Deuterium labeling was manually assigned to residue positions within the protein by first optimizing consensus in deuterium content of overlapping peptide probes, followed by further clustering of labeled amides together in the center of unresolved regions, so that a consensus map was generated. The deduced 10-second exchange maps for each of the 21 proteins, and their consensus maps are summarized in Figure 4-1; the detailed deuterated fragment data is shown in Figure 4-5.
(Supplemental Material). The duration of labeling (10 seconds) was calculated to be sufficient to selectively deuterate primarily freely-solvated amides (41, 42). This was confirmed by first fragmenting reference proteins with pepsin to yield unstructured peptides, followed by deuterium-exchange labeling of the resulting peptide mix for 10 seconds at pH 7.0, 0°C as above and, then, quenching and subjecting the mixture to DXMS analysis, but without repeated proteolysis. Under these conditions, all peptides were saturation-labeled with a 10-second period of on-exchange (data not shown).

DXMS correctly localized disordered regions in control proteins with known 3-D structures. Interpretation of the exchange maps of the *T. maritima* proteins was guided by the expectation of two patterns of fast exchange labeling: structurally stable, but well solvated, rapidly exchanging residues (one to three contiguous residues) versus labeling of longer stretches of sequence (four or more residues) indicative of disorder. It was presumed that three contiguous amino acids was likely the smallest number needed to complete a structurally stable turn on the surface of a protein. The percent of each protein's residues that rapidly labeled in stretches of four or more residues is indicated as “DXMS%” in TABLE 1.

The structure of *T. maritima* thy1 protein TM0449 has been determined to 2.25 Å (44). Its exchange map demonstrated two segments (≥ 4 residues in each) with rapid exchange, labeled A (Phe 31- Glu 38) and B (Ser88- lys 93), and several isolated rapidly exchanging amides in groups of 3 or less, scattered throughout the sequence (Figure4-2A). Both of the rapidly-exchanging segments corresponded closely to regions of disorder in the crystal (Phe 32- Glu 38 and Ser 89- Ser 94, Figures4-2B) confirming the
ability of DXMS data to detect and localize such disordered regions. Interestingly, these regions also appear to be involved in the binding of the enzyme substrate and adopt a structured conformation after binding ligand (44). This suggests that DXMS can also provide some localized prediction of substrate and cofactor binding sites. This raises the caution that even focused deletion of unstructured regions always carries the potential to remove regions critical to biological function. Similar comparisons were performed for other proteins with known structures (data not shown) with regions of internal disorder typically mapping to loop or extended solvent-accessible regions.

Poorly crystallizing T. maritima proteins contain substantial disorder. The exchange map for T. maritima GroES heat shock protein TM0505 demonstrated rapid exchange for three segments containing four or more contiguous rapidly-exchanging residues, which together constitute 16% of its sequence (Figure 4-3). While this T. maritima protein had previously produced only poorly diffracting crystals, it is a close homolog of the GroES heat shock protein of M. tuberculosis, for which crystal structures were available as the GroES heptamer, and as a complex (GroELS) with the GroEL subunit (45, 46). When the T. maritima residues with rapid exchange are mapped on the M. tuberculosis structures, they predominantly localize to disordered residues in GroES that make contact with the GroEL binding surface.

The exchange map for the conserved hypothetical protein TM1816 (Figure 4-1 and Supplemental Material) is dominated by several substantial regions of disorder, constituting 17.7% of its residues. This protein was a unique example where a structure was obtained from a target exhibiting substantial disorder. The poorly crystallizing
proteins TM1171, TM0160, TM1706, TM1733 and TM1079 exhibit, for substantial
portions of their sequence, rapidly exchanging stretches of 4 or more residues (13.9%,
12.1%, 11.5%; 6.6% and 5.7% respectively, Figure 4-1 and Supplemental material).
TM0160, TM1171, and TM1172 had disorder primarily at the carboxy-terminus;
(Figure 4-1 and Supplemental Material). These targets offer a straightforward route to
domain optimization by simple deletion of the disordered carboxy-terminus. The
optimization of two of these targets is described below.

Disorder-depleted constructs of T. maritima proteins preserve ordered structure.
Truncation mutants of TM0160 and TM1171 proteins were prepared (Figures 4-4A and
4-4B), in which the carboxyl-terminal disordered region(s) of both proteins were deleted.
The fragmentation patterns produced by pepsin often exhibited preferences for sites near
exchange-defined stretches of disorder. We therefore produced several truncated
constructs to each full-length protein, in part guided by the location of the "preferred"
pepsin cut sites, and for both TM0160 and TM1171. Deletions were designed solely on
the basis of DXMS experimental data. The truncations expressed well as a soluble
protein. Full-length TM0160, and its longest truncated version (D3), were on-exchanged
variously for 10, 100, 1,000, and 10,000 seconds at 0 °C on ice, exchange-quenched and
subjected to comparative DXMS analysis as described above. The resulting 10-second
exchange maps for full-length protein and the D3 truncated version (Figure 4-4C) had
virtually identical 10-second patterns, and detailed analysis of the longer exchange times
demonstrated that D3 had a stability profile identical to that of the TM0160 full-length.
Similarly, each of the four TM1171 truncated constructs expressed well as soluble
protein, and had DXMS stability maps identical to that of the TM1171 full-length protein in the corresponding sequence regions (data not shown).

Deletion constructs of two T. maritima proteins show marked improvement in crystallization. Full-length TM0160 and the D3 truncation were submitted for crystallization trials (Table 1). A total of 480 commercially available crystallization solutions were screened at 4°C and 20°C as described in the Methods section. From multiple protein preparations and crystallization attempts the full-length protein showed marginal crystals (inadequate for diffraction experiments) for only 3 of 2400 total attempts. In contrast using the same 480 crystallization solutions, 76 crystal hits were obtained for the truncated constructs from 1920 attempts. Crystals from the TM0160 D3 truncation mutant had better morphology than did the few crystals obtained with the full-length construct and diffracted well. Ultimately, a 1.9 Å dataset from selenomethionine-incorporated protein enabled determination of the TM0160 3-dimensional structure, which represents a novel fold (to be presented elsewhere). Similarly, the TM1171 and truncations were subjected to crystallization trials. Whereas the TM1171 full-length protein again showed very marginal crystallization propensity (5 out of 2400 attempts), each of the four TM1171 deletion constructs showed marked improvement in crystallization success with the TM1171- D4 construct ultimately resulting in a 2.1 Å dataset that was used to determine its 3-dimensional structure (to be presented elsewhere). It should be noted that well-diffracting crystals were obtained for DXMS-designed deletion constructs in both native and selenomethionine forms.
4.5 DISCUSSION

These studies have shown that DXMS analysis can reliably detect and localize disordered regions within an otherwise structured protein. Stability profiles were determined for 21 *T. maritima* proteins that had previously been subjected to crystallization studies (Table 1). Twelve proteins crystallized readily in >1% of the conditions tested. Four of the remaining nine poorly-crystallizing proteins had a high fraction (>10%) of their sequence in disordered regions suggesting this as a potential cause of the poor behavior. Most importantly, our present instrumentation allowed determination of the DXMS-protein stability profiles at speeds matching the needs of HT Structural Genomics.

We have also established that successful strategies to selectively delete disorder from protein constructs can be readily discerned from DXMS stability profiles. Furthermore, we have shown that DXMS can rapidly and reliably assess the fidelity of preservation of full-length structure in truncations. While several bioinformatic approaches to construct design can be used with well-characterized protein folds, DXMS-guided construct re-design offers a particular advantage in the study of proteins that have novel folds. DXMS data directly localizes disorder to specific amino acid residues in the target protein regardless of overall fold structure, allowing greatly refined truncation definition. Unlike NMR methods, which can also provide exchange data, DXMS requires only microgram amounts of soluble protein and data acquisition and analysis can be performed in a rapid timescale. In the present investigation, the total time elapsed for
data acquisition and analysis (both fragmentation maps and deuteration study) was two weeks, and a total of 100 ug of each protein was used.

Finally, these results establish that DXMS stability profile-guided construct design can produce derivatives of poorly crystallizing proteins that crystallize and diffract well. In each of two attempts, we have succeeded in producing diffraction quality crystals of truncated constructs of _T. maritima_ full-length proteins that had behaved poorly in several crystallization attempts, and have confirmed that these truncations preserved full-length exchange rate patterns, indicating that they had retained full-length structure with high fidelity. Taken together, these results indicate that DXMS is a valuable tool for structural genomics efforts.

### 4.6 ACKNOWLEDGMENTS

Chapter 4 is a reprint of the material as it appears in Proc. Natl. Acad Sci U SA, 2004 with coauthors J.S. Kim, H.E. Klock, R.C. Stevens, I.A. Wilson, S.A. Lesley, and V.L. Woods, Jr. I thank Walter Englander, David Wemmer, and Philip Bourne for their support and guidance in this DXMS application, Linda Brinen, Glen Spraggon and Ashley Deacon for providing diffraction screening data, Dan McMullan, Kevin Rodrigues, Juli Vincent and Eileen Ambing for protein purification and crystallization studies, and Andreas Kreusch for help with figures. This work was supported by NIH grants CA099835 (VLW) and by NIH Protein Structure Initiative grant GM-99-009 from the National Institute of General Medical Sciences (www.nigms.nih.gov), GM 062411 (IAW), and by grants from the University of California BioStar and LSI programs, grants
S97-90, S99-44, L98-30 (VLW), with the matching corporate sponsor for these grants being ExSAR Corporation, Monmouth Junction, NJ. VLW has an equity interest in ExSAR Corporation and SWE is a consultant to ExSAR Corporation.
Figure 4-1: 10-second deuteration results for 21 proteins analyzed by DXMS. Amino acid lengths varied from 76 to 461 residues. Dark regions indicated fast exchanging amides and clear regions indicate stretches of no exchange. Regions of four or more fast exchanging amides are circled. Corresponding boundaries for fast exchanging amides are displayed in the Supplemental Material. (Figure taken from Pantazatos, D.P et. al. 2004).
Figure 4-2A: 10-second amide hydrogen/deuterium exchange map for TM0449. The horizontal blue bars are the protein’s pepsin-generated fragments that had been produced, identified, and used as exchange rate probes in the subsequent 10-second deuteration study. The number of deuterons that went on to each peptide in 10 seconds is indicated by the number of red residues in each peptide. Deuterium labeling was manually assigned to residue positions within the protein by first optimizing consensus in deuterium content of overlapping peptide probes, followed by further clustering of labeled amides together in the center of unresolved regions (with vertical bars indicating the range of possible location assignments), generating the consensus map at the top, in which two extensive segments are seen to be deuterium labeled: 1 (Phe 31- Glu 38) and 2 (Ser88-lys 93). B. The electron density of the crystal indicates two regions of disordered sequence, corresponding to the segments 1 and 2. (C, D) Detailed electron density maps are shown, in which density is not visualized between the Phe 31 to Glu 39 and Ser 88 to Ser 95 regions of the TM0449 3-D structure (47). DXMS-determined disorder constitutes 6.4% of this protein’s sequence. Figure taken from Pantazatos, D.P et. al. 2004).
Figure 4-3: The on-exchange map of TM0505 indicates three internal segments (A, B, and C) of rapidly exchanging amides. The internal segments are mapped onto the crystal structure of the GroES protein homolog of TM0505. The M. tuberculosis GroEL subunit is shown in blue and the heptamer complex of M. tuberculosis GroES subunits is shown in gray. The homologous location of rapid exchange sites in the T. maritima protein are indicated in red. Disorder constitutes 16.3% of this protein's sequence. (Figure taken from Pantazatos, D.P et. al. 2004).
Figure 4-4: The on-exchange maps of TMTM1171 and TM0160: TM1171 (A) and TM0160 (B) show substantial C-terminal disorder (circled sequences). Four truncated constructs of each protein were made by eliminating the C-terminal regions (D1-D4). (C). Repeat DXMS analysis demonstrates that deletion constructs of TM0160 preserve the core full-length structure. Full-length TM0160, and its longest truncation (D3), were on-exchanged variously for 10, 100, 1,000, and 10,000 seconds at 0º C, exchange-quenched and subjected to comparative DXMS analysis as described above. The resulting comprehensive exchange maps for full-length (upper panel) and D3 truncated (lower panel) had virtually identical patterns (10 second exchange time shown) (Figure taken from Pantazatos, D.P et. al. 2004)
TABLE 4-1: Description of T. maritima proteins studied, as classified by crystallization history. Computational predictions (SEG%) (48) and the portion of each protein’s sequence found to be present in high-exchange rate stretches of primary sequence (four or more rapidly exchanging contiguous residues; DXMS%) are given as a percentage of total residues. The primary location of the DXMS-identified rapidly-exchanging regions is indicated. The number of unique crystallization tests is indicated along with the number of tests showing crystal hits or crystals of sufficient size to mount for diffraction screening. The percentage of total tests that led to crystals is indicated. Those targets showing less than a 1% hit rate are considered poorly crystallizing. The number of crystals screened for diffraction and the best resolution are indicated where data are available. (Table taken from Pantazatos, D.P et. al. 2004).
<table>
<thead>
<tr>
<th>Target</th>
<th>Structure</th>
<th>SEG%</th>
<th>DXMS%</th>
<th>Location</th>
<th>Screened</th>
<th>Hit</th>
<th>Mountable</th>
<th>%Crystallized</th>
<th>Screened</th>
<th>Resolution(Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TM0665</td>
<td>1J6N</td>
<td>11.2</td>
<td>3.1</td>
<td></td>
<td>1920</td>
<td>247</td>
<td>244</td>
<td>25.6</td>
<td>7</td>
<td>2.2</td>
</tr>
<tr>
<td>TM1056</td>
<td>pending</td>
<td>17.7</td>
<td>0.0</td>
<td></td>
<td>3360</td>
<td>175</td>
<td>153</td>
<td>9.8</td>
<td>45</td>
<td>1.8</td>
</tr>
<tr>
<td>TM0064</td>
<td>1J5S</td>
<td>3.0</td>
<td>1.1</td>
<td></td>
<td>2880</td>
<td>114</td>
<td>146</td>
<td>9.0</td>
<td>31</td>
<td>1.8</td>
</tr>
<tr>
<td>TM1464</td>
<td>pending</td>
<td>10.1</td>
<td>1.3</td>
<td></td>
<td>2400</td>
<td>71</td>
<td>98</td>
<td>7.0</td>
<td>68</td>
<td>2.6</td>
</tr>
<tr>
<td>TM1080</td>
<td>1O1X</td>
<td>3.9</td>
<td>2.6</td>
<td></td>
<td>2400</td>
<td>119</td>
<td>26</td>
<td>6.0</td>
<td>32</td>
<td>3.5</td>
</tr>
<tr>
<td>TM0449</td>
<td>1KQ4</td>
<td>2.6</td>
<td>6.4</td>
<td></td>
<td>1152</td>
<td>42</td>
<td>16</td>
<td>5.0</td>
<td>16</td>
<td>2.3</td>
</tr>
<tr>
<td>TM0486</td>
<td></td>
<td>5.7</td>
<td>0.0</td>
<td></td>
<td>3552</td>
<td>139</td>
<td>29</td>
<td>4.7</td>
<td>13</td>
<td>3.9</td>
</tr>
<tr>
<td>TM0542</td>
<td>pending</td>
<td>9.3</td>
<td>2.9</td>
<td></td>
<td>2592</td>
<td>47</td>
<td>49</td>
<td>3.7</td>
<td>56</td>
<td>2.8</td>
</tr>
<tr>
<td>TM1733</td>
<td></td>
<td>10.0</td>
<td>6.6</td>
<td>internal</td>
<td>1920</td>
<td>33</td>
<td>22</td>
<td>2.9</td>
<td>29</td>
<td>3.4</td>
</tr>
<tr>
<td>TM1158</td>
<td>1O1Y</td>
<td>2.5</td>
<td>2.6</td>
<td></td>
<td>2880</td>
<td>32</td>
<td>39</td>
<td>2.5</td>
<td>62</td>
<td>2.2</td>
</tr>
<tr>
<td>TM0269</td>
<td>1J6R</td>
<td>2.8</td>
<td>4.0</td>
<td></td>
<td>4608</td>
<td>55</td>
<td>33</td>
<td>1.9</td>
<td>11</td>
<td>2.1</td>
</tr>
<tr>
<td>TM1764</td>
<td></td>
<td>29.3</td>
<td>n.d.</td>
<td></td>
<td>480</td>
<td>4</td>
<td>2</td>
<td>1.3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>TM1816</td>
<td>1O13</td>
<td>11.0</td>
<td>17.7</td>
<td>internal</td>
<td>2016</td>
<td>2</td>
<td>21</td>
<td>1.1</td>
<td>9</td>
<td>2.0</td>
</tr>
<tr>
<td>TM0505</td>
<td></td>
<td>21.1</td>
<td>16.3</td>
<td>internal</td>
<td>3744</td>
<td>12</td>
<td>10</td>
<td>0.6</td>
<td>8</td>
<td>6.3</td>
</tr>
<tr>
<td>TM0212</td>
<td></td>
<td>21.3</td>
<td>0.0</td>
<td></td>
<td>1152</td>
<td>5</td>
<td>1</td>
<td>0.5</td>
<td>1</td>
<td>8.6</td>
</tr>
<tr>
<td>TM0320</td>
<td></td>
<td>30.4</td>
<td>0.0</td>
<td></td>
<td>1152</td>
<td>4</td>
<td>1</td>
<td>0.4</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>TM1171</td>
<td></td>
<td>10.3</td>
<td>13.9</td>
<td>C-term</td>
<td>2400</td>
<td>4</td>
<td>1</td>
<td>0.2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>D1</td>
<td></td>
<td>16.4</td>
<td>5.9</td>
<td></td>
<td>2880</td>
<td>41</td>
<td>33</td>
<td>2.6</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>D2</td>
<td></td>
<td>16.5</td>
<td>5.9</td>
<td></td>
<td>2880</td>
<td>66</td>
<td>9</td>
<td>2.6</td>
<td>16</td>
<td>2.3</td>
</tr>
<tr>
<td>D3</td>
<td></td>
<td>15.3</td>
<td>5.6</td>
<td></td>
<td>1920</td>
<td>24</td>
<td>10</td>
<td>1.8</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>D4</td>
<td>pending</td>
<td>14.9</td>
<td>5.6</td>
<td></td>
<td>1920</td>
<td>12</td>
<td>14</td>
<td>1.4</td>
<td>1</td>
<td>2.1</td>
</tr>
<tr>
<td>TM1706</td>
<td></td>
<td>17.9</td>
<td>11.5</td>
<td>internal</td>
<td>1440</td>
<td>3</td>
<td>0</td>
<td>0.2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>TM1172</td>
<td></td>
<td>11.5</td>
<td>3.5</td>
<td>C-term</td>
<td>2208</td>
<td>3</td>
<td>1</td>
<td>0.2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>TM0913</td>
<td></td>
<td>7.5</td>
<td>2.2</td>
<td></td>
<td>4320</td>
<td>5</td>
<td>2</td>
<td>0.2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>TM1079</td>
<td></td>
<td>16.3</td>
<td>5.7</td>
<td>internal</td>
<td>1920</td>
<td>3</td>
<td>0</td>
<td>0.2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>TM1773</td>
<td></td>
<td>10.4</td>
<td>n.d.</td>
<td></td>
<td>1440</td>
<td>1</td>
<td>1</td>
<td>0.1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>TM0160</td>
<td></td>
<td>15.5</td>
<td>12.1</td>
<td>C-term</td>
<td>2400</td>
<td>2</td>
<td>1</td>
<td>0.1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>D3</td>
<td>pending</td>
<td>11.6</td>
<td>2.5</td>
<td></td>
<td>1920</td>
<td>37</td>
<td>39</td>
<td>4.0</td>
<td>3</td>
<td>1.9</td>
</tr>
<tr>
<td>TM0855</td>
<td></td>
<td>9.1</td>
<td>n.d.</td>
<td></td>
<td>1920</td>
<td>0</td>
<td>0</td>
<td>0.0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
4.7 SUPPLEMENTAL MATERIAL

