Chapter 9
Conclusion
9.1 Integrative Biology of post genome era

Biology is now in a transition stage from a data-rich science to a knowledge-based science due to large-scale influx of data from various experimental and computational studies. Major biological databanks that store genomic and protein sequence, protein structure and protein-protein interaction data are witnessing an unprecedented growth over the past decades due to low-cost, high-throughput experimental platforms, bioinformatics algorithms and powerful computational methods. Novel bioinformatics protocols, improved algorithms and translational approaches are required for integration, analysis and interpretation of heterogeneous datasets [2, 634]. Among the various data categories that observed exponential growth over the past decades, protein domains are considered as highly important. Protein domains form the core, conserved element of proteins involved in a variety of functional roles as a crucial mediator between the genome and functional interactions within the cell. Deep analysis of proteins domains using computational methods helped to improve the understanding of protein evolution, protein folding and the functional integrity of protein networks [1, 14, 32].

9.2 Relevance of identification and analysis of proteins domains in post-genome biology

This thesis entitled as “Identification and analysis of domains in proteins” explores three categories of protein domains: “sequence domains”, “structural domains” and “functional or interacting domains” organized into 9 Chapters. Thesis content is broadly divided into three major sections within a common framework of bioinformatics databases, tools, algorithms and computational methods developed for the identification and analysis of protein domains. Chapter 1 provides a detailed background of protein domains and computational methods available for analysis of protein domains based on the literature survey of the core theme of the thesis. Chapter 2 is dedicated for sequence domains and discusses a novel bioinformatics data mining protocol and associated data resource for the identification of best representative member of a protein domain family. Chapters 3-6 are focused on an important structural mechanism “3D domain swapping” observed in an array of crystal structures. Chapter 3 presents a new curation method “structure-based literature curation” and employed this method to identify the key features of proteins involved in 3D domain swapping from literature. This information is compiled into a database and is provided in public domain for the access of the structural bioinformatics community. Chapter 4 reports the implementation of a support vector machine (SVM) based algorithm for the classification of a protein structure in to swapping or non-swapping categories using features derived from sequence
and structure data. Chapter 5 presented another novel application of machine learning algorithm to utilize sequence features for the prediction of 3D domain swapping from the sequence data using RandomForest approach. The detailed meta-analysis approaches implemented for the in-depth analysis of proteins involved in 3D domain swapping are discussed in Chapter 6. Chapter 7 provides an overview of the implementation of a structural bioinformatics tool for the analysis of higher order residue interactions in structural domains. Chapter 8 focuses on interacting domains and in this Chapter, I discuss a novel approach of “interacting sequence space” for utilizing the protein-protein interaction data for screening of functional patterns from local interaction space of proteins.

9.3.1 A brief introduction of protein domains from the perspective of sequence, structure and interaction

Post-genome era is witnessing enhanced annotations of protein sequences using computational methods. Chapter 1 provides a detailed overview of protein domains and the classification of protein domains into three distinct classes such as: sequence domains, structural domains and interacting domains (See Figure 1.3). The introduction Chapter provides a detailed overview of the three classes from the perspective of research questions discussed in this thesis. Introduction Chapter also provides a brief overview of various bioinformatics protocols, data resources, tools and integrative approaches that are currently available for the analysis of protein domains [9, 10, 285, 295] and provides reasoning for why novel and improved methods are required for identification and analysis of protein domains.

9.3.2 A novel data mining approach for the identification of Best Representative PSSM Profiles (BRPs) from protein domain families

Chapter 2 explores the sequence domains and proposes a novel data mining method for the identification of a best representative member a protein domain family. Due to exceptional growth of sequence databases, protein domain families are gathering large number of members within short span of time (See Figure 1.9). It is often a difficult task to identify a best representative member of a protein domain family and sequence searches were being done in an ad-hoc fashion using a normal reference sequence. Using sequence data, protein domain families could be related to each other at broad levels that group them as families or superfamilies. These relationships are harder to detect at the sequence level due to high evolutionary divergence. Typically, a protein domain family from Pfam database may have hundreds to thousands of protein sequences in their seed and independent datasets. This
Chapter discusses a novel data mining method developed to identify best representative profiles (BRP) [297] or best representative sequence (BRS) [635] from protein domain families. BRP is generated from the reference sequence that encapsulates all the important information of a diverse or highly similar set of sequences in a protein family into one single profile. The method is applied to the Pfam version 22 and the method successfully obtained best representative for 91.4% of protein domain families. BRS is defined as the reference sequence used to generate the BRP of a given family. The chapter provides the details about the availability of an associated data resource, 3PFDB, developed using the large amount of data generated during the analysis of best representative members. The data were used to compile database of BRPs and also discuss various features of the database and their application in bioinformatics analysis. The database 3PFDB is provided in the public domain and accessible from URL: http://caps.ncbs.res.in/3pfdb

