CHAPTER 1

Introduction
1.1 INTRODUCTION

Proteins, one of the fundamental entities of the cell, are the most diverse and efficient molecular machineries which drive the numerous molecular functions required continuously for the seamless maintenance of any living being. Evolutionary constraints bring in various protein sequence and structural modifications which tweak their functional abilities, thus making them fit to adapt to various environments and lifestyles of diverse species. Structurally similar proteins could diverge at the sequence level by undergoing mutations which poses a computational challenge to discern such relationships given mere sequence information. This difficulty becomes an even more challenging task in this era of genome sequencing, where the protein sequence repertoire is expanding exponentially unlike the already established structural-functional space. While experimental biochemical assays, structure prediction tools like X-rays and NMR (Nuclear Magnetic Resonance spectroscopy) help in the determination of structure, function and characterization of proteins, it is expertly aided by bioinformatics approaches like functional annotation by homology, comparative modeling, docking and molecular dynamics approaches.

The basic unit of protein structure, termed as protein domain, is an independent folding unit and bears a discrete component of the overall biological function [1–4]. Functional diversity of these domains among various organisms has been enabled by evolutionary changes, such as amino acid substitutions or insertions or deletions, occurring in these protein domains. The ability of some protein folds to tolerate large changes in sequence and length has been noted earlier and such length changes have been caused during evolutionary drifts [5]. Protein domains in primary structural databases such as PDB (Protein Data Bank) [6] has been grouped according to structural hierarchy such as protein folds, superfamilies, families and species in derived databases like CATH (Class, Architecture, Topology, Homologous superfamily) [5] and SCOP (Structural Classification of Proteins) [7,8]. There are also secondary databases like PASS2 (Protein Alignments organized as Structural Superfamilies), which follows the SCOP hierarchy and provide highly accurate structure-based sequence alignment for protein domain superfamilies [9–12].

Within this structural classification of protein domains, domains that have similar tertiary structure and retain generally similar functions, despite low sequence identity, are grouped together as a superfamily. In this thesis, the dataset at the superfamily level [Figure 1.1] has been utilized to extract variations in length observed within the conserved structural scaffold among various superfamily members belonging to different taxa [Figure 1.2]. Within protein
superfamilies, some members are close relatives having >40% sequence identity and perform similar functions but other members may be distantly related, having <30% sequence identity (referred to as “remote homologues”), but still possess common fold and function [13,14]. Detection of such widely diverged remote homologues from sequence space cannot easily be performed using direct sequence homology and often require sequence profile based approaches. Inclusion of these distant protein relatives is essential as they often contain length variations and provides a more comprehensive dataset for overall analyses.

1.2 COMPUTATIONAL RESOURCES USED FOR PROTEIN SCIENCE

1.2.1 Databases: Organization of biological and bioinformatics data through curation and maintenance of bioinformatics databases is crucial for efficient analysis of protein structures, sequences and domains. These resources contribute to the classification and analysis of protein sequences from genome projects to their respective class of protein domains, proteins families and superfamilies. Knowledge-based computational approaches are essential in the present era of huge biological data turnout to provide initial prediction and hypothesis about protein function and mechanisms which will be helpful for experimental biologists.

1.2.1.1 Protein Data Bank (PDB): It contains information about experimentally determined structures of proteins, nucleic acids, their complexes and assemblies [6]. The RCSB PDB curates and annotates PDB data on a weekly basis and provides an array of tools and additional resources for the analysis of structural data. Presently, there are 101741 biological macromolecular structures available (on 19 July, 2014) with more than 90000 structures elucidated using X-ray crystallography, 10453 derived using NMR spectroscopy, 804 derived by electron microscopy and rest using various methods viz., hybrid methods, neutron diffraction, electron crystallography, fiber diffraction, infrared spectroscopy and fluorescence methods.

1.2.1.2 Structural Classification of Proteins (SCOP): It is a comprehensive evolutionary classification of proteins with known structures and provides detailed description of relationships between them [8]. The classification is performed hierarchically with protein domains with similar sequence and structures (having common ancestry) being grouped into protein families, those with lower sequence similarity, conserved structural similarity and far evolutionary relationships being grouped into superfamilies. Further, based on their geometrical relationships they are grouped into folds and finally based on the secondary structural composition they are grouped into the highest levels of classes. It is similar to CATH [3] and acts as a dictionary for tracing structural similarity based on evolutionary aspects within protein structures. It has been
widely used in developing and evaluating sensitive sequence search algorithms and also in functional annotations of new proteins in the genomic era.