Detailed Methods

Sample processing for establishment of protein fragmentation probe maps. Vials with frozen samples were stored at –80°C until transferred to the dry ice-containing sample basin of the cryogenic autosampler module of the DXMS apparatus. Samples were individually melted at 0°C, then injected (45 µl) and pumped through an immobilized pepsin column (0.05% TFA, 250µl/min, 16 seconds exposure to pepsin; 66 µl column bed volume, coupled to 20AL support from PerSeptive Biosystems at 30 mg/ml). Pepsin-generated fragments were collected onto a C18 HPLC column, eluted by a linear acetonitrile gradient (5 to 45 % B in 30 minutes; 50 µl/min; solvent A, 0.05% TFA; solvent B, 80% acetonitrile, 20% water, 0.01% TFA), and effluent directed to the mass spectrometer with data acquisition in either MS1 profile mode or data-dependent MS2 mode. Mass spectrometric analyses used a Thermo Finnigan LCQ electrospray ion trap type mass spectrometer operated with capillary temperature at 200°C or an electrospray Micromass Q-Tof mass spectrometer, as previously described (31-37). The Sequest software program (Thermo Finnigan Inc) identified the likely sequence of the parent peptide ions. Tentative identifications were tested with specialized DXMS data reduction software developed in collaboration with Sierra Analytics, LLC, Modesto, CA. This software searches MS1 data for scans containing each of the peptides, selects scans with optimal signal-to-noise, averages the selected scans, calculates centroids of isotopic envelopes, screens for peptide misidentification by comparing calculated and known centroids, then facilitates visual review of each
averaged isotopic envelope allowing an assessment of "quality" (yield, signal/noise, resolution), and confirmation or correction of peptide identity and calculated centroid (31-37).

**On-exchange deuteration of proteins.** After establishment of fragmentation maps for each protein, amide hydrogen exchange- deuterated samples of each of the 24 proteins were prepared and processed exactly as above, except that 5ul of each protein stock solution was diluted with 15 ul of Deuterium Oxide (D$_2$O), containing 5mM Tris, 150mM NaCl, pD (read) 7.0, and incubated for ten seconds at 0º C on melting ice before quench and further processing. Data on the deuterated sample set were acquired in a single automated 30- hour run, and subsequent data reduction performed on the DXMS software. Corrections for loss of deuterium-label by individual fragments during DXMS analysis (after “quench”) were made through measurement of loss of deuterium from reference protein samples that had been equilibrium- exchange- deuterated under denaturing conditions as previously described (31-37). The total time elapsed for data acquisition and analysis (both fragmentation maps and deuteration study) was two weeks, and a total of 100 ug of each protein was used to complete the study. The personnel performing the data acquisition and reduction part of the study were unaware of the identity or crystallization histories of the proteins while data were being acquired and processed. For subsequent comparative analysis of the exchange rates of amide hydrogens within truncated protein constructs vs. their full-length forms, both proteins were contemporaneously on- exchanged as above, but quenched at varying times (10,
30, 100, 300, 1000, 3000, 10,000, and 30,000 seconds), and further processed as above, employing the fragmentation maps established for the full-length protein.

**Equipment configuration.** The equipment configuration consisted of electrically-actuated high pressure switching valves (Rheodyne), connected to two position actuators from Tar Designs Inc., Pittsburgh, as described previously (31-37). A highly modified Spectraphysics AS3000 autosampler, partially under external PC control, employed a robotic arm to lift the desired frozen sample from the sample well, then automatically and rapidly melted and injected the sample under precise temperature control (31-33, 35-37). The autosampler basin was further thermally insulated and all but 20 vial positions were filled with powdered dry ice sufficient to keep samples colder than -45° C for 18 hours. Four HPLC pumps (Shimadzu LC-10AD) were operated by a Shimadzu SCL-10A pump controller. One produced forward flow over the pepsin column, another backflushed the protease column after sample digestion (0.05 % aqueous TFA), and two delivered solvents to a downstream HPLC column for gradient elution (A: 0.05 % aqueous TFA; B: 80% acetonitrile, 20% water, 0.01% TFA; 1 x 50 mm C18 Vydac # 218MS5105, pH 2.3). Valves, tubing, columns and autosampler were contained within a refrigerator at 2.8° C, with pepsin and HPLC columns immersed in melting ice. The timing and sequence of operation of the DXMS apparatus fluidics were controlled by a personal computer running an in-house written LabView-based program, interfaced to solid-state relays (digital input/output boards, National Instruments), controlling pumps, valve actuators, and MS data acquisition (31-37).
Figure 4-5: Summary of DXMS analysis of 21 T. Maritima proteins.
Figure 4-5: Summary of DXMS analysis of 21 *T. Maritima* proteins. (Continued).
Figure 4-5: Summary of DXMS analysis of 21 T. Maritima proteins. (Continued).
Figure 4-5D: Summary of DXMS analysis of 21 T. Maritima proteins. (Continued).
4.8 REFERENCES


CHAPTER 5

ON THE USE OF DXMS TO PRODUCE MORE CRYSTALLIZABLE PROTEINS – STRUCTURES OF THE THERMOTOGA PROTEINS TM0160 AND TM1171

5.1 ABSTRACT

The structure of two *T. maritima* proteins, a conserved hypothetical protein (TM0160) and a transcriptional regulator (TM1171) have now been determined at 1.9Å and 2.3Å resolution, respectively, as part of a large scale structural genomics project. Our first efforts to crystallize full length versions of these targets were unsuccessful. However, analysis of the recombinant purified proteins using the technique of Enhanced Amide Hydrogen/Deuterium Exchange Mass Spectroscopy (DXMS) revealed substantial regions of rapid amide deuterium hydrogen exchange, consistent with flexible regions of the structures. Based on these exchange data, truncations were designed to selectively remove the disordered C-terminal regions, and the resulting daughter proteins showed greatly enhanced crystallizability. Comparative DXMS analysis of full-length protein vs. truncated forms demonstrated complete and exact preservation of the exchange rate profiles in the retained sequence, indicative of conservation of the native folded structure. This study represents the first demonstration of the use of the DXMS method for salvaging intractable crystallization targets. The structure of TM0160 represents a new fold and highlights the use of this
approach where any prior structural knowledge is absent. The structure of TM1171 represents an example where the lack of a substrate/cofactor may impair crystallization. The details of both structures are presented and discussed.

5.2 INTRODUCTION

Structural genomics initiatives that attempt to elucidate structures for an entire proteome are currently ongoing (Lesley et al., 2002). Coupled with this endeavor is the determination of structures for which very little biochemical or structural information is known. Such structures are often classified as “hypothetical proteins” as they have no significant match in sequence comparison searches with proteins of known function. This situation presents a unique problem to structural genomics, as most structures to be analyzed are biochemically characterized. Therefore, crystallographers must rely on structure prediction algorithms for insight into expression construct design and analysis. Problematic proteins may require modification or the addition of a substrate/cofactor to permit crystallization, yet little can be predicted based on existing structural information or primary sequence beyond features like sequence complexity using programs, such as SEG (Wootton and Federhen, 1993) or the vast array of secondary structure prediction algorithms (Barton, 1995) (and references within). It is generally accepted that inherent disorder within proteins can prevent crystallization by inhibiting the formation of stable crystal contacts and thereby, reduce the probability of nucleation.

Although predictive algorithms of disorder and domain boundaries are useful in providing a basis for experimental design, an analytical method that is independent of
structural prediction is necessary in the case of novel protein folds or weakly conserved structures. One of the most powerful techniques to provide protein dynamics prediction is NMR spectroscopy (Wand, 2001). A number of technical obstacles arise in applying this approach in a large-scale structural effort due to sample preparation requirements and the allowable target size. In addition, precise localization of disorder by NMR requires substantial and time-consuming data analysis which is contrary to the necessity for screening of multiple targets. Limited proteolysis coupled to mass spectrometry is another preferred approach (Cohen et al., 1995). Proteolysis, however, may clip internal loops leading to destabilization and proteolysis of structured regions. A rapid and non-disruptive method for characterizing protein flexibility with amino acid-level resolution is therefore, required.

The DXMS method (Englander et al., 2003; Hamuro et al., 2002a; Hamuro et al., 2002b; Hamuro et al., 2003; Pantazatos et al., 2004; Woods Jr., 2001; Woods Jr. and Hamuro, 2001) provides an attractive alternative to these approaches by coupling the labeling of flexible and solvent-exposed regions in the native protein with simple and sensitive detection and analysis. By using the DXMS method, we have rapidly and precisely identified regions of disorder and selectively deleted them from constructs resulting in a marked improvement in crystallization propensity.

The *Thermotoga maritima* proteome is actively being pursued as a structural genomics target by the Joint Center for Structural Genomics (JCSG). As part of this effort, screening of the entire proteome for crystallizability was undertaken (Lesley et al., 2002). While the majority of those proteins which were expressed in soluble form could be crystallized using automated nano-scale crystallization screens (Santarsiero et
al., 2002), approximately 20% of the soluble proteins did not produce any significant crystal hits in this initial attempt. Two such proteins are TM0160 and TM1171. The former is classified as a hypothetical protein without any analogous structural or functional data from homologues, whilst the latter is a transcriptional regulator with some structural homologues and clearly defined domains annotated by databases such as SCOP (Murzin et al., 1995) and Pfam (Bateman et al., 2002). These two proteins, therefore, represent two classes which are readily addressable by DXMS analysis. The first class contains novel proteins with little or no structural information available. For construct design, these proteins are typically analyzed for predicted secondary structure and for regions of low complexity from primary sequence. The second class, includes proteins for which structural information is available, but where flexibility induced by the absence of substrates/cofactors or inherent flexibility between domains makes the selection of constructs difficult or ambiguous. We describe here the first use of DXMS analysis to salvage unsuccessful crystallization targets from each of these classes and the successful outcome that resulted in high resolution crystal structures.

5.3 METHODS

Cloning and mutations.

Full-length DNA fragments encoding amino acids 1-181 of TM0160 and amino acids 1-201 of TM1171 were cloned in-frame into the expression vectors pMH2T7 and pMH1, respectively, between restriction sites Pml I and Psi I. Truncated DNA fragments encoding amino acids 1-141, 1-145, 8-141, 8-145 of TM0160 and incorporating a small 7 residue C-terminal epitope tag and amino acids 1-125, 1-129,
11-125, 11-129 of TM1171 were cloned in-frame into the expression vector pMH4 between the restriction sites Pml I and Pac I. All DNA fragments were created by PCR amplification from genomic *Thermotoga maritima* DNA (ATCC) using pfuTurbo polymerase (Stratagene). The full-length TM0160 amplicon used 5’ primer (5’-ttgaggaaggcatgggtgaa-3’) and 3’ primer (5’-actttctcttcctctcttctc-3’). The full-length TM1171 amplicon used 5’ primer (5’-gtggatctgaaaaaactgcttcc-3’) and 3’ primer (5’-gattctatcatggttcaaaggattt-3’). The four primers used for the TM0160 truncations were (1) 5’-atgaggaaggcatgggtgaa-3’, (8) 5’-actctggcgctcgatagag-3’, (141) 5’-ctcttaattaagtcgccgcaactcgatagagttctcc-3’, and (145) 5’-ctcttaattaagtcgccgtttactctccacaactcgataga-3’. The four primers used for the TM1171 truncations were (1) 5’-atggatctgaaaaaactgcttcc-3’, (11) 5’-aaagtgatcgtgttcagaaaaggt-3’, (125) 5’-ctcttaattaagtcgccggaagaagctttctcagaaatacc-3’, and (129) 5’-ctcttaattaagtcggttgggtgagggagaaggtgac-3’. All cloning junctions were confirmed by sequencing.

