The availability of completely sequenced genomes for a number of organisms provides an opportunity to understand the molecular basis of physiology, metabolism, regulation and evolution of these organisms. Significant understanding of the complexity of organisms can be obtained from the functional characterization of repertoire of proteins encoded in their genomes. Computational approaches for recognition of function of proteins of unknown function encoded in genomes often rely on ability to detect well characterized homologues.

Homology searches based on pair-wise sequence comparisons can reliably detect homologues with sequence identity more than 30%. However, detecting homologues characterized by sequence identity below 30% is difficult using these methods. Distant homology relationship can be established using profiles or position specific scoring matrices, which encapsulate information about structurally and functionally conserved residues. These conserved residues imply high constraints at a particular amino acid residue site due to their involvement in structural stability, enzymatic activity, ligand binding, protein folding or protein–protein interactions. In addition, information on three dimensional structures of proteins also aid in detection of remote homologues, as tertiary structures of proteins are conserved better than the primary structures of proteins. The gross objective of the work reported in this thesis is to employ various sensitive remote homology detection methods to recognize relevant functional information of proteins encoded mainly in pathogenic organisms.

Since proteins do not work in isolation in a cell, it has become essential to understand the in vivo context of functions of proteins. For this purpose, it is essential to have an understanding of all molecules that interact with a particular protein. Thus, another major area of bioinformatics has been to integrate protein-protein interaction information to enable better understanding of context of functional events. Protein-protein interaction analysis for host-pathogen can lead to useful insight into mode of pathogenesis and subsequent consequences in host cell.

Chapters 2-6 of the thesis discuss the sequence and structural characteristics along with remote evolutionary relationships and functional implications of uncharacterized proteins encoded in genomes of following pathogens: Helicobacter pylori, Plasmodium falciparum and Leishmania donovani. The Chapters 6-8 discuss mainly various sequence, structural and functional aspects of protein kinases encoded in genomes of various prokaryotes and viruses.
Chapter 1 discusses background information and literature survey in the areas of homology detection and prediction of protein-protein interactions. The growth of genomic data and need for processing genomic data to infer context of various functional events have been highlighted. Different approaches to recognize functions of proteins (experimental as well as computational) have been discussed. Various experimental and computational approaches to detect/predict protein-protein interactions have been mentioned.

Chapter 2 discusses recognition of non-trivial remote homology relationships involving proteins of *Helicobacter pylori* and their implications for function recognition. *H. pylori* is microaerophilic, Gram negative bacterial pathogen. It colonizes human gastric mucosa and is a causative agent of gastroduodenal disease. The pathogen infects about 50% of the human population. It can lead to development of Mucosa-associated lymphoid tissue lymphoma. About 10% of the infected population develop gastric or duodenal ulcer and approximately 1% develop gastric cancer. *H. pylori* has been classified as class I carcinogen by WHO. Pathogen is characterized by type IV secretion system.

The complete genomic sequences of three widely studied strains including 26695, J99 and HPAG1 of *Helicobacter pylori* are available. According to the genome analysis, the number of predicted open reading frames in strain 26695, J99 and HPAG1 are 1590, 1495 and 1536 respectively. Out of predicted *H. pylori* proteins from 26695, J99 and HPAG1 strains, numbers of proteins with no functional domain assignments in Pfam database (Protein family database) are 453, 357 and 400 respectively. There are proteins in different strains of *H. pylori* genomes where one part of the protein is associated with at least one protein domain of known function and hence preliminary indication of their functions is available whereas rest of the region is not associated with any function. There are 772, 803 and 790 such segments in proteins from strains 26695, J99 and HPAG1 respectively with at least 45 residues with no functional assignment currently available. Sensitive remote homology detection methods have been employed to establish relationships for 294 amino acid sequences and results have been grouped into 4 categories. Results of homology detection have been further confirmed by studying conservation of amino acid residues which are important for functioning of the proteins concerned.

(i) Remote relationship has been established involving protein domain families for which no bonafide member is currently known in *H. pylori*. For example: DNA binding protein domain (Kor_B) has been assigned to a *H. pylori* protein at sequence identity of 20%.
Study involving secondary structure prediction and conservation of amino acid residues confirms the results of homology detection methods.

