Chapter 1
A brief introduction to protein domains from the perspective of sequence, structure and interaction

Publications from this chapter:

1.1 Introduction

Bioinformatics is now in a transition stage, from a data-centric component of biology to a knowledge-based science [1-3]. This transition is consistent with a more pro-active role of bioinformatics approaches in supporting experimental strategies based on a priori information. The complexity of biological systems requires efficient integration of data, tools and protocols to extract relevant information. Computational approaches using high-throughput data integration in bioinformatics have provided new insights into several biological problems. Bioinformatics approaches that utilize available data to generate well-defined, experimentally verifiable data are clearly needed to translate in silico predictions into testable hypothesis. Integrative approaches can provide new avenues to understand molecular systems and aid the design of new experiments to identify interesting molecular players. Large-scale data integration, data mining, machine learning methods and integrated computational approaches in bioinformatics could accelerate such endeavors in genomics, proteomics, metabolomics, network biology, systems biology and translational research studies [4-15].

In the second decade of 21st century, life science is now at the crossroads of large-scale data influx from high-throughput experiments. Such data are being generated at an accelerated rate, efficient and novel bioinformatics solutions are being developed to uncover biologically valid, functional information from such huge datasets. Rapid increase in data generation has resulted in a decline in functional characterization due to practical limitations in characterizing every other gene or protein sequence. This resulted in the accumulation of ‘unknown’ or ‘hypothetical’ genes with no known functional roles in mediating cellular mechanisms [4]. It is practically impossible to validate the available sequence data by means of biochemical experiments for confirmation of the likely associations. In this scenario, bioinformatics approaches can play an important role as a filter for recognizing potential gene products that can represent new fold or a novel function or to associate a sequence with an existing protein fold or biological function [16-22]. Computational approaches enable the recognition of putative gene products of a family and to rationally design mutation experiments [23-28]. Along with rapid incoming data, the availability of various computational resources to analyze the data has also increased. Bioinformatics databases, web servers, web services, software libraries and application programming interfaces are currently available for the integration, analysis and interpretation of biological data.
Bioinformatics based algorithms; approaches and analysis have made extensive contributions to the understanding of various aspects in biology [29, 30].

Protein domains are an important paradigm where bioinformatics contribution was highly significant. Bioinformatics approaches contributed to the understanding of evolutionary aspects, prediction of function and classification of proteins. For example protein sequence classification (protein domains, protein families, sub-families, protein superfamilies) [31] and protein structural classification (domain, family, superfamily, folds and super-folds) are two areas where bioinformatics contributed to the understanding of protein structure and function. [22, 32-39].

1.2 Proteins
Protein structures are elementary units of form and function in living organisms. Sequence, structure, function and evolutionary properties of proteins can be studied using different approaches [40-45]. Protein sequencing methods can be used to identify the sequential order of amino acids. 3-dimensional structure can be elucidated by purification, expression, cloning and crystallization studies using X-ray crystallography or Nuclear Magnetic Resonance technique. The term “protein” refers to organic molecule encoded in the coding regions of genome, translated and transcribed in to polypeptides, destined to perform specific functional roles in the cell (Figure 1.1). Human genome consists of three distinct sequence regions known as coding regions, non-coding regions and regulatory regions. Coding regions refers to the genomic region composed of deoxyribonucleic acid (DNA) molecules, which can be transcribed in to a messenger RNA (mRNA) and translated in to a single protein molecule. Non-coding regions are the genomic regions that do not encode a protein sequence. Regulatory regions in the human genome refers to the genomic region that can regulate (as in increase, decrease or initiate) the expression of a particular gene due to the binding of regulatory proteins like transcription factors or other molecular cues. A single protein molecule is expressed in a cell after the multiple steps of transcription of mRNA from genomic DNA, splicing of mRNA, translation, and post-translational modifications. Proteins are composed of twenty standard amino acids. Post-translational modifications add chemical specificity to the protein chain to perform the specific function. Typically, a protein performs its function inside a cell by interacting with other micromolecules (ions, chemical molecules, etc) and macromolecules (DNA, various types of ribonucleic acid (RNA) and proteins)
within the cellular environment. Attaining the specific 3-dimensional structure and fold is crucial for a protein to perform the particular function.

1.3 Protein structure

Structural properties of proteins can be comprehensively explained using the hierarchical concept of primary, secondary, tertiary and quaternary structures [40, 41]. Primary structure refers to the sequence of a protein characterized by protein sequencing methods. Primary structure of a protein can be experimentally characterized using protein sequencing experiments like Mass spectrometry or Edman’s degradation method [46]. Apart from the sequencing methods, sequence of a protein can be derived from the sequence of genomic DNA by translation of genomic DNA using bioinformatics workflow systems. Due to lower cost in genome sequencing and availability of various complete genome sequences and metagenomics projects, a larger repertoire of protein sequences are currently generated using bioinformatics method and later validated using experimental approaches. The latter approach, fuelled by genome sequencing initiative, provided a deluge of protein sequence data and initiated better understanding of evolutionary perspective. [22, 47-49]

Secondary structure refers to the ordered units of structural elements within the protein sequence. They are local sub-structures made of α helix, β sheets and loops. Secondary structure can be observed during structure characterization experiments (X-Ray or NMR) or it can be derived from bioinformatics experiments. A variety of secondary structure prediction algorithms are available for the elucidation of basic secondary structure from protein sequence. Super-secondary structure (SSS) [50] refers to the ordered assembly of several secondary structure elements observed during the nucleation step of the protein folding mechanism. β-hairpins, α-helix hairpins, and β-α-β motifs are examples for super-secondary structure [51]. Tertiary structures refer to the fully folded state of protein molecules. Protein should attain the native fold [43, 52] to participate in the specific molecular function or biological process [36]. Folding of a protein sequence in to tertiary structure is driven by various factors like non-covalent interactions, hydrogen bonding, ionic interactions, electrostatic interactions, hydrophobic interactions and van der Waals interaction. Tertiary structure of proteins can be experimentally characterized using X-ray crystallography or NMR spectroscopy techniques. Computational approaches like homology modeling, fold recognition and ab-initio protein-modeling techniques are also available to predict or understand the fold of a protein structure. Primary and tertiary structure of crambin
is provided in Figure 1.2 Several proteins perform their respective function by forming protein complexes. A *protein complex* is an assembly of individual polypeptide chains. *Quaternary structure* is the assembled form of individual polypeptide chains or subunits in the multimeric conformation. In general, *protein oligomers* are formed by polymerization of protein monomeric sub-units into dimmers or higher oligomers [53, 54]. Protein oligomers can be broadly classified into two classes: *homo-oligomers* where the monomers are of subunits of same protein where as *hetero-oligomers* are formed by oligomerization of subunits from different proteins. Several molecular functions rely on interactions facilitated by such protein oligomers with other molecular players in the cell [31, 34, 45, 46, 49, 55-61].