The method and the database discussed in Chapter 2 can be utilized in various scenarios where best representative members of protein domains are required for effective annotation and analysis of protein domains. Machine learning algorithms often employs PSSMs to define evolutionary features. PSSMs archived in 3PFDB are ideal for such studies. PSSMs from 3PFDB can also be used as an input for RPS-BLAST searches. HMM models included in the database can be used for annotations of protein sequences.

9.3.3 3DSwap: Curated knowledge base of proteins involved in 3D domain swapping

Availability of structural domains datasets is limited in comparison with the sequence or interacting domains due to various experimental limitations in obtaining protein structures. Such domains are crucial to understand the structural features that mediate the function. Chapters 3, 4 and 5 discuss an interesting structural phenomenon called “3D domain swapping”, observed in different proteins. 3D domain swapping is studied in detail using various approaches like literature-based structural curation, data integration, data mining, machine learning algorithms and meta-analysis approaches. 3D domain swapping is a structural phenomenon observed in a variety of proteins as a mechanism for dimer or higher oligomeric formation. 3D domain swapping phenomenon was explored from a bioinformatics perspective in this thesis. 3D domain swapping is a mechanism by which two or more protein molecules form a dimer or higher oligomer by exchanging an identical structural element between the protein chains. In domain swapping, native intramolecular interactions were replaced by their intermolecular counterparts. It is one of the most interesting structural
features that have been observed in an array of crystal structures. This process emerges as an ubiquitous mechanism for homo-oligomer formation in many completely unrelated protein systems. Domain swapping is an interesting phenomenon not only due to its ‘unity in diversity’ factor; but also due to variety of molecular functions mediated by this swapped macromolecule. The integrated information using bioinformatics analysis, literature curation and manual inspection of protein structures involved in 3D domain swapping, has been compiled into a database of 293 proteins “3DSwap - Knowledgebase of 3D domain swapping in proteins” (K. Shameer et. al, manuscript submitted to Database: The Journal of Biological Databases and Curation). In order to generate the resource information for this database, a combination of integrative database searches, structure analysis and literature curation were required.

Chapter 3 provides a detailed overview of literature-based protein structure curation strategies and data integration approaches used to develop a knowledge-base of proteins reported to be involved in 3D domain swapping for oligomeric assembly. All the available proteins reported to have involved in this mechanism is obtained from literature and bioinformatics database approaches and reported in a new database: “3Dswap”. For the better understanding of different structural and functional aspects of 3D domain swapping, these terms were introduced in this Chapter; a set of four new terms are also introduced viz. ‘primary inter-domain interface’, ‘secondary inter-domain interface’, ‘secondary minor region’, ‘secondary major region’. A new quantitative method called “extent of swapping” is proposed to quantify the effect of swapping in proteins structures. This Chapter further explains various features and tools incorporated in 3Dswap database. 3Dswap is available in the public domain from the URL: http://caps.ncbs.res.in/3dswap

3DSwap knowledge base is a primary attempt to develop a database based on 3D domain swapping. As more and more structures are currently being reported to be associated with 3D domain swapping, the database content will be essential for both structural biology and functional studies. Further, the curation method can be utilized to study similar group of proteins involved in unique structural phenomenon.
9.3.4: Support vector machine-based algorithm for prediction of 3D domain swapping using features derived sequence and structural data

