2014

STRUCTURAL STUDIES OF INTERFERON REGULATORY FACTOR 4: A MOLECULAR PERSPECTIVE OF ITS REGULATORY MECHANISM

Soumya Govinda Remesh
Virginia Commonwealth University

Follow this and additional works at: https://scholarscompass.vcu.edu/etd
Part of the Biochemistry, Biophysics, and Structural Biology Commons

© The Author

Downloaded from https://scholarscompass.vcu.edu/etd/3572
STRUCTURAL STUDIES OF INTERFERON REGULATORY FACTOR 4: A MOLECULAR PERSPECTIVE OF ITS REGULATORY MECHANISM

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Physiology & Biophysics at Virginia Commonwealth University.

by

SOUMYA GOVINDA REMESH
Master of Science, Virginia Commonwealth University, August 2010
Bachelor of Pharmacy, University of Mumbai, India, May 2008

Director: Dr. Carlos R. Escalante
Assistant Professor, Department of Physiology & Biophysics

Virginia Commonwealth University
Richmond, Virginia
September 2014
Acknowledgement

The past four years have been an emotionally and physically demanding roller coaster ride. What has kept me sane? Definitely the love and the support of so many people, good food & coffee!!! I have made some amazing friends here. My roommate, Dr. Khushboo Sharma who went through the grind with me. We are so in sync it’s scary. My amazing group of friends (my neighbors in 812, people of 501 & all those with a monthly pass to 501 😊).

My mum, my dad & my sister (also my source of unlimited entertainment) and my extended family of uncles, aunts and cousins have been a constant source of encouragement and have unconditionally loved and supported me. Well its all part of their job description!! My husband who is an extremely supportive and caring partner and marrying him has been one of the best decisions of my life. My mother-in-law who is nothing short of a second mother to me.

Current lab members – Faik who has known me for six years now is like a father to me. Francisco is a very special and dear friend who has seen me through so many tough times. Clayton was in the lab for a short period of time but has definitely become a good friend (also a coffee lover- brownie points!!) Vishaka, whose arrival made my life easier since my boss now had more than one person to annoy. Past lab members – Kanni (my fellow mallu – I was treated with amazing south Indian cuisine quite frequently), Maria (such energy levels and zeal for life!!!!) and Rahul (a good friend – also a very happy person and provider of more fantastic food). Members of the Liu lab (my buddies especially during nights and weekends) – Jiao (I have learnt a lot by just watching her work – so diligent and methodical & an amazing human being), Xueli (very sweet and happy person – also very supportive and a good cook), Hungtao, Mel (good guys and always supportive and encouraging). Past members like Ping and Ray who would drive me back home if it was too late. 😊

Members of my committee – Dr. Peterson who is always enthusiastic about science and has encouraged me to do good work. Qinglian who would have long chats with me about practically anything. Dr Samso, who always provided me with useful scientific inputs and is also a fun person to hang out with outside of work too. Dr. Wang who also encouraged me and had positive feedback. All other people on the second floor of MMRB (Divya, Jeremy, Vidisha and others) who have all been very helpful and would ensure that there was food on the ping-pong table 😊.
And of course, Carlos the boss. It has been a learning experience for both of us with respect to developing the project, communication and patience. He is a very fair and supportive boss and he really treats you like family, which makes for a good and happy working environment. All in all it has been a blast..!!!
# Table of Contents

<table>
<thead>
<tr>
<th>Acknowledgements</th>
<th>ii-iii</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Tables</td>
<td>vii-viii</td>
</tr>
<tr>
<td>List of Figures</td>
<td>viii-xiii</td>
</tr>
<tr>
<td>Abstract</td>
<td>xiv-xiv</td>
</tr>
<tr>
<td>1 Eukaryotic Transcription Machinery</td>
<td>01-26</td>
</tr>
<tr>
<td>General Introduction</td>
<td>01</td>
</tr>
<tr>
<td>Eukaryotic Transcription</td>
<td>02</td>
</tr>
<tr>
<td>Families of transcription factors</td>
<td>04</td>
</tr>
<tr>
<td>Architecture of transcription factor DNA binding sites</td>
<td>08</td>
</tr>
<tr>
<td>Activation of Transcription factors in Eukaryotic cells</td>
<td>09</td>
</tr>
<tr>
<td>Interferon regulatory factor family</td>
<td>12</td>
</tr>
<tr>
<td>Role of IRF4 in development of the host immunity</td>
<td>21</td>
</tr>
<tr>
<td>IRF4 in progression of diseases</td>
<td>23</td>
</tr>
<tr>
<td>2 Crystal structure and small-angle X-ray scattering studies</td>
<td>27-74</td>
</tr>
<tr>
<td>of interferon activation domain of IRF4</td>
<td>27</td>
</tr>
</tbody>
</table>
Results .................................................................................................................30
Discussion ...........................................................................................................73

3 SAXS solution structure of full length IRF4 and the role of C-terminal auto-inhibitory region and N-terminal 20 residues in the activation process........75-104
   Introduction .......................................................................................................75
   Results .............................................................................................................79
   Discussion .......................................................................................................102

4 Structural studies of the putative linker of IRF4...........................................105-121
   Introduction .....................................................................................................105
   Results ............................................................................................................108
   Future directions .............................................................................................120

5 Conclusion and Future Directions .................................................................121-125
Appendix..........................................................................................................126-146
   A X-ray crystallography ..................................................................................145-134
   B Small Angle X-ray Scattering studies.........................................................134-139
   C Analytical Ultracentrifugation ..................................................................139-142
   D Fluorescence Anisotropy ...........................................................................143-145
E  Tobacco Etch Virus-protease purification ......................................... 145-146

References ................................................................................................................. 165
List of Tables

Chapter 1
Table 1: Salient features of the different members of IRF family of transcription factors. ..........................................................15

Chapter 2
Table 1: Data collection statistics for IRF4IAD1 of data collected at home source. ..........................................................................................................................40
Table 2: Data collection statistics for IRF4IAD1 of data collected at X-4A beamline ..........................................................................................................................43
Table 3: Refinement statistics for IRF4IAD1 of data collected at X-4A beamline ..........................................................................................................................44
Table 4: SAXS parameters for IRF4IAD1 and IRF4IAD2. .........................................................65

Chapter 3
Table 1: Experimental and theoretical sedimentation coefficients for of individual constructs and sphere of equal mass .............................................................................84
Table 2: Dissociation constants for the different constructs .................................................................................................................................88
Table 3: R_g, D_max and Porod exponent of the different constructs ........................................93
Table 4: NSD comparing the averaged Gasbor and Dammin models of the different constructs ..............................................................................................................95
Table 5: Experimental and theoretical sedimentation coefficients of different constructs and calculated for Gasbor models .............................................................................................................96

Chapter 4

Table 1: Data collection statistics of IRF4ΔNC with 17bp DNA.................................111

List of Figures

Chapter 1

Figure 1: Different structural motifs found in DNA binding proteins..........................05

Figure 2: Four basic mechanisms of activation of transcription factors ....................10

Figure 3: General mechanism of activation of IRF3 and IRF5 .................................16

Figure 4: Schematic representation of mechanism of activation of IRF3 and IRF5 ..........................................................18

Figure 5: Schematic representation of full length IRF4 including individual domains..................................................................................................................19

Figure 6: General mechanism of activation of IRF4 .................................................20

Figure 7: Currently accepted mechanism of activation of IRF4.................................21

Figure 8: Importance of IRF4 in development of immune cells................................23
Chapter 2

Figure 1: Schematic representation of IRF4 and its binding partners .................................28
Figure 2: SDS-gel electrophoresis result of purification of IRF4IAD1 ..........................32-33
Figure 3: Images of crystals of IRF4IAD1 .................................................................34
Figure 4: SDS-gel electrophoresis result of purification of IRF4IAD2 .........................36-37
Figure 5: Representative diffraction image of IRF4IAD1 with high mosaicity ...............39
Figure 6: Single representative diffraction image of IRF4IAD1 collected from the edges of single crystal .................................................................42
Figure 7: Ramachandran plot of IRF4IAD1 after refinement. .................................................46
Figure 8: Sedimentation velocity profile of IRF4IAD1 .................................................47
Figure 9: Overview of IRF4IAD1 .................................................................................48
Figure 10: 2-D representation of the crystal structure of IRF4IAD1 ..............................49
Figure 11: Superimposition of subunits A and B of IRF4IAD1 .............................................50
Figure 12: Comparison of crystal structures of MH2 domain of Smad3 and IRF4IAD1 .................................................................52
Figure 13: Superimposition of IRF3IAD and IRF4IAD1 ......................................................53
Figure 14: Sequence alignment of IRF3IAD and IRF4IAD1 ..............................................55
Figure 15: Superposition of IRF5IAD and IRF4IAD1 .......................................................56
Figure 16: Surface representation of IRF4IAD1 ...............................................................58
Figure 17: Surface electrostatics in three different views of IRF4, IRF5 and IRF3 .......................................................... 60-61

Figure 18: Helix bundle of IRF4IAD1 .......................................................... 63

Figure 19: P(r) distribution function of IRF4IAD1 ............................................. 65

Figure 20: SAXS molecular envelope of IRF4IAD1 ............................................. 66-67

Figure 21: Comparison of experimental and theoretical scattering profile of IRF4IAD .......................................................... 68

Figure 22: Two models used as an input for BILBOMD ........................................ 70

Figure 23: Comparison of experimental scattering profile of IRF4IAD1 compared to the theoretical scattering profile generated for Model 1........................................ 71

Figure 24: Two different views of manual fitting of MES model obtained from BILBOMD using Model 1 fit into the SAXS molecular envelope........................................ 72

Figure 25: Comparison of experimental scattering profile of IRF4IAD1 compared to the theoretical scattering profile generated for Model 2........................................ 73

**Chapter 3**

Figure 1: Cartoon representation of full-length IRF4 in open & closed state ..........77

Figure 2: Sequence alignment of the N-terminal 74 residues of IRF4 with IRF3, IRF5 and IRF7 ........................................................................ 78

Figure 3: Schematic representation of the four different constructs of IRF4 ..........79
Figure 4: SDS-page gel electrophoresis results for the purification of IRF4_{21} ......... 81-82
Figure 5: Sedimentation velocity profiles of the different IRF4 constructs ..................... 83
Figure 6: Fluorescence saturation curve for the different constructs................................. 85-87
Figure 7: SAXS scattering profile for different constructs............................................. 89-91
Figure 8: P(r) distribution function for the different constructs overlaid
for comparison ................................................................................................................ 92
Figure 9: Kratky plots for the different constructs overlaid
for comparison ................................................................................................................ 93
Figure 10: SAXS molecular envelope of individual constructs....................................... 97
Figure 11: Flow chart representing the protocol used to generate models
in CORAL ....................................................................................................................... 99
Figure 12: Two distinct models generated using CORAL.............................................. 100
Figure 13: Supcomb based fitting of the two different models generated from CORAL
into the SAXS envelope of IRF4ΔNC............................................................................ 101
Figure 14: Superposition of crystal structure of IRF4 DBD and
its NMR structure ........................................................................................................ 104

Chapter 4

Figure 1: Cluster of crystals of IRF4ΔNC with 17bp DNA
and diffraction pattern ................................................................................................ 109
Figure 2: Single crystals of IRF4ΔNC with 17bp DNA
and diffraction pattern..................................................................................................110

Figure 3: Western blot analysis of the crystals of IRF4ΔNC ........................................112

Figure 4: SDS-gel electrophoresis result of different constructs of IRF4
with varying regions of the linker.................................................................................. 113-115

Figure 5: Western blot analysis of thrombin proteolysis of IRF4ΔN............................116

Figure 6: Purification of the fragments (IRF4 20-201 & IRF4 202-450) on a
phenyl sepharose column .......................................................................................... 117

Figure 7: Scattering profile of IRF4 20-201 ..................................................................118

Figure 8: Scattering profile of IRF4 202-450 .................................................................119

**Conclusion and future directions:**

Figure 1: Schematic representation of the new proposed mechanism
of activation of IRF4 ..................................................................................................119

**Appendix**

Figure 1: Schematic representation of diffraction of a crystal & the
collected diffraction pattern .........................................................................................126

Figure 2: Schematic representation of an asymmetric unit, a unit cell &
their arrangement in a crystal ...................................................................................127

Figure 3: Bravais lattice system ..................................................................................128

Figure 4: Patterson auto-correlation function ................................................................130

Figure 5: Pair distribution function ............................................................................135
Figure 6: Standard Kratky plot for proteins of different shapes ........................................137
Figure 7: P(r) function of proteins of different shapes .......................................................138
Figure 8: Double sector centerpiece and a representative sedimentation profile ..........140
Figure 9: Optical system of Beckman Optima XL-A Analytical ultracentrifuge ..........141
Figure 10: Schematic representation of a fluorescence anisotropy experiment ..........144
Interferon (IFN) regulatory factor family member 4 (IRF4) is a transcription factor that serves specific roles in transcriptional regulation of IFN responsive genes and is indispensable in B- & T-cell differentiation. IRF4 like the other members of the family has two major domains- the N-terminal DNA binding domain (DBD) essential for its recognition and binding to the Interferon Stimulated Response Element DNA sequence and a C-terminal Interferon activation domain (IAD) thought to maintain IRF4 in an auto-inhibited inactive state and is also critical in its activation. A putative unstructured linker
connects the DBD and IAD. Activation in most members of the IRF family requires phosphorylation to induce homo and hetero-dimerization. In contrast, IRF4 functions primarily through ternary complex formation involving different proteins including PU.1 and MyD88. The IRF4IAD has a C-terminal auto-inhibitory region (AIR) that has been proposed to physically impede the DBD from interacting with DNA in the absence of its binding partner.

To understand the activation mechanism in molecular detail we determined the crystal structure of the IAD of IRF4 and also performed small-angle X-ray scattering (SAXS) studies. Our data reveals that the surface electrostatics of IAD and presence of additional loops confers exclusivity to IRF4 in the IRF family. SAXS studies suggest that the AIR is structured and makes interactions with the putative linker. We also performed analytical ultracentrifugation studies, fluorescence anisotropy binding experiments and SAXS studies on full-length IRF4 as well as on constructs where the first 20 residues, exclusive to IRF4 or the AIR were removed. We observe that the first 20 residues are critical in decreasing the binding affinity of full-length IRF4 to DNA. In addition, the putative linker of IRF4 connecting the N- and C-termini appears to be a folded domain and interacts with AIR.

Also, overall full-length IRF4 appears as an elongated molecule and the N- and the C-terminal domains are arranged on either ends of full-length IRF4. Moreover, there are no signs of huge conformational changes in the protein during the activation process. Taken together, based on our data we propose that there is no auto-inhibited state for IRF4.
Furthermore, it is the binding affinity of full-length IRF4 that is increased in the presence of its binding partner most likely through modest conformational changes.
Chapter 1
Eukaryotic Transcription Machinery

1.01 General Introduction:

Genes may range in number from 4000-6000 in single-celled prokaryotic or lower eukaryotic organisms to about 20-25,000 in higher eukaryotes including humans (Pugh, 2001), (Gonzaga-Jauregui et al, 2012). Growth, development & survival of all organisms depends on the accurate temporally and spatially regulated transcription of the information present in DNA into RNA that is in-turn translated to proteins. Regulation of transcription is the primary mode to control tissue specific gene expression as well as stimulus specific gene activity. These processes require numerous regulatory proteins or transcription factors, especially for regulation of eukaryotic genes. Nucleoprotein complexes formed between different transcription factors and specific DNA sequences mediate transcription activation by protein-protein and protein-DNA interactions (Latchman, 2008), (David S. Pederson, 1994).

Not surprisingly, the transcription machinery of eukaryotic organisms is more complex as compared to prokaryotic organisms. In prokaryotes, RNA polymerase along with the promoter selectivity exhibiting σ factors function as most important members of the transcription regulation machinery. Since in prokaryotes all genes are transcribed by a single type of RNA polymerase, distinct sigma factors provide the second level of regulation to the transcription system such that only promoters of specific genes are
upregulated selectively (Clancy, 2008). A two-step equation summarizes the prokaryotic transcription complex formation as follows;

\[ R + P \xrightleftharpoons{K_B} R P_c \xrightarrow{k_f} R P_o \]

where, free RNA polymerase (R) and promoter (P) form a rapid and reversible closed complex \((R P_c)\) with a thermodynamic dissociation constant of \(K_B\) that isomerizes to form the open complex \((R P_o)\) at a rate constant \(k_f\) (Kingston and Green, 1994). The isomerized open complex is the transcriptionally active form.

1.02 Eukaryotic Transcription:

Eukaryotic transcription machinery is more extensive and uses a plethora of transcription factors in different permutations to regulate different genes. It requires the orchestrated binding of several different kinds of transcription factors to different DNA sequences sometimes even several base pairs upstream or downstream of the transcriptional start site. Generally, the basal transcription machinery composed of RNA polymerase II, Transcription Factor IIA (TFIIA), TFIIB, TFIID, TFIIE, TFIIF, TFIIH and TFIIJ are sufficient to ensure low-level or basal transcription of genes \emph{in vitro}. Additional, increased activation of eukaryotic gene expression is mediated through the binding of ‘activators’ like GAL4 or Sp1 to specific promoter sequences just upstream of the transcription start site. Owing to their function, activators are typically composed of a DNA binding domain (DBD) that binds to the specific DNA sequences and an activation domain that is essential for transcriptional activation. Activators are known to interact with basal transcription
factors and affect the equilibrium association constants ($K_a$) of one or multiple steps of the pre-initiation complex assembly formed between the basal transcription factors and the activators. Specifically, they may reduce the dissociation rate of bound transcription factors and ensure that an increased proportion of the DNA template has bound pre-initiation complex during transcription. Furthermore, activators may also induce conformational changes to the assembled pre-initiation complex, ensure complex stability or outcompete repressive chromatin components that negate activated gene transcription. ‘Enhancer proteins’ like HMG1 or IRF3 are also activators that bind to DNA enhancer sequences several base pairs away from the promoter sequence. Enhancers regulate transcription through looping of the DNA and bringing the enhancer complex in close proximity to the pre-initiation complex on the promoter. There is yet a third class of transcription factors that regulate activator-dependent transcription without directly binding to DNA called ‘co-activators’. Co-activators either direct the recruitment of the transcription apparatus by activators or function as chromatin remodeling or modifying enzymes (Eg: Histone acetyltransferase). Together with the basal transcription machinery, the activators, the enhancers and the co-activators constitute the complex transcription apparatus in eukaryotes (Harvey Lodish, 2000), (Naar et al, 2001).