### 1.4 Protein Domains

A protein domain is a part of protein sequence and structure that can evolve, function, and exist independently of the rest of the protein chain [62-64]. Regulators of G protein signaling domain (RGS), SH3 domain (Src homology 3), protein kinase domain (Pkinase) and basic leucine zipper domain (bZIP) are few examples of protein domains. Protein domain forms a compact three-dimensional structure and often can be independently stable and folded. Proteins domains can be defined as conceptual framework to describe the conserved elements in proteins based on sequence or structural homology [65, 66]. It can be utilized for function association of the large number of gene products realized from whole genome sequencing projects [2, 67-71]. Homology is a fundamental principle in biology, which refers to the similarity between two sequences or structures due to descent from a common ancestor. Homology, inferred by sequence similarity, is usually a reason for transfer of function annotation from pre-existing domain families to gene products [72-74]. Conserved elements in protein sequences are generally called *domains or motifs* [75] based on the number of residues that constitute the conserved elements. Domains and motifs are generally distinguishable in a sequence based on conservation observed in sequence alignments. From a biological viewpoint, conserved elements are retained by evolution to retain the functional role imparted by the proteins. Fundamental difference between a domain and motif is in their lengths. From sequence perspective protein domains are the segment of contiguous sequence of amino acids. Sequence domains vary in length from 40 to 350 amino acids and average size of protein domains is considered as 150 residues. Sequence motifs are usually smaller in size with size less than 25 amino acids in length. Short peptide motifs are also called *linear motifs* (*peptide recognition motif* is an example of linear motifs) [76, 77]. Structural motifs are three-dimensional structural (for example helix-turn-helix motif [78, 79], Greek-key motif
Protein sequence domains are the conserved elements in a protein sequence or structure that can fold and perform its function independently with respect to the rest of the protein chain. Based on the type of data used to define domains, protein domains can be classified as sequence domains, structure domains and interacting domains (Figure 1.3). Sequence domains are generally derived from sequence alignments of related protein sequences. Structural domains refer to the three-dimensional structure of a conserved protein domain with experimentally characterized three-dimensional coordinates deposited in public protein structure database like Protein Data Bank (PDB) (Figure 1.4). A structural domain is defined as a compact, globular substructure with more interactions within the domain itself than with the rest of the protein. From a structural perspective, protein domains are conserved structural segments composed of multiple secondary structure units, which can fold independently [31, 49, 56].

Protein domains can exist as single domain and multi-domain forms [85, 86]. Ordered arrangement of protein domains in a protein is called as the protein domain architecture or protein domain organization [87-89]. For example: Supra-domains refer to a peculiar class of domain architecture where the two-domain and three-domain combinations are found recurrently in different proteins with different partner domains [90, 91]. Protein domain architecture is important for the structural integrity and maintaining the function of a protein. Assigned regions refer to the continuous stretch of sequence where a domain is observed. Apart from conserved regions like domains of functional motifs, proteins may also have unassigned regions [92], which are regions with no known domain association. Graphical representation of domain architecture is provided in Figure 1.5. Promiscuous domains are a subset of protein domains that show tendency to participate in a variety of protein domain architectures [86, 93]. It plays a major role in mediating protein–protein interactions, protein binding, and signal transduction pathways inside the cell. Based on conserved domains and evolutionary relationships, proteins can be grouped into a protein family [94]. A protein family consists of homologous proteins with similar functional role. For example, all proteins with a ‘regulators of G-protein signaling (RGS)’ protein domain are classified into RGS family [95, 96]. Similar protein families are combined into protein superfamily [31, 49] and superfamily members can be related to each other based on distant relationships in sequence or structure level. Concept of protein domains, protein family and protein superfamily are
applicable to both sequence domains and structural domains. In addition, exclusive classification schema like *subfamilies* [97], *subdomains* [98], *clans* [99] and *folds* [100-103] are defined for the effective understanding and classification of proteins. Subfamily refers to a subset of a protein sequence or structure family, where the subset can be grouped based on increased sequence similarity or structural similarity. Protein subdomain is defined as a segment of a protein domain. Protein fold is the three-dimensional arrangement of secondary structural elements (SSE) into a unique tree dimensional topology [100-103]. Current version of Structural Classification of Proteins (SCOP) database that provides statistics of available folds indicates that release 1.75 consists of 38221 PDB Entries (released on 23 Feb 2009) with 110800 domains and 1195 folds. Smaller number of folds was attributed to the evolutionary restrictions on fold space. Clan is a sequence-based concept introduced in Pfam database to indicate group of similar protein domains [99, 104-106].

A protein domain mediates its function by interacting with various molecules inside the cell [107]. Such interactions can be broadly classified into protein-nucleic acid interactions, protein-small molecule interactions and protein-protein interactions [108-112]. In the recent years, the advent of high-throughput experimental methods and prediction algorithms enabled the generation of large-scale protein-protein interaction data from various model organisms. Interactions between two proteins are generally called as *protein-protein interactions*, where as molecular interaction between two distinct protein domains are known as ‘*domain-domain interactions*’ [113-118] and the protein domains participating in such interactions are called as ‘*interacting domains*’. Interacting domains are proteins domains, which may participate in a variety of molecular interactions like protein-protein interactions and domain-domain interactions to perform a specific function. Domain-domain interactions can be either intra-domain interaction (where domains within a same protein chain interacts with each other) or inter-domain interaction (where domains from two different protein chains interact with each other) [119]. Availability of large-scale interaction data initiated network based analysis of proteins to understand the global properties of interaction networks related to various diseases and molecular mechanisms. Domain-domain interaction data are currently available based on interaction data derived from protein complexes deposited in PDB and various computational methods for predicting such interactions [107, 113, 114, 118, 120-122]. Graphical representation of protein-protein interaction and domain-domain interaction is provided in Figures 1.6 and 1.7.
1.5 3D domain swapping

Many cellular functions rely on interactions between protein pairs and are mediated by proteins in oligomeric conformations [40]. Although there are many possible mechanisms for oligomeric formation, 3D domain swapping has been proposed as an important structural phenomenon in mediating interactions. Several molecular functions rely on interactions facilitated by such protein oligomers with other molecular players in the cell. 3D domain swapping is a protein structural phenomenon observed in an ensemble of protein dimers or higher oligomers, where two or more protein chains form a dimer or higher oligomers by exchanging an identical structural element between the monomer. 3D domain swapping mechanism was employed as a generic way to describe the evolution of proteins from monomeric to oligomeric conformation [123-125]. 3D domain swapping is defined as a mechanism for forming oligomeric proteins from their monomers by exchanging identical or similar subunits [126]. The swapped region can be an entire domain or a helix or β-strand or loop regions. Protein structures involved in 3D domain swapping phenomenon are distinct from the rest of the oligomers due to the signature-swapping phenomenon. Yet, they are extremely diverse based on their primary sequence and secondary structures and belong to different protein domain families and structural classes [123, 126-128]. Two-hallmark feature of proteins in swapped conformation is the swapped region and hinge region. Swapped region is the region in a particular chain of protein swapped in to the adjacent chain of the protein in oligomeric state. In swapped region the intermolecular interactions will be replaced by intramolecular interactions. Hinge region is the small segment of residues mostly in loop conformation that divides the structural core and swapped region. Although domain swapping is an important mechanism for controlling multi-protein assembly, it has also been suggested as a possible mechanism for protein misfolding and aggregation. Protein structures in swapped conformations are reported to initiate pathological conformations in prion proteins and human cystatin C [128]. They are reported to aggregate same type of proteins to generate aberrant structures. For example, amyloidogenic proteins like cystatin C and prion proteins have been shown to form dimers by exchange of sub-domains of the monomeric proteins. 3D domain swapping phenomenon is interesting not only due to its pathological conformation factor; it is also important due to a wide range of functions mediated by the proteins in swapped conformation. It has been reported as a mechanism for dimer formation in odorant binding proteins and has also been proposed as a possible mechanism for fibril formation. Several well-studied examples for domain swapping events have been reported. For example, bovine seminal ribonuclease is a natural domain-swapped dimer that has special biological
properties, such as cytotoxicity to tumor cells. Barnase, a domain swapped trimer, is an enzyme that acquires enzymatic activity by cyclic domain swapping. For example, diphtheria toxin [126], RNase [129-131], histamine H1 receptor [132] spectrin (cytoskeleton), antibody fragments, human prion protein (implicated in various types of transmissible neurodegenerative spongiform encephalopathy) [133, 134], human cystatin C (implicated in amyloidosis and Alzheimer’s disease) [128, 135] and SH3 domains (important molecule in signal transduction) [136] are shown to be having 3D domain swapped segments with crucial functional roles. Figure 1.8 depicts the structure of bovine seminal ribonuclease in 3D domain swap conformation.