**Protein expression and purification.**

Full-length and truncated TM0160 and TM1171 clones were expressed in *E. coli* DL41 from plasmids based on the expression vectors pMH2T7 and pMH4, respectively. These vectors encode a 12-amino acid tag consisting of the first 6 amino acids of thioredoxin and 6 histidine residues placed at the N-terminus to enhance expression and to allow for rapid affinity purification. Protein expression was performed in a defined medium containing 150 mg/L Selenomethionine (for crystallization trials). Expression was induced by the addition of 0.15% arabinose for 3 hours, and lysozyme was added at
the end of fermentation to a final concentration of 250 µg/ml. Bacteria were lysed by sonication after a freeze-thaw procedure in Lysis Buffer (50 mM Tris pH 7.9, 50 mM NaCl, 10 mM imidazole, 0.25 mM Tris (2-carboxyethyl)phosphine hydrochloride (TCEP)), and cell debris pelleted by centrifugation at 3600 x g for 60 minutes. The soluble fraction was applied to a nickel chelate resin (Amersham Biosciences) previously equilibrated with Lysis Buffer. The resin was washed with Wash Buffer (50 mM potassium phosphate pH 7.8, 0.25 mM TCEP, 10% v/v glycerol, 0.3 M NaCl, 40 mM imidazole), and protein was eluted with Elution Buffer (20 mM Tris pH 7.9, 10% (v/v) glycerol, 0.25 mM TCEP, 300 mM imidazole). Buffer exchange was performed to remove imidazole from the eluate, and the protein in Buffer Q (20 mM Tris pH 7.9, 5% (v/v) glycerol, 0.25 mM TCEP) containing 50 mM NaCl was applied to a Resource Q column (Amersham Biosciences) previously equilibrated with the same buffer. Protein was eluted using a linear gradient of 50 to 500 mM NaCl in Buffer Q, and appropriate fractions were pooled. Protein was buffer exchanged into Crystal Buffer (20 mM Tris pH 7.9, 150 mM NaCl, 0.25 mM TCEP) and concentrated for crystallization assays to 20 mg/ml by centrifugal ultrafiltration (Millipore).

**Protein fragmentation probe maps.**

Aliquots of each protein were adjusted to a concentration of 10 mg/ml in Tris-Buffered Saline (5mM Tris, 150mM NaCl, pH 7.0; TBS), and all subsequent steps performed at 0º C, on melting ice. In a 4º C cold room, 5µl of each solution was further diluted with 15µl of TBS in a microtiter plate employing multi-channel pipetters for simultaneous manipulation. Thirty microliters of a stock “exchange quench” solution
(0.8% formic acid, 1.6 M GuHCl) were then added to each sample (final concentration
0.5% formic acid, 1.0 M GuHCl), the samples were transferred to auto-sampler vials,
and frozen on dry ice within one minute after addition of quench solution, as previously
described (Hamuro et al., 2002a; Hamuro et al., 2002b; Mathews et al., 2003; Woods-Jr.
and Hamuro, 2001). Vials with frozen samples were stored at –80º C until transferred
to the dry ice-containing sample basin of the cryogenic auto-sampler module of the
DXMS apparatus. Samples were individually melted at 0º C, then injected (45 µl) and
pumped through protease columns (0.05% TFA, 250 µl/min, 16 second exposure to
protease). Proteolysis used immobilized pepsin (66 µl column bed volume, coupled to
20AL support from PerSeptive Biosystems at 30 mg/ml). Protease-generated
fragments were collected on a C18 HPLC column, eluted by a linear acetonitrile
gradient (5 to 45 % B in 30 minutes; 50 µl/min; solvent A, 0.05% TFA; solvent B, 80%
acetonitrile, 20% water, 0.01% TFA), and the effluent directed to the mass spectrometer
with data acquisition in either MS1 profile mode or data-dependent MS2 mode. Mass
spectrometric analyses used a Thermo Finnigan LCQ electrospray ion trap type mass
spectrometer operated with capillary temperature at 200º C or an electrospray
Micromass Q-Tof mass spectrometer, as previously described (Hamuro et al., 2002a;
Hamuro et al., 2002b; Mathews et al., 2003; Woods Jr. and Hamuro, 2001). The
Sequest software program (Thermo Finnigan Inc) identified the likely sequence of the
parent peptide ions. Tentative identifications were tested with specialized DXMS data
reduction software developed in collaboration with Sierra Analytics, LLC, Modesto,
CA. This software searches MS1 data for scans containing each of the peptides, selects
scans with optimal signal-to-noise, averages the selected scans, calculates centroids of
isotopic envelopes, screens for peptide misidentification by comparing calculated and known centroids, then facilitates visual review of each averaged isotopic envelope allowing an assessment of "quality" (yield, signal/noise, resolution), and confirms or corrects the peptide identity and calculated centroid (Hamuro et al., 2002a; Hamuro et al., 2002b; Mathews et al., 2003; Woods Jr. and Hamuro, 2001).

On exchange deuteration of proteins.

After establishment of fragmentation maps for each protein, amide hydrogen exchange- deuterated samples of each of the 24 proteins were prepared and processed exactly as above, except that 5μl of each protein stock solution was diluted with 15 μl of Deuterium Oxide (D₂O), containing 5mM Tris, 150mM NaCl, pH (read) 7.0, and incubated for ten seconds at 0º C before quench and further processing. Data on the deuterated sample set were acquired in a single automated 30-hour run, and subsequent data reduction performed on the DXMS software, as previously described (Hamuro et al., 2002a; Hamuro et al., 2002b; Mathews et al., 2003; Woods Jr. and Hamuro, 2001). Corrections for loss of deuterium-label by individual fragments during DXMS analysis (after “quench”) were made through measurement of loss of deuterium from reference protein samples that had been equilibrium- exchange- deuterated under denaturing conditions, as previously described (Hamuro et al., 2002a; Hamuro et al., 2002b; Mathews et al., 2003; Woods Jr. and Hamuro, 2001). The total time elapsed for data acquisition and analysis (both fragmentation maps and deuteration study) was two weeks, and a total of 100 μgs of each protein was used to complete the study. The personnel performing the data acquisition and reduction part of the study were unaware
of the identity or crystallization histories of the proteins while data were being acquired and processed. For subsequent comparative analysis of the exchange rates of amide hydrogens within protein constructs vs. their full-length parental forms, both proteins were contemporaneously on-exchanged as above, but quenched at varying times (10 sec, 30 sec, 100 sec, 300 sec, 1000 sec, 3000 sec, 10,000 sec, and 30,000 sec), and further processed as above, employing the fragmentation maps established for the protein.

Equipment configuration.

The equipment configuration consisted of electrically actuated, high pressure switching valves (Rheodyne), connected to two position actuators from Tar Designs Inc., Pittsburgh, as described previously (Hamuro et al., 2002a; Hamuro et al., 2002b; Mathews et al., 2003; Woods Jr. and Hamuro, 2001). A highly modified Spectraphysics AS3000 autosampler, partially under external PC control, employed a robotic arm to lift the desired frozen sample from the sample well, then automatically and rapidly melted and injected under precise temperature control (Hamuro et al., 2002a; Hamuro et al., 2002b; Mathews et al., 2003; Woods Jr. and Hamuro, 2001). The auto-sampler basin was further thermally insulated and all but 20 vial positions were filled with powdered dry ice sufficient to keep samples colder than -45° C for 18 hours. Four HPLC pumps (Shimadzu LC-10AD) were operated by a Shimadzu SCL-10A pump controller. One produced forward flow over the protease columns, another back-flushed the protease pepsin column after sample digestion (0.05 % aqueous TFA), and two delivered solvents to a downstream HPLC column for gradient elution (A: 0.05 % aqueous TFA; B: 80% acetonitrile, 20% water, 0.01% TFA; 1 x 50 mm C18 Vydae #
Valves, tubing, columns and auto-sampler were contained within a refrigerator at 2.8°C, with protease and HPLC columns immersed in melting ice. The timing and sequence of operation of the DXMS apparatus fluidics were controlled by a personal computer running an in-house written LabView-based program, interfaced to solid-state relays (digital input/output boards, National Instruments), controlling pumps, valve actuators, and MS data acquisition (Hamuro et al., 2002a; Hamuro et al., 2002b; Mathews et al., 2003; Woods Jr. and Hamuro, 2001).

Crystallization

Crystals of both proteins were screened for in a 96 well, sitting drop vapor diffusion format, using 480 commercially available crystallization conditions (Hampton Research, Emerald Biostructures) at 20 and 4°C. 200nl of protein were added to an equal volume of crystallization reagent. Subsequently, 25 of the 960 conditions produced mountable crystals for TM0160, which grew from both high and low molecular weight PEGs at pH’s centered around 7.0. Those crystals used to determine the structure and collect higher resolution data were obtained with Hampton crystal screen #41: 10 % isopropanol, 20 % PEG 4000; 0.1 M HEPES pH 7.5 at 4 and 20°C, whilst mutant Cys50Ala crystals were grown from the Hampton PEG/ion screen #06: 0.2 M Sodium Chloride, 20% w/v PEG 3350, pH 6.9 at 4°C; all crystals screened were isomorphous and indexed in a primitive monoclinic crystal system. The crystals were
cryo-cooled in liquid nitrogen after adding 20% glycerol to the mother liquor as a cryoprotectant.

Three crystal forms of TM1171 were obtained. The crystal form used to determine and refine the structure belonged to spacegroup P6₁22 and was crystallized from 2.0M ammonium sulfate in a sodium citrate buffer (pH 5.5) at a temperature of 4°C. Crystals were also cooled to liquid nitrogen temperature with a cryoprotectant of 15% PEG 200. Two other crystal forms were screened for diffraction, in spacegroups I4/I4₁ and P2₁, but they diffracted poorly (to around 4.0Å) and were not used in any further experiments. Full-length TM0160 and TM1171 proteins that had been freshly expressed and purified were subjected to crystallization trials contemporaneously with the truncation constructs, and again demonstrated very poor crystallizability (data not shown).

Data Collection and Structure Solution

Data for a TM0160 SAD experiment were collected at beamline 5.0.2 of the Advanced Light Source (ALS Berkeley) to a resolution of 2.4Å, at a wavelength of 0.97635Å, corresponding to the Selenium edge as determined by an X-ray fluorescent scan (Table1). In all, 240º of data were collected using an inverse beam strategy so that Friedel mates were collected in 15º wedges close in time. Further, higher resolution, data were collected at beamline 5.0.3 of the ALS at a wavelength of 1.0Å to a maximum Bragg spacing of 1.9Å (Table1).
All data were reduced and scaled using the HKL2000 package (Otwinowski and Minor, 1997). The substructure of 4 seleniums (2 per molecule in the asymmetric unit) were found with Solve (Terwilliger and Berendzen, 1999) which was also used to derive initial phases and along with Resolve to refine the phases via solvent flattening, averaging and automated model building (Terwilliger, 1999; Terwilliger, 2001a; Terwilliger, 2001b) (Table 1), then followed by manual rebuilding and refinement with ‘O’ (Jones et al., 1991) and Refmac (Murshudov et al., 1997). After incorporation of the higher resolution data, automated water building with ARP/wARP (Lamzin and Wilson, 1993) was carried out. All other crystallographic manipulations were carried out with the CCP4 program suite (Collaborative et al., 1994). The final model has an $R_{cryst}$ and $R_{free}$ of 19.8 % and 25.3 %, respectively, with no residues in disallowed regions of the Ramachandran plot (Table 1). The C-terminus was traced to residue 150 in molecule A and 141 in molecule B. A number of regions were disordered and did not have significant electron density; 11 residues at the N-terminus could not be interpreted, constituting all but one of the N-terminal tag residues, as well as loop regions between 107 and 115 in molecule A and 108 to 113 in molecule B. All of these regions corresponded to regions of high mobility in DXMS.

The Cys50Ala mutant data were collected on an in-house RUH3R (Rigaku, MSC) source incorporating Osmic mirrors and a RaxisIV4++ image plate detector to a resolution of 2.8Å. As the crystal was isomorphous to the wild type, after modifying the mutated cysteine residues the model was positioned into the crystal by rigid body refinement with Refmac5 (Murshudov et al., 1997) and followed by two rounds of refinement and manual model building with ‘O’ (Jones et al., 1991) (Table1)
Data for TM1171 were also collected at beam-line 5.02 of the ALS to a resolution of 2.4 Å at a wavelength of 0.97972 Å corresponding to the Selenium edge as determined by an X-ray fluorescent scan. 120° of data were collected using an inverse beam strategy collecting wedges of 10° close in time. Data were processed and the structure determined by similar procedures to that of TM0160 (Table 1). The resultant structure had excellent stereo-chemical properties with all residues in favored regions of the Ramachandran plot; the final $R_{\text{cryst}}$ and $R_{\text{free}}$ for the model converged at 19.7% and 25.3, respectively (Table 1). With the exception of the twelve residues of the N-terminal tag, all residues of the construct could be traced in the electron density map (Table 1).

Coordinates for TM0160 wild type have been deposited in the PDB database as 1O5Y and the TM0160 Cys mutant as PDB entry 1SJ5 and TM1171 entry as 1O5L.

5.4 RESULTS

Domain definition in the absence of structural information.

Deuterium exchange maps were generated initially for the full-length TM0160 and TM1171 proteins (Chapter 4 Figure 4-4). This initial mapping was performed with a 10 second labeling reaction which was previously demonstrated to be sufficient to allow identification of rapidly exchanging and, therefore, likely disordered regions (Pantazatos et al., 2004). The TM0160 map indicates that a region of rapid exchange is located in the C-terminus of the protein (residues 146-156 and 163-175). The amino acid complexity of this region is somewhat low with significant stretches of acidic amino acids. Sequence alignments with 16 of the closest sequence homologues identified three regions of completely conserved amino acids at positions 31-34, 54-61
and 112-127 (Figure 5-1B). Then sequence conservation decreases substantially from residue 134 also corresponding to the region of increased exchange-rate (Figure 5-1A and B). The peptide fragmentation map used to identify rapid-exchange sites also indicated a preferential proteolytic cleavage at residue 141 by the relatively non-sequence specific protease pepsin. Combining the exchange, sequence alignment, and proteolysis information, we chose position 145 to define the C-terminus of our TM0160 daughter construct. The N-terminus was left intact as there was a general absence of DXMS data for this region. This could indicate that this region is particularly sensitive to proteolysis; however, this region was visible in the final electron density map and appears to be well-ordered.

The coding region from positions 1 to 145 was cloned and expressed. The resulting purified protein was re-evaluated by DXMS to determine if there were any substantial changes in the exchange pattern indicative of any gross structural changes as a result of the truncation. Parent TM0160, and its daughter truncation, were on-exchanged variously for 10, 100, 1,000, and 10,000 seconds at 0 °C. The exchange pattern for both the parental full-length TM0160 and the daughter construct are virtually identical in the homologous regions (Pantazatos et al., 2004). Furthermore, both parental TM0160 and the daughter construct behaved as dimers when evaluated by size-exclusion chromatography (data not shown). These results indicate that the DXMS-defined deletion appears to be properly folded.

Unlike TM0160, TM1171 has homologues with known three-dimensional structures and its domain definitions (Figure 5-1C), using the Pfam database (Bateman et al., 2002), enable the sequence to be split into two subdomains: a cyclic-nucleotide
binding domain (residues 17-111) and a bacterial transcriptional regulatory Crp domains (165-196) (binding DNA via a Helix-turn-Helix HTH motif). The DXMS data predict substantial disorder in the region linking the nucleotide binding domain to the Crp, helix-turn-helix DNA binding domain based on sequence alignments Figure 5-1B. This disorder may be suggestive of inter-domain flexibility between the DNA and nucleotide binding domains. Such flexibility may disappear upon binding to a regulatory sequence or may allow interaction with RNA polymerase.