(ii) Remote relationship has been established involving *H. pylori* hypothetical proteins and protein domain families, for which paralogous members are present in *Helicobacter pylori*. For example, Cytochrome_C, an electron transfer protein domain could be associated with a *Helicobacter pylori* protein sequence which shows a sequence identity of 14% with sequences of bonafide cytochrome C.

(iii) “Missing” metabolic proteins of *H. pylori* have also been recognized. For example, Aspartoacylase (EC 3.5.1.15) catalyzes deacetylation of N-acetylaspartic acid to produce acetate and L-aspartate. This enzyme in aspartate metabolism pathway has not been reported so far from *H. pylori*. A remote evolutionary relationship between a *H. pylori* protein and Aspartoacylase domain has been established at sequence identity of 17% thus filling the gap in this metabolic pathway in the pathogen.

(iv) New functional assignments for domains in *H. pylori* sequences with prior assignment of domains for the rest of the sequences have been made. For example, DNA methylase domain has been assigned to C-terminal region of *H. pylori* protein which already had Helicase domain assigned to the N-terminal region of the protein.

All these information should open avenues for further probing by carrying out experiments which will impact the design of inhibitor against this pathogen and will result in better understanding of pathogenesis of this organism in human.

*Chapter 3* describes prediction of protein–protein interactions between *Helicobacter pylori* and the human host. A lack of information on protein-protein interactions at the host-pathogen interface is impeding the understanding of the pathogenesis process. A recently developed, homology search-based method to predict protein-protein interactions is applied to the gastric pathogen, *Helicobacter pylori* to predict the interactions between proteins of *H. pylori* and human proteins *in vitro*. Many of the predicted interactions could potentially occur between the pathogen and its human host during pathogenesis as we focused mainly on the *H. pylori* proteins that have a transmembrane region or are encoded in the pathogenic island and those which are known to be secreted into the human host.
Synopsis

By applying the homology search approach to protein-protein interaction databases DIP and iPfam, in vitro interactions for a total of 623 H. pylori proteins with 6559 human proteins could be predicted. The predicted interactions include 549 hypothetical proteins of as yet unknown function encoded in the H. pylori genome and 13 experimentally verified secreted proteins. A total of 833 interactions involving the extracellular domains of transmembrane proteins of H. pylori could be predicted.

Structural analysis of some of the examples reveals that the predicted interactions are consistent with the structural compatibility of binding partners. Various probable interactions with discernible biological relevance are discussed in this chapter. For example, interaction between CFTR protein (NP_000483) and multidrug resistance protein (HP1206) has been predicted. The structure of the CFTR intracellular domain is known in the homomeric form and consists of five AAA transport domains in tandem (PDB code 1XMI). Out of the five identical subunits, two subunits (the B chain and the E chain in the PDB structure) have been selected. The structure of multidrug resistance protein of the pathogen based on the B chain (sequence identity 32%) of the template has been modeled. This exercise suggests that interface residues in the model are congenial for interaction. This makes the structural complex feasible in in vitro conditions and suggests that the pathogen protein may compete for occupancy with the host protein.

Chapter 4 describes recognition of Plasmodium-specific protein domain families and their roles in Plasmodium falciparum life cycle. Malaria in humans is caused by the parasites of intracellular, eukaryotic protozoan of apicomplexan nature belonging to the genus Plasmodium. Out of five species of Plasmodium, namely, P. falciparum, P. ovale, P. vivax, P. malariae and P. knowlesi which infects human, P. falciparum causes lethal infection. P. falciparum proteins have diverged extensively during the course of evolution. Pathogen genome is rich in A+T composition which accounts for about 80% of the genome. Proteins encoded in Plasmodium genome are often larger than the homologous proteins from other organisms due to presence of low complexity regions.

