Chapter 2

 Contributions of Intrinsic Disorder in Functionality of Structural Domains
2.1 INTRODUCTION

2.1.1 Prevalence of disorder across functional categories

Experimentalists have shown that functions of few proteins depend on the native disordered state. In addition, the prediction of intrinsic disorder in proteins has also been revolutionary in changing views relating to structure-function paradigm. In a classical study by Dunker and Obradovic’s group (121-123), Swiss-Prot terms were used to identify enriched functions, biological processes, ligand and Post Translational Modifications (PTMs) which are frequently associated with intrinsic disorder (121). Since experimentally identified IDRs are relatively less in number, prediction driven annotations are used to find significantly over or under-represented annotational, functional or structural categories.

The functions which have very frequently been associated with ID are cell differentiation, cell division, protein transport, mRNA processing, apoptosis and regulation (123) (also see Fig 1.1). With biological processes like apoptosis, cell division, signal transduction and regulation of cell cycle, IDPs are bound to make impact in cancer research. Other than this, there are several disease associated proteins which exhibit intrinsic disorder and are being further studied from therapeutics point of view (124-130).

2.1.2 Role of intrinsic disorder in interaction and assembly

IDRs have a role in interaction of proteins with multiple partners which is an advantage for functional promiscuity (131;132), regulation (133-139), cross-talk between networks (140-142) and emergence of novel functions (143;144). The interaction through binding depends on interface and structural properties of the disordered region and its partner, which may or may not be unfolded. IDPs are known to primarily interact through molecular recognition features, hence playing an important role in PPI networks (145-147). Relation between disorder content and its relation with degree of interaction suggests that IDRs facilitate interactions. Though importance of IDR in interaction is now established, but the role of location of disorder across the protein is not much explored. One of the studies (148) focussing on finding abundance of intrinsic disorder across four
protein regions, viz. N-terminal, C-terminal, interdomain and domain region showed that IDRs were more likely to be found in the terminal regions (N and C-terminal) followed by the inter-domain and were least found to coexists with domains. There are other studies like the one focused on a fungal hydrophobin which emphasizes on the fact that specific disordered structural motifs can influence protein assembly (149). But does IDR provide advantage to certain motifs or domains is still unanswered to a great extent. In the current work, I have attempted to answer this question using coiled coils as the basis of my study.

2.1.3 Structural motifs favoring disorder

Protein functionality is often associated with structural domains or motifs. It might sound counter to the conventional belief, but few structural domains have been found to be highly associated with intrinsic disorder (unstructured). Initially, disordered proteins were believed to be ones with random coil-like structure. But with use of various techniques including Circular Dichroism (CD) significant amount of secondary structures coexisting with intrinsic disorder was brought to light.

As functional role of IDRs became evident, the significance of short disordered stretches interrupting classical secondary structures were also seen in PDB database with approximately 27% residues in α-helical conformation, 12% in β-strands, and approximately 48% in irregular conformations (26).

In 2007, study by Vucetic et al. (122) showed structural domains associated with ordered as well as disordered structural domains. The domains which were most highly associated with disorders consists of zinc-finger, transit peptides, homeobox, signal anchors, collagen etc. with coiled coil being most significantly associated with disorder. Since, the prevalence of disorder in coiled coils is proven; it becomes interesting to study the relevance of disorder in this structural motif.

2.1.4 Coiled coil motifs

The coiled-coil motifs comprises of two or more a-helices twisted around one another (150;151). It is one of the most frequently found interacting motifs in nature
These versatile structural motifs harbor signature heptad repeats of a seven residue pattern denoted ‘‘abcdefg’’ with residues \( a, d \) forming the hydrophobic core and \( e, g \) usually charged or polar (Fig 2.1). Structurally and functionally, coiled coils play important roles in transcription control, association and organization of complexes, chromosomal and cell cycle maintenance etc. (154;155). They are also vital for the fusion of viral and cellular membrane and hence are being widely studied in HIV (156-159).

