A study on structural conservation of intra-chain domain-domain interfaces: learning for modeling interfaces

Rivi Verma
PH12107

A thesis submitted for the partial fulfillment of the degree of Doctor of Philosophy

Department of Biological Sciences
Indian Institute of Science Education and Research Mohali
Knowledge city, Sector 81, SAS Nagar, Manauli PO, Mohali 140306, Punjab, India.

June 2021

Dedicated to my admiring parents,

and caring husband

Declaration

The work presented in this thesis has been carried out by me under the guidance of Dr. Shashi Bhushan Pandit at the Indian Institute of Science Education and Research Mohali. This work has not been submitted in part or in full for a degree, a diploma, or a fellowship to any university or institute. Whenever contributions of others are involved, every effort is made to indicate this clearly, with due acknowledgement of collaborative research and discussions. This thesis is a bonafide record of original work done by me and all sources listed within have been detailed in the bibliography.

Rivi Verma

In my capacity as the supervisor of the candidate's thesis work, I certify that the above statements by the candidate are true to the best of my knowledge.

Dr. Shashi Bhushan Pandit

Acknowledgements

This thesis is a culmination of several years of my life. I would like to convey my sincere thanks to everyone who helped me reach here and move forward.

Foremost, I would like to express my heartfelt gratitude to my advisor Dr. Shashi Bhushan Pandit for his continuous encouragement of my Ph.D. study and research, for his patience, motivation and immense knowledge. His generous support and exceptional guidance helped me throughout the stages of this work and writing of this thesis. His scientific inputs and true mentorship over the years has helped me venture into various aspects of Computational Biology. I sincerely appreciate his faith in my capabilities that helped me in completion of the thesis work. Thank you for your unwavering support.

I am also grateful to my thesis committee members, Dr Kuljeet Singh Sandhu and Dr. Rajesh Ramachandran, who reviewed my work from time to time and provided insightful guidance throughout the research process.

I would like to thank Dr. N. Sathyamurthy, the founding and former director of IISER, Mohali for providing me an opportunity to do Ph.D. in this prestigious institute. I am also thankful to Prof. Debi Prasad Sarkar, the Director, IISER Mohali for kindly permitting me to use the excellent infrastructure for carrying out my research work. I would also like to acknowledge the faculties and students of the Department of Biological Sciences for providing a great learning platform by organizing seminars and knowledge sharing sessions. Special thanks to Prof. Anand Kumar Bachhawat and Prof. Somdatta Sinha for their constant encouragement during some of the difficult times which helped me stay focused.

These acknowledgements would not be complete without mentioning my B.Sc and M.Sc teachers, Dr. Nandita Bachhawat for teaching me some very important lessons both professionally and personally and motivated me to pursue science as a research career, Prof. F.S. Nandel for making protein structural studies interesting and easy to understand, Dr. Veena Puri, Dr. Rupinder Kaur, Sangeeta mam and Meenu mam for shaping my career during the initial years of Bioinformatics.

This work would not have been materialized without the financial support from Council of Scientific and Industrial Research-University Grants Commission (CSIR-UGC), Centre for Protein Science, Design and Engineering (CPSDE), Ministry of Human Resource and Development (MHRD) and Science and Engineering Research Board (SERB), India. I am also thankful to International Society for Computational Biology (ISCB) for giving me an opportunity to present my research work at ISMB Conference, Chicago.

I also want to thank all the members of the Computational Biology group, past and present at IISER-Mohali for their part in the inspiring working environment that existed in our lab. Thank you for many informal chats about my work and the wider field as well as making it such an enjoyable place to be. I would like to especially thank Nitish, Ashutosh and Priya for helpful suggestions in the early stages of my research and Preeti for her assistance in the common junctions of our works. A special mention must go to Dr. Nidhi for her valuable insights and contribution in finalizing my research paper and of course not to forget my all time coffee buddy. I am also thankful to Deeksha (her quick-witted nature always brought laughters), Srishti, Santosh (a cheerful person to be with), Swapnil, Arash, Ken, Harpreet (for helping with R packages), Amarinder, Meenakshi, Yachna, Nitika, Monica, Jui, Mohan, Rupesh, Purva, Shabduli and Paras for the homely environment and friendly vibes in the lab. Special thanks to Paras (Verma ji) for his undue help whenever I get stuck, be it helping with python codes, connectivity problems of laptop from home, ordering food, peeling apples ☺ and coffee breaks. Thank you to each one of you for all the happy times we enjoyed together.

Life would not have been the same without several good friends I made at IISER-Mohali. Deepest gratitude for the help and support are extended to the following persons who in one way or another have contributed in making this thesis possible. Meenakshi, Aastha, Nitika, Yachna, and Navjot mam, thank you guys for always keeping my back. I will forever cherish our memories of lunch times, coffee breaks, late night walks, movie times, parties and silly discussions. Thank you all for being there and making this journey more beautiful. I am also grateful to Isha for her valuable suggestions in my research work. I thank Sandeep, Samridhi, Ayushi, Poulami, Mayank and TJ for being great buddies. You all will be missed.

Immeasurable appreciation and profound gratitude to my parents for their love, prayers, care, and sacrifices, for educating and preparing me for my future. Your belief in me has given me the strength to reach for the stars and chase my dreams. My younger brother deserves my wholehearted thanks as well. And most of all for my steadfast husband, Param, for his enduring love, for believing in me long after I would lose belief in myself, and for sharing my wish to reach the goal of completing this task, but caring enough to love me even if I never achieved it. I would also likely to express my gratitude to my parents in-law for their unfailing emotional support. This thesis cannot be completed without acknowledging my precious bundle of joy, my four month old son Shaurya, to teach me the power of faith and true meaning of joy in life.

Abbreviations

3-D	Three-Dimensional
ASA	Accessible Surface Area
CASP	Critical assessment of methods of protein structure prediction
CAPRI	Critical assessment of prediction of interactions
CATH	Class Architecture Topology Homologous
CD-HIT	Cluster Database at High Identity with Tolerance
DDI	Domain-Domain Interface
DNA	Deoxy Ribonucleic Acid
EBI	European Bioinformatics Institute
EC number	Enzyme Classification Number
FFT	Fast Fourier Transform
FM	Free Modeling
GO	Gene Ontology
HMM	Hidden Markov Model
HMMer	Hidden Markov Modeller
iAlign	Interface Alignment
ISS	Interface Similarity Score
LPC	Ligand Protein Contact
MT	Mutant type
NDB	Nucleic Acid Database
NMR	Nuclear Magnetic Resonance
PDB	Protein Data Bank
Pfam	Protein Families
PPI	Protein-Protein Interface
PQS	Protein Quaternary Structure
RCSB	Research Collaboratory for Structural Bioinformatics
RMSD	Root Mean Square Deviation
RNA	Ribonucleic Acid
SCOP	Structural Classification Of Proteins
SIFTS	Structure Integration with Function, Taxonomy and Sequences
TBM	Template-based modeling
TM-Score	Template Modeling Score
UniProt	Universal Protein Resource
URL	Uniform resource locator
VMD	Visual molecular Dynamics
WT	Wildtype

Synopsis

In the post-genomic era, there has been significant progress in elucidating the function of individual protein to gain insights into cellular processes. However, proteins rarely act alone rather these assemble in small to large complexes to perform intricate biological functions such as signaling, transcription, replication, trafficking and other processes. A recent study suggests that ~80% of proteins interact with other proteins/ligands for their functions. Thus, for a detailed system-level understanding of cellular machinery requires knowledge of physical/functional interactions of proteins with other molecules. Towards this, years of extensive experimental and computational efforts have led to a compendium of protein-protein interactions (PPIs) or 'interactome'. Moreover, analyses of protein tertiary structural complexes have deciphered the physiochemical features of interaction between proteins as well as conservation of interaction interfaces among PPIs. Proteins can be delineated into modular regions or domains based on sequence or structure. These are usually described as functional or evolutionary units of proteins. Having multiple domains in protein facilitates complex biological function as well as enhances its efficiency by providing scaffold for functional modules. Most proteins in eukaryotes have multidomain proteins. The structural domains have been studied extensively for their occurrence, domain combinations and evolution. The function of many multidomain proteins is known to depend on the relative spatial orientation of domains. Additionally, it has been found that interaction interfaces of intra-chain domains play an important functional role such as allosteric regulation, substrate recognition, and folding/stability of proteins. Thus, understanding the physiological properties of intra-chain domain-domain interactions interfaces (DDIs) and their evolutionary conservation can provide insights into the function, engineering, and modeling of multidomain proteins. Moreover, it can also assist in understanding the evolution of interfaces. The thesis work describes a systematic and comprehensive study on the conservation of intra-chain domain-domain interfaces in multidomain proteins. Here, domains being compared vary from identical in sequence, closely related, distantly related, or completely unrelated. Further, these intra-chain interfaces were compared with inter-chain domain interfaces to enable the generation of a combined domain interface library. The CATH structural domain definitions are used in this study. Since this work spanned over years, we have different CATH release versions used across chapters. The summary of chapters is described below:

Analysis of sequence and structural properties of intra-chain domain-domain interaction interfaces (DDI's) and conservation of interfaces across multiple tertiary structures of the same protein

The interactions between proteins are mediated through their surfaces referred as interaction interfaces. Over years, extensive structural studies on the interaction interfaces have provided molecular detailed properties and suggested mutation sites (hot spots), which can disrupt protein interaction and its biological function. Despite this, analyses of interacting interfaces of domains in a protein (intra-chain) or between two proteins (inter-chain) are rather limited. In this chapter, we have compared physiochemical properties of DDIs with PPIs to find similarities between these intra- or inter-chain interfaces. Further, we investigated the extent of structural variation in domain interfaces of the same protein by analyzing intra-chain domain geometry and interfaces among structures of the same protein. In the following section DDIs, will refer to intra-chain or stated otherwise. For the comparison of physiochemical features of interfaces, we constructed a non-redundant dataset of 5137 DDIs from multidomain proteins and compared it with already known non-redundant set 1514 of PPIs. The atomic contacts are used to define interface residues as those having at least one heavy atom of a domain/protein within a distance of 4.5 Å of a heavy atom from another domain/protein. The analysis of structural features such as solvent accessible surface area, hydrogen bonds, disulphide bonds, interface size, and secondary structure content showed that interfaces share similar properties except for the interface size of PPIs is found to be larger with the relatively large non-polar solvent accessible surface area than DDIs. There were no differences in amino acid propensities between these intra- and inter-chain interfaces. Thus, suggesting in general interfaces are similar between DDIs and PPIs. To study the extent of structural variation among domain interfaces of the same proteins, we constructed a dataset of 1489 non-redundant (at 70% sequence identity) multidomain proteins having a minimum of two experimentally determined tertiary structures. Since there is a possibility that two domain interfaces of the same protein share no common residue, we defined a superset of interacting residues of a domain pair in a protein as the union of all interfacial residues of a domain pair in multiple structures of the same protein. The metric to assess structural variation among domain pairs is the root mean square deviation (rmsd) of union of interfacial residues that is computed subsequent to optimal superposition of two protein structures. The mean rmsd of interfacial residues is 1.27Å suggesting that domain interfaces are in general conserved among structures of the same protein. Moreover, most (81%) of these have rmsd less than 1Å. The relative geometry was also found to be conserved (95%) among domains. The analyses of domain interfaces with large rmsd showed that these domains undergo conformational changes upon ligand/DNA/RNA binding that involves domains. We also compared structures of wild type and mutant proteins and domain interfaces were not involved in large structural changes. This study showed that in general though domains remain invariant in their interaction, proteins known to bind ligands could involve domains in conformational change.

Structural conservation of domain-domain interfaces (DDI) and their geometry in multidomain proteins at varying levels of structural hierarchical classification in CATH

The domains sharing conserved structural features are categorized into family, superfamily, fold and class levels in structural domain classification databases. In this chapter, we analyzed whether structurally related domains (at a given level of structural relatedness) form similar interfaces. This would provide the extent of interface structural conservation between two interacting domain pairs involving closely or distantly related domains. The domain interfaces were aligned using iAlign, which is one of the best structural matching programs for interfaces. The metric of assessment for structural alignment is Interface Similarity score (IS-score), which varies from 0 to 1, with 1 being perfect structural alignment. iAlign gives p-value as the statistical significance of the alignment. We constructed a dataset of pairs of interacting domains such that two domains in a pair share structural relatedness only to a given hierarchical level in CATH. These pairs of interacting domains were aligned using iAlign and interface structural similarity of interfaces was evaluated using IS-score. Apart from this, we also compared relative domain orientation by aligning the best-superposed domain followed by finding required angle rotation and translation to superpose the other two domains. The conserved inter-domain geometry is considered with those having rotation < 20˚ and translation < 5Å. The dataset of domains (1320) sharing a maximum of 35% sequence identity (‘S’ level in CATH) resulted in mean IS-score of 0.7 (99% of these have p-value < 0.05). This shows that domain related at family level mostly result in similar interfaces. On the contrary, CATH homology related domains showed an average IS-score of 0.42 with a bimodal distribution with many domains pairs showing low IS-score. Thus, suggesting distantly related domains (homology/superfamily level) does not necessarily conserve interfaces. This could be because of constraints on domains to maintain functional form rather than conserve interfaces. At the fold and class level, as expected, average IS-score is 0.21, which is close random interface IS-score. We observed similar interfaces among class/fold related domains despite low or no similarity in domain. The analyses of inter-domain geometry showed domain pairs at family level have higher conservation of domain geometries than homology related domain pairs. We investigated functional constraints on domain interfaces of enzymes by analyzing correlation between domain interface similarity and function overlap as assessed by EC number or GO terms. The result showed that in general, interface similarity and function are not highly correlated. A high interface similarity does not lead to a similar function of two domains. We had restricted the comparison of DDIs for domains related at 'S' or homology level of CATH.

Understanding structural relatedness of domain-domain interfaces among sequence/structurally unrelated domains

In the previous chapter, we observed that unrelated domains can have structurally similar interfaces and in related work on protein-protein interactions it has been found that PPI interfaces exhibit structural similarity despite no structural/sequence relationship between aligned protein structures. This prompted us to investigate whether structural degeneracy of interfaces is a general feature among any interface. This was analyzed by structurally aligning intra-chain domain interfaces of unrelated domains. Further, these were also compared with inter-chain domain interfaces. In multidomain proteins, the interaction interface between two domains can be spatially constrained by linker region. To investigate this effect on domain interface degeneracy, we constructed three intra-chain non-redundant (40% sequence identity) domain datasets as: a) consecutive continuous domains (1511) as spatially constrained; b) non-consecutive continuous domains, (1046) as no constraints; and c) consecutive and discontinuous domains (512) as constrained by more than one linker. Since our objective is to detect similar interfaces formed by domains pairs having no structural or sequence relatedness, we generated a list of dissimilar domain pairs for each member in the dataset. The domains in dissimilar domain pairs do not share any sequence (e-value > 1) and structural similarity as assessed by TM-score (<0.4) as well as topology/fold level relationship. The domain interfaces were aligned using iAlign and the closest interface was identified as assessed by IS-score. For consecutive continuous domain dataset, the closest interfacial match has a mean IS-score of 0.307, suggesting that domain interfaces are not random as 0.2 is the mean IS-score of random interfaces. Importantly, most of these (88%) have a statistically significant IS-score. The same was observed in the other two datasets as well. Further analyzing similar interfaces revealed that this is observed due to limited possible ways of packing secondary structures and flat interface region. Next, to discern whether inter and intra-chain interfaces share similar property of interface similarity, we compared intra-chain with inter-chain domain interfaces (non-redundant dataset of 1464 domains) using the same criteria as employed in DDIs analysis. The best structural match of inter-chain interfaces among intra-chain interfaces showed that the mean IS-score is 0.311 and ~86% could find a statistically significant match to intra-chain interface. Thus, suggesting domain interfaces are structurally degenerate. Next, we investigated the connectivity of domain-domain interface structural space using a directed graph at a given IS-score. This analysis shows that DDI structural space is highly connected as ~84% of all directed interface pairs are at most separated by the eighth neighbor at IS-score of 0.26 and the largest connected component consists of 83% of interfaces. Hence, interface structural space is highly connected and degenerate.

A method to improve ranking of docked domain structures using interface constraints

From the analysis in previous chapters, we found that domain interfaces are structurally conserved considering the best structural matches based on IS-score. In this chapter, we exploited interface similarity to identify near-native interfaces among docked domain structures. In our approach, we find all possible docked poses between two domains using rigid-body docking and this list of docked complexes is ranked by IS-score by aligning them to the template interface library. This has potential application in modeling multidomain structures of proteins. In this study, we constructed a benchmark dataset of 1375 proteins from CATH v 4.1. The interacting domains were extracted and docked using rigid body docking program Z-dock. Based on Z-dock ranking, in 67% of proteins first rank docked complex is close to the native domain-domain structure. However, here domains are extracted from experimentally determined structures and the performance of first ranked docked complex would not be in the case of modeled structures. Thus, we explored re-ranking of docked complexes by aligning interface of docked complexes to interface template library of 389 proteins. This template library is constructed by clustering interfaces using IS-score as similarity metric. Since the alignment of interfaces is computationally expensive, we used additional filters to remove futile or incorrect docked complexes. These filters were protein globularity and spatial distance between the last residues of domains. These filters significantly reduced the docked complexes for alignment. The IS-score based rank of docked complexes obtained by aligning them with template interface library showed a remarkable enrichment in identifying near-native domain orientations in 67% to 90% of proteins in the top 20 best docked complexes. Thus, the interface alignment can potentially provide near native inter-domain geometries, which can be improved using other modeling tools in predicting structures of multidomain proteins.

List of publications arising from this work:

1. Verma Rivi, and Shashi Bhushan Pandit. "Unraveling the structural landscape of intrachain domain interfaces: Implication in the evolution of domain-domain interactions." PLoS ONE, vol. 14, no. 8, 2019, p. e0220336.

Table of contents

All datasets used in Chapters 2-5 are deposited in github under the repository:

https://github.com/riviverma/thesis-md-proteins/

Chapter 1

Review of literature

1.1 Essential biomolecules of forms, livings, and life

In the realm of evolution of life, proteins can truly be regarded as the Nature's "beasts of burden". The "protein" is a hetero-polymer of L-amino acids encoded by the coding region of genome that is involved in almost all cellular functions. In the past decade, genome sequencing of thousands of organisms have provided their genetic "blueprint", which could provide the 'part list' involved in various biological processes. This can facilitate understanding molecular basis of complex biological processes that usually begin with function association to proteins (Koonin & Galperin, 2003; McGuire et al., 2020). Moreover, genomics has also enabled many high-throughput experimental studies such as microarray, RNA-seq and mass-spectrometry, which provide the mRNA/protein expression of genes in a given cellular condition (Díez et al., 2012). These experimental studies in conjunction with genomic information can give a system-level understanding of biological phenomenon and importantly, has revolutionized the study of complex human diseases (McGuire et al., 2020).

Proteins perform diverse range of functions such as enzymes, transporters, signaling, structural component, hormones, immunity, and storage (Nelson, et al., 2008). Enzymes catalyze biochemical reactions both in anabolic or catabolic processes to maintain stable metabolism state as well as homeostasis of essential metabolites in cell. The structural proteins maintain the integrity of cellular components such as cytoskeleton. Hormones control or regulate specific physiological processes such as growth, metabolism and reproduction. The signaling proteins assist in transmission and/or amplification of signal generated in response to external or internal stimuli in a cell such as neuronal signaling in neurons (Nelson et al., 2008). Proteins adopt three-dimensional (3-D) structure to perform these diverse functions. This was realized with the experimental atomic structure determination of sperm whale Myoglobin in 1960as well as structures of many other proteins (Kendrew et al., 1958). This has led to the formulation of sequence-structure-function paradigm, which states that protein sequence determines its structure and structure determines the protein function. However, recent studies have found aberration to this paradigm wherein proteins without adopting unique three-dimensional structure performs their function. These are known as intrinsically disordered regions or disordered proteins, which gave alternate paradigm of disorder-function paradigm (Fuxreiter, 2018).

The availability of protein three-dimensional structures provides insights into the molecular basis of its function as well as crucial structural features such as ligand binding sites, protein interaction sites, and flexible regions. These features could be exploited in the rational designing of drug against the lead drug target (Mandal et al., 2009). Moreover, several structure-function relationships were studied because protein tertiary structure can reliably provide function of protein (Pascual-García et al., 2010). Apart from these, comparison of protein structures gives an evolutionary perspective (Chothia, 1992). Below, we briefly discuss the fundamentals of protein structure, classification and other structural features.

1.1.1 Introduction to protein structure

Traditionally, protein structure can be described based on increasing complexity at four hierarchical levels: primary, secondary, tertiary and quaternary structure. The *primary structure* is the linear representation of amino acid sequence that can be either obtained from experimental sequencing of protein or derived from *in-silico* translation of open reading frame in genomic DNA. The *secondary structure* is local conformation of protein backbone or simply local 3-D structures, such as α-helices, β-sheets, and turns. These secondary structure forms distinct *super secondary structure*, which is assembly of secondary structural elements for instance, β-hairpin, β-α-β motif, observed in proteins. *Tertiary structure* refers to completely folded and energetically stable state of the protein that represents 3-D arrangement of secondary structure (Branden & Tooze, 1999; Nelson et al., 2008). The structure is stabilized by a number of favorable interactions such as hydrogen bond, hydrophobic interactions, electrostatic interactions, salt-bridges, and disulfide bonds (Figure 1.1). The tertiary structure of protein associate with other polypeptide proteins in specific geometry and spatial orientation of these is referred to as *quaternary structure*. The protein oligomers can be of 2 types: *homo-oligomer* consists of monomeric units of same polypeptide chain, whereas *hetero-oligomer* is formed of two or more different polypeptide units.