Organism specific families are important as they play roles in peculiar life style of an organism. If the organism is a pathogen, then these family members may play roles in pathogenesis. Inhibiting these specific proteins is unlikely to interfere with host system as no homolog may be present in host. In the present work we identify Plasmodium specific protein families and their role in different stages of life cycle of the pathogen.
A total of 5086 amino acid sequences (full length sequences/fragments of proteins) show homology only with amino acid sequences from *Plasmodium* organisms and hence are *Plasmodium*-specific. These *Plasmodium*-specific amino acid sequences cluster into 106 *Plasmodium*-specific families (≥2 members per family). 14 *Plasmodium*-specific protein domain families with known physico-chemical properties are observed. These *Plasmodium*-specific protein domain families are involved in various important functions such as rosetting and sequestering of infected erythrocytes, binding to surface of host cell and invasion process in life cycle of pathogen. Also, 89 new *Plasmodium*-specific protein domain families have been recognized. Analysis of various aspects of members of *Plasmodium*-specific proteins domain families such as their potential to target apicoplast, protein-protein interaction, expression profile and domain organization has been performed to derive relevant information about function. New *Plasmodium* specific domain families for which no function can be associated could provide some insight into much diverged *Plasmodium* species. These proteins may play role in parasite-specific life style. Experimental work on these *Plasmodium*-specific proteins might fill the gaps of less understood physiology of this parasite.

**Chapter 5** presents genome-wide compilation of low complexity regions (LCR) in proteins. An in-depth analysis of the nature, structure, and functional role of the proteins containing low complexity regions in *Plasmodium falciparum*, was undertaken given the high prevalence of LCRs in the proteome of this organism. Low complexity regions and repeat patterns have been recognized in proteins encoded in 986 genomes (68 archaea, 896 prokaryotes and 22 eukaryotes).

Low complexity regions have been classified into following three categories:

a) Composition of LCRs: (i) LCRs can be stretches of homo amino acid residues (ii) LCRs can be stretches of more than one amino acid residue type.

b) Periodicity of amino acids in LCRs: Certain amino acid residues can be observed at certain specific periodicity in proteins.

c) Repeat patterns: Certain motif of amino acid residues are repeated in protein. 850 *Plasmodium falciparum* proteins are observed to have at least one repeat pattern where the repeating unit is at least 5 amino acid residues long.

Statistical analysis on single amino acid residue repeats indicate that occurrence of stretches of homo amino acid residues is not a random event. Studies on recognition of functions, protein-
protein interactions and organization of tethered domain(s) in proteins containing LCR suggest that these proteins are part of variety of functional events such as signal transduction, enzymatic processes, cell differentiation, pyrimidine biosynthesis, fatty acid biosynthesis and chromosomal replication. Representations of low complexity regions of *Plasmodium falciparum* in protein data bank suggest that LCRs can take conformation of regular secondary structure (apart from disordered regions) in 3-D structures of proteins.

Chapter 6 describes sequence analysis, structural modeling and evolutionary studies of *Leishmania donovani* hypusine pathway enzymes. *Leishmania* is an eukaryotic kinetoplastid protozoan parasite which causes leishmaniasis in humans. Hypusine is a non standard polyamine-derived amino acid N\(^{ε}-(4\text{-amino-2-hydroxybutyl})\) lysine and is named after its two structural components, hydroxyputrescine and lysine. The eukaryotic translation initiation factor 5A (eIF5A) is the only cellular protein containing hypusine. Synthesis of hypusine is critical for the function of eIF5A and is essential for eukaryotic cell proliferation and survival.

Formation of hypusine is the result of a two step post-translational modification process involving enzymes (i) deoxyhypusine synthase (DHS) (ii) deoxyhypusine hydroxylase (DOHH). DHS, the first enzyme involved in hypusine pathway catalyzes the NAD-dependent transfer of the butylamino moiety of spermidine (substrate) to the ε-amino group of a specific lysine residue of eIF5A precursor and generates deoxyhypusine containing intermediate. DOHH, the second enzyme in same pathway catalyzes the hydroxylation of deoxyhypusine-containing intermediate, generating hypusine-containing mature eIF5A.

Two putative deoxyhypusine synthase (DHS) sequences DHS34 and DHS20 have been identified in *Leishmania donovani*, by Professor Madhubala and coworkers (Jawaharlal Nehru University, New Delhi) with whom the work embodied in this chapter was done in collaboration.

Detailed comparison of DHS34 sequence from *Leishmania* with human DHS protein indicated conservation of functionally important residues. 3D structural modeling studies of protein suggested that residues around the active site were absolutely conserved. NAD binding regions are located spatially closer, however, one NAD binding region was observed in a large (225 amino acid residues long) insertion. Based on these observations, DHS34 was predicted to have enzymatic activity. Experimental studies done by our collaborators confirmed preliminary results of computational analysis.
Based on sequence and structural analysis of DHS20 and DOHH proteins, DHS20 and DOHH were proposed to be catalytically inactive and active respectively. Experimental studies on these proteins supported results of computational analysis.

