Significant advancements in genome sequencing techniques and other high-throughput initiatives have resulted in the availability of complete sequences of genomes of a large number of organisms, which provide an opportunity to study detailed biological information encoded therein. Identification of functional roles of proteins can aid in comprehension of various cellular activities in an organism, which is traditionally achieved using techniques pertaining to the field of molecular biology, protein chemistry and macromolecular crystallography. The established experimental methods for protein structure and function determination, although accurate and resourceful, are laborious and time consuming. Computational analyses of sequences of gene products and exploration of evolutionary relationships can give clues on protein structure and/or function with reasonable accuracy which can be used to direct experimental studies on proteins of interest, effectively. Moreover, with growing volumes of data, there has been a growing disparity in the number of well-characterized and uncharacterized proteins, further necessitating the use of computational methods for investigating evolutionary and structure-function relationships.

The remarkable progress made in the development of computational techniques (Chapter 1) has immensely contributed to the state-of-the-art biological sequence analysis and recognition of protein structure and function in a reliable manner. These methods have largely influenced the exploration of protein sequence-structure-function space.

One of the relevant applications of computational approaches is in the understanding of functional make-up of human pathogens, their complex interplay with the host and implications in pathogenesis. In this thesis, sensitive profile-based search procedures have been utilized to address various aspects in the context of three pathogens- *Mycobacterium tuberculosis*, *Plasmodium falciparum* and *Trypanosoma brucei*, which are causative agents of potentially life-threatening diseases. The existing drugs approved for the diseases, although of immense value in controlling the disease, have several shortcomings, the most important of them being the emergence of drug resistance that render the current treatment regimens futile. Thus, the identification of practicable targets and new drugs or new combination therapies become an important necessity. Analyses on structural and functional repertoire of proteins encoded in the pathogenic genomes can provide means for rational identification of therapeutic intervention strategies.

This thesis begins with the computational analyses of proteins encoded in *M. tuberculosis* genome. *M. tuberculosis* is a primary aetiological agent of tuberculosis in humans, and is
responsible for an estimated 1.5 million deaths every year. The complete genome of the pathogen was sequenced and made available more than a decade ago, which has been valuable in determination of functional roles of its gene products. Yet, functions of many *M. tuberculosis* proteins remain unknown. Computational prediction of protein function is an on-going process based on ever growing information made available in public databases as well as the introduction of powerful homology recognition techniques. Hence, a continuous refinement is essential to make the most of the sequence data, ensuring its accuracy and relevance. With the use of multiple sequence and structural profile-based search procedures, an enhanced structural and functional characterization of *M. tuberculosis* proteins, totalling to 95% of the genome was achieved (Chapter 2). Following are the key findings.

- Domain definitions were obtained for a total of 3566 of 4018 proteins. Amino acid residue coverage of >70% was achieved for 2295 proteins which constitute more than half of the proteome.
- Domain assignments were newly identified for 244 proteins with domain-unassigned regions. Structure prediction for these proteins corroborated all the remote homology relationships recognized using profile-based methods, enhancing the reliability of the predictions.
- Comparison on domain compositions of proteins between *M. tuberculosis* and human host, revealed presence of pathogen-specific domains that are not homologous to proteins in human. Such proteins in *M. tuberculosis* are mainly virulence factors involved in host-pathogen interactions such as immune-dominance and aiding entry and survival in human host macrophages, hence forming attractive targets for drug discovery.
- Putative structural and functional information for proteins with no recognizable domains were inferred by means of fold recognition and an iterative profile-based search against sequence database.
- Attributing putative structures and functions to 955 conserved hypothetical proteins in *M. tuberculosis*, 137 of which are reportedly essential to the pathogen, provide a basis to re-investigate their involvement in pathogenesis and survival in the host.
- Proteins with no detectable homologues were recognized as *M. tuberculosis* H37Rv-specific, which can serve as promising drug targets.
- An attempt was made to identify porin-like proteins in *M. tuberculosis*, considering MspA porin from *M. smegmatis* as a template. The difficulty in recognition of putative porins in *M. tuberculosis* is indicative of novel outer membrane channel proteins, not characterized yet, or high representation of ion-channels, symporters and transporters to compensate for the functional role of porins. In addition, MspA-like proteins were not readily recognized in other slow-growing mycobacterial pathogens that are known to infect human host, apart from *M. tuberculosis*. This indicates probable acquisition of physiological adaptations, i.e. absence of porins, to confer drug-resistance, in the course of their co-evolution with human hosts.
Evolutionary relationships recognized between sequence (Pfam) and structural (SCOP) families aided in association of potential structures and/or functions for 55 uncharacterized Pfam domains recognized in *M. tuberculosis*. Such associations deliver useful insights into the structure and function of a protein housing the uncharacterized domain.