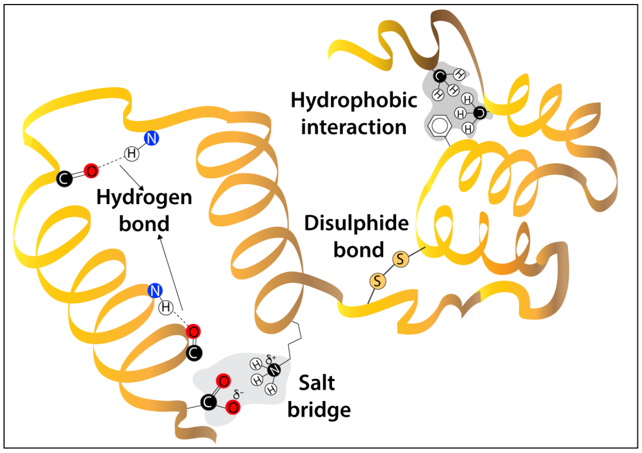


Figure 1.1 Interactions stabilizing tertiary structure of the protein. Interactions include covalent linkages such as disulphide bonds, non-covalent interactions like Salt bridges, hydrogen bond, hydrophobic interactions.

1.1.2 Experimental determination of protein tertiary structure

The atomic structure of the protein can be determined by following methods: X-ray crystallography, nuclear magnetic resonance (NMR), fiber diffraction and electron microscopy (EM). Of these, X-ray crystallography is most often employed method for structure determination. Below we briefly discuss these methods of structure determination.

1.1.2.1 X-ray crystallography

The x-ray crystallography relies on the scattering of x-rays by the electrons in the protein molecule.In this method, first the crystal of pure protein is obtained that is subjected to x-rays for recording diffraction data. The ability to determine relative coordinates of atoms is feasible only from diffraction data of protein crystals because similar structural motifs forming individual unit cell are periodically arrangedin a crystal. The diffracted rays depending on input direction show relative interference and its intensity depends on the arrangement of atoms in a unit cell (Wlodawer et. al, 2008). This diffraction data is used for constructing electron density maps, thus providing the coordinate of protein atoms. Among various ways to assess the quality of protein structure determined using x-ray diffraction data, the most common metrics are resolution and R-factor of the structure (Branden & Tooze, 1999). Resolution is a measure of quality of diffraction data (electron density maps) collected from the protein crystal, and it serves as an indirect measure of the precision to which the three-dimensional co-ordinates have been determined. The resolution of a structure is expressed in Angstroms (Å). The detailed atomic structure depends on the resolution of structure (Table 1.1). The R factor is a global measureof electron density fit of the calculated protein structure matches with experimental data. Typically, a protein structures are reported to have an R factor of around 0.2 (Laskowski, 2003;Wlodawer et al., 2008). An R factor between 0.4-0.6 can be obtained from a completely disordered structure. The limitation of x-ray crystallography is the ability to obtain stable crystals of pure proteins, which depends on many factors such as pH, solubility etc. Apart from obtaining a diffracting crystal, solving phase problem in structure determination remains a challenge.

Table 1.1 Protein structure resolution and possible interpretation (Minor, 2007)**.**

Resolution(in Å)	**Interpretation**
>4.0	Secondary structure elements can be determined.
3.0 – 4.0	Side chains are not resolved. Random main chain can be visible.
2.5 – 3.0	Major structural features such as α helices and β sheets are clearly distinguishable, but many side chains may not be resolved.
2.0 – 2.5	Number of sidechains in wrong rotamer is considerably less. Water molecules and small ligands become visible.
1.5 – 2.0	Individual side chains resolved to define specific conformers
<1.5	Individual atoms in a structure can be resolved.

1.1.2.2 Nuclear Magnetic Resonance (NMR)

The limitation of obtaining crystals to determine structure is overcome in NMR method, which can elucidate structure of globular proteins in aqueous solution. NMR is based on the spin quantum property of nuclei/proton that changes orientation (spin states) on application of external magnetic field at a resonant frequency (Marion, 2013). This resonant frequency depends on the chemical surrounding of proton/nuclei with spin property. Thus, this measurement can describe chemical nature of nuclei as well as their spatial distances with other protons. The 2D, 3D NMR spectroscopy provides inter-proton distances, which along with stereochemical constraints are used to compute the 3D structure of proteins (Sugiki et al., 2017). Since the distance measurements are slightly imprecise, it is difficult to obtain a unique structure satisfying all observed distance restraints. Consequently, structure obtained from NMR is represented using ensemble that are consistent with experimentally observed constraints. The advantage of NMR is its ability to model protein dynamics and understand flexibility of protein regions (Kovermann et al., 2016; Narayanan et al., 2017). The limitation of NMR is that the protein should remain soluble in high concentration and it is not feasible for large proteins (≈ >25kDa) (Clore & Gronenborn, 1998; Sugiki et al., 2017).

1.1.2.3 Electron Microscopy

The electron microscopy technique allows taking high-resolution images of (biological/non-biological) samples. The flash-freezing of protein solution prevents damage due to electron beam and is used produce microscopic images of individual molecule using electron microscopy (cryo-EM). Thus, produced images can be used to produce EM maps to fit protein models and generate medium to low-resolution protein structures. The EM maps allow the fitting of the atomic-resolution of individual components (domains, proteins, sub-complexes) into the lower resolution density of whole assembly (Rossmann et al., 2005; Topf & Sali, 2005). The rigid body fitting results ingeneration of the atomic structures of an entire complex (Chacón & Wriggers, 2002; Fabiola & Chapman, 2005).

EM is mostly employed for determining the structure of large multi-protein assemblies. The main advantage of EM is that it requires small quantities of sample to image protein complexes in their physiological environment. Recently, a database *Electron Micrcoscopy Data Bank* at PDBe (*https://www.ebi.ac.uk/pdbe/emdb/index.html/*) is developed that serves as a repository for multi-protein complexes (Patwardhan, 2017). According to the latest statistics report, it has released ~10,000 electron microscopy density maps. With each passing year, this method is achieving to determine near atomic-resolution structures (less than 4Å), which has been indicated in the latest release of the database as shown below in Figure 1.2. Few of the protein complex structures listed in the database includeYeast RNA polymerase I elongation complex (resolution 3.42Å), Leviviridae PP7 coat protein dimer capsid (resolution 2.89Å), Helicobacter pylori urea channel in open state (resolution 2.7Å) and many more.

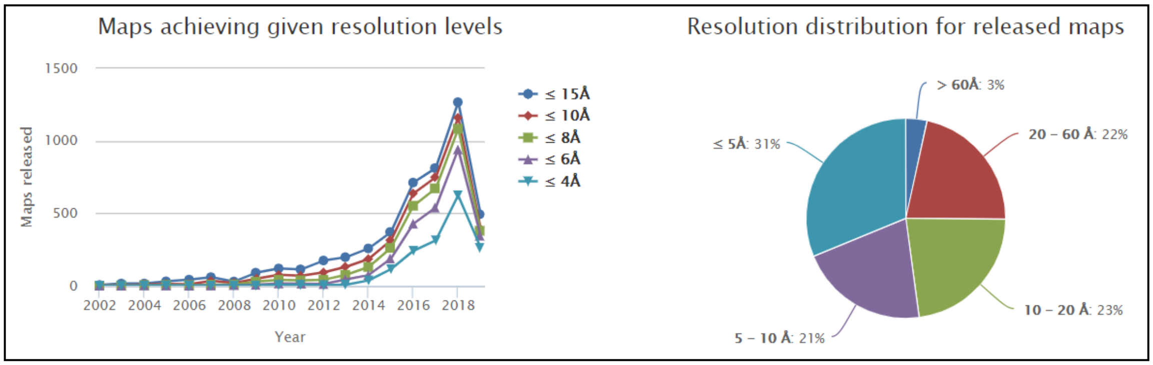


Figure 1.2Summary of growth of EM maps released with their resolutions (Figure and data source: *EMDB database*).

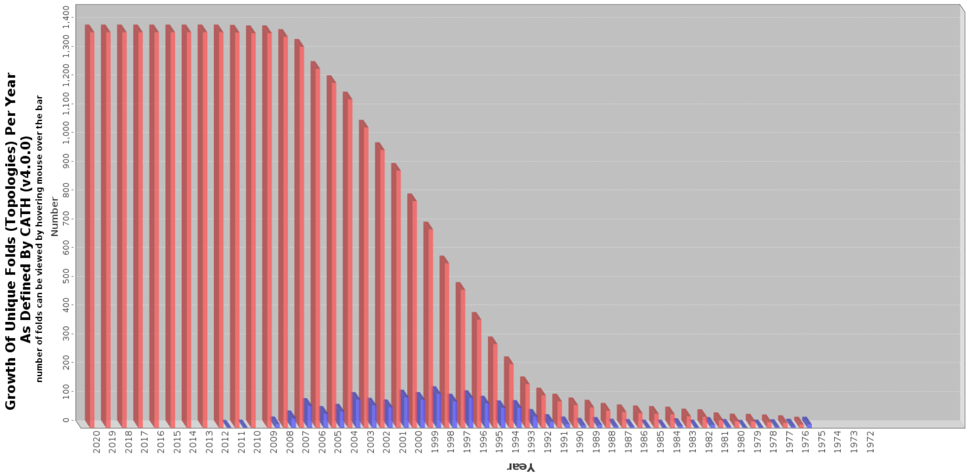
1.1.2.4 Protein Data Bank (www.rcsb.org/)

Protein structures determined using experimental methods are deposited in the central repository database, known as Protein Data Bank (PDB) (Berman et al., 2000; Bernstein et al., 1977). PDB is one of the earliest community-wide databases of biological data started in 1971 at the Brookhaven National Laboratory (Bourne & Weissig, 2003). It constitutes the largest database of solved high-resolution structures of monomeric as well as multimeric complexes bound to proteins, chemical compounds, metal ions, ligands, and nucleic acids (DNA, RNA). According to the current release of PDB (2019), there are 134588 X-ray solved structures, 12578 from NMR and 3015 using EM as experimental method. In last decade, there has been a tremendous growth in the number of protein structures deposited in PDB (Figure 1.3).

1.1.2.5 Structural Genomics Initiative

In the early 2000's a worldwide initiative Structural Genomics Initiative (SGI) of National Institute of Health (NIH) was established to decrease the ever increasing large gap between number of protein sequences, generated due to genome sequencing, and number of protein structures (Burley & Bonanno, 2002). Moreover, function of many of these genomic proteins is unknown that could be predicted using tertiary structure of protein (experimental or predicted). Since the number of sequences without known structure is overwhelming, it is not feasible to determine structure of all protein sequence. Therefore, it is essential to utilize computational prediction of structure that relies on the ability to detect protein having tertiary structure (template) for modeling of protein without known structure (Yan & Moult, 2005). The sequences can be clustered to identify representative sequence and its experimental structure can serve as template for members of a cluster. Overall, it is expected that structural genomics efforts would provide structure of representative sequence as template for modeling other genomic sequences (Levitt, 2009; Yan & Moult, 2005). Thus, structural genomics is making a great contribution in expansion of protein universe and providing new drug targets for designing drugs against fatal diseases (Grabowski, et al., 2016).

Figure 1.3 Annual growth of Protein Data Bank. The red bars show the total number of structures deposited to PDB annually and blue bars are the number of unique folds for CATH_4.0.0 added per year. (Modified and adapted from the following data source: *http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=fold-cath*).



1.2 Protein function annotation

The function of protein can be described at levels from their cellular to molecular role that usually are challenging to compare across organisms. This necessitated for a unified description of gene function and led to a collaborative initiative of Gene Ontology (GO) project in 1998 (Ashburner et al., 2000). The initial work on model organisms *viz.* yeast, Drosophila and mouse developed common schema to classify gene function that provide a comparable description across organisms. The GO is a structured, precisely defined, common and a controlled vocabulary (ontologies) to describe the roles of genes. These ontologies are arranged as a directed graph; where nodes represent the GO terms and connection between these nodes represent the specific annotation between two GO terms. Thus, making GO annotation of gene products computationally manageable, transferable and comparable across species. The functions of gene products of any organism are defined with three aspects of biological domains:

Cellular component: describes the component of the cell (or anatomical structure) where the gene product is located (Ribosome, Nuclear membrane, Proteasome etc).

Biological process: refers to the large biological process accomplished by molecular function of single or multiple genes. For example, broader terms include signal transduction, cell growth and maintenance; whereas, specific terms include translation, cAMP biosynthesis etc.

Molecular function: refers to the biochemical activities (catalytic or binding) of a gene product occur at molecular level. Broader terms include enzymes, ligand, and narrow functional terms are toll-receptor ligand.

The GO database is available at *http://www.geneontology.org/*along with exhaustive collection of various tools utilizing GO.

1.3 Enzyme Commission

The Enzyme Commission (EC) number is a hierarchical numbering system to classify enzymes based on their cognate chemical reactions (Webb, 1993). The enzyme is described using four levels EC numbers. The first number indicates the type of reaction, the second and third number indicates the chemistry that occurs, and the last number indicates the specificity of the substrate. There are seven major classes of enzymes based on the top-level of EC number and are listed in Table 1.2. Although EC numbers are computationally tractable, still these are inadequate in classifying non-enzymatic proteins or describing the cellular role of gene products.

1.4 Protein-protein interactions

Proteins do not work in isolation, rather these interact both functionally or physically to perform biological processes. A protein-protein interface can be defined as the physical interaction and has been studied extensively as these are crucially for the biological function (Bonetta, 2010). The proteins can control the flow of information in a given network, both from within and between biological processes. Moreover, it has been realized that protein interactions are important component for organism complexity and has been highly conserved in evolution (Bolser & Park, 2003; Park & Bolser, 2001). The loss or aberration in protein-protein interaction could lead to diseases (Alberts, 1998; Eisenberg et al., 2000). Therefore, implying that protein interaction are important and the interaction surfaces are under natural selection pressure to be conserved than non-interface regions (Caffrey et al., 2004).

Table 1.2 Major classes of enzymes based on first level EC number

Top level of EC	Enzyme class	Reaction catalyzed
1	Oxidoreductases	Catalyze oxidation/reduction reactions; transfer of H and O atoms or electrons from one substance to another
2	Transferases	Transfer of a functional group from one substance to another. The group may be methyl-, acyl-, amino- or phosphate group
3	Hydrolases	Formation of two products from a substrate by hydrolysis
4	Lyases	Non-hydrolytic addition or removal of groups from substrates. C-C, C-N, C-O or C-S bonds may be cleaved
5	Isomerases	Intramolecular rearrangement, i.e. isomerization changes within a single molecule
6	Ligases	Join together two molecules by synthesis of new C-O, C-S, C-N or C-C bonds with simultaneous breakdown of ATP
7	Translocases	Catalyze the movement of ions or molecules across membranes or their separation within membranes

Since studying protein-protein interactions (PPIs) are fundamental to understand flow of information in biological processes, there have been both experimental high-throughput and computational efforts to generate information on PPIs and map them at the genome level (Janin et al., 2008; Kim et al., 2004). The experimental methods used to study protein recognition mechanisms have contributed invaluably to identify and characterize protein-protein interfaces (Sharan & Ideker, 2006). With development of experimental methods and generation of wealth of information, it has been possible to document physical protein interactions, within a cellular system, commonly referred to as interactome or protein-protein interaction networks (Jeong et. al., 2001; Rual et al., 2005; Sanchez et al., 1999). This provides a network view of PPIs in a system, where the *nodes* represent the protein molecules and the interaction between these proteins is represented as network *edges*. The accumulation of interaction data has facilitated to study topological properties of such network as well as suggestion of scale-free nature of network (Barabási & Albert, 1999). According to scale-free behavior, most of the nodes in the network are connected sparsely, whereas, few nodes make most of the connections (also called hub nodes/proteins). These hub nodes interactions constitute a small fraction of whole interactome. In a study, the quaternary fold was estimated using non-redundant dataset of dimeric protein complexes from PDB (Garma et al., 2012) using a new scoring function 'rTM-score', which measures the similarity between individual monomers as well as their relative orientation in the complex. Taking rTM-score > 0.5 as a cut-off, 3629 quaternary families were clustered into 1761 quaternary folds. The largest cluster is comprised of 47 protein complexes from RNA polymerase family. Interestingly, approximately, 60% of the clusters are orphans with no other structure exist in PDB. These are regarded as single complex clusters and represented as nodes with black color in Figure 1.4. Therefore, efforts are being made to enhance the knowledge in predicting biologically relevant protein-protein interactions sites in order to fill the gap between known protein structural folds and protein complexes.

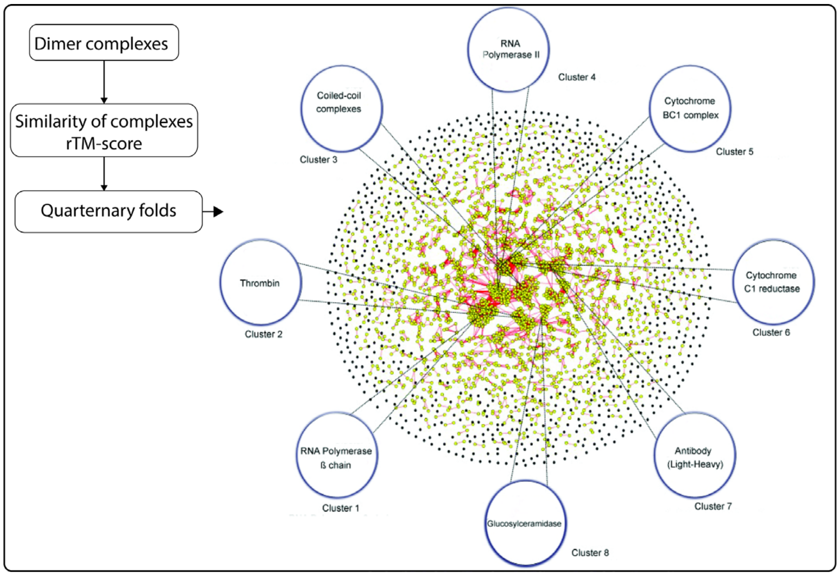


Figure 1.4 A graphical representation of non-redundant protein complexes at 90% sequence identity. The sequence of each dimer is mapped to their Pfam database. In a graph, black nodes are single complexes without any structural match. Yellow nodes represent the ones, which are connected atleast by one edge (Adapted and modified from Garma et al., 2012).

The 3-D structures of protein complexes along with experimental methods such as alanine scanning mutagenesis, mass spectrometry-based approaches have provided detailed information onPPI interfaces. Moreover, these have been used to understand the principles of protein-protein recognition. Briefly, initial work on PPIs has analyzed contribution of physical and chemical features of interaction residues in protein recognition (Chothia & Janin, 1975). Lawrence and group (Lawrence & Colman, 1993; McCoy et al., 1997) analyzed the roles Electrostatic Complementarity (EC) and shape correlation index in PPI interfaces. Several studies investigated physicochemical properties of interfaces between subunits of proteins or domains of proteins (Argos, 1988) with attempts to understand their geometric characteristics using spline function (Harder & Desmarais, 1972; Meinguet, 1979). Later, the work of Jones introduced surface patch method (Jones & Thornton, 1997), which assisted in determining the parameters contributing to the protein-protein interactions. In subsequent studies interface of PPI was dissected into core and rim region based on solvent accessible surface area and various features were characterized for these regions (Bahadur et al., 2003; Chakrabarti & Janin, 2002). Some recent studies have employed geometrical and topological methods, such as Voronoi diagrams, to study protein interfaces (Cazals et al., 2006).

1.4.1 Characteristics of protein- protein interfaces

The structural features of PPIs have been studied in detailed over decades of research that has been summarized in various reviews (Berggård et al., 2007; De Las Rivas & Fontanillo, 2010, 2012; Lehne & Schlitt, 2009; Perkins et al., 2010; Zinzalla & Thurston, 2009). We give brief overview of structural characteristics of protein-protein interaction interfaces as structures provide molecular details,which determine specificity in protein-protein recognition (Skrabanek et al., 2008). Moreover, identification of conserved interaction sites is important defining molecular description of PPI network, metabolic pathways and development of drug targets (Zinzalla & Thurston, 2009).

The initial work of Chothia and Janin in 1975 analyzed three protein complexes to examine the characteristics of protein interaction interfaces (Chothia & Janin, 1975). Their study found that interfaces are tightly packed consisting of mostly hydrophobic residues and shape complementarity is important for interaction (Chothia & Janin, 1975). Later, many studies extended analyses of PPIs on large dataset and included other structural feature to characterize interfaces. These were focused to determine distinguishing interface features from rest protein surface that could serve as key determinants to develop tools for predicting protein-protein interaction sites. Later, work of Jones and Thornton proposed a method to find strength of binding in protein complexes based on size of interface (Jones & Thornton, 1996). Here, the size of an interface is calculated as the difference of solvent accessible surface area between the complex and the separated components. This is useful in differentiating biological contacts from crystal contacts because more surface area gets buried on complex formation in large interfaces, which could be likely in case of biologically relevant complexes. Since, solvent accessible area can be related with the hydrophobic energy of de-solvation, it has been observed that in homodimers, which form tightly packed complexes are rich in hydrophobic residues, showed a linear relationship between their solvent accessible surface area and molecular weight (Samanta et al., 2002).