1.2.1.3 PASS2: This database is the fourth update of a derived structural database and has been used as the seed dataset in this thesis [12]. It is in correspondence with SCOP “superfamily” definitions, but considers only distantly related proteins (less than 40% sequence identity between its members) having a common evolutionary ancestor. It also provides highly accurate structure-based sequence alignments for protein domain superfamilies.

1.2.1.4 PFAM: Pfam (Protein family database) (Finn et al., 2008) is a highly comprehensive collection of protein domains and families which are represented as multiple sequence alignments and as profile Hidden Markov Models. Pfam families are divided into two categories, Pfam-A and Pfam-B. Pfam-A families consists of a curated seed alignment with a small set of representative members of a family, profile HMMs built from the seed alignment and an automatically generated full alignment containing all protein sequences of a family. HMMPFAM module from HMMER package, which reads a single sequence and compares it against all the HMMs for significantly similar sequence matches, is used for creation of Pfam-HMMs. Pfam method spans entire domains, including conserved motifs as well as less conserved regions like insertions and deletions. HMM-profile method allows variable conservation and insertions/deletions to be dealt in a fairly robust way [15,16].

1.2.1.5 Protein Information Resource (PIR): It is one of the earliest public integrated resource for genomic and proteomic research [17–19]. It assists researchers in the identification and interpretation of protein sequence information from 1984 and has collaborated with EBI and SIB to form UniProt in 2002. It offers a wide variety of resources mainly oriented to assist the propagation and standardization of protein annotation: PIRSF [19], iProClass [20], iProLINK [21]. Presently, it hosts iProClass which offers descriptions of all UniProt proteins and has highly informative links to databases of protein family, function, pathway, interaction, modification, structure, genome, ontology, literature and taxonomy.

1.2.1.6 Phylogeny and ALIngment of homologous protein structures (PALI): This database contains structure-based sequence alignments (by using STAMP) and dendrograms (by using PHYLIP) of homologous protein domains of various protein families [22]. Alignments of protein domains of known three-dimensional structure from PALI which are integrated with homologous sequences from UniProt database are also available for every family present in PALI database. All the entries in are in direct correspondence with SCOP 1.75C and has been annotated with the SCOP id at family level (e.g. globins annotated as a.1.1.2). The present version contains 2220 multi-member families and 1582 orphans (single-member families) consisting of about ~20,000
domains along with 2,00,000 pair-wise and 2220 multiple structural alignments for all multimember families. It forms a useful resource to help in analysing the relationship between sequence and structure variation at a given level of sequence similarity as well as forming a benchmarking dataset to evaluate performance of sensitive remote homology search algorithms.

1.2.2 Sequence Search Analysis: Divergence of two sequences from a common ancestor can be easily detected by identifying the similarities and dissimilarities between sequences with respect to the ancestral sequence. Earlier researchers used dynamic programming (Needleman and Wunsch, 1970; Smith and Waterman, 1981) for aligning the two sequences, where similarities and dissimilarities (match and non-match) were represented as numerical values obtained from substitution scoring matrices, which contained information about the likelihood of residue exchange. Popular scoring matrices which were determined by the nature of sequence compared included PAM [23], BLOSUM [24], Gonnet [25] and JTT matrices [26]. Recently, heuristic approaches like BLAST and FASTA are used for their usage of minimal computation time and comparable accuracy. Further, usages of profile-based searches and Hidden Markov Models (HMM) have also been employed for increasing sensitivity and specificity in sequence search areas.

1.2.2.1 FASTA: It is a heuristic DNA and protein sequence alignment software package designed for initially for protein sequence similarity [27]. The algorithm efficiently identifies regions of similar sequence and then scores the aligned identical and differing residues in those regions by means of an amino acid replacability matrix.