1.6 Computational approaches to study protein domains

Sequence data and the size of sequence databases are rapidly increasing in exponential rate in this post-genome era (Figure 1.9) [3, 62]. Genomes and metagenomes are sequenced at a rapid rate with efficient sequencing technologies, faster algorithm and rapidly reducing sequencing cost. These efforts are generating a huge data-to-information-to-knowledge inference paradigm in biology due to extensive efforts required for the functional characterization of proteins encoded in the sequenced genomes using biochemical experiments [31, 32, 49, 63, 64, 94, 100, 106, 137]. This trend is visible from the comparison of sequence statistics available in UniProtKB/Swiss-Prot (524420 protein structures as on January 2011) with the UniProtKB/TrEMBL (12788857 protein sequences as on January 2011). TrEMBL is having more members compared to Swiss-Prot, but the Swiss-Prot annotations are based on curated data, where as TrEMBL annotations are based on in-silico annotations [7, 8]. With the advent of plethora of sequence data, efficient computational approaches and analysis pipelines are being developed to deal with next-generation genome sequencing and further downstream analysis [138]. After the sequencing efforts, a primary approach is the annotation of the proteins involving a homology search using Basic Local Alignment Search Tool (BLAST) suite of programs to identify remote homologs [139]. BLAST is a widely used tool in molecular biology related research projects. BLAST algorithm which use a substitution matrices like Block Substitution Matrix (BLOSUM-80 or BLOSUM-63), Point Accepted Mutation (PAM-30 or PAM-70) etc to perform dynamic programming based heuristic search to find true matches between the query sequence and sequence database to identify highest-scoring segment pairs (HSPs). Such sequence search approaches can be enhanced further with the application of sequence search techniques based on sequence alignment and sequence profiles, which can be used to connect a new sequence
with a known sequence. Identification and analysis of specific protein domain families reported in integrated protein domain databases like SMART [140, 141], Pfam or Interpro [19] are subsequent step after sequence database search. From a structural perspective bioinformatics tools are employed to study the structural aspects of protein like structure-based sequence alignment, identification of functional sites, identification of structural motifs, binding pockets to study structural properties (secondary structure [142], electrostatics properties [143], hydrogen bonds [144], disulphide bonds [145], hydrophobic interactions, ionic interactions, aromatic–aromatic interactions, aromatic–sulphur interactions and cation–π interactions [146] etc), and to study protein docking studies (protein-ligand, protein-protein or protein-nucleic acid docking) [147, 148] and molecular dynamics [109, 149-152]

1.7 Bioinformatics Resources for the identification and analysis of protein domains

Bioinformatics databases, tools and methods play an important role in the identification and classification of proteins based on protein domains and it also introduced several new concepts to understand protein domains with better clarity. Post genomic era is showing unprecedented growth in the molecular biology databases and it is impossible to experimentally characterize every other protein-coding gene identified by sequencing projects. Knowledge based computational approaches are important in the current scenario to provide initial prediction results about the function of proteins which can help the experimental biologist to design the experiments. Effective integration of various types of data will also help in identifying new connections that could be characterized by experimental approaches. Availability of bioinformatics resources like databases, web servers, webservices, software libraries and tools are important for the efficient analysis of protein domains [62, 64, 99]. Integration of available and new resources can also be used to understand new aspects of protein domains. Various bioinformatics databases dedicated for the classification and analysis of protein domains are available and contributed to the identification and classification of protein sequences from genome projects to their respective class of protein domains, protein families and superfamilies based on bioinformatics methods like sequence searches and profile based alignment methods. Bioinformatics link directory list 124 resources in the category of ‘Domain and Motifs’. A concise overview of selected protein domain based bioinformatics tools and resources employed in the various chapters of this thesis are provided here. A detailed account of 1052 bioinformatics tools and resources related to various aspects of protein analysis is available elsewhere. (See
1.7.1 BLAST (version 2.2.17)
The Basic Local Alignment Search Tool (BLAST) is a group of programs designed to perform heuristic search to find homologs of sequence based on local similarity between sequences. The program compares protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help to identify members of gene families based on remote homology. BLAST uses a scoring matrix (such as BLOSUM) to find high scoring matches for a given query sequence in a target database and use them as seeds for joining into large alignment by the dynamic programming. BLAST also performs some pre-processing of the query by filtering out low-complexity regions and discarding words unlikely to form high scoring pairs. BLAST package contains several distinct programs designed for specific applications. For example BLASTP compares protein sequences to sequence databases and calculates the statistical significance of matches.

1.7.2 HMMER suite (version 2.2)
HMMER is a suite of programs that implement profile hidden markov models (HMM) principles for sequence analysis. HMMER package contains several programs designed around profile HMMs. For example HMMER includes programs to query profile(s) against a sequence database (hmmssearch), samples sequence from a profile HMM (hmmemit) align sequences to a profile HMM (hmmlalign), generate profile HMM from from multiple sequence alignments (hmmbuild) and search a sequence against a database of HMM profiles like Pfam database (hmmpfam).

1.7.3 ScanProsite (version 1.17)
ScanProsite is a useful utility that allows the scanning of individual sequences using database of PROSITE pattern [153]. This tool can be used to understand the functional motifs or patterns like phosphorylation site, glycoosylation site, myristoylation site etc encoded in a given protein sequence. A dedicated Perl script ps_scan.pl (Version 1.34) can be used to scan a given sequence against the database of patterns in a given PROSITE data file (Release 20.11).
1.7.4 PSIPRED (version 2.5)
PSIPRED [154] is a popular, powerful and computationally intensive protein secondary structure prediction program. PSIPRED requires a database of non-redundant sequence to derive secondary structure from the sequence homologues. NCBI-NR database is filtered to remove low-complexity regions, transmembrane regions and coiled coil regions using ‘pfilt’ program from the PSIPRED package (Version 2.5).

1.7.5 DISOPRED (version 2.1)
DISOPRED is a tool to predict disorder regions in protein sequences [155]. DISOPRED requires a database of non-redundant sequence to derive secondary structure from the sequence homologs. NCBI-NR database is filtered to remove low-complexity regions, transmembrane regions and coiled coil regions using ‘pfilt’ program from the DISOPRED package (Version 2.1).