A score to measure the fit of interacting interface was introduced called as surface complementarity ('Sc') score, which varies from 0 (no fit) to 1 a perfect fit (Lawrence & Colman, 1993). Using this score, it found that antigen-antibody complex have poor fit (lower 'Sc' score) than enzyme-inhibitor complex (Lawrence & Colman, 1993). It was suggested that poor surface complementarity in antigen-antibody complexes possibly helps antibody to recognize a vast diversity of new epitopes on the antigens. Later, analyses showed that antigen-antibody interfaces are relatively planar than enzyme-inhibitor complexes (Decanniere et al., 2001; Jones & Thornton, 1996). The enzyme-inhibitor complexes show less planar interfaces as the catalytic residues involved in catalyzing a given reaction are located in clefts on enzyme surface. The feature of shape complementarity has been used to find the best fit between two interacting proteins using rigid- body or flexible-body searches in docking studies (Gabb et al., 1997; Lawrence & Colman, 1993; Shoichet & Kuntz, 1991).

The electrostatic interactions at the protein-protein interface contribute greatly towards the specificity of the interaction. In a study, it has been observed that the rate of protein association correlates directly with the electrostatic energy of proteins in a complex (Selzer & Schreiber, 1999; Sheinerman et al., 2000). Salt bridges are also a type of electrostatic interactions, which provides stability to protein complexes in harsh environments (Kumar et al., 2000). The significance of hydrogen bond at interfaces was examined (Janin & Chothia, 1990) showing that the number of hydrogen bonds are on average same between tightly packed enzyme-inhibitor complex and antigen-antibody complexes. The hydrogen bonds are known to confer specificity to hydrophobic interactions at protein-protein interface (Fersht, 1987).

Other studies on PPI have focused on residue composition and inter-residue contacts at interfaces (Glaser et al., 2001; Ofran & Rost, 2003). Miyazawa and Jernigan developed a method to extract inter-residue potentials from frequencies of contacts between different residues at protein interfaces (Miyazawa & Jernigan, 1985). Further, the existence of "hot-spot" residues were speculated to make the major contribution in the protein complex formation (Keskin et al., 2005). Several studies have confirmed the importance of hydrophobic forces in mediating protein-protein interactions, however, in a recent study; they have claimed that it is the hydrophilic interactions, which promote protein-protein associations (Ben-Naim, 2006).

1.4.2 Classification of protein-protein interfaces

Based on various criteria, protein-protein interactions can be classified into following types:

1.4.2.1 Based on interaction strength

Protein complexes can be classified as obligate and non-obligate depending on whether each protomer can exist independent of the other. The complexes having proteins, which cannot exist as independent stable structures are obligate complexes otherwise they are non-obligate (Jones & Thornton, 1996). The obligate protein complexes have large interfaces and strong binding energy with high shape complementarity and tightly packed interfaces. The non-obligate protein complexes are formed temporarily (for example, enzyme- inhibitor complex) under certain physiological conditions like phosphorylation. The interfaces of obligate complexes have relatively high preponderance of non-polar amino acids and non-obligate complexes tend to be more hydrophilic in nature (Jones & Thornton, 1996; Teichmann, 2002). It has also been observed that obligate complexes evolve at slower rates and are subject to constraints forcing compensatory mutations to a much greater extent than the transiently interacting ones which evolve at a fast rate (Mintseris & Weng, 2005).

1.4.2.2 Based on composition of protein complex

Protein complexes composed of identical protein chains are termed as *homo-oligomers* and those complexes formed between non-identical proteins chains are termed as *hetero-oligomers*. The homo-oligomers can further be classified as *isologous and heterologous* complexes (Monod et al., 1965). Isologous have same set of interacting residues from each interacting partner, whereas in heterologous homo-oligomer involve different set of residues from each protomer. Isologous interfaces are known to give rise to finite number of oligomeric complexes as compared to heterologous which is responsible for indefinite polymerization (for example, actin filaments). Structurally or functionally obligate interactions are usually permanent, whereas non-obligate interactions may be transient or permanent (Acuner Ozbabacan, et al., 2011).

1.4.2.3 Depending upon the lifetime of the complex

Permanent interactions are usually very stable and exist only in complexes. *Transient interactions* associate and dissociate *in vivo*. These can be weak transient interactions which form and break continuously and strong transient interactions which require molecular trigger to shift the oligomeric equilibrium (Acuner Ozbabacan et al., 2011).

1.4.2.4 Depending upon the specificity of the interaction

Protein-protein interactions can be specific, multi-specific or non-specific. In nature, most of the protein interactions are highly specific, for instance, antigen-antibody interactions. The multi-specific interactions can be such as between a serine protease and its inhibitor. And non-specific interactions are rare in nature, for example, binding of major histocompatibility complex with antigens (Teichmann, 2002).

1.4.2.5 Depending upon the biological relevance

It has been known that not all protein-protein interactions determined through X-ray crystallography methods are biologically relevant. The crystallization artifacts can cause incorrect protein-protein interactions. Hence, care should be taken while analyzing any further property of protein-protein interactions.

1.4.2.6 Depending upon the timing and spatial distribution of binding sites on protein surface

In a protein-protein interaction network, if different protein partners bind simultaneously to the same hub protein and this hub protein further possesses a unique binding site for its partners, the interaction is said to be simultaneously possible. Usually, the proteins involved in this kind of interaction are products of co-expressed genes and generally are obligate in nature. The other type of interaction is known as mutually exclusive where; protein partners are not co-expressed and bind at different times or location using the same interface. These interactions are generally transient in nature (Kim et al., 2006).

1.4.3 Protein-protein interaction databases

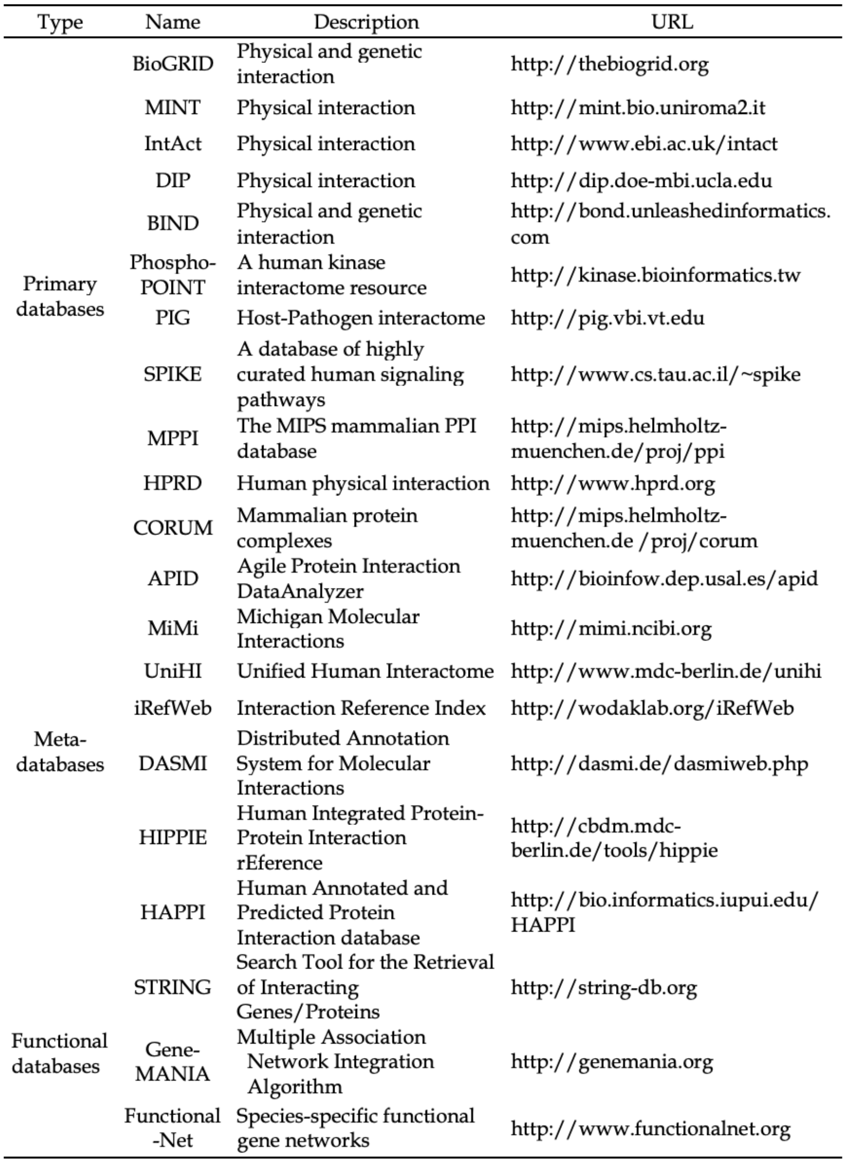
An enormous amount of data generated by computational and experimental methods are deposited in primary databases, where these are manually curated. Some of these databases integrate to form consortium for sharing and improving curated data. These meta-databases help removing the redundancy and other inconsistencies in the data (Turinsky et al., 2010). Apart from these, there are species specific databases documenting interactions (Lee et al., 2010, 2011). Some of commonly known PPI databases are listed in Table 1.3.

1.4.4 Prediction of protein-protein interactions and PPI as drug targets

Since experimental methods are usually time-consuming and technically challenging, computational methods have been proved to be useful in predicting PPI and also reliably identifying interfacial residues. Such prediction can be broadly classified into (i) knowledge based and (ii) docking based. The knowledge-based methods rely on detecting homologs of experimentally known PPIs, whereas, docking methods rely on geometric models and shape complementarity measures between protein structures. With the current expansion of PDB, target proteins are dominantly modeled using template-based approaches such as modeling based on physical processes of folding is a daunting task. Some of the known knowledge based methods available publically are mentioned in the following Table 1.4.

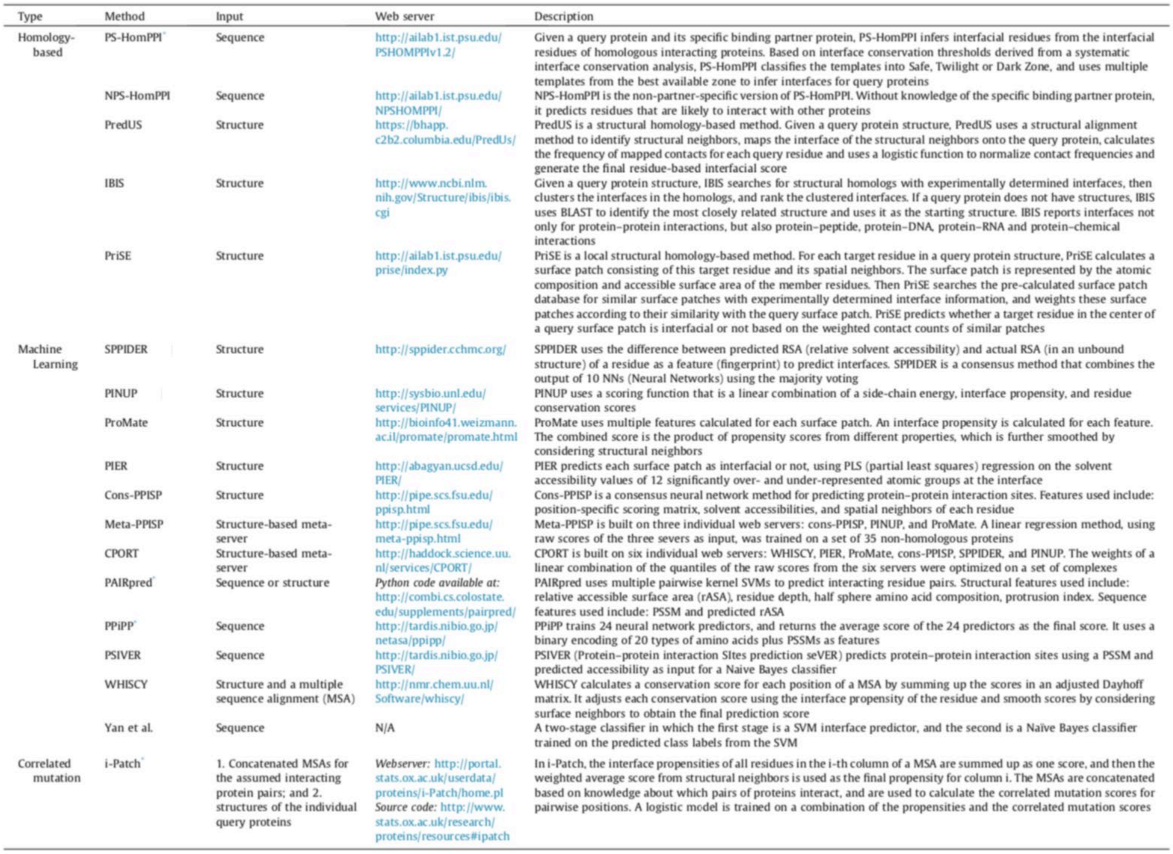
As has been discussed before, protein–protein interactions are essential to mediate various physiological processes in the cell and its aberrant activities could lead to pathological conditions (Cohen & Prusiner, 1998; Loregian et al., 2002; Selkoe, 1998). This suggests that specific PPIs can be targeted for drug development (Makley & Gestwicki, 2013; Ozdemir et al., 2019). Therefore, it is crucial to identify protein-protein interactions, which can be used for rational drug design process. The cellular processes apart from being regulated by environmental conditions can also be influenced by external compounds (Eyster, 1998; Furukawa et al., 2002; Klemm et al., 1998; Markus & Benezra, 1999). The focus is to find small molecules targeting PPI with high affinity and can regulate PPI that are referred as 'PPI modulators' affecting protein-protein interaction through stabilization of inhibition (Zinzalla & Thurston, 2009). However, there are challenges in identification of such modulators, mostly, due to topology of protein-protein interfaces such as flat in nature in comparison to other binding sites having clefts (Jin et al., 2014), lack of small molecule binding sites, false positive interfaces and diversity of protein-protein interfaces (Arkin & Wells, 2004; Gurung et al., 2017). Several computational modeling and molecular biology techniques have been developed to address these challenges (Cheng et al., 2007; Huang & Jacobson, 2010; Jin et al., 2014) and modulators could be designed to either destabilize protein-protein interaction or to inactivate protein complexes by locking complexes in a non-functional state.

Table 1.3 List of protein-protein interaction databases (Adapted from(Jung et al., 2012)



Type	Name	Description	URL
Primary databases	BioGRID	Physical and genetic interaction	http://thebiogrid.org
	MINT	Physical interaction	http://mint.bio.uniroma2.it
	IntAct	Physical interaction	http://www.ebi.ac.uk/intact
	DIP	Physical interaction	http://dip.doe-mbi.ucla.edu
	BIND	Physical and genetic interaction	http://bond.unleashedinformatics.com
	Phospho-POINT	A human kinase interactome resource	http://kinase.bioinformatics.tw
	PIG	Host-Pathogen interactome	http://pig.vbi.vt.edu
	SPIKE	A database of highly curated human signaling pathways	http://www.cs.tau.ac.il/~spike
	MPPI	The MIPS mammalian PPI database	http://mips.helmholtz-muenchen.de/proj/ppi
	HPRD	Human physical interaction	http://www.hprd.org
	CORUM	Mammalian protein complexes	http://mips.helmholtz-muenchen.de /proj/corum
Meta-databases	APID	Agile Protein Interaction DataAnalyzer	http://bioinfow.dep.usal.es/apid
	MiMi	Michigan Molecular Interactions	http://mimi.ncibi.org
	UniHI	Unified Human Interactome	http://www.mdc-berlin.de/unihi
	iRefWeb	Interaction Reference Index	http://wodaklab.org/iRefWeb
	DASMI	Distributed Annotation System for Molecular Interactions	http://dasmi.de/dasmiweb.php
	HIPPIE	Human Integrated Protein-Protein Interaction rEference	http://cbdm.mdc-berlin.de/tools/hippie
	HAPPI	Human Annotated and Predicted Protein Interaction database	http://bio.informatics.iupui.edu/HAPPI
Functional databases	STRING	Search Tool for the Retrieval of Interacting Genes/Proteins	http://string-db.org
	Gene-MANIA	Multiple Association Network Integration Algorithm	http://genemania.org
	Functional-Net	Species-specific functional gene networks	http://www.functionalnet.org

Table 1.4 Representative protein–protein interface prediction methods (Adapted from (Xue et al., 2015)



Type	Method	Input	Web server	Description
Homology-based	PS-HomPPI*	Sequence	http://ailab1.ist.psu.edu/PSHOMPPIv1.2/	Given a query protein and its specific binding partner protein, PS-HomPPI infers interfacial residues from the interfacial residues of homologous interacting proteins. Based on interface conservation thresholds derived from a systematic interface conservation analysis, PS-HomPPI classifies the templates into Safe, Twilight or Dark Zone, and uses multiple templates from the best available zone to infer interfaces for query proteins
	NPS-HomPPI	Sequence	http://ailab1.ist.psu.edu/NPSHOMPPI/	NPS-HomPPI is the non-partner-specific version of PS-HomPPI. Without knowledge of the specific binding partner protein, it predicts residues that are likely to interact with other proteins
	PredUS	Structure	https://bhapp.c2b2.columbia.edu/PredUs/	PredUS is a structural homology-based method. Given a query protein structure, PredUS uses a structural alignment method to identify structural neighbors, maps the interface of the structural neighbors onto the query protein, calculates the frequency of mapped contacts for each query residue and uses a logistic function to normalize contact frequencies and generate the final residue-based interfacial score
	IBIS	Structure	http://www.ncbi.nlm.nih.gov/Structure/ibis/ibis.cgi	Given a query protein structure, IBIS searches for structural homologs with experimentally determined interfaces, then clusters the interfaces in the homologs, and rank the clustered interfaces. If a query protein does not have structures, IBIS uses BLAST to identify the most closely related structure and uses it as the starting structure. IBIS reports interfaces not only for protein–protein interactions, but also protein–peptide, protein–DNA, protein–RNA and protein–chemical interactions
	PriSE	Structure	http://ailab1.ist.psu.edu/prise/index.py	PriSE is a local structural homology-based method. For each target residue in a query protein structure, PriSE calculates a surface patch consisting of this target residue and its spatial neighbors. The surface patch is represented by the atomic composition and accessible surface area of the member residues. Then PriSE searches the pre-calculated surface patch database for similar surface patches with experimentally determined interface information, and weights these surface patches according to their similarity with the query surface patch. PriSE predicts whether a target residue in the center of a query surface patch is interfacial or not based on the weighted contact counts of similar patches
Machine Learning	SPPIDER	Structure	http://sppider.cchmc.org/	SPPIDER uses the difference between predicted RSA (relative solvent accessibility) and actual RSA (in an unbound structure) of a residue as a feature (fingerprint) to predict interfaces. SPPIDER is a consensus method that combines the output of 10 NNs (Neural Networks) using the majority voting
	PINUP	Structure	http://sysbio.unl.edu/services/PINUP/	PINUP uses a scoring function that is a linear combination of a side-chain energy, interface propensity, and residue conservation scores
	ProMate	Structure	http://bioinfo41.weizmann.ac.il/promate/promate.html	ProMate uses multiple features calculated for each surface patch. An interface propensity is calculated for each feature. The combined score is the product of propensity scores from different properties, which is further smoothed by considering structural neighbors
	PIER	Structure	http://abagyan.ucsd.edu/PIER/	PIER predicts each surface patch as interfacial or not, using PLS (partial least squares) regression on the solvent accessibility values of 12 significantly over- and under-represented atomic groups at the interface
	Cons-PPISP	Structure	http://pipe.scs.fsu.edu/ppisp.html	Cons-PPISP is a consensus neural network method for predicting protein–protein interaction sites. Features used include: position-specific scoring matrix, solvent accessibilities, and spatial neighbors of each residue
	Meta-PPISP	Structure-based meta-server	http://pipe.scs.fsu.edu/meta-ppisp.html	Meta-PPISP is built on three individual web servers: cons-PPISP, PINUP, and ProMate. A linear regression method, using raw scores of the three severs as input, was trained on a set of 35 non-homologous proteins
	CPORT	Structure-based meta-server	http://haddock.science.uu.nl/services/CPORT/	CPORT is built on six individual web servers: WHISCY, PIER, ProMate, cons-PPISP, SPPIDER, and PINUP. The weights of a linear combination of the quantiles of the raw scores from the six servers were optimized on a set of complexes
	PAIRpred*	Sequence or structure	*Python code available at:* http://combi.cs.colostate.edu/supplements/pairpred/	PAIRpred uses multiple pairwise kernel SVMs to predict interacting residue pairs. Structural features used include: relative accessible surface area (rASA), residue depth, half sphere amino acid composition, protrusion index. Sequence features used include: PSSM and predicted rASA
	PPiPP*	Sequence	http://tardis.nibio.go.jp/netasa/ppipp/	PPiPP trains 24 neural network predictors, and returns the average score of the 24 predictors as the final score. It uses a binary encoding of 20 types of amino acids plus PSSMs as features
	PSIVER	Sequence	http://tardis.nibio.go.jp/PSIVER/	PSIVER (Protein–protein interaction SItes prediction seVER) predicts protein–protein interaction sites using a PSSM and predicted accessibility as input for a Naive Bayes classifier
	WHISCY	Structure and a multiple sequence alignment (MSA)	http://nmr.chem.uu.nl/Software/whiscy/	WHISCY calculates a conservation score for each position of a MSA by summing up the scores in an adjusted Dayhoff matrix. It adjusts each conservation score using the interface propensity of the residue and smooth scores by considering surface neighbors to obtain the final prediction score
	Yan et al.	Sequence	N/A	A two-stage classifier in which the first stage is a SVM interface predictor, and the second is a Naïve Bayes classifier trained on the predicted class labels from the SVM
Correlated mutation	i-Patch*	1. Concatenated MSAs for the assumed interacting protein pairs; and 2. structures of the individual query proteins	*Webserver:* http://portal.stats.ox.ac.uk/userdata/proteins/i-Patch/home.pl *Source code:* http://www.stats.ox.ac.uk/research/proteins/resources#ipatch	In i-Patch, the interface propensities of all residues in the i-th column of a MSA are summed up as one score, and then the weighted average score from structural neighbors is used as the final propensity for column i. The MSAs are concatenated based on knowledge about which pairs of proteins interact, and are used to calculate the correlated mutation scores for pairwise positions. A logistic model is trained on a combination of the propensities and the correlated mutation scores

The physiochemical characteristics of protein-protein interfaces were analyzed to determine druggability of PPIs. It has been demonstrated that druggable PPIs have predominant hot spots (London et al., 2013), which can be regarded as continuous small peptide contributing most to the interface (London et al., 2010). The modulators can be designed to target PPIs by stabilizing/inhibiting protein complex formations (Arkin & Whitty, 2009; Mullard, 2012; Thiel et al., 2012). Mostly, inhibitors of a PPI are designed so that small molecule competes for the intersection site, which are usually hot spots or allosteric site (Figure 1.5). In allosteric inhibition mechanism, binding of a small molecule bind at a site distant from the interface leads to conformational change resulting in inability of protein to form complexes (Shangary & Wang, 2009; Yin & Hamilton, 2005). In PPI stabilization, the small molecules bind to protein monomer and stabilize the complex by increasing their natural binding affinity (Figure 1.5c).