Since their discovery in 1970s, extensive research in the field has enabled understanding of the general principles of site-specific recognition of DNA sequences by transcription factors. It is now well known that the peptide backbone or side chains of transcription factors make hydrogen bonds with specific nucleotide bases, especially purine bases in the major groove or salt-bridges with the phosphodiester backbone. It is also well established
that transcription factors bind as homo- or heteromers to ensure site-specific and high affinity binding and frequently bind cooperatively to the DNA often perturbing the overall shape of the double helix (Eg: ternary complex-IRF4/PU.1/DNA or Reb1/DNA complex). Nevertheless, there is a need for a comprehensive investigation to tease apart the characteristic features of individual transcription factor complexes to gain insight into the regulation mechanism of specific nucleoprotein complexes.

1.03 Families of transcription factors:

Structural studies of transcription factors have enabled the classification of these DNA binding proteins based on the commonly occurring structural motifs and secondary structures that are used for site-specific binding to the DNA double helix, (Pabo and Sauer, 1992).

*Helix-turn-helix (HTH)*:

A common feature among the crystal structures of the first DNA-binding proteins like λ Cro protein (Anderson et al, 1981), (Ohlendorf et al, 1982) DNA-binding domain of λ repressor (Pabo and Lewis, 1982) and *E.coli* CAP proteins (McKay and Steitz, 1981) was the occurrence of a motif composed of a α-helix, a turn and second α-helix (HTH motif). HTH motif has since been found in structures of several transcription factors like the lac and trp repressors (Kaptein et al, 1985), (Schevitz et al, 1985). It is worth noting that the HTH motif cannot fold and function by itself and is a part of a bigger domain. Nevertheless, the role of the HTH motif in DNA recognition is extremely important &
frequently involves side-chains of residues of the HTH making site-specific hydrogen bonds or polar interactions with the major groove of DNA and extensive hydrogen bond as well as electrostatic interactions between the protein & the DNA backbone.

**Figure 1.** Different structural motifs found in DNA binding proteins. **a.** Helix-turn-helix motif (2P81) **b.** Zinc finger motif (1SP1) **c.** Leucine zipper motif (1T2K) **d.** Helix-loop-helix motif (2LFH)
Zinc finger:

Another major structural motif found in DNA binding proteins is the zinc-finger motif. Eg: MYST family of histone acetyltransferases (2L43) and estrogen receptor (1HCQ). These are small protein domains of repeats of 20-30 residues with closely placed cysteines and histidines that coordinate zinc atoms to generate a robust and stable structure. Zinc finger motifs are sub-divided based on the number of zins coordinated per domain as well as number of cysteines and histidines involved in the coordination Eg: C2H2 fold group, the gag knuckle or the treble clef and several others. Zinc fingers are involved in numerous cellular processes like transcription, translation, cell proliferation and apoptosis. Zinc finger domains recognize DNA through an $\alpha$-helix and several zinc fingers need to bind the DNA sequence simultaneously at different positions to ensure tight binding.

Leucine zipper:

A third motif is that of the leucine zipper such as that found in enhanceosomes of ATF-1/c-Jun (Cyclic-AMP-dependent transcription factor / Transcription factor AP-1) complex with IRF3 on interferon beta enhancer (1T2K). Leucine zippers are the DNA binding domains of basic leucine-zipper superfamily. Monomeric zippers are long $\alpha$-helices that bind DNA as homodimers or heterodimers. Each monomer has an N-terminus that is basic & interacts with the major groove of DNA while the C-terminus is amphipathic with every seventh residue being a leucine that forms hydrophobic interactions with the C-terminus of another leucine zipper monomer to form coiled coils. The hydrophobic portion of the
monomer is thus buried and the hydrophilic side is solvent exposed. bZIP transcription factors are important mediators of stress responses, cytokine stimulation and affect several developmental processes.

_Helix-loop-Helix (HLH):_

This is another motif that functions as a dimer and are closely related to the leucine zipper motifs. HLH motifs contain two helices linked by a loop region. One helix is basic and recognizes and binds a hexanucleotide E-box (CACGTG) in the major groove of DNA. Dimerization occurs through a hydrophobic second helix and the loop and these regions also make some minimal interactions with DNA. HLH motif is present in MyoD transcription factors important in muscle differentiation (1MYD) and in CLOCK proteins important in maintaining the circadian rhythm (1R8J).

In addition to the aforementioned transcription factor families that recognize DNA primarily through α-helices, there are some prokaryotic regulatory proteins that use beta sheets in the DNA recognition process. Repressors like MetJ (Methionine repressor-1MJM) and Arc (involved in the switch between lysis & lytic cycle in phages-1PAR) use an anti-parallel β-sheet make major contacts with DNA with few additional contacts rendered by some α-helices.
1.04 Architecture of transcription factor DNA binding sites:

Transcription factor binding sites (TFBS) are usually 10-30 nucleotides in length with only a few nucleotides conserved that are essential for binding by specific residues of the transcription factors. Furthermore, there is a built-in level of lack of specificity such that different transcription factors can bind to a single site (Ziegler-Heitbrock et al, 2003). The sites have also been fine-tuned to ensure timely dissociation of the transcription factors or disassembly of complexes at the end of transcription (Werner, 2009). In addition to sequestration of the transcription factors in the cytoplasm in the absence of any induction signal for gene expression, methylation of DNA at the cognate sites (Douet et al, 2007) (Kim et al, 2003) and more tight packing of DNA following DNA deacetylation (Ling et al, 2007) constitute the major modes to ensure context-dependent binding of transcription factors. Physical vicinity of binding sites and their relevance to functionality of transcription factors has been frequently studied in great detail in several bioinformatics studies. It was shown by Smith et al, that homotypic cluster of enhancer sites with 1-8 copies of the binding site of a single transcription factor can but not always amplify the strength of the enhancer for example in case of transcription factors that function as heterodimeric complexes (Smith et al, 2013). The origin of such homotypic clusters with several weak sites and few strong sites is evolutionarily more favorable most likely without significant change in the functionality of the site (He et al, 2012). Smith et al also showed that heterotypic sites composed of two or three to four different binding sites for distinct transcription factors are stronger than homotypic sites. Multimerization of single type or
different types of transcription factors on such clusters is of significance to ensure potent transcription. Although there is a line of thought suggesting that such tandem repeats of DNA may serve as ‘decoys’ binding to transcription factors in inhibiting gene expression (Lee and Maheshri, 2012).

1.05 Activation of Transcription factors in Eukaryotic cells:

In eukaryotic cells, modulating the activity of pre-existing transcription factors rather than inducing its expression at the time of transcriptional regulation is a rapid & flexible means to ensure efficient control of gene expression. With the exception of very few signaling molecules that can also act as transcription factors like lactoferrin (He and Furmanski, 1995), most signaling molecules indirectly produce an active state of the previously inactive transcription factor. Generally, there are four basic mechanisms by which the transcription factors may be activated (Figure 2.).
Figure 2. Four basic mechanisms of activation of transcription factors

a. Regulation by binding of ligand (L)

b. Regulation by protein-protein interaction.

c. Regulation by protein modification [Eg: Phosphorylation (P)]

d. Regulation by protein degradation or processing (Inactive protein is activated by degradation of an inhibiting protein or some region of the inactive factor to release the active form.)
**Regulation by ligand binding:**

Transcription factors may be activated by direct binding of a signaling molecule. This binding event induces a conformational change in the transcription factor and ensures a rapid response to an activation signal. The identity of the signaling molecule ranges from a metal ion in case of induction of yeast metallothionein gene that responds to copper (Thiele, 1992) to oxygen molecules that ensures activation of target genes by transcription factors like Yap1 only under high oxygen concentration (Wood et al, 2004) and even steroid hormones in the activation of glucocorticoid receptor by

**Regulation by protein-protein interaction:**

Certain transcription factors remain in the cytoplasm in an inactive state bound to another inhibitory protein. One of the most well characterized example of such transcription factors is NFκB that activates transcription in mature B-cells or cells treated with lipopolysaccharides (LPS) or phorbol esters and not in pre-B cells (Yamamoto and Gaynor, 2004), (Hoffmann et al, 2006). The inactive NFκB present in the dimer of NFκB/IκB becomes active after IκB is degraded following an upstream signal.
Regulation by protein degradation and processing:

Degradation of IκB is an important step in the activation of NFκB as mentioned above. There are other instances of regulation of transcription factors by protein degradation and processing for example, the activation of sterol regulatory element binding protein (SREBP) in response to changes in cholesterol levels (Brown and Goldstein, 1997).

Regulation by protein modification:

Post-translational modifications (PTMs) are central to the activity of several transcription factors. Several such PTMs have been documented including phosphorylation of serine/threonine residues, acetylation of lysine residues and addition of ubiquitin. Gene activation is frequently dependent on such PTMs to ensure regulation by active transcription factors. Phosphorylation by specific kinases is a common mode of activation of most members of IRF family of transcription factors. Eg: Phosphorylation of IRF3, IRF5 & IRF7. The mechanism has been discussed in detail below.

1.06 Interferon regulatory factor family (IRF family):

Interferons (IFN) are cytokines that are important members of cell signaling in the host immune system against viral infection. Of the two types of IFNs, type I IFNs (IFNα and IFNβ) are produced by several different types of cells like B cells, T cells, macrophages, fibroblasts and endothelial cells and type II IFNs (IFNγ) are produced by natural killer cells (NK cells) & activated T cells.
Proteins of the IRF family are transcription factors that mediate viral and bacterial induced IFN signaling pathways. IRF family members are important regulators of not only host defense mechanism but also cell growth and apoptosis. Promoters of genes regulated by IRFs have an interferon consensus DNA binding sequence called IFN stimulation responsive elements (ISRE). Members of the IRF family have a 115 amino acid N-terminal DNA binding domain (DBD) with five conserved tryptophan residues, three of which aid in binding of DBD to DNA. Exclusivity in function is bestowed to individual members of IRF family their by cell specific expression and by the presence of a C-terminus transactivation domain that can interact with other members of the IRF family or other transcription factors.

There are nine members in the murine IRF family- IRF-1 through IRF-9. Some members also have different names such as Pip/ICSAT for IRF4, ICSBP for IRF8 and ISGF3γ/p48 for IRF9. Except for IRF-4 and IRF-8, which are expressed specifically in cells of the immune system like dendritic cells and macrophages, the other members of the IRF family are ubiquitously expressed in various different cells in eukaryotes. Salient features of individual members are tabulated below (Table 1). IRF family members are important players in the development and execution of the host defense machinery. IRFs stimulate the expression of IFN responsive genes in addition to also regulating IFN expression itself. IFN acts as a signaling molecule to relay information downstream of the attacking pathogen. In addition, IRFs have been implicated in tumor development (IRF-2) as well as tumor suppression (IRF-1). Post-translational modifications like acetylation (IRF-1 or
IRF-7), phosphorylation (IRF-3 or IRF-7) and ubiquitination (IRF-7) are important in activation of IRFs.
<table>
<thead>
<tr>
<th>IRF</th>
<th>Structural characteristics</th>
<th>Distribution Signal dependent modifications</th>
<th>Role in host defense</th>
<th>Role in growth/ development</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRF1</td>
<td>DBD + regulatory domain</td>
<td>Ubiquitous, inducible by IFN, Acetylation</td>
<td>Stimulates expression of IFN responsive genes &amp; proinflammatory cytokines, Promotes Th1 responses</td>
<td>Tumor suppressor</td>
</tr>
<tr>
<td>IRF2</td>
<td>DBD + regulatory domain</td>
<td>Ubiquitous, transcribed after virus infection, Acetylation</td>
<td>Represses Type I IFN responsive genes by antagonizing IRF1 &amp; IRF9</td>
<td>Promotes cell/tumor growth</td>
</tr>
<tr>
<td>IRF3</td>
<td>DBD + IAD + auto-inhibitory domain</td>
<td>Ubiquitous, phosphorylated by IKKe &amp; TBK1</td>
<td>Activates Type I IFN genes transcription, Stimulates IFN responsive genes</td>
<td>Stimulates apoptosis in macrophages upon bacterial infection</td>
</tr>
<tr>
<td>IRF4</td>
<td>DBD + IAD + auto-inhibitory domain</td>
<td>Expresses in lymphocytes, myeloid cells, dendritic cells</td>
<td>Stimulates B-cell specific genes, Inhibits pro-inflammatory cytokines</td>
<td>Required for T-cell-B-cell and dendritic cell differentiation, Oncogenic in myeloma</td>
</tr>
<tr>
<td>IRF5</td>
<td>DBD + IAD + auto-inhibitory domain</td>
<td>Ubiquitous, phosphorylated by IKKe &amp; TBK1</td>
<td>Induces expression of Type I IFN, proinflammatory cytokines upon viral infection</td>
<td>Suppresses oncogene-induced transformation, Required for DNA damage induced apoptosis</td>
</tr>
<tr>
<td>IRF6</td>
<td>DBD + IAD</td>
<td>Constitutive in skin</td>
<td>Required for keratinocyte differentiation</td>
<td>Required for skin, limb, craniofacial development</td>
</tr>
<tr>
<td>IRF7</td>
<td>DBD + IAD + auto-inhibitory domain</td>
<td>Ubiquitous, phosphorylated by IKKe &amp; TBK1</td>
<td>Activates Type I IFN genes transcription</td>
<td>-</td>
</tr>
<tr>
<td>IRF8</td>
<td>DBD + IAD + auto-inhibitory domain</td>
<td>Expressed in macrophages, B cell, dendritic cells</td>
<td>Stimulates macrophage, B-cell specific genes, Stimulates Type I IFN &amp; proinflammatory cytokines in macrophages &amp; dendritic cells</td>
<td>Required for macrophage, B-cell &amp; dendritic cell development</td>
</tr>
<tr>
<td>IRF9</td>
<td>DBD + IAD</td>
<td>Ubiquitous, inducible by IFNγ</td>
<td>Stimulates IFN responsive genes</td>
<td>Mediates Type I IFN induction of p53</td>
</tr>
</tbody>
</table>
Activation of IRF family of proteins:

Activation of most members of the IRF family requires phosphorylation of specific serine residues in the C-terminus of the IAD involved in auto-inhibition that induces conformational changes & generates a functional dimer (Figure 3). The best-studied member of the IRF family IRF3 is expressed ubiquitously and constitutively in different organs and is localized in the cytoplasm as an inactive monomer (Lin et al., 1999). Virus-induced, IKKe or TBK1 mediated phosphorylation of specific serine-threonine clusters leads to dimerization, nuclear localization followed by activation of IFNβ gene (Lin et al., 1998; Servant et al., 2001; Yoneyama et al., 1998; Sharma et al., 2003; Sankar et al., 2006). Phosphorylation of corresponding residues in IRF5 and IRF7 has also been shown to be important for activation of the protein following induction of expression upon IRFα/β mediated viral infection (Marie et al., 2000; Chen et al., 2008).

**Figure 3.** General mechanism of activation of IRF3 & IRF5. The activation is mediated through C-terminal phosphorylation induced dimerization.
Structural details of the auto-inhibitory region in IRF3 were revealed by the crystal structure of the IAD of IRF3 (Qin et al., 2003). Several phosphorylation clusters have been identified in the C-terminal extension of IRF3 such as Ser385-Ser386 cluster, Ser396-Ser398 cluster or Ser402-Thr-404-Ser405 cluster (Yoneyama et al., 1998). Alanine mutations of these residues produce an IRF3 protein that is incapable of activation. Conversely, phosphomimetic aspartate mutations of Ser396-398 cluster and Ser402-Thr-404-Ser-406 cluster yield a constitutively active mutant. The crystal structure of auto-inhibited IRF3-IAD illustrated a system of intermolecular interactions between the N- and C-terminal helices that buries a hydrophobic pocket stabilizing a four-helix bundle involved in maintaining its inactive state. A co-crystal structure of IRF3 bound to co-activator CREB binding protein (CBP) revealed that the hydrophobic auto-inhibitory elements of IRF3 also provided the binding site for CBP (Qin et al., 2005). Later, the crystal structure of a constitutively active phosphomimetic mutant of IRF5 (S430D) was solved by Chen et al (Chen et al., 2008). The structure showed that the S430D mutation caused the unfolding of some key autoinhibitory elements that exposes the hydrophobic pocket that is now available to generate a functional dimer. The structure revealed that upon phosphorylation the auto-inhibitory region flips out as compared to the auto-inhibited inactive conformation captured for IRF3. Together structural and several biochemical studies have aided in detailing the mechanism of activation of the IRF3, IRF5 and IRF7, which is shown in the schematic below (Figure 4).
Interestingly, IRF4 (also called PU.1 interacting protein (Pip) or Interferon consensus sequence binding protein for activated T-cells (ICSAT) is quite different from other members in multiple ways. IRF4 along with IRF8 are the only two members of the IRF family that express more restrictively in the macrophages and dendritic cells and to some extent in cardiocytes (Jiang et al, 2013) and adipocytes (Eguchi et al, 2011; Fabrizi et al, 2014) as compared to the ubiquitous cell expression profile of the other members.