Deoxyhypusine synthase (DHS) and Deoxyhypusine hydroxylase (DOHH) are key proteins conserved in the hypusine synthesis pathways of eukaryotes. Because they are highly conserved, they could be coevolving. Comparison of the genetic distance matrices of DHS and DOHH proteins reveals that their evolutionary rates are better correlated when compared to the rate of an unrelated protein such as Cytochrome C. This indicates that they are coevolving, further serving as an indicator that, even non-interacting proteins that are functionally coupled, experience correlated evolution. However, this correlation does not extend to their tree topologies.

**Chapter 7** provides a classification scheme for protein kinases encoded in genomes of prokaryotic organisms. Overwhelming majority of the Ser/Thr protein kinases identified by gleaning archaeal and eubacterial genomes could not be classified into any of the well known Hanks and Hunter subfamilies of protein kinases. This is owing to the development of Hanks and Hunter classification scheme based on eukaryotic protein kinases which are highly divergent from their prokaryotic homologues. A large dataset of prokaryotic Ser/Thr protein kinases recognized from genomes of prokaryotes has been used to develop a classification framework for prokaryotic Ser/Thr protein kinases.

Traditional sequence alignment and phylogenetic approaches have been used to identify and classify prokaryotic kinases which represent 72 subfamilies with at least 4 members in each. Such a clustering enables classification of prokaryotic Ser/Thr kinases and it can be used as a framework to classify newly identified prokaryotic Ser/Thr kinases. After series of searches in a comprehensive sequence databases, it is recognized that 38 subfamilies of prokaryotic protein kinases are associated to a specific taxonomic level. For example 4, 6 and 3 subfamilies have been identified that are currently specific to phylum proteobacteria, cyanobacteria and actinobacteria respectively. Similarly, subfamilies which are specific to an order, sub-order, class, family and genus have also been identified. In addition to these, it was also possible to identify organism-diverse subfamilies. Members of these clusters are from organisms of different taxonomic levels, such as archaea, bacteria, eukaryotes and viruses.

Interestingly, occurrence of several taxonomic level specific subfamilies of prokaryotic kinases
Synopsis

contrasts with classification of eukaryotic protein kinases in which most of the popular subfamilies of eukaryotic protein kinases occur diversely in several eukaryotes. Many prokaryotic Ser/Thr kinases exhibit a wide variety of modular organization which indicates a degree of complexity in protein-protein interactions and the signaling pathways in these microbes.

Chapter 8 focuses on recognition, classification of protein kinases encoded in genomes of viruses and their implications in various functions and diseases. Protein kinases encoded by viral genomes play a major role in infection, replication and survival of viruses. Using traditional sequence homology detection tools, sequence alignment methods and phylogenetic approaches, protein kinases were recognized. 646123 protein sequences from 35799 viral genomes (including strains) have been used in this analysis. Protein kinases are identified using a combination of profile-based search methods such as PSI-BLAST, RPS-BLAST and HMMER approaches.

Based upon sequence similarity over the length of catalytic kinase domains, 479 protein kinase domains recognized in 244 viral genomes have been clustered into 46 subfamilies with minimum sequence identity of 35% within a subfamily. Viral protein kinases are encoded in genomes of retro-transcribing viruses or viruses which possess double stranded DNA as genetic material. Based on the available functional information present for one or more members of a subfamily, a putative function has been assigned to other members of the subfamily. Information regarding interaction of viral protein kinases with viral/host protein has also been considered for enhancing understanding of function of kinases in a subfamily.

Out of 46 subfamilies, 14 subfamilies are characterized by various functions. Kinases belonging to UL97, US69, UL13 and BGLF subfamilies are virus specific. For 7 subfamilies, nearest neighbors are from well characterized eukaryotic protein kinase groups such as AGC, CAMK and CDK. Out of 25 new uncharacterized subfamilies observed in this analysis, 13 subfamilies are virus specific. Different subfamilies have been characterized by various functions which are crucial for viral infection such as synthesis of structural unit, replication of genetic material, modification of cellular components, alteration in host immune system, competing with cellular protein for efficient usage of host machinery.