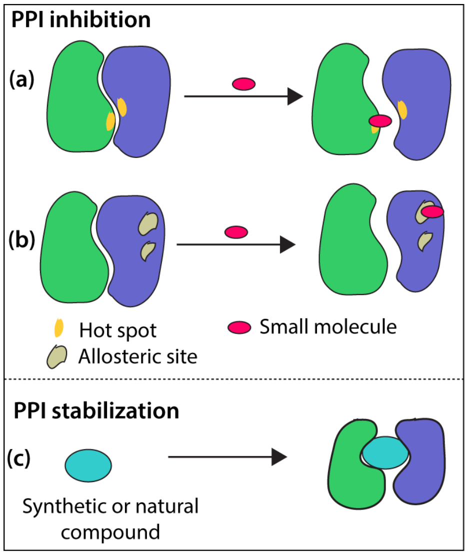


Figure 1.5 PPI inhibitions by (a) hot-spot approach, (b) allosteric mechanism and (c) PPI stabilization. The two protein monomers are represented in blue and green color. Hot spots are shown in yellow (Zinzalla & Thurston, 2009).

1.5 Protein domains

The term 'domain' was coined in 1960's and in 1970's researchers observed that there are certain patterns in protein sequence/structures appeared to repeat as sequence motifs or substructure within a structure. The structure of hen egg white lysozyme structure (Phillips, 1966) showed the existence of distinct substructure in lysozyme protein and found that there is an interior hydrophobic and somewhat hydrophilic surface in the substructures. Later, Cunningham et al. (Cunningham et al., 1971) in a separate study on immunoglobin proteins could identify distinct regions, which were referred to as domains. It was also hypothesized that these regions have evolved through evolutionary events such as gene duplication/translocation. Further, Wetlaufer (Wetlaufer, 1973), who first examined multiple proteins and compile a list of their domains. Based on this, he suggested that structural independence was largely due to rapid self-assembly of these distinct regions unlike Cunningham's work, which suggested separate genetic control for such regions. Subsequently, it was proposed that protein domains are independent folding units that form the basic 'building blocks' of proteins in evolution and architecture (Blake et al., 1967; Wetlaufer, 1973). Another similar study also proposed that protein domains are structurally self-sufficient in terms that if cleaved from protein backbone, the domain would retain their three-dimensional geometry and often their function (Levitt & Chothia, 1976). With availability of multiple structures, the most accepted domain concept is based on the globularity or compactness of the proteins, which assumes that the atomic interactions within domains are stronger than between domains. Structural domain such as protein kinase domain (Pkinase), SH3 domain (Src homology 3), and leucine zipper domain (bZIP) can exist as an independent entity in case of single domain proteins or they may exist in combination with other domains in multidomain proteins.

A large proportion of proteins coded in diverse set of organisms are made up of more than one domain where these form functionally or structurally distinct modules in multidomain proteins (Campbell & Baron, 1991). It has largely been accepted that multidomain proteins have evolved through multiple events of duplication and adaptive changes from single-domain proteins (Vogel et al., 2004; Vogel et al., 2005). It has been also proposed that modular multidomain proteins are produced by exon-shuffling during evolution (Patthy, 1996). Using an existing domain repertoire, duplication and shuffling of domains led to the emergence of numerous unique and novel functions (Vogel et al., 2005). It has been suggested that multiple domains can provide structural stability and functional advantages to proteins.

1.5.1 Types of protein domains

The protein can be delineated into domains, which depending on the concept of domain definitions can be of various types such as sequence, structural, functional, evolutionary and mobility (Majumdar et al., 2009; Postic et al., 2017). The widely accepted and well-defined types of domains are sequence and structural domains. For completeness, we provide definitions of various types of domains:

Sequence domains – are purely defined on the basis of conservations of residues over significant length of alignment and hence, can be detected using sequence similarity measures. These are often found in combination with other sequence domains and are well characterized in Pfam database (Finn et al., 2014).

Structural domains – are defined on the basis of compactness, globularity and presence of hydrophobic core. It is assumed that atomic interactions are stronger within a domain than between domains. Structural domains are classified in SCOP (Murzin et al., 1995) and CATH database (Orengo et al., 1997).

Functional domains –are defined as having independent function associated with a region of protein. These domains are capable of carrying out an activity such as enzymatic function.

Evolutionary domains –are defined as primary unit of evolution, propagated by recombination, shuffling, transposition etc.

Folding domains –are defined as independent region capable of folding independently or atleast possess a folding nucleus that can initiate a folding process.

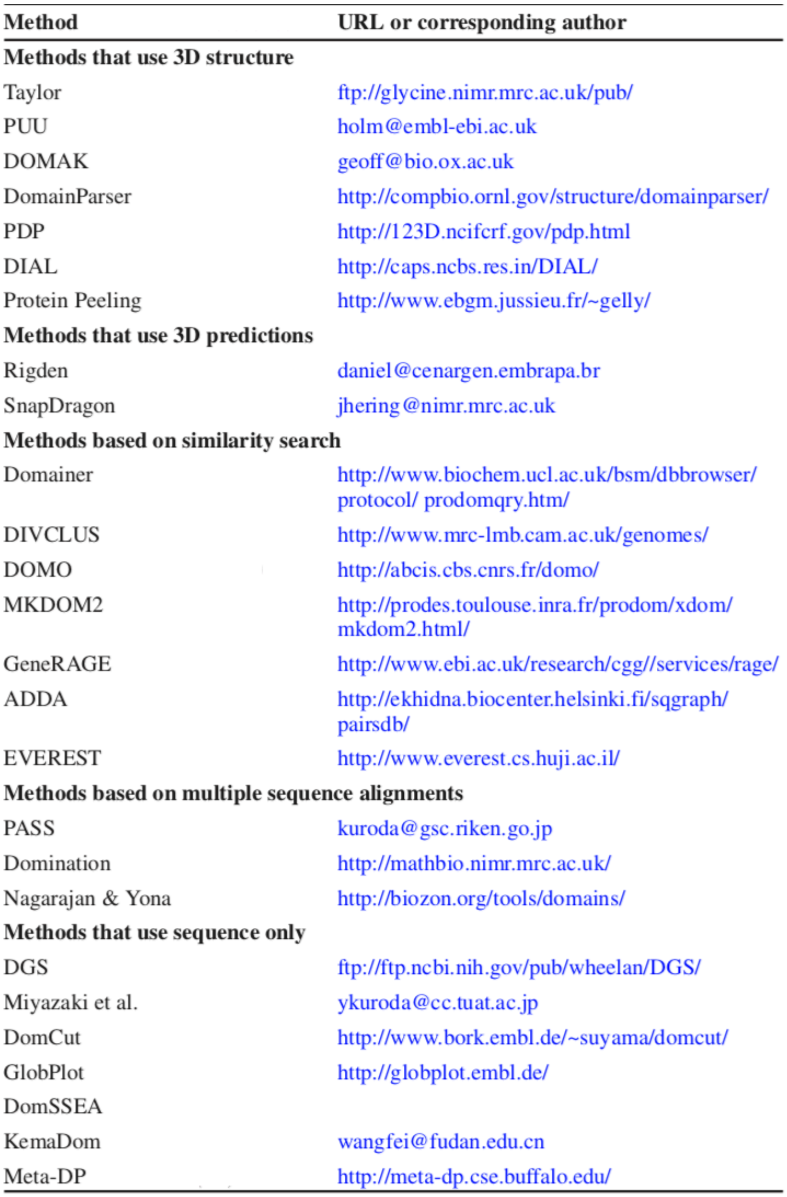
Mobility domains –are defined as domains with high correlated mobility, which rearrange during evolution of protein. These domains physically interact with many other domains or bind to different types of molecules (Basu et al., 2009).

It is important to note that the above domain definitions do not necessarily agree with each other. For instance, a compact protein structure does not correspond to a functional unit of protein and therefore, it is possible that proteins can have different valid annotations depending upon the basis used for domain annotation. In essence, domains are considered as a fundamental structural, functional and evolutionary unit of proteins. Many databases are dedicated for depositing as well as retrieval of well-annotated protein domains.

1.5.2 Structural domains

Structural domains can be broadly defined as a compact region of protein that is often, but not always, consist of continuous segment of amino acid sequence and is usually capable of folding, stable enough to exist on its own. Alternatively, it is compact, local and semi-independent units of protein structure (Richardson, 1981). The identification of protein domains is the essential step for protein structure determination and functional annotations. Several methods are known to predict domains using information from either structure or sequence of the proteins and some of them are mentioned in Table 1.5. The significance of predicting domains in proteins is to identify new putative members from hypothetical proteins and subsequently classify them in their corresponding protein domain family. Additionally, the domain identification methods will help in annotating genes with unknown function in newly sequenced genomes.

Table 1.5 List of databases of domains and methods for their prediction or identification (Ingólfsson & Yona, 2008)



Method	URL or corresponding author
Methods that use 3D structure	
Taylor	ftp://glycine.nimr.mrc.ac.uk/pub/
PUU	holm@embl-ebi.ac.uk
DOMAK	geoff@bio.ox.ac.uk
DomainParser	http://compbio.ornl.gov/structure/domainparser/
PDP	http://123D.ncifcrf.gov/pdp.html
DIAL	http://caps.ncbs.res.in/DIAL/
Protein Peeling	http://www.ebgm.jussieu.fr/~gelly/
Methods that use 3D predictions	
Rigden	daniel@cenargen.embrapa.br
SnapDragon	jhering@nimr.mrc.ac.uk
Methods based on similarity search	
Domainer	http://www.biochem.ucl.ac.uk/bsm/dbbrowser/ protocol/ prodomqry.htm/
DIVCLUS	http://www.mrc-lmb.cam.ac.uk/genomes/
DOMO	http://abcis.cbs.cnrs.fr/domo/
MKDOM2	http://prodes.toulouse.inra.fr/prodom/xdom/ mkdom2.html/
GeneRAGE	http://www.ebi.ac.uk/research/cgg//services/rage/
ADDA	http://ekhidna.biocenter.helsinki.fi/sqgraph/ pairsdb/
EVEREST	http://www.everest.cs.huji.ac.il/
Methods based on multiple sequence alignments	
PASS	kuroda@gsc.riken.go.jp
Domination	http://mathbio.nimr.mrc.ac.uk/
Nagarajan & Yona	http://biozon.org/tools/domains/
Methods that use sequence only	
DGS	ftp://ftp.ncbi.nih.gov/pub/wheelan/DGS/
Miyazaki et al.	ykuroda@cc.tuat.ac.jp
DomCut	http://www.bork.embl.de/~suyama/domcut/
GlobPlot	http://globplot.embl.de/
DomSSEA	
KemaDom	wangfei@fudan.edu.cn
Meta-DP	http://meta-dp.cse.buffalo.edu/

1.5.2.1 Continuous and discontinuous domains

As mentioned before, structural domains can be made of continuous segment or be formed from more than one segment. Evolution has led to different forms of domain arrangements across the sequence of proteins that can be classified by their connectivity (Das & Smith, 2000). These arrangements go beyond sequential permutations of entire domain structures and may affect the organization of domains. Based on the connectivity, there are two main classes in which domains can be divided: *continuous domains* and *discontinuous domains*. The continuous domains are composed of an uninterrupted stretch of amino acids in the associated coding sequence, which can be represented as $1 < i < N$. The continuous part of protein chain has a tendency to form islands of ordered structures (Henry et al., 2013; Munoz & Eaton, 1999), which may serve as nuclei for protein folding. The discontinuous domains have coding sequence interrupted by subsequences, which encode alternative structures, such as inserted domains. To represent a discontinuous multidomain protein, consider there are discontinuous segments consisting of residues $1 \leq i \leq N_1$ and $N_2 \leq i \leq N$, and another continuous domain, consisting of residues $N_1 < i < N_2$. In this way, discontinuous segments of domain can be connected more than one inter-domain linker (Figure 1.6). The discontinuous domain may not be an independent folding unit, but may depend on the continuous domain.

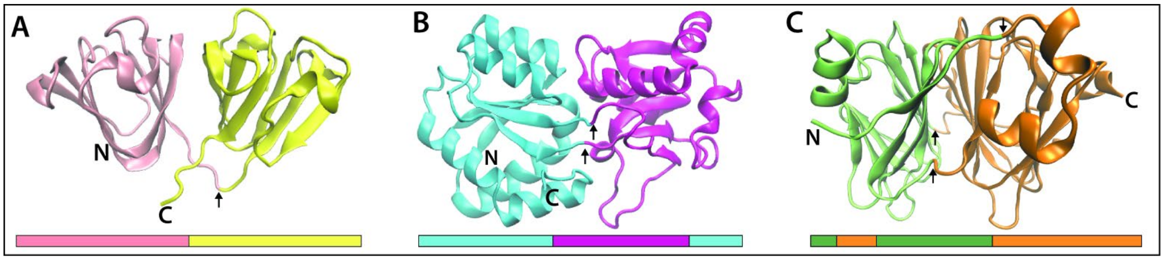


Figure 1.6 Examples of two-domain proteins with different topological complexities. (A) Human γD-crystallin (PDB ID: 1hk0A) having two independently foldable domains connected by one linker. (B) Bacterial solute binding protein ModA (PDB ID: 1atgA), which has two domains of which one domain (shown in blue) is discontinuous domain. (C) 5-keto-4-deoxyuronate Isomerase structure (PDB ID: 1xruA) consists of two domains and both are discontinuous domains. The N and C terminals of structure are shown with positions of linkers shown by black arrows. The linear arrangement of structural domains is shown in color corresponding to same color as structural domains.

1.5.3 Classification of structural domains

Even though, analysis of individual protein structure can reveal a great deal of information, over years it has been realized that a comprehensive view of proteins can be understood from comparing multiple proteins and investigating the evolutionary relationships among them. In order to perform such systematic analysis requires a system of classifying proteins into structurally related groups, which can be used to infer function as well. The field of structural classification aims to identify and characterize these domain structures by grouping them into a hierarchical manner using relationships among groups of protein based on their sequence or structural similarities.

Among many attempts to divide the protein structure universe into defined regions, two databases *viz.* SCOP and CATH, have been useful in classification of structural domains and have been maintained over decades. These are briefly described below.

1.5.3.1 Structural Classification of Proteins (SCOP) (http://scop.mrc-lmb.cam.ac.uk)

The SCOP database provides comprehensive information on structural and evolutionary relationships of protein domains (Murzin et al., 1995). Subsequent to identification of domains, mostly manually, these are classified hierarchically at four levels (Figure 1.7). The most fundamental level of classification is family, which are grouped into superfamily. These are classified into fold and then into classes. Each of these levels is defined as follows:

Family – Protein domains classified at this level have clear evolutionary relationships among each other. The sequence identity shared between proteins is usually high (>30%). Proteins are grouped based on sequence identity, structure similarity and functional similarity (for example, Globin family).

Superfamily – Protein families having similar structures or functions are grouped together into a superfamily. The sequence identities are usually low, however, structures are similar. The proteins related at superfamily are suggested to have a common evolutionary origin.

Fold – This level group superfamilies based on core protein structure. Fold level is defined as "the same secondary structure elements in the same arrangement with the same topological connections" (Bourne & Weissig, 2003). Proteins with same folds may differ in secondary structure elements in the periphery of the structure. There is no evidence of common evolutionary origin but independently evolving proteins may end up in having a similar fold due to physiochemical constraints, which favor specific secondary structure arrangements.

Class – The highest level of classification is class. It is defined by the content and organization of secondary structural elements of the domains. The predominant classes are: "all α" domains composed mainly of α-helices; "all β" domains consist of majorly β sheets; "α/β" domains having β sheets surrounded by α helices; "α + β" domains having regions separated by both helices and sheets.

The SCOP database is extended into a resource database called ASTRAL Compendium (http://astral.berkeley.edu). This database aids in providing tools to analyze protein domain structures, which are classified in the SCOP database. It documents two types of domain sequences: a. ATOM sequence record of SCOP domain boundaries and b. domain sequence generated from SEQRES record of PDB file (Brenner et al., 2000). Apart from this, ASTRAL database offers criteria to select sub-sections of SCOP domains based on different sequence identity levels (Chandonia et al., 2004). SUPERFAMILY database uses SCOP superfamilies to build hidden Markov Models to annotate proteins in many genomes (Gough et al., 2001).

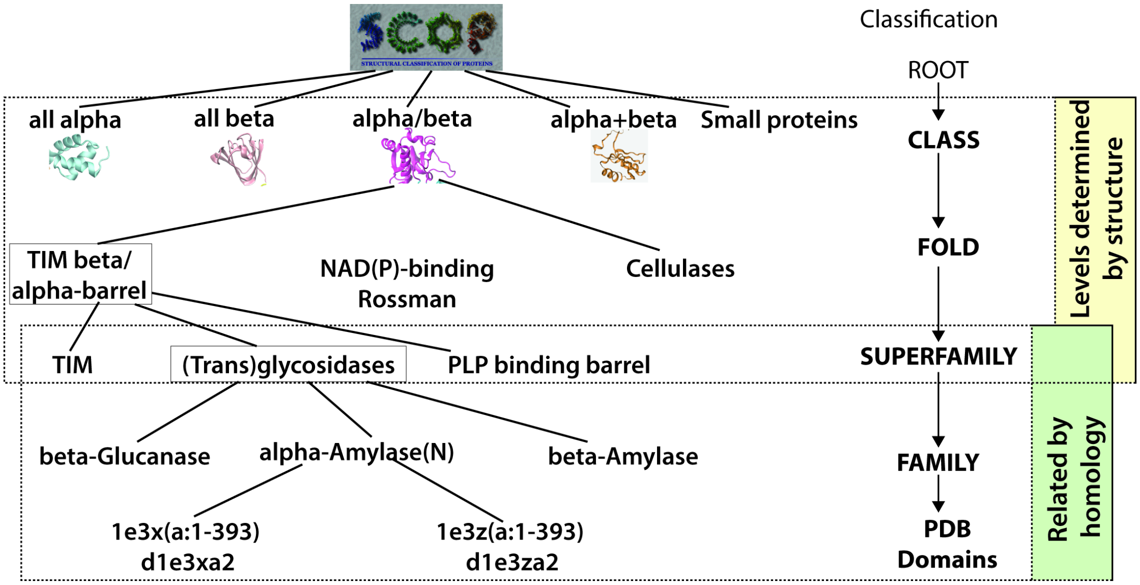


Figure 1.7 Hierarchy of protein classification in SCOP database (Modified from *http://compbio.berkeley.edu/people/ed/SeqCompEval/*)

In 2013, SCOPe (SCOP-extend) was developed, which is an extension to v1.75 (last released version) of SCOP database (Fox et al., 2013). It focuses on the classification of new PDB structures in SCOP-1.75 by utilizing automated methods. It also corrects errors in SCOP classification and maintains the accuracy as that of manually curated protein structures. Similar to SCOP, it integrate ASTRAL database and also update it regularly. The current release of SCOPe-v2.07 includes 92665 PDB entries and 294450 domains.