**Figure 4.** Schematic representation of activation of IRF3 and IRF5. IAD of IRF5 undergoes phosphorylation induced dimerization, followed by binding to CBP. IRF5 binds to its recognition site of DNA as a dimer. (adapted from (Chen et al, 2008)). IRF3 and IRF7 also have phosphorylation sites in their C-terminal extension and are believed to undergo activation in a similar manner.
Murine IRF4 contains 450 residues and has a DNA binding domain (DBD) of ~119 residues and an IRF association domain (IAD) spanning 212 amino acids (Figure 4). The last 30 residues at the C-terminus have been suggested to be involved in maintaining the protein in an auto-inhibited state and conferring low DNA binding ability to the full length protein. We will refer to this region as the auto-inhibitory region (AIR).

![Figure 5. Schematic representation of full length IRF4 including individual domains. Represented in cyan is the DBD, in yellow, the putative linker, in pink the IAD and in violet the AIR as well as the N-terminal 20 residues.](image)

In addition, IRF4 is the only member of the family comprising of a 20 amino acid long N-terminus extension. Surprisingly, IRF4 is known to interact with multiple different binding partners including the leucine-zipper heterodimer BATF-JunB (Li et al, 2012), STAT6, (Gupta et al, 1999) as well as other members of the IRF family with the most well characterized binding partner being PU.1 (an ETS family member) (Pongubala et al, 1993; Eisenbeis et al, 1993; Perkel and Atchison, 1998; Yee et al, 1998). Although, phosphorylation induced homo-dimerization as a mode of activation for IRF4 has been suggested (Biswas et al, 2010b), ternary complex formation between IRF4, its binding
partner and DNA to relieve the auto-inhibited inactive state of IRF4 remains the better-established mechanism of activation for IRF4 (Figure 5).

![Diagram](image)

**Figure 6.** General mechanism of activation of IRF4. The schematic here is not suggestive of hetero-dimerization between IRF4 and its binding partner in the absence of DNA. The sequence of binding of the different macromolecules and stability of hetero-dimers of IRF4 with individual binding partner has not been fully investigated yet.

The well-accepted mechanism of activation of IRF4 involves interaction of IRF4 with a binding partner. This interaction is thought to induce conformational changes in the AIR, which translates into extensive conformational changes in the full-length protein. The full-length protein is thought to pivot around a flexible linker that connects the DBD and IAD to go from a ‘closed’ to an ‘open’ conformation. It has been suggested that the AIR physically interacts with DBD preventing it from binding to the ISRE consensus sequence. Shown in Figure 6 is the schematic representation of the currently accepted mechanism of activation of IRF4 (Brass et al, 1996; 1999).
1.07 Role of IRF4 in development of the host immunity:

Being exclusively expressed in immune cells IRF4 plays key roles in the development of the host defense system (Figure 5). In addition to binding to ISRE IRF consensus sequence, the ability of IRF4 to partner with several different proteins through protein-protein interactions extends the range of genes whose regulation is mediated by IRF4. IRF4 plays a pivotal role in the plasma cells differentiation from pre-B cells after they produce high-affinity antibodies shown using transgenic mice with conditional deletion of Irf4 (Klein et al., 2006). The reach of IRF4 in the development of host immunity is widespread. IRF4 is also important in the differentiation of CD4+ dendritic cells that link the innate and adaptive immune systems by functioning as antigen presenting cells (Suzuki et al., 2004; Tamura et al., 2005). Although IRF4 has a dispensable role in activation of CD4+ T-cells, it has been shown to be important in differentiation of CD4+ T-cells like Th2 and...
Th17 cells (Biswas et al, 2010a). IRF4 was also implicated in the process of cell expansion and differentiation of effector CD8⁺ T-cells, which additionally required a robust T-cell receptor (TCR) signaling (Yao et al, 2013). Further, IRF4 also has indirect roles in mediating the host immune system. Toll-like receptor (TLR) mediated signaling central to the activation of the host immune system involves interaction with MyD88 adaptor molecule (Medzhitov et al, 1998). IRF5 and IRF7 interact with MyD88 to ensure induction of proinflammatory cytokines & type I IFNs (Honda et al, 2004, 2005; Kawai et al, 2004; (Takaoka et al, 2005). IRF4 interacts with MyD88 and negatively regulates the TLR signaling by competing with IRF5 and not with IRF7 (Negishi et al, 2005).
Owing to its diverse interactome, IRF4 is now known to be a major player in the progression of multiple disease states like multiple myeloma, ulcerative colitis and certain other auto-immune diseases like type 1 diabetes, multiple sclerosis and systemic lupus.
erythematous (reviewed in Xu et al, 2012). In case of type 1 diabetes there are no animal models for studying the development of these auto-immune diseases but deficiency and lack of IRF4 in murine models have been shown to improve the prognosis in multiple sclerosis and lupus, respectively.

*Achilles heel of multiple myeloma:*

IRF4 has been called an ‘Oncogenic Biomarker’ (reviewed in Ning, 2013). As a regulator of cell cycle, apoptosis & cell proliferation IRF4 transcriptionally regulates several genes like *Blimp1, Bcl6* and *Myc*. Like the other members of the family such as IRF2 and IRF7, IRF4 also possesses oncogenic and anti-apoptotic properties and IRF4 driven regulatory mechanisms are of significance in hematological malignancies like multiple myeloma (MM), activated B-cell like Diffuse Large B-Cell Lymphoma (ABCDLBCL) and other proliferative disorders. Over expression of IRF4 in these malignancies serves as a diagnostic tool in understanding the progression of the disease. In MM, IRF4 serves as a key component in maintenance and progression of the malignancy.

Multiple myeloma is malignancy of terminally differentiated B-cells or plasma cells. In 2008, Shaffer et al (Shaffer et al, 2008) illustrated the oncogenic role of IRF4 in multiple myeloma. Due to the indisputable role of IRF4 in B-cell differentiation as well as lethality to myeloma cell lines shown with small hairpin RNA based IRF4 inhibition, it has been called the ‘Achilles’ heel by the authors. The authors also identified the target of IRF4 in activated B-cells and myeloma to be MYC. Moreover, an autoregulatory circuit between MYC and IRF4 was described in myeloma cell lines important in the sustenance of the
disease. Furthermore, a 50% knockdown of IRF4 mRNA is sufficient to kill myeloma cells while mice lacking one allele are phenotypically normal. Thus, being a master regulator of multiple myeloma IRF4 presents itself as a viable target for drug design.

**IRF4 in ulcerative colitis**

IRF4 has a well-established role in the development & differentiation of dendritic and T- and B-cells (De Silva et al, 2012; Yao et al, 2013). These cells have direct role in the pathogenesis of autoimmune diseases like Crohn’s disease or ulcerative colitis and hence a correlation of the pathogenesis of these diseases with IRF4 is of significance and is well characterized. Proinflammatory cytokines like IL-6 have long been known to be play important role in these inflammatory bowel diseases (IBD) but the molecular mechanism of their production had remained elusive. It was shown by Mudter et al (Mudter et al, 2008) that in patients with IBD there was significant over-expression of IRF4 especially in lamina propria of CD3+ pro-thymocytes. Trinitrosulfonic acid or oxazolone did not induce severe colitis in IRF4 deficient mice and was directly correlated to the decreased production of IL-6 in the mucosa. Thus, IRF4 was identified as a key regulator of IL-6 production as well as pathogenesis of T-cell dependent ulcerative colitis and a probable therapeutic target in ulcerative colitis.
IRF4 in cardiac hypertrophy:

Notwithstanding the incontrovertible role of IRF4 in immune system, its role in cardiac hypertrophy was recently highlighted by Jiang et al. (Jiang et al., 2013). Cardiac hypertrophy or increase in cardiomyocytes in response to decrease in cardiac output and hemodynamic overload has a complicated etiology involving several signaling pathways of G-protein coupled receptors, mitogen-activated protein as well as other members of IRF family like IRF1 and IRF3 (Rockman et al., 2002; Molkentin and Dorn, 2001; Heineke and Molkentin, 2006). In case of IRF4, a transgenic mouse model was subjected to aortic banding and it was delineated that it was overexpression of IRF4 that increases the pressure overload of the heart. The end result is cardiac dysfunction, fibrosis and cardiac hypertrophy. Acting as a transcription factor IRF4 was shown to bind to its specific response element in the promoter region of cAMP response element binding protein (CREB) resulting in up-regulation of CREB. IRF4 may well serve as a selective target for development of novel therapeutics for cardiac hypertrophy.
Chapter 2

Crystal structure and small-angle X-ray scattering studies of interferon activation domain of IRF4

2.01 Introduction:

The IAD of IRF family of proteins not only contain the auto-inhibitory elements to maintain the proteins in an inactive state but more importantly these domains are involved in interaction of IRF proteins with their respective binding partners. The IAD of IRF4 is significant in this context. IRF4 is known to bind DNA poorly in the absence of its binding partners. IRF4 transcriptionally regulates different genes by forming ternary complexes with multiple binding partners binding to DNA composite sites comprising of an IRF DNA recognition sequence adjacent to sequence recognized by the partner protein (Figure 1).

Several such binding partners have been identified including PU.1, Spi-B, NFAT (nuclear factor of activated T cell) and E47 (reviewed in De Silva et al, 2012). The IAD of IRF4 serves as the primary domain involved in interaction with these binding partners and additionally the DBD also provides some degree of interaction. A region composed of residues 150-340 of IRF4 was determined to be important in interaction of NFAT (Nuclear factor of activated T-cells) with IRF4 (Rengarajan et al, 2002; Gupta et al, 2001).
**Figure 1.** Schematic representation of IRF4 and its binding partners. 

**a.** Schematic representation of the regions of IRF4 known to interact with its various binding partners. 

**b.** Schematic representation of additional binding partners of IRF4.
Furthermore, residues 140-207 and 300-420 of IRF4 were determined to be important in interaction of E47 (E2A immunoglobulin enhancer-binding factors E12/E47) with IRF4. DBD of IRF4 was also shown to be important in forming the ternary complex with E47 for maximal transcriptional synergy (Nagulapalli and Atchison, 1998). There are more examples of binding partners of IRF4 that are being discovered but IAD of IRF4 presents as the domain required for the protein-protein interaction.

Although, there are some reports of putative phosphorylation sites identified in the C-terminus of IAD of IRF4, dimerization as a mode of activation has not been established for IRF4 yet (Biswas et al, 2010b). The crystal structure of the ternary complex of IRF4 DBD with the DBD of PU.1 bound to the composite site of immunoglobulin-λ enhancer is the only atomic level structural information available about this member of IRF family (Escalante et al, 2002). It highlighted the cooperativity between the two proteins through their respective DBDs, which is also important to the activation process. In addition, a putative helical region in the IAD from residues 398-400 was identified by Brass et al to be important in interaction with PU.1 with Lysine399 described as the key residue for the interaction (Brass et al, 1999).

Lack of structural and biophysical characterization of IRF4 IAD leaves a great void in our understanding of this unique member of IRF family of transcription factors. Sequence alignment in the IADs as well as AIR extension of the three proteins, namely IRF3, IRF5 and IRF4 is quite low. Thus, although there may be an overall similarity in the fold of the three proteins, IRF4 presumably has characteristic traits enabling it to function in conjunction with a binding partner. One of our important goals is to delineate structurally
the molecular details unique to IAD of IRF4 to better understand the interaction of IRF4 with its binding partner.

In an attempt to crystallize the IAD of IRF4 we cloned, expressed and purified the construct composed of residues 238-450. This construct proved to be non-ideal for crystallization since it aggregated at a concentration of above 2mg/mL. Successful, crystallization trials generally require a minimal protein concentration of 10mg/mL. The protein was dialyzed in several different buffers including high concentration of reducing agents (upto 10mM beta-mercaptoethanol or 2-3mM TCEP) and amino acids like L-Arginine (upto 400mM) with little success in achieving crystallization hits. Hence, we expressed and purified another construct composed of residues 238-420, which was amenable to crystallization trials. To obtain structural information on the AIR, we utilized solution X-ray scattering studies on IRF4 IAD 238-450

2.02 Results:

2.02.1 Purification and crystallization:

IRF4IAD1:

Mouse IRF4 C-terminus construct 238-420 (IRF4IAD1) was subcloned into pET-15b TEV vector (pet15TEV_NESG (EvNO00338203) from DNASU plasmid repository, the vector has a Tobacco Etch Virus-protease cleavable site following the 6-His site). The gene of interest was PCR amplified using the appropriate primers, restriction digested using two restriction enzymes NdeI and BamH1 to ensure directional cloning, ligation and transformation into XL10-Gold Escherichai coli cells. Positive clones were sequenced and
subsequently transformed into BL21-pLysS* E. coli expression cells. An overnight pre-culture was grown in Luria-broth media with 100µg/mL ampicilin and 25µg/mL chloramphenicol antibiotics at 37°C. Next day, 4L of autoclaved LB media was inoculated with the pre-culture (20mL per liter), cells were grown at 37°C, induced with 0.5mM IPTG at OD ~0.6 and harvested after 4hrs. The cells were flash-frozen, thawed and resuspended in 25mM Tris-base pH 8.0, 500mM NaCl, 2mM TCEP, 5mM Imidazole, 10% glycerol, sonicated and centrifuged at 20000 rpm for 30 mins. Protein was eluted from a Ni-NTA column with 300mM Imidazole and the Histidine (His) tag was removed using the home grown and purified Tobacco Etch Virus (TEV)-protease (details of purification in appendix). The cleaved protein was further purified on Ni-NTA column to remove the His-tag as well as the TEV-protease followed by gel filtration on a Superdex75 column in 25mM Tris-base pH 8.0, 300mM NaCl, 2mM TCEP and 5% glycerol. Purified protein was concentrated to 16mg/mL (0.753mM) and stored at -80°C in 60µL aliquots.
Figure 2. SDS-gel electrophoresis result of purification of IRF4IAD1. 

a. SDS-gel electrophoresis result of purification of IRF4IAD1 after a Ni-NTA column. 

b. SDS-gel electrophoresis result IRF4IAD1 after cutting with TEV-protease
Figure 2. SDS-gel electrophoresis result of purification of IRF4AD1 c. SDS-gel electrophoresis result of purification of IRF4AD1 after second Ni-NTA column. d. SDS-gel electrophoresis result after purification of IRF4AD1 after a gel filtration column.
Initial factorial screens of IRF4IAD1 (at 0.753mM) with Hampton and Emerald crystallization screens yielded thin needle shaped crystals in Emerald Wizard Screen I condition 37 (2.5M NaCl, 0.1M Imidazole pH 8.0) with hanging drops at 20°C. The initial crystals grew as a bunch and could not be separated (Figure 3 a.).

Figure 3. Images of crystals of IRF4IAD1 a. Initial needle shaped clusters grown at 20°C. b. Thin plate crystals obtained after changing the precipitant from NaCl to KCl. c. Single plate crystals obtained after changing the temperature to 4°C and precipitant.
The precipitant salt was changed from NaCl to KCl and the reservoir concentration was reduced to 2M KCl, which resulted in thin plate crystal clusters (Figure 3 b). Further optimization of the condition along with a change in crystallization temperature to 4 °C resulted in single thin plate crystals in surprisingly much lower salt concentration of 1.5-1.7M KCl, 0.1M Imidazole pH 8.0 as compared to the original condition at 20 °C (Figure 3 e).

**IRF4IAD2:**

Mouse IRF4 C-terminus construct 238-450 (IRF4IAD2) was subcloned into pET-15b TEV by a procedure similar to that mentioned above using the appropriate primers. IAD was also grown at 37°C until the OD reached ~0.6, induced with 0.5mM IPTG and harvested after 4hrs. Following Ni-NTA based affinity purification detailed above the His-affinity tag of IAD was removed using thrombin (1Uthrombin/1U protein) (Figure 4 a and b). Thrombin was precipitated with p-aminobenzamidine-agarose beads by continuous shaking for 45 mins. Finally IAD was gel purified on a Superdex-75 column in 25mM Tris-base pH 8.0, 500mM NaCl, 5mM BME, 10% glycerol (Figure 4 c and d). Protein was concentrated to 1mg/mL (0.04mM) and stored in 100µL aliquots since the protein precipitated at a concentration above 2mg/mL.
Figure 4. SDS-gel electrophoresis result of purification of IRF4IAD2. 

a. SDS-gel electrophoresis result of purification of IRF4IAD2 after Ni-NTA column. 

b. SDS-gel electrophoresis result of IRF4IAD2 after cutting with thrombin protease.
Figure 4. SDS-gel electrophoresis result of purification of IRF4AD2 c. SDS-gel electrophoresis result of purification of IRF4AD2 after second Ni-NTA column. d. SDS-gel electrophoresis result of purification of IRF4AD2 after gel filtration column
2.02.2 *Data collection and structure determination*:

To check the diffraction of the crystals we collected diffraction data on one of the thin plate crystals from the clusters that were obtained after initial optimization (Figure 5. b).