Figure 2.1: Helical wheel diagram of coiled coil motif. The residues a-g represents one helix and a’-g’ the other. Within the heptad repeat, a-a’ and d-d’ exhibit hydrophobic interactions while e-g’ and e’-g exhibits the ionic interactions [Adopted from Mason et al., 2004 (155)]

With such a versatile structural motif, it becomes important to know if disordered coiled coils are associated with particular functions or cellular compartments. Coiled coil proteins in eukaryotes are double of that found in prokaryotes (160). Since bacterial proteome has limited number of coiled coil proteins, the study was conducted on human proteome. Apart from eukaryotes, coiled coils are also crucial for prokaryotes especially in host–pathogen interactions specifically because most of the coiled coil proteins are secretory (160).
Mycobacterium tuberculosis is known to survive and replicate in macrophages dodging the innate immune system of the host (161-163). I further explored the host proteins which might be contributing in establishment and survival of M. tuberculosis in humans and found that Coronin, which is crucial for pathogen survival in microphages, belongs to the disordered coiled coil category and it has been presented as case study.

2.2 MATERIAL & METHODS

2.2.1 Human coiled coil data set

A set of 1968 human proteins annotated as bearing coiled-coil region were retrieved from the UniProt database (164) along with their Gene Feature Files (GFF). This was used as the data and knowledgebase for the study. The human variation and their associated disease information was retrieved from “Human polymorphisms and disease mutations” release 08-Feb-2011 (UniProt).

2.2.2 Disorder prediction and gene ontology analysis

Similar to the above stated approach, intrinsic disorder prediction was performed on 1968 coiled coil bearing proteins and intrinsically disordered regions (score >= 0.5 and length of consecutive amino acid >=30) were identified using IUPred (63) and verified using DISPro (69). The disordered coiled coils were found by mapping the coordinates of IDR with coiled coils.

883 proteins were found to be disordered by both algorithms, IUPred as well as DISPro (Fig. 2.2). These proteins were categorized into 2 categories - one which had proteins harboring “disordered coiled coils” (DisCCs) and other set of proteins with “disorder outside coiled coil” (DOCCs). Proteins which lacked potential IDRs belonged to the ordered coiled coil set (OCCs) (Fig 2.2).

Differential Gene ontology (GO) enrichment analysis of cellular components, processes and molecular functions in the three sets was done using GOEAST (165). It uses hyper-geometric test as a default method to identify statistically enriched or
depleted ontologies. Multi-GOEAST (165) functionality was used to compare the three sets of coiled coils.

![Workflow depicting the categorization of proteins as disordered coiled coils (DisCCs), ordered coiled coils (OCCs) and disorder outside coiled coil domains (DOCCs)](image)

**Figure 2.2: Workflow depicting the categorization of proteins as disordered coiled coils (DisCCs), ordered coiled coils (OCCs) and disorder outside coiled coil domains (DOCCs)**

### 2.2.3 Amino acid composition

Disordered and ordered coiled-coil regions were extracted from protein sequences based on their UniProt coordinates. DisCC regions comprised of the predicted disordered coiled coils and the ordered coiled-coil region form DOCCs and OCCs.
were combined to form a set of regions called All Ordered Coiled coils (AOCCs).
Amino acid composition of these two set of region was calculated using the freely accessible web server PROFEAT (166;167). The mean value of the percentage composition for each amino acid was retrieved.

### 2.2.4 Oligomerization states and low complexity regions enrichment

Oligomerization states for the coiled-coil regions of disordered coiled coil and ordered coiled coil was predicted using Multicoil (168). The predictions were run on the two set of regions – DisCC and AOCC (collective set of DOCCs and OCCs). The program predicts probability of oligomerization states for a coiled-coil region based on training set of dimeric and trimeric coiled coils.

Low complexity regions (LCRs) in the coiled-coil regions of DisCC and AOCCs sets were predicting using Seg (ftp://ftp.ncbi.nih.gov/pub/seg/seg). It implements the method of Wootton & Federhen (169) for identifying compositionally biased regions in the amino acid sequence. Default parameters of Seg were used to predict LCR. The length of LCRs was calculated for each coiled-coil region for both the sets.

