ORDER AND DISORDER IN PROTEINS

by

ASLI ERTEKIN

A dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in

Computational Biology and Molecular Biophysics

Written under the direction of

Dr. Gaetano T. Montelione

And approved by

____________________________
____________________________
____________________________
____________________________
____________________________

New Brunswick, New Jersey

OCTOBER, 2011
ABSTRACT OF THE DISSERTATION

Order and Disorder in Proteins

by ASLI ERTEKIN

Dissertation Director:

Dr. Gaetano T. Montelione

In contrast to the general view that proteins should have a specific 3D structure in solution for their activity, there are many proteins which do not have a folded “native” structure for a big portion of their sequence. While these intrinsically disordered regions are essential for protein function, they cause problems in efforts for determining the 3D structures for the folded domains. It has been shown that the removal of the disordered domains improved the structure determination success both by X-ray crystallography and by NMR. As part of Northeast Structural Genomics (NESG) effort I worked on identifying the disordered and flexible parts of the protein using Hydrogen/Deuterium Exchange with Mass Spectroscopy (HDX-MS) analysis for construct optimization for high-throughput structure determination. Using this method I also studied human Smad3, which is an important part of the TGF-β-signaling pathway; and provided the first experimental data on structural features of the linker domain. During my training, I also studied human Deleted in Oral Cancer (DOC-1) protein, which was one of the proteins I studied by HDX-MS for construct optimization. We determined the solution structure of
the folded region of DOC-1, which was shown to be important in cell-cycle regulation and cancer biology; and I also studied structure-function relations. Additionally, we studied the solution structure of Methionine Sulfoxide Reductase B from Bacillus subtilis, an important protein for reversing oxidative damage in cells, by NMR as a part of methods development studies for NMR for large proteins.
ACKNOWLEDGEMENTS

The work I described here was made possible with the help and support of many people. Especially because with the beginning of my Ph. D. journey I have been learning something new at every step I took.

I am grateful to all the people in the Montelione lab, who were very helpful for every need I had for all my projects. Especially Haleema Janjua, who helped me incredibly on the CDK2AP1 project, was there when the project and I needed help. I learned a lot through our studies on the finicky protein, CDK2AP1. I would like to give special thanks to my highly knowledgeable research professors Tom Acton and Rong Xiao, who were always there when I was confused and I needed help with my projects.

I had three mentors throughout my studies, GVT Swapna, Jim Aramini and Paolo Rossi. I started to learn about NMR with Swapna with many hours of sitting by the instrument and scribbling pulse sequences on any paper available around. With great help from Jim, I solved my first structure, CDK2AP1. He always seemed more than happy to help me, and I abused it for sure. Paolo had the unfortunate luck to have his desk next to mine. He had to be the first person I bounced my questions off, about anything remotely related. And he worked with me on a very challenging protein structure, MsrB. Thanks to all my mentors for their patience and help.

And my advisor, the curator of this great lab with wonderful people and wonderful projects, Guy Montelione. I cannot find enough words to explain my gratitude for everything he did for the past 6 years. He gave me the opportunity to work on really
exciting and challenging projects. With every project I grew up a little more as a scientist. He has been a big inspiration, from every meeting we had, I would come out thinking “Hell yeah! Let’s do science!” I respect the way he does science, I respect the way he shapes science, I respect him infinitely. I am just happy to be able to work with him and witness how a great mind thinks.

I also want to thank to my dearest friend, Elif, being of one the biggest support here, away from home, since the first day I arrived, and my roommate, Meric, for all the fun times we had. Many special thanks to my boyfriend, Mehul, who was there with me, from start to end.

Finally, I want to dedicate my work here to my family, my mom, Betul, and my sister, Ege, for sacrificing the youngest family member for science. Because I know how hard it is to be away from family. I love them both.
# TABLE OF CONTENTS

ABSTRACT ......................................................................................................................... ii

ACKNOWLEDGEMENTS ................................................................................................. iv

TABLE OF CONTENTS ...................................................................................................... vi

LIST OF TABLES ................................................................................................................ x

LIST OF FIGURES ............................................................................................................. xi

LIST OF ABBREVIATIONS ............................................................................................... xv

INTRODUCTION ............................................................................................................... 1

1. HYDROGEN-DEUTERIUM EXCHANGE FOR PROTEIN CONSTRUCT OPTIMIZATION ......................................................................................................................... 4

   1.1 Introduction .............................................................................................................. 4

   1.2 Methods and Materials .......................................................................................... 10

       1.2.1 Hydrogen/Deuterium Exchange with Mass Spectroscopy ......................... 10

       1.2.2 Experimental Setup and Standard Protocol ................................................. 11

       1.2.3 HDX-MS data analysis ................................................................................ 13
3.2.2 Structure Determination ................................................................. 36

3.2.3 Polycistronic Co-expression of CDK2 and CDK2AP1 .................. 38

3.2.4 Size Exclusion Chromatography for Binding Studies ............... 39

3.2.5 IKKE Phosphorylation of full-length CDK2AP1 ......................... 40

3.3 Results ................................................................................................. 40

3.3.1 HDX-MS Analysis ....................................................................... 40

3.3.2 CDK2-AP1(61-115) is a dimer in solution ................................ 42

3.3.3 Solution Structure of CDK2AP1 (61-115) ................................. 45

3.3.4 C105A mutation does not disrupt DOC-1(61-115) dimerization ... 53

3.3.5 CDK2AP1 is phosphorylated by IKKe ......................................... 54

3.3.6 CDK2:CDK2AP1 interaction ....................................................... 58

3.4 Discussion .......................................................................................... 62

4. SOLUTION NMR STRUCTURE OF PEPTIDE METHIONINE SULFOXIDE
REDUCTASE FROM BACILLUS SUBTILIS ......................................................... 65

4.1 Introduction ....................................................................................... 65

4.2 Methods and Materials .................................................................. 67
4.2.1 Protein Cloning, Expression and Purification............................... 67

4.2.2 NMR data collection and structure calculation................................. 68

4.3 Results............................................................................................. 69

4.3.1 Samples Used for Data Collection.................................................. 69

4.3.2 Chemical Shift Assignments............................................................ 71

4.3.3 The Solution structure of MsrB ...................................................... 75

4.3.4 MsrB Dynamics ............................................................................. 80

4.3.5 X-ray Crystal Structure................................................................. 83

4.4 Discussion and Conclusions ............................................................... 86

REFERENCES ........................................................................................ 90

A. APPENDIX.......................................................................................... 108
LIST OF TABLES

Table 1.1 List of NESG targets studied by HDX-MS by Seema Sharma......................... 8

Table 1.2 List of the proteins studied with HDX-MS for construct optimization ............ 16

Table 3.1 Summary of NMR and structural statistics for human CDK2AP1(61-115)..... 46

Table 4.1 Structure calculation statistics and quality scores for MsrB......................... 77
LIST OF FIGURES

Figure 1.1 NESG Process of Optimizing Sequences of Partially Disordered Proteins for NMR Structure Analysis and/or Crystallization......................................................... 7

Figure 1.2 $^{15}$N-HSQC spectra for NESG targets ER553, LkR15, SaR32 and VpR68 for full-length (a), and optimized (b) constructs................................................................. 9

Figure 2.1 The peptides used for HDX-MS analysis of Smad3. ........................................... 24

Figure 2.2 HDX-MS results on the full-length human Smad3 protein................................. 25

Figure 2.3 a) Sequence alignment of Smad3 and Smad2. b) DisMeta disorder consensus prediction results for Smad3 and Smad2. ................................................................. 27

Figure 3.1 Multiple sequence alignment of CDK2AP1 with homologues, by ClustalW. 32

Figure 3.2 HCPIN Interactome for (a) CDK2AP1 and (b) CDK2. ....................................... 34

Figure 3.3 HDX-MS data for CDK2AP1 (NESG ID: HR3057). ............................................ 41

Figure 3.4 Overlay of 600 MHz $^1$H-$^{15}$N HSQC spectra at 298 K of full-length (red) and 61-115 construct (blue) of CDK2AP1 ........................................................................ 42

Figure 3.5 (a) Static light scattering chromatogram for CDK2AP1(61-115) in the presence of DTT. (b) Rotational correlation time vs MW ......................................................... 44

Figure 3.6. Solution structure of residues 61-115 of CDK2AP1........................................... 49
Figure 3.7 Disulfide mapping by MALDI-TOF of sample without (a) and with (b) DTT.. ................................................................................................................................. 51

Figure 3.8 Overlay of 600 MHz (a) $^1$H-$^{15}$N HSQC and (b) $^1$H-$^{13}$C HSQC spectra of CDK2AP1(65-115) with and without DTT ........................................................................................................ 52

Figure 3.9 Overlay of 600 MHz $^1$H-$^{15}$N HSQC spectra of wild-type and C105A mutant 54

Figure 3.10 The LC/MS chromatograms of peptides after trypsin digestion of samples before and after phosphorylation reaction.. .............................................................................................. 56

Figure 3.11 MS/MS spectrum of QLLSDYGPPSpLGYYQTGQGNSQVPQSK peptide. S46 is identified as the only phosphorylated site after IKKε phosphorylation reaction... 56

Figure 3.12 Overlay of 600 MHz $^1$H-$^{15}$N HSQC spectra of wild-type and phosphorylated full-length CDK2AP1 .................................................................................................................. 57

Figure 3.13 Size exclusion chromatography for CDK2, CDK2AP1(65-115) and mixture. ...................................................................................................................................................... 59

Figure 3.14 Size exclusion chromatography for CDK2, CDK2AP1 and mixture........ 60

Figure 3.15 SDS-GEL of the fractions from co-purification assay for CDK2: CDK2AP1 co-expression system. .......................................................... 61

Figure 3.16 Size exclusion chromatography for CDK2, CDK2AP1(S46p) and mixture. 62
Figure 4.1 Projection of HNcaCO spectra for (a) double-labeled ($^{13}$C,$^{15}$N)-sample and (b) perdeuterated triple-labeled ($^{13}$C,$^{15}$N2H) sample of *B. subtilis* MsrB ........................................ 71

Figure 4.2 The 800MHz $^{15}$N-HSQC spectrum for MsrB at 25 °C ........................................ 73

Figure 4.3 Secondary chemical shift histograms for $\text{C}_\beta$ atoms in folded proteins for Proline residues ......................................................................................... 75

Figure 4.4 Solution NMR structure of major conformational state of fully-reduced MsrB ........................................................................................................................................ 78

Figure 4.5 Stereo image of the active site of MsrB. .......................................................... 79

Figure 4.6 Backbone $^{15}$N relaxation measurements at 600 MHz, 25°C .................. 80

Figure 4.7 The solution structure of MsrB, the residues with low hetNOE values are colored red, the residues with high $R_2$ rates are colored blue ................................................. 83

Figure 4.8 The superimposition of 2.6 Å X-ray crystal structure (PDB ID: 3E0O) and sparse-constraint NMR structures for MsrB from *B. subtilis*. .................................................. 84

Figure 4.9 Secondary shifts of $\text{C}_\alpha$ atoms for N-terminal residues ......................... 86

Figure A.1 HDX-MS results for NESG target BfR218. ................................................. 108

Figure A.2 HDX-MS results for NESG target ER554 .................................................... 109

Figure A.3 HDX-MS results for NESG target HR2891 ................................................. 110
Figure A.4 HDX-MS results for NESG target HR2921 ........................................ 111

Figure A.5 HDX-MS results for NESG target HR3018 ........................................ 112

Figure A.6 HDX-MS results for NESG target HR3070 ........................................ 113

Figure A.7 HDX-MS results for NESG target HR3074 ........................................ 114

Figure A.8 HDX-MS results for NESG target HR3153 ........................................ 115

Figure A.9 HDX-MS results for NESG target HR3159 ........................................ 116

Figure A.10 HDX-MS results for NESG target SmR84 ........................................ 117

Figure A.11 HDX-MS results for NESG target SpR36 ........................................ 118

Figure A.12 HDX-MS results for NESG target SvR375 ........................................ 119
LIST OF ABBREVIATIONS

CDK2: Cyclin-dependent kinase 2

CDK2AP1: CDK2 associated protein 1

DOC-1: Deleted in oral cancer 1

DTT: Dithiothreitol

FA: Formic acid

HDX-MS: Hydrogen/deuterium exchange with Mass Spectrometry

HPLC: High performance liquid chromatography

HSQC: Heteronuclear single quantum correlation

IPTG: Isopropyl β-D-1-thiogalactopyranoside

MCS: Multiple cloning site

MES: 2-(N-morpholino)ethanesulfonic acid

MH1, MH2: Mad-homology domain 1,2

MOPS: 3-(N-morpholino)propanesulfonic acid

MsrB: Methionine sulfoxide reductase

NESGC: Northeast Structural Genomics Consortium
NMR: Nuclear magnetic resonance

NOE: Nuclear Overhauser Effect

NOESY: Nuclear Overhauser Effect Spectroscopy

PDB: Protein data bank

PSI: Protein Structure Initiative

PSVS: Protein structure validation suite

RDC: Residual dipolar coupling

RMSD: Root mean square deviation

TCEP: tris(2-carboxyethyl)phosphine

TGF-β: Transforming growth factor
INTRODUCTION

Proteins are dynamic entities. They fluctuate between similar, but distinct, conformational states. This conformational plasticity is fundamental to their biological function in the cell, and contributes to both the thermodynamic and kinetic aspects of their functions. The strong link between enzyme dynamics and function has been studied extensively on different systems [1-4]. Many of these studies demonstrate that conformational states spanned during native state dynamics of the protein often correspond to the predominant conformations they attain in their functionally active states [4].

In contrast to the general view that a protein should have a specific 3D structure in solution for its activity there are many proteins which do not have “native state” structure. Rather, they exhibit “intrinsic disorder”, which is often fundamental to their function. In contrast to the enzymes mentioned above, these proteins require higher level plasticity and flexibility for their function in the cell. It is observed that these intrinsically disordered proteins are especially enriched in cell signaling pathways and in transcriptional and translational regulation [5]. The intrinsic disorder in proteins is advantageous in protein signaling, providing readily available post-translational modification sites, and conformational flexibility that is often required in order to provide increased promiscuity in binding partners. Conformational flexibility of intrinsically-disordered proteins may also provide important entropic contributions in determining the affinities of protein-protein interactions [6-11].
In my thesis, I focus on understanding the structural and dynamic features of proteins. For this I studied my favorite proteins with different experimental methods including Hydrogen/Deuterium Exchange with Mass Spectrometry (HDX-MS) and Nuclear Magnetic Resonance Spectroscopy (NMR).