TM0160 and TM1171 deletion constructs show marked improvement in crystallization efficiency.

The TM0160 full length parent has been extensively evaluated for crystallization. Despite multiple screening attempts of 480 crystallization conditions, only 3 marginal hits were obtained from 2400 individual crystallization tests. By contrast, for the TM0160 deletion mutant, 78 hits were obtained from 1920 individual tests including numerous crystals of sufficient size and quality for diffraction studies. An almost identical result was experienced with TM1171 where only 5 marginal crystal hits were observed from 2400 individual crystallization tests. The DXMS guided construct produced 3 different crystal forms from 19 crystallization conditions that resulted in mountable crystals (Pantazatos et al., 2004).

The structure of TM0160

The dimer of TM0160 forms a wedge, each monomer being of basic triangular shape of size 70Å x 40Åx40Å (Figure 5-2B and C). From a Dali search (Holm and
Sander, 1993), no significant matches were found for TM0160 suggesting that it possesses a new fold. The topology diagram of the protein (Bond, 2003; Westhead et al., 1999) is shown in Figure 5-2A. The monomer is composed of an eight-stranded, distorted β sheet consisting of a four-stranded, anti-parallel β sheet (B1,B2,B3,B8), and a four-stranded mixed β sheet (B4,B5,B6,B7). The sheets are intercalated by 3 short α helices (H1,H3, H4), whilst a longer 11 residue α helix (H2) forms the central core of the dimer interface. A short helix (H5) at the C-terminus of monomer A is formed largely from the C-terminal epitope tag and marks the beginning of the highly flexible C-terminus removed from the wild type protein (Figure 5-1A,B). This helix is somewhat stabilized by crystal contacts absent from its equivalent location in molecule B.

**Interchain disulphide and binding interface**

Of interest is the interchain disulfide bridge between the two units in the dimer. In nature, the reducing environment of *T. maritima* would not seem to allow this arrangement. The monomers occlude a surface area of approximately 2400 Å$^2$ (calculated using the Lee and Richards algorithm (Richards, 1977) with a probe radius of 1.4 Å) upon binding, which is one quarter of the surface area of each individual monomer. The binding interface itself is primarily formed around the molecular two fold axis from three leucine residues and one valine residue. This, combined with the lack of conservation of this cysteine residue in related sequences (Figure 5-1B), suggests that this disulfide may have arisen by genetic drift. However, a recent study has suggested that disulfide bonds may be much more common than expected for some
prokaryotic microbes (Mallick et al., 2002). To investigate this phenomenon a Cys50-Ala Mutant was constructed (see Materials and Methods). Crystals were obtained from essentially the same conditions and a data set was collected to 2.9Å on an in-house rotating anode source. The crystals were isomorphous and the resultant dimer structure was essentially identical apart from the missing sulfur atoms at the position of the disulfide bridge. This would suggest that the covalent interaction is not necessary for the formation of the TM0160 dimer, but does not exclude the possibility that it provides additional stability to the oligomer of the thermophilic protein.

Genomic Information

TM0160 is located in a region of the chromosome containing several proteins of unknown function. However, one proximal gene, TM0161 is annotated as a geranyl transferase enzyme (Nelson KE et al., 1999). An examination of the sterol biosynthesis pathway for *T. maritima* indicates that many of the enzymatic activities surrounding geranyl transferase do not have gene assignments. We evaluated the neighboring genes for TM0160 homologs from 19 other genomes. Annotated activities from proximal genes which are potentially co-transcribed included: oxido reductases (Nostoc, Thermosynechococcus), endonuclease III (Aquifex), DNA helicase II (Synechocystis), glycine dehydrogenase (Mycobacterium), glycine cleavage P (Mycobacterium), uncharacterized ACR cofactors (Thermoanaerobacter), phosphoribosyl AMP cyclohydrolase and imidizole glycerol-phosphate synthase (Halobacterium), protein-L-isoaspartate (D-aspartate) O-methyltransferase (Methanocarcina), ribose-5-phosphate isomerase and glutamate-1-semialdehyde 2,1, aminomutase. The preponderance of
enzymes involved in amino acid metabolism may indicate a putative role in this process for TM0160.

DXMS analysis shows the C-terminal region of full-length TM0160 to be disordered which may account for its apparent interference in crystallization. One potential reason for this disorder is the lack of a protein binding partner. We attempted to identify such a potential interaction through a two-hybrid protein interaction screen (Fields et al., 1999). Full-length, truncated, as well as the deleted C-terminus were evaluated; in each case, the fusion constructs demonstrated self-activation in the two-hybrid screen and could not be pursued for novel interactions.

Putative Active Site

In an attempt to locate similar active site geometries in the protein, a rigorous search was performed of all clusters of 3 and 4 putative active site residues in the TM0160 dimer. A cluster of active residues was defined as the subset of all non-hydrophobic residues grouped within 15Å of each other. The coordinates of the putative “active sites” were then submitted to SPASM (Kleywegt, 1999). Of the 1849, three-residue combinations and 2090 four-residue combinations searched, none produced any hits reminiscent of a known active site. Submission of the coordinates to the PINTS server (Stark et al., 2003) also produced no hits of any significance. This analysis suggests that, if TM0160 is an enzyme, then it will likely possess a novel enzymatic activity and mechanism.

The TM0160 structure, when combined with a sequence alignment of homologous sequences, however, can give considerable insight into the possible
location of its active site (Figure 5-1B, 5-3A and B). A high degree of sequence conservation (Figure 5-3A) occurs around a large groove situated at the thick end of the wedge, which represents the largest cavity in the molecule. This area is centered around the molecular two fold axis, which may in part account for its sequence conservation. However, it also extends far into the pocket suggesting evolutionary conservation independent of the formation of a dimeric structure. This pocket contains some unaccounted for electron density, too ambiguous to trace but clearly not a network of water molecules. Of particular note in the pocket is His58 centered around the two fold which is coupled to Asp 115 via a possible proton shuttling mechanism allowing the histidine to co-ordinate a putative water molecule (Figure 5-3B). This sort of chemistry may indicate a region of possible active site chemistry. The only other potential proton donor would be Thr 57 from the other subunit in the dimer which is a highly conserved residue (5-1B). In cases where a substitution of this residue occurs, it is most often replaced with an equally viable serine residue (Figure 5-1B), which could then produce a putative catalytic triad.

This putative active site may also be indicated by the DXMS data, where Leu56 and Thr57 are indicated as regions of high exchange, indicating considerable solvent accessibility, that could also be indicative of an active site, as seen for another T. maritima protein (TM0449) where ligand binding stabilizes the active site (Mathews et al., 2003; Pantazatos et al., 2004).

The structure of TM1171 cNTP domain
TM1171 belongs to the CRP family and is believed to be a transcriptional regulator. As representatives of this family have been previously determined, the structure was not expected to have a novel fold, but was distant enough from other sequence homologues (highest sequence identity 19%, PDB ID, 1O3S) to expect that structure determination would be more successful by MAD/SAD. In other members of the CRP family, the structure consists of two domains: a cyclic nucleotide binding domain (cNTP), situated at the N-terminus, responsible for dimerization and binding cNTP’s and a C terminal helix-turn-helix (HTH_CRP) cAMP regulatory domain responsible for DNA binding. The connection between the two domains is defined by a long α helix (20-30 residue) which could be assigned to either of the two domains but in itself is structural disparate to each. The fold of the truncated version of TM1171 consists of two four stranded anti-parallel β sheets (B1,B8, B3, B6, and B2, B7,B4,B5) forming a jelly-roll sandwich topology (Figure 5-4A and B) and is classified as a double-stranded β-helix by SCOP (Murzin et al., 1995). This sandwich is terminated by two C-terminal α helices, the latter being a 25 residue helix which forms the dimerization interface for the molecule (H4). Two other helical turns are formed, the first being a 5- residue helix at the N-terminus (H1) and the second a 3-residue alpha helical turn bridging β strands 6 and 7 (H2) (Figure 5-4A and B). In comparison to its two closest structural homologues, the E. coli catabolite gene activator (Parkinson et al., 1996) (PDB code 1RUN) and Listeria monocytogenes Listeriolysin regulatory protein (PDB code1OMI ) TM1171 has a root mean square deviation of 1.74 and 1.92 Å on 111 and 97 aligned Cα atoms, respectively (Calculation performed with STAMP (Russell and Barton, 1992)).
TM1171 Dimer

As a putative transcription regulator TM1171 is expected to bind to a specific sequence of DNA as a dimer, through its two C-terminal domains that are absent in the truncated TM1171 and are characteristic of the CRP family (Parkinson et al., 1996) (Figure 5-5). The dimer interface is provided by the interaction of the two-fold symmetric H4 helices (Figure 5-4). Upon binding, the interface occludes a surface area of 3708 Å² (calculated using the Lee and Richards algorithm and a probe radius of 1.4 Å (Richards, 1977)) representing 28 % of the available surface area and is formed from a cluster of 9 hydrophobic residues pairing up with their equivalent counterparts around the molecular/crystallographic two fold axis.

cNTP binding region

The cyclic nucleotide binding site, situated between the two β sheets, helix 2 and helix 4, is structurally conserved relative to homologous structures (Figure 5-1B). The long helix (H4) is rotated by approximately 20º in TM1171 relative to the other C-NTP structures that contain both domains (Figure 5-6). However, superposition of TM1171 with other cAMP binding proteins co-crystallized with bound cAMP shows that a movement of the loop containing residues 63-66 occludes the volume occupied by cAMP in the other crystal structures suggesting that in TM1171 either cAMP binds in a different conformation or the binding of cAMP is accompanied by a conformational change. Some residual electron density is present in the TM1171 electron density maps indicating the acquisition of a bound nucleotide during expression. The exact identity
of this electron density could not be unambiguously assigned due to its poor quality and therefore the nucleotide was not modeled.

Interpreting the TM1171 DXMS data in light of the structure

TM1171 is involved in DNA binding, the primary interactions for which reside in the C-terminal domain, while the N-terminal domain is responsible for binding cAMP or other cyclic nucleotides. The long coiled-coil helix between the two domains forms the dimer interface. It is interesting to note that the closest sequence homologues to TM1171 (Figure 5) were crystallized in the presence of DNA which presumably stabilize the dimer by bridging the two monomers. The exception is the structure from *Listeria monocytogenes*, (1OMI [PDB]) where the entire protein was crystallized and the structure determined without a DNA substrate, although the average B-values were relatively high ~78.0Å².

Combining these homologue structural data with the DXMS data suggests that the C-terminal Crp HTH domain is intrinsically flexible relative to the N-terminal domain until stabilized by the binding of a specific DNA sequence. It seems reasonable that the presence of DNA decreases the number of conformational degrees of freedom between the two domains thus increasing the chance of forming a crystal lattice (Figure 5-5).

Comparison with SEG Analysis

It is important to compare the DXMS experimental results with those obtained by primary sequence computational analysis. We used the SEG program to look at low
complexity regions of the primary sequence (Wootton and Federhen, 1993). From analysis of the sequences presented in Figures 5-1B and 5-1C, the low complexity regions have been shaded red in sequence 2, whilst the DXMS comparisons are shaded green in sequence 1. For TM1171, the regions of low complexity given by SEG represent the loop regions between the turn regions of the penultimate helix and the long C-terminal long helix residues 94-109, that forms the dimer interface. This region connects the dimerization helix (H5) to domain 1 and is ordered in the crystal as indicated by the electron density. This would suggest that, if SEG was utilized in the absence of structural homology information in preparing the constructs, the designed domain would be smaller and possibly more compact, but would remove the dimerization helix. The DXMS analysis indicated that the region to initiate the cut would be the loop connecting the dimerization helix and domain 2 of the molecule (Figure 5-5), a region likely to be flexible in the absence of its DNA substrate.

On the other hand, the computational prediction from TM0160 is relatively accurate (Figure 5-1B) where SEG predicts that there is a disordered region in the C-terminus but places the start at residue 169, rather than the residue (162) that DXMS predicts (Figure 5-1B). SEG also suggests the position of the disordered loop 106-112 which exhibits no discernable electron density, and indicates that the start of this region is only one residue from that suggested by DXMS (Figure 5-1B).

5.5 DISCUSSION

DXMS provides an experimental means to analyze local protein flexibility and a specific means to design more “crystallizable constructs”. Here two proteins which, in
their full length states, were resistant to crystallization attempts are used to demonstrate the DXMS utility. The first, TM0160 is a novel fold and was truncated at its C-terminus to yield viable crystals. The second, TM1171 is a transcriptional regulator protein which probably requires its DNA substrate to form a stable structure. The designed construct for TM1171 excised a sub-domain from the C-terminus which would probably inhibit crystallization. These results demonstrate that DXMS can provide a simple and rapid means to give meaningful data as to where to terminate/separate domains to provide more stable and ordered constructs, in cases where little is known of the protein structure or function.

The structure of TM0160 unravels reveals another unique fold which displays a bacterial inter-chain disulfide bond which are now being found in other examples of bacterial proteins (Mallick et al., 2002). The structure has not revealed the exact function of the gene primarily because so little is known about the host organism and this protein or its homologues. The position of a putative active site can nevertheless be proposed from conserved residues in homologous family members, some unaccountable electron density in the large putative binding cavity that contains residues that could exhibit some interesting chemistry, such as protease activity. The DXMS technique itself may also lend itself to predicting areas of ligand binding. Although the exact position and function of the protein’s active site will only be unambiguously determined by experimental verification, which is now ongoing, this approach has narrowed down the search.
Chapter 5 is a reprint of the material as it appears in *Protein Sci*, 2004, Spraggon, G., D.P. Pantazatos, H.E. Klock, I.A. Wilson, V.L. Woods, Jr. I was one of the primary researchers, and was responsible for experimental design, execution, and interpretation of all DXMS data in the work. The co-authors listed in this publication either directed and supervised the research which forms the basis for this chapter, or assisted in data acquisition/analysis and manuscript writing. We thank Walter Englander, David Wemmer, and Philip Bourne for their support and guidance in this DXMS application, Dan McMullan, Kevin Rodrigues, Juli Vincent and Eileen Ambing for protein purification and crystallization studies and Peter Schultz for continued support. Data were collected at the Advanced Light Source at the Lawrence Berkeley National laboratory on beamlines 5.0.2 and 5.0.3 we would like to thank all of the staff of these beamlines for their continued support. This work was supported by NIH grants CA099835 (VLW) and by NIH Protein Structure Initiative grant GM-99-009 from the National Institute of General Medical Sciences (www.nigms.nih.gov), GM 062411 (IAW), and by grants from the University of California BioStar and LSI programs, grants S97-90, S99-44, L98-30 (VLW), with the matching corporate sponsor for these grants being ExSAR Corporation, Monmouth Junction, NJ. VLW has an equity interest in ExSAR Corporation.
### TABLE 4-2: Summary of Data Collection and Refinement Statistics for TM0160 and TM1171