1.7.6 MALIGN (version 4.0)
MALIGN is an efficient multiple sequence alignment program that search for search for minimum length claudograms [156]. MALIGN accepts a set of sequence in PIR format and can use one of the 13 pre-defined scoring matrices or a custom matrices to generate an alignment with percentage identities, normalized alignment scores (NAS) and distances based on NAS scores. In MALIGN, multiple sequence alignment topologies are constructed and improved through branch swapping of claudograms. Each alignment topology yields a multiple alignment and claudograms are constructed. The most parsimonious claudogram is then assigned as the final multiple sequence alignment.

1.7.7 JOY (versions 4.0 and 5.0)
JOY is a structural bioinformatics package for enhanced annotation and analysis of protein sequence and structures [144]. The program and its associated tools (psa, hbond, sstruc) can be used to identify various structural features and use them in an intuitive 2D illustration.

1.7.8 Entrez / NCBI Protein
Entrez protein is an integrated protein sequence repository that consists of nucleotide sequence translations from annotated coding regions from GenBank, DNA Databank of Japan (DDBJ) and EMBL nucleotide sequence database as well as sequences from Third Party
Annotation databases (TPA), SwissProt, Protein Information Resource (PIR), RefSeq and PDB.

1.7.9 RefSeq
NCBI’s reference sequence (RefSeq) database is a curated non-redundant collection of sequences representing genomes, transcripts and proteins. RefSeq is an integrated resource that uses data from multiple sources, to represent the current description of sequences and annotation features like coding regions, conserved domains, tRNAs, sequence tagged sites (STS), variation, references, gene and protein product names, and database cross-references. Sequence is reviewed and features are added using a combined approach of literature and database curation.

1.7.10 Uniprot
Uniprot is a comprehensive, high-quality and freely accessible resource of protein sequence and functional information. More than 99% of the protein sequences provided by UniProtKB are derived from the translation of the coding sequences (CDS), which have been submitted to the public nucleic acid databases EMBL-Bank/GenBank/DDBJ databases of International Nucleotide Sequence Database Collaboration (INSDC). UniProt is divided in to three sections: UniprotKB, UniRef and UniParc. UniprotKB is an integrated resource based on SwissProt and TrEMBL. SwissProt is a comprehensive protein sequence database with manually annotated and reviewed entries, where as TrEMBL entries are automatically annotated and is not reviewed. UniprotKB entries are characterized by the protein sequence with additional information such as name (nomenclature), taxonomic classification and citation. Additional aspects are provided, if available: such as protein attributes, general annotations, gene ontology annotations, binary interactions, sequence features and associated references. SwissProt data are further enriched by database cross-reference to various databases (sequence, structure, interaction, genome annotation, phylogenetic, interaction, enzyme, gene expression, protein family, protein domain and pathway database) and is beneficial to study specialized aspects of protein sequences or families. UniRef databases provide clustered sets of sequences from UniProt Knowledgebase. UniParc is a comprehensive and non-redundant database that contains most of the publicly available protein sequences. UniParc avoid redundancy by storing each unique sequence only once and giving it a stable and unique identifier (UPI) making it possible to identify the same protein from different source databases. The UniProt Metagenomic and Environmental Sequences
(UniMES) database is a repository specifically developed for metagenomic and environmental data [157, 158].

1.7.11 SwissProt
SwissProt is a non-redundant, manually curated protein sequence database with high quality annotation. A UniProtKB/Swiss-Prot entry is manually reviewed and provides high quality annotation and non-redundant protein sequence database, which brings together experimental results, computed features and scientific conclusions [159, 160].

1.7.12 TrEMBL
TrEMBL computer-annotated supplement of Swiss-Prot that contains all the translations of EMBL nucleotide sequence entries not yet integrated in Swiss-Prot. TrEMBL entries are unreviewed and contains protein sequences associated with computationally generated annotation and large-scale functional characterization [160].

1.7.13 Protein Information Resource (PIR)
PIR is an integrated public bioinformatics resource to support genomic, proteomic and systems biology research and scientific studies. It is the oldest curated protein resource established in 1984 by the National Biomedical Research Foundation (NBRF) that includes comprehensive well-annotated non-redundant data. The annotation based on structural, functional and experimental data extracted from literature. Various public databases cross-refer to PIR to integrate the quality annotations in PIR [161].

1.7.14 Protein Data Bank (PDB)
The PDB archive contains information about experimentally determined structures of proteins, nucleic acids, and complex assemblies. As a member of the wwPDB, the RCSB PDB curates and annotates PDB data and provides a variety of tools and additional resources for the analysis of structural data. As of 01/18/2011 PDB archives 65430 biological macromolecular structures (proteins, nucleic acids and protein/nucleic acid complexes). Where 61438 structures are elucidated using X-RAY crystallography, 8734 structures derived using NMR spectroscopy, 342 are based on electron microscopy, 30 are derived using hybrid approaches. Apart from these major categories, 151 structures are derived from methods like fiber diffraction, neutron diffraction, solution scattering, electron crystallography, infrared spectroscopy and fluorescence transfer [162].
1.7.15 InterPro
InterPro is an integrated database of predictive protein "signatures" used for the classification and automatic annotation of proteins and genomes. InterPro is an online resource that provides information about sequence classification at superfamily, family and subfamily levels and provide information about the occurrence of functional domains, repeats and important sites. InterPro adds in-depth annotation, including GO terms, to the protein signatures. Interpro records information from various databases like ProDom (for sequence-clusters built from UniProtKB using PSI-BLAST), PROSITE (for functional patterns or motifs), HAMAP (for profiles to annotate microbial proteome) PRINTS (for fingerprints, which are groups of aligned, un-weighted Position Specific Sequence Matrices (PSSMs)), PANTHER, PIRSF, Pfam, SMART[140, 141], TIGRFAMs, Gene3D and SUPERFAMILY (for hidden Markov models (HMMs)) [163-165].

1.7.16 Pfam
The Pfam database is a large collection of protein domain families, each represented by multiple sequence alignments and hidden Markov models (HMMs). Pfam database is divided into two levels depending up on the quality of the families as Pfam-A and Pfam-B. Pfam-A is derived from the UniprotKB derived sequence database ‘Pfamseq’. Each Pfam-A family consists of a curated ‘seed alignment’ containing a small set of representative members of the family, profile Hidden Markov Models (profile HMMs) built from the seed alignment, and an automatically generated full alignment, which contains all detectable protein sequences belonging to the family as defined by profile HMM searches of primary sequence databases. Pfam-B families are un-annotated and lower quality automated alignments generated automatically from the non-redundant clusters from ADDA database [99, 104-106, 166]

1.7.17 SMART
SMART (Simple Modular Architecture Research Tool) is a comprehensive database of protein domains. The database provides Hidden Markov Models (HMMs) derived from high quality manually curated alignments for each family, thereby facilitating search of protein domains and domain architectures in sequence databases. Normal SMART, the database contains Swiss-Prot, SP-TrEMBL and stable Ensembl proteomes. In Genomic SMART, only the proteomes of completely sequenced genomes are used; Ensembl for metazoans and
Swiss-Prot for the rest. Current version of Genomic SMART provides information from 630 genomes. The protein database in Normal SMART has significant redundancy, even though identical proteins are removed. If using SMART to explore domain architectures or to find exact domain counts in various genomes, Genomic mode is more appropriate. The numbers in the domain annotation pages will be more accurate, and there will not be many protein fragments corresponding to the same gene in the architecture query results [141, 167].