The functional inferences drawn for *M. tuberculosis* proteins based on the predictions can provide valuable basis for experimental endeavours in understanding mechanisms of pathogenesis and can significantly impact anti-tubercular drug discovery programmes.

An interesting outcome benefitted from the exercise of exploring relationships between Pfam and SCOP families, was the identification of evolutionary relationship between a Pfam domain of unknown function DUF2652 and class III nucleotidyl cyclases. A detailed investigation was undertaken to assess this relationship (Chapter 3). Nucleotidyl cyclases synthesize cyclic nucleotides which are critical second messengers in signalling pathways. The DUF2652 family predominantly comprises of bacterial proteins belonging to three lineages- Actinobacteria, Bacteroidetes and Proteobacteria. Thus, recognition of evolutionary relationship between these bacterial proteins and nucleotide cyclases is of particular interest due to the indispensability of cyclic nucleotides in regulation of varied biological activities in bacteria. Use of fold recognition program suggested presence of nucleotide cyclase-characteristic topological motif (βaaββαβ) in all the members of the DUF2652 family. Detailed analyses on structural and functional features of the uncharacterized set of bacterial proteins corresponding to 50 bacterial genomes, using profile-based alignments, revealed presence of key features typical of nucleotidyl cyclases, including metal-binding aspartates, substrate-specifying residues and transition-state stabilizing residues. Depending on the features, 20 proteins of Actinobacteria lineage, predominantly mycobacteria, of unknown structure and function were identified as putative nucleotide cyclases, 23 proteins of Bacteroidetes lineage were associated with guanylyl cyclases, while 8 uncharacterized proteins of Proteobacteria were recognized as nucleotide cyclase-like proteins (7 adenylyl and one guanylyl cyclase). Sequence similarity-based clustering of the predicted nucleotide cyclase-like proteins with established nucleotide cyclases indicated the apparent evolutionarily distinctness of the subfamily of class III nucleotidyl cyclases predicted. Furthermore, analysis of evolutionarily conserved gene clusters of the predicted nucleotide cyclase-like proteins indicated functional associations that support the predictions on their participation in cellular signalling events. The inferences made can be experimentally investigated further to ascertain the involvement of the uncharacterized bacterial proteins in signalling pathways, which can help in understanding the pathobiology of pathogenic species of interest.

The next objective was the recognition of biologically relevant protein-protein interactions across *M. tuberculosis* and human host (Chapter 4). *M. tuberculosis* is well known for its ability to successfully co-evolve with human host in terms of establishing infection, survival and persistence. The current knowledge on the mechanisms of host invasion, immune evasion and
persistence in the host environment can be attributed, and is limited, to the experimental studies pursued by numerous groups. Chapter 4 presents an approach for computational identification of biologically feasible protein-protein interactions across *M. tuberculosis* and human host. The approach utilizes crystal structures of intra-organism protein-protein complexes which are transient in nature. Identification of homologues of host and pathogen proteins in the database of known protein-protein interactions, formed the initial step, followed by identification of conserved interfacial patch and integration of information on tissue-specific expression of human proteins and subcellular localization of human and *M. tuberculosis* proteins. In addition, appropriate filters were used to extract biologically feasible host-pathogen protein-protein interactions. This resulted in recognition of 386 interactions potentially mediated by 59 *M. tuberculosis* proteins and 90 human proteins. A predominance of host-pathogen interactions (193 protein-protein interactions) brought about by *M. tuberculosis* proteins participating in cell wall processes, was observed, which is in concurrence with the experimental studies on immuno-modulatory activities brought about by such proteins. These set of mycobacterial proteins were predicted to interact with diverse set of host proteins such as those involved in ubiquitin conjugation pathways, metabolic pathways, signalling pathways, regulation of cell proliferation, transport, apoptosis and autophagy. The predictions have the potential to complement experimental observations at the molecular level. Details on couple of interesting cases are presented in the chapter, one of which is the probable mechanism of immune evasion adopted by *M. tuberculosis* to inhibit lysozyme activity in macrophages, and second is the mechanism of nutrient uptake from host. The set of *M. tuberculosis* proteins predicted to mediate interactions with host proteins have the potential to warrant an experimental follow-up on probable mechanisms of pathogenesis and also serve as attractive targets for chemotherapeutic interventions.