Crystals diffracted to 2.4Å at the home source (Molecular Structure Corporation (MSC) X-Stream Cryogenic Crystal Cooler with a Rigaku IV ++ image plate detector, and a Rigaku MicroMax-007 X-Ray source (copper) fitted with MSC Varimax Confocal optics operating at 40 kV and 20 mA). A representative image of an oscillation frame is shown in Figure 5. The mother liquor with 25% glycerol served as the cryo-protectant. Initial indexing indicated that the crystals were in monoclinic space group and 180° data were collected. A total of 380 frames were collected from φ 0 to 190 at 0.50° intervals at a detector distance of 160mm. Data collection statistics for the data collected at home source is provided in Table 1. Scaling and averaging of this data was followed by multiple unsuccessful attempts to obtain a solution, initial phases and a starting model for IRF4 IAD using IRF5 IAD as a search model for molecular replacement in either ccp4 (Winn et al, 2011) or Phenix (Adams et al, 2010). We concluded that the high mosaicity of the collected data indicated by spot overlap (inset Figure 5) might be causing errors in indexing, scaling and averaging. Thus, a single solution was not achieved using molecular replacement. Furthermore, the quality of the crystal used here was not optimal as it was broken off from the cluster of crystals shown in (Figure 3 b).
Figure 5. Representative diffraction image of IRF4IAD1 with high mosaicity. Inset shows closely spaced diffraction spots indicating high mosaicity in the diffraction data.
After optimization of the original condition to obtain single crystals, the crystals were harvested at 4°C into the reservoir-cryo solution (1.5M-1.7M KCl, 0.1M Imidazole and 25% glycerol used as a cryoprotectant) and flash-frozen after about 15-20 seconds of soaking in the cryo-solution. Diffraction data were collected for multiple single thin-plate crystals at beamlines X4A and X25 at National Synchrotron Light Source (NSLS). Initial attempts to collect diffraction data at X4A beamline were unsuccessful since indexing did

<table>
<thead>
<tr>
<th><strong>Data collection</strong></th>
<th><strong>IRF4IAD1</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Space group</td>
<td>C2</td>
</tr>
<tr>
<td>Cell dimensions</td>
<td></td>
</tr>
<tr>
<td>$a, b, c$ (Å)</td>
<td>151.4 84.2 45.6</td>
</tr>
<tr>
<td>$a, b, g$ (°)</td>
<td>90 94.14 90</td>
</tr>
<tr>
<td>Number of reflection</td>
<td></td>
</tr>
<tr>
<td>Total measured</td>
<td>39597</td>
</tr>
<tr>
<td>Unique</td>
<td>17356</td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>2.5</td>
</tr>
<tr>
<td>$R_{merge}$</td>
<td>9.6 (29)</td>
</tr>
<tr>
<td>$I/\sigma I$</td>
<td>6.7 (2.7)</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>87.5 (88)</td>
</tr>
</tbody>
</table>

**Table 1.** Data collection statistics for IRF4IAD1 of data collected at home source. Statistics for the last shell is shown in brackets.
not provide any acceptable spacegroup (Figure 8). We attempted annealing and also soaking the crystals in the reservoir-cryo solution for longer periods of time with little success. We observed that high degree of mosaicity in the crystals was causing problem indexing and further processing. We reasoned that although the crystals appeared single to the naked eye they probably were multiple crystals stacked up on each other. We then used the micro-focus beam available at both the beamlines to collect data sets from the edges of the thin plates crystals where the crystals had a higher probability of being single. Of about 50 crystals that were screened between the beamlines X4A and X25 only four complete data sets with low-mosaicity could be collected between 2.8–2.4Å resolution. A representative image of the diffraction data is shown in Figure 8. But at resolutions above 2.6Å the completeness of the data sets was quite low and hence the data set 2.6Å was only used for refinement. Data collection statistics of data set that was used for structure solution (data set at 2.8Å) are provided in Table 2. The data sets could be readily indexed and gave a spacegroup of P222₁.
Figure 6. Single representative diffraction image of IRF41AD1 collected from the edges of single crystals.
<table>
<thead>
<tr>
<th>Data collection</th>
<th>IRF4IAD1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Space group</td>
<td>P222₁</td>
</tr>
<tr>
<td>Cell dimensions</td>
<td></td>
</tr>
<tr>
<td>$a, b, c$ (Å)</td>
<td>45.5  84.9 149.9</td>
</tr>
<tr>
<td>$a, b, g$ (º)</td>
<td>90  90  90</td>
</tr>
<tr>
<td>Number of reflections</td>
<td></td>
</tr>
<tr>
<td>Total measured</td>
<td>342484</td>
</tr>
<tr>
<td>Unique</td>
<td>15281</td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>2.8</td>
</tr>
<tr>
<td>$R_{merge}$</td>
<td>10 (15.6)</td>
</tr>
<tr>
<td>$I/\sigma I$</td>
<td>8.5 (13.5)</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>83 (78)</td>
</tr>
</tbody>
</table>

**Table 2.** Data collection statistics for IRF4IAD1 of data collected at X-4Abeamline. Statistics for the last shell is shown in brackets.
2.02.3 **Structure determination:**

The crystal structure of the transactivation domain of IRF4IAD1 was determined at 2.8Å using molecular replacement and later refined to 2.6Å. Refinement statistics for IRF4IAD1 is provided in (Table 3.).

<table>
<thead>
<tr>
<th>Refinement</th>
<th>IRF4IAD1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resolution (Å)</td>
<td>2.6</td>
</tr>
<tr>
<td>Number of reflections</td>
<td>17122</td>
</tr>
<tr>
<td>Rwork/ Rfree</td>
<td>22.19 / 27.95</td>
</tr>
<tr>
<td>Non hydrogen atoms</td>
<td>Protein</td>
</tr>
<tr>
<td></td>
<td>2744</td>
</tr>
<tr>
<td>B-factors</td>
<td>Protein (Å2)</td>
</tr>
<tr>
<td></td>
<td>36</td>
</tr>
<tr>
<td>R.m.s deviations</td>
<td>Bond lengths (Å)</td>
</tr>
<tr>
<td></td>
<td>Bond angles (°)</td>
</tr>
<tr>
<td>Ramachandran plot quality</td>
<td>Most favored (%)</td>
</tr>
<tr>
<td></td>
<td>Generally allowed (%)</td>
</tr>
<tr>
<td></td>
<td>Disallowed (%)</td>
</tr>
</tbody>
</table>

*Table 3.* Refinement statistics for IRF4IAD1 of data collected at X-4A beamline.
The structure was solved by molecular replacement using the structure of IRF5 IAD as a search model in the program Phaser in ccp4 ((Adams et al, 2010), (McCoy et al, 2007)), a software suite for automated structure solution. The data was cut off to 2.8Å during the search to generate a starting model with an acceptable R-factor of less than 50. The starting model was then exported to Phenix (Adams et al, 2010) and refinement of the structure was performed in Phenix.refine. Manual model corrections including loop building, changing the side chain and peptide orientations and addition of missing residues and ions was performed after every round of refinement using the improved map and the Ramachandran plot (please see appendix -X-ray crystallography glossary; sub-heading: Refinement for details). Initial 4-5 rounds of refinement were performed using the refinement strategy that included simulated annealing and real space refinement.

For subsequent 10 rounds of refinement we used a strategy including the non-crystallographic symmetry (NCS) in the refinement. In addition, TLS refinement (please see appendix section of X-ray crystallography (sub-heading: Refinement) for an explanation of this refinement strategy) was also used in these rounds of refinement.

A web server was used for generating an input file for TLS motion determination to be used in Phenix.refine (Merrit, 2006). After multiple rounds of refinement and addition of water molecules to the structure a final $R_{factor}$ of 27.95% with the final $R_{free}$ of 22.19% was obtained. A final Ramachandran plot obtained for the structure is shown in Figure 9.
Figure 7. Ramachandran plot of IRF4IAD1 after refinement.
2.02.4 *X-ray structure of IRF4IAD1*:

The X-ray structure of IRF4IAD1 contains two polypeptide subunits per asymmetric unit. The dimeric arrangement of IRF4IAD1 has no physiological significance but is the result of favorable crystal packing. Sedimentation analysis of IRF4IAD1 reveals a monomeric state of the protein with no suggestion of possibility of oligomerization (Figure 8).

![Sedimentation velocity profile of IRF4IAD1](Figure 8). Sedimentation velocity profile of IRF4IAD1. It shows a single peak for the protein with an $S_{w,20}$ of 2.0.
Figure 9. Overview of IRF4AD1. It represents the four alpha helices, ten beta sheets and the associated loops.
**Figure 10.** 2-D representation of the crystal structure of IRF4AD1. The various secondary elements are clearly visible including the loops that connect them.
The subunit A of the X-ray structure of IRF4IAD1 (Figure 9 and 10) is described hereunder. It is composed of four alpha helices (labeled α1-α4) that surround the core of beta sheets (labeled β1-β10). Helices α1, α3 and α4 form a helix bundle that is located towards the upper end of the structure. There are two beta hairpins in the structure are formed by β1-β2 and β5-β6. β3 and β4 are connected by loop Lβ3-β4 of moderate length while another such loop Lβ4-α2 connects beta sheet β4 to helix α2. We also observe two quite long loops Lβ7-β8 and Lβ9-β10 that connect beta sheets β7-β8 and β9-β10, respectively.

**Figure 11.** Superimposition of subunits A and B of IRF4IAD1
Subunits A and B of IRF4IAD1 superimposed with a root mean square deviation (rmsd) between the Cαs of 0.56Å using the program Lsqman (Kleywegt and Jones, 1994) indicating that the two subunits are quite similar (Figure 11) (Lsqman was used to calculate all of the mentioned rmsds). The electron density for loops Lβ7-β8 and Lβ9-β10 is missing in subunit B although two extra residues one on either N- and C-termini are visible in subunit B and not in subunit A. In addition, Lβ7-β8 in subunit B has some secondary structure characteristic shown by the presence of a small beta sheet. The largest difference between the two subunits is in the position of helix α1 (rmsd of 3.72Å) between the Cαs of the first 12 residues of the N-terminus. This is indicative of appreciable degree of flexibility of helix α1. The corresponding helix in IRF3IAD is superimposable between the two subunits. This feature seems to be inherent to IRF4.

It is worth noting that IRF4IAD1 contains the MH2 fold of the Smad protein family shown in Figure 12. Overall, IRF4IAD1 has a contour resembling a claw and although there is similarity between the shape of IRF4IAD1 and MH2 domain of Smad3 (1MJS), MH2 domain of Smad3 lacks the helix bundle seen in IRF4IAD1, is less bulky and has shorter loops. Smad3 also has a Ser-X-Ser motif that gets phosphorylated by its binding partner, an activated type II tissue growth factor receptor β (TGFβ receptor) (Huse et al, 2001), (Wu et al, 2000), (Wu et al, 2001). IRF4IAD1 lacks such phosphorylation sites although it binds to multiple binding partners.
Figure 12. Comparison of crystal structures of MH2 domain of Smad3 and IRF4AD1. a. MH2 domain of Smad3 (1MJS)  b. IRF4AD1
2.04.5 Comparative analysis of IRF4IAD1 with IADs of other members of the IRF family:

IRF4IAD1 and IRF3IAD:

![Superimposition of IRF3IAD (green) and IRF4IAD1 (cyan)](image)

**Figure 13.** Superimposition of IRF3IAD (green) and IRF4IAD1 (cyan)
The rmsd between the C\(\alpha\) atoms of IRF4IAD1 (182 residues) and IAD of IRF3 (193 residues, auto-inhibitory region was removed for the superimposition) is \(~1.35\text{Å}\). There are three regions that show major differences between the two structures (Figure 13). There is a \(5.18\text{Å}\) rmsd between helix \(\alpha1\) of subunit A of IRF4IAD1 with the corresponding 11 C\(\alpha\) atoms of IRF3 (6.2Å rmsd between subunit B of IRF4IAD1 and corresponding 11 C\(\alpha\) atoms IRF3). Loop \(L_{\beta7-\beta8}\) of IRF4IAD1 is more structured in the IAD of IRF3 forming two beta-sheets and is also in a different orientation (rmsd of 12.24Å and 9.88Å between subunit A and subunit B and the corresponding 14 C\(\alpha\) atoms in IRF3, respectively). But the major structural difference of physiological relevance is in loop \(L_{\beta9-\beta10}\) of IRF4IAD1 (rmsd of 9.66Å and 9.74Å between subunit A and subunit B and the corresponding 14 C\(\alpha\) atoms in IRF3, respectively). This loop contains the critical residue lysine399 flanked by residues 398-400 on its either side. This region was proposed to form an \(\alpha\)-helix in previous studies (Brass et al, 1999) but we show here that it is a highly flexible linker that forms a positive patch (see below in electrostatics) for making key electrostatic interactions with phosphorylated residues of partnering proteins like phosphorylated PU.1. The corresponding loop in IAD of IRF3 has some helical character to it, which makes it more rigid and thus has lesser degrees of freedom.

It is worth noting that the 30 residue C-terminal extension in the IAD of IRF4 has no appreciable sequence similarity to the auto-inhibitory region of IRF3 (Figure 14). It is also much shorter and lacks residues that may undergo phosphorylation as compared to the key phosphorylation sites identified for IRF3. The residues that undergo phosphorylation in IAD of IRF3 are labeled in red in Figure 16. Secondary structure of IAD of IRF3 is depicted on
Putative phosphorylation serine residues in IRF4 that have regulate IL-17 and IL-21 production in mice are Serine 446 and Serine 447 (Biswas et al, 2010b).

**Figure 14.** Sequence alignment of IRF3IAD and IRF4IAD1. In red are shown the phosphorylation sites of IRF3IAD. In blue are shown the putative phosphorylations sites in IRF4IAD1. In cyan are shown the hydrophobic residues of IRF3IAD found in helix α5. Corresponding residues are absent for IRF4IAD1 (described below under in subsection 2.04.6)
IRF4IAD1 and IRF5IAD:

The rmsd between the Cα atoms of IRF4IAD1 (182 residues) and IAD of IRF5 (auto-inhibitory region was removed for the superimposition, 199 residues) is 1.13Å (Figure 15).

Figure 15. Superposition of IRF5IAD and IRF4IAD1

The rmsd between the Cα atoms of IRF4IAD1 (182 residues) and IAD of IRF5 (auto-inhibitory region was removed for the superimposition, 199 residues) is 1.13Å (Figure 15).
Similar to IRF3IAD, the IAD of IRF5 also has a more structured loop that corresponds to Loop $L_{\beta_9-\beta_{10}}$ of IRF4IAD1, which is critical for interaction with a phosphorylated partner protein PU.1 (rmsd of 8.21Å and 7.22Å between subunit A and subunit B and the corresponding 12 C$\alpha$ atoms in IRF5, respectively). Furthermore, the helix in the constitutively active IAD of IRF5 is non-superimposable with the corresponding helix $\alpha 1$ of IRF4IAD1. There is a 7.5Å rmsd between helix $\alpha 1$ of subunit A of IRF4IAD1 with the corresponding 10 C$\alpha$ atoms of IRF5. Loop $L_{\beta_7-\beta_8}$ shows an rmsd of only 3.33Å and 2.91Å between the subunit A and subunit B and the corresponding 10 C$\alpha$ atoms of IRF5, which is in contrast to what is observed for IRF3. The loop is buried and is inaccessible in the functional dimer of IRF5. Additionally, there is also negligible sequence similarity between the auto-inhibitory elements of IAD of IRF4 (last 30 residues) and those of IRF5 (last 36 residues) and there are no corresponding phosphorylation sites in IRF4IAD1 as compared to the several sites available for IRF5. Instead there are threonine residues at some of the serine phosphorylation sites for example, Threonine 421 corresponding to Serine 421 and Threonine 435 corresponding to Serine 436.
2.04.6 Surface electrostatics of IRF4IAD1:

A surface representation of IRF4IAD1 is shown in Figure 16.

Figure 16. Surface representation of IRF4IAD1
Although the overall structure of IRF4IAD1 is almost completely superimposable with the IADs of IRF3 and IRF5, the surface residues of IRF4IAD1 are quite different than the other IADs. This is seen in Figure 17 showing the different views of the various IADs. In panel a of Figure 17, three different views of the surface electrostatics of IRF4IAD1 is shown. The original view of IRF4IAD1 is shown in inset. Encircled in yellow is the loop Lβ9-β10 with K399 shown previously to be important in interaction with phosphorylated PU.1. As can be seen the region is highly positively charged and may explain its functional significance in making electrostatic interactions with negatively charged phosphate group in PU.1.

The different views of surface electrostatics of IRF5 IAD and IRF3 IAD are shown in Panel b and c of Figure 17, respectively. As pointed earlier the region corresponding to loop Lβ9-β10 of IRF4IAD1 has some helical character in either IADs and may not be suitable for interacting with a binding partner. In addition, in an auto-inhibited state in IAD of IRF3 (and IAD of IRF5), the loop corresponding to loop Lβ9-β10 of IRF4IAD1 (pointed by an arrow in inset in Panel a) is oriented such that it may impede interaction with any incoming partnering protein. Further, in the activated state IRF5 (and IRF3 IAD) functions as a dimer and loop Lβ9-β10 is buried at the dimer interface. Also, worth noting is the apparent bulkiness of IRF5IAD as compared to IRF4IAD.
Figure 17. Surface electrostatics in three different views of IADs of IRF4, IRF5, IRF3 a. IRF4IAD1 b. IRF5IAD
Also, encircled in black in Panel A is the region that forms the helix bundle in IRF4IAD1. The surface charge of helix bundles of IRF4IAD1, IAD of IRF5 and IRF3 are quite different. While the helix bundles of IRF5IAD and IRF3IAD are highly negatively charged, that of IRF4IAD1 is less charged and has some degree of hydrophobicity. Several hydrophobic residues hold the helix bundle intact in IRF3IAD that buries the autoinhibitory elements including residues from helix α1, α3, α4 and additionally from α5.

