Chapter 5

SUMMARY, CONCLUSION
& FUTURE DIRECTIONS
5.1 OVERVIEW

Globally, India and China bear the maximum burden of tuberculosis (WHO report, 2012). *Mycobacterium tuberculosis* - the causative agent of this disease has been one of the most widely studied pathogen, but the way it survives and grows resistance against the defense mechanism of the host and to therapeutics is an area of active research. In the current study, the intrinsically disordered proteins are explored for their potentiality as drug targets. I studied IDPs, which are an interesting class of proteins known to carry out biological functions in natively disordered state, from a drug target perspective. IDPs are rich in highly charged and low hydrophobic regions (12) and can either be completely disordered or might bear natively disordered regions. Almost 30-50% of the eukaryotic proteins are predicted to be IDPs. The functionality and mode of interactions of these proteins have been of particular interest because of its promiscuity. Studies suggest a correlation between intrinsic disorder, protein function and expression (3;16;229). Proteins associated with several diseases including cancer, cardiovascular and neurodegenerative diseases are particularly enriched in IDPs (75;136). IDR's are better suited for inhibition by small molecules as they exhibit low affinity interactions and highly specific binding (see Chapter I). Few drug targets having an intrinsically disordered region at their binding interface have been reported (230). These regions exhibit disordered-to-ordered transitions during the interaction with their substrates. Computational methods to predict disordered regions in proteins have advanced rapidly. IDR's have the potential to be pursued as novel therapeutic targets.

5.2 SUMMARY & CONCLUSIONS

Intrinsic structural disorder in proteins does not mean functional disorder; rather the disordered regions/proteins are crucial in the functioning of biological systems. The mechanism of coupled binding and folding for protein-protein interactions are crucial and I tried to implement *in silico* approaches to identifying essential proteins that could be promising targets for inhibition by small molecules. Targeting the
intrinsically disordered region has the potential to enhance the discovery for new drug like molecules.

Thus, the role of intrinsically disordered region in the target protein was studied. IDPs which could be potential drug target against *M. tuberculosis* were identified, followed by predicting the important disordered residues which can act as potential inhibition sites for therapeutics.

5.2.1 Gaining insight into biological function of intrinsically disordered essential proteins

5.2.1.1 Prevalence of disorder across structural motifs

In an early work (122), motifs highly associated with intrinsic disorder were identified and it was found that coiled coils are most frequently associated with intrinsic disorder (11;122). Prevalence of intrinsically disordered coiled coils is established. The study was aimed at gaining insight about the relevance of intrinsic disorder across the coiled coil proteins. In the current work, relevance of these prevalent motifs (coiled coil) was analyzed.

5.2.1.2 Categorization of coiled coil proteins

A sequence level analysis of human coiled coils was done to find out if this is universally true for all coiled coils. 1968 human proteins annotated as coiled coil were analyzed for presence of disordered regions. Two parallel methods of disorder prediction were implemented – namely IUPred and DISpro. As stated earlier, IUPred predicts the disorderness score based on inter-residue interaction matrix. While, DISpro is a neural network based method which works on a background set of curated disordered proteins. Disorder prediction and coordinate mapping of coiled coils and IDRs resulted in categorization of 598 proteins harboring disordered coiled coils as DisCCs and 285 proteins having disordered regions outside coiled coils as DOCCs. I also found 444 coiled coil proteins which were devoid of disordered region (OCCs).
5.2.1.3 Enrichment of disordered coiled coil providing functional insight

GOEAST, which is a powerful tool to study the enrichment patterns of gene ontology (GO) from a given set of genes/proteins, was used to identify enrichment of these three categories of coiled coils (DisCCs, OCCs and DOCCs) across various cellular components. Performing the cellular component analysis using multiple set GO comparison utility of GOEAST, I found that OCCs are enriched in structural components of the extracellular space including fibrinogen complex and laminin complex along with membrane trafficking complexes and exocysts.

On the contrary, DisCCs were found to be exclusively over-represented in proteins involved in actin filament, lamellipodium, cell junction, macromolecule complexes, ciliary rootlet, perinuclear region of cytoplasm and nucleolus. Most of the cellular components, listed prior, contribute to motility and mechanical integrity of the cell. DOCCs were found to be enriched in cellular components including kinetochore-microtubule complexes, ruffle and midbody (172).