As part of NIH Protein Structure Initiative (PSI) Northeast Structural Genomics (NESG) effort, I worked on identifying the disordered and flexible parts of proteins using HDX-MS analysis (Chapter 1). These data were obtained primarily for construct optimization needed for high-throughput structure determination, but in some cases have also provide important insights into structure-dynamics-function relationships. I studied thirteen NESG protein targets under this project.

Additional to these NESG targets I also studied human Smad3, a very important protein in TGFβ-signaling, regulating many key cellular processes including proliferation, differentiation, adhesion, apoptosis, and immune suppression. Smad3 is composed of two well-folded domains, MH1 and MH2, which are connected by a linker domain. The structural features of MH1 and MH2 domains were determined experimentally but there were no experimental data revealing the structural features of the linker domain. In Chapter 2, I describe my HDX-MS studies on full–length human Smad3, where I present the first experimental data on structural characterization of the linker region, which is largely intrinsically disordered.

I next focused on structure-function studies of the cyclin-dependent kinase 2 associated protein 1 (CDK2AP1, NESG ID: HR3057), which I had initially studied in Chapter 1. This protein, corresponding to gene doc-1 (deleted in oral cancer-1), is a
cyclin-dependent kinase 2 inhibitor, one of the regulators of cell cycle and has important role in cancer biology. HDX-MS studies revealed that N-terminal 60 residues are intrinsically disordered. CDK2AP1 is also involved in the Mi-2/NurD transcriptional-regulation complex, an important complex regulating epigenetic gene expression. The details of the structural features of the C-terminal ordered region and new findings are explained in Chapter 3.

In the final part of my thesis work, I pursued structural and dynamics studies of a biologically important enzyme, methionine sulfoxide reductase B (MsrB). MsrB has the critical function of reducing oxidized methione (methionine sulfoxide) residues in proteins back into methionine, and hence plays a role in the aging process. We solved the 16 kDa MsrB structure using “sparse constraint” approach, using limited NMR distance constraints that could be obtained on a perdeuterated protein sample. The resulting structure is very similar to an independently determined 2.6 Å crystal structure of the same protein. The differences observed between our sparse-constraint NMR structure and the crystal structure provide guidance on areas than require improvement in order to make this sparse-constraint approach robust for accurate protein NMR structure determination. The details of this structure are described in Chapter 4.
1. HYDROGEN-DEUTERIUM EXCHANGE FOR PROTEIN CONSTRUCT OPTIMIZATION

1.1 Introduction

In contrast to general view of proteins exhibiting a certain equilibrium native structure, it has been recently acknowledged that many proteins evolved to lack any kind of ordered equilibrium structure for either the entire of the protein or for a significant portion [12-14]. Based on disorder predictions on entire proteomes of more than 30 organisms suggested that 6-33% bacterial proteins contained regions <40 residues predicted to be disordered. For eukaryotes this prediction increases to 35-51% of the proteome [15]. The increase in the occurrences of disorder in higher organisms can be linked to the complexity of the protein-protein interaction networks [16]. The plasticity of these proteins is especially advantageous in providing different binding sites for multiple binding partners, providing flexible linkers, reorganizing the connected functional domains, providing different binding sites for different interactions, and providing entropic contributions to protein-protein interactions. These regions often include the post-translational modification sites, allowing solvent accessibility [12-16].

Throughout the last decade the structural genomics groups have been working on determining protein structures as well as exploring ways of high-throughput structure determination. In studying protein structures with X-ray crystallography, obtaining well-diffracting protein crystals has been a serious bottleneck. Even though the unstructured regions found in proteins may be functionally important, it has been observed that these highly flexible regions may prevent formation of high-quality crystals by inhibiting
formation of stable crystal contacts, making the crystal formation thermodynamically unfeasible and/or disrupting the homogeneity of the protein ensemble due to high susceptibility to protease cleavage. Thus the problem of obtaining diffraction quality crystals could be partially overcome by removing these unstructured regions of the protein.

In earlier studies, crystallization success was shown to improve by in-situ limited proteolysis by addition of minute amounts of proteases into crystallization media to enable cleavage of disordered tails [17, 18]. In many cases limited proteolysis with Mass Spectroscopy (LP-MS) was also used to identify the disordered regions and to allow engineering of protein constructs lacking the problematic region(s) [19]. However, cleavage at internal loops, rather than N- or C-terminal disordered “tails” may result in the wrong interpretation of the data.

In the last decade it was demonstrated crystallization success for structure determination for partially disordered proteins by X-ray crystallography could be improved by identifying disordered regions of the protein by Hydrogen/Deuterium Exchange Mass Spectroscopy (HDX-MS) [20-22]. This method proved to be very effective, requiring very small quantities of protein (micrograms) and minimal data analysis, thus making it perfect tool for high-throughput studies of partially disordered proteins.

Nuclear Magnetic Resonance (NMR) spectroscopy is an alternative tool, complementary to X-ray crystallography, for structure determination, providing high-resolution solution structures as well as information on dynamic characteristics and
interactions of proteins and other biomolecules. Determination of 3D structures of proteins with partially unfolded regions can be very complicated for NMR as well, since these segments cause high-intensity, overlapping peaks in NMR spectra. Different NMR experiments could be used to identify the flexible regions [23-25], however the amount of protein and extensive labor required for data analysis makes these methods highly unfavorable. In this study it is shown that the adaptation of the HDX-MS protocol for high-throughput structural studies by NMR can be highly effective for providing sample with better properties for rapid and accurate determination of resonance assignments and structures. Where appropriate, these assignments and/or structures of the ordered “core” regions of partially-disordered proteins provides a solid starting point for NMR studies of the full-length protein or studies of complexes using either NMR or X-ray crystallography.

As part of the technology development of the Northeast Structural Genomics Consortium (NESGC) project, a process has been developed to identify and remove disordered or flexible N- and C-terminal regions of proteins in order to provide samples that are more amenable to high-throughput NMR analysis and crystallization [26]. First, disordered regions of target proteins are identified using the Dis-Meta server, which is a collection of different disorder prediction tools available [27-38]. If a consensus of these multiple bioinformatics methods is observed, suggesting limited number of disordered regions at one or both ends of the polypeptide chain, new constructs are made based on this information. When the bioinformatics methods fail to provide a consensus result, disordered regions are identified experimentally by HDX-MS experiments [26]. New
constructs are designed, excluding the highly flexible or disordered residues on either terminal region, based on these experimental data and secondary structure predictions (Figure 1.1).

![Diagram of NESG Process](image)

**Figure 1.1 NESG Process of Optimizing Sequences of Partially Disordered Proteins for NMR Structure Analysis and/or Crystallization.**

A set of NESG targets, which were studied by HDX-MS, analyzed earlier by Seema Sharma [26], is given in Table 1.1 and Figure 1.2. Six proteins, for which the disorder prediction results were inconclusive, have been studied by HDX-MS. For four of these proteins the structures of the optimized constructs were determined by NMR and deposited in the Protein Data Bank (PDB IS’s: 2K1S, 2K3D, 2K5D, 2JZ5, respectively). These results suggest that by incorporating HDX-MS into pipeline we have achieved a high salvage-success rate of ~70%.
Table 1.1 List of NESG targets studied by HDX-MS by Seema Sharma.

<table>
<thead>
<tr>
<th>NESG ID</th>
<th>Result Summary</th>
<th>Status of New Constructs</th>
</tr>
</thead>
<tbody>
<tr>
<td>ER553</td>
<td>Disordered N-term</td>
<td>NMR In PDB</td>
</tr>
<tr>
<td>SaR32</td>
<td>Disordered C-term</td>
<td>NMR In PDB</td>
</tr>
<tr>
<td>LkR15</td>
<td>Disordered N- and C-term</td>
<td>NMR In PDB</td>
</tr>
<tr>
<td>VpR68</td>
<td>Disordered N-term</td>
<td>NMR In PDB</td>
</tr>
<tr>
<td>HR2951</td>
<td>Disordered C-term</td>
<td>Poor Exp/ Sol</td>
</tr>
<tr>
<td>DrR44</td>
<td>Disordered C-term</td>
<td>Poor Exp/ Sol</td>
</tr>
</tbody>
</table>
Figure 1.1 $^{15}$N-HSQC spectra for NESG targets ER553, LkR15, SaR32 and VpR68 for full-length (a), and optimized (b) constructs
1.2 Methods and Materials

During my dissertation I continued working on HDX-MS part of this project. We carried out HDX-MS studies for thirteen NESG protein targets in order to experimentally identify the intrinsically-disordered regions in these proteins. Results on these 13 proteins, along with a summary of the methodological improvements made relative to the published experimental protocol [26] in order to make it more suitable for a broader set of proteins are presented in this Thesis chapter.

1.2.1 Hydrogen/Deuterium Exchange with Mass Spectroscopy

Acidic hydrogen in proteins such as –OH, –NH₂, –SH, peptide amide hydrogen continuously and reversibly interchange with the hydrogen in water. The hydrogen of –OH, –NH₂, and –SH groups exchange very rapidly, which makes the real-time observation of this reaction very difficult. However, the exchange rates for the amide hydrogen were shown to be highly dependent on local environment, such as secondary and tertiary structures and solvent accessibility. Based on the observations and calculations, the amide hydrogen exchange rates for random coil were found to be typically in 10-1000 sec range [39]. In a folded polypeptide chain the amide proton, exchange half times as long as weeks could be observed, due to the stability of the peptide and protection of the amide group from the solvent. Thus the measurement of real-time amide hydrogen exchange is a very sensitive and unique probe for protein structural studies.
The real-time $^1$H/$^2$H exchange reaction can be achieved by isotope labelling, which can be monitored by mass spectrometry [40]. The isotope incorporation to backbone amides can be achieved by diluting the protein solution with buffer with high deuterium content (>99%), yielding high concentrations of $^2$H$_2$O in the reaction medium, forcing the exchange to proceed in one direction. At different preset time points the reaction is quenched, sample is subjected to a proteolysis reaction, then the proteolysis products are separated by high-performance liquid chromatography and analyzed by mass spectrometry. The increase in the peptide mass, due to exchange of hydrogen with deuterium, is measured by the comparison of the centroids of the mass envelopes of the peptide of interest corresponding to before and after the exchange reaction.

1.2.2 Experimental Setup and Standard Protocol

Protein $^1$H/$^2$H exchange experiments were conducted following the methods described by Spraggon et al [22]. A 5 μl aliquot of protein sample (~ 25 - 50 μg, in preferred buffer for structural studies) was mixed with 15 μl of deutarated buffer and incubated on ice for set time points before being quenched by the addition of 30 μl of quench solution (Buffer A: 0.8% formic acid and 1.6 M Guanidine HCl) and was frozen immediately on dry ice. For the 0 time point experiment, 5 μl of protein sample was mixed 15 μl of buffer in H$_2$O and quenched and analyzed in the same way as the $^1$H/$^2$H exchanged samples.

For the correction of back exchange, a completely exchanged sample was prepared as described by Hamuro et al [41]; 5 μl of the protein sample was mixed with 15
µl of 0.5% formic acid in $^2$H$_2$O for complete unfolding reaction and incubated at room temperature for 24 h. The sample was then quenched and analyzed using the same conditions as other $^1$H/$^2$H exchange experiments.

High Performance Liquid Chromatography (HPLC) solvent bottles, connection lines including the sample loop, the pepsin column and the analytical column were all kept on ice. For analysis, frozen samples were thawed on ice and manually injected into a pre-column (66 µl bed-volume, Upchurch) packed in-house with immobilized pepsin (PIERCE) at a 50 µl / 30 sec flow-rate and followed by an injection of 100 µl 0.1% formic acid in 1 min into a 200 µl sample loop. By valve switching, the sample loop was brought online with the C18 HPLC column (Discovery, BioWide Pore C18-3, 5 cm X 2.1 mm 3um, Supelco). After a three minute wash step with 200 µl/min of 0.1% formic acid, the digested peptides were separated by a linear acetonitrile gradient of 2-50% in 17 min at 200 µl/min and analyzed downstream by an electrospray-linear ion-trap mass spectrometer (LTQ, Thermo). For measurement of the mass shift in $^1$H/$^2$H exchange experiments, MS was set to perform full-scan in the range of 300-2000 in profile mode for the entire HPLC run. For peptide identifications, the same sample was processed the same way as the 0 time samples. The mass spectrometer was set to perform one full-scan MS in the m/z range 300-2000 followed by zoom scans of top 5 most intense ions and MS/MS of multiply charged ions. Dynamic exclusion conditions were set to exclude parent ions that were selected for MS/MS once in 30 sec, and the exclusion duration was 60 sec.
1.2.3 **HDX-MS data analysis**

Acquired MS/MS data was searched using Sequest software against a homemade protein sequence database composed of 83095 entries. The search parameters were set to use no enzyme and parent peptide tolerance of +/- 2 amu and fragment ion tolerance of +/-1 amu. The search results yield the list of the identified proteolysis product peptides and the assignment of each of these peptides to corresponding spectral peaks.

The percent deuterium incorporation for each peptide is calculated based on the shift observed on the centroid of the mass peak based on the formula Zhang & Smith [40]:

\[
D_t = \frac{m(\text{labeled}, t) - m(\text{unlabeled})}{m(\text{control}) - m(\text{unlabeled})} \times 100,
\]

where \( m(\text{labeled}, t) \) corresponds to mass peak centroid for the exchange reaction for duration of \( t \), \( m(\text{control}) \) and \( m(\text{unlabeled}) \) correspond to fully deuterated and non-deuterated cases.

1.2.4 **Protein Cloning, Expression and Purification**

All proteins used for the \(^1\text{H}/^2\text{H}\) exchange mass spectrometry experiments and NMR analysis were expressed, cloned and purified based on methodologies previously published by our laboratory [42]. Briefly, the full-length gene constructs of proteins of interest were cloned into modified pET15 expression vectors containing an N-terminal purification tag (MGHHHHHHHSH) or pET21 expression vectors containing a C-terminal affinity tag (LEHHHHHH) [43]. All vectors were transformed into codon enhanced BL21
(DE3) pMGK *E. coli* cells, which were cultured at 37 °C in MJ minimal medium [44]. Protein expression was induced at reduced temperature (17 °C) by IPTG (isopropyl-β-D-thiogalactopyranoside). Expressed proteins were purified using an AKTAexpress (GE Healthcare) two-step protocol consisting of HisTrap HP affinity and HiLoad 26/60 Superdex 75 gel filtration chromatography. Sample purity was confirmed using SDS-PAGE and MALDI-TOF mass spectrometry.

1.3 Results

1.3.1 Modifications to Existing Protocol

Obtaining full coverage of all residues in protein sequence in HDX-MS experiments by sequencing experiments is important to map the dynamic properties throughout the protein. For the proteins, where the original protocol (see Section Methods and Materials) did not yield good coverage, different approaches were explored for completing the residue coverage.