Figure 17. Surface electrostatics in three different views of c. IRF3IAD
Residues of α1, α3 and α4 are conserved between IRF3IAD and IRF4IAD1 including L247 in α1, L413 in α4 and L365 and T361 in α4 of IRF4IAD1. Residues corresponding to V391, L393 and I395 in helix α5 of IRF3IAD are not found in IRF4IAD1 (see sequence alignment Figure 14) Leu329 in the helix α3 of IRF3IAD is conserved in all other members of IRF family. The corresponding residue in IRF4IAD1 is Leu368. Mutation of this residue to proline resulted in no ternary complex formation with phosphorylated PU.1 (Meraro et al., 1999). The corresponding residues in IRF8IAD and IRF9IAD were also shown to be important in IFN signaling as well (Levi et al., 2002; Meraro et al., 1999). L368 is shown in red in Figure 18 along with some of other conserved hydrophobic residues that form the helix bundle. Since, this residue is conserved across all of the members of IRF family it may be critical in maintaining the helix bundle and protein-protein interaction although most likely not being directly involved in the process. Since, IRF4IAD1 does not have the residues corresponding to the hydrophobic residues of helix α5 in IRF3 (see sequence alignment in Figure 14), L368 may be particularly the most critical residue in maintaining the helix bundle. Mutation of L368 may lead to loss in the integrity of the helix bundle and may affect the overall domain in its ability to form a ternary complex with phosphorylated PU.1.
2.02.7 Small angle X-ray scattering profiles (SAXS) of IRF4IAD1 and IRF4IAD2:

SAXS data collection:

Samples for SAXS data collection were prepared by purifying them on a Superdex-75 column pre-equilibrated in 25mM Tris-base pH 8.0, 500mM NaCl, 5mM BME, 10% glycerol right before data collection for both IRF4IAD1 and IRF4IAD2. Data were collected at the undulator-based beamline, X9 at National Synchrotron Light Source part of Brookhaven National Laboratory (Allaire and Yang, 2011). SAXS data were collected using a MAR165 CCD area detector, which was located at distance around 3.5m from the...
sample. The energy of the X-ray beam was \( \sim 2 \text{ keV} \). Additionally, X9 beamline is also equipped with a Photonic Science CCD detector useful to measure Wide Angle X-ray scattering (WAXS). Together, the two detectors allowed for data to be collected in the \( q \) range of \( 0.005 - 2.0 \, \text{Å}^{-1} \) where \( q = 4\pi \sin(\theta/2)/\lambda \), \( \theta \) is the scattering angle, and \( \lambda \) is the wavelength. The sample solution is run through a vacuum-sealed 1mm capillary eliminating background scattering.

Data was collected on for the IRF4IAD1 at three different concentrations of 1mg/mL, 2.2mg/mL and 4.4mg/mL. Data was collected on for the IRF4IAD2 at two different concentrations of 0.8mg/mL and 0.9mg/mL. Each sample was exposed for 60 seconds and buffer subtraction was performed at the end of data collection.

**SAXS data processing:**

Buffer subtraction and data processing is done using a Python-based packaged developed by the beamline scientist. The package allows users to convert 2D images into one-dimensional SAXS profiles. WAXS data, which also part of the subtracted data is primarily used to study the crystallinity of polymers and was eliminated during further data processing. The buffer subtracted SAXS data were used for subsequent merging, trimming and model building. Merging, trimming and scaling were performed using PRIMUS including calculation of radius of gyration (\( R_g \)). GNOM was used to determine the \( P(r) \) distribution function to obtain \( D_{max} \) for the molecule (Figure 19). SAXS parameters is
provided in Table 4. Subsequently, 10 ab initio models were built using both GASBOR and DAMMIN for IRF4IAD1 and IRF4IAD2. To build an averaged most typical model from these low-resolution ab initio models, DAMAVER suite of programs was used followed by a final run of DAMMIN to generate a final model of IRF4IAD1 and IRF4IAD2.

**SAXS parameters:**

<table>
<thead>
<tr>
<th>Protein</th>
<th>Radius of gyration (R_g)</th>
<th>D_{max}</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRF4IAD1</td>
<td>20</td>
<td>68</td>
</tr>
<tr>
<td>IRF4IAD2</td>
<td>22</td>
<td>77</td>
</tr>
</tbody>
</table>

**Table 4.** SAXS parameters for IRF4IAD1 and IRF4IAD2

**Figure 19.** P(r) distribution function of IRF4IAD1
Averaged IRF4IAD1 *ab initio* model:

The crystal structure of IRF4IAD1 was manually fit into the averaged SAXS molecular envelope in Pymol, a molecular visualization system (Figure 20) (Pymol)

**Figure 20.** SAXS molecular envelope of IRF4IAD1 (view 1)
Figure 20. SAXS molecular envelope of IRF4IAD1 (view 2)
Comparison of the experimental and model scattering curves

Additionally to determine the accuracy of our SAXS model, we used a web-server (FoXS) (Schneidman-Duhovny et al, 2013), (Schneidman-Duhovny et al, 2010) to calculate a theoretical SAXS profile for the atomic resolution X-ray structure of IRF4IAD1 and compared its experimental SAXS profile. We obtained a $\chi$ value of $\sim 0.2$ indicating an almost complete agreement between the theoretical and experimental SAXS profiles. This is also indicated below in a graphical representation.

![Graph showing comparison of experimental and theoretical scattering profiles](image)

**Figure 21.** Comparison of experimental and theoretical scattering profile of IRF4IAD1
Averaged IRF4IAD2 ab initio model

In order to determine the conformation of the AIR, we generated ab initio averaged model from the scattering data collected for IRF4IAD2. The $R_g$ and $D_{\text{max}}$ values were slightly larger than IRF4IAD1 suggesting that this region will not be an unfolded loop (Table 4). We generated models of IRF4IAD2 with different conformations of the AIR and compared the experimental scattering profile with the one calculated for the different models. Two separate models were built – one using the X-ray structure of IRF4IAD1 as the template to generate Model 1 and another using the X-ray structure of IRF3IAD as a template to generate Model 2. The two models are shown in Figure 22. Models were built using the Modeller interface of Chimera (Sali and Blundell, 1993). Note the orientation and position of the last 30 residues (highlighted in red in each model) (Figure 22). In model 1 the last 30 residues are more extended while in model 2 those residues are folded back on to the helix bundle similar to helix $\alpha_5$ of IRF3IAD. We used these individual models as an input pdb in BILBOMD (Pelikan et al, 2009), a web server allowing molecular dynamics based conformational sampling with integrated structure validation using FoXS to generate Minimum Ensemble Search (MES) models. The experimental scattering profile of IRF4IAD2 was used for structure validation.
The MES model generated using Model 1 as an input pdb in BILBOMD best fit the experimental data shown in Figure 23. The chi value ($\chi$) for the fit was 0.32. The MES model was manually fit into the averaged \textit{ab intio} model in Pymol shown in Figure 24.

\textbf{Figure 22.} The two models used as an input for BILBOMD. Note the different orientation of the last 30 residues shown in red.
Figure 23. Comparison of experimental scattering profile of IRF4IAD1 compared to the theoretical scattering profile generated for Model 1.
Figure 24. Two different views of manual fitting of MES model obtained from BILBOMD using Model 1 fit into the SAXS molecular envelope.
The MES model generated using Model 2 as an input pdb in BILBOMD did not fit the experimental data well shown in Figure 25. The chi value ($\chi$) for the fit was 0.69. This indicates that the C-terminal extension of IRF4IAD2 does not fold like its auto-inhibited counterpart of IRF3.

**Figure 25.** Comparison of experimental scattering profile of IRF4IAD1 compared to the theoretical scattering profile generated for Model 2.

2.03 Discussion:
Although, overall the IADs of IRF4, IRF3 and IRF5 are superimposable, the crystal structure of IRF4IAD1 has helped shed light on the unique features of IRF4. The N-terminal helix α1 in IRF4IAD1 is quite flexible in spite of forming a helix bundle with helices α3 and α4. Surface electrostatics of the helix bundle of IRF4IAD1 shows that the bundle is not as charged as those of IRF3 and IRF5. Yet the hydrophobic residues that form the helix bundle are quite conserved except for the residues corresponding to helix α5 of IRF3. L368 of helix α3 in IRF4IAD1 seems to be critical in maintaining the helix bundle in IRF4. Mutation of this residue results in a mutant incapable of forming the ternary complex with PU.1 and DNA (Meraro et al., 1999). Disruption of the helix bundle may be the reason for the inability of IRF4 to form the ternary complex though direct interaction between L368 and PU.1 is highly unlikely. In addition, K399 in loop Lβ9-β10 sits in a pocket of highly positively charged residues. The residue is critical in formation of the ternary complex with phosphorylated PU.1 (Brass et al., 1999). The pocket may be involved in making electrostatic interactions with the negatively charged phosphate of PU.1.

SAXS molecular envelope of IRF4IAD2 suggests that the residues forming the AIR are structured. Yet their position is most likely not similar to helix α5 of IRF3IAD, which folds back onto the helix bundle and buries the auto-inhibitory elements. Based on our SAXS molecular envelopes, AIR seems to be positioned away from the helix bundle and towards the center of the full-length IRF4 making interactions with regions of the putative linker.
Chapter 3

SAXS solution structure of full length IRF4 and the role of AIR in the activation process

3.01 Introduction:

Transcription factors are multi-domain proteins with each domain having specific functions in the gene regulation process. Multi-domain proteins can have increased stability and also can have specific cooperative functions. Knowledge of the 3D structures of individual domains although valuable presents only small part of the story pertaining to the regulatory mechanism. For example, in the case of human nuclear hormone receptor structural studies were performed on the isolated DNA and ligand-binding domains but the cooperative interactions between the domains that modulates the properties of the full-length receptor was not demonstrated until the structure of the full length peroxisome proliferator-activated receptor (PPAR-γ) in complex with retinoid X-receptor (RXR-α) was solved by Chandra et al (Chandra et al, 2008). The crystal structure of the complex provided information about missing extensions in the two proteins for example in the C-terminus of RXR-α, which the authors suggest, is critical to RXR-α in forming heteromeric complexes with different binding partners. Human mitochondrial transcription factor A (TFAM) is another example of a multi-domain transcription factor where a flexible linker connects two high-mobility group (HMG) protein domains which becomes structured and is involved in interaction with DNA once bound to the ligand mitochondrial light strand promoter (LSP) DNA (Rubio-Cosials et al, 2011).
In case of IRF family of proteins, the crystal structure of individual DNA binding domains (DBD) and interferon activation domains (IAD) are available for several members of the family. But structural information of full-length proteins is lacking and hence, there is a void in the understanding of the characteristic regulatory mechanism. In case of IRF3 and IRF5, structural information about the interferon activation domain in its auto-inhibited and phosphorylation induced activated state respectively, described major conformational changes occurring upon phosphorylation (Qin et al, 2003), (Chen et al, 2008). Although the structures of the full-length proteins were unavailable phosphorylation induced dimerization presented to be a viable mode of activation.

In case of IRF4, the proposed and currently accepted mechanism involves huge conformational changes in the full-length protein wherein the DBD and IAD pivot around a flexible putative linker (Brass et al, 1999). The authors suggested that the auto-inhibitory region (AIR) in the full-length protein physically impedes the DBD to generate a closed more globular state. Although, structural information is available for IRF4 DBD (Escalante et al, 2002) and IAD (Chapter 1), even low-resolution structural information about how these domains are arranged in the full length of protein is unavailable. This lack of information limits a full mechanistic understanding of activation of IRF4. Secondary structure predictions of the putative linker region suggest the region to be unstructured but this could primarily be a result of lack of availability of any structural information of the region. It is worth pointing out here again that the AIR at the C-terminal of IRF4 was proposed to keep the full-length protein in the closed state. Furthermore, the protein was proposed to undergo huge conformational changes in the presence of its binding partner
going between an ‘open’ more extended state in the activated form and a ‘closed’ more globular state (Figure 1). Biophysical techniques like analytical ultracentrifugation (AUC) as well as small angle X-ray scattering (SAXS) provide excellent methods to access such conformational changes if present.

Sequence alignment of different members of IRF family of proteins (like IRF3, 5 &7) reveals that IRF4 is the only member with an N-terminal 20 amino acid extension (Figure 1). Cartoon representation of full-length IRF4 in open & closed state. According to the currently accepted mechanism such conformational changes may occur during the activation process.
We cloned, expressed and purified full-length IRF4 protein with and without the first 20 and/or last 30 residues (Figure 3). We performed AUC and SAXS studies on individual proteins to obtain the overall shape and size of the different IRF4 constructs to evaluate the conformation of the protein. We also performed fluorescence anisotropy binding experiments to determine the effect of the N-terminus and C-terminus extensions on the dissociation constants ($K_d$) of the four different constructs. We performed these experiments with a 21bp DNA with the ISRE IRF4 binding site adjacent to the PU.1 binding site. The $K_d$ determined here also aided in explaining the role of the N-terminal 20 amino acid residues & the AIR in the gene activation process.
3.02 Results:

3.02.1 Purification of the different constructs:

IRF4\textsubscript{wt}:

Mouse full-length IRF4 construct 1-450 (IRF4\textsubscript{wt}) was subcloned into pET-15b TEV using the appropriate primers by a procedure similar to that mentioned in the results section of chapter 2 in purification of IRF4IAD1. Positive clones of IRF4\textsubscript{w} were transformed into BL21-pLysS\* \textit{E. coli} expression cells and an overnight pre-culture were grown using the appropriate antibiotics stress. IRF4\textsubscript{wt} was grown at 37\textdegree C until the OD reached ~0.6, induced with 0.5mM IPTG & harvested after 6hrs. The cells were flash-frozen, thawed & resuspended in 25mM Tris-base pH 8.0, 500mM NaCl, 2mM TCEP, 5mM imidazole, 10%
glycerol, sonicated and centrifuged at 20000 rpm for 30 mins. Protein was eluted from a Ni-NTA column with 300mM Imidazole and the Histidine (His) tag was removed using the home grown and purified Tobacco Etch Virus (TEV)-protease. The cleaved protein was further purified on Ni-NTA column to remove the His-tag as well as the TEV-protease. Further purification of the cleaved protein was done using a phenyl sepharose with 25mM Tris-base pH 8.0, 1M NaCl, 1mM TCEP, 1mM EDTA as the binding buffer and a similar buffer with 25mM NaCl as the elution buffer. The last polishing step of the purification was done with gel filtration on a Superdex-75 column in 25mM Tris-base pH 8.0, 300mM NaCl, 2mM TCEP & 5% glycerol. The protein was concentrated to about 10mg/mL (0.196mM) & stored at -80°C in 100µL aliquots.

IRF4ΔC, IRF4ΔN and IRF4ΔNC were cloned, expressed and purified using a similar protocol described above. Yield of ~4-5mg of protein was obtained per litre of the culture for each construct.
Figure 4. SDS-page gel electrophoresis results for the purification of IRF4<sub>wt</sub> a. Ni-NTA column purification b. TEV-protease cleavage c. Second nickel column purification
**Figure 4.** SDS-page gel electrophoresis results for the purification of IRF4<sub>wt</sub> d. Phenyl Sepharose column purification e. Gel purification
3.02.2 *Analytical Ultracentrifugation studies*:

We performed analytical ultracentrifugation studies on the four different constructs to assess their state of oligomerization if any and to detect any conformational changes between the different constructs. Sedimentation velocity profiles of the different IRF4 constructs are shown in Figure 5. Experimental and theoretical sedimentation coefficients of sphere of equal mass are shown in Table 1.

![Figure 5](image_url)

**Figure 5.** Sedimentation velocity profiles of the different IRF4 constructs. The profiles are overlaid to compare their sedimentation coefficients. Total area under each cover is comparable. IRF4<sub>wt</sub> and IRF4ΔC may have slight oligomerization, which is in fast equilibrium.
The sedimentation coefficients for the various constructs show that the open-closed model of activation for IRF4 is incorrect. The removal of AIR in IRF4ΔC does not result in drastic changes in the sedimentation coefficient. It is also worth noting that the proteins appear to be elongated (compare the $S_{w,20}$ of a sphere of equal mass with the experimental $S_{w,20}$).

<table>
<thead>
<tr>
<th>Protein</th>
<th>Experimental $S_{w,20}$</th>
<th>Theoretical $S_{w,20}$ for a sphere</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRF4$_{wt}$</td>
<td>3.4</td>
<td>4.9</td>
</tr>
<tr>
<td>IRF4ΔC</td>
<td>3.0</td>
<td>4.74</td>
</tr>
<tr>
<td>IRF4ΔN</td>
<td>3.1</td>
<td>4.7</td>
</tr>
<tr>
<td>IRF4ΔNC</td>
<td>3.2</td>
<td>4.5</td>
</tr>
</tbody>
</table>

*Table 1.* Experimental and theoretical sedimentation coefficients of individual constructs and sphere of equal mass. $S_{w,20}$, corrected for water at 20°C. Theoretical $S_{w,20}$ is calculated for a sphere of equal mass.

The sedimentation coefficients for the various constructs show that the open-closed model of activation for IRF4 is incorrect. The removal of AIR in IRF4ΔC does not result in drastic changes in the sedimentation coefficient. It is also worth noting that the proteins appear to be elongated (compare the $S_{w,20}$ of a sphere of equal mass with the experimental $S_{w,20}$).

### 3.02.3 Fluorescence anisotropy experiments:

To understand the role of the N-terminal 20 amino acid residues and the AIR in the gene activation process mediated by IRF4 we performed fluorescence anisotropy binding experiments titrating each of the different constructs against a constant concentration of a 21bp 5’-fluorescein labeled DNA (F-DNA) (5’-FAM/ TAT TTT CCT TCA CTT TGG TTT-3’). Excitation maxima of 498nm & emission maxima of 526nm were used for the experiment corresponding to the values reported for fluorescein tag. The fluorescence

84
saturation curves of fraction of protein bound to the 21bp F-DNA (Y) against increasing concentration of protein (X) is represented in Figure 6 for individual constructs.