One of the major shortcomings with current treatment regimen for tuberculosis is the emergence of multidrug (MDR) and extensively drug-resistant (XDR) strains that render first-line and second-line drug treatments futile. This entails a need to explore target space in *M. tuberculosis* as well as explore the potential of existing drugs for repurposing against tuberculosis. A drug repurposing strategy i.e. exploring within-target-family selectivity of small molecules, has been implemented (Chapter 5) to contribute towards time and cost-saving anti-tubercular drug development efforts. With the use of profile-based search procedures, evolutionary relationships between targets (other than proteins of *M. tuberculosis*) of FDA-approved drugs and *M. tuberculosis* proteins were investigated. A key filter to exclude drugs capable of acting on human proteins substantially reduced the chances of obtaining anti-targets. Thus, total of 130 FDA-approved drugs were recognized that can be repurposed against 78 *M. tuberculosis* proteins, belonging to the functional categories- intermediary metabolism and respiration, information pathways, cell wall and cell processes and lipid metabolism. The catalogue of structure and function of *M. tuberculosis* proteins and their involvement in host-pathogen protein-protein interactions compiled from chapters 2 and 4 served as a guiding tool to explore the functional importance of targets identified. Many of the potential targets identified have been experimentally shown to be essential for growth and survival of the
pathogen earlier, thus gaining importance in terms of pharmaceutical relevance. Polypharmacological drugs or drugs capable of acting of multiple targets were also identified (92 drugs) in the study. These drugs have the potential to stand tolerance against development of drug resistance in the pathogen. Comparative sequence and structure-based analysis of *M. tuberculosis* proteins homologous to known targets yielded credible inferences on putative binding sites of FDA-approved drugs in potential targets. Instances where information on binding sites could not be readily inferred from known targets, potentially druggable sites have been predicted. Comparison with earlier experimental studies that report anti-tubercular potential of several approved drugs enhanced the credibility of 74 of 130 FDA-approved drugs that can be readily prioritized for clinical studies. An additional exercise was pursued to identify prospective anti-tubercular agents by means of structural comparison between ChEMBL compounds and 130 FDA-approved drugs. Only those compounds were retained that showed considerably high structural similarity with approved drugs. Such compounds with minor changes in terms of physicochemical properties provide a basis for exploration of compounds that may exhibit higher affinities to bind to *M. tuberculosis* targets. The set of approved drugs recognized as repurpose-able candidates against tuberculosis, in concert with the structurally similar compounds, can significantly impact anti-tubercular drug development and drug discovery.

The next part of the thesis focuses on *Plasmodium falciparum*, an obligate intracellular protozoan parasite responsible for malaria. The parasite genome features unusual characteristics including abundance of low complexity regions and pronounced sequence divergence that render protein structure and function recognition difficult. The parasite also manifests remarkable plasticity in its metabolic organization throughout its developmental stages in two hosts-human and mosquito; thus obtaining an exhaustive list of metabolic proteins in the parasite gains importance. Considering the utility of multiple sensitive profile-based search approaches in enhanced annotation of *M. tuberculosis* genome, a similar exercise was employed to recognize potential metabolic proteins in *P. falciparum* (Chapter 6). A total of 172 metabolic proteins were identified as participants of 78 metabolic pathways, over and above 609 proteins known to participate in *P. falciparum* metabolism. Pathway holes, where evidence on metabolic step exists but the catalysing enzyme is not known, have also been addressed in the study, several of which have been suggested to play an important role in growth and development of the parasite during its intra-erythrocytic stages in human host.

A subsequent objective was the recognition *P. falciparum* proteins potentially capable of remodelling erythrocytes to suit their niche (Chapter 7). Exploitative mechanisms are brought about by the parasite to remodel erythrocytes for growth and survival during intra-erythrocytic stages of its life-cycle, the understanding of which is limited to experimental studies. To achieve physicochemically viable protein-protein interactions potentially mediated by proteins of human erythrocytes and *P. falciparum* proteins, a structure-influenced protocol, similar to the one demonstrated in Chapter 4, was employed. Information on subcellular localization and protein expression is crucial especially for parasites like *P. falciparum*, which reside in
heterogeneous environmental conditions at different stages in their lifecycle. Inclusion of such data aided in extraction of 208 biologically relevant protein-protein interactions potentially mediated by 59 *P. falciparum* proteins and 30 erythrocyte proteins. Host-parasite protein-protein interactions were predicted pertaining to several major strategies spanning intra-erythrocytic stages in *P. falciparum* pathogenesis including- gaining entry into the host erythrocytes (category: RBC invasion, protease), redirecting parasitic proteins to erythrocyte membrane (category: protein traffic), modulating erythrocyte machinery (category: rosette formation, putative adhesin, chaperone, kinase), evading immunity (category: immune evasion) and eventually egress (category: merozoite egress) to infect other uninfected erythrocytes. Elaborate means to analyse and evaluate the functional viability of a predicted interaction in terms of geometrical packing at the interfacial region, electrostatic complementarity of the interacting surfaces and interaction energies is also demonstrated. The protein-protein interactions, thus predicted between human erythrocytes and *P. falciparum*, have the potential to provide a useful basis in understanding probable mechanisms of pathogenesis, and indeed in pinning down attractive targets for antimalarial drug discovery.