The optimum peptide length for HDX-MS analysis, after pepsin digestion, is 7-15 residues. It is not possible to get exchange data for shorter peptides, due to high exchange rates at the terminal regions, and the long peptides do not give good enough resolution to accurately map the dynamics for a given residue along the protein sequence. The digestion duration of the protein can be modified for the cases where the observed peptide lengths are outside of the desired range. Additionally, different digestion times can be explored for the cases where low residue coverage is observed by the original
protocol. Exposing the protein to pepsin for different durations may result in different set of peptides, which may increase the overall residue coverage.

It was observed that the data collection mode of the spectrometer also affects the peptides detected and identified peptides by the instrument and the software, respectively. In the original protocol, the instrument is set to do a full-scan for the m/z range of 300-2000 followed by zoom scan for the five most intense ions and the corresponding MS/MS scans for these ions. In this original protocol, zoom-scans before each MS/MS scan take additional time, which may cause low populated ions to be depleted until their MS/MS scans are carried out, thus they will not be detected and identified by the receiver. Hence, we observed improved sequence coverage for several cases when the zoom-scan is not detected during data collection.

The unfolding conditions are also important in sequence coverage and quality of the data. For several cases it was observed that some proteins were resilient to denaturating conditions of the quench solution (Buffer A). This results in low sequence coverage as well as non-accurate estimation corrected $^1$H/$^2$H exchange levels. Hence, we suggested two additional denaturation conditions at quenching stage, where Buffer A does not give satisfactory results:

**Buffer A:** 1.60 M GuHCl, 0.8% FA

**Buffer B:** 2 M Urea, 0.8% FA

**Buffer C:** 1 M Urea, 1 M TCEP
For the cases we studied it was concluded that buffer B offered more efficient protein denaturation, thus for the rest of the study buffer B is selected as the first choice for sequencing experiments.

1.3.2 Results of HDX-MS Analysis on NESG Target Proteins

For this project I analyzed HDX-MS data for thirteen NESG protein targets. These results are being used to design new constructs, some of which have been brought through the NESG sample production pipeline. These results are summarized in Table 1.2. Specific results for target HR3057, cyclin-dependent kinase 2 associated protein (CDK2AP1), a cancer-associated human protein, are presented in Chapter 3 and results for the rest of the targets are shown in Appendix A. The percent deuteration is calculated as described in Section 1.2.3 and represented as heat map where color coding is given in the inset of each figure. The residues are colored “white” where no data was available.

<table>
<thead>
<tr>
<th>NESG ID</th>
<th>Number of res</th>
<th>Result Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>BfR218</td>
<td>91</td>
<td>Ordered</td>
</tr>
<tr>
<td>ER554</td>
<td>202</td>
<td>Flexible loops, tails</td>
</tr>
<tr>
<td>HR2891</td>
<td>96</td>
<td>Disordered at C-term</td>
</tr>
<tr>
<td>HR2921</td>
<td>134</td>
<td>Disordered N- and C- terms</td>
</tr>
<tr>
<td>HR3018</td>
<td>313</td>
<td>Flexible loops</td>
</tr>
<tr>
<td>HR3057</td>
<td>115</td>
<td>Disordered at N-term</td>
</tr>
<tr>
<td>HR3070</td>
<td>307</td>
<td>Disordered at N-term</td>
</tr>
<tr>
<td>HR3074</td>
<td>261</td>
<td>Ordered</td>
</tr>
<tr>
<td>HR3153</td>
<td>340</td>
<td>Disordered loop</td>
</tr>
<tr>
<td>HR3159</td>
<td>160</td>
<td>Disordered at N- and C-term</td>
</tr>
<tr>
<td>SmR84</td>
<td>158</td>
<td>Disordered at N- and C-term</td>
</tr>
<tr>
<td>SpR36</td>
<td>172</td>
<td>Ordered</td>
</tr>
<tr>
<td>SvR375</td>
<td>140</td>
<td>Disordered at C-term</td>
</tr>
</tbody>
</table>
The HDX-MS analysis of targets BfR218, HR3074, HR3018 and SpR36 revealed that these proteins were highly ordered (Figure A.1, Figure A.7, Figure A.5, Figure A.11). Low dispersion in amide chemical shifts, as seen for these cases, may be observed for unstructured proteins or in helical proteins. The secondary structure predictions by two different prediction softwares (PROFsec [45] and PSIPRED [46]) suggest strongly that BfR218, HR3074 and HR3018 are highly helical structures. SpR36 is also predicted to be mostly helical by PSIPRED. Hence the $^{15}$N-HSQC profile observed for these three targets in screening can be explained by the predicted helical content.

HDX-MS data for ER554 (Figure A.2), indicated that the structure is highly ordered while slightly higher flexibility is observed for N-terminal 20 residues and for residues ~125-145, which may correspond to a loop region. HR2891, on the other hand, was found to be highly disordered at the C-terminus, where only ~30-40 residues on the N-terminal region found to have an ordered structure. For HR2921 we also observe highly flexible N- and C-terminal tails, where two predicted helices around residues ~30-50 and ~60-80 make the ordered structure (Figure A.4). The HDX-MS data reveal that HR3070 has a highly accessible N-terminal region composed of ~30 residues and possibly a dynamic loop at residues ~60-90 (Figure A.6). For HR3159 we observe a highly flexible internal loop for ~100 residues (~70-170), which makes up ~1/3 of the protein (Figure A.9). This outcome is interesting and suggests a detailed analysis on the sequence. The disordered region identified by HDX-MS analysis may correspond to an internal loop within a single folded structure or it may correspond to a loop between two domains of the protein, each composed of residues ~1-70 and ~175-340, respectively.
The data for HR3159 indicate that the protein is highly ordered, however higher level of solvent accessibility is observed for ~30 residues on each terminal region of the protein compared to the core (Figure A.9). For SmR84 no disordered region could be identified by HDX-MS experiments (Figure A.10). Finally, for SvR375 it is observed that N-terminal region of the protein shows high levels of protection from solvent exchange while the C-terminal regions shows higher solvent accessibility, even though we do not observe high exchange rates as one would observe for disordered proteins (Figure A.11).

1.4 Discussion

Once HDX-MS data for each protein is obtained they are studied carefully to identify the possible disordered regions. For the cases, where increased flexibility is observed at the terminal regions new constructs are designed to exclude these flexible regions. Since the resolution of the data is at the level of 5-10 residues, multiple constructs are proposed with different termini, excluding different sizes of peptides at each end. Since the internal loops may contribute to the stability of the structure, those residues, which were found to be highly flexible but not located at the terminal regions, are kept intact during designing new constructs.

These proteins discussed here were studied by HDX-MS because the predictions from the DisMeta server were not conclusive, mostly due to lack of consensus between different prediction tools used within DisMeta. For these cases the HDX-MS studies were necessary to identify the disordered regions accurately. For some cases, as in BfR218, the consensus indicated that there was no disordered region in the protein, which was confirmed by the HDX-MS studies. However, for some cases the consensus points to
interleaved disordered regions, as in HR3153, for which a very different disorder patterns was identified based on the experimental data. This indicates one should use caution when a clear consensus between different prediction tools is not achieved.

Here I described an important use of the HDX-MS tool in identifying the disordered regions of the protein for planning a future construct for structure determination and summarized the findings in Table 1.1. This low cost and relatively non-invasive method can also be used in identifying the protein-protein interaction sites and allosteric effects; it can provide insights on structural changes and dynamics accompanying post-translational modifications and interactions with other molecules, and it can also be used to understand equilibrium folding/unfolding mechanisms [41, 47-49].
2. THE INTERDOMAIN LINKER REGION OF HUMAN SMAD3 IS INTRINSICALLY-DISORDERED

2.1 Introduction

Transforming growth factor beta (TGF-β) is an important cytokine, which along with the other related proteins, regulates many key cellular processes including proliferation, differentiation, adhesion, apoptosis, and immune suppression [50]. In the TGF-β pathway, TGF-β-activated membrane receptor protein kinases transfer the ligand-binding signal to intracellular substrate proteins, Smads, which in turn propagate the signal into the nucleus where they function as transcription factors [51-54]. There are eight identified Smad proteins, which are classified, based on both sequence and structure, into three groups: receptor-regulated Smads (R-Smad), common Smads (Co-Smad) and inhibitory Smads (I-Smad). Smad2 and Smad3 from the R-Smad family are activated by direct phosphorylation by the TGF-β receptor kinase at their SSXS phosphorylation motif at the C-terminal tail. Activated Smad2 and Smad3 proteins homooligomerize, form complexes with co-Smad Smad4, and then accumulate in the nucleus where they regulate transcription of target genes [51-54].

Smad3 is composed of two highly conserved N- and C- terminal domains, referred to as MAD Homology 1 (MH1) and MAD Homology 2 (MH2) domains, respectively. The MH1 domain contains the DNA binding domain and transcription factor binding sites as wells as phosphorylation sites for protein kinase C (PKC) and GSK3 (glycogen synthase kinase 3) [51-57]. It may also be phosphorylated by the Ca^{++}- and calmodulin-dependent kinase II (CamKII) [58]. The MH2 domain is responsible for
Smad-pathway activation by phosphorylation at the SSXS site by the TGF-β type I receptor after its recruitment to the receptor complex through interaction with Smad anchor for receptor activation (SARA) [51-54, 59]. It is also important for cross-talk between different pathways, and mediates interactions with other Smad molecules to form homo- and hetero-complexes and interactions with other transcription factors [51-54]. The MH2 domain is also essential for Smad3 transcriptional activity. These two well-folded domains are connected by a linker region with high proline content [51-54]. The linker region is important in oligomerization prior to nuclear transport [60], and it is also essential for transcriptional activation [61, 62]. Furthermore, it contains demonstrated as well as suspected phosphorylation sites for multiple kinases, such as members of the cyclin-dependent kinase (CDK) family, members of the mitogen-activated protein kinase (MAPK) superfamily including ERK (extracellular-signal regulated kinase), JNK (c-Jun N-terminal kinase) and p38, GSK3, G protein-coupled receptor kinase 2 (GRK2), and CamKII [58, 63-82]. Phosphorylation of the linker region by the various kinases differentially affects Smad3 activity in a context-dependent manner. The linker region also contains a recognition site for Nedd4L, an E3 ubiquitin ligase, which can lead to the degradation of the Smad3 protein in response to TGF-β [66, 83].

Three-dimensional X-ray crystal structures are available for the MH1 (residues 1-139) domain bound to DNA [55] and for the MH2 (residues 232-425) domain in the apo-state [84], in complex with the Smad binding domain of SARA [84], and in a phosphorylated state in complex with Smad4[85]. To date, however, there have not been
any reports on structure of full-length Smad2 or Smad3 proteins; in particular there is no experimental information about the structure of the linker regions. The high proline content of the linker region suggests lack of ordered structure, but this is not certain in the absence of structural or biophysical studies on the full-length protein. Since multiple phosphorylation sites within the linker region are targets for a number of kinases, understanding the structural features of this part of the protein is critical for understanding the functional role of the interdomain linker.

Amide hydrogen-deuterium exchange (HDX) studies provide an important method for characterizing stable hydrogen-bonded structures in proteins [86]. The exchange rates for amide hydrogens are highly dependent on solvent accessibility and local structural environment. Slower exchange rates are observed for amide hydrogens involved in hydrogen bonds within secondary and tertiary structures proteins compared to the intrinsically disordered regions, due to the reduced solvent accessibility; and continuous stretches of backbone amides with fast exchange rates are observed for intrinsically-disordered regions of proteins, including interdomain flexible linker regions. The analysis of amide HDX rates with Mass Spectrometry (HDX-MS) provides a reliable method for assessing such features of a protein [40, 87, 88]. In this study, we report results of HDX-MS analysis for full-length Smad3, providing the first experimental evidence that the linker region is intrinsically-disordered and largely solvent accessible in full-length human Smad3.
2.2 Materials and Methods

The sample preparation and HDX-MS analysis were carried out as explained in Chapter 1.

2.3 Results

The results of HDX-MS analysis for human Smad3 are summarized in Figure 2.2. The color coding is shown in the inset of the figure, and the interdomain linker region is shown within the brackets. This HDX-MS analysis includes data from 81 multiple charged peptide ions from 53 peptide fragments; out of these data 12 peptide ions from 9 fragments provide coverage for the interdomain linker region. For some parts of the protein, data coverage was not achieved because the mass spectrometer could not detect any peptides corresponding to these regions, or no reliable HDX-MS data could be obtained for the detected peptides. These parts are color-coded in white.
Figure 2.1 The peptides used for HDX-MS analysis of Smad3. The secondary structures for the MH1 and MH2 domains are also indicated, based on structural data obtained from the Protein Data Bank (ID: 1OZJ for the MH1 domain and 1MJS for the MH2 domain). The linker region is indicated with brackets.
Figure 2.2 HDX-MS results on the full-length human Smad3 protein. The color coding, as indicated in the inset, describes the average percentage of deuteration levels, derived by multiple overlapping peptide fragments, of each residue after either 10 or 100 second exposure to deuterium; white color coding is used where no data are available. The linker region is indicated with brackets, and proposed Proline-directed phosphorylation sites in the linker are labeled with purple stars, and the non-proline-directed GRK2 site with a green star. The secondary structures for the MH1 and MH2 domains are also indicated, based on structural data obtained from the Protein Data Bank (ID: 1OZJ for the MH1 domain and 1MJS for the MH2 domain). Polypeptide segments exhibiting rapid amide exchange are indicated in red and orange, while segments exhibiting slow amide exchange rates are indicated in blue and green. Some green color-coded segments show the same color when exposed to deuterated solvent for 10 seconds versus 100 seconds; the deuteration levels for these segments were in fact increased when exposed for 100 seconds compared to 10 seconds, but they were still in the same color range.
As shown in Figure 2.2, high deuteration levels are observed even at the shortest 10-second exposure to solvent deuterium for the interdomain linker region, demonstrating that the backbone amides in this region of full-length Smad3 are solvent exposed. This continuous segment of amide protons with fast HDX rates, including residues 140-185, indicates an intrinsically-disordered interdomain linker. The C-terminal part of the linker exhibits somewhat lower exchange rates, suggesting lower solvent accessibility, and potentially some transient local structure. The SSXS motif in the C-tail also shows high deuteration levels at short HDX exchange times, consistent with the notion that the C-tail is accessible for phosphorylation by the TGF-β receptor. Although the sequence coverage of the data is not 100%, the available data are adequate for qualitative assessment of the solvent accessibility and structural flexibility for the entire protein. These results reveal that the overall structure of the Smad3 protein features well-ordered structures in MH1 and MH2 domains, connected by an intrinsically-disordered linker.