Wilcoxon rank sum test is performed on the oligomerization state probability as well as proportion (percentage length) of LCRs in the coiled coils. Two separate tests were done- the first test is based on a p-value for the null hypothesis that probability of a coiled coil to attain dimeric and trimeric states of oligomerization in DisCCs and AOCCs is the same, the second test was done with the null hypothesis that the proportion of LCRs in coiled coils in ordered and disordered coiled coil (DisCCs and AOCCs) is the same. The tests were performed using Wilcox.test function of R which implements Wilcoxon rank sum test on provided vectors.

### 2.3 RESULT & DISCUSSION

#### 2.3.1 Disordered human coiled coils

Intrinsic disorder is high in low complex sequences and repeats (170;171). There are specialized domains which transit from disordered state in due course. One such structural domain which is highly associated with intrinsic disorder is the coiled coil
domain. Since the mycobacterium proteome consists only of four proteins bearing the coiled coil domain, a study was carried out on human proteome which comprises of hundreds of such proteins. Usually, coiled coils are frequently disordered as monomers and become folded upon association and formation of quaternary structure.

Primarily there were two questions I wanted to address: a) how frequently coiled coils are intrinsically disordered and b) how are they differentially enriched across gene ontologies (functions, cellular components and biological processes). I categorized the coiled coil proteins into three sets—disordered coiled coils (DisCCs), ordered coiled coils (OCCs) and disorder outside coiled coil (DOCCs) and performed in silico analysis to identify significantly and differentially associated ontologies with the three categories. The disorder predictions and mapping of disordered and coiled coil coordinates gave three sets—598 proteins harboring “disordered coiled coils” (DisCCs), 285 proteins with “disorder outside coiled coil” (DOCCs) and 437 proteins which lacked potential IDRs categorized as "ordered coiled coil" set (OCCs). These three sets of proteins were used to get GO enrichment data. 488 out of 598 DisCCs, 375 out of 437 OCCs and 256 out of 285 DOCCs had associated gene ontology information (Fig 2.2) and were further analyzed for differential enrichment.

2.3.2 Gene ontology enrichments across coiled coil categories

Cellular component analysis was done using multiple set GO comparison utility of GOEAST. OCCs were found to be enriched in structural components of the extracellular space (Fig 2.3) including the fibrinogen complex and laminin complex along with membrane trafficking complexes like SNARE complexes and exocysts. On the other hand, DisCCs to be exclusively over-represented in proteins involved in components associated with motility and mechanical integrity of the cell including actin filament, lamellipodium, cell junction, macromolecule complexes, ciliary rootlet, nucleolus etc. (Fig. 2.3) Also, DOCCs were rich in cellular components including kinetochore–microtubule complexes, ruffle and midbody (Fig. 2.3). This set of coiled coil proteins showed high association with biological processes including regulation of calcium ion transport. The details of the gene ontology analysis in respect to specific class of proteins like motor and skeletal proteins are discussed in
Chapter 2

the published manuscript (172). A list of cellular component ontologies enriched in the three categories of coiled coil proteins along with the p-value is listed in ST.1

Figure 2.3: Enrichment of different categories of coiled coil proteins in various cellular components as found by gene ontology analysis. The disordered coiled coils (DisCCs) are shown in red, ordered coiled coils (OCCs) in green and disorder outside coiled coils (DOCCs) in blue.

2.3.3 Enrichment across skeletal and motor proteins

Skeletal proteins form the mechanical basis of the cytoskeleton. The eukaryotic filamentous cytoskeleton (including microtubules, actin containing microfilaments and intermediate filaments) are often associated with motility and cell division (173).

The lamellipodium and actin cytoskeletal proteins were found to be significantly enriched in DisCCs. The latter are known to play an important role in cell shape determination, motility, cytokinesis and interactions (173). Similarly, lamellipodium is essential for motility, membrane domain organization, substrate adhesion and phagocytosis (174). The DisCC proteins were seen to be involved in mechanical integrity of cell, adhesion, motility regulation, recruitment and organization.