5.2.1.4 Comparison of amino acid abundance and oligomerization states

As expected, DisCCs are particularly enriched in charged residues including glutamic acid (E) and arginine (R), hydrophilic residues viz. glycine (G) along with aromatic residues like phenylalanine (F) and tyrosine (Y). On the contrary, all ordered coiled coils were found to be enriched in hydrophobic residues like alanine (A), leucine (L) and methionine (M). This analysis was conducted using PROFEAT (261). Multicoil (168) was used to predict dimeric and trimeric states of the coiled-coil region for DisCCs and AOCCs separately. The difference in the distribution probability scores of dimerization between the two sets was found to be significant with a p-value << 1e -10 as compared to the less significant difference in probability of the trimeric state p-value =0.01345 (Wilcoxon test).
5.2.1.5 Coronin: a case study

Coiled coils are crucial for host–pathogen interactions. Mycobacterium tuberculosis is known to survive and replicate. Coronin 1 (also known as p57 or TACO), one of the DisCC proteins, is known to mediate survival of *M. tuberculosis* in host cell (179) and is associated with phagosomes of live bacterium. These proteins help survival of the pathogen in phagosome by escaping association of phagosome with lysosomes (179). But despite aiding the pathogen in surviving in the macrophages, it has been retained by the cells possibly because the protein is involved in multifunctional regulation of cytoskeleton, membrane trafficking and calcium signaling. It has been shown to utilize the coiled coil region to form stable trimer which is essential for its functionality. The proteins bearing disordered region are capable of interacting with multiple partners. It is suggested 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 (172).

5.2.2 Comparative analysis of XDR, MDR and H37Rv strain IDPs and identification of change of binding pocket

5.2.2.1 Genome assembly and annotation of ten acquired Extrapulmonary & Pulmonary samples

Extrapulmonary form of tuberculosis has remained less studied because of the complexity of sample collection. Current efforts are being made to get an insight into the variation between pulmonary (PTB) and extrapulmonary TB (EPTB). The current work aims at identifying variations across EPTB and PTB sample. As already stated intrinsically disordered proteins are highly evolving, it would be interesting to find if intrinsically disordered proteins contribute to the difference in pathogenicity and survival strategies of drug resistant strains of EPTB and PTB. Short reads of one PTB and nine EPTB samples were acquired. These reads were assembled using *denovo* assembly pipeline – SOAPDenovo (195). The insert size and kmer for the assembly was tested on one of the samples by varying the values iteratively and observing its effect on N50 and no. of
scaffolds formed. 45 kmer and 1000 insert size were used. *M. tuberculosis* H37Rv was taken as the reference to arrange scaffolds in optimal orientation.

### 5.2.2.2 Genomic level variations across samples

Once the ten genomes were drafted, the variable regions were identified. The PPE/PE family of proteins which also happen to be highly disordered proteins (221;222) along with the phi-Rv (phage) proteins were found to be most varying.

### 5.2.2.3 Variation observed in intrinsically disordered proteins

Alongside the genomic comparisons based on the draft genomes, the unused contigs and their products were also compared. Results indicated presence of PlcD harbouring contigs which was present across all our samples except the ethambutol resistant samples and H37Rv. As already reported in the literature this region is prone to the insertion site of IS6110 and RvD2 deletions (262). The PlcD protein is flanked by Glycosyl transferase, Molybdopterin oxidoreductase, cutinase, MmpL family protein, hypothetical proteins etc most of whose functions are yet to be reported. I also observed deletion of the IDR in the PTB sample. The IDRs were conserved in the rest EPTB samples and had highest similarity to *M. africanum* (MAF00010) and *M. CDC1551* (MT179/9) counterparts. The IDR which is located at the C-terminal also has predicted potential binding sites which could be of functional relevance. Further work is being carried out in our laboratory to identify DNA and RNA level variations between EPTB and PTB.

### 5.2.3 Identification of intrinsically disordered proteins which could be potential drug targets for *Mycobacterium tuberculosis*.

#### 5.2.3.1 Disoderness of the *M. tuberculosis* proteome

Earlier studies have shown that an average of 33% eukaryotic, 4.2% of eubacterial and 4.2% of archaean proteins consists of IDPs (66;222). For identification of IDPs in *M. tuberculosis*, two independent tools were used. The first tool used was IU Pred
which predicts intrinsically disordered regions (IDRs) based on pairwise interaction energy between neighboring amino acids. The second tool used was DISPro which takes into account the evolutionary information profiles, predicted secondary structure and solvent accessibility of residues. This approach ensured that the set of proteins is likely to be free of any biased prediction of IDPs. Those proteins which harbored an intrinsically disordered stretch of consecutive 30 or more amino acids were considered as IDPs. IUPred predicted 586 and DISPro predicted 276 proteins to be IDPs. An intersection of 222 proteins were the high confidence IDPs predicted by both the methods.

5.2.3.2 Essentiality of IDPs

Earlier in 2003, Sassetti et al. had identified the proteins which are required for optimal growth of *M. tuberculosis* by performing Transposon Site-Hybridization (TraSH) (205). The method had identified 614 proteins to be essential for *M. tuberculosis*. 46 essential IDPs were segregated by comparing the two protein sets.