Application of machine learning methods in bioinformatics for pattern discovery and data mining is gaining more significance due to the generalization capability of such methods and relevance in biological problems. Chapters 3 and 4 provided detailed overview of the implementation of soft computing-based prediction models, using support vector machines and Random Forests, to predict 3D domain swapping from structure and sequence-derived features. Chapter 4 describes about approaches to use data compiled in 3DSwap for the development of a new machine learning algorithm for the prediction of 3D domain swapping using sequence and structure derived features. 3-Dimensional domain swapping is a mechanism where two or more protein molecules form higher order oligomers by exchanging identical or similar subunits. While 3-Dimensional Domain swap mechanism can be detected from three-dimensional structures, it remains a formidable challenge to derive common sequence or structural patterns from the proteins involved in swapping. We have developed a SVM-based classifier to predict domain swapping events using a set of features derived from sequence and structural data [298]. The SVM classifier was trained on sequences derived from 220 protein structures involved in domain swapping and sequences derived from 220 different protein structures not known to be involved in swapped conformation or related to proteins involved in swapping phenomenon. The testing was performed on 63 domain swapping and 63 non-domain swapping sequences. We obtained 76.33% accuracy from training and 73.81% accuracy from testing. The approach provides insights into a set of features, which upon experimental validation may potentially help to understand the mechanisms mediating 3D domain swapping in proteins.

9.3.5: 3Dswap-pred: Machine learning algorithm for the prediction of 3D domain swapping from protein sequence using Random Forests

Chapter 5 provides details about the prediction model developed using sequence-derived features to classify a given protein sequence as “3D domain swapping” or “non-swapping” classes using sequence-derived features. A new algorithm was developed for the prediction of the structural phenomenon “3D domain swapping” from sequence data using Random Forests. The prediction method achieved an overall accuracy of 63.78% accuracy with sensitivity of 58.82% and specificity of 64.86% with the independent dataset. The methods which rely on features derived from sequence “3dswap-pred” were implemented as a web server. Results from the feature selection method used in the algorithm shows that hybrid
features like “sequence-derived fusion features” can be utilized as an effective manner of dealing with various classification problems in biology. This Chapter also explains, in detail, about the RandomForest algorithm, features, feature selection method, assessment of prediction model and future directions to improve the sequence based prediction model. An online bioinformatics server to predict whether a given sequence can undergo 3D domain swapping is implemented using the prediction model. 3Dswap-pred is server is available in the public domain from the URL: http://caps.ncbs.res.in/3dswap-pred/ (K. Shameer et. al, manuscript accepted in Protein & Peptide Letters, 2011). 3dswap-pred could be useful to identify putative proteins, which may involve in 3D domain swapping using sequence data. Machine learning algorithms described in Chapters 5 and 6 could be further improved by introducing different statistical models for feature ranking, feature selection and classification.

9.3.6 Meta-analysis of proteins involved in 3D domain swapping

Chapter 6 provided a detailed account of the meta-analysis of proteins involved in 3D domain swapping (K. Shameer et. al, manuscript submitted to BMC Structural Biology). Statistical tests were used to identify the significance of various categories of data related to sequence and structure of proteins involved in swapping. For example, Gene Ontology (GO) based enrichment were performed for all entries using ‘biological process’, ‘molecular function’ and ‘cellular component’. Detailed analysis was performed in the level of Pfam domains, Pfam clans, SCOP class, fold, superfamily and family. Different segments of proteins involved in 3D domain swapping were also analyzed using secondary structure elements, solvent accessibility and Ooi numbers. Further detailed functional enrichment using Gene Ontology and KEGG pathways were performed using human proteins involved in 3D domain swapping. New insights and therapeutic role of 3D domain swapping in various diseases beyond neurodegenerative diseases were observed (K. Shameer and R. Sowdhamini; Manuscript submitted to Trends in Biochemical Sciences). ‘Database-wide enrichment analysis’ is a new method introduced in the thesis to perform functional enrichment analysis for a group of proteins involved in a common biological mechanism (K. Shameer and R. Sowdhamini; Manuscript submitted to Journal of Biomedical Semantics). This Chapter provides a comprehensive view of various properties of proteins involved in 3D domain swapping from the perspective of functional motifs, domains, Gene ontology, KEGG pathways and disease ontology. Meta-analysis of proteins involved in 3D domain swapping provided new insights in to functional properties of proteins involved in 3D domain
swapping. 3D domain swapping has received much attention in the context of prions and neurodegenerative diseases, due to its role in the functional regulation, formation of higher oligomers, protein misfolding, aggregation etc. Our meta-analysis using human proteins indicated that human proteins involved in 3D domain swapping is part of several biochemical pathways and disease categories.