1.2.2.2 Basic Local Alignment Search Tool (BLAST): It is a group of programs utilized to perform heuristic searches in sequence/structure databases to find homologues based on local similarity between sequences. It uses a scoring matrix (BLOSUM, PAM) to find local high scoring matches for a query protein sequence in a target database and attempts to optimize a specific similarity measure by aligning the target and query sequences. It also calculates the statistical significance of matches in the form of terms like E-value (expect value). The Expect value (E-value) is a parameter that describes the number of hits one can "expect" to see by chance, when searching a database of a particular size. BLAST can be used to infer functional and evolutionary relationships between sequences as well as in identifying members of gene families based on remote homology searches.

1.2.2.3 PSI-BLAST: It is a sensitive profile-based sequence search program, which is useful for identifying distant relationships from divergent families of proteins [28,29]. Firstly, a database search is initiated using a query sequence where the identified homologues are used to create sequence profile and the homologues are further used to re-build profile to identify distant
homologues. The entire procedure is iterated till it gets converged and no new homologue is detected. The accuracy of the homologues identified is measured by E-value, alignment score and sequence identity with the query. The h-value specifies which hits are used to build the profile for the next iteration of PSI-BLAST. These types of database searches, using position-specific scoring matrices called profiles, offer better detection of distant relationships than simple BLAST searches.

1.2.2.4 Cascade PSI-BLAST: Sequences diverge more rapidly than their structures and this property of structural comparison is utilized to link remote homologies. Proteins whose sequences have sequence similarities of <30% are termed as remote homologues and can be connected using the transitivity concept of intermediate sequences where a third sequence, whose sequence characteristics are intermediate between the two matched, is used [30,31]. The assumption in such methods is that high scores for a sequence match between the first and third sequence and between the second and third sequence implies a relationship between the first and second sequence even though their own match score is low. Park and co-workers described this approach in PDB-ISL search method which relates two distantly related proteins by collecting homologues that match both with high significant scores from a large database [32]. Profile-based sequence searches are employed to detect remote relationships between protein sequences. Cascade PSI-BLAST (Bhadra et al., 2006; Sandhya et al., 2005) is an approach to detect remote relationships between proteins by implementing intermediate sequences. Here, sequence searches are initiated using PSI-BLAST, which involves multiple rounds of iterations. This is referred as ‘first generation’ of Cascade PSI-BLAST. All the homologues identified in the first generation are considered as query sequence to initiate second generation of Cascade PSI-BLAST, to recognize homologous sequences which were not considered earlier.

1.2.2.5 HMMER: It is a software package which uses hidden Markov Models (HMMs) and profile-HMMs for sequence analysis [16,33]. It has an array of modules which are used for query profiles against sequence databases (hmmsearch), building profile HMMs from multiple sequence alignments (hmmbuild), align sequences to profile-HMMs (hmmalign) and identifying homologus protein sequences against a database of HMM profiles like Pfam database (hmmpfam) and detecting domain patterns. The version 3 of HMMER package has improved upon speed of profile-HMM searches and is comparable in computation time to BLAST package.

1.2.3 Protein Domain assignment: Protein domains diverge and are often duplicated during evolution thus implying that many proteins domains can be grouped into families having similar function [34]. Protein domains are known to carry out diverse functions spanning from catalysis
during metabolism to cell-cell recognition and hence their identification from newly sequenced genomes will aid in easy functional annotation of proteins. Other than domain identification from known structures of proteins, identification of domains at sequence level relies on the detection of global-local sequence alignments between a target sequence and domain sequences from known databases such as Pfam[34], CDD and SMART.

1.2.3.1 CDD: Conserved Domain Database (CDD) is a protein annotation resource that consists of well-annotated multiple sequence alignment models for domains and full-length proteins [35]. They are represented as position-specific scoring matrices (PSSMs) and uses RPS-BLAST for efficient and sensitive identification of conserved domains in protein sequences. It can be used to explicitly define domain boundaries, aid in finding domain architectures and provide insights into sequence-structure-function relationships.

1.2.4 Multiple sequence alignment: Protein sequences which show similarity in function and structure exhibit specific residue conservation patterns indicating evolutionary significance. It is essential to recognize such patterns of residue conservation and can be detected by creating multiple sequence alignments from many protein sequences. Position-specific information about residue usage, conservation, co-evolution and correlation can be indicated from a multiple sequence alignment. These alignments are also fundamental to similarity searches, structure modeling, function prediction, motif detection and phylogenetic analysis. Algorithms used for creation of multiple sequence alignment include widely known progressive alignment [36] and iterative refinement alignment.