1.5.3.2 Class, Architecture, Topology, Homology (CATH) (http://www.cathdb.info/)

Similar to SCOP hierarchical classification, CATH also classifies domain structures (Orengo et al., 1997) with some additional hierarchical levels. CATH uses a semi-automatic procedure to identify domains and classify into various levels based on composition and packing of secondary structure elements. The four hierarchical levels defined in CATH are (Figure 1.8):

Homologous Superfamily (H) – The structural domains having clear evolutionary relationship based on the similarity in their structure, sequence or function are classified into homologous superfamily.

Topology (T) – This is analogous to the *Fold* in SCOP and groups structures based on topology of their core regions, that is, if they share overall shape and connectivity of the secondary structures in the domain core. As has been in Fold of SCOP, structures within a topology can have varying structural decorations to the common core.

Architecture (A) – The topologies are grouped in same architecture based on the overall shape of the domain structure as determined by their secondary structures but ignoring the connectivity between secondary structures. This classification is performed manually.

Class (C) – This is the highest level of classification and it represents the gross content of secondary structures of the domain (automatic process). The four classes in CATH are: mainly alpha, mainly beta, alpha beta and few secondary structures.

In CATH, the homologous superfamily (H) is further divided into subfamilies by clustering based on pairwise sequence identities from Needleman and Wunsch algorithm (Needleman & Wunsch, 1970). These clusters are namely S, O, L, and I depicting domains in clusters having atleast 35%, 60%, 95% and 100% sequence identity respectively. The final level, D, is used as counter to make sure that each domain is represented uniquely in the CATH database. Thus, each H is further classified at S, O, L, I and D levels of classification.

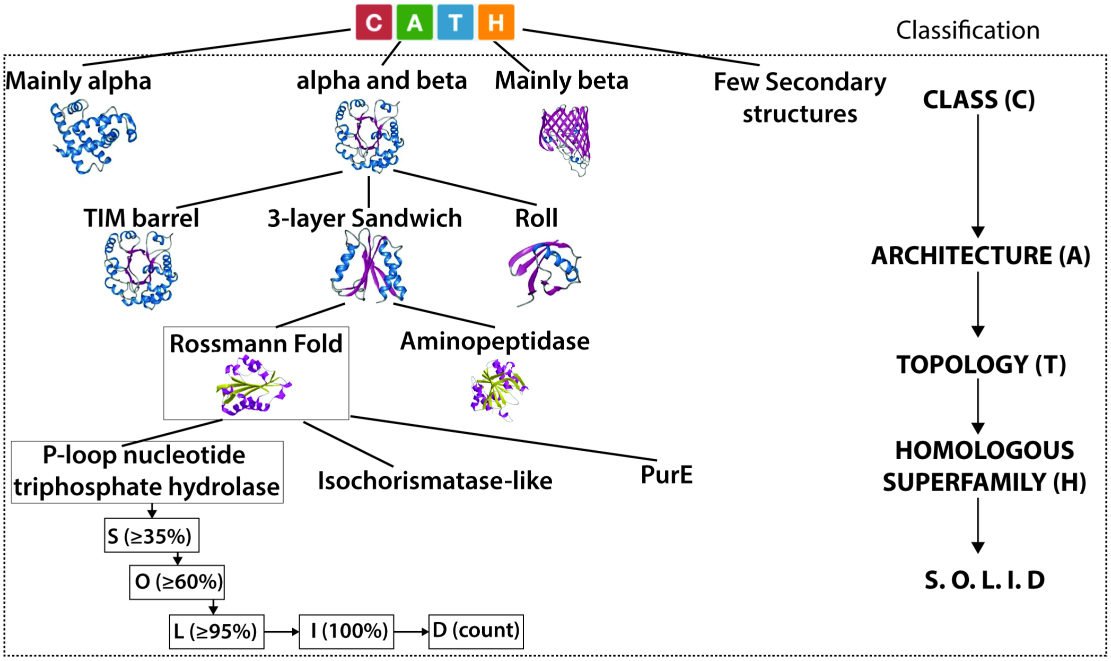


Figure 1.8 Schematic representation of CATH database hierarchy.

Through decade of research, both SCOP and CATH have been established as the gold standard databases in the field of protein structure research. These have been used in protein structure prediction and classification, assessment, various machine learning approaches. However, the differences and inconsistencies in both hierarchies could lead to unavoidable issues during training and benchmarking phases of protein structural studies. In total, there is approximately 70% overlap between definitions of the domain SCOP and CATH (Csaba et al., 2009). Table 1.6 summarizes current statistics of SCOP and CATH databases.

Table1.6 Showing population of different hierarchical levels in the databases

CATH hierarchy	CATH v. 4.20 (number of members)	SCOPe 2.07 (number of members)	SCOP hierarchy
Class	4	7	Class
Architecture	41	--	--
Topology	1391	1243	Fold
Homologous superfamily	6119	2044	Superfamily
--	--	4955	Family

1.6 Protein structural space

The representation of all possible proteins often referred to as 'protein universe' (Levitt, 2009; Taylor, 2020). As described before, sequence of amino acids adopt regular secondary structures such as α-helices and β-sheets, which are predominant in protein structures. These secondary structures are connected by loops or turns. Some amino acids are selectively more prevalent in one class over the other (Chou & Fasman, 1974) and make their prediction in protein sequences more straightforward. The secondary structure organizes into tertiary structure, which is next level of structural complexity. The tertiary structure of the protein is defined to represent the secondary structure with interconnectivity among them that is a unique conformation of protein in 3D space called "fold or topology". Although, theoretically it is possible for a protein sequence to adopt any fold or have various possible arrangement of secondary structures (Holm & Sander, 1996), yet it has been shown that structures of proteins have limited and repeated folds (Chothia, 1992). This is most likely because proteins are restricted by numerous constraints including structure, interactions, function and biophysical characteristics and hence the resultant structural space used in nature is surprisingly small. It has been fascinating to study structural space to understand evolutionary relationships between proteins despite no significant sequence similarity.

1.6.1 Nature of protein structural space

The structural space of protein can be defined as collection of experimentally determined protein structures, which could be structural domains or complete protein structures (Taylor, 2007, 2020; Sadreyev et al., 2009). The distribution of entities in any spaces can be described by distance separating them. Using some distance measures, either the space can be viewed as disparate if the large distances separate enclosing entities and these cannot be reached from each other, or the space can be of continuous nature consisting of region with density gradation where regions of high density can communicate with each other through intermediate regions of low density. Similarly, the protein space can be described as discrete or newly accepted a continuous space (Taylor, 2020). In the following sections, we have described both views of structural space.

1.6.1.1 Discrete nature of fold space

The belief that protein structural space is divided into distinct folds or discreetness of protein structure space originated early when limited structures were available and it was observed that structures showed relatedness to each other and not so much to other folds. For instance, hemoglobin (Perutz et al., 1960) and myoglobin (Kendrew et al., 1958) structures were found to be similar in their structure and hence classified into the same protein family. Similarly, some proteins were found similar to known structural folds and classified with it, otherwise they were deemed to be new fold. In absence of quantitative measure to assess structural relatedness, the structural space got sparsely populated with unrelated structural folds. Figure 1.9 shows few examples of these earlier fold prototypes used to classify other protein structures (Sadreyev et al., 2009).

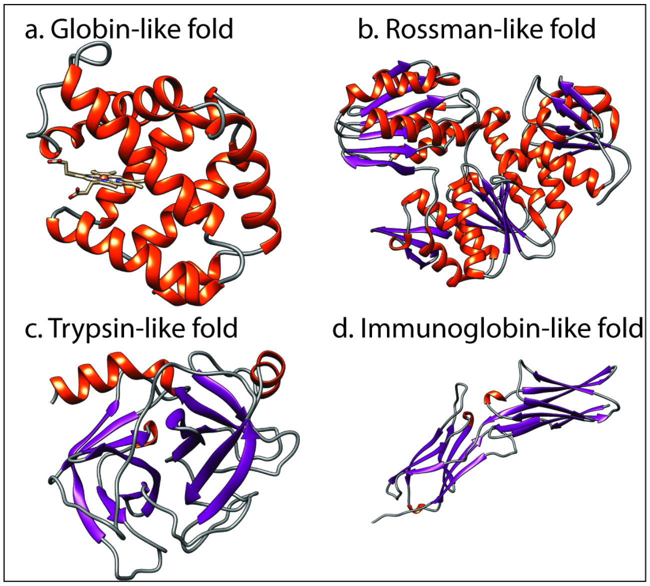


Figure 1.9 Examples of earliest known and abundant fold types. A) Globin-like fold (PDB ID 3SDH); B) Rossmann-like fold (PDB ID 2JFG); C) Trypsin-like fold (PDB ID 1AQ7); D) Immunoglobulin-like fold (PDB ID 1VCA). Structures are colored by secondary structures: helices as orange, strands as magenta and coils as gray (Modified from Sadreyev et al., 2009).

Some of the earlier studies suggested that the space is discontinuous for single domain proteins (Holm & Sander, 1997; Hou et al., 2005). They performed all-against-all comparisons of structures from PDB and claimed that in case of proteins with single domain, structural space is made of distinct, non-overlapping folds. Since domain is regarded as conserved region and basic unit of evolution in the protein, which led to the assumption that fold space is disparate in nature.

Moreover, such observations laid the basis for the development of two very well known protein classification databases: SCOP (Murzin et al., 1995) and CATH (Orengo et al., 1997). These databases group proteins into different fold types based on their evolutionary relationship (homology), function and topology. An analysis on the available protein structural folds has shown that structures occupy only four regions of the thinly populated fold space (Hou et al., 2003). The four regions correspond to the arrangement of secondary structural elements, which formed the basis of two popular protein structural databases.

It is known that proteins are marginally stable molecules (Giver et al., 1998; Privalov & Khechinashvili, 1974; Ruvinov et al., 1997; Savage et al., 1993; Vogl et al., 1997) and only infinitesimal fraction of stable fold are observed to carry out is function in an ocean of unstable conformations (Lupas & Koretke, 2008). The external factors like mutations may drift proteins from their stable state of proteins that will either be eliminated by natural selection or constrained if it is bearable by fold (Edwards & Deane, 2015). Therefore, one can argue that discreteness is a result of evolutionary pressure along with constrained thermodynamic stability of protein, which prevents the movement between folds (Choi & Kim, 2006; Lupas & Koretke, 2008).

1.6.1.2 Continuous nature of fold space

Later, the work of Shindyalov and Bourne (Harrison et al., 2002; Shindyalov & Bourne, 2000) demonstrated that there is continuity in fold space shown by some topologies in secondary structure class. Moreover, suggesting that many paths exists between folds and protein structure can move between them. In a similar study, structural similarities were observed among SCOP folds (Yang & Honig, 2000). Recently, many studies have found that structural space is continuous and dense based on structural similarity of single domain proteins (Kolodny et al., 2006; Pascual-García et al., 2009; Sadowski & Taylor, 2010; Skolnick et al., 2009). Further, it has been suggested that continuous nature of structural space is due to similar packing of secondary structure (Skolnick et al., 2009). The continuous nature is purely caused by structural relatedness, which involves the arrangement of secondary structure elements to form a stable state (Sadreyev et al., 2009) and need not have any direct implications on protein evolution (Skolnick et al., 2009).

1.6.1.3 Dual nature of structural space

With structural alignments of many proteins, an alternate view of structural space has been proposed, that is, it dual in nature showing both discrete and continuous. As a consequence, distribution of protein points in the fold can be visualized as regions of high density and low density (Sadreyev et al., 2009). With limited number of structures determined the structures were mostly distinct creating isolated clusters of islands giving a discrete picture. However, with more structures region between islands of structures started getting populated making the distribution look more continuous (Figure 1.10). This suggests that the structure space mostly continuous having few region of preferred conformations. The discrete and continuous paradigms about the fold space complement each other and provide important insights into evolutionary and structure-function landscapes respectively.

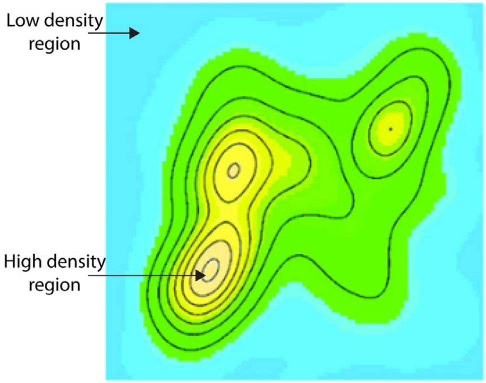


Figure 1.10 Contour plot of estimated probability density for clustered protein structures based on geometric similarities. Contour plot colored from blue (low density) to orange (high density) regions. When few protein structures were considered, they tend to cluster around yellow region, however, as more and more protein structures were determined they started to occupy green-blue area making the fold space to look more continuous (Modified from (Sadreyev et al., 2009)).

1.7 Evolution of multidomain proteins

The protein domain can either exist as a single domain protein (Jaenicke, 1987) or it can combine with different domains of the same protein and to become a part of multidomain protein (Doolittle, 1995; Rossmann et al., 1974). Considering function associated with each domain (Ponting & Russell, 2002), in evolution multidomain protein allow extending the function of proteins by bringing domains of varying functions. In such context, proteins are also considered as unit of evolutionary unit (Murzin et al., 1995). The linear arrangement (Figure 1.11) of different domains in a protein chain is called *domain order* or *domain architecture or domain organization* (Apic et al., 2001a; Kummerfeld & Teichmann, 2009; Bjorklund, et al., 2005; Bornberg-Bauer et al., 2005). Such domain organization could be helpful in cladistics analysis by providing unique evolutionary markers (Koonin et al., 2000). During evolution, with increasing organismal complexity are also associated with proteins acquiring new functions by domain combination, where domains undergo various recombinations with each other to result in complex multidomain architectures (Koonin et al., 2000). The multidomain architectures is advantageous as it can increase efficiency of cellular processes as functional modules lie close to each other (Yu et al., 2019; Enright et al., 1999; Marcotte et al., 1999); the rate and ability of proteins to fold spontaneously into a spatial substructure and utilizing mechanisms to avoid misfolding (Garbuzynskiy et al., 2013; Han et al., 2007); reconfiguration of domain assemblies by rearrangement of existing domains to perform new functions (Vogel et al., 2004).

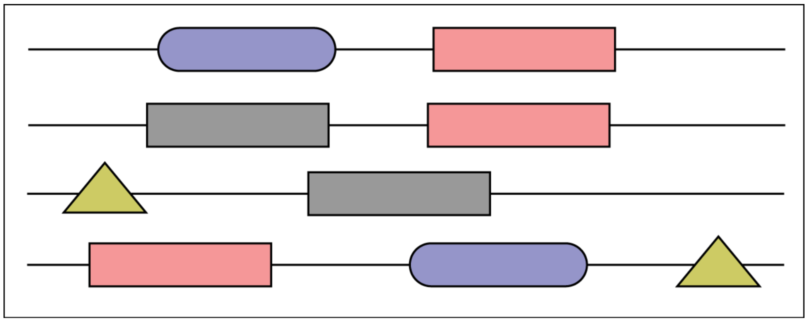


Figure 1.11 Schematic shows representation of domain architectures. Figure shows cartoon representations of linear organization of domains.

Previous studies have shown that new domain combinations in course of evolution mostly occur through non-allelic homologous recombination of nearby genes (Buljan et al., 2010). Such events (Figure 1.12) include divergence, duplication, insertion/ deletion, fusion/ fission of genes and their gene products (Chothia & Gough, 2009; Ekman et al., 2005; Fong et al., 2007; Kummerfeld & Teichmann, 2005; Vogel et al., 2005; Weiner et al., 2006; Weiner & Bornberg-Bauer, 2006). Among these mechanisms, gene fusion preceded by duplication and recombination dominantly mediated the process of domain gains in proteins. It is also noteworthy that domain fusion events more likely to occur at amino and carboxyl termini of proteins (Marsh & Teichmann, 2010). In addition, it has also been observed that small fraction of eukaryotic proteins also gain new domains through intronic recombination mechanism (Patthy, 1996).

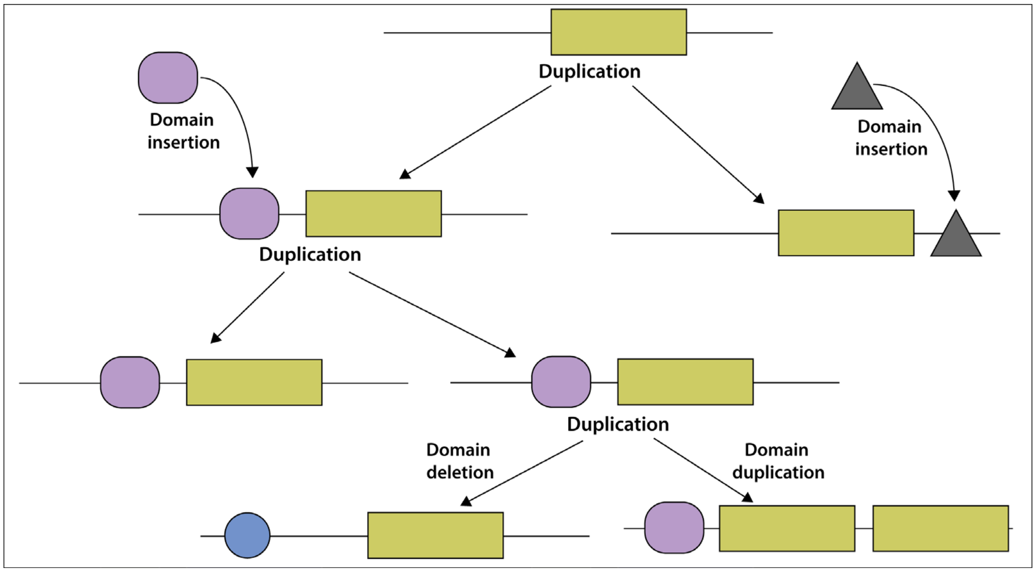


Figure 1.12 Schematics of various events in evolution of multidomain proteins (Modified from Kannan & Wheeler, 2012).

Generally, domain architectures predominantly appear to originate only once but if they found to span many species, it consequently indicates their common origin (Doolittle, 1995). Domains are often associated with a specific function, but domain fusion ensue novel proteins with complex and diverse functions either forming new inter-domain functional sites or incorporating domains with separate functions (Apic & Russell, 2010; Bashton & Chothia, 2007; Han et al., 2007). This process of alternate domain combinations from a limited set of existing units potentially creates a significant functional diversity and discovers novel proteins (Levitt, 2009; Marsh & Teichmann, 2010; Moore et al., 2008). Regardless of this, some domain architectures evolved independently due to functional constraint or by random chance (Forslund et al., 2008). These types of proteins offer new insights into structure function and convergent evolution (Gough, 2005).

Lastly, it is the stability of the novel protein that will determine the fate of the new domain architecture, whether it will continue to be preserved in nature or not (Marsh & Teichmann, 2010). Noticeably, it has been observed that certain domains combine with multiple different domains, these are regarded as ‘promiscuous’ domains, while, other domains combine with one or few domains (Basu et al., 2008). These findings taken together implies that a strong purifying selection constraints act on evolution of domain architectures which leads to specific domain interaction from originally evolved non-specific or promiscuous interactions (Basu et al., 2009).

1.7.1 Inter-domain linkers

Many studies analyzed the organization of structural domains in multidomain proteins with properties of segment connecting domains, which is called as inter-domain linker (IDL). These linker regions can vary in their structure, size and composition (George & Heringa, 2002c; Bhaskara, et al., 2013) and serve as covalent link between domains that affect folding, stability and domain-domain orientations (Robinson & Sauer, 1998; Bhaskara et al., 2013). The linker regions are known to modulate the function of the proteins by allowing the tethered domains to communicate such as inter-domain linker harbors functional sites/active sites (Bashton & Chothia, 2002; Wei et al., 2001). Apart from composition of linkers, their length is an important feature as changes in its length affects domain stability and inter-domain orientations (Bhaskara et al., 2013; Robinson & Sauer, 1998; van Leeuwen et al., 1997; George & Heringa, 2002). Moreover, IDLs have been shown to play key role in maintaining inter-domain cooperative interactions and as scaffold prevents unfavorable interactions between folding domains. The knowledge of linker properties will help in designing fusion protein engineering (Bhaskara et al., 2013). Previous studies have categorized linkers as: helical (rigid) and non-helical (soft) linkers. It was observed that soft linkers foundin hinge regions and are rich in glycine residues making these highly flexible (Ikebe et al., 1998). Because of their flexibility these can easily break and form contacts with adjacent domains and aid in catalytic events. The rigid linkers are observed to be rich in proline residues (Adzhubei & Sternberg, 1994), which act as spacers and keep domains apart to prevent unfavorable interaction during folding process (Briggs & Smithgall, 1999; George & Heringa, 2002c; Gokhale et al., 1999; Ikebe et al., 1998). As rigid linkers act spacers, these are called as ‘molecular rulers’ because these behave as ‘metric’ function, for instance to keep distance between domains or depth of binding pocket (Wriggers et al., 2005). Such linker usually includes stable α-helical structures (Johnson et al., 2003).