**Figure 6.** Fluorescence saturation curve for the different constructs a. Fluorescence saturation curve for IRF4_{wt}.
Figure 6. Fluorescence saturation curve for the different constructs b. Fluorescence saturation curve for IRF4ΔC
c. Fluorescence saturation curve for IRF4ΔN
The dissociation constants ($K_d$) for individual constructs calculated using the one site binding equation

$$Y = B_{\text{max}} \cdot X / (K_d + X)$$

are reported in Table 4. $B_{\text{max}}$ here is the maximal specific binding and the dissociation constant is a ratio of $k_{\text{off}}/k_{\text{on}}$ ($k_{\text{off}}$ – off rate of binding and $k_{\text{on}}$ – on rate of binding) for a protein-ligand binding event in equilibrium. Dissociation constants for the different constructs are shown in Table 2.
Surprisingly, we found that the removal of the N-terminal 20 amino acids residues results in a substantial improvement in the ability of the protein to bind the 21bp F-DNA indicated by a marked decrease in $K_d$. Removal of the AIR does not result in any appreciable change in $K_d$. This result is in direct contrast of what would be expected if the currently accepted model of the activation process were accurate. The mechanism was proposed based on gel-shift assays performed on truncated proteins. Using the crystal structure of IRF4IAD1 as a guide, we assessed the constructs that were used for the gel-shift assays and have concluded that for structural integrity of the truncated proteins where AIR was removed were not intact and the results obtained in the gel-shift assays may be erroneous.

3.02.4 SAXS studies:
SAXS curves for each of the four constructs are depicted in Figure 7. SAXS profile for each construct was analyzed for possibility of aggregation in the sample, which is indicated by an upward turn in the low q region of a scattering profile (region encircled in the scattering profile of IRF4\textsubscript{wt} in Figure 7 a.). None of the samples showed any signs of aggregation & thus, further analysis of each sample could be carried out successfully.

![SAXS scattering profile for different constructs](image)

**Figure 7.** SAXS scattering profile for different constructs a. SAXS scattering profile for IRF4\textsubscript{wt}. Note that there is no sign of aggregation in the low q region encircled in black.
Figure 7. SAXS scattering profile for different constructs

b. SAXS scattering profile for IRF4ΔC.

c. SAXS scattering profile for IRF4ΔN.
Plots for pair-distribution function (P(r) function) are presented in Figure 8. The P(r) distribution function is used to calculate the maximum length of the scattering molecule. P(r) distribution function of each of the four constructs have been overlaid here for comparing the difference in $D_{\text{max}}$ between the different constructs.
Kratky plots useful in determining the extent of unfoldeness in proteins were also plotted and were helpful in comparing the extent of flexibility between the different constructs (Figure 9).
The basic structural parameters that describe the individual constructs; radius of gyration (Rg), maximal dimension (D_{max}) and Porod exponent are tabulated in Table 3.

<table>
<thead>
<tr>
<th>Protein</th>
<th>R_g (Å)</th>
<th>D_{max} (Å)</th>
<th>Porod exponent</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRF4_{nt}</td>
<td>39</td>
<td>127</td>
<td>2.2</td>
</tr>
<tr>
<td>IRF4ΔC</td>
<td>38</td>
<td>123</td>
<td>2.7</td>
</tr>
<tr>
<td>IRF4ΔN</td>
<td>35</td>
<td>121</td>
<td>2.9</td>
</tr>
<tr>
<td>IRF4ΔNC</td>
<td>34</td>
<td>118</td>
<td>3.2</td>
</tr>
</tbody>
</table>

**Table 3.** R_g, D_{max} and Porod exponent of the different constructs
Based on the P(r) distribution functions, the corresponding D\textsubscript{max} and the Kratly plot, we conclude that IRF4\textsubscript{wt} is present in an extended conformation in the DNA unbound state and not in a globular conformation as would be expected if the currently accepted model were true. Furthermore, we do not observe any large conformational changes in the SAXS profiles and associated parameters of the different constructs. An increase in Porod exponent from 2.2 for IRF4\textsubscript{wt} to 3.2 for IRF4\Delta NC is indicative of systematic decrease in flexibility between the proteins.

The molecular shape for individual constructs was determined using the programs available in ATSAS software suite for \textit{ab-initio} reconstruction (ATSAS). Using the GNOM (Svergun, 1992) output file, which specifies the D\textsubscript{max} for the protein, ten models were generated in two different \textit{ab-initio} reconstruction programs, DAMMIN (Svergun, 1992) and GASBOR (Svergun \textit{et al}, 2001) for each construct individually. Each set of the ten models from DAMMIN and GASBOR were averaged separately using SUPCOMB (Svergun, 2001). The averaged model for each construct from either of the two programs was run through DAMMIN one last time using the averaged model as the input pdb to be fit into the SAXS profile (.dat file). Normalized spatial discrepancy (NSD) parameters were calculated (reported in Table 4.) to compare the models generated from either of the two programs for each construct and increased our confidence in the models.
We used the experimental sedimentation co-efficient values generated by the AUC studies as a parameter to increase our confidence in the *ab-initio* models. We used US-SOMO (Rai *et al.*, 2005), (Brookes *et al.*, 2010b), (Brookes *et al.*, 2010a) in Ultrascan III software to calculate a theoretical sedimentation co-efficient for the final averaged bead models generated from either DAMMIN or GASBOR for the individual constructs. These were compared to the experimental sedimentation coefficients obtained for the different constructs (Table 5).

<table>
<thead>
<tr>
<th>Protein</th>
<th>Normalized Spatial discrepancy between Gasbor averaged &amp; Dammin averaged models</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRF4&lt;sub&gt;wt&lt;/sub&gt;</td>
<td>0.699</td>
</tr>
<tr>
<td>IRF4ΔC</td>
<td>0.693</td>
</tr>
<tr>
<td>IRF4ΔN</td>
<td>0.7</td>
</tr>
<tr>
<td>IRF4ΔNC</td>
<td>0.673</td>
</tr>
</tbody>
</table>

*Table 4.* NSD comparing the averaged Gasbor and Dammin models of the different constructs. (NSD of less than one is indicative of a good correlation between the different models being compared.)
We observed a good correlation between the theoretical S-value calculated based on the SAXS \textit{ab-intio} model and the experimental S-values for the individual constructs (The partial specific volume of the protein & the viscosity of the buffer were corrected for before calculating the S-values) Table 5.

Superimposition of IRF4\textsubscript{wt} on IRF4\textDelta NC is shown below in Figure 10a. It can be seen that there are no significant differences between the overall shape when both the N- and the C-termini of the protein are removed. Interestingly, IRF4\textDelta C shows the presence of a broad density (encircled in red) that is absent in both IRF4\textDelta N and IRF4\textDelta NC (Figure 10 b and c). This region may correspond to the N-terminal 20 residues in the absence of the AIR indicating that there may be some indirect interaction between the N- & the C-terminal residues presumably through the linker that is lost when the AIR is removed.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Experimental $S_{w,20}$</th>
<th>Theoretical $S^*_{w,20}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRF4\textsubscript{wt}</td>
<td>3.4</td>
<td>3.2</td>
</tr>
<tr>
<td>IRF4\textDelta C</td>
<td>3.0</td>
<td>3.0</td>
</tr>
<tr>
<td>IRF4\textDelta N</td>
<td>3.1</td>
<td>3.2</td>
</tr>
<tr>
<td>IRF4\textDelta NC</td>
<td>3.2</td>
<td>3.2</td>
</tr>
</tbody>
</table>

\textbf{Table 5.} Experimental and theoretical sedimentation coefficients of individual constructs and calculated for Gasbor models. $S_{w,20}$, corrected for water at 20\textdegree C. Theoretical $S_{w,20}$ were calculated for the SAXS envelopes obtained from Gasbor.
Figure 10. SAXS molecular envelope of individual constructs a. Superimposition of IRF4<sub>wt</sub> (green) on IRF4ΔNC (white). b. Superimposition of IRF4ΔC (yellow) on IRF4ΔNC (white) c. Superimposition of IRF4ΔN (blue) on
We used the program CORAL (Petoukhov, 2012) to perform rigid body modeling to fit the individual DBD and IAD against the SAXS profile and into the *ab-intio* model generated for IRF4ΔNC. The two domains were manually moved into the SAXS molecular envelope of IRF4ΔNC until there was visual satisfaction of the fit. Next we fixed the initial coordinates of the individual domains fit through visual inspection for a run of CORAL to generate Model 1 (Figure 12). We also ran CORAL by not fixing the initial coordinates of the individual domains to generate Model 2. The loops utilized for fitting the putative linker connecting the two domains linker are chosen from a pre-defined library of loops. Finally, simulated annealing is utilized to find the optimal orientations of the atomic structures and approximate conformations of loops (Figure 11).
Figure 11. Flow chart representing the protocol used to generate models in CORAL.
The most probable orientation of the two domains, DBD & IAD is likely to be in-between the two representative models presented in Figure 12. In Model 1 there are minimal clashes between the N- & the C- terminal domains and the putative linker and the $R_g$ of model 2 of 30Å is close to the experimental $R_g$ of 36Å.

**Figure 12.** Two distinct models generated using CORAL
The individual models (Model 1 & Model 2) were fit into the SAXS molecular envelope of IRF4ΔNC using the program Supcomb (Svergun, 2001). The fit is shown in Figure 13. Model 1 fits into the envelope better; we calculated an NSD of 1.167 between model 1 & the envelope. We obtained a slightly higher NSD of 1.259 between Model 2 & the SAXS envelope.

**Figure 13.** Supcomb based fitting of the two different models generated from CORAL into the SAXS envelope of IRF4ΔNC.
The two domains are most likely located at the either ends of IRF4ΔNC suggested by the two models generated using CORAL. This can be extrapolated to IRF4\textsubscript{wt} since the two proteins do not differ considerably in their overall shapes. In addition, the model also suggests the orientation of the DBD & IAD as facing the centre of IRF4ΔNC. This may also be fairly accurate since the different constructs do not show long extensions on either ends. Interestingly, docking of the DBD & IAD into the SAXS molecular envelope of IRF4ΔNC suggests that the putative linker that joins the N-terminal DBD with the C-terminal IAD must be folded. The physical space between the DBD & IAD as well as the \( R_g \) and \( D_{\max} \) of IRF4ΔNC reinforce this observation. Nevertheless, it does not rule out the possibility that the region has some degree of flexibility yet there appears to be an overall domain characteristic to the region. The region may not be an independently folding domain & seems to require the DBD & IAD to maintain it in the folded state. This was suggested by our inability to successfully express & purify the constructs with the DBD and different lengths of the putative linker (see Results section of Chapter 4). As pointed before we did not observe conformational changes corresponding to distinct open & closed states between the different constructs. An average NSD of 0.688 was calculated comparing the SAXS molecular envelopes of the four constructs further suggesting the envelopes have subtle and not extensive changes in structure.

### 3.03 Discussion:

The currently accepted model of gene regulation by IRF4 lacks any support from structural studies performed on the full-length protein. Based on the low-resolution SAXS
models as well as the analytical ultracentrifugation studies presented here it is clear that the protein does not undergo huge conformational changes when AIR is removed and is present in solution in an extended conformation in the DNA unbound state. Furthermore, although the putative linker may have some flexibility to it, based on our SAXS models the region appears more structured than previously reported and may be rechristened as a ‘novel’ domain.

To understand the inhibitory mechanism of the N-terminal extension on binding, we superimposed the crystal structure of DBD of IRF4 on the solution NMR structure (2DLL) of the DBD (Figure 14).

The region is unstructured and may physically impede the IRF4DBD from binding to the 21bp DNA. In other words, the flexible N-terminal residues may decrease the binding affinity of full-length IRF4. Thus, although IRF4wt may bind DNA, the protein-DNA complex may not be stable. It is worth noting that IRF4 is the only member of the IRF family that has this N-terminal extension. Furthermore, IRF4 along with IRF8 are the only members that have a high degree of promiscuity with respect to their binding partners. Also, the crystal structure of the ternary complex between IRF4/PU.1/21bpDNA pointed to the cooperativity between the two proteins through residues V111, L116 and D117 in IRF4DBD and R222 and K223 in PU.1DBD. Hence, the cooperativity between IRF4 and
its binding partner may play a role in stabilizing the binding of IRF4\textsubscript{wt} to DNA and may explain the need for a binding partner for IRF4 in its regulatory mechanism. Interestingly, when the SAXS molecular envelope of IRF4\textsubscript{ΔNC} is integrated with the crystal structure of the ternary complex of DBD of PU.1, IRF4 and DNA it becomes obvious that the additional interaction between PU.1 and IRF4 would require conformational changes in both the proteins to successfully regulate transcription. The crystal structure of the ternary complex between PU.1, IRF4 and DNA will surely provide further insight into the mechanism of regulation.

Figure 14. Superposition of crystal structure of IRF4 DBD on its NMR structure. Flexible N-terminus loop is shown in red.
Chapter 4

Structural details of the putative linker of IRF4

4.01 Introduction:

Spacers of short amino acid sequences, frequently called linkers are generally thought to primarily to separate domains in multi-domain proteins. However, the most important role of linkers is to ensure that undesirable interactions between the individual domains are prevented. Thus, linkers may have some degree of rigidity conferred to them (George and Heringa, 2002; Reddy Chichili et al, 2013). Length of linkers can range from 3-40 amino acids. The length of the linkers is also critical & has been shown to affect protein stability and domain orientations (van Leeuwen et al, 1997; Robinson and Sauer, 1998). For example, decreasing the length of the linker connecting the sub-domains of phosphorylated smooth-muscle myosin leads to cessation of actin translocation activity (Ikebe et al, 1998). The conformational activation of α1 intergrins is another example where the length of the linker is shown to be critical (Weitz-Schmidt et al, 2011). The C-terminal (C-linker) connects the α domain to the β domain forming the α/β heterodimeric integrins. The length of the C-linker was demonstrated to be critical in maintaining negative regulation of the α domain by the βI domain. Linkers are also important in maintaining cooperative interactions between domain (George and Heringa, 2002). For example, it was shown for Src family of kinases that mutation of a critical residue (Trp260Ala) in the linker leads to uncoupling of the catalytic and regulatory domains ((LaFevre-Bernt et al, 1998), (Briggs...
and Smithgall, 1999)). The coupling ensures that the catalytic activity of the protein is repressed. Linkers may also serve as regions that directly interact with a ligand and provide additional binding energy to ensure robust binding of the ligand. For instance, the structure of human PAX6 PD (paired DNA binding domain) DNA complex showed that the glycine rich linker of human PAX6 makes minor groove interaction with the DNA ((Mishra et al, 2002; Xu et al, 1999). More recently, linkers have been documented to be involved in oligomerization events of specific proteins. In case of hsp70, Aprile et al (Aprile et al, 2013) showed that oligomerization of hsp70 involves specific interaction between the interdomain linker of one molecule of hsp70 with substrate-binding domain of another molecule of hsp70. Another instance of linker-mediated oligomerization is shown for Rep68 structural protein of Adeno Associated Virus-2 (AAV-2). Herein, an aromatic amino acid, Tyrosine224 was shown to be critical in ensuring oligomerization and infectivity of AAV-2 virus (Zarate-Perez et al, 2012).

Linkers may be broadly classified into two categories - helical and non-helical (George and Heringa, 2002). Helical linkers are often poly-glycine linkers and are often also referred to as soft peptide linkers (Wriggers et al, 2005). These linkers have been frequently used in protein engineering experiments to fuse two domains to allow for some degree of flexibility conferred to the fusion proteins. Non-helical linkers or often also called ‘molecular rulers’ are frequently poly-proline linkers. They have been referred to as molecular rulers because these hard linkers may be used to predict end-to-end distances and hence depths of binding pockets. Interestingly, proline residues are found to be the terminal residues in linking different domains. This may be because of the inability of
proline to form any regular secondary structure and thus ensure rigid separation of domains. The *trans* conformation of proline residues maintains rigidity while neighbor dependent change in conformation of proline residue from *trans* to *cis* confers flexibility to these linkers. Other residues that are frequently found in linker regions glutamine, arginine, glutamate, and serine (Wootton and Drummond, 1989).

In IRF4, the putative linker is much longer, composed of 100 residues. It has about 17% of proline residues, which would categorize it as a non-helical linker. Hence, the putative linker is also expected to have high rigidity and may lack any appreciable secondary structure. Previously accepted mechanism for IRF4 activation suggested the linker to be highly flexible. Intriguingly, there is also a report suggesting post-translational regulation of IRF4 wherein peptidyl-prolyl isomerase was shown to catalyze the *trans* to *cis* isomerization of proline residues in the putative linker region (Mamane *et al*, 2000). The isomerization may render some flexibility to the linker. The results obtained from the SAXS studies of full-length IRF4 that suggested the possibility of a structured linker, thus we used X-ray crystallography and SAXS studies to characterize the linker region.

We performed crystallization trials of IRF4ΔNC with a 17bp DNA, which later proved to be crystals of the degradation product of IRF4ΔNC. We also attempted to express and purify various different constructs of IRF4 with the DNA binding domain and different regions of the linker. SAXS scattering profile of IRF420-201 & IRF4202-450 showed aggregation of the sample and further analysis of the sample to generate *ab-initio* SAXS molecular envelopes was not carried out.
4.02 Results:

4.02.1 Crystallization of full-length IRF4 with a 17bp-DNA:

Crystallization of full-length IRF4 was attempted using two constructs IRF4ΔN (20-450) and IRF4ΔNC (20-420). Crystallization trials were performed simultaneously in the presence and absence of DNA. Different lengths of DNA from 17-21bp were utilized for the trials. In addition to a wealth of information provided by a crystal structure of full-length IRF4 in presence of DNA, the DNA as a ligand can minimizes the different probable conformations of the full-length protein and makes the protein more favorable for hits in crystallization trials. The two different constructs were mixed with the individual DNAs in molar ratio of 1:1.2 (protein:DNA) and concentrated upto ~16-20µg/mL (350-430µM of protein concentration). Hanging drop crystallization trials were performed with the available factorial screens from Hampton and Emerald Biosciences at 20°C.