1.2.4.1 CLUSTAL W: It is a commonly used progressive multiple sequence alignment method which has improved upon the Feng and Doolittle’s progressive algorithm without compromising with its speed and efficiency [37]. It exploits the fact that homologous sequences are evolutionarily related and builds up multiple alignments progressively by adding pairwise alignments of closely related sequences followed by gradual addition of distantly related sequences according to a branching order of a phylogenetic guide tree. The caveat of the algorithm is that unlike dynamic programming it cannot produce an optimal alignment and hence errors made in pairwise alignments, which feature as gaps tend to stay on further along the sequence addition procedure thus shows limited success in aligning divergent sequences.

1.2.4.2 MUSCLE (MUltiple Sequence Comparison by Log-Expectation): It is a highly efficient program to create multiple sequence alignment of protein sequences [38]. Its speed and accuracy is better than the other programs viz. T-Coffee, MAFFT, ClustalW and can align thousands of sequences with ease. Its iterative refinement algorithm includes fast distance estimation using kmer counting, progressive alignment using a new profile function (the log-expectation score),
and refinement using tree-dependent restricted partitioning. It can also be used to append new sequences to already existent multiple sequence alignment with ease. It follows the iterative refinement method which involves repeatedly dividing the aligned sequences into sub-alignments and realigning the sub-alignments to correct the errors made in progressive/initial alignment thus yielding enhanced alignment quality.

1.2.4.3 MAFFFT: It is a multiple sequence alignment program which offers a range of methods, L-INS-i (accurate), FFT-NS-2 (fast; for alignment of <~10,000 sequences), etc. which follow strategies like a) the progressive method, b) the iterative refinement method with the WSP score, and c) the iterative refinement method using both the WSP and consistency scores [39]. It works in a Unix-type of environment, is very fast and scalable but requires large CPU time for aligning unrelated and long genomic DNA sequences.

1.2.5 Phylogeny: It is the study of the history of the evolution of a species or group, especially in reference to lines of descent and relationships among broad groups of organisms. The evolutionary history inferred from phylogenetic analysis is depicted as branching, tree-like diagrams, representing the pedigree of the inherited relationships of the organisms. It also helps in understanding the relationships between an ancestral sequence and its descendants and finding evolutionary ties between organisms along with divergence between the groups of organism that share a common ancestor. These evolutionary relationships between sequences can be represented through clustering patterns in phylogenetic trees and is commonly used in comparative genomics, cladistics, and bioinformatics. Phylogenetic analysis involves four steps: alignment of sequences, determination of substitution model, tree building and tree evaluation.

1.2.5.1 MEGA: MEGA (Molecular Evolutionary Genetics Analysis) [40] is an integrated suite of tools for statistical analyses of DNA and protein sequence data from an evolutionary standpoint and mainly is used for sequence alignment, inferring phylogenetic trees. It allows comparative analysis of sequence data for reconstructing the evolutionary histories of species by inferring the nature and extent of selective forces which shape the evolution of genes and species. The 5th version of MEGA allows construction of trees using different methods like Neighbor-joining, UPGMA, Maximum parsimony and Maximum likelihood.

1.2.6 Databases for functionally important residues: FireDB and Catalytic Site Atlas (CSA) are two of the most comprehensive and detailed repository of known functionally important residues, wherein the former brings together both ligand binding and catalytic residues in one site and the latter provides the hand-annotated catalytic residue annotation for enzymes [41,42].
1.2.7 Comparative/Homology modeling: Comparative modeling techniques can reliably provide three-dimensional structure of a protein with accuracy comparable to a low-resolution experimentally determined structure. The steps required in comparative modeling are: a) Template identification by searching the structural databases like PDB, SCOP, b) Target-template alignment by using pairwise sequence alignment tools like CLUSTALW. The difficulty in accurate alignment arises at this point when the identity between target-template is <30% (twilight zone), c) Building of three-dimensional models of the target protein using MODELLER and iTASSER programs, and d) Evaluation of the generated models using Verify3D, PROCHECK to check for good stereochemistry and model quality with respect to existent crystal structures [43,44].