<table>
<thead>
<tr>
<th>Protein</th>
<th>TM0160 (phasing)</th>
<th>TM0160 (refining)</th>
<th>TM0160 (C38A)</th>
<th>TM1171 (SeMet)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Space Group</td>
<td>P2₁</td>
<td>P2₁</td>
<td>P2₁</td>
<td>P6₁22</td>
</tr>
<tr>
<td>Unit Cell Parameters (Å)</td>
<td>a=43.82Å, b=51.87Å, c=73.62Å, β=97.31°</td>
<td>a=43.51Å, b=51.07Å c=73.97Å, β=97.39°</td>
<td>a=44.00Å, b=52.14Å c=73.62Å, β=97.64°</td>
<td>a=b=62.62Å, c=166.78Å</td>
</tr>
<tr>
<td>Wavelength (Å)</td>
<td>0.97635</td>
<td>1.0</td>
<td>1.54</td>
<td>0.97972</td>
</tr>
<tr>
<td>Resolution Range (Å)</td>
<td>50.0-2.3</td>
<td>50.0-1.9</td>
<td>50.0-2.8</td>
<td>50.0-2.3</td>
</tr>
<tr>
<td>R_{sym} (in highest resolution shell)</td>
<td>0.046(0.3.2)</td>
<td>0.067(0.34)</td>
<td>0.079(0.65)</td>
<td>0.082(0.32)</td>
</tr>
<tr>
<td>No. Unique Refs (observed)</td>
<td>19573 (351594)</td>
<td>24410 (398636)</td>
<td>8171 (24075)</td>
<td>9367 (73376)</td>
</tr>
<tr>
<td>Completeness (%) (Highest shell)</td>
<td>98.1(95.7)</td>
<td>95.0 (92.4)</td>
<td>98.2(98.9)</td>
<td>99.4(99.2)</td>
</tr>
<tr>
<td>Highest Resolution Shell (Å)</td>
<td>2.43-2.31</td>
<td>1.97-1.9</td>
<td>2.9-2.8</td>
<td>2.43-2.31</td>
</tr>
<tr>
<td>Mean I/σ(I)</td>
<td>25.2 (2.9)</td>
<td>19.6(1.5)</td>
<td>12.0(3.1)</td>
<td>22.3(4.2)</td>
</tr>
<tr>
<td>No. of Se sites</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>Model and Refinement Statistics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of reflections (total)</td>
<td>-</td>
<td>23110</td>
<td>7762</td>
<td>8737</td>
</tr>
<tr>
<td>No of reflections (test)</td>
<td>-</td>
<td>1214</td>
<td>378</td>
<td>443</td>
</tr>
<tr>
<td>R_{cryst} (R_{free} *) †</td>
<td>-</td>
<td>0.198 (0.253)</td>
<td>0.238(0.295)</td>
<td>0.197(0.253)</td>
</tr>
</tbody>
</table>
TABLE 4-2: continued

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No. protein atoms</td>
<td>-</td>
<td>2175</td>
<td>2173</td>
<td>2135</td>
</tr>
<tr>
<td>No. water atoms</td>
<td>-</td>
<td>242</td>
<td>0</td>
<td>252</td>
</tr>
<tr>
<td>Stereochemical Parameters</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rmsd bonds Å</td>
<td>-</td>
<td>0.019</td>
<td>0.017</td>
<td>0.017</td>
</tr>
<tr>
<td>rmsd angles º</td>
<td>-</td>
<td>1.67</td>
<td>1.57</td>
<td>1.503</td>
</tr>
<tr>
<td>Average isotropic B-</td>
<td>-</td>
<td>41.4</td>
<td>24.8</td>
<td>47.4</td>
</tr>
<tr>
<td>value (Å²)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ESU based on R&lt;sub&gt;free&lt;/sub&gt; (Å)§</td>
<td>0.157</td>
<td>0.448</td>
<td>0.215</td>
<td></td>
</tr>
</tbody>
</table>

† $R_{\text{factor}} = \frac{\sum |I_i - \langle I_i \rangle|}{\sum |I_i|}$ where $I_i$ is the scaled intensity of the $i$th measurement, and $\langle I_i \rangle$ is the mean intensity for that reflection.

* $R_{\text{free}}$ = as for $R_{\text{cryst}}$, but for 5.0% of the total reflections chosen at random and omitted from refinement.

§ Estimated overall coordinate error (Ottinowski and Minor, 1997; Tickle et al., 1998)
Figure 5-1A: DXMS time exchange data and sequence alignments for TM0160 and TM1171. Amino acid sequences for truncated TM0160 and TM1171 constructs. Amino acids indicated to be flexible by DXMS are shaded in cyan, those indicated by SEG are shaded in yellow, and those predicted to be disordered by both programs are shaded green. Residues removed from the wild-type sequence are colored red; the beginning and the end of the excision is labeled with a left and a right arrow, respectively. (Figure taken from Spraggon, G., D.P. Pantazatos et al. 2004)
Figure 5-1B: Sequence Alignment of TM1060 and its closest homologues. Figure produced by T_COFFEE (Notredame et al., 2000). Sequence 1 and 2 represent TM0160 the green regions of sequence 1 define the positions of proposed flexible regions determined by DXMS analysis whilst the red regions of sequence 2 defining those predicted by SEG. Secondary structure elements are defined below, α-helices represented as blue tubes and β-strands as magenta arrows. Absolutely conserved regions are shaded yellow. Alignment resulted with (e score (<1.e-10). (Figure taken from Spraggon, G., D.P. Pantazatos et al. 2004.).)
Figure 5-1C: Structural alignment of TM1171 and its two closest structural homologues. The sequence of the two homologs, L. monocytogenes (1OMI [PDB]) and E. coli (1RUN [PDB]) are numbered with reference to TM1171. Conserved structural regions are contained within boxes; conserved hydrophobic residues are masked in yellow, while totally conserved residues are shaded grey. Residues beyond 126 that were not in the construct are colored red and are aligned only by sequence. Secondary structure elements are defined as in Figure 5-1B. Figures were produced with STAMP (Russell and Barton 1992 and ALSCRIPT (Barton 1993).
**Figure 5-2A,C: Structure of TM0160.** (A) Topology diagram of the overall fold of TM0160. The long mixed $\beta$-sheet is shaded cyan. The dimerization helix (H2) is shaded yellow, while the highly mobile C-terminal epitope tag helix (H5) in molecule A is shaded red; all other helices are shaded green. The picture was generated by TOPS (Westhead et al. 1999) and Topdraw (Bond 2003). (C) Two orthogonal ribbon diagram representation of the TM0160 dimer. The interchain disulfide is depicted in a ball-and-stick representation and sits on the molecular twofold displayed as an arrow in the top diagram and as an oval in the bottom. The ribbon is colored from blue to green in subunit A and green to red in subunit B. The figure was generated using Bobscript (Kraulis 1991; Esnouf 1997) and Raster3d (Meritt and Murphy 1994). (Figure taken from Spraggon, G., D.P. Pantazatos et al. 2004)
Figure 5-2B,D: Structure of TM0160. (B) Stereo diagram of the TM0160 monomer generated by VMD (Humphrey et al. 1996). C\& atom numbering is every 20 residues. (D) Representative 2Fo-Fc electron density. The electron density of the region around the molecular twofold axis details the interchain disulfide bond. The electron density map is contoured at 1.5 standard deviations above the mean. (Figure taken from Spraggon, G., D.P. Pantazatos et al. 2004)
Figure 5-3: Putative active site region for TM0160. Regions of residue conservation as determined by the sequence alignment in Figure 1B. Residues are colored from red representing 0% conservation to green at 100%. Those residues that are 100% conserved are also displayed as a ball-and-stick representation. Conservation is calculated as the percent of conserved residues among the 17 sequences displayed in Figure 1B. The figure was generated with Bobscript (Kraulis 1991; Esnouf 1997) and Raster3d (Meritt and Murphy 1994). (B) Close-up of putative active site region of TM0160 as defined in A. Molecule A is colored cyan, while molecule B is colored yellow. The putative water molecules coordinated to His 58 are colored red. Interactions <4.0 Å are represented as dashed black lines. The picture was generated by Bobscript (Kraulis 1991; Esnouf 1997) and Raster3d (Meritt and Murphy 1994) (Figure taken from Spraggon, G., D.P. Pantazatos et al. 2004)
Figure 5-4: Overall Structure of TM1171. (A) Ribbon representation of the dimer of TM1171 cNTP domain. Molecule A is colored from blue to green from N to C-terminus, whilst molecule B is represented from green to red over the same range. The dimer was produced by rotating one subunit around a crystallographic two fold axis, which is represented by an arrow. (B) Stereo trace of TM1171 dimer. Residues are labeled every 20 residues. Figures generated with VMD (Humphrey et al., 1996).
Figure 5-5: Comparison of TM1171 with *E. coli* transcription regulator. Ribbon diagram of *E. coli* transcription regulator in complex with its DNA substrate (Parkinson et al. 1996). (Figure taken from Spraggon, G., D.P. Pantazatos et al. 2004)
Figure 5-6: Superposition of TM1171 cNTP domain with its counterpart in E. coli (PDB code 1RUN [PDB]). TM1171 is colored red and 1RUN [PDB] is colored yellow. The overall rmsd between the two domains is 1.74 Å over 111 aligned Cα residues; the dimerization helix is rotated relative to its counterpart in 1RUN [PDB] by about 20° (Figure taken from Spraggon, G., D.P. Pantazatos et al. 2004)
5.7 REFERENCES


CHAPTER 6

METHODS FOR THE DETERMINATION OF PROTEIN THREE-DIMENSIONAL STRUCTURE EMPLOYING HYDROGEN EXCHANGE ANALYSIS TO REFINE COMPUTATIONAL STRUCTURE PREDICTION.

6.1 ABSTRACT

We propose a new approach to protein 3-D structure determination that combines computational prediction methods with high-quality model-constraining information provided by amide hydrogen-deuterium exchange mass spectroscopy. If successful, this new approach will significantly accelerate the pace of structure elucidation, a timely result with genomics poised to present us with dramatically increased numbers of human disease-relevant targets for structure determination.

Considerable evidence has established that the exchange rates of the peptide amide bond hydrogens within proteins are exquisitely dependent upon protein structure and thermodynamic stability. Thus peptide amide hydrogen exchange rate measurements are well-suited to provide constraints for protein structure prediction. We (Dennis Pantazatos, Dr. Virgil Woods, UCSD) have developed a number of enhancements to amide hydrogen/deuterium exchange-mass spectrometry
(which we term DXMS) that allow the exchange rates of all of a protein's peptide amide hydrogens (its exchange rate "fingerprint") to be determined in days. Furthermore, one of my mentors, Dr. Vincent Hilser, (University of Texas, Galveston) pioneered the development of computational approaches (the COREX algorithm) capable of reliably calculating amide hydrogen exchange rate fingerprints from actual or presumed protein structure(s). Additionally, we can make structural predictions for proteins employing the Rosetta algorithm of Dr. David Baker (University of Washington) another of my mentors. Rosetta is widely considered one of the most successful ab-initio predictive methods available. Also, my mentor Dr. Halbert White (UCSD) has expertise in advanced methods for correlative analysis of patterns within “noisy” data. Finally, my mentor Phillip Bourne (UCSD) has expertise in the utilization of supercomputing platforms, protein structure determination, and large-scale data integration, most particularly in the context of the Protein data Bank (PDB).

From these elements a method for refining structure prediction is made possible by combining exchange rate measurements made with small quantities of a protein (using DXMS), with the ability to calculate exchange rates of amides within proposed structures for the protein (using COREX):

A. Multiple structures (100-1000) are predicted for a target protein using any of a variety of individual methods or multiple methods, including Rosetta.

B. COREX, run on supercomputing platforms, is employed to calculate the exchange rate that each of the protein's several peptide amides would exhibit in each proposed structure, establishing each prediction's virtual exchange rate "fingerprint".
C. Microgram quantities of the protein are produced and studied by DXMS technology to measure amide hydrogen exchange rates, establishing the protein's true exchange rate fingerprint.

D. The predicted structures that have COREX- predicted rate fingerprints that most closely match the true fingerprint, determined by DXMS experiment, are identified and discerned from structure predictions with less accurately -matching fingerprints. We term this approach the DXMS –Rosetta-COREX filter.

My studies to date have been directed at developing and implementing the several components of the method, and testing its overall feasibility. I have had much success in the application of large scale supercomputing to the generation of Rosetta structural predictions, COREX calculations of exchange rate fingerprints from actual and virtual (Rosetta) structures on a large scale, and in the development of algorithms and methods of data acquisition that generate high- quality exchange rate fingerprints from DXMS experimental data (HR- DXMS).

We presently can produce remarkably accurate predictions that are at the threshold where prediction accuracy variance becomes highly correlated with hydrogen exchange rate “fingerprint” variance. Further studies, outlined in Chapter 7, Conclusions and Future Directions, will likely move prediction accuracy into the correlated regime, thus rendering the structure determination method operable.
6.2 INTRODUCTION.

Three-dimensional determination of protein structure is required for a fundamental understanding of how proteins participate in human disease, and can provide an effective guide to the rational design of therapeutics to clinically important targets. The pressing need for this information contrasts with the agonizingly slow pace of present high-resolution structure determination methods. To address this shortcoming we propose a new approach to protein structure that combines computational predictive methods with experimentally determined constraints of exceptional utility: peptide amide hydrogen/deuterium exchange rate experimental data acquired by advanced mass spectrometric techniques (DXMS) we have pioneered. We have devised a simple approach in which such constraints are rapidly acquired and applied to the refinement of the output of virtually any protein structure predictive method. We have found the Rosetta structure prediction method of David Baker to be exceptionally well-suited to this approach.

**What is the status of computational structure prediction?**

The prediction of the 3-D structure of a protein from protein sequence remains a holy grail of molecular biology. A key issue not known a priori, of course, is whether the target (the name we use for the sequence of unknown structure) has a new fold or is close to an existing fold. If it takes one of the existing folds, which is the most suitable fold among the known folds (fold recognition)? If the sequence is homologous (over 30% sequence identity) to an existing structure can we use comparative (homology)
modeling to define a template and can we align the target to the template? Alternatively what if no relationship to a known structure exists, can the structure be modeled from first principles (ab initio)? These are weighty questions indeed. The international structural genomics projects will begin to address these questions by providing good templates for the majority of protein fold space (there will be gaps because of the present intractability of certain protein families, for example, membrane proteins). If it is this simple why do we need this proposal for experimentally augmenting the prediction process, but rather wait until fold space is more completely defined? The answer is that even with a better coverage of fold space we will not solve many of the ambiguities that exist in selecting the correct fold and aligning the template to that fold.

Perhaps the best measures of current success are afforded through the Critical Assessment of Structure Prediction (CASP) experiments. The targets for these double-blind assessments of protein structure prediction methods are NMR and X-ray protein structures comprising one or more domains either determined and not published or anticipated to be determined in time for review. Groups competing in CASP experiments make a series of predictions of the targets' 3-D structures based on the protein sequence and submit those results to a server for independent and comparative review. The CASP experiments were run in 1994, 1996, 1998 and 2000, with CASP5 run in the summer of 2002.

Progress through CASP4 is summarized by Venclovas et.al and is best described as incremental (3). In CASP 4 and 5 the Rosetta algorithm of David Baker was often the most successful method for de novo protein structure prediction(4-6). Using only primary sequence information, successful Rosetta predictions yield models with typical
accuracies of 3-6Å Ca RMSD to the experimentally determined structures for substantial segments of 60 or more residues. In such low to moderate accuracy models of protein structure, the global topology is correctly predicted, the location and arrangement of secondary structure elements is generally correct, and functional residues are frequently clustered to an active site region. An extension of the Rosetta method to incorporate limited NMR-derived experimental constraints generally yields structures of higher overall accuracy, often 2-3 Å RMSD over the entire protein (7, 8).

**Peptide Amide Hydrogen Exchange Rates Are Exquisitely Dependent Upon Protein Structure.** Peptide amide hydrogen-exchange techniques have proven to be increasingly powerful methods by which protein dynamics, domain structure, regional stability and function can be studied (9-12). Peptide amide hydrogens are not permanently attached to a protein, but continuously and reversibly interchange with hydrogen present in water. The chemical mechanisms of the exchange reactions are understood, and several well-defined factors can profoundly alter exchange rates, most particularly the extent to which a particular exchangeable hydrogen is exposed (accessible) to water (9, 13-15). The exchange reaction proceeds efficiently only when a particular peptide amide hydrogen is fully exposed to solvent.