The emergence of drug resistance against all known antimalarial agents, currently in use, necessitates discovery and development of either new antimalarial agents or unexplored combination of drugs that may not only reduce mortality and morbidity of malaria, but also reduce the risk of resistance to antimalarial drugs. In an attempt to contribute towards the same, Chapter 8 explores the established concept of within-target-family selectivity of small molecules to recognize antimalarial potential of the approved drugs. Eighty six FDA-approved drugs, predominantly constituted by antibacterial agents, were identified as feasible candidates for repurposing against 90 *P. falciparum* proteins. Most of the potential parasite targets identified are known to participate in housekeeping machinery, protein biosynthesis, metabolic pathways and cell growth and differentiation, and thus are pharmaceutically relevant. During intra-erythrocytic growth of *P. falciparum*, the parasite resides within the erythrocyte, within a protective encasing, known as parasitophorous vacuole. Hence a drug, intended to target a parasite protein residing in an organelle, must be sufficiently hydrophilic or hydrophobic to be able to permeate cell membranes and reach its site of activity. On the basis of lipophilicity of the drugs, a physical property determined experimentally, 57 of 86 FDA-approved drugs were recognized as feasible candidates for use against *P. falciparum* during the course of blood-stages of infection, which can be prioritized for antimalarial drug development programmes.

The final section of the thesis focuses on the protozoan parasite *Trypanosoma brucei*, a causative agent of African sleeping sickness (Chapter 9). This disease is endemic to sub-Saharan regions of Africa. Despite the availability of completely sequenced genome of *T. brucei*, structure and function for about 50% of the proteins encoded in the genome remain unknown. Absence of prophylactic chemotherapy and vaccine, compounded with emergence of drug-resistance renders anti-trypanosomal drug discovery challenging. Thus, considering the utility of frameworks established in earlier chapters for recognition of protein structure, function and drug-targets, similar steps were undertaken to understand functional repertoire of the parasite
and use drug repurposing methods to accelerate anti-trypanosomal drug discovery efforts. Structures and functions were reliably recognized for 70% of the gene products (5894) encoded in *T. brucei* genome, with the use of multiple profile-based search procedures, coupled with information on presence of transmembrane domains and signal peptide cleavage sites. Consequently, a total of 282 uncharacterized *T. brucei* proteins could be newly coined as potential metabolic proteins. Integration of information on stage-specific expression profiles with *Trypanosoma*-specific and *T. brucei*-specific proteins identified in the study, aided in pinning down potential attractive targets. Additionally, exploration of evolutionary relationships between targets of FDA-approved drugs and *T. brucei* proteins, 68 FDA-approved drugs were predicted as repurpose-able candidates against 42 potential *T. brucei* targets which primarily include proteins involved in regulatory processes and metabolism. Several targets predicted are reportedly essential in assisting the parasite to switch between differentiation forms (bloodstream and procyclic) in the course of its lifecycle. These targets are of high therapeutic relevance, hence the corresponding drug-target associations provide a useful resource for experimental endeavours.

In summary, this thesis presents computational analyses on three pathogenic genomes in terms of enhancing the understanding of functional repertoire of the pathogens, addressing metabolic pathway holes, exploring probable mechanisms of pathogenesis brought about by potential host-pathogen protein-protein interactions, and identifying feasible FDA-approved drug candidates to repurpose against the pathogens. The studies are pursued primarily by taking advantage of powerful homology-detection techniques and the ever-growing biological information made available in public databases. Indeed, the inferences drawn for the three pathogenic genomes serve an excellent resource for an experimental follow-up. The set of protocols presented in the thesis are highly generic in nature, as demonstrated for three pathogens, and can be utilized for genome-wide analyses on many other pathogens of interest.

The supplemental data associated with the chapters is provided in a compact disc attached with this thesis.