Disorder prediction methods are also reasonably reliable at identifying intrinsically-disordered regions of proteins. The DisMeta server (http://www-nmr.cabm.rutgers.edu/bioinformatics/disorder/) uses 10 different disorder prediction methods to provide a consensus analysis of intrinsically-disordered regions of proteins. DisMeta results for human Smad3 are shown in Figure 2.3b. The predicted results agree well with the experimental HDX-MS data; i.e. the interdomain linker region and the C-terminal tail segment containing the SSXS phosphorylation motif are intrinsically-disordered, while the MH1 and MH2 domains are predicted to be well-ordered. Similar
Figure 2.3 a) Sequence alignment of Smad3 and Smad2. The secondary structure elements are shown for Smad3, and the linker regions are shown in brackets. Proposed phosphorylation sites within the linker are boxed. b) DisMeta disorder consensus prediction results for Smad3 and Smad2. The value plotted for each residue represents the number of disorder prediction servers that predict that residue to be intrinsically-disordered. The corresponding secondary structure elements are indicated. The secondary structures of Smad3 MH1 and MH2 domains and Smad2 MH2 domain are based on published results in PDB (ID: 1OZJ, 1MJS, 1KHX, respectively). The Smad2 MH1 domain secondary structure elements are based on the Smad3 MH1 domain structure.
results, including an intrinsically-disordered interdomain linker, are predicted for the homologous human Smad2 protein (Figure 2.3b).

**2.4 Discussion**

Intrinsically-disordered interdomain linkers have been observed in several other proteins [73, 89-93] and appear to be particularly common in multidomain proteins of higher eukaryotic proteins [6]. Functional roles of such intrinsically-disordered regions may include (i) providing structure plasticity between domain orientations allowing for forming alternative complexes with different binding partners; (ii) allowing solvent access for phosphorylation and other post-translational modifications, and (iii) providing an entropic contribution that modulates the binding affinity in specific protein-protein interactions [6-11]. Though it is not yet clear whether all of these potential roles are important for the functionally-important interdomain linker region of Smad3, these data provide the first experimental evidence for an intrinsically-disordered interdomain linker in the R-Smad family of proteins.

Our findings are especially important to shed light on the structural features of the linker region. There are four proline-directed phosphorylation sites in the linker region, T179, S204, S208, S213, which are phosphorylated by a variety of kinases [63-74, 78-81]. Our results show that the N-terminal part of the linker region from amino acid 140 to amino acid 185 is highly solvent exposed. T179 is located in this region. T179 is phosphorylated by G1 CDKs, CDK2 and CDK4, in a cell cycle-dependent manner, and the phosphorylation of T179 and other sites inhibits the antiproliferative function of Smad3 [63, 78, 79, 94-96]. T179 is also phosphorylated by ERK in response to EGF
treatment [68]. Furthermore, T179 is phosphorylated by members of the CDK superfamily after TGF-β treatment [64-66]. T179 can also be phosphorylated by JNK and p38 in the presence of TGF-β [69-72, 82]. Importantly, TGF-β-induced phosphorylation of T179 provides the major site for Smad3 binding to Pin1, a proline isomerase [97]. Pin1 promotes TGF-β-induced migration and invasion of cancer cells [97]. The tumor promoting function of Smad3 and Smad2 in later stages of cancer is linked with their interaction with Pin1, which is overexpressed in many cancers [98, 99].

The S204 site is located in a part of the linker region that has slower amide exchange rates than the N-terminal region of the linker, which may indicate some transient structure in this part of the linker. Residue S204 is phosphorylated by ERK in response to EGF, and by GSK3 in response to TGF-β [64, 68, 74]. A recent study shows that GSK3 phosphorylation of S204 can increase Smad3 binding to the E3 ubiquitin ligase Nedd4L [100], which can lead to Smad3 degradation [66, 83]. Residue S208 is phosphorylated by ERK in response to EGF and by members of the CDK family, JNK, and p38 in the presence of TGF-β [64-66, 68-73]. Residue S213 is phosphorylated by CDK2 and CDK4 in a cell cycle-dependent manner [63]. While our study is not able to reveal structural features for the S208 and S213 due to the peptides containing these two sites were not detected by mass spec, we expect that the phosphorylation sites at S208 and S213 are also solvent exposed.

In addition to the proline-directed kinases, the Smad3 linker region is also phosphorylated by GRK2 [76]. Phosphorylation by GRK2 inhibits the TGF-β receptor kinase to phosphorylate the SSXS-motif in the C-terminal tail [76]. The GRK2
phosphorylation site has been mapped to S157 in the Smad3 linker region [77]. Residue S157 is within the intrinsically-disordered segment 140-185 of the linker region (Figure 2.2).

In summary, the linker region of Smad3 is under dynamic regulation by various kinases. Our study provides the first insights into the structural features of the linker region in the context of the full-length Smad3 protein. We have shown that the linker region, especially the N-terminal part of the linker region containing residues 140-185, is intrinsically disordered, and has a high solvent accessibility. This feature allows T179, which is a site for multiple proline-directed kinases, and S157, which is the site for GRK2, to be readily phosphorylated. Perhaps more significantly, the intrinsically-disordered linker region of Smad3 (and also of Smad2) provides structural plasticity, allowing for multiple interdomain orientations of the MH1 and MH2 domains, which may be essential to allow different interactions with alternative binding partners.
3. SOLUTION NMR STRUCTURE OF THE FOLDED C-TERMINAL
DOMAIN OF HUMAN CYCLIN DEPENDENT KINASE 2 ASSOCIATED
PROTEIN 1 (CKD2AP1)

3.1 Introduction

CDK2AP1 (p12\textsuperscript{DOC-1}) is a well known, highly conserved tumor suppressor
protein, corresponding to gene doc-1 (deleted in oral cancer) (Figure 3.1). It was first
identified by subtractive hybridization experiments using hamster oral cancer cells, where
it was observed that the gene product is absent or reduced in the malignant cells. When
transformed keratinocytes are transfected with doc-1, its expression inhibits cell growth
[101]. A ubiquitously expressed, highly conserved homolog of this gene, with 88%
cDNA identity and 98% protein sequence identity, has also been identified in different
human tissues [102, 103]. In three malignant human oral keratinocyte cell lines analyzed,
the expression of the doc-1 gene was absent or reduced beyond detection, similar to
previous results on hamster models [102]. These early findings suggested an important
role for CDK2AP1 protein in carcinogenesis.
In recent years, the importance of CDK2AP1 in various cancer cell lines has been intensely investigated. Differential expression of CDK2AP1 was observed for two phenotypes of human colorectal cancer cells, micro-satellite stable and unstable cell lines [104]. In another study of 180 gastric cancer tissues, negative expression of CDK2AP1 was reported for 78% of the tissues, which was directly correlated with more advanced tumor stages and invasion [105]. *doc-1* gene therapy on mouse models of head and neck squamous cell carcinoma (HNSCC) resulted in significant reduction in the weight, size and growth of tumors compared to control systems, suggesting CDK2AP1 as a new therapeutic target [106]. Ectopic expression of *doc-1* in transfected malignant keratinocyte hamster models was observed to elevate the number of apoptotic cells,
implying a possible role for CDK2AP1 in apoptotic pathways [107]. In addition to its role in carcinogenesis, it was observed that doc-1 was one of the 216 genes enriched in stem cells, making it an important marker defining “stemness” of the cell [108]. Furthermore, it was recently observed that CDK2AP1 is associated with embryonic development and low levels of this protein lead to embryonic lethality [109].

Westernblot and immunoblot experiments on cell lysates using GST-tagged CDK2AP1 reveal three putative binding partners (Figure 3.2), DNA polymerase α/primase (DNApol α/primase), deleted in oral cancer one related protein (DOC-1R, CDK2AP2), and cyclin-dependent kinase 2 (CDK2), where DNApol α/primase and CDK2 are important in cell cycle regulation. In vitro analyses have shown that the first six amino acids of CDK2AP1 interact with DNApol α/primase, the only protein that can initiate de novo DNA replication [110]. It was observed that the CDK2AP1:DNApol α/primase interaction negatively regulates DNA replication at the level of initiation, suggesting a new mechanism for S-phase regulation [102, 111]. However, there have been no subsequent reports in the literature to support this data or shed light on the details of this interaction.

DOC-1R, which is a close homolog of CDK2AP1 with 57% sequence identity, was reported as one of the binding partners of CDK2AP1. [112]. It was reported that DOC-1R protein was a substrate of MAP kinase and was important in microtubule organization during meiotic maturation [113]. However, there have not been any further studies elucidating the function of this protein in the cell and the details of its interaction with CDK2AP1.
Figure 3.2 HCPIN Interactome for (a) CDK2AP1 and (b) CDK2. Each node represents a protein, which interacts with the protein of interest, and the circles around the nodes indicate if the structure for that protein or a close homolog is available [114].

CDK2AP1 has also been observed to interact specifically with free, non-phosphorylated CDK2, making it the only known specific inhibitor of CDK2 [115]. CDK2 regulates G1-to-S phase transition of the cell cycle through its complexes with cyclin A and cyclin E [116]. It was observed that CDK2AP1 inhibits CDK2 kinase activity by sequestering the inactive monomer, preventing formation of complexes with cyclin A and cyclin E, and/or by directing it to the proteosome degradation pathway [115]. Thus, DNA replication is inhibited by CDK2AP1 by inhibition of CDK2 and/or inhibition of DNApol α/primase. This scenario is consistent with the previous observation that transfected malignant keratinocytes experience significantly reduced cell growth and reversion of phenotype [101]. Overexpression of CDK2AP1 consistently results in alteration of the cell cycle profile, with increased G1 phase and reduced S phase [115].
More recently, stable isotope labeling with aminoacids in cell culture (SILAC) study on NUCleosome Remodeling and histone Deacetylase (Mi-2/NuRD) chromatin remodeling complex revealed that CDK2AP1 is a subunit of Mi-2/NuRD complex [117]. Mi-2/NuRD is one of the ATP dependent chromatin remodeling complexes which plays a critical role in epigenetic gene regulation [118]. This multi-domain complex, with variable subunits, changes epigenetic markers by removing methyl and acetyl groups on DNA, thus changes the positions of nucleosome, making the DNA available for transcription factors and DNS repair proteins [118]. CDK2AP1 was discovered to be a bona fide subunit of Mi-2/NuRD complex; however the molecular function of CDK2AP1 within this complex is still unknown.

Although the importance of CDK2AP1 in cell cycle regulation and tumor formation has been clearly established, the structural features of this protein have not been reported to date. Here, we present the first report of the three-dimensional solution structure of CDK2AP1 (UniProtKB/Swiss-Prot ID, CDKA1_HUMAN; NESG ID, HR3057; hereafter referred to as CDK2AP1(61-115)). The N-terminal 60 residues of the dimeric full-length CDK2AP1 are ‘intrinsically-disordered’, on the basis of hydrogen/deuterium exchange with mass spectroscopy (HDX-MS) analysis [26]. The solution structure of C-terminal region of CDK2AP1, including residues 61-115, determined by NMR methods is a homodimeric four-helical bundle. Significantly, CDK2AP1 also contains a predicted phosphorylation site for IκB kinase epsilon (IKKε), Ser46 [119]. Here we show that IKKε phosphorylates CDK2AP1 in vitro at the predicted
Ser46 site, which is located in the intrinsically-disordered N-terminal region of the protein structure.

3.2 Materials and Methods

3.2.1 Sample preparation

Full-length CDK2AP1, CDK2AP1 (61-115), CDK2AP1 (61-115)-C105A, and CDK2 were cloned and purified based on methodologies previously published by our laboratory [42]. Briefly, the full-length gene constructs of proteins of interest were cloned into modified pET15 expression vectors containing an N-terminal purification tag (MGHHHHHHHSH) [43]. All vectors were transformed into codon enhanced BL21 (DE3) pMGK E. coli cells, which were cultured at 37 °C in MJ minimal medium [44] containing (\(^{15}\)NH\(_4\))\(_2\)SO\(_4\) as the sole nitrogen source for uniformly \(^{15}\)N enrichment, and 100% or 5% \(^{13}\)C-glucose as sole carbon source for uniform or 5% \(^{13}\)C enrichment, respectively. Protein expression was induced at reduced temperature (17 °C) by IPTG (isopropyl-\(\beta\)-D-thiogalactopyranoside). Expressed proteins were purified using an AKTAexpress (GE Healthcare) two-step protocol consisting of HisTrap HP affinity and HiLoad 26/60 Superdex 75 gel filtration chromatography. Sample purity was confirmed using SDS-PAGE and MALDI-TOF mass spectrometry.

3.2.2 Structure Determination

Samples of uniformly \(^{13}\)C, \(^{15}\)N- and 5%-\(^{13}\)C, \(U^{15}\)N-enriched human CDK2AP1 (61-115) and CDK2AP1 (61-115) -C105A for NMR structure determination were concentrated to 0.7 to 0.9 mM in 95% H\(_2\)O/5% \(^2\)H\(_2\)O solution containing 20 mM MES,
200 mM NaCl, 10 mM DTT, 5 mM CaCl$_2$ at pH 6.5. Analytical gel filtration with static scattering data demonstrates that both full-length and truncated CDK2AP1 are dimeric in solution under the conditions used in the NMR studies. All NMR data were collected at 25°C on Varian INOVA 600 and Bruker AVANCE 800 NMR spectrometers equipped with 5-mm cryoprobes, processed with NMRPipe [120], and visualized using SPARKY [121]. The rotational correlation times for 5%-13C, U-15N-enriched CDK2AP1 (61-115), and CDK2AP1 (61-115)-C105A are calculated from measurements of average $^{15}$N $T_1$ and $T_2$ relaxation times using 1D $^{15}$N-edited relaxation experiments [122, 123]. Complete $^1$H, $^{13}$C, and $^{15}$N resonance assignments for CDK2AP1 (61-115) were determined using conventional [124] triple resonance NMR methods, respectively, and deposited in the BioMagResDB (BMRB accession number 16808). Stereospecific isopropyl methyl assignments for all Val and Leu residues were deduced from characteristic cross-peak fine structures in high resolution 2D $^1$H-$^{13}$C HSQC spectra of 5%-13C,100%-15N CDK2AP1 (61-115) [125]. Resonance assignments were validated using the Assignment Validation Suite (AVS) software package [126]. $^1$H-15N heteronuclear NOEs were measured with gradient sensitivity-enhanced 2D heteronuclear NOE approaches [123, 127]. The dimer interface in CDK2AP1 (61-115) was identified using X-filtered $^{13}$C-NOESY experiments [128], on a 1:1 sample of $^{13}$C,$^{15}$N- enriched and unlabeled (natural abundance) protein. One-bond N-H$^N$ and N-C’ residual dipolar couplings, D$_{NH}$ and D$_{NC'}$, were obtained on 200 μL [$U$-5%-13C,100%-15N]- and [$U$-100%-13C,100%-15N]-[CDK2AP1 (61-115)] aligned in 7% positively charged compressed gel using standard protocols. The RDCs were determined from $^1$J($H^N$-$N^H$) scalar couplings measured from
an interleaved pair of 2D $^1$H-$^{15}$N TROSY and HNCO-TROSY acquisitions on isotropic and aligned samples.