1.7.18 ProDom
ProDom is a database of protein domain families automatically generated from the UniProtKB. Clustering homologous segments using a procedure called MKDOM2 that utilize recursive PSI-BLAST searches populates a protein domain family in PRODOM database. Each entry in the PRODOM databases provides a multiple sequence alignment of homologous domains and a family consensus sequence [168, 169].

1.7.19 PROSITE
PROSITE is a database that consists of documentation entries describing protein domains, families and functional sites as well as associated patterns and profiles to identify them. In the latest version of PROSITE 1599 documentation entries, 1308 patterns, 912 profiles and 902 ProRule [170]. PROSITE is complemented by ProRule, a collection of rules based on profiles and patterns, which increases the discriminatory power of profiles and patterns by providing additional information about functionally and/or structurally critical amino acids. It is apparent, when studying protein sequence families, that some regions are better conserved than others during evolution. These regions are generally important for the function of a protein and/or for the maintenance of its three-dimensional structure. By analyzing the constant and variable properties of such groups of similar sequences, it is possible to derive a signature for a protein family or domain, which distinguishes its members from all other unrelated proteins. Protein signature can be used to assign a newly sequenced protein to a specific family of proteins and thus to formulate hypotheses about its function. PROSITE currently contains patterns and profiles specific for more than a thousand protein families or domains. Each of these signatures comes with documentation providing background information on the structure and function of these proteins [75, 171, 172].

1.7.20 Conserved Domain Databases
National Centre for Biotechnology Information (NCBI)-Conserved Domain Database (CDD) is a collection of multiple sequence alignments representing conserved protein domains. It is populated with data imported from Pfam [173], SMART [140, 141] and Cluster of Orthologous Groups (COG), thus, one of the most comprehensive repositories of protein domains. Protein query sequences may be scanned against position-specific scoring matrices (PSSMs) derived from the representative conserved domain alignments, employing reverse position-specific BLAST (RPS-BLAST), a variant of the PSI-BLAST program [139]. A useful accessory called Conserved Domain Architecture Retrieval Tool (CDART) is employed to scan for protein sequences with similar domain architectures [174, 175].

1.7.21 Structural Classification of Proteins (SCOP)
Structural Classification of Proteins (SCOP) is a comprehensive repository of all proteins of known structure according to their evolutionary and structural relationships. Protein domains in SCOP are grouped into species and hierarchically classified into families, superfamilies, folds and classes. SCOP data has been instrumental in the development of sensitive sequence based structure prediction algorithms and embedded in several structural bioinformatics tools and established as significant component in function annotation of new proteins in the post-genome era [103, 176, 177].

1.7.21 CATH
CATH is a manually curated classification of protein domain structures. Each protein has been parsed into structural domains and assigned into homologous superfamilies (groups of domains that are related by evolution). This classification procedure uses a combination of automated and manual techniques, which include computational algorithms, empirical and statistical evidence, literature review and expert analysis. The CATH database is a hierarchical domain classification of protein structures in the Protein Data Bank (PDB, Berman et al. 2003). Only crystal structures solved to resolution better than 4.0 angstroms are considered, together with NMR structures. All non-proteins, models, and structures with greater than 30% “C-α only” are excluded from CATH. Protein structures are classified using a combination of automated and manual procedures. There are four major levels in this hierarchy: Class, Architecture, Topology (fold family) and Homologous superfamily. Each level is described below, together with the methods used for defining domain boundaries and assigning structures to a specific family [178, 179].
1.7.22 Catalytic Site Atlas
Catalytic Site Atlas (CSA) is a database that documents enzyme active sites and catalytic residues in enzymes of 3D structure based on a classification of catalytic residues. Current version of CSA contains 251776 catalytic sites derived [180]

1.7.21 STRING
STRING is a database of known and predicted protein-protein interactions. STRING archives protein-protein data derived using six approaches (neighborhood, gene fusion, co-occurrence, co-expression, experiments, databases and text mining. STRING can be used to understand functional protein association networks and the data can be utilized to study functional networks in genome scale. Current version of STRING integrates functional level protein-protein interaction data for 2,590,259 proteins from 630 organisms [181-183]. Various curated, protein-protein interaction databases (Human Protein Reference Database (HPRD) [184, 185], Database of Interacting Proteins (DIP) [186], BioGRID [187] etc) are currently available, but STRING is a convenient integrated resource that integrate data from multiple resources based on genomic context, high-throughput experiments conserved coexpression data and data from various protein-protein interaction databases and predicted interactions using text mining algorithms. STRING derive functional network information from multiple approaches, still every single interaction is scored using a confidence score. This gives a higher advantage to filter specific interactions based on biological context (for example human protein-protein interaction with a confidence score >0.7 from experimental approach) and thus you can reduce the false positive rate. Another interesting aspect of STRING is the predicted interactions that are not reported in DIP or HPRD; such predicted interaction provides interesting connections that may lead to new biological insights.

1.7.23 DOMINE
DOMINE database offers a large of collection experimentally verified and predicted domain-domain interaction at the level of Pfam [99, 104, 106] or Interpro domains [188, 189]. Experimentally verified with inter-domain and intra-domain interactions are derived from protein structures deposited in PDB and mapped via iPfam. In addition to structure-derived interactions, DOMINE integrates data from 14 resources (iPfam [190], 3did [191, 192], ME, RCDP, P-value, Interdom [193], DPEA, PE, GPE, DIPD, RDFF, K-GIDDI, Insite, DomainGA, DIMA). Domain-domain interactions in DOMINE are scored based on three different confidence levels such as HCP (High-confidence prediction), Medium Confidence
Prediction (MCP) and Low Confidence Prediction (LCP). Current version of DOMINE contains 26,219 domain-domain interactions out of which 6,634 (gold-standard positives) are inferred from PDB entries (the union of the sets of interactions from iPfam [190] and 3did [191]), and 21,620 are predicted by at least one out of the 13 computational approaches [188, 189].

1.8 Pathway Databases
Biological pathways provide a global view of multi-step biochemical reactions, which are connected together by contributions by various biological molecules (proteins, nucleic acids and small molecules) [194]. Metabolic pathways and signal transduction pathways are typical examples of biological pathways. Pathguide [195], a database of biological pathway database provides information about 325 biological pathway related resources and molecular interaction related resources. KEGG Pathway database [196], Reactome [197], WikiPathways [198] are some of the widely used biological pathway databases.