**Fully solvated amides have very fast exchange rates.** Peptide amide hydrogens that are freely accessible to water exchange at their maximal possible rate, with an average half-life of exchange of approximately one second at 0 °C and pH 7.0. (16, 17). The precise rate of exchange of a particular fully-solvated amide can vary more than thirty-
fold from this average rate, depending upon the identity of the two amino acids flanking
the amide bond. Exact exchange rates expected for fully solvent-exposed amide
hydrogens can be reliably calculated from knowledge of the temperature, pH, and
primary amino acid sequence involved. (16, 17). In the course of our hydrogen
exchange studies (discussed below), proteins are “exchange-quenched” and then
extensively proteolyzed. The amides in the resulting random-coil peptides are
maximally solvated, and their exact exchange rates (under these quench conditions) can
be calculated (16, 17). We utilize this knowledge of each amide’s exchange rate under
quench conditions to correct for losses of exchanged deuterium from the peptides
during liquid chromatography-mass spectrometry (lc-ms) processing and, most
importantly, to measure exchange rates within regions of the protein not sufficiently
resolved by overlapping fragmentation alone.

When all amides in a substantial stretch of primary sequence exchange at the
maximal rate, it usually indicates that the sequence is unstructured (1). In Chapter 4 we
have described the use of DXMS to quickly identify and localize such rapidly
exchanging unstructured stretches of sequence within otherwise well-structured
proteins, and have demonstrated that truncated protein constructs depleted of such
unstructured regions exhibit superior crystallization properties (1). We have recently
recognized that the same method also rapidly identifies and localizes the fully-solvated
amides of amino acids that are on the protein surface in structured regions, and that this
information can be used to dramatically refine structural predictions in a high
throughput manner (1). Future studies will implement and evaluate our ability to
improve aggregate Rosetta structural predictions with this information, prior to the use of the predictions in the DAMS- Rosetta- COREX method.

**Amides sequestered from solvent have slowed exchange rates.** In a structured protein, most peptide amide hydrogens exchange slower (up to \(10^9\)-fold slower) than the maximal, fully solvated exchange rate, as they are not efficiently exposed to solvent water most of the time, the single exception being the set of structured, but well-solvated amides on the surface of a protein. As further discussed below, protein structure is not static, but best considered as an ensemble of transiently unfolded states-the native state ensemble. Amide hydrogen exchange occurs only when a particular transient unfolding event fully exposes an amide to solvent. The ratio of exchange rates for a particular amide hydrogen, in the folded vs. random coil states is referred to as the exchange protection factor, and directly reflects the free energy change in the atomic environment of that particular hydrogen between unstructured and structured states of the protein. In this sense, amide hydrogens can be treated as atomic-scale sensors of highly localized free energy change throughout a protein and the magnitude of free energy change reported from each of a protein's amides in a folded vs. unfolded state is precisely equal to \(-RT \ln(\text{protection factor})\) (10). In effect, each peptide amide's exchange rate in a folded protein (when measured) directly and precisely reports the protein's structure and thermodynamic stability at the individual amino acid scale (9, 10). The most important element of this approach to structure determination is that we treat this aggregate exchange rate data for a protein as a “fingerprint” that is uniquely linked with, and identifies, the protein’s 3-D structure.
Development of High Resolution, High Throughput Peptide Amide Hydrogen/Deuterium Exchange-Mass Spectrometry (DXMS). Deuterium exchange methodologies coupled with Liquid Chromatography Mass Spectrometry (LCMS) presently provide the most effective approach to perform hydrogen exchange studies of proteins larger than 30 kDa in size (12, 18-28). Building upon the pioneering work of one of my mentor’s (V. Woods) collaborators – Walter Englander, and David Smith (11, 12, 23, 26, 29)– Woods has developed and implemented a number of improvements to their methodologies, automation and experimental equipment that have significantly improved throughput, comprehensiveness, and resolution. Woods terms this method enhanced Deuterium Exchange-Mass Spectrometry (DXMS) (1, 30-48). As described below, it is a methodology well suited to provide data to rapidly and effectively refine computational structure determination.

Three years ago one of my mentors, V. Woods, began to make DXMS available for basic research at UCSD. The productivity in these collaborations has been remarkable, with thirteen investigators, applying DXMS variously to issues of protein structure, protein-protein and protein-small molecule interactions, protein allosteric (conformational) change, and protein folding, with seven published articles (1, 33, 36, 37, 45, 46, 48) and four issued U.S. patents (30-32, 47). These studies demonstrate the power of DXMS to rapidly and precisely define protein domain organization, for example that of the recently discovered cAPK anchoring protein, DAKAP-2, (Susan Taylor, UCSD); follow the changes in regional stability of proteins before and after they are bound to protein or nucleotide binding partners (Csk. Pat Jennings and Joseph
Adams, UCSD); probe the binding of cAPK regulatory subunits to the cAPK catalytic subunit or to cyclic nucleotides, (Susan Taylor, UCSD); localize and measure free energy transmission through hemoglobin while undergoing conformational change, (Walter Englander, U. Penn). Additional work in progress includes studies with Douglas Cines, U. Penn, (Urokinase-Urokinase receptor binding); Palmer Taylor, UCSD (acetylcholinesterase); Edward Dennis, UCSD, (phospholipase A2); David Wemmer, UC Berkeley (prion proteins); and Ruby MacDonald, Northwestern U. (spectrin).

My studies with investigators at the Joint Center for Structural Genomics (Ian Wilson, Ray Stevens, The Scripps Research Institute; Scott Lesley, Genomics institute of Novartis) have demonstrated the ability of DXMS-derived insights into the presence of disordered regions in proteins to guide the design of well-crystallizing protein constructs, in a high throughput manner (Chapter 4,5). This work is of particular importance to the present proposal, as it affords high throughput methods to rapidly define surface amino acids in proteins, information that can dramatically improve the accuracy of Rosetta predictions.

**COREX: Development of Reliable Methods for Calculating Amide Hydrogen Exchange Rates from Protein Structures.** Under native conditions, proteins are not unique structures, but actually ensembles of conformational states. This observation led to the development of the COREX algorithm by Dr. Vincent Hilser, a computational tool that utilizes the high-resolution structure as a template to generate a large ensemble of incrementally different conformational states(49-53). COREX has been shown to
predict exchange rates with remarkable accuracy and precision when tested against available NMR-derived experimental data, suggesting that the calculated ensemble captures the general features of the actual ensemble, and thus provides a realistic physical description of proteins, reviewed in (54). This algorithm performs the following two computational tasks:

1. The high resolution X-ray or NMR structure (or a presumed structure) is used as a template from which a large ensemble of conformations (>10^5) is generated.

2. The relative enthalpy, $\Delta H_i$, and entropy, $\Delta S_i$, are calculated for each conformation using a surface area-based parameterization of the energetics (55-63) and the resultant Gibbs energy change, $\Delta G_i$, is used to calculate the probability of each state.

The most important aspect of the ensemble-based approach is the ability to calculate the probability for each residue to be in a folded or an unfolded conformation (49-52). These residue-specific probabilities are used to define residue stability constants $K_{f,j}$, as:

$$K_{f,j} = \frac{\Sigma P_{f,j}}{\Sigma P_{nf,j}}$$

where $\Sigma P_{f,j}$ and $\Sigma P_{nf,j}$ are the summed probabilities of all states in the ensemble in which residue $j$ is either folded or unfolded, respectively. According to equation 1 residues with high stability constants will be folded in the majority of highly probable states, while residues with low constants will be unfolded in those states. The importance of equation 1 to the current proposal is that the stability constant can, in many cases, be directly compared to the exchange rates (protection factors) obtained from DXMS experiments. To validate COREX we implemented within it the ability to
derive exchange rate values from the calculated ensemble behavior, and demonstrated that such COREX-predicted rates were remarkably accurate when compared with existing NMR-determined rate data (49-53). Thus the fundamental ability of COREX to calculate accurate hydrogen exchange rate fingerprints from 3-D structures has been firmly established.

**The DXMS-COREX Filter and the Rosetta Algorithm.** We propose that a simple method for refining computational structure prediction (the “DXMS-COREX filter”) is made possible by the combination of the ability to rapidly acquire comprehensive amide exchange rate data on small quantities of protein (with DXMS), and the ability to use the COREX algorithm to calculate the exchange rates of the amides within multiple (thousands) of alternative structures proposed for a protein. Assessment of the degree to which the predictions’ COREX-calculated exchange rates matched the experimental rates would allow identification of accurate structural predictions. The Rosetta algorithm of David Baker is ideally suited for use with these elements.

**Rosetta.** The Rosetta method of *de novo* protein-structure prediction is based on the assumption that the distribution of conformations available to any short segment of the chain is determined largely by the local sequence. To approximate the conformational space available to each segment, sets of 3-mer and 9-mer fragments for each position along the chain are extracted from the protein-structure database based on the sequence-profile similarity and secondary-structure predictions. Compact structures are then
assembled by randomly combining these fragments using a Monte Carlo simulated annealing search. The fitness of individual conformations with respect to non-local interactions is evaluated using an energy function derived from observed distributions in known protein structures. The energy function favors hydrophobic burial and strand pairing, and disfavors steric clashes. For each target sequence, thousands of possible structures (termed “decoys” in the Rosetta literature) are generated with this protocol (64).

As presently employed, the population of proposed structures is automatically filtered and then refined in a full-atom protocol that adds on all side-chains and hydrogen atoms and performs a coupled Monte Carlo minimization of the backbone and side-chain conformations. In addition to the energy function described above, the full-atom energy function includes Lennard-Jones and pairwise solvation potentials, as well as several statistical potentials for side-chain atom pairs, side-chain rotamers, and hydrogen bonds. The accuracy of the Rosetta full-atom energy function has been demonstrated recently by the experimental verification of a computationally designed novel fold (65).

Despite the relative success of the Rosetta method, there is a pressing need for methods to: 1. generate “better”, more accurate Rosetta predictions; 2. discriminate between accurate and less accurate predictions in Rosetta populations; and 3. identify correctly predicted sub-regions within overall less accurate Rosetta predictions. As described below, it is our expectation that the first two of these needs will be robustly met with proper application of DXMS-derived “surface-residue” experimental constraints to be developed in future studies, and the third by the use of advanced
pattern matching techniques brought to the project by my mentor Halbert White, as described below.

**Formulation of the “DXMS-Rosetta-COREX method is as follows:**

1. Microgram quantities of the protein are produced and studied by DXMS technology to measure amide hydrogen exchange rates, establishing the protein's true exchange rate fingerprint.

2. A simple analysis of a portion of this rate information allows precise identification of the protein’s surface-disposed peptide amides (typically 10-20% of them) that have very fast exchange rates, indicating that they are always in full contact with solvent water in the protein, and therefore are on its surface.

3. Multiple (thousands) of alternative structures are predicted/proposed for the target protein using the Rosetta algorithm, with the computations performed in a manner that takes advantage of the DXMS-derived knowledge of the identity of the surface-disposed amides, greatly improving the accuracy of predictions and speeding calculations; **DXMS surface-constrained Rosetta.**

4. The COREX algorithm is used to construct virtual hydrogen exchange rate fingerprint for each of the several DXMS-refined Rosetta-proposed structure(s) for a target protein.

5. These calculated fingerprints are compared to the true DXMS-determined rate fingerprint by employing advanced methods for such comparisons that are sensitive to partially correct structures, and the structural predictions with calculated exchange rate fingerprints that most closely match the experimentally
determined fingerprints identified.

The principal virtues of this approach are its simplicity, and the ease with which DXMS data can be rapidly obtained despite protein idiosyncrasies. Most of the experimental technique is performed under conditions that suppress the unique features of individual amino acids - the use of acid pH, denaturants, and non-specific proteases, making the same basic DXMS methods universally applicable to proteins that have dramatically differing properties under native conditions. Crystallization is not required, protein size is not a limiting factor, and the amount of protein needed for DXMS study is far less than required for crystallography or NMR. Dramatic increases in structure determination throughput and economy are expected with the successful development of DXMS-Rosetta-COREX. The most expensive part of the necessary instrumentation, the mass spectrometer, is now affordable at the departmental-level within most academic institutions (~ $200k): compare this to the present cost and inefficiency of robotic nano-crystallization, diffraction data acquisition, and data analysis- not to mention the cost of a Synchrotron light source!
6.3 RESULTS

1. **DXMS Methodology.** Detailed descriptions of the several enhancements that constitute DXMS are presented in seven articles (1, 33, 36, 37, 45, 46, 48) and four U.S. patents issued to the PI (30-32, 47). The technique has an initial exchange-labeling step performed under entirely physiologic conditions of pH, ionic strength, and buffer salts (Figure 6-1A); and a subsequent localization step performed under non-native, exchange-"quench" conditions (Figure 6-1B). The labeling is performed by simply adding deuterated water to a solution of the protein. During this on-exchange incubation, deuterium exchanges onto the several amides of the protein. As labeling progresses, aliquots are exchange-"quenched" by shifting the protein to conditions (low pH, and temperature) that dramatically slow the rate of exchange, effectively locking in place the attached deuterium. Undesired "off-exchange", or loss of label after establishment of sample quench, can be minimized by holding samples at very low temperatures (-80°C) until they are melted (at 0°C) and further processed as below. This process has been automated with the development of a cryogenic autosampler within our DXMS apparatus (30-33, 37, 46) (See Figure 6-1B).

In the second step, we determine the amino acid sequence location and amount of attached deuterium (Figure 6-1B). Under "quench" conditions, the protein sample is (automatically) first optionally denatured, optionally disulfide-reduced, and then proteolyzed by solid-phase pepsin into overlapping fragments of ~3-15 amino acids in size. It is to be emphasized that this is high-throughput, exhaustive but not limited
proteolysis, with typical digestion times being of the order of 20 seconds. The digests are then subjected to rapid high performance liquid chromatography (HPLC) separation (5-10 minute gradients), and directly analyzed by electrospray-ion trap or time of flight (TOF) mass spectrometry performed under conditions adapted to amide hydrogen exchange studies.

The extent of pepsin digestion is finely tuned with the goal of generating multiple overlapping fragments of the protein. When desired, we achieve even finer fragmentation with additional acid-reactive proteinases. Fragmentation is followed by rapid sequence-identification, performed first with undeuterated protein under quench conditions, followed by assessment of deuterium label bound to each fragment generated from deuterated protein samples. The differences in deuterium content between peptides with overlapping sequences are used to further sub-localize and quantify attached deuterium label. Integrated automation of fluidics, including sample preparation (functional deuteration), sample storage and injection (cryogenic autosampler), solid-state proteolysis, liquid chromatography, and mass spectrometry allows rapid, continuous data acquisition, typically with one sample processed every 20 minutes. With these enhancements, we now complete a high resolution, comprehensive DXMS analysis of a protein in two weeks, and can process 10 proteins simultaneously (1).

2. DXMS exchange rate fingerprints. MS scans containing the numerous peptides of interest are individually isolated from the mass-intensity lists, processed to optimize signal-to-noise ratios, and then the geometric centroids of the isotopic envelopes of each
peptide determined and recorded. Calculation of the difference in weight between the measured centroid of the deuterated peptide and the centroid for the same peptide without deuterium allows determination of the amount of deuterium on each peptide at the time of MS measurement. These data manipulations are now automatically performed by specialized data reduction software that is the result of a University of California-funded, two-year collaborative software writing effort with Sierra Analytics, LLC. Modesto, CA. (33, 36, 42). This, and the HR-DXMS de-convolution software described below, results in the determination of exchange rates for the majority of the amides in a target protein (2). The two-dimensional matrix formed by annotating each peptide bond amide within a protein with its hydrogen exchange rate, as measured under conditions native for the 3-D structure of the protein, constitutes its "exchange rate fingerprint".