For the truncated CDK2AP1 (61-115) structure determination, initial structure calculations were performed by Cyana 3.0 [129, 130], using peak intensities from 3D simultaneous CN-NOESY ($\tau_m = 120$ ms) [131], dihedral angle constraints computed by TALOS ($\varphi \pm 20^\circ$; $\psi \pm 20^\circ$) [132], and N-H$^N$ and N-C’ RDC constraints. The 20 structures with lowest target function out of 100 in the final cycle were further refined by restrained molecular dynamics in explicit water using CNS 1.2 [133, 134], using the final NOE derived distance constraints, TALOS dihedral angle and RDC constraints. Final refined ensemble of 20 structures was deposited into the Protein Data Bank (PDB ID, 2KW6). Structural statistics and global structure quality factors, including Verify3D [135], ProsaII [136], PROCHECK[137], and MolProbity [138] raw and statistical Z-scores, were computed using the PSVS 1.3 software package [139]. The global goodness-of-fit of the final structure ensembles with the NOESY peak list data were determined using the RPF analysis program [140].

3.2.3 Polycistronic Co-expression of CDK2 and CDK2AP1

To study the interaction between CDK2 and CDK2AP1 we created an artificial polycistronic co-expression system utilizing our pET15_NESG T7-based expression vector. Briefly, CDK2 was PCR amplified with a forward primer allowing in frame fusion with the pET15_NESG N-terminal 6X-His tag (MGHHHHHHHSH). The reverse primer contained a 25 bp sequence (reverse complement) comprised of a T7 gene 10 Translational Enhancer and a canonical Shine-Dalgarno Ribosome Binding Site (5’ -
ATGTATATCTCCTTTTAAAGTTAA – 3’). Full-length CDK2AP1 was PCR amplified using a forward primer containing the 25 bp translational control elements and a reverse primer containing a region of overlap with 3’ end of the pET15_NESG multiple cloning site (MCS), this open reading frame (ORF) does not impart a 6X-His Tag. Each ORF retained their native stop codon. CDK2 and CDK2AP1 were PCR amplified using the primers described above and the correct sized bands visualized and isolated by agarose gel electrophoresis. Following gel extraction, the two PCR products were cloned into pET15_NESG using Infusion-based (Clontech) ligation independent cloning. More specifically, this was achieved using the CDK2 region of overlap with the pET15_NESG 6X-His tag, the 25 bp region of overlap containing the translational control elements added on to the 3’ end of CDK2 and 5’ end of CDK2AP1, and finally the region of overlap with the vector’s 3’ end of the MCS added on to CDK2AP1. The resulting expression vector construct contains open reading frames for both CDK2 and CDK2AP1 under the control of a single T7 promoter and termination sequence. Thus following IPTG induction, T7 RNA polymerase will transcribe a single message containing the two ORFS (CDK2 and CDK2AP1) each with their own translational control sequences, allowing the proteins to be co-expressed from the same transcript. This co-expression strategy (in close proximity) often allows for the isolation of protein complexes that have proven recalcitrant in other co-expression or co-purification systems.

3.2.4 Size Exclusion Chromatography for Binding Studies

The binding between CDK2AP1 (truncated, full-length and phosphorylated full-length) and CDK2 were studied by size exclusion chromatography. 50 µM of CDK2 and
100 µM of CDK2AP1 construct were mixed in buffer conditions of 50 mM MOPS, 100 mM NaCl, 10 mM MgCl2 at pH 7.4, and incubated for minimum of 1 hour. After incubation the 50 µl of the mixture is injected onto a Superdex 75 10 mm/30 cm gel-filtration column (GE Healthcare) with a 100-µL loop at 0.5 mL/min using an AKTA Explorer system (GE Healthcare) at 4 °C. The eluting protein was detected by UV absorbance at 280 nm. The absorbance profile of the mixture is compared with those of the injections of CDK2 and CDK2AP1 constructs alone.

3.2.5 IKKE Phosphorylation of full-length CDK2AP1

For phosphorylation studies IκB kinase epsilon (IKKε) is bought from Invitrogen (Catalog number: PV4875). The reaction solution is prepared as 0.1 mM of full-length CDK2AP1 in 50 mM Tris (pH 7.5), 200 mM NaCl, 10 mM CaCl2, 5mM β-glycerolphosphate, 2.5 mM ATP. For the reaction 4 µl of the IKKε is injected and it was incubated at 17°C for 48 hrs. Extent of phosphorylation was confirmed by LC/MS/MS experiments.

3.3 Results

3.3.1 HDX-MS Analysis

Residues comprising unstructured or intrinsically-disordered regions in a protein can be readily identified by hydrogen/deuterium exchange mass spectrometry (HDX-MS), since the exchange rates of these amide hydrogens are significantly higher than those in structured regions due to surface accessibility and lack of hydrogen bonding [86]. Studying the dynamics with HDX-MS provides a low cost and high-throughput way
of quantifying amide proton exchange rates in polypeptide segments of proteins [21]. Moreover, construct optimization involving the elimination of unstructured elements identified by HDX-MS has been shown to yield significant improvement in the success of both crystallization experiments for X-ray crystallography studies and protein structure determination by solution NMR methods [21, 22, 88].

Full-length CDK2AP1, was analyzed by HDX-MS. The backbone amide groups in the entire N-terminal 60 residues of the protein exhibit rapid HDX rates characteristic of intrinsically-disordered polypeptide segments (Figure 3.3). Based on this information a truncated construct corresponding to residues 61-115 was designed for NMR structural studies. The overlapping highly dispersed peaks from $^1$H-$^{15}$N HSQC spectra for full-length and truncated CDK2AP1 (Figure 3.4) verifies that the removal of the disordered region does not disrupt the overall structure of ordered regions of the protein.

Figure 3.3 HDX-MS data for CDK2AP1 (NESG ID: HR3057). The secondary structures are shown based on the NMR structure described here.
3.3.2 **CDK2-AP1(61-115) is a dimer in solution**

Functional analysis of CDK2AP1 on contact inhibited diploid cells suggests that it can occur in the cell in both in monomeric and dimeric forms [141]. However, in contact inhibition phase, it was observed that the dimer content in the cell increases with concomitant decrease in monomer content. In the same study, mutational analysis indicated that replacing a single cysteine residue at position 105 by alanine residue...
disrupts dimerization and CDK2 inhibition. This study also suggested that the dimer is the active form of CDK2AP1, and is stabilized by formation of an inter-chain disulfide between C105 residues of each chain of the dimer [141].

We have analyzed the oligomerization state of CDK2AP1(61-115) using analytical gel filtration with static light scattering and $^{15}$N NMR relaxation experiments. The molecular weight measured by static light scattering experiments corresponds to the dimer molecular weight (Figure 3.5a). The rotational correlation time ($\tau_c$), corresponding to the time required for a molecule to rotate 1 radian, is directly proportional to its molecule size. The $\tau_c$ vs molecular weight plot of known monomeric proteins demonstrates that $\tau_c$ is directly proportional to the molecular mass of the globular proteins smaller than 25 kDa (Figure 3.5b). We observe that $\tau_c$ for CDK2AP1(61-115) highly deviates from the general trend and it corresponds to molecular weight of dimeric form of CDK2AP1 (61-115) (Figure 3.5b). Thus, both techniques firmly establish that CDK2AP1 (61-115) is a dimer in solution under the conditions used in this study. Based on these results and X-filtered NOESY data, the solution NMR structure of CDK2AP1 (61-115) was solved as a dimer (see below).
Figure 3.5 (a) Static light scattering chromatogram (blue) for CDK2AP1 (61-115) in the presence of DTT. The sample was injected onto an analytical gel filtration column (Protein KW-802.5, Shodex, Japan; flow rate, 0.5 ml/min, 4 °C) with the effluent monitored by refractive index (black trace; Optilab rEX) and 90° static light scattering (blue) detectors. The magenta data points reflect the estimated molecular weight of the molecule, and the dashed line indicates theoretical molecular weight for the dimer. (b) Rotational correlation time vs MW plotted for known monomeric proteins (blue), CDK2AP1 (61-115) (red) and CDK2AP1 (61-115)-C105A (green) constructs. CDK2AP1 samples were in 20 mM MES buffer at pH 6.5, containing 0.02% NaN₃, 10 mM DTT, 5 mM CaCl₂, 200 mM NaCl, 1x Protease Inhibitors, 10% D₂O, and 50 µM DSS, at 25 °C.
3.3.3  **Solution Structure of CDK2AP1 (61-115)**

Structural statistics for the dimeric solution NMR structure of CDK2AP1(61-115) are summarized in Table 3.1. The structure was determined using 1827 NOE-based distance constraints contacts, 72 RDC data and 182 dihedral constraints obtained based on the backbone chemical shift values. These provided some 18.3 constraints per residue, and 3.6 constraints per residue for long range interactions. The average backbone root mean square deviation of ordered residues in the final ensemble is calculated as 0.6 Å. Backbone dihedral angle analysis indicates that 99.3% of the residues fall in the most favored regions of the Ramachandran plot. Overall, the structure quality scores, including RPF scores comparing the structure with the unassigned NOESY peak list (summarized in Table 3.1) indicate a high quality solution NMR structure for the dimeric C-terminal region of CDK2AP1.
Table 3.1 Summary of NMR and structural statistics for human CDK2AP1(61-115)

<table>
<thead>
<tr>
<th>Completeness of resonance assignments</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Backbone (%)</td>
<td>97.45</td>
</tr>
<tr>
<td>Side chain (%)</td>
<td>84.43</td>
</tr>
<tr>
<td>Aromatic (%)</td>
<td>100</td>
</tr>
<tr>
<td>Stereospecific methyl (%)</td>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Conformationally-restricting constraints</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance constraints</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1827</td>
</tr>
<tr>
<td>intra-residue ($i = j$)</td>
<td>560</td>
</tr>
<tr>
<td>sequential ($</td>
<td>i - j</td>
</tr>
<tr>
<td>medium range ($1 &lt;</td>
<td>i - j</td>
</tr>
<tr>
<td>long range ($</td>
<td>i - j</td>
</tr>
<tr>
<td>distance constraints per residue</td>
<td>16.6</td>
</tr>
<tr>
<td>Dihedral angle constraints</td>
<td></td>
</tr>
<tr>
<td>Number of constraints per residue</td>
<td>182</td>
</tr>
<tr>
<td>Number of long range constraints per residue</td>
<td>3.6</td>
</tr>
</tbody>
</table>

| Residual constraint violations          |  |
| Average number of distance violations per structure |  |
| 0.1 – 0.2 Å                           | 4.05 |
| 0.2 – 0.5 Å                           | 0.4  |
| > 0.5 Å                               | 0    |
| average RMS distance violation / constraint (Å) | 0.01 Å |
| maximum distance violation (Å)         | 0.30 Å |
| Average number of dihedral angle violations per structure |  |
| 1 – 10°                               | 1.85 |
| > 10°                                 | 0    |
| average RMS dihedral angle violation / constraint (degree) | 0.22° |
| maximum dihedral angle violation (degree) | 4.80° |

| RMSD from average coordinates (Å)       |  |
| backbone atoms                         | 0.6 Å |
| heavy atoms                            | 1.2 Å |

| Ramachandran statistics for ordered residues (Richardson lab Molprobity) |  |
| most favored regions (%)              | 99.3% |
| additional allowed regions (%)        | 0.7%  |
| disallowed regions (%)                | 0.0%  |

| Global quality scores^c                 | Raw / Z-score  |
| Verify3D                                | 0.41 / -0.80  |
| ProsaII                                 | 1.07 / 1.74   |
| Procheck(phen-psi)                      | 0.54 / 2.44   |
| Procheck(all)                           | 0.42 / 2.48   |
| MolProbity Clash                       | 13.76 / -0.84 |
Table 3.1 continued

<table>
<thead>
<tr>
<th>RPF Scores</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>0.961</td>
</tr>
<tr>
<td>Precision</td>
<td>0.868</td>
</tr>
<tr>
<td>F-measure</td>
<td>0.913</td>
</tr>
<tr>
<td>DP-score</td>
<td>0.738</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RDC Statistics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of D_{NH} constraints</td>
<td>36</td>
</tr>
<tr>
<td>R</td>
<td>0.913 ± 0.011</td>
</tr>
<tr>
<td>Q_{rms}</td>
<td>0.133 ± 0.009</td>
</tr>
<tr>
<td>Number of D_{NC} constraints</td>
<td>36</td>
</tr>
<tr>
<td>R</td>
<td>0.988 ± 0.001</td>
</tr>
<tr>
<td>Q_{rms}</td>
<td>0.115 ± 0.006</td>
</tr>
</tbody>
</table>

^aStructural statistics were computed for the ensemble of 20 deposited structures
^bOrdered residue ranges \([S(\phi) + S(\psi) > 1.8]\): residues 62-112
^cCalculated based on PSVS v1.4 program

Each protomer of CDK2-AP1(61-115) is composed of two helices, including residues 61-82 (α1) and 84-112 (α2), connected by a type II β-turn to form a hairpin helix motif. These hairpin helices dimerize to form an anti-parallel four helix bundle (Figure 3.6) that has twofold rotational symmetry about an axis nearly parallel to the helix axes and vertical in Figure 3.6. The four-helical bundle structure is predominantly stabilized by the hydrophobic interactions involving numerous hydrophobic residues at the interface. While this fold is relatively rare, similar dimer protein folds have been reported for a microtubule binding protein (PDBID: 1WU9) and a phycobilisome degradation protein (PDBID: 1OJH), which are significantly different in sequence, and do not appear to be homologous proteins.