Microscopic needle shaped clustered crystals were obtained in Emerald Biosciences condition 0.1M MES pH6.5, 12% PEG 20K for the protein complex with a 17bp DNA (5’-AAG GAA GTG AAA CCA GG -3’) (Figure 1). The condition showed high degree of precipitation and phase-separation and the crystals appeared from the precipitated drops in about 5-7 days. An optimized condition of a slightly higher percentage of 15% PEG20K in 0.1M MES pH 6.0 showed slight improvement in the appearance of the crystals. These bunches of crystals were sent to Argonne National Laboratory GM/CA beamline, which is equipped with a microfocus beam that can be utilized to collect data on microcrystals. Some of the crystals diffracted to about 5Å and the diffraction pattern indicated that the crystals were of protein and/or protein-DNA complex and not of salt.
Figure 1. Cluster of crystals of IRF4ΔNC with 17bp DNA and diffraction pattern a. Cluster of crystals of IRF4ΔNC with 17bp DNA b. Image of the single crystals being shot using a micofocus beam at Argonne National laboratory (Illinois) c. A representative diffraction image showing few spots
To increase the size of the crystals and to obtain single crystals, an additive screen was performed around the optimized condition. 40-80mM cesium chloride dramatically improved the crystals and resulted in single crystals measuring about 100 microns (Figure 2).

Figure 2. Single crystals of IRF4ΔNC with 17bp DNA and diffraction pattern a. Single crystals of IRF4ΔNC with 17bp DNA b. A representative diffraction image.
The crystals diffracted to ~ 2.9Å and could be indexed with a P222 spacegroup. The crystals were still considerably small and a complete data set could not be collected. The data set was collected at the X25 beamline at NSLS. The data collection statistics are provided in Table 1.

<table>
<thead>
<tr>
<th>Data collection</th>
<th>IRF4ΔNC / 17bp DNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Space group</td>
<td>P222</td>
</tr>
<tr>
<td>Cell dimensions</td>
<td></td>
</tr>
<tr>
<td>$a, b, c$ (Å)</td>
<td>39.1, 41.3, 114.8</td>
</tr>
<tr>
<td>$a, b, g$ (°)</td>
<td>90 90 90</td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>2.9</td>
</tr>
<tr>
<td>$R_{merge}$</td>
<td>6.8 (13.6)</td>
</tr>
<tr>
<td>$I/\sigma I$</td>
<td>13.5 (5.4)</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td><strong>48.75 (13.3)</strong></td>
</tr>
</tbody>
</table>

**Table 1.** Data collection statistics of IRF4ΔNC with 17bp DNA. Completeness for the data set was lower than 50%
Surprisingly, the unit cell parameters of the crystals were smaller than would be expected for IRF4ΔNC of molecular weight ~46kDa. We investigated whether the crystals were of a degradation product of IRF4ΔNC. An SDS-page gel indicated that the crystals were indeed degraded products of IRF4ΔNC and the molecular weight of the degradation product was ~28kDa. Using the molecular weight of the degradation product as a parameter, we next investigated whether the crystals were of the N-terminal DBD domain plus some region of the putative linker (NTD) or of the C-terminal IAD (both have a molecular weight ~28kDa). Western blot of the crystals using a primary antibody against the N-terminal DBD confirmed that the degraded product contained the N-terminal DBD along with some region of the putative linker.

![Figure 3](image_url)  
**Figure 3.** Western blot analysis of the crystals of IRF4ΔNC. The crystals were confirmed as having the DBD.

We will further try to improve the condition to be able to collect a better data set to determine the crystal structure of the putative linker. To this end, we designed several
different constructs of DBD along with some additional residues from the linker. We aimed at expressing, purifying & crystallizing these constructs.

4.02.2 Purification of protein constructs for additional crystallization attempts:

Expression of different protein constructs with IRF4DBD plus varying regions of the putative linker were attempted in BL21\textsuperscript{*}pLysS cells. All the constructs IRF4 1-237, IRF4 20-225, IRF4 20-231 and IRF4 20-237 did not yield more than 4mg/L of the culture after the first immobilized-metal affinity chromatography (Figure 4). IRF4 20-237 was further purified on a gel-filtration column although the purity of the protein did not improve (Figure 4. b.). About 500µg of total purified protein was obtained from 1L culture. Protein was detected in inclusion bodies for each of the constructs and change in induction temperature to 18°c did not improve the yield in soluble fraction.

\textbf{Figure 4.} SDS-gel electrophoresis result of different constructs of IRF4 with varying regions of the linker. \textbf{a.} SDS-gel electrophoresis result of IRF4 1-237 expressed in BL21\textsuperscript{*}pLysS cells. Over expressed protein is seen in inclusion bodies (shown by arrow)
Figure 4. SDS-gel electrophoresis result of different constructs of IRF4 with varying regions of the linker. b. SDS-gel electrophoresis result of IRF4 20-237 expressed in BL21*pLysS cells. Left panel shows the fractions obtained after purification on a Ni-NTA column & right panel shows the fractions obtained after purification gel filtration column.

Figure 4. SDS-gel electrophoresis result of different constructs of IRF4 with varying regions of the linker. c. SDS-gel electrophoresis result of IRF4 20-231 expressed in BL21* pLysS cells. Over expressed protein is seen in inclusion bodies as well as very little protein is seen in soluble fraction (shown by arrow).
Expression of IRF4 20-225 was attempted in BL21pLysS Codon plus cells (RIPL cells) with a slight increase in expression level (Figure 4. d). The protein was further purified after removing the His-tag on a gel filtration Superdex-75 column pre-equilibrated in 25mM HEPES pH 7.5 150mM NaCl, 1mM TCEP. We obtained ~ 1mg/L of purified IRF4 20-225 after the final gel filtration polishing step.

**Figure 4.** SDS-gel electrophoresis result of different constructs of IRF4 with varying regions of the linker d. SDS-gel electrophoresis result of IRF4 20-225 expressed in BL21pLysS Codon plus cells. Over expressed protein is seen in inclusion bodies as well as very little protein is seen in soluble fraction (shown by arrow)
4.02.3 *Thrombin fragments of IRF4ΔN*:

Using peptidecutter from Expasy, a pseudo-thrombin protease cleavage site was determined after residue 201 of IRF4\textsubscript{wt}. Since, we had little success with expression and purification of the above-mentioned constructs we attempted to obtain fragments of IRF4ΔN (IRF4 20-201 & IRF4 202-450), purify the protein & perform crystallization trials on the individual fragments separately.

**Figure 5.** Western blot analysis of thrombin proteolysis of IRF4ΔN. The fragment IRF4 20-201 showing the presence of the DBD of IRF4 is shown here (red arrow)
We took advantage of this site and performed thrombin cleavage on IRF4ΔN purified to homogeneity as per the procedure previously described (Results section of Chapter 3). The protein was desalted into thrombin cleavage buffer (25mM Tris-base pH 8.0, 200mM NaCl, 10% glycerol) and thrombin was added at 1U per mg of protein and the cleavage reaction continued at room temperature for 4-6hrs. The reaction was stopped with protease inhibitor p-aminobenzamidine-agarose beads by continuous shaking for 45 mins. The two fragments were separated on a phenyl-sepharose column (Figure 6) followed by a final step of gel filtration into 25mM HEPES pH 7.5, 150mM NaCl, 1mM TCEP.

**Figure 6.** Purification of the fragments (IRF4 20-201 & IRF4 202-450) on a phenyl sepharose column. The two fragments separated as two peaks in the elution profile. Uncut IRF4ΔN was separated from IRF4 202-450 on the gel filtration column.
4.02.4 SAXS studies of IRF4 20-201 & IRF4 202-450:

SAXS data were collected as previously mention in the results section of Chapter 3 though the Mail-in service at the Advanced Light source at Berkeley. Both IRF4 20-201 & IRF4 202-450 were buffer exchanged into 25mM Tris-base pH 8.0, 300mM NaCl, 2mM TCEP, 5% glycerol before shipping the samples for data collection.

The SAXS scattering profile of IRF4 20-201 showed an upward turn at low q range indicative of aggregation and was not processed any further (Figure 7).

![IRF4 20-201](image)

**Figure 7.** Scattering profile of IRF4 20-201. An upward turn is seen in the low q range (encircled in black)
The SAXS scattering profile of IRF4 202-450 also showed an upward turn at low q range indicative of aggregation and was not processed any further (Figure 8).

Figure 8. Scattering profile of IRF4 202-450. An upward turn is seen in the low q range (encircled in black)

We will attempt to send these samples for SAXS data collection in different buffers with different salts as well as reducing agents.
Crystallization trials of IRF4 202-450:

Crystallization trials of IRF4 202-450 were attempted using the Gryphon crystallization robot. The protein could be readily concentrated to ~ 10mg/mL and factorial screens were set up using commercially available crystallization screens from Midwest Center for Structural Genomics. Most of the trays showed clear crystallization drops & no sign of any crystals.

We reason that the concentration of the protein was in the under saturated portion of a crystallization phase diagram (Figure. 10). We will increase the protein concentration to ~20mg/mL or even higher (limited only by visualization of precipitation) and retry crystallization of IRF4 202-450. At higher concentration the protein should enter the supersaturated and induce crystal formation.

4.03 Future directions:

Structural information about the putative linker is pivotal in understanding the activation of IRF4. Since we obtained diffracting crystals for IRF4ΔNC with 17bp DNA we will focus our efforts in getting a better data set even though we have shown that the crystals are of a degradation product of IRF4ΔNC. Further, since we have purified IRF4 20-225 (construct with DBD along with some portion of the linker), we will perform fluorescence anisotropy with a 21bp F-DNA (composite site containing both IRF4 and PU.1 binding sites) to assess the effect of the linker on the DNA binding ability of IRF4. In addition, we will perform circular dichroism experiments on IRF4 202-450 to get secondary structure
information about the linker region and compare the results obtained with circular
dichroism spectra of IRF4IAD1 & IRF4IAD2. We will also perform SAXS studies on
IRF4 20-201 & IRF4 202-450 in different buffers. Crystallization trials will also be
attempted for IRF4 20-201 with DNA as well as for IRF4 202-450.

Although, crystallization experiments can be quite challenging, the information obtained
from circular dichroism and fluorescence anisotropy experiments proposed here we will be
able to obtain a more complete picture of the putative linker and its role in the activation
process of IRF4.
Conclusion and future directions

IRF4 is a unique member of the IRF family of transcription factors. It regulates transcription of various genes by forming ternary complexes with different transcription factors, presumably through its IAD. Our structural studies of IRF4IAD1 highlighted several salient features that make IRF4 unique in the IRF family. The N-terminal helix $\alpha_1$ - involved in forming a helix bundle - is quite flexible and may provide flexibility to the full-length protein. This helix bundle is stabilized by hydrophobic interactions of residues L413 from helix $\alpha_4$ and L368 from helix $\alpha_3$ with additional interactions provided by A246 of helix $\alpha_1$. The surface electrostatics of IRF4IAD1, especially around the helix bundle is more hydrophobic than the helix bundles of IRF3 & IRF5. Lastly, we identified loop $\beta_9$-$\beta_{10}$ that accommodates K399 in a pocket of positively charged residues. The flexibility of this loop is a unique feature of IRF4IAD1 that may enable its interaction with phosphorylated PU.1. SAXS studies of IRF4IAD2 revealed that the AIR is most likely folded and probably interacts with a region in the linker region. If the AIR were interacting with the DBD as previously proposed, we would expect these residues to be unfolded & completely extended. The AIR also lacks phosphorylation sites corresponding to those of IRF3 and IRF5 and the sequence alignment of this region is also quite low. Thus, the AIR appears to quite different in IRF4 and may provide additional exclusivity to IRF4 in its regulation process.
Our fluorescence anisotropy binding studies identified the first 20 residues unique to IRF4 and not the AIR to be important in decreasing the binding affinity of full-length IRF4 to DNA. These residues most likely impede the recognition helix from making specific hydrogen bond interactions with nucleic acid bases in the major groove of DNA. We show through analytical ultracentrifugation (AUC) studies of full-length IRF4 that the protein occurs as a single species in solution with a sedimentation coefficient of ~3.4S indicating that the protein is not globular but extended. Furthermore, our SAXS studies of the full-length protein corroborate with this result and show no dramatic differences in the overall shape of the protein in the presence or absence of the AIR. SAXS studies also suggest that the putative linker is most likely a folded domain. This result is in direct contrast to what it would be expected if the previously proposed mechanism of inactive-close and active-open states and conformational changes therein were accurate. Porod exponent obtained from our SAXS studies of IRF4wt of ~2.2 is indicative of appreciable degree of flexibility in the molecule. Removal of the N-terminal 20 residues or the AIR results in systematic decrease in the flexibility in IRF4 with an increase in Porod exponent to 3.3 for IRF4ΔNC. Moreover, there are subtle yet visually perceivable differences in the overall shape of IRF4 in the presence and absence of the first 20 residues or the AIR.

Taken together, we have come up with a different mechanism of activation for IRF4 than what was previously suggested shown in the schematic below. We propose that full-length IRF4 occurs as an extended and a flexible molecule in solution, which can bind DNA. There is no auto-inhibited closed state for the protein, but the N-terminal 20 residues of IRF4 decreases its binding affinity to DNA. The binding event is most likely transient and
hence not effective. Interaction of IRF4 with a partner protein like phosphorylated PU.1 ensures stable complex formation and prevents IRF4 from falling off the DNA. Following the ternary complex formation transcriptional regulation of the effector gene ensues.

**Figure 1**: Schematic representation of the new proposed mechanism of activation of IRF4
There are several questions that still remain unanswered about the process of activation of IRF4, for example: is the off rate of the binding event reduced during the complex formation? What is the sequence of binding events during the assembly of the ternary complex? A more rigorous examination of the ternary complex formation may be performed using Surface Plasmon Resonance (SPR) & Isothermal Calorimetric (ITC) techniques. More importantly, structural details of the putative linker region needs to be determined using a combination of biophysical techniques like AUC, SAXS & X-ray crystallography. Circular dichroism techniques may be used to gain some insight into the secondary structure of AIR as well as the putative linker. Furthermore, structural studies of full-length IRF4 in conjunction with its various binding partners, although challenging, seems like a viable future direction. It will provide molecular details of different surfaces of interaction between IRF4 & its binding partners. We could then understand the peculiarity of assembly of the hetero-complexes of IRF4 with its partner and in turn shed light on the promiscuity of this transcription factor.

A detailed mechanistic picture of the activation process of IRF4 may be exploited in designing therapeutic end products for diseases modulated by IRF4 like multiple myeloma, certain auto-immune disease and in cardiac hypertrophy.
Appendix

A. X-ray crystallography:

X-ray crystallography is an experimental technique that employs X-rays of wavelength in Angstrom scale (~10^{-10} m) such that atomic level information of the scattering protein molecules can be determined. X-ray crystallography utilizes crystals made of regularly spaced lattices of protein molecules. The electron clouds of individual atoms scatter the X-rays and thus a diffraction pattern is obtained. The three-dimensional structure of the protein under investigation can be reconstructed using from the diffraction data and additional phase information.

Figure 1. Schematic representation of diffraction of a crystal & the collected diffraction pattern. (Ugleingo, 2013)
**A.01 X-ray crystallography basics:**

**Unit cell:**

A unit cell is the simplest repeating unit composed of all the structural and symmetry information to reproduce the crystal lattice by translation along the cell edge.

**Asymmetric unit:**

An asymmetric unit is the smallest unit composed of all the structural information to reproduce the unit cell by symmetry operations.

**Figure 2.** Schematic representation of an asymmetric unit, a unit cell & their arrangement in a crystal. Image from (Shuchismita Dutta)
Bravais lattices:

Translating the unit cell contents in 3-D creates lattice systems called the Bravais lattice system that are 14 in number (Figure 3). A combination of rotations and translations results in a distinct pattern of symmetry elements that form the space groups each of which belongs to a Bravais lattice system.

Figure 3. Bravais lattice system (picture)
Structure determination:

To determine the 3-D crystal structure of any protein information regarding the distribution of the electron density that is related to the atomic positions in the unit cells needs to be deciphered from the diffraction data. The electronic density function is given by,

\[ \rho(x,y,z) = \frac{1}{V} \sum_{hkl} F_{hkl} e^{-2\pi i (hx + ky + lz)} e^{i\phi_{hkl}} \]

where, \( F_{hkl} \) are structure factors (the quantity expressing both amplitude & \( \phi_{hkl} \) phase of the reflection corresponding to the diffracted beam in reciprocal space with indices \( hkl \) of real space \((x,y,z)\) coordinates that represents atomic positions) and \( V \) is the cell volume. To solve for \( \rho(x,y,z) \) information about the amplitude, position \( (hkl) \) and phase of the diffracted beam is necessary. The amplitude of structure factors and positions are easily obtained from individual reflections. Intensities of diffraction spots and \( F_{hkl} \) are related as follows;

\[ I^2 = F_{hkl} \]

But information regarding the phase of the diffracted beam \( (\phi_{hkl}) \) is lost. There are multiple ways to obtain the ‘phase information’. A Patterson function like the one shown in Figure 4 is used to obtain the phase information.
Figure 4. Patterson auto-correlation function (Putnam et al, 2007)
Multiple/Single Isomorphous replacement (MIR/SIR):
Diffraction data is collected for native and heavy atom derivatives of crystals & the positions of the heavy atoms are identified based on the difference in intensity between the data sets from an isomorphous Patterson map. Hence, structure may be solved.