The reliable detection of domain boundaries usually rely on atomic coordinates of experimentally determined or predicted 3-D structures (George & Heringa, 2002; Holm &

Sander, 1994; Islam et al., 1995; Marsden et al., 2002; Siddiqui & Barton, 1995; Taylor, 1999; Wernisch et al., 1999). The knowledge of conserved regions in sequences and evolutionary information could further aid in identification of domain boundaries (George & Heringa, 2002; Gouzy et al., 1997; Gracy & Argos, 1998; Sonnhammer & Kahn, 1994). Such as CHOP algorithm utilizes both domain boundary information from structural and sequence (Pfam-A) domains to delineate proteins into domain-like fragments (Liu & Rost, 2004b). Its improved version (CHOPNet) rely on neural networks with additional evolutionary and predicted tertiary structure predicted domain boundaries (Liu & Rost, 2004a). An alternative approach for detecting domain boundaries is to identify interdomain linkers (Bae et al., 2005). Most linker identification methods use predicted secondary structure, propensity of amino acids, or a combination of the two (Miyazaki et al., 2002; Tanaka et al., 2003). For example, Q-linkers occur in a variety of bacterial regulatory and sensory transduction proteins at the boundaries of functionally distinct domains. These are usually 15-20 residues long, and are not conserved among homologous proteins. These adopt coil structure with preference of amino acidsArg, Gln, Ser, Glu, and Pro.

1.8 Domain-domain interaction interfaces

Since the protein domain is considered as a fundamental unit of protein, protein-protein interactions can further be described as interaction between domains or ‘domain-domain interactions’ (Björkholm & Sonnhammer, 2009). More appropriately, these are called as inter-chain domain interaction to distinguish from interacting domains of a multidomain protein that are referred as intra-chain domain interaction (Park et al., 2001) (Figure 1.13). The knowledge of inter/intra-chain domain interactions has been primarily derived from available protein structures in PDB apart from computational methods to predict such interactions (Prieto & Rivas, 2010; Zhao et al., 2008). The domain interaction interfaces can host functional (both catalytic and binding) sites of proteins for example in ATPases, the catalytic and effector functions being separate part of different domains (Ito et al., 2003; Janin & Wodak, 1983). Moreover, analyses of PPI at domain level can give molecular insights as well as crucial in detecting previously unrecognized protein-protein interactions, protein docking, hot-spot residues, and development of new drugs (Apic et al., 2001a; Aytuna et al., 2005; Betel et al., 2007; Shoemaker et al., 2006; Yellaboina et al., 2011).

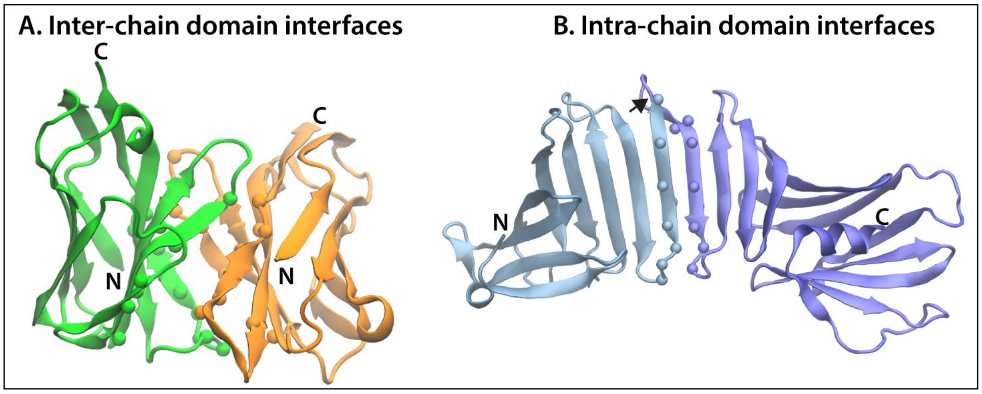

A. Inter-chain domain interfaces
C
C
N
N
B. Intra-chain domain interfaces
N
C

Figure 1.13 Representation of inter-chain and intra-chain domain-domain interfaces. Domain interfaces between chains shown in A) where inter-chain interfaces is formed between chains C and D of PDB ID: 3mff and B) Intra-chain domain interfaces of two domain protein (PDB ID: 1ospO). The protein chains C and D of 3mff are colored green and orange respectively. The domains 1 and 2 of 1ospO are colored light and dark blue respectively. The interfacial residues are shown in vander Waals sphere representation.

The computational analyses of intra-chain domain interfaces have shown that these are mostly hydrophobic, with a degree of hydrophobicity intermediate between obligate and non-obligate protein complexes (Argos, 1988; Jones et al., 2000). The interface residues are relatively more conserved in comparison to solvent exposed residues (Littler & Hubbard, 2005). The intra-chain domain interfaces show remarkable differences in surface area, which ranges from small interfaces that allow restricted inter-domain motion by IDL to much larger interfaces where little inter-domain motion (Bhaskara et al., 2013).

As has been mentioned before, arranged order of domains or ‘domain architecture’ in multidomain has been suggested to be strongly conserved across different organisms (Apic et al., 2001b; Bashton & Chothia, 2002; Han et al., 2007; Vogel et al., 2004). The analyses on arrangement of protein domain families in multidomain protein have shown that some domain families interact with only one or two families whereas some families (for example, “P-loop containing nucleotide triphosphate hydrolase”) interact with many other domains (Apic et al., 2001a; Iyer et al., 2004). Such “promiscuous domains” are a major source of functional novelty (Basu et al., 2008). It has been found that domains that interact with a limited domains partners, typically, interact with the same interface, while domains that interact with multiple different partner domains are usually observed making use of different interfaces (Littler & Hubbard, 2005).

Previous studies on conservation of intra-chain interacting domains have found that domains in close homologous proteins (typically sequence identity 30-40% or higher) interact similarly (Aloy et al., 2003) suggesting orientation of interacting domains tends to be evolutionary conserved. The study on classical Rossmann superfamily domain combination with 8 catalytic superfamilies showed that relative domain geometry is conserved in superfamily-superfamily pairs. However, the same is not conserved between two superfamilies (Bashton & Chothia, 2002). The geometrical relationship between domains was not conserved when domains sequential order is reversed. Subsequent studies on the extent of conservation of domain-domain geometry and molecular structure of interface among homologous two-domain proteins have shown that ~60% of pairs conserve their geometry and interface and ~38% of pairs have variable geometrics and interface (Han et al., 2006). Interestingly, variable geometry and interface can be found even in homologous structures. Another study has noted that usually the relative positioning of two superfamily-related domains in unrelated proteins are not similar. These suggest that domain orientations in 3D may be mostly affected by functional restraints (Rekha et al., 2005).

The intra-chain domain interactions in multidomain protein have been suggested to be important for stability, and folding (Arviv & Levy, 2012; Bhaskara & Srinivasan, 2011; Flaugh et al., 2005; Han et al., 2007; Levy, 2017). Although it is assumed that domains in multidomain proteins follow the same folding principles as of single domain proteins, however, unlike single domain proteins, which lack domain-domain interactions, the length of linker and nature of domain interface impact folding of protein. Some proteins show cooperative behavior among domains where folding of one domain is influenced by the other domain (Batey et al., 2005, 2006). In other proteins, where the interaction between domains is weak, interfaces are small and loosely packed, domains fold independently (Han et al., 2007; Scott et al., 2002).

1.9 Tertiary structure comparison methods

The structural alignments of proteins are essential to detect distantly related sequences (remote homologs). Apart from this, alignments can be used for function prediction of a new protein by detecting regions of local and global similarity to a protein with known function (Carugo, 2006).

1.9.1 Different measure of domain interface structural similarity

1.9.1.1 Root Mean Square Deviation (RMSD)

It is the most commonly used metric while quantifying the similarities between superimposed atomic coordinates. The superposition of protein structures between structurally equivalent positions is performed using Kabsch algorithm (Kabsch, 1976). Usually, proteins are represented using trace of C-α atoms. The RMSD is computed from Euclidean distance of equivalent C-α atoms of superposed coordinates. RMSD is calculated by the equation given below:

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N}\left(x_i^b - x_i^a\right)^2 + \left(y_i^b - y_i^a\right)^2 + \left(z_i^b - z_i^a\right)^2}{N}}$$

where, x,y,z are the coordinates of atoms with a and b denoting coordinates from 2 structures; N is total number of superposed Cα-atoms

RMSD is expressed in unit of Å. The RMSD can be calculated over different subsets C-α atoms, or all heavy atoms of the protein or for small molecules such as ligands bound to these protein structures or for interfacial residues. Usually, RMSD calculated over the full length of the protein or domain is referred as 'global RMSD' and other calculated for a subset of atoms could be referred as 'local RMSD'. The RMSD for interfacial residues is referred to as 'interface RMSD'.

Although RMSD is an often used measure to assess the structural similarity between protein structures, the value of RMSD is function of protein/domain/interface length. For example, a small set of residues having large deviation can lead to large RMSD value despite deviations arising from a local region.

1.9.1.2 Template Modeling Score (TM-score)

Since length dependence is a disadvantageous feature of RMSD as a measure of structural similarity, a length independent measure TM-score was developed for assessment (Zhang & Skolnick, 2004). Though TM-score was developed to assess quality of modeled structure, it has been useful metric to compare structures. The issue with RMSD is that all residues are weighted evenly during calculations and therefore high RMSD values depict more sensitivity towards local structure deviation rather than to the global topology. Whereas, TM-score is a variation of the Levitt-Gerstein (LG) score (Gerstein & Levitt, 1998), which weights shorter distances between corresponding residues more strongly than longer distances which makes the score more sensitive towards global fold topology. TM-score is given by the equation:

$$TM - score \ = max\left(\frac{1}{L}\left[\sum_{i=1}^{L_{ali}}\frac{1}{1 + \ {d_i^2}\Big/{d_o^2}}\right]\right)$$

where max is the maximum value after optimal superposition, L is the length of target/native structure, L_{ali} is the length of the aligned residues to the template structure i.e. the number of equivalent residues in two proteins, d_i is the distance between the i^{th}pair of residues and d_0 is a scaling factor. The d_0 is defined as:

$$d_o = 1.24\sqrt[3]{L - 15} - \ 1.8$$

which is an approximation of the average distance of corresponding residue pairs of random related proteins. This makes it length independent measure. TM-score ranges from 0 to 1, where 1 indicates a perfect match between two structures. TM-score also signifies the quantitative correspondence with fold/topology classification of the predicted/model structure. Based on this, TM-score of 0.5 between two protein structures implies that these structures related at the level of fold (Xu & Zhang, 2010).

1.9.1.3 Interface Similarity Score (IS-score)

This score as its name suggests, gives a measure of the extent of structural similarity between two protein-protein interfaces. Usually, the structural similarity of PPI interfaces is derived upon aligning the interacting structures individually and computing the similarity of structurally equivalent interfacial residues. Such an approach is guided by global alignment of individual structures and would not be able to detect structural similarity of interfacial regions (Gao & Skolnick, 2010a). It has been observed that despite protein structures show higher global similarity, some of the proteins differ in their interaction modes. To identify and measure interface structural similarity, a program interface alignment (iAlign) was developed that aligns interface and measure similarity using IS-score (Gao & Skolnick, 2010a). Unlike other structural comparison methods known which rely only on geometric matches, IS-score also includes conservation of contact pattern among residues between interfaces. In this way, IS-score offers a more reliable metric to provide better insights into interface alignments and is defined as:

IS-score = (S + s0) / (1 + s0), where

$$S = \frac{1}{L_Q}\max\left[\sum_{i=1}^{Na} f_i /(1+\frac{d_i^2}{d_0^2})\right],$$

where, L_Q is the length of the query interface; N_a is alignment length between query and template; d_i is distance (in Å) between Cα residues of aligned pairs; f_i is contact overlap defined by$f_i \equiv (c_i/a_i + c_i/b_i)/2$; where a_i and b_i are number of interfacial contacts of template and query interfaces at i^{th} position in the alignment respectively, and c_i is number of overlapping interfacial contacts at the same i^{th} position; d_0 is given by

$$d_0 \equiv \begin{cases} 1.24(L_Q - 15)^{1/3} - 1.8\, For\ sequential\ alignment \\ 0.7(L_Q - 15)^{1/3} - 0.1\, For\ non - sequential\ alignment \end{cases}$$

To make score S length independent, it is normalized with s_0, which is given by $s_0 \equiv 0.18 - 0.35/{L_Q}^{0.3}$. The normalized S score is referred to as IS-score, which has the maximum score of one for perfect alignment between two identical structures (Gao & Skolnick, 2010). Based on random protein-protein complexes, p-value is calculated for IS-score and IS-score with p-value < 0.05 suggests that two aligned interfaces are significantly similar to some biological relevance.

1.9.1.4 Inter-domain geometry

Previously, there were several studies to analyze the different arrangements and combinations of domains within multidomain proteins (Apic et al., 2001a; Bashton & Chothia, 2002; Gough, 2005). It has been learned that when different domains carry out a function in a protein, they just not only interact but also undergo some domain motions to orient themselves in a particular conformation in 3-D space which forms the final functional interaction interface (Gerstein et al., 1993). In one of this work, measure to inter-domain geometry was proposed based on a simple concept of center of gravity (CoG) to describe the relative orientation of domains in two domain homologous proteins. This involves superposing one pair of domains followed by calculating the translation and rotation required to superimpose the CoG's of other pair of domain. This quantifies as inter-domain geometry (Figure 1.14). The geometry is regarded as conserved if the required rotation < 20° and translation < 5Å to optimally superimpose CoG's. In our work, to calculate the difference in orientation of interaction interface between intra-chain domain pairs, we have used the best TM-score to select the first pair of domains and subsequently, other pair was used to calculate the rotational and translational movements.

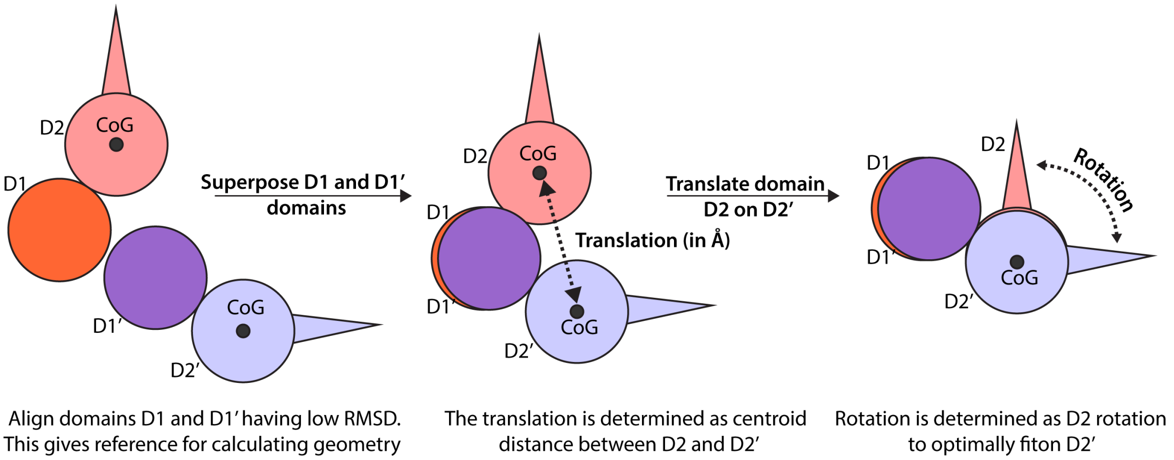

D2
CoG
D1
D1'
CoG
D2'
Superpose D1 and D1' domains
D2
CoG
D1
D1'
Translation (in Å)
CoG
D2'
Translate domain D2 on D2'
D1
D2
Rotation
D1'
CoG
D2'
Align domains D1 and D1' having low RMSD. This gives reference for calculating geometry
The translation is determined as centroid distance between D2 and D2'
Rotation is determined as D2 rotation to optimally fiton D2'

Figure 1.14 Schematics to show calculation of interdomain geometry in 2-domain proteins. Figure shows steps in calculating interdomain geometry following the method of Han (Han et al., 2006). Two domains of each protein are colored differently. The domains D1 and D2 of one protein are colored as orange and pink respectively. The equivalent domains D1' and D2' (structually related) in other protein are colored as violet and light blue respectively (Modified from (Han et al., 2006)).

1.9.1.5 Dihedral angle

In order to compute inter-domain geometry another measure was developed based on psuedo diherdral angles. This angle (χ = -180 to +180) is calculated between two contiguous domains using a center of mass and terminal residues (Cα C-terminal of a 1st domain and N-terminal domain of 2nd domain) of the domains (Figure 1.15) (Bhaskara et al., 2013). The smallest dihedral angle is taken and the difference ($\Delta\chi$) between this angle will classify the protein pairs either as conserved ($\Delta\chi \leq 30°$) and not conserved ($\Delta\chi > 30°$).

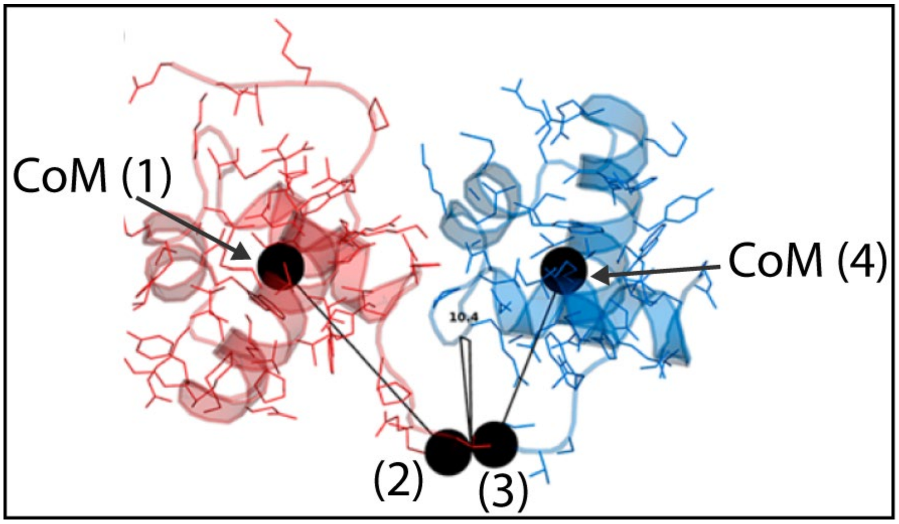


Figure 1.15 Representation of four points (coordinates) for calculating the dihedral angle. The four points: (1) center of mass of domain 1; (2) last Cα of domain 1; (3)Cα of N-terminal residue of domain 2; (4) center of mass domain 2. These points are shown as black color filled circles in the figure (Modified from (Bhaskara et al., 2013)).

1.10 Modeling of multidomain proteins

The number of known protein sequences has increased exponentially with the success of an expanding array of genome sequencing projects. Since knowledge of 3-D structure of proteins can give insights into molecular detail of function, it is important to determine their structures. Despite structural genomics efforts, determination of protein structure has not been able kept the same growth pace as sequences. Presently, the number of single domain protein structures out numbers multidomain proteins in PDB, which has ~32% of multidomain structure removing redundancy (Xu et al., 2015). This can be explained by limitation of experimental approaches to determine structures of large size and having inter-domain motion of multidomain protein. In many instances, this is overcome by cleaving protein at domain boundaries to obtain stable individual domains for structure determination (Savitsky et al., 2010). Owing to the gap between known protein sequences and known structures due to the limitations of experimental methods in solving the protein structure, the development of computational approaches would be helpful in generating models of good quality that can be used in various experimental/computational studies.

Broadly, computational approaches for protein structure prediction can be classified as:(i) Template based modeling (TBM) and (ii) free modeling- FM (*de novo/ ab initio*). TBM refers to the modelling of target protein sequence based on structural templates derived from threading/fold recognition, whereas FM or *ab initio* predict structure of target protein relying on general structural features and does not use templates (Kryshtafovych et al., 2011). TBM methods have been most widely because of reliability and accuracy of models. The TBM involves identification of template using sequence-based approaches, threading or fold recognition that is followed by using one or more templates for homology modeling or fragment assembly approach to predict tertiary structures (Pandit & Skolnick, 2010). It is crucial in TBM to: (i) to reliable identify correct templates and (ii) ability to refine the template structure closer to that of the native structure.A systematic study on dependence of modeling accuracy using MODELLER on template quality showed that modeling is highly dependent on evolutionary distance between target and template (Fiser, 2010). The template-target having sequence identity >50% usually results in model with RMSD within 1Å. Similarly, when the target sequence has sequence identity between 30 and 50% to the templates, models have core region (~85%) within 3.5 Å, with the errors mainly in loop regions and tails (Eswar et al., 2006). The modeling accuracy drops for twilight zone sequence identity (<30%) (Chung & Subbiah, 1996; Eswar et al., 2006). Recently developed methods such as TASSER/I-TASSER (Roy et al., 2010; Zhang et al., 2005; Zhang & Skolnick, 2004a) and ROSETTA (Wollacott et al., 2007) are able to address these issue of low sequence identity templates where predicted structure is close to native than to template structure. Most of these studies are performed on single domain protein modeling.