1.9 Gene Ontology and Gene Ontology Annotations
Gene Ontology is an important, community driven bioinformatics initiative started in conjunction with the release of the first draft of human genome with the aim of standardizing the representation of gene and gene product attribute across species and major biological databases. Gene Ontology (GO) project provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data from GO Consortium members and tools to access and analyze the ontology and annotation data [199]. GO produces sets of explicitly defined, structured vocabularies that describe biological process, molecular function and cellular component of gene products. A biological process is a series of events accomplished by one or more ordered assemblies of molecular functions (for example: pyrimidine metabolic process or alpha-glucoside transport). Cellular component is a component of a cell, but it may be part of some larger object; or an anatomical structure (e.g. rough endoplasmic reticulum or nucleus) or a gene product group (e.g. ribosome, proteasome or a protein dimer). Molecular function describes activities, such as catalytic or binding activities, that occur at the molecular level (catalytic activity, transporter activity, or protein binding). The GO project has developed three structured controlled vocabularies that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner. As of ontology version 1.1725, retrieved on 24/01/2011, 33407 ontology terms (where 20188 belongs to the biological
process category, 2796 terms in cellular components and 8933 terms of molecular functions) have been stored. The ontologies are structured as directed acyclic graphs (DAG), which are similar to hierarchies but differ in that a more specialized term (child) can be related to more than one less specialized term (parent). Gene Ontology Annotation (GOA) provides high-quality GO annotations to proteins in UniProtKB, NCBI and Ensembl [200]. A GO annotation consists of a GO term associated with a specific reference that describes the work or analysis upon which the association between a specific GO term and gene product is based. Gene Ontology annotations are generally accompanied by evidence code to indicate how the annotation to a particular term is supported. GO or GOA does not provide a direct way to derive function of sequence or structural domains. Protein based ontologies like Protein-Feature Ontology [201], Protein Ontology [202], Sequence Ontology(SO) [203] and several database identifier based mapping resources are introduced to deal with this lacunae. External2go provides inter-database mapping of GO terms via to database identifiers provide a convenient way to integrate GOA with protein domains. GO project provide mapping to various protein databases for pfam2go (mapping of GO terms with Pfam identifiers), interpro2go (mapping of GO terms with InterPro identifiers), smart2go (mapping of GO terms with SMART identifiers), prosite2go(mapping of GO terms with PROSITE identifiers) are some of the mapping available from external2go.

1.10 Bioinformatics approaches for the identification of domains

Integrated approaches in bioinformatics have become an important step in the process of knowledge discovery in life science. Analysis of proteins based on the evolutionarily conserved protein domains offers a distinct advantage to understand the possible functional and structural role of proteins. Function association can be performed using fast and effective sequence searches. Protein domain based approach can be employed to identify new putative members of protein domain family from hypothetical proteins and enhanced annotation of genes with unknown function [31, 32, 49, 63, 64, 94, 100, 106, 137]. A typical computational analysis workflow of a protein domain level analysis begins with homology searches and further analysis using the alignment. Protein domains can be analyzed in different level using sequence or structural data. Various tools are available to analyze the domains encoded in the sequence or structure data. Ensembles of bioinformatics resources are available for the analysis of protein domains. Identification of sequence domains is computational intensive approach are generally error prone due to the higher degree of homology between sequence members and careful assessment of statistical parameters are required to delineate between
false positive and true positives. Several algorithms and approaches are currently available for the identification of protein domains from a given sequence. HMMPFAM that used Pfam database and hidden markov modeling approach is an example of sequence domain identification algorithm. Hmmpfam is part of HMMER package, which can be used to search a database in the format of profile hidden markov models (such as Pfam [99, 104, 106]) to identify statistically significant domains that match to the query sequence.

The Conserved Domain Architecture Retrieval Tool (CDART) performs similarity searches of the NCBI Entrez Protein database based on domain architecture, defined as the sequential order of conserved domains in proteins. The algorithm finds protein similarities across significant evolutionary distances using sensitive protein domain profiles rather than by direct sequence similarity. Proteins similar to a query protein are grouped and scored by architecture. CD-Search interface provide an interface to search the Conserved Domain Database with protein query sequences. It uses RPS-BLAST, a variant of PSI-BLAST, to search a set of pre-calculated position-specific scoring matrices (PSSMs) with a protein query. The results of CD-Search are presented as an visual representation with annotation of protein domains on the user query sequence and can be visualized as domain multiple sequence alignments with embedded user queries. High confidence associations between a query sequence and conserved domains are shown as specific hits.

Several structural bioinformatics algorithms are developed to delineate structural domains based on different properties of protein structure. Domain Identification ALgorithm (DIAL) is a seminal attempt to develop a structural bioinformatics algorithm that defines domains boundaries using various structural properties [204, 205]. DIAL considers Secondary structure data derived from SSTRUC program from JOY package [144], which implements the algorithm of Kabsch and Sander to define secondary structures [206]. $C_\alpha$ distances between secondary structures are represented in the form of average values termed "proximity indices" and these indices are used to perform clustering using KITSCH to understand secondary structure organization in the form of dendrograms. Specific nodes in these proximity-derived dendrograms are considered as tertiary structural clusters of the protein. A ratio of the average proximity indices to the average of all proximity indices in the structure, weighted for the aggregation of small subclusters and termed the disjoint factor, is
employed as a discriminatory parameter to identify automatically clusters representing individual domains.

1.11 Machine learning approaches for analysis of protein domains

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. Soft computing based data mining techniques like support vector machines (SVM) [207], Random Forest [208], probabilistic modeling [209], association rule mining [210], clustering techniques [211-213], evolutionary computation (EC) [214], fuzzy logic (FL) [215], artificial neural networks (ANN) [216], genetic algorithms (GA) [217], logistic regression [218], nearest neighbor [218], decision tree (DT) [219], self-organizing maps (SOM) [220], swarm intelligence (SI) [221], Swarm Particle Optimization [222] and approaches that combine several of the aforementioned methods to achieve better prediction accuracy are employed to solve specific problems in bioinformatics [122, 223-234]. Earlier studies have indicated that machine learning based approaches and associated web servers were useful for analyzing various properties of protein structure and function. Machine learning approaches are implemented to identify and predict various aspects of protein domains or specific class of proteins involved in particular molecular function. For example, learning approaches are applied for the prediction of neural network based secondary structure prediction [235], drug-target interaction networks [236], prediction of odorant binding proteins [237], prediction of MHC binding peptides [238], prediction of metabolic stability of proteins [239], protein sub-cellular location prediction [225], protein pathway networks prediction [240, 241], tight turns [242], predicting the network of substrate-enzyme-product triads [243], protein structural class prediction [244-247], protein quaternary structural attribute [248], membrane proteins and their types [44], functional class of enzymes [45], prediction of trans-membrane helices in proteins [249] G-Protein coupled receptor (GPCR) classes [250-253], different types of protease [254], protein cleavage site prediction [255-257], signal peptide [258], prediction of secondary structures [259, 260], protein 3D structure prediction based on sequence alignment [52] prediction of cyclin protein sequences [261, 262], prediction method for virulent proteins in bacterial pathogens [263], sequence-based prediction of DNA-binding residues in DNA-binding proteins [264], fold level classification [265], structure based prediction of protein-protein interaction [266] etc are some of the examples of the application of various machine learning techniques for specific biological applications.
1.12 Bioinformatics Software libraries

Bioinformatics software libraries provide convenient toolkit for developing programs for the analysis of protein domains. Bio-* toolkit libraries (BioPerl [267], BioJava [268], BioPython [269], BioRuby [270] & BioPHP) are currently important part of large-scale bioinformatics projects [267, 271]. Bio-* libraries provide programming interface for access, retrieval and parsing of various bioinformatics programs to analyze protein domains [267-270, 272]. These software libraries are used for developing programming pipelines and bioinformatics workflows described as part of this thesis work.