We have developed computational methods for the de-convolution of aggregate, time-dependent peptide deuteration data to specific exchange rates for each of the amide hydrogens within a protein. This method, which we term “High Resolution, residue-specific determination of amide hydrogen exchange rates from DXMS data” (HR-DXMS), employs an algorithm centered on use of a two-phase numerical technique, linear programming (LP) for an initial rate estimation, followed by a nonlinear least squares fit (NLS). Essential to success of the method is the derivation and incorporation of residue-specific corrections for deuterium loss during “back-exchange” in contrast to the use of “peptide-average” loss corrections usually employed in hydrogen exchange data analysis. The method can also make use of additional hydrogen exchange data, obtained by systematically varying the duration of the usually
deleterious “off-exchange”, performed under quench conditions, to allow resolution of individual amide rates within protein regions not sufficiently resolved by enzymatic fragmentation alone. A detailed description of the algorithm and software is included in the Supplemental Material section of Chapter 2.

Figure 6-2 is an example of validation studies of HR-DXMS de-convolution, in which we: 1) Started with the exchange rate “fingerprint” for horse cytochrome c, that had been previously determined by sparse NMR data (Figure 6-2, black line) (66); 2) Computationally generated simulated DXMS deuterated-fragment datasets from this experimental rate map, employing proteolytic fragmentation densities and numbers of sampled on-exchange times typical for our high fragmentation-density DXMS studies; 3) supplemented the dataset with 20% random error in fragment deuteration level to conservatively simulate experimental noise, (which is typically 5-10% in our studies) and; 4) processed the dataset with the HR-DXMS program. The resulting deconvoluted fingerprint (Figure 6-2, grey line) closely matches that of the reference protein, even with this high level of simulated error included. De-convolutions of datasets employing modestly higher fragmentation densities, and multiple off-exchange times under quench conditions have nearly identical maps with native cytochrome c.

3. Implementation of computational economics. Our overarching objective is to be able to provide DXMS-Rosetta-COREX as a practical resource to the academic community. This requires that the cost of the necessary computations be affordable by the average investigator- hence our emphasis on the development
and implementation of more efficient approaches to the calculations. We have accomplished this by use of supercomputing resources, primarily the now decommissioned Blue Horizon at the SDSC Supercomputer center in San Diego, and the Condor Flock of networked computers at the University of Wisconsin. Both COREX and Rosetta have been installed and intensively used on platforms, operating in both full and Monte-Carlo sampling modes. While COREX calculations are computationally intensive, they lend themselves well to massive parallelization. We were awarded large allocations of computing time on a competitive basis through the National Partnership for Advanced Computational Infrastructure (NPACI), and have performed extensive preliminary studies, some of which are presented below. This experiment has indicated that substantial computational economies can be readily implemented into DXMS-Rosetta-COREX.

An example of this is our recently completed study of the elastic protein spectrin, where we found that running COREX in a sparse Monte Carlo sampling mode markedly decreased the time required to compute rate fingerprints from structures. It also demonstrates the degree to which COREX-calculated fingerprints, computed from a known 3-D structure, agree with DXMS-derived data, even when very significant time-saving computational approximations are made. This study (which incidentally proposes a novel molecular mechanism for spectrin elasticity) is included as Chapter 2. We acquired high-density fragmentation DXMS data on a 221 aa two-tandem-repeat construct of chicken brain α-spectrin (16th-17th repeats), with the protein deuterated for 10 to 3000 seconds at room temperature (Figure 6-3A). We then processed the resulting data by HR-DXMS to produce the construct's actual DXMS experimental rate
fingerprint (black line, Figure 6-3B), and then calculated the exchange rate fingerprint of the \( \alpha \)-spectrin construct from its known crystal structure (67) employing COREX, operating on Blue Horizon in Monte Carlo mode, with a window size of 8 amino acids and a total sampling of 8000 partially unfolded states (grey line, Figure 6-3B). There is close agreement between the true, experimentally-determined (DXMS) rate fingerprint and the Monte Carlo-COREX fingerprint calculated from the known 3-D structure. Most of the differences are in areas where there is relatively sparse peptide probe overlap: increased fragmentation and use of varying off-exchange times under quench conditions will have a beneficial effect on such areas of experimental fingerprints. Remarkably, this analysis was accomplished with a run time \(~24,500\)-fold faster than would have been required for a complete, high-resolution analysis of all unfolding states.

4. **Proof-of-Principle Studies of the DXMS-COREX filter.** The central tenant of our method is that accurate structural predictions for a protein will have hydrogen exchange-rate fingerprints that closely match that of the actual protein’s structure. As a direct test of this, we applied DXMS-COREX to the structural predictions submitted for target T0102 (Bacteriocin AS-48, PDB 1E68) in the CASP 4 experiment. COREX was run in full-sampling mode against each of the 87 structural models submitted for the complete structure of this target, employing an 8 aa residue window to produce 3,856 partially folded states. The analysis was performed on 2 nodes (sixteen 375 GHz cpu’s) on Blue Horizon at the San Diego Supercomputing Center and took a total of 2 hours to run. Since experimental hydrogen exchange data on T0102 was not available in order to
produce an exchange rate profile we made the assumption that the theoretical profile produced by COREX would accurately predict the hydrogen exchange profile of the protein. Therefore the crystal structure of target T0102 was also analyzed by COREX and its “experimental” rate fingerprint calculated. The RMSD between the protection factor fingerprint of each prediction and the fingerprint derived from the crystal structure was calculated.

In Figure 6-4A, the 87 predictions are ranked (horizontally) by degree of RMSD agreement between prediction rate fingerprints and the structure's COREX rate fingerprint, with the positions of some of the best (most accurate) structural predictions (as ranked by CASP using alpha-carbon RMSD) indicated by arrows. In Figure 6-4B, the predictions are ranked (horizontally) by 3-D structural accuracy using the longest contiguous segment of the backbone of the protein within a 3 Angstrom deviation (LGA) with the positions of some of the best exchange-rate "fingerprint fits" indicated by arrows. As seen, the sum of the COREX residuals (residue-specific RMSD values) is lowest for the most accurate structures. Overall, there is strong correlation between prediction accuracy and exchange-rate fingerprint fit, at least for the best predictions using two different parameters for measuring structural similarity. The anomalous behavior seen with two of the predictions likely reflects the inability of these measures such as global rate fingerprint RMSD to detect partially correct structures. This analysis was repeated for the CASP4 protein T0125 with a window size of 10, and again, the top structure predictions had the closest agreement in exchange rate fingerprints (data not shown).
5. **Performance of Rosetta predictions in the DXMS-COREX filter.** We next performed a detailed analysis (Figure 6-5.) of the relationship between variance in protein 3-D structure prediction accuracy and variance in COREX- calculated exchange-rate fingerprints for prediction vs. true structures. Six proteins were studied (RecA C-terminal domain, Bovine pancreatic trypsin inhibitor, chicken brain α-spectrin, immunoglobulin binding protein (1gb1), gamma-delta resolvase and Ribosomal protein S1, RNA-binding domain. Hundreds of “partially” correct versions of each of the six structures were computationally generated from the true structures by partial sidechain rearrangement (using the residue mutation function of SwissPdb viewer (68), and even less accurate variants generated by full sidechain rearrangement, produced using an in-house algorithm that randomly generated sidechain torsion angles, rejecting any conformations with hard sphere clashes. A publicly available version of Rosetta (69) available at http://depts.wshington.edu/bakerpg was used to generate one thousand Rosetta backbone predictions for each of these proteins, each of which had a single sidechain conformation assigned arbitrarily by the program Scap (70), as the version of Rosetta we used did not directly enable side chain placement. COREX was run against these several permuted structures and Rosetta predictions to determine their exchange-rate fingerprints, and calculated the structural variance between these and the true, native structure, by a variety of means including all-atom RMSD, C-α-RMSD; and calculated the variance between exchange rate fingerprint as RMSD. The results of this analysis demonstrated that low amide exchange rate fingerprint variance was strongly correlated with structural accuracy, once sufficiently accurate predictions were produced.
Figure 6-5 shows representative results for the protein 1gb1 where fingerprint variance is plotted against structural variance, with side-chain rearrangements of the native structure shown as small closed diamonds, and Rosetta predictions (with Scap side-chain placements) as large open diamonds. We then took one of the better Rosetta backbone predictions we had made (Figure 6-5, circled large diamond), generated 100 side-chain rotamer structural variants of it with our in-house program, ran COREX against each, recalculated structural and fingerprint variance, and plotted these values (small open diamonds, Figure 6-5). We found that the resulting Rosetta predictions, while slightly more structurally accurate, had exchange-rate fingerprints that were markedly closer to the native structure than the fingerprint of the Rosetta prediction from which they were generated (circled, Figure 6-5). Simply varying the method used for side-chain placement on Rosetta backbones had substantially improved the structure/rate fingerprint variance behavior of Rosetta predictions, indicating that even more substantial improvements would be forthcoming with state-of-the-art application of Rosetta.

6. DXMS-constrained Rosetta- High throughput DXMS identification of amino acids on the surface of proteins to improve Rosetta predictions. We have recently realized that a simple analysis of a small portion of the DXMS-derived experimental rate information allows unambiguous identification of the set of a protein’s peptide amides (typically 10-20% of them) that are always in full contact with solvent water in the protein, and therefore are on its surface; the very most rapidly exchanging amides. These “very fast-amides” are hydrogen-bonded to solvent water rather than the protein
most of the time, and may variously be present in surface- oriented structured regions (short loops, in kinks in \( \alpha \)-helices, and in edge strands of \( \beta \)-sheets) or disordered stretches of sequence. If available, knowledge of the identity of these residues can be readily incorporated into the actual Rosetta structure backbone and sidechain calculation process, resulting in much improved accuracy of Rosetta predictions and speeding calculations. Furthermore, we realized that a DXMS data acquisition approach we recently employed in an “unrelated” project (described below) could readily provide this uniquely valuable “surface residue” experimental information in a high-throughput manner making it readily available to improve Rosetta predictions, thus likely ensuring their operability in DXMS-COREX. Future studies will focus on implementation and validation of what we now term **DXMS-constrained Rosetta**. Given the central role we now envision for Rosetta, we refer to this structure determination engine as **DXMS-Rosetta-COREX**.

In Chapter 4 we presented our study in collaboration with our colleagues at the Joint Centers for Structural Genomics (JCSG), were we tested the hypothesis that DXMS could be used to localize long stretches (4 or more contiguous residues) of rapidly exchanging amino acid sequence in proteins, as these likely represented “disordered” regions in the protein. We identified, and then engineered-out, these disordered regions to see if crystallization success was improved for use in x-ray crystallographic studies.

In summary DXMS analysis was successfully performed on 21 *Thermatoga maritima* proteins provided to us by the JCSG with various crystallization and diffraction characteristics. Data acquisition was performed in a single 30-hour run, and reduction of
the data to exchange rate maps was completed in two weeks, with resulting localization and prediction of several unstructured regions within the proteins. When compared with those *Thermotoga* targets of known structure, the DXMS-method correctly localized small regions of disorder. DXMS analysis was then correlated with the propensity of such targets to crystallize and was further utilized to define truncations that might improve crystallization. Truncations that were defined solely on the basis of DXMS analysis demonstrated greatly improved crystallization, and were successfully used to obtain high-resolution structures for two proteins that had previously failed all crystallization attempts(1).

Figure 6-6 shows the ten-second amide hydrogen/deuterium exchange map for one of the proteins studied, TM1079. The brief, 10 second deuteration employed allowed selective labeling of the most rapidly exchanging amides in the protein. The horizontal dark blue bars are the protein’s pepsin-generated fragments that had been produced, identified, and used as exchange rate probes in the subsequent 10-second deuteration study. The number of deuterons that went on to each peptide in 10 seconds is indicated by the number of red residues in each peptide. Two segments (circled) each contained five contiguous highly solvated amides, indicating that they are likely disordered (1). Also seen were a large number of relatively isolated, highly solvated amides scattered throughout the sequence, constituting 13% of all amides in the protein. These represent highly solvated amides on or near the surface of the structured regions of the protein (1).

Figure 6-7 shows the deuteration results for all of the 21 proteins that were analyzed in this study, whose amino acid lengths varied from 76 to 461 residues. Dark
regions indicated fast exchanging amides and clear regions indicate stretches of no exchange. Likely disordered regions of four or more fast exchanging amides are circled. While the circled stretches of sequence were the focus of our study, we now recognize that the isolated, (single to triple) rapidly exchanging amides, that are peppered throughout these sequences, are fast-exchanging structured residues on the surfaces of the proteins. Their facile identification will be of great benefit in refining Rosetta structural predictions for subsequent use in DXMS-Rosetta-COREX. The design of this study was biased towards the detection of large stretches of rapidly exchanging sequence. In our future studies we will employ modifications that will ensure high sensitivity detection and localization of these structured, but rapidly exchanging surface-amides, and expect that 10-20 % of a typical protein’s amides will prove to be so surface-localized.

7. Advanced methods for amide hydrogen exchange rate fingerprint comparison. There are three principal sources of error and inaccuracy in the operation of DXMS-Rosetta-COREX: 1. Inaccuracies in the DXMS experimental data; 2. Inaccuracies in the COREX calculations, arising from algorithmic insufficiency and sampling errors and, 3. Inaccuracy in Rosetta structure predictions. The project will greatly benefit from exchange-rate “fingerprint” comparison methods that are tolerant to these sources of error. Future studies will focus on two approaches:

a) Comparison methods that optimize the ability of DXMS-COREX to identify Rosetta predictions that contain accurately-determined structures for sub-regions of a
target protein, but are otherwise globally inaccurate. Our collaborator Halbert White will develop and implement methods by which the exchange rate comparisons between DXMS-data and COREX exchange-rate calculations are performed in a Rosetta-prediction sub-volume specific manner.

b) Comparison methods which ensure that accurate Rosetta predictions will be identified by DXMS-COREX even when occasional outliers are imbedded in the DXMS data or COREX computational results. In calculating the goodness of fit between DXMS-determined exchange-rate fingerprints and COREX-produced rate fingerprints for predictions, we have used two measures: mean absolute error (mae) and root mean squared error (rmse). In our experiments it has been evident that extreme outliers are a common feature; that is, there are typically a few amides for which the difference between the experimental and COREX rates is quite large relative to the differences for other amides. Mean absolute error and root mean squared error are ideally suited to accommodating random variation in errors arising from the double exponential and normal distributions respectively, but can be seriously adversely affected by outliers.

The adverse effects of outliers have been well studied in the statistics literature, and Hal White has made a number of contributions to this literature with particular attention to measuring goodness of fit in the presence of outliers (71, 72). His research has focused on both the serious adverse consequences of outliers for standard measures such as mae and rmse, as well as on the use of more sophisticated approaches to measuring goodness of fit that are robust to the presence of outliers. Future studies will apply these more sophisticated measures to the fingerprint-matching problem, as well as to validate and refine these methods. Please see (72) pg. 555 for a presentation of the
mathematical formalisms that will be employed. In particular, we expect that Tukey’s biweight function may provide substantial protection against the adverse effects of outliers (73, 74).