1.2.7.1 iTASSER (Iterative Threading ASSEmbly Refinement): It is an integrated automated platform for three-dimensional structure and function prediction of proteins from a given amino acid sequence [45]. Using the technique of fold recognition, structural templates are selected and utilized for reassembling structural fragments using replica-exchange Monte Carlo simulations.

1.2.7.2 MODELLER: It is one of the most widely used tools for homology modeling of three-dimensional structure of proteins [46]. It uses restraints on spatial structure of amino acid sequence of related structures or ligands for modeling. The three-dimensional model is obtained by optimization of molecular probability density function by variable target function procedure in Cartesian space that employs methods of conjugate gradient and molecular dynamics with simulated annealing.

1.2.8 Molecular Dynamics simulations: It is the study of motions of biomolecules in a given system where in vivo conditions occurring within a system is replicated computationally using classical force fields which in turn follows Newton’s laws of physics. Simulations provide the ultimate detail regarding individual particle motions as a function of time [47]. It connects the macroscopic characteristics of a system or state functions to their microscopic description through the ergodic hypothesis. Generalized force fields like AMBER99sb-ildn, CHARMM, GROMOS or OPLS etc. uses different test cases to optimize the parameters involved and maybe biased [48–51]. For instance, AMBER is more focused on nucleic acids and DNA-protein complexes. To accurately represent the solvated system, an “explicitly solvent” water models (SPC, TIP3P, TIP4P) are introduced [52]. Molecular dynamics approach has been applied on many systems like liquids and even complex biomolecules like proteins where it can provide the ultimate details of motional phenomena [53]. GROMACS is a widely used tool to simulate molecular dynamics approach on a bio molecular system [54]. It has a plethora of force fields to select from, large number of analysis programs and is very user-friendly. It is in public domain
and distributed under the GNU General Public License. Some of the basic analyses performed viz. RMSD, RMSF, correlation analyses and essential dynamics of proteins have been applied widely on many systems [55–57]. While Root Mean Square Deviation (RMSD) of the protein with respect to its starting coordinates provides pointers regarding average amount of change occurring in the whole structure, Root Mean Square Fluctuations (RMSF) provides a per-residue or per-atom fluctuations during the simulation. Programs like VMD [58] and PyMOL [59] allow visualization and creation of simulation movies. Macromolecules like proteins are very diverse and have complex mechanisms and functional dynamics which cannot be elucidated using static crystal structures but can be estimated by molecular dynamics simulations. This facilitates usage of this approach in areas spanning from activity of enzymes to conformational transition dynamics of proteins [60,61].

Steps carried out in molecular dynamics simulations:

- Solvation of system in water box, adjustment of ion concentration and application of force fields
- Minimization of solvent using steepest descent method followed by minimization of system
- Equilibration of system at NVT and NPT until the energy values reach convergence
- Production runs of the equilibrated system for fixed amount of time are carried out, where snapshots at certain intervals of time steps are taken. The resultant data/trajectory forms the basis of further analysis of macroscopic properties and conformational changes occurring in the system.

1.2.9 Docking: It is a method to predict the preferred mode of interaction (orientation) of one molecule to another when bound to form a stable complex. It is very useful to understand the individual molecular interactions occurring between all the processes inside a living organism. Docking of protein and small molecules (ligand), often used for rational drug design, finds the optimized fit between them. Widely used and robust program for protein-ligand docking is AUTODOCK [62], which uses Monte Carlo simulated annealing, genetic algorithm and Lamarckian genetic algorithm for the automated docking. Protein-protein docking is slightly more complicated and top-ranking programs in CAPRI (Critical Assessment of PRedicted Interactions) competitions [63] like GRAMM-X [64] and ClusPro 2.0 [65] are often used. While GRAMM-X program utilizes the GRAMM Fast Fourier Transformation methodology by employing smoothed potentials, refinement stage, and knowledge-based scoring, ClusPro utilizes the rigid-body docking which generates the putative structures and further subject them to filtering and clustering methods to generate accurate docked poses.
Other sequence and structural analysis tools used in this thesis are JOY [66], CDHIT [67], PSIPRED [68], DISOPRED2 [69], MEGA [70], PRODAT loop database of SYBYL, GraProStr [71], PoreWalker [72], FATCAT [73] and DaliLite [74].