Multiple/Single wavelength Anomalous Diffraction (MAD/SAD):
The difference in anomalous scattering intensity inherent to elements with higher orbitals naturally present in protein molecules (Eg: Sulfur) or artificially incorporated into protein molecules (Eg: Selenium) is used to determine the positions of these anomalous scatters from anomalous difference Patterson maps. The information is then used to determine the structure of the protein under investigation.

Molecular replacement (MR):
Initial phase information is calculated from a model protein with appreciable sequence similarity to the protein currently under investigation. The model protein is repositioned onto the current diffraction data with as good an agreement as possible. Calculated phases are combined with the experimental data carefully so that minimal bias from the model protein is imposed onto the structure of the new protein.
A.03 Refinement:

Refinement is a process of improving the electron density map after the initial solution for the electron density function is obtained by adjusting the positions of atoms to minimize the discrepancies between the observed and calculated amplitudes using stereochemistry and electric and chemical bond restraints. There are several cycles of refinement that are run to ultimately generate acceptable atomic models based on the original diffraction data.

A.04 X-ray crystallography glossary – important terms (Rhodes, 2000):

For data collection statistics:

\[ R_{\text{merge}} = \frac{\Sigma_h \Sigma_i |I_i - \langle I\rangle|}{\Sigma_h \Sigma_i I_i} \]

\( R_{\text{merge}} \) is the measure of agreement between multiple non-symmetry related measurements of the same reflections and is given by;

\( I_i \) is the \( i \)th intensity measurement of reflection \( h \), and \( \langle I\rangle \) is the average intensity from multiple observations. Acceptable values for \( R_{\text{merge}} \) are frequently below 10%.

Redundancy:

Redundancy is given by:

\[ \text{(Number of measured reflections) / (Number of unique reflections)} \]

Higher redundancy values (>3) are usually acceptable.
I/\sigma(I):

Higher signal to noise ratio depicted by I/\sigma(I) (> 2 in the highest resolution shell) is highly recommended.

For refinement:

R_{factor}:

It is a measurement of the agreement of the crystallographic model obtained from the calculated intensities and those from the original diffraction data for individual reflections. R_{factor} is given by;

\[ R_{\text{factor}} = \frac{\sum |F_{\text{obs}}| - |F_{\text{calc}}|}{\sum |F_{\text{obs}}|} \]

Final R_{factor} in the range of 0-0.3 are acceptable and a low R_{factor} value is indicative of good model quality.

R_{free}:

R_{free} is calculated similar to the way that R_{factor} is calculated using a small set of randomly selected reflections (test set) that is not used during the refinement process. Frequently R_{free} is only slightly greater than R_{factor} and an acceptable difference between the two values is not more than 6%. This strategy of using R_{free} values in addition to R_{factor} helps in minimizing bias in the refinement process that may result in incorrect calculation of models in subsequent refinement runs.
TLS Motion Determination (TLSMD) (Merrit, 2006):

This is an additional strategy in the refinement process that may be used during refinement runs. A crystal structure is analyzed for evidence of flexibility within the domain or between domains etc. Individual chains of the structure are partitioned into segments modeled as rigid bodies. The segments are evaluated for TLS (Translation/Libration/Screw) vibrational motion. Multiple trials of partitioning the chains is scored against the atomic displacement parameters ("B-factor values").

B. Small Angle X-ray scattering (SAXS) (Putnam et al, 2007):

Although, X-ray crystallography provides unparalleled information at atomic resolution about proteins there are several pitfalls to the technique including intense sample preparation, requirement for crystals and possibility that the solution state of the protein is different from that in a crystal even though protein crystals have appreciable solvent content. SAXS becomes a very useful technique in such cases where obtaining diffractable crystals is a serious problem. SAXS can provide low-resolution models of proteins and frequently aid in designing protein constructs that may be amenable to crystallization. Further the modest sample preparation and data collection specifications makes SAXS a highly attractive technique in general and specifically when solution state studies of proteins need to be performed. It is also worth noting that SAXS is an extremely useful technique to monitor gross conformational changes, large unstructured regions and
determine the overall size of the protein more accurately than the frequently employed gel filtration studies.

*B.01 SAXS basics:*

Unlike crystallography where a complex Patterson function that assimilates directional inter-atomic distances (See Figure 4 in previous section) is used, the pair-distribution function utilized in SAXS studies resolves equivalent distances within a scattering unit without any information retained of the directionality (Figure 5). SAXS measurements generate scattering profiles that are plotted with scattering intensity \( I(q) \) as the Y-axis and momentum transfer \( q = 4\pi\sin\theta/\lambda \), where \( 2\theta \) is the scattering angle ) as X-axis. SAXS is a

![Pair distribution function](image)

**Figure 5.** Pair distribution function (Putnam *et al*, 2007)
difference technique and hence the protein sample has to be in the exact buffer, which is used to obtain a scattering profile for buffer alone. This is achieved by either dialyzing the protein extensively against the same buffer or performing gel-filtration runs prior to data collection.

B.02 SAXS glossary – important terms:

Radius of gyration (R_g):

R_g, distribution of the mass of a macromolecule around its centre of gravity, is the square root of the averaged squared distance of each scatterer from the centre of the macromolecule. It may be described in the equation;

\[ \ln I(q) = \ln I(0) - \frac{R_g^2}{3} * q^2 \]

Guinier analysis plot of \( \ln I(q) \) vs \( q^2 \) is a straight line at low \( q \) and is used to estimate \( R_g \). A non-linear dependence of \( \ln I(q) \) vs \( q^2 \) is indicative of presence of aggregation. SAXS data is collected at multiple different concentrations above 1mg/mL to determine if there is any aggregation in the sample that can have deleterious effects to further analysis of the data.

I(0):

I(0) or intensity measured at \( q=0 \) is determined by extrapolation of Guinier analysis plot. It is an important parameter to determine the molecular weight of the protein sample & functions as another parameter to check the quality of the collected data.
Kratky plot:

The Kratky plot is an important tool to estimate the folding of protein samples. It is a plot of $q^2I(q)$ as a function of $q$. For folded domains the plot appears like a bell shaped parabola while elongated and unstructured macromolecules appear to plateau as a function of the extent of unfoldedness (Figure 6).

![Kratky plot](image)

**Figure 6.** Standard Kratky plot for proteins of different shapes (Putnam et al, 2007)

Pair-distribution function $P(r)$:

$P(r)$ function describes a distribution of all equivalent paired sets of distances within macromolecules. $P(r)$ function is described by;

$$I(q) = \int p(r) * \frac{\sin (q.r)}{q.r} \, dr$$
where \( r \) is a measure of distance. Slight variations in the function are indicative of conformational changes in the macromolecule. Globular molecules have a \( P(r) \) with a single peak while multidomain proteins have multiple peaks and more elongated proteins can have multiple peaks with long tail at larger \( r \) values (Figure 7).

**Figure 7.** \( P(r) \) function of proteins of different shapes (Putnam *et al*, 2007)

### B.03 Solution structure modeling:

*Ab-initio* shape reconstruction is frequently carried out using SAXS data. Using the \( P(r) \) distribution function and information generated from the scattering profile low-resolution SAXS envelopes of proteins are generated. Multiple different programs may be used to do so including DAMMIN or GASBOR and results from either compared with each other. Further, multiple runs of the same program is performed which is essentially simulated annealing using a bead-model in DAMMIN and chain-like ensemble of dummy residues in
GASBOR that best fits the scattering data. The SAXS envelopes generated from these multiple runs are averaged using SUPCOMB suite of programs to generate unbiased non-over fitted models of proteins in solution. If atomic structure of the protein under investigation is known then theoretical SAXS profile from the atomic structure of the protein may be compared with the experimental SAXS profile of the protein to investigate any differences between the crystal & solution structures. If atomic structure of only individual domains of full-length protein is known then each may be fitted into the SAXS envelope. Further, using programs like BUNCH, CRY SOL, CORAL and CREDO missing loops or domains may be built into or adjoining the atomic structure using SAXS scattering profile as a guide.

C. **Analytical Ultracentrifugation (AUC) (Ralston, 1993):**

Analytical ultracentrifugation is yet another biophysical technique to access the shape and size of a macromolecules. Variants of the technique can be used to investigate conformational changes in proteins, oligomerization state of proteins as well as visualize complexation with various ligands.

The basic principles involved in analytical ultracentrifugation experiment are described hereunder. A solute suspended in a solvent that begins to move along a radial path when centrifugal force is applied (as is experienced by the solute if it is centrifuged at high speed) experiences three forces during its sedimentation process, namely gravitational force, buoyant force and frictional force. For a time scale of $10^{-6}$ seconds the three forces balance each other out since frictional force is a retarding force. A term, sedimentation
coefficient (s) has thus been defined based on the above considerations that depend on the physical properties of the particle. It has units of seconds & is given by,

\[ s = \frac{v}{\omega^2 r} \]

or velocity (v) of the particle per unit gravitational acceleration (\( \omega^2 r \)). It depends on the molar weight of the sedimenting particle and is inversely proportional to fractional coefficient.

A schematic representation of a sedimenting boundary (of sedimenting particles) is shown in Figure 8.

**Figure 8.** Double sector centerpiece and a representative sedimentation profile (Ralston, 1993)
In a general experiment, the sample is sedimented at high speed and at every couple of seconds the reference and the sample sectors are imaged. A schematic diagram of the Beckman Optima XL-A analytical ultracentrifuge is shown in Figure 9.

**Figure 9.** Optical system of Beckman Optima XL-A Analytical ultracentrifuge. (Ralston, 1993)
The optical system of an AUC instrument measures differences in absorbance or light scattering between the reference sector and the sample sector. There are two optical systems, one that measures interference (based on the light scattered by the samples) from the sample and one that measures absorbance from absorbing molecules from the sample (say protein molecules that absorb at 280nm). There are two major experiments that are performed – sedimentation velocity experiments and sedimentation equilibrium experiments.

In sedimentation velocity experiments, a homogenous solution is subjected to a high angular velocity in the instrument that leads to depletion of sample molecules from the meniscus and formation of a sharp boundary that separates the depleting sample molecules from the homogenous concentration of the sedimenting solute. Integrating and processing the data to generate a size-distribution (c(s)) vs sedimentation coefficient curve & hence obtain the sedimentation coefficient of the solute particle.

In a sedimentation equilibrium experiment, a small volume of a homogenous sample solution is subjected to lower angular velocity in the instrument. As the sedimentation process proceeds with increased concentration of solute at the bottom of the cell, diffusion acts as an opposing force. The concentration of the solute increases exponentially towards the bottom of the cell once equilibrium is reached between the two forces. This procedure enables one to determine the molecular weight of the sedimenting particle by measuring concentrations at different points in the equilibrium curve.
D. Fluorescence Anisotropy (Invitrogen, 2006):

Fluorescence anisotropy is used to determine the dissociation constants of a binding event say,

\[ P + L \rightleftharpoons PL \]

where, \( P \) is the protein, \( L \) is the ligand & \( PL \) is the protein-ligand complex. Plane polarized light is used to excite fluorescent ligands. Small fluorescent ligands tend to tumble too quickly and depolarize during their fluorescence lifetimes (time between excitation and emission). When bound to a macromolecular protein the fluorescent ligand rotates less rapidly and tends to emit polarized light. The intensity of the emitted light is measured in both parallel and perpendicular planes (Figure 10). Fluorescence anisotropy is given by

\[
\text{Anisotropy} = \frac{I_{II} - I_{\text{per}}}{I_{II} - 2I_{\text{per}}}
\]

where, \( I_{II} \) is the intensity measured in the parallel (vertical) direction to the excitation plane polarized light and \( I_{\text{per}} \) is the intensity measured in the perpendicular (horizontal) direction to the excitation plane polarized light.
Figure 10. **a.** Schematic representation of small and large complexes and their differences in anisotropy. **b.** Schematic representation of general detection protocol of a fluorescence polarization experiment (Invitrogen, 2006)
Binding isotherms are generated by titrating increasing concentration of the macromolecule against a constant concentration of the ligand till no change in anisotropy is observed or in other words saturation is achieved. Further, using a one site binding equation for a ligand-macromolecule interaction that obeys this relationship a dissociation constant ($K_d$) is calculated. One site binding equation is given as,

$$Y = \frac{B_{max} \times X}{K_d + X}$$

where, $B_{max}$ is the maximal specific binding in units of anisotropy or percent ligand bound, $X$ is the concentration of the ligand and $Y$ is increasing units of anisotropy or percent ligand bound.

**E. Tobacco Etch Virus- protease (TEV) purification:**

TEV-protease plasmid with an N-terminal His-tag and S219V mutation that prevents the autolysis of the protein improving its yield by almost 100-fold was a kind gift from Dr. Takeharu Kawano (Kapust *et al*, 2001). The plasmid is Ampicillin resistant & an overnight culture was grown overnight with ~100μg/L final concentration of Ampicillin. Next day, 10-20ml of the overnight culture was added per litre of culture. The protein was grown at 37°C till OD ~0.6 was reached. The culture was induced with 1mM IPTG & the temperature was reduced to 30°C and the cells were harvested after about 4-6hrs of shaking, pelleted and frozen at -80°C. The cells were thawed and run on a Ni-NTA column equilibrated in 50mM sodium phosphate dibasic pH 8.0, 200mM NaCl, 10% glycerol, 5mM imidazole and eluted with 250mM imidazole. The eluted protein was concentrated
and further purified on a gel filtration column pre-equilibrated in 25mM sodium phosphate pH 7.5, 100mM NaCl, 10% glycerol. The fractions with TEV-protease (molecular weight ~25kDa) were pooled, concentrated to 1.5mg/mL and flash frozen and stored at -80°C. About 45-50mg of total protein was obtained per litre.
References:


picture W. http://www.cmmp.uel.ac.uk/~ijf/3c25/bravais.gif


Read more: http://www.ehow.com/how_8725889_cite-pymol.html.


Shuchismita Dutta RKG, and Catherine L. Lawson. Looking at Structures: Introduction to Biological Assemblies and the PDB Archive.


VITA

SOUMYA GOVINDA REMESH

Place of birth: Thrishur, Kerala, India
Date of birth: April 28, 1985
Citizenship: India

EDUCATION:

Doctor of Philosophy

Structural studies of Interferon Regulatory factor 4:
A molecular perspective of its regulatory mechanism

Department of Physiology and Biophysics
Virginia Commonwealth University (VCU)

Advisor: Dr. Carlos Escalante Assistant Professor
Department of Physiology and Biophysics VCU

2010-2014
**Master of Science**  
**2008-2010**  
*Insights into the catalytic mechanism of retro-aldol cleavage of β hydroxy amino acids by L-threonine aldolase*  
Department of Medicinal Chemistry  
Virginia Commonwealth University (VCU)  
*Advisor: Dr. Martin Safo*  
*Associate Professor*  
*Department of Medicinal Chemistry*  
*VCU*

**Bachelor of Pharmacy**  
**2004-2008**  
University of Mumbai  
Mumbai Education Trust’s  
Institute of Pharmacy, Mumbai-400050  
India

**EXPERIENCE:**

**Graduate Research Assistant**  
**2010-2014**  
Department of Physiology and Biophysics  
Virginia Commonwealth University (VCU)  
Advisor: Dr. Carlos Escalante (<escalante@vcu.edu>)

**Research Assistant**  
**2009-2010**  
Department of Physiology and Biophysics  
Virginia Commonwealth University (VCU)  
Advisor: Dr. Martin Safo (<msafo@vcu.edu>)

**RESEARCH SKILLS:**

1. **Crystallography**  
- Crystallization: Trained in use of Gryphon and Minstrel for setting up and monitoring of crystallization trials  
- Data Collection: In-house and at Synchrotron beam lines  
- Data Processing, Structure Solution and Refinement: Dtrek, hkl2000, Phenix, ccp4, Coot, Pymol, Refmac and Chimera

2. **Other Biophysical Techniques**  
- Analytical Ultracentrifugation: Analysis using Sedfit, Sedphat programs  
- Small-Angle X-ray Scattering: Data Collection at beamlines  
- Analysis of SAXS data & generation of ab-initio models: Primus and ATSAS software suite
3. **Cloning and Protein Purification**
- Design and cloning of different protein constructs to be used for protein purification and other biochemical experiments, site-directed mutagenesis
- Expression and purification of different proteins: Experience primarily in bacterial expression system with working knowledge of mammalian expression system
- Trained in using Akta FPLC purification system for purification of proteins.

4. **Biochemical Experiments**
- Use of SDS-Page gel electrophoresis and UV spectroscopic studies (analysis using Origin)
- Western blot experiments
- Circular Dichroism

**TRAINING:**

1. Trained to use the beamline X-4C and X-25 at National Synchrotron Light Source at Brookhaven National laboratory for X-ray crystallographic experiments.
2. Trained to use the beamline X9 at NSLS at BNL to perform small angle X-ray scattering (SAXS) experiments.
3. Attended SAXS workshops
   - UC Berkeley (Oct, 2013): 4th annual SIBYLS bioSAXS workshop, Advance Light Source (ALS) at Lawrence Berkeley National Laboratory, Berkeley, CA

**PUBLICATIONS:**

Based on work done during Master of Science in Pharmaceutical Sciences

1. *di Salvo Martino, Remesh SG et al, On the catalytic mechanism and stereospecificity of L-threonine aldolase; FEBS J, 2013*  
   (Second author with equal contribution, structural work was part of my Master’s thesis project)
2. *A second untitled manuscript is in preparation* (First author)

Based on work done during my PhD in Biomedical Sciences

1. *Manuscripts in preparation to publish my graduate work*
2. *Abstract Title: I woRk diFfently 4- Mechanism of activation of auto-inhibited IRF4*
   Abstract was published in Journal of Federation of American Societies for Experimental Biology (FASEB journal, April 2014, Vol. 28, Supplemental LB282)