1.13 Web servers in Bioinformatics

Web servers play a pivotal role in providing easy to use bioinformatics software to the community via web interfaces. A typical web server requires limited computational resource in the client side to perform computationally intensive tasks through the Internet. Web servers are relevant in the analysis of macromolecule structures: for example, in the latest release of Bioinformatics Links Directory [10], 856 resources are reported under the category of proteins that explains the importance and availability of web-based tools for protein sequence and structure analysis from the bioinformatics community [273].

1.14 Bioinformatics web services

Data interpretability is a key element of bioinformatics. Interpretability between tools and applications can improve the integration of available resources. Web services provide a unified methods and protocols for data and resource interoperability. Web service is a collection of protocols and standards used for exchanging data between applications or systems using Hypertext Transfer Protocol (HTTP) and eXtensible Markup Language (XML) serialization [274-277]. Software applications written in various programming languages and running on various platforms can use web services to exchange data over computer networks like the Internet in a manner similar to inter-process communication on a single computer. Interoperability between systems (e.g., between Java and Python, or Windows and Linux applications) is enabled through open technology standards like Simple Object Access Protocol) or REST (REpresentational State Transfer). Data from major biological databases are currently available via Web services or Web API. Web services enable the access of the raw data from public database via programmatic access. A normal web-based bioinformatics database will be available only through traditional web interface; this is often a bottleneck for bioinformatics data analysis. Bioinformatics based web services are highly useful in such
scenarios to access the required data via API. Various public bioinformatics resources like NCBI, EBI, STRING, KEGG, RCSB-PDB are providing programmatic access to the data and interoperability via web services [278, 279]. These web-services are used to perform various data integration task associated with this chapters discussed in this thesis.

1.15 Integration of bioinformatics resource for identification and analysis of protein domains

Understanding of protein domains as structural, functional and evolutionarily conserved units has a strong relationship with the inception of computational approaches in biology [31, 32, 49, 63, 64, 94, 100, 106, 137]. Margaret Dayhoff’s initiative to catalogue the protein sequence data [31, 56, 280] fuelled the data archive revolution in biology, which scaled into new heights by the availability of low-cost sequencing technologies. Integration of bioinformatics resources is now possible due to the open access nature of scientific resources in bioinformatics. Apart from widespread adoption of software and databases open access to the bioinformatics also enabled the integration of resources can provide new avenues to understand molecular interactions and aid the design of new experiments to identify interesting molecular players. Large-scale data integration, data mining and semantic approaches in bioinformatics could accelerate such endeavors. Integrated approaches in bioinformatics have become an important step in the process of knowledge discovery in life science [15]. Several new bioinformatics protocols and resources are developed for the analysis of protein domains based on the integration of tools and databases.

1.15.1 Integrated Web Server (IWS)

Integrated Web Server (IWS) is an integrated bioinformatics is a compilation of 40 different databases, web servers and web interface for various programs related to protein sequence and structure analyses clustered as ten modules. IWS is developed as part of this thesis as an easy-to-use web server, to carry out various levels of protein domain analysis using sequence and structure data. WS provides various tools and database related to protein sequence and structure analysis classified into 10 different modules. IWS provides the tools and database under 10 different modules: Database and Servers, Sequence Retrieval and Search, Alignment, Sequence Analysis, Secondary Structure Prediction, Structure Analysis, Protein Modeling and Structure Validation, Sequence-Structure analysis, Phylogeny and Fold Recognition. Some of the major programs and databases available from IWS are PSI-BLAST [69, 139], CASCADE PSI-BLAST [69], PHYLIP [281], SEQPLOT [9], JOY [144], MODIP
[282], SCANMOT [283], MODELLER [284], HARMONY [285], PASS2 [286, 287], DSDBASE [282] etc. More than 40 bioinformatics resources for protein sequence and structure analysis is available from IWS. (Figure 1.10) illustrates a flowchart that explains about different databases and tools available from IWS. IWS is running on a CentOS-Apache server. Front-end of IWS is developed using HTML, Perl script, CGI script, and Java scripts. Back-end is a combination of different programs developed using different languages like FORTRAN, C library (GD), C++, and Perl. IWS is an ideal example of integrating various bioinformatics resources under a common theme of analysis [9].

1.15.2 SEQPLOT
SEQPLOT is a web-based graphing tool developed as part of this thesis to visualize amino acid indices based properties of a protein sequence. SEQPLOT can be used to generate AAINDEX based plots using query sequence and window size. SEQPLOT can simultaneously generate 3 different plots using 516 amino acid indices from AAINDEX database (Figure: 1.11) [9]. SEQPLOT was further integrated into 3DSwap Knowledgebase (See Chapter 3 and [9]) and PeptideMine server (See Chapter 8 and [288]) developed as part of this thesis.

1.15.3 Prediction of Unassigned REgions in proteins (PURE)
PURE (Prediction of Unassigned REgions in proteins) is an integrated method for assigning domains to the unassigned regions in protein sequences using effective sequence search methods. PURE protocol utilizes the concept of intermediate sequence search (ISS) with domain assignment using sequences from homologous space of unassigned sequence. PURE method assigns a Pfam domain to a given unassigned region with the help of connecting sequences using a frequency-based percentile. An indirect connection between the query and distantly related domain is established through a powerful procedure using PSI-BLAST hits, which are individually routed through a rigorous hmmpfam (from HMMER suite) search against Pfam database. Globplot [289], Disopred2 [155], Pepcoils [290], Tmap [290], PSIPRED [154], Scanprosite [153, 291], PSI-BLAST [139], CD-HIT[211] and Hmmpfam [292, 293], PSI-BLAST output is enabled using MView [294], BioPerl and Bio::Graphics [267]. A graphical overview of the PURE method is provided in Figure 1.12. PURE Server examines unassigned regions for the presence of disordered regions, coiled coils, transmembrane helices, appropriate extent of predicted secondary structural content and presence of homologous sequences before the assignment of probable structural domains.
PURE is an example of integrating various tools to design a new application that can help in formulating and testing new hypothesis related to protein domains [10, 295].

1.15.4 HARMONY
HARMONY is a protein structure validation tool developed to assess the quality of a protein structural domain using substitution and propensity scores derived from homologous sequence space. Protein structure validation is an important step in computational modeling of protein domains and structure elucidation of protein domains. Stereochemical assessments of protein structural domains examine internal parameters such as bond lengths and Ramachandran (φ, ψ) angles. Gross structure prediction methods such as inverse folding procedure and structure determination especially at low resolution can sometimes give rise to models that are incorrect due to assignment of misfolds or mistracing of electron density maps. Such errors are not reflected as strain in internal parameters. HARMONY is a structural bioinformatics method that examines the compatibility between the sequence and the structure of a protein by assigning scores to individual residues and their amino acid exchange patterns after considering their local environments. Local environments are described by the backbone conformation, solvent accessibility and hydrogen bonding patterns [285]. HARMONY server is developed to validate protein structural domain using the HARMONY method using protein structure files as the input. Scores are mapped on the structure for subsequent examination that is useful to also recognize regions of possible local errors in protein structures. HARMONY algorithm is developed as a web based software for the quality of protein structure and integrating HARMONY algorithm with tools like PSI-BLAST, JOY package and molecular visualization toolkit MOLSCRIPT. An example output generated using MOLSCRIPT embedded in HARMONY server is provided in Figure 1.13. HARMONY output which consists of 3 parts 1) HARMONY scores of the query protein on a calibration plot using known structures 2) Graphical plot which provides the smoothened scores between query sequence in comparison with the reverse sequence 3) MOLSCRIPT [296]. Harmony output for PDB IDs 1ABP, 1ABE and 1CRN are provided in Figure 1.14. HARMONY web server is an elegant example of integrating various applications around an algorithm to develop a new structural bioinformatics tools [285].