Unlike single domain modeling, multidomain protein structure prediction not only requires individual domains to model accurately but also has to reliably predict inter-domain orientation and/or domain-domain interfaces. The latter part is challenging and can be considered to be part *ab initio* modeling, which can be divided into two approaches: (a) docking approach – structure prediction of multidomain through domain assembly is treated as docking problem (Cheng et al., 2008; Inbar et al., 2005; Zhou et al., 2019) (b) iteratively sample the degrees of freedom of the linker rather than of two domains (Wollacott et al., 2007).

1.10.1 Comparative modeling of multidomain proteins

The multidomain structure can be predicted using the most common approach of comparative modeling provided a reliable template could be identified for aquery sequence encompassing most of it. Thus, in this both modeling of individual domains and their relative orientations will be based on the template structure. However, the most common issue is the observed variability of inter-domain geometries between homologous proteins (Aloy et al., 2003; Han et al., 2006). Therefor, prediction of domain orientation purely based on homologous relationship may not result in accurate prediction of multidomain proteins structure, especially their interfaces. Below, we describe alternate methods of multidomain structure predictions.

1.10.2 Docking approach

In absence of a template for the full length of a target multidomain protein sequence, docking of predicted or experimentally solved domain structures can be used to predict tertiary structure (Halperin et al., 2002). In this approach, domains are identified in multidomain proteins followed by modeling of domains separately. Subsequently, domains are docked to generate multiple conformations with an objective to assemble domains.

The docking approaches have been utilized to identify optimally interacting molecular structures where two separate molecular structures, receptor and ligand, are used as an input and usually all possible protein surface interaction of receptor is probed using ligand to identify most likely interacting interface (Huang, 2015; Inbar et al., 2005; van Zundert et al., 2016). The surface can be described using geometric shape descriptor or a grid. Docking has been preferred method for identifying interfaces in protein-protein interactions that can be implemented for docking domains (Inbar et al., 2005; Lise et al., 2006).

Primarily, docking involves two main steps: (i) conformational space search, and (ii), ranking of potential solutions (Halperin et al., 2002) (Figure 1.16). The first step generates a large number of putative bound conformations by sampling the interacting conformational space. During this sampling, backbone and side-chain motions are ignored to reduce large search space that is also called as "rigid body" docking (Huang, 2015). The docking methods often improve computational complexity by combining structures based on shape complementarity simplifications. Among many approaches known for simplifications of molecular surfaces, the most commonly employed is fast Fourier transform (FFT) approach (Katchalski-Katzir et al., 1992). The docked conformations are ranked using scoring function, which are benchmarked to identify native like interfaces and can discriminate between native and non-native docked conformations. These usually involved knowledge based potentials, shape complementarity and electrostatic interactions.

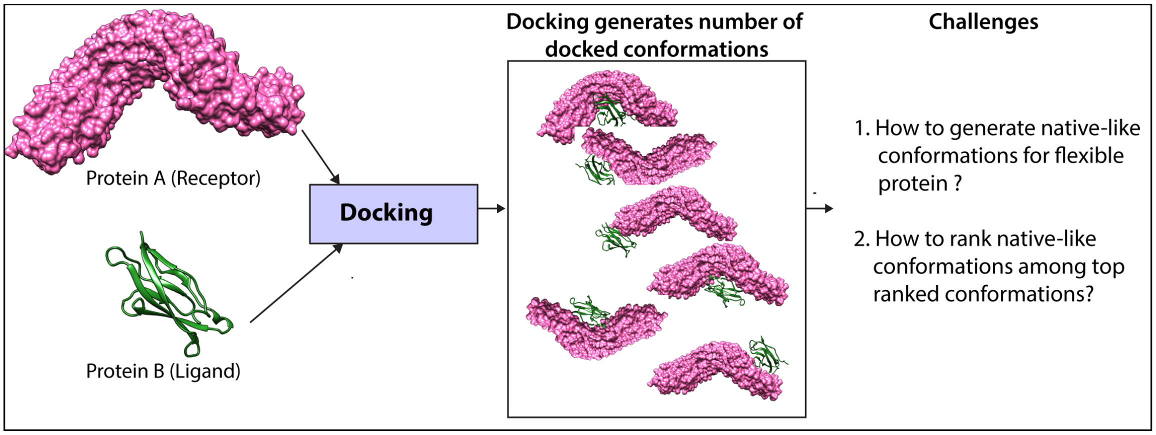

Protein A (Receptor)
Protein B (Ligand)
Docking
Docking generates number of docked conformations
Challenges
1. How to generate native-like conformations for flexible protein ?
2. How to rank native-like conformations among top ranked conformations?

Figure 1.16 Schematic shows docking procedure and two major challenges of docking (Modified from Xue et al., 2015)

Despite recent developments in the international community docking competition–CAPRI (Critical assessment of prediction of interactions) (Lensink & Wodak, 2013), docking still suffers from disadvantages such as rigid body docking ignores flexibility of local interacting regions, particularly in cases where changes in protein conformation occur upon binding (Figure 1.16) (Bonvin, 2006; Zacharias, 2010). Another challenge is development of a robust scoring function to rank docked complexes. While current functions can potentially identify near-native models however, they are not sufficiently accurate and hence, model scoring is an active field of research (Kastritis & Bonvin, 2010; Lensink & Wodak, 2013).

For modeling multidomain proteins, Cheng(Cheng et al., 2008) proposed ranking of the results of rigid body docking using additional restraints derived from the domain linker conformations found in the PDB. They benchmarked their method ona set of 542 linker regions from highly-resolved X-ray structures ranging from 2 to 29 residues. They calculated the end-to-end distance as the distance between the Cα atoms of N-and C-terminal residue for each linker and summarized the data in a length-dependent manner through the mean and standard deviation. They successfully demonstrated that rigid-body docking approach along with energy scoring and linker-based restraints are proved to be highly useful for modelling domain-domain interactions (Cheng et al., 2008).

1.10.3 Domain assembly approach

The domain assembly approach involves modeling of a multidomain protein structure from separate domains, by making use of the knowledge that domains in multidomain proteins are connected via the chain (Cheng et al., 2008; Wollacott et al., 2007; Xu et al., 2014; Zhou et al., 2019). In this approach, binding modes are considerably reduced compared to docking by taking the linker between domains as a tether. A number of approaches have been proposed, which typically keep the individual domain structures unchanged, and alter the inter-domain linker conformation to sample available tethered motions. Those with the lowest pseudo-energy scores are then taken as the final solution from a variety of generated models, similar to model selection in docking.

Recently, an approach is developed called MultiDomain Assembler (MDA) (Hertig et al., 2015) which begins by finding the close non-overlapping templates for the query sequence from BLAST and consequently map the local alignments between template and target onto the target sequence. Depending on the length of inter-domain gaps in the alignment from the previous step, the initial model is built by placing the individual templates at relative distances in order to avoid steric clashes. Finally, MODELLER is used to build any missing regions of the linker and to resolve inter-domain interactions and packing.

Another method known to assemble domains within a multidomain protein is *Ab Initio* Domain Assembly (AIDA) (Xu et al., 2014; Xue et al., 2015) which is a fast energy minimization method guided by *ab initio* folding potential (Xu & Zhang, 2012). In this method, the initial full-length model generated incorporates linker regions modeled based on predicted secondary structure types from PSIPRED (Jones, 1999). In order to sample the range of possible motions, torsion angles of linker backbone are subsequently perturbed. By minimizing the energy functions, the final model is generated, which includes terms to score both the linker conformation and the resulting inter-domain interactions. This approach can be extended to assemble proteins with discontinuous domains. It is available at http://ffas.burnham.org/AIDA/.

A similar approach demonstrated on two domain proteins using ROSETTA method (Wollacott et al., 2007). The starting structures consisted of two domains with a fully extended conformation of the linker. Initially, the linker's conformational space was sampled using a low-resolution search, with the chain represented as the centroids of the backbone and side chain. Following this, more intensive refinement is done after residue side chains are restored within the linker via further small random backbone changes.

1.10.4 Incorporation of predicted interface features in modeling

Docking with knowledge of potential interfacial residues can improve the identification of native like protein-protein interaction docked poses. The sequence conservation feature can be used as interface is relatively more conserved (Littler & Hubbard, 2005). The conserved polar residues form the hot spot regions, which may indicate putative binding sites located at the interface (Hu et al., 2000). Previous protein-protein docking methods have exploited such information to improve prediction (Duan et al., 2005; Oliva et al., 2013). The docking methods with experimentally determined interfacial data (chemical shift perturbation or mutagenesis experiments) such as HADDOCK (de Vries et al., 2010; Dominguez et al., 2003) have been successful in restricting the docking search space.

Similar to guided docking, scoring functions can be improved by including sequence and interface features to improve prediction of interfacial regions. Such features can be used with machine learning approaches to predict binding surfaces. Lise (Lise et al., 2006) used different interfacial features such as residue-pair potentials, shape complementarity, interface propensity, correlated mutations and residue conservation to select native multidomain structures from a set of docking generated models. Similarly, analysis on domain- domain interfaces using both sequence and structural features to train a classifier that can predict intra-molecular domain interfaces (Bhaskara et al., 2014). Docking approach in conjunction with machine learning approaches to predict residues at the interface can improve interface prediction accuracy. If multiple residues at the interface can be known by prediction methods, machine learning methods can be used to filter multiple docking poses with known residues at their docked interface (Li & Kihara, 2012).

While in the recent years, many computational approaches have been developed which are known to have significant contributions toward the discovery and understanding of domain-domain interactions. For the accurate structure prediction of multidomain prediction apart from reliable modeling of individual domains, their relative orientations and interfaces is essential. Moreover, the structural space of interfaces can be explored with the development of a structural alignment of interfaces to understand evolutionary aspect and functional restraint of intra-chain domain interfaces. Below, we briefly outline the thesis work.

1.11 Thesis outline

The objective of the thesis is to perform systematic analyses intra-chain domain interfaces and propose a methodology for multidomain protein structure modeling with improved interface prediction. In this work, we have described systematic and comprehensive analyses on the conservation of intra-chain domain-domain interfaces in multidomain proteins. The domains being compared vary from identical in sequence, closely related, distantly related, or completely unrelated. These can assist in understanding structural and evolutionary constraints on two interacting domains. We have used CATH structural domains and atomic distance criteria were used to define the interface.

First (Chapter 2), we have investigated the physiochemical properties of intra-chain domain-domain interfaces and compared these with protein-protein interaction interfaces. Further, we analyzed the extent of domain interface structural variation inmultiple structures of the same protein to understand factors influencing changes in interfaces. We also studied the variation in domain orientation using inter-domain geometry. After studying the general structural features of domain-domain interfaces, next (Chapter 3), we have investigated whether structurally related domains (at a given level of structural relatedness) would form similar interfaces. This could give structural conservation of interfaces between two interacting domain pairs involving closely or distantly related domains. Here, we have used structural similarity as assessed by IS-score obtained after optimal structural superposition of interfaces. In essence, this would provide the effect of structural similarity of domain pairs on similarity of their interfaces. Additionally, we have analyzed functional constraints on domain interfaces of enzymes by analyzing correlation between domain interface similarity and function overlap as assessed by EC number or GO terms. The previous analysis found similar interfaces in completely unrelated domains that prompted us to investigate whether domain interfaces exhibit structural degeneracy. For this (Chapter 4), we structurally aligned intra-chain domain interfaces of unrelated domains using iAlign and used IS-score to find structural redundancy of interfaces. Further, intra-chain domain interfaces were also compared with inter-chain domain interfaces from protein-protein interactions. This showed that a combined template interface library could be constructed by including both intra/inter-domain interfaces. In the last Chapter 5, we studied whether domain interface similarity could be used to find near native interfaces among various possible docked poses obtained from simple rigid body docking of domains using Z-DOCK. In our approach, all possible interactions obtained from rigid body docking domains and this list of docked complexes is ranked by IS-score by aligning them to template interface library. This has potential application in modeling multidomain structures of proteins. Figure 1.17 summarizes thesis main objectives.

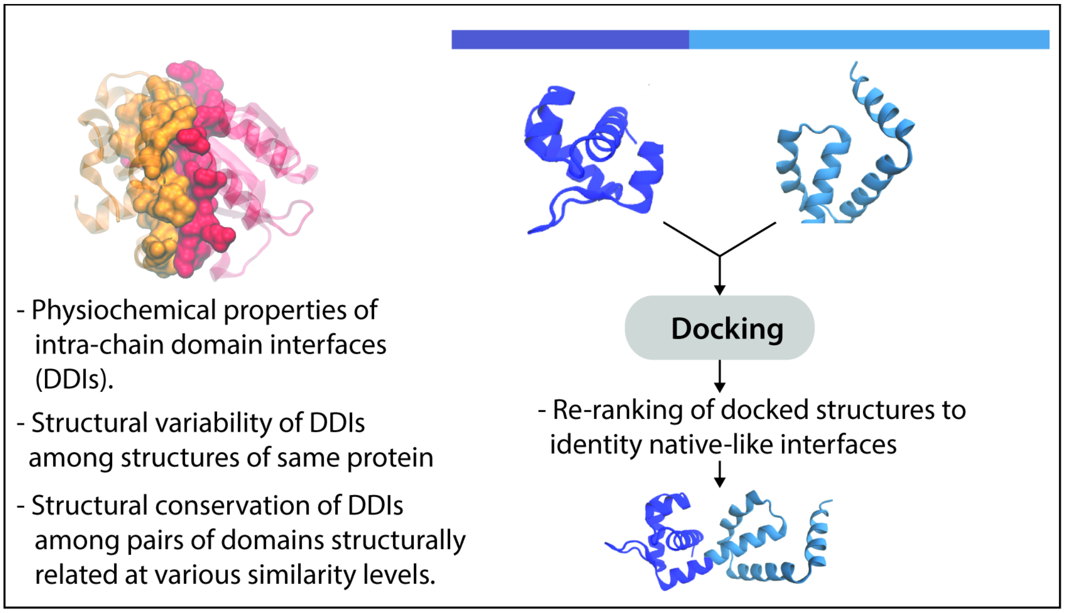

- Physiochemical properties of intra-chain domain interfaces (DDIs).
- Structural variability of DDIs among structures of same protein
- Structural conservation of DDIs among pairs of domains structurally related at various similarity levels.
Docking
- Re-ranking of docked structures to identity native-like interfaces

Figure 1.17 Overall aim of the thesis

Chapter 2

A detailed analysis of sequence and structural properties of intra-chain domain-domain interaction interfaces (DDI's)

2.1 Introduction

Proteins rarely perform their molecular function in isolation. In fact, proteins are involved in physical interactions with other proteins, DNA/RNA, and chemical compounds (ligands) to mediate their function in a biological process (Bronowska, 2011; De Las Rivas & Fontanillo, 2012). Since inter-molecular interactions are essential for protein function, several studies have focused on determining the molecular basis of protein-protein or protein-ligand interactions (Hernández-Santoyo et al., 2013). Moreover, such detailed insights could be used for designing or identifying known chemical compound inhibitors to abrogate either protein-protein or protein-ligand interactions that can serve as lead for therapeutic discovery (Lionta et al., 2014; Taylor et al., 2008; Zinzalla & Thurston, 2009). Previous work in recognizing the molecular basis of inter-molecular interactions have found that among many features, shape complementarity between interaction patch on protein surface with the interacting protein/ligand is an important contributing factor (Connolly, 1986; Keskin et al., 2016; Scott et al., 2016). Thus, a proper shape complementary between interacting partners is important in differentiating and assessing genuine interactions (Lawrence & Colman, 1993). Apart from this, it has been observed that protein interaction surface patches are usually large and in general interactions are favored by a number of forces such as hydrogen bonding, ionic interactions, Van der Waal's forces, disulfide bridges and hydrophobic packing (Yang et al., 2016). There have been significant progress in understanding inter-molecular interactions at atomic level by analyzing sequence and/or structural features to uncover the mechanism of metabolic and signal transduction networks (Bahadur & Zacharias, 2008; Caffrey et al., 2004; Chakrabarti and Janin, 2002; Chakravarty et al., 2013; Chothia & Janin, 1975; Gaines et al., 2018; Glaser et al., 2001; Hou et al., 2017; Jones and Thornton, 1996; Keskin et al., 2008; Lo Conte et al., 1999; Nooren & Thornton, 2003; Ofran & Rost, 2003; Wodak & Janin, 1978; Yan et al., 2008). Many of these are also exploited in prediction of residues that participate in protein–protein interactions, which can aid in experimental studies (Brender & Zhang, 2015; Ortiz et al., 1999; Rao et al., 2014) to disrupt protein-protein interactions or improve existing networks. The protein-protein interaction network provides a way of information transfer required during biological processes.

Since proteins are composed of domains, it is also important to understand the communication within protein involving domains necessary for regulation and function of protein (Sistla et al., 2005). Initial studies on characterizing domain-domain interactions were carried in early 2000's with small dataset that analyzed and compared the physiochemical properties of domain-domain interfaces with protein subunit interfaces (Argos, 1988; Jones et al., 2000). It was observed that in general nature of domain interfaces is intermediate between permanent and non-obligate protein interfaces. Further, the work of Jones and Thornton (Jones et al., 2000) found that protein-protein and domain-domain interactions in multidomain proteins are similar in terms of physical and geometrical properties. These were shown to have paramount implication in domain swapping mechanism, which is known for the formation of oligomeric proteins from monomeric and results in the transition of inter-domain interaction sites to inter-subunit sites (Bennett et al., 1995; Schlunegger et al., 1997). This is seemingly feasible because interaction sites share common characteristic features (Jones et al., 2000).

In this work, we have extended the study on physiochemical comparison of intra-chain domain-domain interfaces with protein-protein interfaces by performing the same on a large dataset. Next, to gain insight into structural variability of intra-chain domain interfaces of proteins, we have analyzed the changes in domain-domain interfaces of multidomain proteins of same protein having multiple experimentally determined structures.

This chapter is broadly divided in following two sections: a. Comparison of physiochemical properties between intra-chain and protein-protein interaction (PPI) interfaces and b. Analysis of changes in intra-chain domain interfaces of same protein.

2.2 Comparison of physiochemical properties between intra-chain domain and protein- protein interaction interfaces

In the first section, we have compared several ways to define intra-chain domain interfaces. Subsequently, using these we have compared the physiochemical features between intra-chain domain and PPI interfaces to understand commonalities in these interfaces.

2.2.1 Materials and methods

In our study, we have used structural domain definitions from CATH structural domain database (Orengo et al., 1997). The construction of CATH database involves first delineating domains and subsequently, classification of these into appropriate hierarchical level. Hence, it is possible that domains are delineated for a structure without domain being classified in CATH. For all our analysis, we have used defined and classified structural CATH domains.

2.2.1.1 Construction of non-redundant two-domain dataset

We constructed a two-domain non-redundant dataset in order to compare various methods of defining intra-chain domain interfacial residues. Since most multidomain proteins (46%) documented in CATH domain database are two domain proteins, we constructed database of proteins having two classified continuous domains. The flowchart of dataset generation is shown in Figure 2.1A. Briefly from CATH domain database (v 4.1.0), we extracted two continuous domain protein structures with resolution ≤ 2.5 Å. This resulted in 19,770 proteins, which were made non-redundant at 40% sequence identity using CD-HIT (Li &Godzik, 2006). The atom record sequence of tertiary structure was used for preparing non-redundant dataset. Thus, we obtained a set of 1729 non-redundant two continuous domain proteins. This dataset is referred to as 2-dom-cont-DDI. The dataset *C2-two-dom-cont-DDI* is provided as excel sheet inthe supplementary file Chap02-Dataset.xlsx available in the GitHub repository URL: https://github.com/riviverma/thesis-md-proteins/

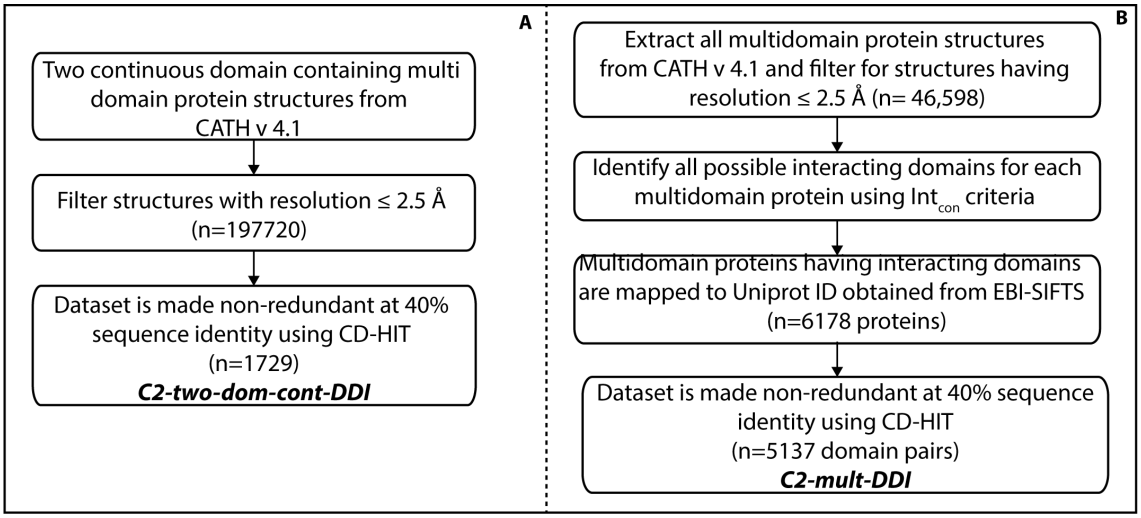

A
Two continuous domain containing multi domain protein structures from CATH v 4.1
Filter structures with resolution ≤ 2.5 Å (n=197720)
Dataset is made non-redundant at 40% sequence identity using CD-HIT (n=1729)
C2-two-dom-cont-DDI
B
Extract all multidomain protein structures from CATH v 4.1 and filter for structures having resolution ≤ 2.5 Å (n= 46,598)
Identify all possible interacting domains for each multidomain protein using Int con criteria
Multidomain proteins having interacting domains are mapped to Uniprot ID obtained from EBI-SIFTS (n=6178 proteins)
Dataset is made non-redundant at 40% sequence identity using CD-HIT (n=5137 domain pairs)
C2-mult-DDI

Figure 2.1 Overview of datasets construction. Schematic flowchart showing steps in generating A) C2-two-dom-cont-DDI and B) C2-mult-DDI datasets.