9.3.7 HORI: A tool for computing Higher Order Residue Interactions in structural domains and its application in protein structure analysis

Folding of a structural domain of a protein into its three-dimensional structure is influenced by both local and global interactions within a protein. Higher order residue interactions, like pairwise, triplet and quadruplet ones, play a vital role in attaining the stable conformation of the protein structure. It is generally agreed that higher order interactions make significant contribution to the potential energy landscape of folded proteins and therefore it is important to identify them to estimate their contributions to overall stability of a protein structure. Protein structural domains maintain their folds using a variety of interactions. A web server called “HORI” is developed for the computation of higher order residue interactions in protein structural domains. The server can be used for the analysis of pairwise, triplet and quadruple interactions in the structural domains. Chapter 7 provides details about a new bioinformatics web server developed for the computation of higher order residue interactions and further explains its features and application in the analysis of structural domains in proteins [300, 301]. Different programs for computation of higher order residue interactions in protein structures can be utilized in different contexts of protein structure analysis. For example, the higher order interactions can be utilized in the identification of alternate binding sites, active site residues (biochemical approaches), to design point mutations (protein engineering) and to recognize appropriate folds for a protein of known sequence (fold prediction). The HORI server is provided in the public domain and available from the URL http://caps.ncbs.res.in/hori. The server will be highly useful for structural bioinformatics community to perform structure analysis using higher order residue interactions. Higher order residue interactions could be also incorporated into new fold recognition methods and algorithms for predicting protein structural features.

9.3.8 PeptideMine: Computational method for designing peptides using functional patterns derived from protein-protein interactomes
Chapter 8 discusses a novel method developed to utilize interacting domains. A functional or interacting protein domain performs its respective function by interacting with other macromolecules and small molecules. Due to high-throughput experimental and computational approaches, organism-level protein-protein interaction and protein domain family based domain-domain interaction data are currently available. A new approach has been developed by integrating 12 bioinformatics resources called “PeptideMine” for the identification and analysis of peptides for computational and experimental protein-peptide binding studies. The method enables search in the “interacting sequence space” of a protein to identify potential peptides based on three different search methods [288]. The method was made available to the public as a web server and currently available for 5 model organisms. The PeptideMine server is available from the URL http://caps.ncbs.res.in/peptidemine.

PeptideMine method is a useful method to find functionally relevant peptides from interactomes using a knowledge-based approach. In-depth analysis could be performed for the analysis of interacting sequence space of protein families to understand important functional cues.

Appendix: In addition to the core Chapters, thesis also discusses an algorithm (STIF [636]) and associated database (STIFDB URL: http://caps.ncbs.res.in/stifdb [637]) developed for the identification of stress-responsive transcription factor sites and transcription factor domains.

9.4 Summary
The thesis provides a detailed account of analysis approaches designed around various aspects of protein domains to obtain new insights and identify new aspects of protein domains using powerful computational methods. Protein domains were analyzed from the perspective of sequence, structure and interaction. Thesis discussed about various novel methods, databases and web servers developed for the identification and analysis of protein domains. This was complemented with analysis of various protein domain families. Thus, this thesis addresses important problems and provides detailed account of several new concepts like best representative members of protein domain families, structure-based literature curation strategy, extent of swapping, identification of higher order residue interactions in quadruple scale and sequence searches in interacting sequence space to find functionally relevant patterns. Meta-analysis approaches used to analyze 3D domain swapping provided new insights into sequence, structure and functional properties of proteins involved in 3D domain swapping.
This thesis also introduces several novel bioinformatics methods for computational analysis of protein domains. A novel data mining method is designed to identify best representative members of protein families. Two novel machine learning algorithms using support vector machines and Random Forests are developed to predict 3D domain swapping. A set of novel hybrid features are introduced and incorporated into these algorithms. A new functional enrichment calculation strategy “database-wide enrichment calculation” is proposed and applied to proteins involved in 3D domain swapping. Novel concept of interacting sequence space is proposed and utilized in ‘PeptideMine’ to identify functional peptides from interactome of a given protein. Three new integrative bioinformatics databases (3PFDB, 3Dswap, STIFDB (See Appendix 1)) and three new web-based bioinformatics tools (3dswap-pred, HORIZ and PeptideMine) were also developed as part of this study. Various computational approaches and bioinformatics methods proposed in this thesis will be recognized as important areas in the coming years and open new avenues for further research in integrative biology.