1.16 Identification and analysis of domains in proteins
The brief discussion on various aspects pertaining to three categories of protein domain
asserts that identification and analysis of various properties of protein domains using computational approaches can improve the understanding of various underlying biological principles. As experimental approaches cannot scale up to the current demand to identify, characterize and analyze protein domains new databases, algorithms, protocols, web servers and analysis approaches are need of the hour. This thesis performed in-depth bioinformatics based analysis of 3 classes of protein domains: “sequence domains”, “structural domains” and “interacting domain”. In this thesis, various methods pertaining to data mining, data integration, machine learning, algorithm development, bioinformatics protocol development, web server development, database development and statistical analysis approaches were employed to analyze 3 classes of protein domains. Chapter 1 provides an overview of protein domains and various aspects of 3 classes of protein domains. A cursory view of computational approaches available for analysis of protein domain and introduces the importance of new bioinformatics protocols, algorithms, databases and analysis strategies required for the identification, classification and analysis of protein domain. Chapter 2 is focusing on first category of protein domains ‘sequence domains’ and explains the importance of sampling protein sequence families to identify a single representative member as a generic problem for large-scale data mining experiments and discuss about a new novel method for identifying the best representative member (sequence or profiles) from protein domain families [297]. The computationally expensive method is then applied to protein domain families in Pfam (version 22 [99]) and the results are compiled into a database. Chapter 3, 4, 5, 6 and 7 are dedicated to structural domains. Chapters 3-6 are focusing on the unique protein structural phenomenon ‘3D domain swapping’. Chapter 3 discusses about a new “structure based literature curation” method and a new curated database of proteins involved in the unique structural phenomenon 3D domain swapping (K. Shameer et. al; manuscript submitted to Database: The journal of biological database and curation). Chapter 4 discuss about the application of Support Vector Machines to develop a new algorithm to predict 3D domain swapping from structure-derived features [298]. Chapter 5 discuss about a new Random Forest based algorithm developed to classify a given sequence as “3D domain swapping” or “non 3D domain swapping” [299]. Chapter 6 discuss about new insights gained from the analysis of sequence, structure and functional aspects of proteins involved in 3D domain swapping using the curated dataset (manuscript submitted to Database: The Journal of Biological Databases and Curation). Chapter 7 explains about a new structural bioinformatics tool, HOURI developed to identify and visualize higher order residue interactions from structural domains and discuss about various application of the
tools. Chapter 8 discuss about a unique method to utilize protein-protein interaction data from the perspective of a novel concept of ‘interacting sequence space’ and discuss various aspects of the concept [300, 301]. An integrated bioinformatics resource using protein-protein interaction, domain-domain interaction, Gene Ontology, protein domain and motif search to identify potential protein-peptide binding associations is developed using the conceptual background and the predictions were shown to have comparable with experimentally verified molecular connections in protein-peptide interactions [288]. Chapter 9 provides a detailed summary of the thesis with possible future directions. Sequence data and the size of sequence databases are increasing at a constant rate in this post-genome era. Genomes and metagenomes are sequenced at a rapid rate with efficient sequencing technologies, faster algorithm and rapidly reducing sequencing cost. These efforts are generating a huge data-to-information-to-knowledge inference paradigm in biology due to extensive efforts required for the functional characterization of proteins encoded in the sequenced genomes using biochemical experiments. Computational approaches discussed in this thesis will be recognized as important areas in the coming years and open new avenues for further research in integrative biology.
Figures of Chapter 1:

Figure 1.1: Conceptual diagram of a eukaryotic genomic region. Figure is prepared by adding additional information using figures from URL: http://en.wikipedia.org/wiki/Gene
Figure 1.2: Primary structure and tertiary structure of Crambin (PDB ID: 1CRN)
Figure 1.3: Schematic representation of various types of domains
Figure 1.4: Representation of sequence and structural domains a) Sequence of AXN1_HUMAN protein (862 residues) b) Conserved domains and motifs in AXN1_HUMAN derived using hmmpfam search c) Secondary structure assigned to the region 74-220 of AXN1_HUMAN using DSSP derived from PDB d) Crystal structure of the region containing...
Figure 1.5: a) Domain architecture of AXN1_HUMAN derived using Pfam 22 and hmmpfam (HMMER 2.2) b) Protein domains, low complexity regions and unassigned regions in AXN1_HUMAN
Figure 1.6: Protein-Protein interaction network of a human protein RGS7BP, interaction data was retrieved from STRING client using Cytoscape
Figure 1.7: Visualization of domain-domain interactions mediated by RGS domain (Pfam ID: PF00615). Interactions between interactants are not shown here. Domain-domain interaction data is obtained from DOMINE (2008 release) [189] and visualized using Cytoscape [302, 303].
Figure 1.8: Structure of Bovine seminal ribonuclease with 3D domain swapping (PDB ID: 11BG [304]) a) Individual chains are colored in cyan and green b) Individual chains are colored in cyan and green. Hinge region is highlighted in red and swapped region is highlighted in coffee-brown color.
Figure 1.9: Growth of sequence and structure databases (retrieved from http://www.genome.jp/en/db_growth.html on January 26, 2011) Used with permissions form GenomeNet database team
Figure 1.10: Tools and databases integrated in IWS (Reproduced with permissions from Bioinformation. 2007; 2(3): 86–90)
Figure 1.11: SEQPLOT output for crambin sequence (PDB ID: 1CRN) using amino acid indices Hydrophilicity scale (Kuhn et al), Hydrophobicity (Jones) and Absolute entropy (Hutchen)
Figure 1.12: HARMONY scores provided for a multi-domain protein (Arabibnose binding protein) and Crambin. Crystal structure of L - Arabinose binding protein has been reported at 2.4Å resolution initially (PDB code 1ABP [305]) and subsequently at a higher resolution of 1.7Å (PDB code: 1ABE [305]). Harmony scores mapped on Crambin structure are provided in lower panel. No significant structural errors were observed on this structure.
Sequence from unassigned region of a protein

Step 1: Sequence filtering
- Trans-membrane domain
- Coiled-coil region

- (Sequence length > 70 & Sec. Structure => 15%)

Step 2: Homology Search & Homology Reduction
- (PSI-BLAST search & CD-HIT)

Step 3: Pfam Domain assignment
- (hmmpfam)

Step 4: Frequency Score calculation
- Domain frequency calculation

Step 5: Assign domain or report unique domain regions

Figure 1.13: Integrated workflow of PURE algorithm