The α1-helix exhibits a bend around residue K75 near the β-turn, deviating from an ideal linear helical structure. This bend is supported by intra-chain interactions of side-chains of residues A87 with T80 and P79 and T80 with L91; and inter-chain interactions of I77 with I95. The electrostatic potential map (Figure 3.6d) indicates that the surface on inter-chain α1-α2 interface has a net positive charge near the loop region, while, to a
lesser extent, a net negative charge is observed near N- and C-terminal regions of the chains. This electrostatic distribution on overall protein provides a strong dipolar character along the axis of the protein, which may be important for its interactions with its binding partners.
Figure 3.6. Solution structure of residues 61-115 of CDK2AP1. Lowest energy ensemble of 20 structures, lowest energy model (b,c), and the electrostatic potential map (d) are shown. In a-c the structures are colored based on the secondary structures, red for α-helix, green for loop regions, in c cysteine residues are shown in stick representation, with carbon atoms shown in green. The electrostatic potential map is calculated by APBS add-on in pymol. The N- and C-terminal regions are labeled in (a) and (c); the orientation of the models in (a), (b) and (d) are the same.
Even though formation of a disulfide bridge between C105 thiol moieties of each chain does not violate the structural constraints used in calculations, the Cβ chemical shift value of C105 (26.198 ppm), which is diagnostic of the oxidation state of the associated thiol group [142, 143] indicate that, under the sample conditions used in this study (with 10 mM DTT as reducing agent), the two cysteines at position 105 are predominantly in the reduced state [143]. Disulfide mapping using MALDI-TOF analysis confirmed that 10 mM DTT is sufficient to fully reduce the Cys residues of to CDK2AP1(61-115) cysteine (Figure 3.7). Moreover, both 1H-15N HSQC (Figure 3.8a) and 1H-13C HSQC spectra (Figure 3.8) of CDK2AP1(61-115) in solution both with and without DTT are almost identical. This implies that, even in the absence of reducing agent (i.e., no DTT), the reduced form of C105 is the dominant state, and the protein forms a dimer in both conditions. Hence, our structure and data do not support previously published results concerning the requirement for disulfide bonding of C105 residues in dimer formation [141].
Figure 3.7 Disulfide mapping by MALDI-TOF of sample without (a) and with (b) DTT. The trypsin digestion product peptides with and without disulfide bond formation are highlighted. No disulfide formation is observed when DTT is added.
Figure 3.8 Overlay of 600 MHz (a) $^1$H-$^{15}$N HSQC and (b) $^1$H-$^{13}$C HSQC spectra of CDK2AP1(65-115) with (red) and without (green) 10 mM DTT in solution, at 25 °C. The samples are in 20 mM MES buffer at pH 6.5, containing 0.02% NaN$_3$, 5 mM CaCl$_2$, 200 mM NaCL, 1x Protease Inhibitors, 10% D$_2$O, and 50 µM DSS.
3.3.4 **C105A mutation does not disrupt DOC-1(61-115) dimerization**

To verify our findings the C105A of CDK2AP1(61-115) mutant was also studied by NMR. Kim et al. have reported that the C105A mutation abolishes dimer formation and the activity of the protein both *in vivo* and *in vitro* [141]. However, comparison of the \( ^{1}\text{H}-^{15}\text{N} \) HSQC spectra of wild-type and the C105A mutant of CDK2AP1(61-115) (Figure 3.9), suggests otherwise. The fact that the chemical shift perturbations are located mostly in the vicinity of the mutation site indicates that the mutation did not cause significant changes in the 3D structure. The \( \tau_{c} \) measurements also supports this finding (Figure 3.5b). This is expected from the results presented above, since the dimer structure is stabilized mainly by hydrophobic interactions. These data further demonstrate that disulfide bonding of C105 is not required for dimer formation.
Figure 3.9 Overlay of 600 MHz $^1$H-$^{15}$N HSQC spectra of wild-type (magenta) and C105A mutant (blue) CDK2AP1(61-115) in 20 mM MES buffer at pH 6.5, containing 0.02% NaN$_3$, 10 mM DTT, 5 mM CaCl$_2$, 200 mM NaCL, 1x Protease Inhibitors, 10% D$_2$O, and 50 µM DSS, at 25 °C. The residue assignments for the wild-type protein are as labeled.

3.3.5 CDK2AP1 is phosphorylated by IKKε

IkB kinase (IKK) complex is a part of the canonical cellular mechanism which regulates propagation of cellular response to inflammation, through inactivation of inhibitors of nuclear factor kappa-B (NF-κB) thus activating NF-κB pathway [144]. A less studied IKK family protein, IKK epsilon (IKKε), which is a Ser/Thr kinase, is shown to regulate the interferon response mechanism, as well the canonical NF-κB pathway [145, 146]. IKKε was also identified as an oncogene, which was found to be over-expressed in 30% of breast cancer cell lines [147-149]. The several binding targets of
IKKε in interferon signaling pathways were identified but no binding partners responsible for cell transformation were identified.

Recently consensus phosphorylation peptide motif for IKKε was described through peptide library scanning studies. A proteome-based bioinformatics search for the identified motif revealed a set of possible phosphorylation targets for IKKε. On a search against Swiss-prot data base the site S46 in CDK2AP1 scored in the top 0.05% sites searched, predicting it as a candidate IKKε target [119].

Here, by in-vitro phosphorylation studies on full-length CDK2AP1 using GST-tag IKKε (purchased from Invitrogen PV4875) we achieved complete phosphorylation at S46 verified by LC/MS-MS analysis. In Figure 3.10 the LC/MS/MS chromatograms for the peptide including the putative phosphorylation site, S46, from samples before and after phosphorylation reaction is shown for both the phosphorylated (m/z ~ 934.8, retention time ~ 36.7 min) and non-phosphorylated (m/z ~ 908.1, retention time ~ 37.7 min). No signal is observed at retention time 37.7 and m/z 934.8 for the control sample, suggesting that the phosphorylation of the peptide only occurs after IKKε reaction. The chromatogram for the non-phosphorylated peptide from reacted sample indicates that the reaction almost reached completion, with very small signal coming from the non-phosphorylated sample. Figure 3.11 shows the MS/MS spectrum for the phosphorylated peptide, showing that the phosphorylation site is S46, but no other side within this peptide.
Figure 3.10 The LC/MS chromatograms of peptides after trypsin digestion of samples before and after phosphorylation reaction. Chromatograms of non-phosphorylated peptide (a and b) and phosphorylated peptide (c and d) are shown. The elution times of non-phosphorylated and phosphorylated peptides are 36.7 and 37.7 mins, respectively.

Figure 3.11 MS/MS spectrum of QLLSDYGPPSpLGYTQGTGNSQVPQSK peptide. S46 is identified as the only phosphorylated site after IKKε phosphorylation reaction.
NMR experiments indicate that no significant structural changes are observed on the 3D structure of the protein due to phosphorylation (Figure 3.12). For the cases where the structures of other candidate IKKε targets are known, putative phosphorylation sites were located at both structured and unstructured regions. For CDK2AP1 the phosphorylation site is located on the disordered region of the protein, as determined by HDX-MS analysis. This suggests that structural flexibility of the substrate may be a structural requirement for kinase activity of IKKε.

Figure 3.12 Overlay of 600 MHz $^1$H-$^15$N HSQC spectra of wild-type (red) and phosphorylated (green) full-length CDK2AP1 in 20 mM MES buffer at pH 6.5, containing 0.02% NaN$_3$, 10 mM DTT, 5 mM CaCl$_2$, 200 mM NaCl, 1x Protease Inhibitors, 10% D$_2$O, and 50 µM DSS, at 25 °C.
3.3.6 CDK2:CDK2AP1 interaction

We showed that the first 60 residues of CDK2AP1 were highly flexible and only the C-terminal 55 residues form a highly ordered tertiary structure and we determined the solution structure of this region. Based on mutation studies it was reported that residues 109-111 were important in CDK2 interaction [115]. The suggested interaction site is included in our CDK2AP1 (61-115) construct. Based on this information we expected to observe an interaction between CDK2 and CDK2AP1 (65-115).

We studied the interaction between CDK2 and CDK2AP1 (65-115) by size exclusion chromatography. In Figure 3.13 size exclusion chromatograms of CDK2, CDK2AP1 (65-115) and stoichiometric mixture of CDK2 and CDK2AP1 (65-115) are shown. Under the same conditions CDK2 elutes at 11.5 ml while CDK2AP1 (65-115) elutes at 13 ml, due to the difference in their molecular weight. If there is an interaction between CDK2 and CDK2AP1 (65-115), due to the increased molecular weight of the complex, the elution volume is expected to be lower than what was observed when pure CDK2 or CDK2AP1 (61-115) was injected to the column. However, when the mixture is injected, two separate peaks are observed for CDK2 and CDK2AP1 (65-115) at the same elution volumes observed for individual injection of each sample; and no new peak is observed corresponding to complex formation. The separation of two proteins is also confirmed by the SDS-gel ran on the elution fractions (data not shown).
Figure 3.13 Size exclusion chromatography for CDK2 (red), CDK2AP1(65-115) (blue) and mixture (green), with injection of 50 µL of protein sample to Superdex 75 10 mm/30 cm column with a flow rate 0.5 mL/min at 4 °C. The expected elution volume (10.2 mL) for the complex is shown with an arrow.

This observation suggests that the disordered N-terminal tail is required for CDK2 and CDK2AP1 interaction. To test this hypothesis we did a similar study for CDK2 and full-length CDK2AP1 (Figure 3.14). In this case CDK2 and CDK2AP1 elution volumes are similar, 11.5 and 12.5 ml, respectively. When injected in stoichiometric amounts, we still do not observe elution of the complex at an earlier elution volume. This again suggests that there is no interaction detected between CDK2 and CDK2AP1, at the affinity level that can be captured by gel-shift assays. However, due very low absorbance observed for full-length CDK2AP1 in this experiment because of its instability in solution we cannot derive definite conclusions on binding solely based on this experiment.
Figure 3.14 Size exclusion chromatography for CDK2 (red), CDK2AP1 (blue) and mixture (green), with injection of 50 µL of protein sample to Superdex 75 10 mm/30 cm column with a flow rate 0.5 mL/min at 4 °C. The expected elution volume (9.75 mL) for the complex is shown with an arrow.

To test this hypothesis further we prepared a co-expression system where both proteins, CDK2 and full-length CDK2AP1, are expressed in the same vector in *E.coli*; in which only CDK2 had Ni-NTA affinity tag. The binding of the two proteins was studied by co-purification with a Ni-NTA affinity column, where the soluble fraction, after sonication and centrifugation of the cells, is loaded to the affinity column and eluted by increasing concentrations of imidazole. Each fraction was collected and analyzed by SDS-Gel (Figure 3.15). In the final gel we observe that CDK2AP1 starts eluting from the column at no or low imidazole concentrations. CDK2 starts to elute at 100 mM, after all CDK2AP1 is depleted in the column by 40mM imidazole wash. The fact that these proteins elute separately from the column indicates that there is no significant interaction between CDK2 and CDK2AP1 to keep CDK2AP1 on the column bound to CDK2.
Figure 3.15 SDS-GEL of the fractions from co-purification assay for CDK2: CDK2AP1 co-expression system. The summary of the co-expression setup is given at the top, the bands corresponding to CDK2 and CDK2AP1 are marked on the right. The flow through (FT) and the imidazole concentrations for each fraction are marked on each column.

Under the conditions we studied the CDK2 and CDK2AP1 proteins we did not detect any interaction between these proteins. These proteins are shown to be important in cell cycle regulation, and post-translational modification is highly utilized in the cell for regulation of protein functions. In the same paper reporting CDK2 and CDK2AP1 interaction, unphosphorylated and monomeric form of CDK2 was reported to be interacting with CDK2AP1 [115]. We have described above that new potential phosphorylation site was identified for CDK2AP1 and we proved the phosphorylation in vitro. It can be suggested that the phosphorylation of CDK2AP1 may provide a regulation of its function in the cell, i.e. its interaction with CDK2. However, we could not observe
any interaction between CDK2 and CDK2AP1 (S46p) in our gel shift analysis (Figure 3.16).

![Size exclusion chromatography](image)

Figure 3.16 Size exclusion chromatography for CDK2 (red), CDK2AP1 (S46p) (blue) and mixture (green), with injection of 50 µL of protein sample to Superdex 75 10 mm/30 cm column with a flow rate 0.5 mL/min at 4 °C. The expected elution volume (9.75 mL) for the complex is shown with an arrow.

### 3.4 Discussion

HDX-MS studies on full-length CDK2AP1 reveal that the N-terminal domain is highly unstructured, and, hence, solution structural analyses were carried out on a truncated construct, CDK2AP1 (61-115), excluding the residues in the disordered region. Static light scattering and $^{15}$N NMR relaxation data indicate that CDK2AP1 (61-115) forms a stable dimer in solution. The solution NMR structure of CDK2AP1 (61-115) is a symmetric homodimer, featuring an anti-parallel four-helical bundle motif. The structure of CDK2AP1(61-115) is mainly stabilized by hydrophobic interactions at the interface. The only cysteine residue in the protein, C105 located at the interface, is in a reduced
state under the conditions studied. Mutation of this cysteine to alanine does not disrupt the structure or dimer formation. It was reported that the protein was observed in monomer form in log-phase and dimerization of these proteins is observed when cell-growth was inhibited by contact. However, under the conditions studied here the light scattering data and size exclusion chromatography suggest that the full-length protein also forms a dimer.

We have also showed that the CDK2AP1 S46 site gets phosphorylated by IKKε by in-vitro studies, as was suggested by bioinformatics search. With this study, the known interactome of CDK2AP1 has been extended by addition of its interaction with IKKε, which is a very important protein for cell defense and an identified oncogene. Moreover, this interaction may provide a means to regulate CDK2AP1 activity or a more complex interaction pathway, which needs to be elucidated by further studies.

We also studied the CDK2 and CDK2AP1 interaction using gel shift assays and co-expression/co-purification assay. The gel shift assays did not yield any data supporting interaction between CDK2 and CDK2AP1 (61-115), CDK2AP1, and CDK2AP1 (S46p). Further binding assay on E.coli co-expression and Ni-NTA co-purification of CDK2 and CDK2AP1 also could not identify any interaction between these two proteins. The suggested interaction was studied and published by a single lab and it was not supported by any other work from other labs. A recent study revealed that CDK2AP1 was co-purifying with Mi-2/NuRD, which is a chromatin remodeling complex [117]. In the same study it was reported that CDK2AP1 shows an affinity for methylated CpGs DNA; however the authors report that could not detect CDK2 as one of the binding partners of
CDK2AP1. There is an extensive literature about the phenotypes of under-expression or deletion of doc-1 gene in different cancer cell types; however our data suggest that more detailed study is required to understand the mechanisms of CDK2AP1 in the cell.
4. SOLUTION NMR STRUCTURE OF PEPTIDE METHIONINE SULFOXIDE REDUCTASE FROM BACILLUS SUBTILIS

4.1 Introduction

Reactive oxidative species (ROS), consisting of singlet oxygen molecule (O$_2$), superoxide anion, hydroxyl radical, hydroxyl ion, hydrogen peroxide and hypochlorite ion, can cause damage in the cell by reacting with the nucleic acids, proteins, amino acids, fatty acids or co-factors. The accumulated oxidative damage leads to age related degenerative diseases [150-152], cancer [153, 154], and aging due to Free-radical Theory [155, 156]. Although ROS can occur due to environmental exposure, the main sources of oxidative species are the by-products of energy production and other cellular activities. As a result, organisms have developed auxiliary mechanisms to prevent or reverse the damages caused by oxidation.