1.3 LENGTH VARIATION IN PROTEIN DOMAINS

1.3.1 Introduction to length variations

Protein domains are structural and/or functional entities of a protein chain and contribute to the overall function of the protein. It is also used as an entity to classify proteins into similar sequence and structural categories. Extensively curated databases like SCOP, CATH classifies protein domains into sequence similar families, structurally similar superfamilies and evolutionarily related fold and classes. Presence of limited number of “folds” known to mankind depicts the tendency of proteins to conserve structural scaffolds, but tolerate high sequence dispersion [75]. Thus, presence of conserved fold in a newly sequenced protein (e.g. triosephosphate isomerase (TIM) barrel, Greek-key immunoglobulin, Rossmann fold) facilitates function annotation. But folds like TIM barrel are functionally diverse and have been used as a result of years of evolution mutating the protein scaffold. The diverse functional repertoire in spite of the limited number of extant protein scaffolds is possible due to the tinkering performed at sequence and structure level which caters to the adaptations in the workings of a protein domain required for organisms leading diverse lifestyles [76] Mutations, which are like signatures of the elapsed time in proteins, are present in many forms viz. point mutations, insertions, deletions and substitutions. Other than evolution, alternative splicing is another contributor to formation of insertions/deletions (indels). Analysis of various indels in PDB also highlighted indels which have been introduced in proteins for better protein crystallization or for structural and functional studies of short peptides [77]. Variation in length caused by these insertions and deletions (indels) within homologous protein domains are studied in this thesis and the terms “indels” and “length variations” are used here interchangeably. An insertion in one sequence of an aligned pair of related sequences implies a deletion in the other, hence these insertions or deletions are referred to as indels in this study [78]. Though variations in protein sequence occur during evolution of homologous proteins but structural conservation is higher as compared to sequence and function. Thus, proteins with similar topology may have only 30% sequence similarity within them but sometimes can have diverse functions [79]. Protein length expansions, which include accumulation of functional motifs and contribute to structural innovation, are observed 40-60% greater in eukaryotes than in prokaryotes [80,81].
1.3.2 General and structural features of length variations (indels) and their impacts on protein structure, function and disease relevance in organism

In diverse protein domain superfamilies, indels are found to participate in the accretion of structural and functional features in the form of internal structural repeats amongst related domains thus leading to increased interacting proteins partners, change in quaternary arrangements, introduction of substrate specificity, increased protein stability and regulation of protein functions [82]. Indels are mostly observed as short stretches of less than 10 amino acids and occurred in mostly loop regions [83]. Studies of domain length variations by Reeves and co-workers of CATH superfamilies showed that 50% of domain was conserved even at low sequences identities of 30% [84]. But within distantly related proteins, changes in form of structural re-orientations and introduction of small structural elements are frequent and are critical in mediating structural and functional variety within homologous protein domains. In lieu of indels, protein structures may adapt to changes in shape of buried residues or to inclusion of hydrophobic residues in the core and still preserve structural integrity [85], while some indels leading to significant local changes in form of distorted turn or rotated buried residues may lead to considerable loss of activity [86]. Also, in case of remotely related proteins like TIM fold and P-loop NTP hydrolases, they adopt and retain similar structural scaffolds despite having as large as two-fold length variations within protein domain superfamilies [87].