6.4 ACKNOWLEDGMENTS

Chapter 6 is taken in part as it appears in US/PCT Patent Application # PCT/US2004/036456, “Methods for the Determination of Protein Three-Dimensional Structure Employing Hydrogen Exchange Analysis to Refine Computational Structure Prediction” filed November 1, 2004. Assignee: Regents, University of California with coauthors V.L. Woods Jr., P. Bourne, V. Hilser, and H. White. I performed all DXMS studies, Rosetta structure predictions, COREX analyses, and correlative studies described in this chapter. I would like to thank my several mentors: Virgil Woods, for assisting in the writing and progress of the grant, Phil Bourne, for assisting in setting up the use of the Blue Horizon supercomputer, Vince Hilser, for providing COREX, David Baker, for his assistance in running Rosetta, and Halbert White, for his advice on the computational development.
Figure 6-1A: Deuterium on exchange: The fragmentations of undeuterated protein samples are determined by “tuning conditions to produce an optimal fragmentation pattern. Samples are deuterated by incubation in deuterated buffer for a period of on exchange times. Deuterated samples are quenched by shifting to pH2.7 and -80°C thereby locking the deuterium in place.
Figure 6-1B: Deuterium localization and quantification: On-exchanged sample is placed in a cryogenic autosampler and rapidly melted at 0°C followed by injection into a denaturing solvent stream at pH 2.7 maintained at 0°C. The sample undergoes proteolysis and fragments are loaded onto a column for rpHPLC elution. Effluent from the column is immediately directed into the mass spectrometer for initial peptide identification and MS2 fragmentation. Data is analyzed for deuterium localization and quantification. Figure 1-2 illustrates a flowchart for the method.
Figure 6-2: an example of validation studies of HR-DXMS de-convolution HR-DXMS deconvolution of Cytochrome-c. A 20% error was incorporate in the generation of the simulated cytochrom c deconvoluted exchange rate fingerprint. Reference rates were obtained from published NMR results [66].
Figure 6-3: high-density fragmentation DXMS data on a 221 aa two-tandem-repeat construct of chicken brain α-spectrin (16th-17th repeats). (A) The spectrin R1617 construct was deuterated for 10 to 300,000 seconds at room temperature and produced a fragmentation map of 200 peptides from a pepsin + FPXIII digestion. (B) Processed the resulting data by HR-DXMS to produced the construct's actual DXMS experimental rate fingerprint (black line), with the calculated the exchange rate fingerprint (red line) of the α-spectrin construct from its known crystal structure (67) employing COREX, operating on Blue Horizon in Monte Carlo mode, with a window size of 8 amino acids and a total sampling of 8000 partially unfolded states.
Figure 6-4A: Structure predictions for CASP4 target protein #102 ranked by the RMSD of the residuals between the COREX calculated rate fingerprint of the predicted structure, and the COREX calculated fingerprint of the crystal structure. The 87 predictions are ranked (horizontally) by increasing RMSD agreement between the rate fingerprints of the predicted structures and the rate fingerprint of the actual crystal structure with the positions of some of the best (most accurate) structural predictions for this target (as determined by CASP4 α-carbon RMSD ranking criteria) indicated by arrows. Top 10% of the best structures as ranked by CASP are clustered in the regions of lower COREX RMSD.
Figure 6-4B: Structure predictions for CASP4 target protein #102 ranked by the RMSD of the residuals between the COREX calculated rate fingerprint of the predicted structure, and the COREX calculated fingerprint of the crystal structure. The 87 predictions are ranked (horizontally) by increasing RMSD agreement between the rate fingerprints of the predicted structures and the rate fingerprint of the actual crystal structure with the positions of some of the best (most accurate) structural predictions for this target (as determined by the number of the longest contiguous segment within 4Å backbone α-carbon deviation (LGA) in CASP4) indicated by arrows. Top 10% of the best structures as ranked by CASP using the number of correctly aligned residues (LGA) with the crystal structure are clustered in the regions of lower COREX RMSD.
Figure 6-5: analysis of the relationship between variance in protein 3-D structure prediction accuracy and variance in COREX- calculated exchange-rate fingerprints for prediction vs. true structures. Six proteins were studied (RecA C-terminal domain, Bovine pancreatic trypsin inhibitor, chicken brain \( \alpha \)-spectrin, immunoglobulin binding protein (1gb1), gamma-delta resolvase and Ribosomal protein S1, RNA-binding domain. Hundreds of “partially” correct versions of each of the six structures were computationally generated from the true structures by partial sidechain rearrangement (using the residue mutation function of SwissPdb viewer (68). Side chaing rearrangements were procuced by SCAP (70).
Figure 6-6: Ten-second amide hydrogen/deuterium exchange map for one of the proteins studied, TM1079. 10 second deuteration employed allowed selective labeling of the most rapidly exchanging amides in the protein. The horizontal dark blue bars are the protein’s pepsin-generated fragments that had been produced, identified, and used as exchange rate probes in the subsequent 10-second deuteration study. The number of deuterons that went on to each peptide in 10 seconds is indicated by the number of red residues in each peptide. Two segments (circled) each contained five contiguous highly solvated amides, indicating that they are likely disordered (Pantazatos et. al. 2004)
Figure 6-7: Deuteration results for all of the 21 proteins that were analyzed using DXMS for surface exposed residues. Amino acid lengths varied from 76 to 461 residues. Dark regions indicated fast exchanging amides and clear regions indicate stretches of no exchange. Likely disordered regions of four or more fast exchanging amides are circled. (Figure Adapted from Pantazatos et al. 2004).
6.5 REFERENCES


47. Woods Jr., V. 2003. Methods for identifying hot- spot residues of binding proteins and small compounds that bind to the same. *In* U.S. Patent # 6,599,707 USA.


Global efforts in structural genomics are currently powered by the need to develop novel therapeutic drugs towards disease and antimicrobial infections. Current technologies and strategies for genome sequencing enable a novel microbial genome to be computed within days, so that even newly arising pathogens can undergo sequencing for moderate costs. The ongoing development of these technologies has sparked a massive effort for increasing HT structure determination efforts through numerous international centers (Figure 7-1) (1). Currently, 67 eubacterial genomes have been completely sequenced, including nearly all the clinically important pathogens (http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/micro.html). Crystallization is the main bottleneck but progress is encouraging as improved predictive and experimental methods for structure prediction are merging. For development of therapeutic drugs for targeted disease, efforts are focused on understanding protein function. In order to understand protein function it is imperative to study protein dynamics and conformational change. While crystallography and NMR provide detailed high resolution structural information, dynamics can be probed by biophysical methods that detect changes in secondary and tertiary structure. Various methods exist for obtaining information on protein structure dynamics but they provide low resolution analysis and do not capture the structural energetics in a quantitative way. These include but are not limited to: florescence spectroscopy, circular dichroism, nonspecific and site specific proteolysis, ultracentrifugation, HPLC mass spectrometry analysis(2), oxidative
footprinting mass spectrometry (3), site-directed and non-specific chemical cross-linking coupled with proteolysis and mass spectrometry(4, 5), and time-solved fluorescence polarization, linear dichroism time resolved x-ray crystallography (6).

In the forefront of this field is the increasing development and use of H/D exchange mass spectrometry(7, 8). The integration of hydrogen/deuterium exchange methods with mass spectrometry has resulted in a powerful platform for high-resolution studies of protein structure, ligand binding, and protein dynamics in solution. The addition of proteolytic cleavage to the coupling of deuterium exchange with HPLC/MS raises the possibility of determining exchange rates of fractionated peptides short enough to allow characterization of the exchange at each amide linkage. Refinements in proteolytic methods, especially the introduction of different proteases able to cleave a variety of linkages, have enabled the extensive fragmentation of a protein to yield the many overlapping peptides necessary to determine the exchange rates of individual amides. DXMS is impressive in its ability to rapidly generate unambiguous information on microenvironment for individual amides, or small groups of amides across the sequence of a protein. Mapping of the exchange rates of individual amides can be viewed as a “fingerprint” that relates to local environment, and ultimately structure. Efforts have been made for using experimental data for constraining structure prediction (Galat, 1996; Xu, 2000; Zheng, 2002; Bowers, 2000; Borbat, 2002; Sharp, 2005). These methods have reported some success using data from NMR (Xu, 2000), solution X-ray scattering (Zheng, 2002), and other biophysical techniques (Borbat, 2002). A more recent study has shown promising results for using solvent accessibility data of sidechains for refining and validating NMR determined structures and models.
generated by the Rosetta server using oxidative footprinting {Sharp, 2005}. Oxidative footprinting, much like H/D exchange is an alternative method of detecting regions of solvent exposure in proteins by using hydroxyl radical production to chemically modify solvent exposed side chains. Analysis by proteolysis and mass spectrometry similar to DXMS provides information on regions of solvent exposure based on side chain reactivity with the hydroxyl radical in solution. A recent review of the progress on the use of oxidative footprinting for detecting solvent accessibly regions is reported by Guan and Chance (3). Oxidative footprinting and DXMS share similarities in that both methods attempt to detect conformational changes in proteins and regions of solvent accessibility using solvent labeling with subsequent proteolysis and mass spec analysis. The advantage that footprinting has is that the possibility of back exchange is eliminated because of the permanent chemical nature of the oxidation reaction. The disadvantages outweigh the advantages. Footprinting uses side chains as probes for solvent accessibility. Side chains are not as sensitive as amide hydrogens for obtaining quantitative information on regional and local flexibility because amide hydrogens probe the thermodynamic nature of the backbone of the structure without the contribution of the side chains. Using side chains as the probes for flexibility and solvent accessible area may result in false positives because regions that are detected as solvent exposed with labeled side chains may not be since amides can be shielded from solvent and still have a solvent exposed side chain. An example is the secondary structure of a triple helical coiled coil such as spectrin. Helical positions of coiled coils are represented by a repeating heptad pattern. Positions of amino acids in the helix can be predicted and are represented by the heptad repeating pattern where the locations of the aliphatic and
hydrophilic amino acids and their side chains show localization on specific sides of the helix. Although these two techniques might seem competitive there is a powerful potential for incorporating amide solvent accessibility data with subsequent sidechain solvent accessibility data for further model refinement and filtering of decoys. This avenue offers opportunities for further investigation. The successful development and emergence of experimental data with structure determination efforts will prove to be a significant contribution towards the determination of 3D protein structures. Furthermore, advances in computational and developing technology will allow bottlenecks and challenges in the further development of this work to be overcome.

My studies have succeeded in their overarching goal of applying DXMS analysis to speed the pace of protein structure determination. First, DXMS-guided crystallographic construct design, described in Chapters 4 and 5, has proven to be an effective and high-throughput method for increasing the pace of protein structure determination by crystallographic means. Our collaborators at the Joint Centers for Structural Genomics employ it as a central tool in their efforts to speed structure determination. Furthermore, the elements of a hybrid computational-experimental method for structure determination has been identified, implemented, and evaluated (Chapter 6). My studies indicate that modest improvements in prediction accuracy, perhaps through future implementation of solvent-constrained Rosetta, are likely to render the method operational. If this proves to be the case, this method may also significantly accelerate the pace of protein structure determination. Finally, high-resolution DXMS studies of the protein alpha-spectrin, made possible with one of the components developed for DXMS-Rosetta-COREX (HR-DXMS), have allowed us to propose a novel mechanism
by which mechanical energy is stored by spectrin, accounting for its unusual elastic behavior (Chapter 2). Submission of this work for publication will follow successful completion of additional validation studies of the HR-DXMS method, including analysis of additional proteins whose exchange rates have been determined using NMR such as BPTI, OMTKY3 and Snuclease, and computational improvements to the method, such as provision of statistical indicators of the reliability of specific rate determinations. We anticipate that publication of our HR-DXMS method will have broad impact on how DXMS data is analyzed and interpreted for studies of protein dynamics and structure.

The development and use of a surface constrained DXMS-Rosetta filter was proposed in chapter 6 in step 3 of the DXMS-ROSETTA-COREX method to rapidly filter thousands of predicted structures based on DXMS determined solvent exposed regions. It is likely that further development and validation of this approach to a preliminary refinement of a prediction set will significantly improve the performance of the DXMS-Rosetta-COREX filter, and should be a principal focus of future work.

It is suggested that the process might proceed as follows:

A. Select target proteins for which David Baker has produced a set of decoys that scored well in CASP5.

B. Choose a reliable algorithm for analyzing amide hydrogen solvent accessibility from these models. GETArea and AEROMOL are examples of algorithms that may determine the percentage of exposed surface amide hydrogens on the crystal structure of the target and the top scoring decoys.

C. Optimize the parameters for calculating the amide hydrogen accessible surface area percentage.
D. Test the effectiveness of filtering the decoys of the selected target proteins using amide H ASA%.

E. Determine if the filtered decoy set contains an enriched set of predictions that closely match the crystal structure.
<table>
<thead>
<tr>
<th>Structural genomics research center</th>
<th>Organisms</th>
<th>URL</th>
<th>Principal institutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Berkeley Structural Genomics Center</td>
<td>Mycoplasma genitalium, Mycoplasma pneumoniae</td>
<td><a href="http://www.ahgenc.org/">http://www.ahgenc.org/</a></td>
<td>Lawrence Berkeley National Laboratory; Stanford University; University of North Carolina</td>
</tr>
<tr>
<td>TB Structural Genomics Consortium</td>
<td>Mycobacterium tuberculosis</td>
<td><a href="http://www.doe-mbi.ucla.edu/TB/">http://www.doe-mbi.ucla.edu/TB/</a></td>
<td>Los Alamos National Laboratory; a collaboration of scientists in six countries</td>
</tr>
<tr>
<td>The Southeast Collaboratory for Structural Genomics</td>
<td>Caenorhabditis elegans</td>
<td><a href="http://www.secg.org/">http://www.secg.org/</a></td>
<td>University of Georgia</td>
</tr>
<tr>
<td>Northeast Structural Genomics Consortium</td>
<td>Drosophila melanogaster, Saccharomyces cerevisiae, C. elegans; human</td>
<td><a href="http://www.nsgc.org/">http://www.nsgc.org/</a></td>
<td>Rutgers University; Columbia University; Cornell University; Hampton Woodard Research Institute; Pacific Northwest National Laboratories; The State University of New York at Buffalo; University of Toronto; Yale University</td>
</tr>
<tr>
<td>The Midwest Center for Structural Genomics</td>
<td>All biological kingdoms</td>
<td><a href="http://www.mcs.g.welch.gov/">http://www.mcs.g.welch.gov/</a></td>
<td>Argonne National Laboratory; Northwestern University; Washington University School of Medicine; University College London; UT Southwestern Medical Center at Dallas; University of Toronto; University of Virginia</td>
</tr>
<tr>
<td>New York Structural Genomics Research Consortium</td>
<td>Human, model organisms</td>
<td><a href="http://www.nysgrc.org/">http://www.nysgrc.org/</a></td>
<td>Rockefeller University; Albert Einstein College of Medicine; Brookhaven National Laboratory; Weill Medical College; Columbia University</td>
</tr>
<tr>
<td>The Joint Center for Structural Genomics</td>
<td>C. elegans</td>
<td><a href="http://www.jcrsg.org/cbi/crcr/prod/htone.html">http://www.jcrsg.org/cbi/crcr/prod/htone.html</a></td>
<td>California-based center</td>
</tr>
<tr>
<td>RIKEN Structural Genomics Initiative</td>
<td>Arabidopsis thaliana</td>
<td><a href="http://www.irms.nik.kog.org/">http://www.irms.nik.kog.org/</a></td>
<td>University of Wisconsin</td>
</tr>
<tr>
<td>Ontario Center for Structural Proteomics</td>
<td>Mouse, Arabidopsis thaliana, Thermus thermophilus</td>
<td><a href="http://www.sci.bik.org">http://www.sci.bik.org</a></td>
<td>The Institute of Physical and Chemical Research, Japan</td>
</tr>
<tr>
<td>ESF Functional Genomics Programme</td>
<td>All biological kingdoms</td>
<td><a href="http://www.functionalgenomics.org.uk">http://www.functionalgenomics.org.uk</a></td>
<td>European Science Foundation</td>
</tr>
</tbody>
</table>

*Abbreviation: TB, tuberculosis.

Figure 7-1: Table listing ongoing structural genomics initiatives. (Figure taken from Scmid M. B. 2002).
7.1 REFERENCES