Methionine is highly susceptible to oxidation by ROS. The oxidation of free or peptide-bound methionine results in a equal mixture of R- and S-epimers of methionine sulfoxides (methionine-S,R-sulfoxides; Met-S,R-SO), due to the achiral sulfur atom [157]. The oxidation of the methionine residues in proteins was shown to impair enzymatic activity and protein function, and to be related to many pathologies [158-160]. As one of the defense mechanisms of the cell against oxidative damage, methionine sulfoxide reductase (Msr) catalyzes the reduction of oxidized methionine (Met-SO) to methionine in all three kingdoms of life. There are two classes of Msr proteins identified, MsrA and MsrB, which specifically function to reduce S- or R- epimers, respectively. Mice lacking msrA gene showed higher sensitivity to oxidative stress with 40% shorter
life span [161]. Also an increased resistance and viability is observed when MsrA and MsrB are overexpressed in oxidative conditions [162-164].

Msr’s are thought to originate from convergent evolutionary processes due to their very similar cellular functions with highly divergent sequence and structure [165]. Extensive biochemical and biophysical studies revealed the catalytic mechanisms of both MsrA [166] and MsrB [167], which are quite similar despite their divergent protein sequences and folds. The proposed reaction mechanism involves the formation of the sulfenic acid at the catalytic cysteine with concomitant release of reduced ligand. The chemically modified enzyme is recycled by formation of intra-chain disulfide bond between the catalytic cysteine (C115 in Bacillus subtilis MsrB) and the recycling cysteine (C65 in B. subtilis MsrB), which is in turn reduced by thioredoxin (Trx) [166, 167].

Three dimensional (3D) structures of MsrB from different organisms have been determined by X-ray crystallography (PDB ID: 3HCG, 3HCH, 3CEZ, 3HCI, 3HCJ, 1L1D, 3E0O and 3CXK). These structures provide structural information on reduced, oxidized and ligand-bound states of the enzyme. The comparison of the crystal structures among different organism indicates that the structure is highly conserved with less than 0.7 Å rmsd among structures from different organisms with an overall fold composed of two β-sheets (seven β-strands) and two short helices at N- and C- terminal regions of the proteins.

Here we present the solution NMR structure of MsrB from B. subtilis (Gene name: msrB, Pfam id: PF01641, SWISS-PROT ID: MSRB_BACSU), as a part of North East Structural Genomics Consortium (NESGC) project. Initial studies indicate that, due
to its molecular weight (16.6 kDa), which is on the larger end of the proteins regularly studied by NMR, and dynamic character of the protein, standard protocols do not provide sufficient data for high resolution structure determination. The solution structure was finally calculated using recently developed techniques developed for solution structure determination for higher molecular weight molecules using sparse constraints from NMR data.

4.2 Methods and Materials

4.2.1 Protein Cloning, Expression and Purification

The msrB gene from *B. subtilis* was cloned into pET21 expression vector with C-terminal hexa-His (LEHHHHHH) purification tag. The expression vector was then transformed into codon enhanced BL21 (DE3) pMGK *E. coli* cells and cultured in MJ9 media. The \([U-5\%-^{13}C, 100\%-^{15}N]-\)labeled samples were prepared using \((^{15}NH_4)_2SO_4\) and \([U-^{13}C]-D\)-glucose as the sole nitrogen and carbon sources in the culture medium [168]. Perdeuterated triple-labeled \((^{13}C,^{15}N,^2H)\), with the methyl groups of Val, Leu, Ile (δ1) selectively protonated, \([U-^2H,^{13}C,^{15}N; ^1H-Ile-\delta1,Leu-\delta,Val-\gamma]-MsrB\) sample was prepared using \((^5NH_4)_2SO_4\) and \([U-^{13}C]-D\)-glucose with addition of \([U-^{13}C_4, 3,3-^2H]\)-α-ketobutyrate (50 mg/L), \([U-^{13}C_5, 3-^2H]\)-α-ketoisovalerate (CIL Inc.) (100 mg/L) in D_2O medium [169]. Initial cell growth is carried out at 37°C and the protein expression was induced by 1 mM isopropyl-β-D-thiogalactopyranoside (IPTG). Expressed proteins were purified by two step purification consisting of HisTrap HP affinity chromatography followed by HiLoad 26/60 Superdex 75 gel filtration chromatography using
AKTAexpress (GE Healthcare). Final samples for NMR experiments contain 0.8-1.0 mM of protein in 20 mM MES buffer at pH 6.5, containing 0.02% NaN₃, 10 mM DTT, 5 mM CaCl₂, 200 mM NaCl, 1x Protease Inhibitors, 10% D₂O, and 50 µM DSS.

4.2.2 NMR data collection and structure calculation

All NMR data were collected at 25°C on Varian INOVA 600 and Bruker AVANCE 800 NMR spectrometers, processed with NMRPipe [120], and visualized using SPARKY [121]. The backbone N, Hᴺ, C’, Cα and Cβ, and ILV residues side chain methyl assignments are used as it was reported earlier [170]. Stereospecific isopropyl methyl assignments for all Val and Leu residues were deduced from characteristic cross-peak fine structures in high resolution 2D ¹H-¹³C HSQC spectra of 5%-¹³C, 100%-¹⁵N MsrB. Distance constraints for structure calculations were acquired from two 3D, ¹³C- and ¹⁵N-edited NOESY spectra and four 3D HSQC-NOESY-HSQC type of spectra (hCCH, hCNH, hNNH, hNCH) [171-173].

Measurement and Analysis of Residual Dipolar Coupling Data - MsrB was first aligned in 13.3 mg/ml Pf1 phage medium (ASLA biotech) [174]. The one bond ¹H-¹⁵N couplings for isotropic and aligned samples of MsrB were measured using an NH J-modulation experiments [175].

Structure Calculations and Structure Quality Assessment - Initial structure calculations were performed by Cyana 3.0 [129, 130], using peak intensities six NOESY spectra, dihedral angle constraints computed by TALOS (ψ ± 20°; φ ± 20°) [132], and NH-HN RDC constraints. The 20 structures with lowest target function out of 100 in the
final cycle were further refined by restrained molecular dynamics in explicit water using CNS 1.2 [133, 134], using the final NOE derived distance constraints, TALOS dihedral angle and RDC constraints. Final refined ensemble of 20 structures was deposited into the Protein Data Bank (PDB ID: 2KZN). Structural statistics and global structure quality factors, including Verify3D [135], ProsaII [136], PROCHECK[137], and MolProbity [138] raw and statistical Z-scores, were computed using the PSVS 1.3 software package [139].

4.3 Results

4.3.1 Samples Used for Data Collection

Protein structure determination by NMR is mostly limited by the size of the proteins. With increasing size of protein molecules the efficiency of magnetization transfer reduces significantly, due to increased relaxation rate of magnetization in the transverse plane [176], which also affects the resolution of the spectra. To reduce the relaxation effects observed in larger proteins different deuteration schemes are introduced, which expanded the size range of proteins for NMR studies [177-180]. Since the gyromagnetic ratio of $^2$H is 6.5 fold less than that of $^1$H, the dipolar effects of $^2$H on relaxation of the surrounding nuclei is much less than that of $^1$H. Hence, perdeuterated or selectively deuterated samples provides substantial increase in the sensitivity and resolution in NMR spectra.

MsrB from B. subtilis is a 141 residue protein with molecular weight of 16.6 kDa is at the upper limit molecular weight for the proteins routinely studied by solution NMR.
methods. Initial studies on structure determination using standard protocols on uniformly $^{13}$C,$^{15}$N labeled protein did not provide good quality data for assignment and structure determination. Thus, 3D structural studies were carried on using triple-labeled with the methyl groups of Val, Leu, Ile (δ1) selectively protonated, $[^{1-2}$H,$^{13}$C,$^{15}$N; $^{1}$H-Ile-δ1,Leu-δ,Val-γ]-MsrB (ILV) sample. Replacing majority of protons in a protein with deuterium provides significant increase in the resolution and sensitivity of the data, as shown in comparison of 2D projections of HNcaCO spectra collected for each sample, in Figure 4.1. The backbone and Ile, Leu and Val side-chain assignments and structure calculation was made possible by use of a perdeuterated triple-labeled ($^{13}$C, $^{15}$N, $^{2}$H) sample, with the methyl groups of ILV labeled as $^{13}$CH$_3$. 
Figure 4.1 Projection of HNcaCO spectra for (a) double-labeled ($^{13}$C,$^{15}$N)-sample and (b) perdeuterated triple-labeled ($^{13}$C,$^{15}$N$_2$H) sample of B. subtilis MsrB in 20 mM MES buffer at pH 6.5, containing 0.02% NaN$_3$, 10 mM DTT, 5 mM CaCl$_2$, 100 mM NaCL, 1x Protease Inhibitors, 10% D$_2$O, and 50 µM DSS.

4.3.2 Chemical Shift Assignments

Although backbone $^{15}$N and $^{13}$C and Ile, Leu and Val $^{13}$C and $^1$H methyl chemical shift values have been determined previously (BMRB ID: 5619) [170], we first verified and corrected these resonance assignments using standard triple-resonance and NOESY
NMR experiments, as described in Material and Methods section. The updated chemical shift list, including the corrections and additional assignments to the existing list, was used for the remainder of the study. These updated resonance assignments for *B. subtilis* MsrB have been deposited in the BioMagResDataBase (BMRB ID: 17008).

In the final list 97.2% of C’, 97.9% of Cα and Cβ, 96.5% of HN and N assignments were complete. Complete side-chain assignments for methyl residues residues of Ile, Val and Leu were determined using 13C-edited NOESY data. Stereospecific assignments of isopropyl methyl groups of Ile and Val residues (17 out of 17) determined based on high resolution 13C-HSQC spectra using a 100% 15N-5% 13C labelled sample.

In the course of determining these resonance assignments, we observed two distinct resonance frequencies for many of the residues in or near the active site of MsrB (Figure 4.2). These multiple resonance frequencies for some atoms of MsrB indicate two different conformational states of MsrB, in or near the enzymatic active site, which are in equilibrium in solution. The peak intensities of the double peaks suggest that these states reflect approximately 65% and 35% of the total population and will be referred to as major and minor states in this text, respectively.
Figure 4.2 The 800 MHz $^{15}$N-HSQC spectrum for MsrB at 25 °C, the residues exhibiting two different resonance frequencies are labeled in blue. Sample in 20 mM MES buffer at pH 6.5, containing 0.02% NaN$_3$, 10 mM DTT, 5 mM CaCl$_2$, 100 mM NaCl, 1x Protease Inhibitors, 10% D$_2$O, and 50 µM DSS.

It was mentioned that the active site of MsrB includes two cysteine residues in proximity to each other, which form a disulfide bond during the enzymatic recycling process following catalysis. We therefore considered the possibility that the multiple states observed for some resonances of the active site may be due to the different oxidation states of these cysteine residues. However, no multiple resonance frequencies are identified for any of the cysteines. Dilution with DTT, a disulfide reducing agent, did not perturb the spectra significantly, or change the ratios of major to minor peaks, suggesting that different oxidation states of these cysteine residues is not the basis of
these multiple states. Rather, the multiple states appear to be due to slow exchange between two or more conformational states of the protein, which differ primarily in structure in or near the active site.

The kind of slow exchange we observe for MsrB often is attributable to cis-trans isomerization around the peptide bond preceding proline residues. We observe two proline residues at the loop near the active site, at the center of the exchanging residues. The greatest chemical shift difference between different states is observed around the residue P109. The chemical shift values of C$^\beta$ of P109 in major and minor conformational states are in agreement with the expected chemical shift values in $trans$ and in $cis$ isomeric states, respectively [181] (Figure 4.3). Hence the observed multiple conformational states observed near the active site are likely to be due to cis-trans isomerization at residue P109.
Figure 4.3 Secondary chemical shift histograms for C\textsubscript{\textbeta} atoms in folded proteins for Proline residues preceded by a trans (blue, left y-axis) and cis (red, right y-axis) peptide bonds. The secondary chemical shift values observed for major and minor conformations are indicated by stars (blue and red, respectively). (Figure adapted from [181])

4.3.3 **The Solution structure of MsrB**

Here we report the solution structure of the major conformational form of *B. subtilis* MsrB determined on a perdeutereated (\textsuperscript{13}C,\textsuperscript{15}N,\textsuperscript{2}H) sample with \textsuperscript{13}C-\textsuperscript{1}H labeled methyl resonances of Ile(\delta), Leu, and Val by ‘minimal constraint’ approach. The structure was studied using sparse constraints from backbone-backbone, backbone-methyl and methyl-methyl distance constraints obtained from NOESY spectra. These data basically provide distance constraints among backbone amides and side-chain methlys from 5 Ile(\delta), 10 Leu and 7 Val residues of MsrB.
The solution structure for this protein has been studied earlier in our lab using sparse constraints under our automated protein fold determination studies [182]. That study focused on rapid structure determination using sparse NMR constraints with medium accuracy, which required minimum user interference. This study indicated that highly automated structure calculation methods using sparse constraints would provide sufficient information for determining the fold of the protein. However, specifically as observed for MsrB, the quality scores for the final structures may be quite poor. Here, we revisited the structure of MsrB, using recent structure calculation tools to improve the quality of the final structure.