5.2.3.3 Essential IDPs with reported interactions

IDPs are known to be hub proteins in protein networks. It has been observed that the assembly of protein complexes is enabled and to an extent promoted by disordered region of the protein (147). 21 proteins which had reported interactions in *M. tuberculosis* were selected from the set of essential IDPs.

5.2.3.4 Proteins with no homologs in Humans and microbiome

Targeting a proteome that has a homologous protein in the host and gut and oral microbiome can give rise to adverse drug reactions (209). In order to avoid these proteins which had significant similarity with either host proteins or the consolidated proteome of microbiome were filtered out of the target dataset. This lead to a final set of 13 proteins which were proposed to be potential drug targets in *M. tuberculosis* (222).
5.2.4 Identification of key residues and binding pockets which can be targeted using standard bioinformatics approaches

5.2.4.1 Generation of complete model for GlmUm\textsubscript{Mt}b

On the basis of available literature and present knowledge about the shortlisted proteins (BioAssay, Literature and patent searches), one of the proteins -GlmU was prioritized for further study. LOMETS server was used to predict the complete structure of the GlmU. The meta-server takes sequence of the protein in fasta format and uses a set of available tools to generate the tertiary structure. It generates 3D models by collecting consensus target-to-template alignments from 9 locally-installed threading programs (FUGUE, HHsearch, PAINT, PPA-I, PPA-II, PROSPECT2, SAM-T02, SPARKS, SP3). The structure generated by SAM T20 method was taken for further analysis. The initial structure was crosschecked using various available approaches like Ramachandran plot and WhatIf (247). Proline puckering was observed and was rectified by minimizing the modelled structure.

5.2.4.2 Identifying highly interacting disordered residues

While the work was in progress, Bioassay result of GlmU inhibitors became freely available via PubChem (254) (AID 1376). This experiment was against the GlmU\textsubscript{Mt}b protein and hence used in current study. The study reports actives against GlmU\textsubscript{Mt}b, but the region of binding was not known. In order to identify which could be the potential binding sites for these compounds, the blind docking approach was implemented using Autodock. The trimeric form of GlmU was taken as the target protein. The resultant top energy docking poses were studied in details to conclude most interacting residues. This yielded a list of key residues which could be potential binding regions on the target’s disordered surface. These highly interacting residues identified from this approach are ILE457, TRP460, ARG463.
5.2.4.3 Highly modulating residues in the disordered tail

Molecular dynamics simulations were run on the trimeric GlmUmtb structure in order to identify the highly modulating residues and transient structures in the disordered region. A 10ns long simulation using Gromacs at normal pH showed appearance of two main transient structures a) a semi helical structure which appeared around the highly interacting residues (457-463 aa) & b) Beta-hairpin like structure towards distal end of the tail (478-493 aa). The earlier identified highly interacting residues attained the semi-helical structure, while the distal B-hairpin structures can act as potential nucleation site for the folding of the disordered region.

5.2.5 Identification of leads that interact with the critical amino acids in the selected targets.

As earlier proposed (260), models were developed for predicting inhibitory activity (IC50) of chemical compounds against DHDPS protein using QSAR and docking techniques. These models were trained on 84 diverse compounds (GlmU inhibitors) which had a 3D structure and reported IC50 value in the earlier mentioned PubChem BioAssay (AID 1376). These inhibitors were docked in the active site of the C-terminal domain of GlmU protein (2O16) (243) using the AutoDock. In order to improve the performance hybrid models were developed using various types of descriptors and achieved high correlation of 0.83/0.70 (r/r2) between predicted and actual IC50 (260). Some molecular descriptors used in this study correlated very well with pIC50. Chemical libraries were screened using these models and 40 potential GlmU inhibitors were predicted (260). These inhibitors can be used to develop drugs against *M. tuberculosis*.

5.3 FUTURE DIRECTIONS

Intrinsically disordered proteins are highly abundant in nature. Many of these IDPs have been associated with human diseases such as neurodegenerative diseases, cancer, amyloidosis etc. These proteins are different from the misfolded proteins as they
natively reside in the disordered state and yet exhibit a wide array of functions. Despite important role of IDPs in interactions and biological processes, their role as therapeutic drug targets against pathogen is not widely explored. This study provides niche for carrying out mutational and targeting experiments to combat tuberculosis.

There are instances where molecules have been designed to inhibit interaction of IDR (MDM2-p53, c-Myc-Max complex), but the potentials of similar approach at a cellular level is yet to be explored.

Apart from the listed intrinsically disordered drug targets, there is also a need to focus on cross-strain variations in IDPs as these proteins are known to be of highly evolving as compared to their ordered counterpart. It would be interesting to see which functional classes of proteins tend show highest degree of variation across pathogenic samples. The mentioned pipeline for identification of intrinsically disordered drug targets can also be implemented across proteome of various pathogens which have ample amount of publically available data.