Occurrences of indels is a continuous process and found frequently in loop regions, especially as “nested forms” into previously inserted regions [80] or observed as sub-structural regions in the minimum core of protein domain scaffold, termed as “structural embellishments”[84]. Analysis of such embellishments in HUP domain superfamily showed that indels are usually located sequentially far apart, but are spatially proximate and form subdomains which further fine-tunes the diverse functions of different members of the same superfamily [88]. Recent studies by various groups had focused on the role played by indels in affecting structure, function [82,89] and oligomeric status [90] of a protein, while emphasizing that fixation of such indels in genome which is highly context-dependent [91]. Small non-hydrophobic insertions were observed to support oligomerization at protein-protein interfaces of higher order oligomers in 12 SCOP families [92]. Introduction of short indels within active-site loops caused emergence of other enzymatic capabilities [93,94] or led to conformational switches, as in the case of C2A domain in Piccolo protein [95]. Another study of highly homologous proteins from PDB showed “plasticity” in protein structures to tolerate short internal indels, along with the fact that most indels are in disordered conformation [77]. Though they may cause structural shift in flanking regions, with a relation between the amount of structural shift and the distance to the indels [96]. Studies performed in enzyme families observed that presence of indels could easily alter
enzymatic activities and even lead to enzyme evolution by sub-functionalization [97–99]. Common consequences of indel insertions have been comprehensively studied by researchers wherein they observe local structural rearrangements, repositioning of backbone side chains, structural shifts in tertiary structures and even probable instability of secondary structural elements occurring at the sites of indels [96,100]. Length variations in B30.2(SPRY) domain of TRIM proteins have been shown to mediate innate resistance to retroviruses in selected primate species and conferring plant disease resistance by presence of large copies of distantly related resistance (R) genes [101,102]. Presence of small indels in protein coding regions are known to impact the functionality of protein and lead to disease conditions [103–105]. Hence, studying length variations in protein domains, from a sequence and structure perspective, and the location and structural properties of length variations across a protein domain is very crucial.

### 1.3.3 Length variation classification at protein domain superfamilies

Protein domain superfamilies group proteins with very low sequence identities but having same structural fold through high conservation of secondary structural elements. These protein domains (across certain protein folds) have the ability to tolerate changes in sequence and length during evolutionary drifts, without changing its folding topology [84]. Tolerance to accretion of length variation (indels) in protein domains is an intrinsic structural attribute and varies within different protein domain superfamilies [82]. Superfamilies which can tolerate large amount of length variations are referred as “length-deviant” whereas the less tolerant ones are referred as “length-rigid” superfamilies [Figure 1.3]. In this thesis, the degree of length variation for every member of domain superfamily from the mean domain size of the superfamily is expressed as the ratio of length difference of each member with respect to its mean domain size.

### 1.4 SCOPE AND SIGNIFICANCE OF STUDY

Insertion/deletion of contiguous stretch of amino acids (indels) can induce structural modifications which can eventually impart functional diversity or affect stability or differences in quaternary arrangements of these protein domains. Analysis of indels amongst sequence homologues of domain superfamilies will provide an early understanding of the evolutionary changes of protein function from a handful of protein folds [106]. Though earlier efforts have been made to study the length variations at whole protein level, it will be interesting to trace these length variations (indels) within protein domain boundaries, given the fact that protein domains are functionally independent modules [83,107,102]. Few studies pertaining to study length variations performed earlier at protein domain level have been carried out using only structural entries. Here in this thesis, I have tried to mine and incorporate the information from the vast sequence space in order to provide a comprehensive analysis. I have studied length
variations within protein domains from sequence, structure and dynamics perspective. The
various chapters are divided in three major parts:

a) Large scale: to identify length variations from sequence and structural domains for all
multi-membered superfamilies in PASS2 database,

b) Intermediate scale: to study their location and structural features in a smaller set of only
enzyme containing superfamilies and

c) Small scale: to perform behavior analysis of a length variation in one protein (TdT) using
dynamics.

This study begins with classification of structural domains in PASS2 database on the basis of the
length variations accumulated by the members of each of the multi-membered superfamilies and
also compares the transition dynamics of member domains in between two PASS2 database
updates (Chapter 2). Next, the focus is on the development of a sequence homologue collection
pipeline which provides stage-specific filters and reliably accumulates remote homologues and
further encourages sampling of length variations along with a low false positive rate (Chapter 3).
Subsequently, the collected homologues are then processed to perform indel identification and
leads to creation and analyses of a comprehensive resource of length variant protein domains at
superfamily level (LenVarDB) (Chapter 4). Further, length variations across enzyme
superfamilies are collated to create a dataset of indel positions and the structural and
environmental features of these indels are analysed by various parameters to evaluate the
functional impact on length-deviant proteins (Chapter 5). Further, molecular dynamics
simulations was carried out on length-variant isoforms of a unique DNA polymerase (TdT) at
different temperatures to understand the conformational changes occurring within a protein
structure due to the presence of an insertion (Chapter 6). Lastly, intermediate sequence search
based algorithm (Cascade PSI-BLAST) was used to detect maximum possible remote relationships
of two serine protease families (rhomboids and subtilisin). Further structural analyses were
carried out to understand the unique domain composition of the rhomboid family of intra-
membrane proteases (Chapter 8). Such a cascaded approach of remote homology search will be
useful to collect distantly related proteins which may have diverse amount of length variations
within them.