In contrast, coiled coils associated with the laminin and fibrinogen complex were enriched in OCCs. These proteins comprise of laminin subunit alpha 1, 2, 3 and beta 3 and fibrinogen like protein 1 and fibroleukin.
Cytoskeletal motor proteins are comprised of three major classes—kinesin, myosin, and dynein. Among these, myosins were found to be enriched in DisCCs. Most of these proteins are involved in microfilament motor activities and regulation. They play an integral role in muscular contraction (175) which might depend on flexible coiled coils. Contrary to kinesins and myosins, dynein complexes were predicted to lack significant length of IDRs and hence were categorized as OCCs.

### 2.3.4 Amino acid level enrichments and its role in oligomerization

Coiled coils have lower complexity than globular proteins. I further explored the difference in sequence-based complexity by filtering the low complexity regions (LCRs) in the ordered and disordered coiled coils. As expected, the abundance of LCRs in DisCCs was higher than those in AOCCs, hence, justifying the role of the disordered region in the flexibility of coiled coils.

![Amino acid composition of DisCCs and AOCCs](image)

**Figure 2.4:** Amino acid composition of DisCCs and AOCCs. The X-axis shows the 20 amino acids and the Y-axis shows the percentage composition of the amino acids in the coiled-coil region. The error bar depicts the standard error of mean.

The difference in the distribution probability scores of dimerization between the two sets was found to be significant with the median of probability of dimer forming coiled coil in DisCC being higher for DisCCs(0.95) as compared to AOCCs (0.78) (supplementary figure SF. 1). It is known that, the heptad repeats have preferred positions for hydrophobic (a and d) and charged (e and g) amino acids. My...
observation further suggests that amongst the hydrophobic residues isoleucine (I) is enriched in DisCCs and alanine (A) and leucine (L) are favored by AOCCs (Fig 2.4). Leucine and isoleucine which have same composition and size, because of their packing geometry, play a major role in oligomerization of coiled coils and studies have shown that presence of isoleucine at both a and d positions is preferred in a trimeric oligomeric state (176;177). In my study, marginal high probability of DisCCS to form trimers over AOCCs was observed which can be because of the high frequency of isoleucines in DisCCs (p-value of the Wilcoxon test = 0.01345). The charged amino acids occupy e and g positions of the heptad repeats. Earlier Kohn et al. (178) observed that increasing the frequency of glutamic acid (E) destabilizes the helical conformation of the dimeric coiled coils, hence pushing conformational state towards random coiled coil. Additionally, the group also showed that the protonation of glutamic acid increases the stability of the coiled coils and as the number of protonated glutamic acid residues increases, the dimer shifts from a less stable to a more stable form. Since DisCCs were found to be enriched in glutamic acids, the dimer formation of such a coiled coil might be triggered by protonation of the residue which in turn is caused by the lowering of pH.

2.3.5 Role in diseases and host-pathogen interaction

Coiled coils are also crucial for host–pathogen interactions. *M. tuberculosis* is known to survive and replicate in humans macrophages evading our innate immune system. I found that Coronin 1 (also known as p57 or TACO) harbors disordered coiled coils. It is known to mediate survival of mycobacteria in the host cell (179) and is associated with phagosomes containing live bacterium. This protein aids in persistence of the pathogen. The pathogen fails to survive in Kupffer cells where coronin is not expressed (163;180). Despite having adverse effects on the host itself, it has been retained by the cells possibly because its role in regulation of cytoskeleton, membrane-trafficking and calcium signaling. Coiled coil motifs are involved in formation of a stable trimer which is essential for its functionality.
Figure 2.5: Structural and functional motifs across coronin 1A protein. The IDR overlaps with the actin-binding and cytoskeletal interacting region of the coiled coil.

As shown in Fig 2.5 the coiled coil region is known to interact with cytoskeleton and also has an actin binding region, actually bears the disordered segment (181-183). This region might be providing flexibility to the protein to perform multiple functions. My observation suggests that experimentally studying point mutations in the disordered region of the coiled coil (424–439) in Coronin 1A (UniProt id P31146) may provide insight into dynamic functioning of this coiled coil protein.