Also, many viral kinases share very high sequence identity (~97%) with their eukaryotic counterpart and represent disease state. For example, a protein kinase encoded in Avian erythroblastosis virus shares 97% sequence identity with catalytic domain of human epidermal growth factor receptor tyrosine kinase. Leucine at position 861 in human protein is substituted
by Gln in cancer conditions; the viral protein kinase sequence possesses Gln at corresponding position and thus represents disease state.

**Chapter 9** provides study of dependency on the ability of 3-D structural features of comparative models and crystal structures of inactive forms of enzymes to predict enzymes by considering protein kinases as case study. With the advent of structural genomics initiatives, there is a surge in the number of proteins with 3-D structural information even before functional features are understood on many of these proteins. One of the useful annotations of a protein is the demarcation of a protein into an enzyme or non-enzyme solely from the knowledge of 3-D structure. This is facilitated by the identification of active sites and ligand binding sites in a protein. In this work, which was carried out in collaboration with Dr Jim Warwicker of Manchester University, UK, an approach developed by Warwicker and coworkers has been used.

In the 3D structure of proteins, the largest clefts are generally considered to be ligand binding sites. This feature along with other sequence alignment independent properties such as residue preferences, fraction of surface residues and secondary structure elements have been considered to differentiate enzymes from non-enzymes. Electrostatic potential at the active site is one of the key properties utilized in this respect. Active sites in enzymes are generally associated with ionizable groups which can take part in catalysis. In addition to the feature of large clefts in enzymes, active site residues are in buried environments and show larger deviation in pKa values than surface residues. The method proposed by Warwicker and co-workers distinguish proteins in to enzymes and non-enzymes considering the electrostatic features at clefts along with the sequence profile of the protein concerned.

Conformation of the inactive state of an enzyme is not congenial to the catalytic function. In an ideal situation, a method should be capable of predicting an enzyme irrespective of whether determined structure corresponds to active or inactive state. Peak potential values have been calculated by using Warwicker program for a set of 15 protein kinases for which 3-D structures are present in active as well in inactive conformations. Comparison of peak potential values calculated for active and inactive conformations suggests that algorithm can differentiate between active and inactive conformations as value for active conformations are generally higher than corresponding values for inactive conformations. However, the peak potential values are high enough for even the inactive conformations to be predicted as enzyme.

Peak potential values calculated for generated homology models of protein kinases (for which crystal structures are already available) at different sequence identities with template sequences
predict protein kinases as enzymes and their peak potential values are comparable to corresponding values for X-ray structures. This suggests that proteins for which there are no crystal or NMR structures yet available and no good template with high sequence identity are present, peak potential values for models generated at low sequence identity can still give insight into probable function of protein as an enzyme.

The enzyme/non-enzyme prediction algorithm was also found to be useful in confirming enzyme functionality using 3-D models of putative viral kinases. Initially, putative function of kinase has been assigned to these viral proteins based solely upon their sequence characteristics such as presence of residues/motifs which are important for activity of the protein. The enzyme recognition method which is not directly sensitive to these motifs confirmed that all the analyzed putative viral kinases are enzymes.

**Chapter 10** presents conclusions of work embodied in the entire thesis. Very briefly, various computational approaches have been used to analyze and understand structural and functional properties of repertoire of proteins of pathogenic organisms. Analysis of uncharacterized protein domain families has helped to understand the functional implications of constituent proteins. Experimental validation of these results can further facilitate unraveling of functional aspects of proteins encoded in various pathogenic organisms.

Apart from studies embodied in the thesis, author has been involved in two other studies, which are provided as appendices.

**Appendix 1** describes comparison of substitution pattern of amino acid residues of protein encoded in *P. falciparum* genome with substitution pattern of corresponding homologous proteins from non-*Plasmodium* organisms. Salient differences have been highlighted.

**Appendix 2** discusses study of bacterial tyrosine kinases with an objective of recognition of all putative protein tyrosine kinases in *E. coli*. Computational study suggests that protein SopA can be a potential tyrosine kinase and this conclusion is being tested experimentally in collaborator’s laboratory.