Some of these insertions/deletions (indels) can occur in close proximity to the active site residues
which can modify the direct functionality of a protein. In other cases, indels can be present on
protein surface without causing major structural and functional changes in the protein. But, their
presence at few positions, like at dimerization surfaces or on protein surfaces which causes
allosteric conformational changes, may have indirect effects. Such features make it interesting to
perform a detailed study on positions of indels in protein structure. Deletion mutants [removal of a stretch of certain amino acids] of enzymes have been performed previously extensively by researchers for finding functionally important residues or for stability studies [108–110]. However, such in vitro studies can be time-consuming and intensive and hence the demand for a tool, which can indicate putative positions where an indel can obstruct or destabilize a protein, arises.

This study differs from the previous studies in the usage of evolutionary guidance from collected sequence homologues in LenVarDB database [111], to identify these length variations (indels). Previous studies had performed pairwise alignments of only available structural entries [77]. Inclusion of sequence homologues is crucial to provide a comprehensive picture about the indels and how any mechanism related to the emergence of protein structures to perform diverse function gets limited by observations from only the extant protein folds. Knowledge about length variations in a newly discovered protein can simplify understanding of its various effects, such as if the presence of indels could hinder substrate entry by formation of a capping loop or create subdomains and form new interfaces leading to diversifying the social skills of a protein domain [Figure 1.4]. This type of study reflects the length variations found in the structural and sequence entries as it is critical to the understanding of its impact as well as the evolution of protein structure and function.

1.5 OPEN QUESTIONS IN THIS FIELD

Study of length variations in protein domains, which have been introduced by indels (insertions/deletions), is important since these indels act as evolutionary signatures in introducing variations in substrate specificity, domain interactions and sometimes even regulating protein stability. Before understanding the structural effects of length variations, their automated and accurate detection from sequence alignments needs to be improved by enhancing sensitivity of sequence alignment programs. Programs which can reliably perform sequence alignments of a vast number of sequences or can even detect alignment errors would be helpful. Further, such length variations, detected for known superfamilies, need to be verified experimentally. Such wet-lab experiments would strengthen the observations and enable better understanding of their impact on the function and activity of the proteins of interest.
FIGURES

Figure 1.1 Description of categories of protein domains in structural databases like SCOP, PASS2. An example of RmlC-like cupin superfamily has been illustrated in terms of superposed superfAMILY members and examples of few family members.

CLASS
All beta proteins [SCOP:48724]

FOLD
Double-stranded beta-helix [SCOP:51181]

SUPERFAMILY
RmlC-like cupins [SCOP:51182]

FAMILY
Hypothetical protein TM1112 [89406] (pdb:1o5u)
Homogentisate dioxygenase [51194] (pdb:1eyb)
Glucose-6-phosphate isomerase, GPI [89403] (pdb:1x82)
Figure 1.2 The database of PASS2.4 has been used as dataset in this thesis where superfamily members are distantly related (<40% sequence identity) to each other. Statistics about number of multi-membered superfamilies and total protein domain members have been depicted.

Number of superfamilies (PASS2.4 [2012]): 1961
Number of multi-membered superfamilies: 731
Number of proteins/domains in multi-membered superfamilies: 7439
Figure 1.3 Illustration of superfamilies at length deviant, length rigid and length normal groups depicted. Length variations obtained marked in green.
Figure 1.4 Variations in length within a superfamily can be as large as two-fold or more as shown by following examples. Structurally conserved scaffold has been marked in pink and the variability in length (indels) obtained from homologues of each superfamily shown in blue-red spectrum.

Lysozyme like superfamily (SCOP id: 53955)

![Lysozyme-like superfamily](image)

- **d1iiza (120)**
  - Antibacterial Protein
- **d1qusa (312)**
  - Lytic trasnsglycosylase

All-alpha NTP pyrophosphatases (SCOP id: 101386)

![All-alpha NTP pyrophosphatases](image)

- **d1vmga (80)**
  - MazG Protein (*Sulfolobus solfataricus*)
- **d1w2ya (226)**
  - dUTPase (*Campylobacter jejuni*)