I further explored the mutation in the proteins belonging to the DisCCs set. Some of these were found to be associated with diseases like amyloidosis type 8, Joubert syndrome type 8, congenital fibrosis of extraocular muscles type 1, hepatocellular carcinoma etc. (Table 2.1). Intrinsically disordered proteins have been associated with amyloid related disorders (75) and likewise mutations in fibrinogen a protein (FIBA) are associated with amyloidosis type 8 were found. The mutations reported are in residue numbers 545 and 574 and exhibit substitution of disorder favoring (Glu and Arg) to order favoring residues (Val and Leu respectively) (184). This finding suggests that a reduction in the flexibility of the coiled coil may result in altering its functionality and hence might be associated with the diseased state. The molecular basis of this disease is still under exploration and the current finding can help in
getting new insight into the mechanism of action of this protein in normal and
diseased state.

Table 2.1: List of proteins bearing non-synonymous mutation in the disordered
coiled coils associated with human diseases

<table>
<thead>
<tr>
<th>UniProt-id</th>
<th>Gene</th>
<th>Mutation</th>
<th>Disease</th>
<th>MIM-id</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1AK53</td>
<td>ESPN</td>
<td>R774Q</td>
<td>Deafness autosomal dominant without vestibular involvement (DFNAWVI)</td>
<td>[MIM:606351]</td>
</tr>
<tr>
<td>P02671</td>
<td>FIBA</td>
<td>E545V</td>
<td>Amyloidosis type 8 (AMYL8)</td>
<td>[MIM:105200]</td>
</tr>
<tr>
<td>P02671</td>
<td>FIBA</td>
<td>R573L</td>
<td>Amyloidosis type 8 (AMYL8)</td>
<td>[MIM:105200]</td>
</tr>
<tr>
<td>Q3SXY8</td>
<td>ARL13B</td>
<td>R200C</td>
<td>Joubert syndrome type 8 (JBTS8)</td>
<td>[MIM:612291]</td>
</tr>
<tr>
<td>Q7Z4S6</td>
<td>KIF21A</td>
<td>M947R</td>
<td>Congenital fibrosis of extraocular muscles type 1 (CFEOM1)</td>
<td>[MIM:135700]</td>
</tr>
<tr>
<td>Q7Z4S6</td>
<td>KIF21A</td>
<td>M947V</td>
<td>Congenital fibrosis of extraocular muscles type 1 (CFEOM1)</td>
<td>[MIM:135700]</td>
</tr>
<tr>
<td>Q7Z4S6</td>
<td>KIF21A</td>
<td>M947T</td>
<td>Congenital fibrosis of extraocular muscles type 1 (CFEOM1)</td>
<td>[MIM:135700]</td>
</tr>
<tr>
<td>Q7Z4S6</td>
<td>KIF21A</td>
<td>R954Q</td>
<td>Congenital fibrosis of extraocular muscles type 1 (CFEOM1)</td>
<td>[MIM:135700]</td>
</tr>
<tr>
<td>Q7Z4S6</td>
<td>KIF21A</td>
<td>R954W</td>
<td>Congenital fibrosis of extraocular muscles type 1 (CFEOM1)</td>
<td>[MIM:135700]</td>
</tr>
<tr>
<td>Q9ULD2</td>
<td>MTUS1</td>
<td>Q1201R</td>
<td>Hepatocellular carcinoma (HCC)</td>
<td>[MIM:114550]</td>
</tr>
</tbody>
</table>

2.4 CONCLUSION

Coiled coil are ideal structures for peptide designing and are also structurally important motif. They are involved in a wide array of biological functions and should be computationally and therapeutically explored. There are several hypotheses regarding stability and folding of coiled coils. Though the heptad repeats guide the interactions between helices, but specificity of interaction may be influenced by other sequence factors. Here, I explored one such factor which is location of intrinsic
disorder across coiled coil bearing proteins of the human proteome. With this data it becomes evident that coexistence of IDRs with structural domains can have direct implications on the preference in oligomerization states.

Another interesting aspect is the biological implications of ordered and disordered coiled coil. Enrichment analysis shows that localization of IDR in a coiled coil protein may influence the subcellular localization and biological functionality of the protein.