In structure calculations, a total of 622 NOE-derived constraints, 85 RDC constraints for N-H\textsuperscript{N} bond vectors, and 235 dihedral angle constraints derived from the backbone chemical shift data using TALOS [132] were used. These correspond to 1.9 long-range restraining constraints per residue, and a total of 6 restricting constraints per residue. Detailed structural statistics and quality scores are given in Table 4.1.
Table 4.1 Structure calculation statistics and quality scores for MsrB

<table>
<thead>
<tr>
<th>Conformationally-restricting constraints</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance constraints</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>622</td>
</tr>
<tr>
<td>intra-residue ((i = j))</td>
<td>46</td>
</tr>
<tr>
<td>sequential ((</td>
<td>i - j</td>
</tr>
<tr>
<td>medium range ((1 &lt;</td>
<td>i - j</td>
</tr>
<tr>
<td>long range ((</td>
<td>i - j</td>
</tr>
<tr>
<td>distance constraints per residue</td>
<td>4.3</td>
</tr>
<tr>
<td>Dihedral angle constraints</td>
<td>235</td>
</tr>
<tr>
<td>Number of constraints per residue</td>
<td>6.0</td>
</tr>
<tr>
<td>Number of long range constraints per residue</td>
<td>1.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Residual constraint violations</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Average number of distance violations per structure</td>
<td></td>
</tr>
<tr>
<td>(0.1 – 0.2 \text{ Å})</td>
<td>30.3</td>
</tr>
<tr>
<td>(0.2 – 0.5 \text{ Å})</td>
<td>9.75</td>
</tr>
<tr>
<td>(&gt; 0.5 \text{ Å})</td>
<td>0.35</td>
</tr>
<tr>
<td>average RMS distance violation / constraint (Å)</td>
<td>0.06 Å</td>
</tr>
<tr>
<td>maximum distance violation (Å)</td>
<td>1.85 Å</td>
</tr>
<tr>
<td>Average number of dihedral angle violations per structure</td>
<td></td>
</tr>
<tr>
<td>(1 – 10^\circ)</td>
<td>28.65</td>
</tr>
<tr>
<td>(&gt; 10^\circ)</td>
<td>1.55</td>
</tr>
<tr>
<td>average RMS dihedral angle violation / constraint (degree)</td>
<td>1.71 (^\circ)</td>
</tr>
<tr>
<td>maximum dihedral angle violation (degree)</td>
<td>20.70 (^\circ)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RMSD from average coordinates (Å)</th>
<th>ordered(^b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>backbone atoms</td>
<td>1.4 Å</td>
</tr>
<tr>
<td>heavy atoms</td>
<td>2.1 Å</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ramachandran statistics for ordered residues (Richardson lab Molprobity)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>most favored regions (%)</td>
<td>94.3%</td>
</tr>
<tr>
<td>additional allowed regions (%)</td>
<td>55.5%</td>
</tr>
<tr>
<td>disallowed regions (%)</td>
<td>0.2%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Global quality scores(^c)</th>
<th>Raw / Z-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verify3D</td>
<td>0.29 / -2.73</td>
</tr>
<tr>
<td>ProsaII</td>
<td>0.42 / -0.95</td>
</tr>
<tr>
<td>Procheck((\phi-\psi))</td>
<td>-0.45 / -1.46</td>
</tr>
<tr>
<td>Procheck(all)</td>
<td>-0.38 / -2.25</td>
</tr>
<tr>
<td>Molprobity clash</td>
<td>18.61 / 1.67</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RDC statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of (D_{\text{NH}}) constraints</td>
</tr>
<tr>
<td>(R)</td>
</tr>
<tr>
<td>(Q_{\text{mm}})</td>
</tr>
</tbody>
</table>

\(^a\)Structural statistics were computed for the ensemble of 20 deposited structures

\(^b\)Ordered residue ranges \([S(\phi) + S(\psi) > 1.8]): residues 4-27,31-33,37-59,64-80,84-95,98-133,135-141

\(^c\)Calculated based on PSVS v1.4 program
Figure 4.4 shows a ribbon representation of the solution structure bundle of MsrB, where two anti-parallel β-sheets formed by β2-β1-β8 and β4-β5-β6-β7-β3, which were flanked by one α-helix formed on N-terminal region and two α-helices on C-terminal region of the protein. The root mean square deviation (rmsd) for all heavy atoms of the protein is calculated as 2.1 Å. The described secondary structural elements are connected by long loop regions, which make up approximately 50% of the protein. When only the ordered residues are considered, the rmsd for the backbone reduces to 1.4 Å. The two β-sheets form the stable core of the protein with 1.0 Å rmsd. This sparse-constraint solution NMR structure of *B. subtilis* MsrB is quite similar to the X-ray crystal structures of homologs that have been published previously [183, 184].

![Solution NMR structure of major conformational state of fully-reduced MsrB](image)

The catalytic cysteine (C115) is located in the β-sheet involving β3 through β7, on strand β7, and the recycling cysteine is located in the β2-β3 loop, between strands β2 and...
β3 and connecting the two β sheets. The distance between the thiol groups of catalytic cysteine (C115), located in strand β7, and the recycling cysteine (C62), located on the loop β2 and β3 is 7.7±2.9 Å (Figure 4.5). As mentioned earlier, in order to recycle the activity of MsrB following Met reduction, formation of a disulfide between C62 and C115 is necessary. Since the average distance between the sulfur atoms engaged in a disulfide bond is 2.05 Å, significantly shorter than the ~ 7.6 ± 3.0 Å distance observed in this fully reduced form of the enzyme, the first step of recycling process requires a conformational rearrangement in order to bring the β2-β3 loop closer to the active β-sheet. The average distance between these groups in crystal structures were 3.8 Å, which still requires conformational rearrangement for recycling.

Figure 4.5 Stereo image of the active site of MsrB, the catalytic and recycling cysteine residues are shown in stick representation and labeled. The residues experiencing slow exchange are colored in red.
4.3.4 MsrB Dynamics

We have also measured backbone $^{15}$N nuclear relaxation data for MsrB, in order to understand the internal dynamics of the protein. $^1$H-$^{15}$N-heteronuclear NOE (hetNOE) and $^{15}$N-relaxation relaxation rates were collected at 600 MHz at 25 °C and plotted in Figure 4.6. Based on the $^{15}$N nuclear relaxation data the rotational tumbling time for MsrB is calculated as 13.8 ns, which is higher than expected (10.8 ns) for this size of protein.

![HetNOE measurements](image)

**Figure 4.6** Backbone $^{15}$N relaxation measurements at 600 MHz, 25°C. $^1$H-$^{15}$N-heteronuclear NOE and $^{15}$N longitudinal ($R_1$) and transverse ($R_2$) relaxation rates are shown. The secondary structures are shown at the top of the figure, location of catalytic and recycling cysteines are shown by black arrows.

HetNOE measurements provide identification of backbone $^{15}$N-$^1$H amide sites which have tumbling rates higher than the rate of tumbling of the overall molecule ($\tau_c \sim 13.8$ ns at 25 °C). These constitute the highly flexible regions of the protein. In hetNOE
measurements for some residues, especially in the loop regions, we observe a decreased hetNOE values. Specifically backbone $^{15}$N-$^1$H amide sites of residues Q30, N31, H36 and E38 between α1 and β1, residue S60 between β2 and β3, residues M85 and I86 between β4 and β5, residues F103 and N110 near the catalytic site have improved flexibility on the < 13.8 ns timescale. Additional to these residues, at the loop regions residue 15 at the N-terminal helix experiences significantly reduced hetNOE value, indicating some flexibility at this location in the helical structure.

The transverse relaxation rates ($R_2$) provide information on dynamics on both the < 13.8 ns and also on the μs-ms timescales. We observe high relaxation rates for residues E72, at β4, and E74 and E76 at the loop connecting β3 and β4, indicating exchange broadening due to intermediate timescale dynamics.

The enzyme plasticity due to thermal energy is the key to catalytic function of enzymes. The link between the thermally driven dynamics at different timescales observed in enzymes with the catalytic function has been established by NMR relaxation studies. It has been observed for several cases that the conformational fluctuations observed in substrate-free enzymes correlate with the fluctuations during catalysis and often constitute the rate limiting step of the overall reaction [3, 4, 185-187].

The NMR relaxation studies and assignments suggest MsrB experiences dynamics in multiple time-scales. Near the active site we observe slow conformational fluctuations in NMR timescale, probably due to cis-trans isomerization at P109. These fluctuations may be related to the reorientation of the loop carrying the recycling cysteine, bringing two cysteines together, to start the recycling process. However, in the
crystal structure a significant difference between the sulfur groups in *cis* and *trans* conformations is not observed. The role of cis-trans isomerization is subject to further investigation.

There are two different forms of the MsrB enzyme, where a subset of MsrBs include a C4 Zn-binding motif, corresponding to loops between β1-β2 and β5-β6, and wider sequence variability for structural stability [188, 189]. The mutational studies incorporating Zn-binding sites on an originally non-binding *N. meningitides* MsrB increased its thermal stability but abolished its recycling by thioredoxin [189]. This indicates that the general flexibility observed in MsrB, especially in the loop regions is essential for continuous activity of the protein.
Figure 4.7 The solution structure of MsrB, the residues with low hetNOE values are colored red, the residues with high $R_2$ rates are colored blue.

4.3.5 X-ray Crystal Structure

As we were working on the solution structure of *B. subtilis* MsrB, an X-ray crystal structure (2.6 Å resolution) was deposited in the Protein Data Bank (PDB ID: 3E0O) [190]. The backbone superimposition of this crystal structure and minimum energy model from our sparse-constraint NMR studies is shown in ribbon representation in Figure 4.8.
Figure 4.8 The superimposition of 2.6 Å X-ray crystal structure (PDB ID: 3E0O) (green) and sparse-constraint NMR (blue) structures for MsrB from *B. subtilis*.

The general fold obtained is very similar (2.3 Å backbone rmsd), while the core formed by two β sheets shows little deviation among two structures with an rmsd of 1.2 Å. The majority of the difference between the crystal structure and the lowest energy model occurs at the loop regions, where the NMR structure was loosely defined due to lack of methyl-methyl contacts; in some cases these loops appear to be flexible in solution, based on $^{15}$N relaxation data described above.

However a significant difference is observed between structures from different methods in the N-terminal helix, consisting of residues 4-23. The X-ray crystal structure exhibits a kink in the helix, around residue L12, which results in two short helices. A
kink is also observed in the corresponding helices of homologues with known structures [183, 184]. In the sparse-constraint NMR structure, we observe a single continuous helix for this region of the protein.

Chemical shifts are very sensitive to the local environment, including both covalent and non-covalent structures, and provide valuable information on secondary structures in proteins [191, 192]. The chemical shift values observed for residues within a secondary structure deviates from the random coil values [193]. For example, the Cα chemical shift values experience on the average 2.6 ppm downfield shift in an α-helical structure, and 1.4 ppm upfield shift in β-sheet structure. The chemical shift index, the variation of the chemical shifts with respect to the random coil values, for the Cα atoms at the N-terminal region is shown in Figure 4.9. The hetNOE data also suggest some flexibility around the location where the kink was observed. While the chemical shift data and hetNOE data are support discontinuity in the helix around residue L12, $^{15}$N-$^1$H RDC data correlates with the NMR (correlation coefficient = 0.978) structure at this region better than X-ray structure (correlation coefficient = 0.871). This discrepancy may be due to ensemble averaging on the measured RDC values, which may not reflect the actual orientation of the bond vectors.
Figure 4.9 Secondary shifts of Cα atoms for N-terminal residues. These data are consistent with a kink in the helix at residues 10 – 11. The helical regions as observed in NMR and X-ray structures are shown above.

4.4 Discussion and Conclusions

The solution NMR structure of MsrB from *B. subtilis* has been determined using a “sparse constraint” approach; using a perdeuterated (\(^{13}\)C\(^{15}\)N\(^2\)H) enriched protein and distance constraints among backbone amide and sidechain methyls from residues Ile, Val, and Leu only. The overall structure forms two antiparallel \(\beta\)-sheets with \(\alpha\)-helices at both terminal regions where the secondary structures are connected by long loop regions, very similar to the structures observed for the homologs from different organisms.

During our studies, and subsequent to completing the NMR structure, an independently-determined X-ray crystal structure was released. The comparison of these structures indicates that while the sparse constraint structure generally agrees with the high-resolution X-ray crystal structure, some differences are observed, mainly around the loop regions, which are within the uncertainty of our NMR structure. The relaxation
experiments also suggest that these loop regions are flexible. Additional to the loop regions, a discrepancy was observed for the helix at the N-terminal region, where a kink at residue L12 breaks the helix into two short helices in the crystal structure. Our analysis indicated that one long helix, as observed in solution structure, is highly correlated with the RDC data, however \(^{13}\)C chemical shift data indicate some distortions from \(\alpha\)-helical structure near residues 12-15. The hetNOE data also indicate some backbone flexibility in the middle of this helix. The comparison of X-ray and NMR structures with the NOESY data used in the calculations indicated that the sparse distance constraints obtained on this perdeuterated sample cannot distinguish between these two conformations at the N-terminal helix. The RDC constraints included in the calculations may be the solution ensemble averaged values, which may not reflect the actual state of the NH-bond vectors. These results suggest that while the sparse constraint approach is valuable for determining overall folds of proteins more work is needed to make this method reliable for defining details of backbone structure and core sidechain packing. For example, inclusion of sophisticated energy force fields, such as the Rosetta force field, may be needed in order to obtain more accurate structures of proteins like MsrB in the 15 – 30 kDa size range which can be addressed by sparse constraints methods.

The multiple resonances observed for a set of residues indicate that the protein exists in two different states which differ in conformation around the enzyme active site. The sparse constraint structure presented here corresponds to the major (trans isomeric state; \(~65\% \) population) conformational state, determined based on the corresponding assignments, NOE contacts and RDC data. Though it was possible to determine backbone
resonance assignments for the minor conformer, efforts to compute structures did not yield a well-defined conformation around the active site due to lack of NOE contacts and RDC data corresponding to the minor state peaks. Again, this kind of problem may be best addressed by combining these sparse constraints, and particularly the chemical shift data obtained for backbone resonances, with more sophisticated energy force fields, such as those provided by programs like Rosetta [194-197].

Based on the reducing agent dilution studies it was concluded that the multiple conformational states do not arise due to different oxidation states of the active cysteine residues. Another explanation would be different isomeric states of the peptide bonds preceding proline residues. At the loop region connecting β2 and β3, where most of the double resonances are observed, two proline residues are located, P107 and P109. The largest chemical shift differences are observed around residue P109 and the Cβ chemical shifts measure at P109 at two different states support cis-trans isomerization at this site, trans state being the major conformational state [181]. In the X-ray crystal structure of this protein in four out of six chains in the asymmetric unit cis-isomeric state is observed at P109. Also in solution NMR structure of a Zn-binding homolog is reported to exhibits cis-isomeric state on a proline corresponding to P109 in structural comparison, supporting our findings.

The relaxation studies also indicated that MsrB experiences structural flexibility at different timescales. Increasing thermal stability by incorporation of Zn-binding sites abolished the recycling by thioredoxin, indicating that the flexibility is essential in continuous activity of the protein. The detailed analysis for understanding exactly what
kind of dynamics is related to the thiredoxin-recycling step is subject to further investigation.
REFERENCES


27. Huang, Y. http://www-nmr.cabm.rutgers.edu/bioinformatics/disorder/.

28. MacCallum. DRIP-PRED *CASP 6 meeting*. Online paper.


Figure A.1 HDX-MS results for NESG target BfR218.
Figure A.2 HDX-MS results for NESG target ER554.
Figure A.3 HDX-MS results for NESG target HR2891
Figure A.4 HDX-MS results for NESG target HR2921
Figure A.5 HDX-MS results for NESG target HR3018
Figure A.6 HDX-MS results for NESG target HR3070
Figure A.7 HDX-MS results for NESG target HR3074
Figure A.8 HDX-MS results for NESG target HR3153
Figure A.9 HDX-MS results for NESG target HR3159
Figure A.10 HDX-MS results for NESG target SmR84
Figure A.11 HDX-MS results for NESG target SpR36
Figure A.12 HDX-MS results for NESG target SvR375