2.2.1.2 Construction of non-redundant domain-domain interaction (DDI) dataset

In order to compare various structural/sequence properties between PPI and intra-chain interfaces we constructed non-redundant DDI dataset. As mentioned before, we used CATH domain database (v4.1.0) to derive domain definition and classification. Here, we considered both continuous (having single segment) and discontinuous (consists of more than one segment) domains. The schematic flowchart is shown in Figure 2.1B. We extracted all multidomain proteins (≥ 2 domains) from CATH and took structures with resolution ≤ 2.5 Å for further processing. From this set, we prepared all possible domain combinations and extracted intra-chain interacting domains based on the interatomic contact criteria (details are given in section 2.2.2). In order to remove trivial redundancy, *i.e.* multiple structures of a given protein sequence, we used mapping of PDB ID (with chain) to Uniprot identifier ("UniProt: The Universal Protein Knowledgebase" 2017) available in EBI-SIFTS database (Velankar et al., 2013). Thus, interacting domains were mapped to Uniprot identifier. For a given Uniprot identifier, we took a representative structure having maximum combined length of two domains. Thus, we obtained a set of 52,849 intra-chain interacting domains. Finally, this dataset was made non-redundant at 40% sequence identity using a protocol (described below), which ensures that at least one of the domains is non-redundant after CD-HIT clustering. Finally, we obtained 5137 non-redundant intra-chain interacting domains (*C2-mult-DDI*). The dataset is provided as excel sheet in the file

Chap02-Dataset.xlsx available at https://github.com/riviverma/thesis-md-proteins/. Below we discuss the protocol used in constructing non-redundant dataset.

Full-length protein sequences are usually used in generating non-redundant DDI datasets. However, such non-redundant dataset does not have information of non-redundancy available at the level of domains because domain boundaries are not used as an input for alignment or extracting words as in case of CD-HIT (Li & Godzik, 2006). Moreover, such approach of using full-length sequence cannot be used for non-consecutive domains, as these are two distinct regions of the protein sequence. In order to generate non-redundant dataset at the level of domains, we have designed a simple method, which considers non-redundancy at the level of domains and ensures that at least one domain has the minimum desired non-redundant level.

In this procedure, we cluster all domains using CD-HIT at 40% sequence identity or desired identity threshold. This results in clusters having domain entries, which have PDBID followed by chain identifier with domain numbers. Each cluster is numbered from 1 to N, where N is the number of clusters. Next, we generate combinations of clusters *i=(1 to N)* and *j=(1 to N)* such that cluster numbers $i< j$ (i and j are cluster numbers), essentially the upper triangular matrix of N cluster combination matrix. For each such combination of cluster, first, the common PDB entries having different domain numbers between two clusters are identified. Then, depending on the number of structures (zero, one or more) identified in previous step, following is performed: a) if there are zero common cluster members (structures), then no domain pair structure is selected, b) if only one common entry exists, then it is taken as representative structure, and c) if there is more than one common PDB entry, then a representative non-redundant structure is selected that has the highest (best) resolution. There are some domains, which are clustered as single unit given the CD-Hit algorithm. We specifically checked for these cases and calculated the sequence identity of this lone domain with its corresponding domains in different proteins. If the sequence identity is > 40%, we merge this domain in the other cluster; otherwise it will be included in the final dataset. We also made sure that we conserve the domain order in a given protein. In cases, where domain order is reversed we selected both the domain pairs in the non-redundant dataset (Figure 2.2).

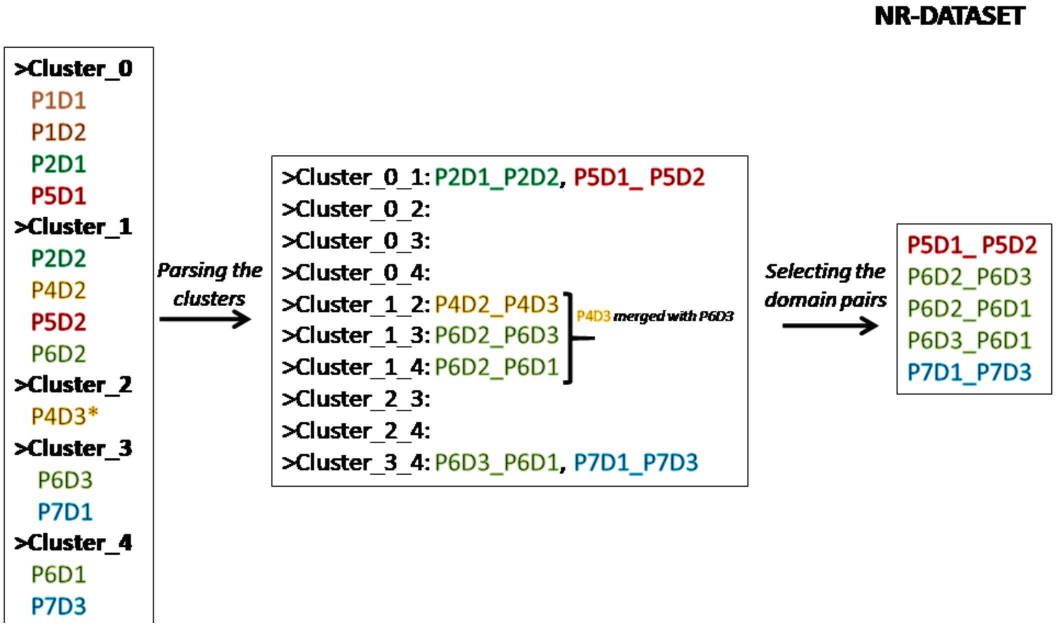


Figure 2.2 Overview of the protocol to generate non-redundant domain- domain interacting pairs. Figure showing representative clusters in an example list of domain entries. P4D3 is a lone member, which shares more than 40% sequence identity with P6D3 and got merged in cluster4. Additionally, in order to conserve the order of domains, both pairs (P6D3_P6D1 and P7D1_P7D3) are considered in the final non-redundant dataset.

2.2.1.3 Protein-protein interaction dataset (PPI dataset)

We used previously described non-redundant PPI dataset of 1517 protein dimers (Gao & Skolnick, 2010b). The proteins are interacting was defined based on interatomic contact criteria, which we have also used in the generation of DDI dataset. The dataset (C2-int-PPI-data) is provided in the file Chap02-Dataset.xlsx at https://github.com/riviverma/thesis-md-proteins/.

2.2.2 Defining domain-domain interactions

In most studies, the identification of protein-protein interaction interfaces in multi-chain tertiary structures rely on criteria such as size of interface, which is solvent accessible area buried upon complex formation; change in residue solvent accessible area between complex and free form of the protomers in an oligomer; and interatomic contact criterion. Using the same set of criteria, we defined intra-chain domain interfaces and compared the variations in interfaces because of interface definitions. The criteria for identifying intra-chain domain interfaces are:

a. Buried surface area (BSA): It is the size of interface measured as solvent accessible surface area (SASA) buried upon complex of two domains together (Chakravarty et al. 2013) in comparison to separated domains.It is given by equation below:

   Buried surface area = $\text{SASA}_{\text{domain1}}$+$\text{SASA}_{\text{domain2}}$ – $\text{SASA}_{\text{(domain1+domain2)}}$

   It has been advocated to use a minimum BSA of ~800Å^2 to distinguish specific or biologically relevant interactions from non-specific PPI (Bahadur et al., 2004; Deremble & Lavery, 2005). Alternatively, a minimum BSA of ~400Å^2 of one chain also has been used in some studies (Yan et al., 2008). Another two methods to identify interacting domains are based on change in residue SASA from accessibilities calculated in an isolated domain and two domains taken together.

b. Interface constitutes a set of residues, which undergo change in >1Å^2 SASA between accessibilities calculated in an isolated domain and two domains (Susan et al., 2000). Here after it is referred to as $\text{Int}_{\text{ASA-1.}}$

c. A residue is defined to be at interface if solvent accessibility is ≥ 10% in isolated domain and the same is ≤ 7% in complex with another domain (Rekha et al., 2005). This is a strict definition to detect interfacial residues and this is referred to as $\text{Int}_{\text{ASA-2.}}$

d. In our study, we have used interatomic contact criterion to define interfacial residues. Based on this, if any heavy atom of a residue in a domain is within 4.5Å of another heavy atom of a residue from another domain, these residues are said to be in contact or lying at interface. These residues constitute a set of interfacial residues. A minimum number of twenty interfacial residues are used to define interaction between two proteins (Deremble & Lavery, 2005; Fischer et al., 2007; Gao & Skolnick, 2010b). We follow the same condition of at least 20 interfacial residues to define intra-chain interacting domains. This is followed to define interacting domains throughout in this thesis. This is referred to as Int_{con} definition of interfacial residues.

2.2.3 Description of features used for comparison of interfaces

We systematically compared various physiochemical, structural features and amino acid propensities of intra-chain domain interfaces with PPI interfaces to find commonalities between them as well as to identify key distinguishing features of either interface. PPI interfaces have been extensively characterized in terms of their physiochemical properties and used in characterizing biologically relevant interfaces or finding hot spots (Elez et al., 2018; Hamon & Morelli, 2013; Macalino et al., 2018; Yan et al., 2008). For comparison of interfaces, we have used following features:

2.2.3.1 Solvent Accessible Surface Area

The solvent accessible surface area is defined as surface area given by the center of spherical probe rolling over a molecule. The probe used is usually water molecule (Lee & Richards, 1971). SASA or simply accessible surface area (ASA) has been regularly used in characterization of protein structures such as in studying protein folding (Auton & Bolen, 2005; Guinn et al., 2013; Miller et al., 1987) implicit solvent effects (Weiser et al., 1999) and in characterizing, distinguishing specific/non-specific and prediction of interfaces (Elez et al., 2018; Jones & Thornton, 1996; Xue et al., 2015; Carugo &Argos, 1997; Dasgupta et al., 1997; Henrick &Thornton, 1998; Janin, 1997; Janin &Rodier, 1995; Ponstingl et al., 2000; Sriwastava et al., 2013; Zhu et al., 2006).

The SASA was calculated using NACCESS (Hubbard and Thornton, 1993), which implements Lee and Richard algorithm. The absolute SASA was used to compute buried surface area as has been defined in section 2.2.2. In order to define whether a residue is buried or exposed in protein structure, we used relative ASA (rASA) as given by NACCESS program. Relative ASA is fraction of accessible surface area of a given amino acid in protein structure to ASA of the same in its conformational expanded state, which is given by A-X-A (X is amino acid and A is Alanine). A residue with rASA>5% is regarded as exposed or surface residue (Miller et al., 1987).

2.2.3.2 Secondary structure content

We obtained the secondary structure elements *viz.* helices, sheets, turns and coils of protein structure from output of STRIDE (Heinig &Frishman, 2004) program. We compared the number and type of secondary structure elements at the interfaces.

2.2.3.3 Hydrogen bonds

A hydrogen bond isone of the most important intermolecular interaction force, which is formed when a hydrogen atom is covalently attached to one electronegative donor shared with another electronegative atom (acceptor) of same or different molecule (Jones &Thornton, 1996). It confers specificity and directionality to intermolecular interactions. The geometrical parameters of hydrogen bonds are mostly derived from crystal structures (Hubbard &Haider, 2010). The analysis of protein-protein interfaces have shown that the hydrogen bond geometry is not optimal and in general is weaker compared to intra-chain hydrogen bonds. Additionally, water molecule mediate hydrogen bonds for non-optimally oriented donor/acceptor atoms (Xu et al., 1997). The contribution from hydrogen bond has been used as one of interaction energies in prediction of protein-protein interactions (Sukhwal &Sowdhamini, 2013). We have used output of STRIDE to identify hydrogen bonds at interfaces.

2.2.3.4 Disulfide bonds

The tertiary structure of proteins is stabilized by numerous covalent and non-covalent interactions. Disulfide bond is a covalent bond between sulphur (S) groups of two cysteine residues. The connection between Sγ of two cysteine residues makes a -Cβ-Sγ-Sγ-Cβ- bond. The disulfide bond can be formed intra-molecularly (within a single polypeptide chain) where they stabilize the tertiary structure or inter-molecularly (between two polypeptide chains) where they are involved in stabilizing quaternary structure of the protein. The distance between sulphur group of two cysteine residues ≤ 2.2 Å is used as a cut-off to define inter-domain disulfide bond (Bhattacharyya et al., 2004). In our analysis, we considered only inter-chain (PPI) or intra-domain disulfide bonds.

2.2.3.5 Amino acid propensity

The interface amino acid propensity can be considered as a measure for preference of amino acid occurrence at interface. This also represents the composition of amino acids and their relative importance at the domain interface. Interface residue propensity is calculated using the equation given below:

$$Residue\ interface\ propensity\ AA_j = \frac{\sum_{k=1}^{N} nAA(I)_{j,k} \Big/ \sum_{k=1}^{N}\sum_{j=1}^{20} nAA(I)_{j,k}}{\sum_{k=1}^{N} nAA(S)_{j,k} \Big/ \sum_{k=1}^{N}\sum_{j=1}^{20} nAA(S)_{j,k}}$$

where, nAA(I)$_{j,k}$ is the number of amino acid (j) in the interface of protein/domain (k); nAA(S)$_{j,k}$ is the number of amino acid (j) on protein surface in protein/domain (k); N is the number of proteins/domains. A residue is considered to be exposed to solvent when rASA> 5%.

Briefly, it is the ratio of amino acid relative occurrences observed at interface and protein/domain surface. An interface residue propensity >1.0 indicates that a residue type has more likelihood to be present at domain interface.

2.2.4 Results

In our work, we have defined two domains as interacting if there are at least 20 interfacial residues identified based on interatomic contact. In the absence of established criteria to identify DDIs, we have compared various known ways of defining interfaces in terms of the number of interfacial residues. This will provide an estimate of overlap region from other approaches.

2.2.4.1 Comparison of different methods to define an interface

2.2.4.1.1 Comparison of buried surface area to the number of interfacial residues

Using the dataset of 1729 two continuous domains (C2-two-dom-cont-DDI), we compared the number of interfacial residues identified using interatomic contacts (Int$_{con}$) with buried surface area for 1729 non-redundant two continuous domain dataset. As is observed in Figure 2.3A, number of interface residues is linearly correlated (r^2=0.98) to buried surface area. Considering at least 20 interacting residues to define interacting domains, ~80% of domain pairs are found to be interacting. Importantly, the BSA of domain pairs having 20 residues at interface is on average ~820 Å^2, which is close to the definition used for biologically relevant protein-protein interfaces (~800Å^2). Further, interface defined using Int$_{con}$ was compared to Int$_{ASA-1}$ and Int$_{ASA-2}$ criteria, which uses change in ASA between free and complex state to define interfacial residues (see methods). The comparison of the number of interfacial residues identified using Int$_{con}$ to Int$_{ASA-1}$ and Int$_{ASA-2}$ are shown in Figures 2.3B and 2.3C respectively. As can be seen, method of

Susan et al., (Int$_{\text{ASA-1}}$) results in slightly more number of interfacial residues as compared to Int$_{\text{con}}$ method with mean (SD) sizes of interface 57.4(28.6) and 52.2 (26.8) respectively. The Int$_{\text{ASA-2}}$, Rekha et al., method is very stringent with mean interface residue of 15.2(11.7).

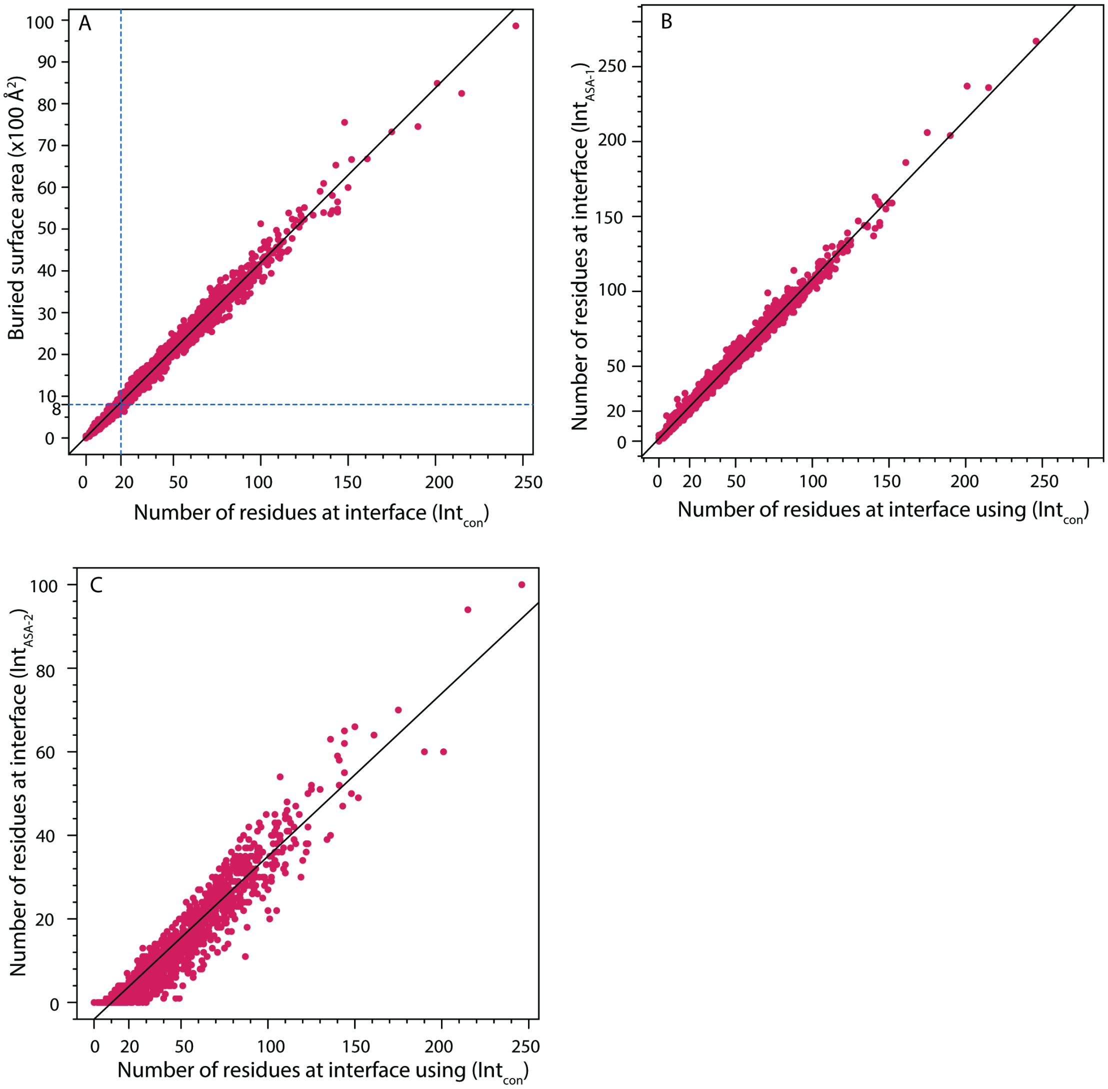


Figure 2.3 Scatter plot showing the comparison of number of interfacial residues of Int$_{\text{con}}$ to other ways of defining interface. Scatter plot showing the number of interfacial residues defined using Int$_{\text{con}}$ to A) Buried surface area (BSA), B) number of interfacial residues as defined in Susan et al., (2000) Int$_{\text{ASA-1}}$ C) number of interfacial residues as defined in Rekha et al., (2005) Int$_{\text{ASA-2}}$. The best-fit line is shown in all panels.

Thus, suggesting Int$_{con}$ method identifies size of interface close to what we observed in biologically relevant PPI. Additionally, in order to define biologically valid interfaces, having atleast 20 residues at interface, results in ~80% of domains as interacting. This means that using a stringent definition of iAlign on domain dataset, we are able to fetch valid interfaces. Henceforth, we have used interatomic contact criteria with at least 20 residues to define interacting domain-domain interfaces.

2.2.4.2 Comparative analysis of physiochemical properties between DDI and PPI

The objective in this work is to explore and possibly identify distinguishing physiochemical, sequence or structural features of domain-domain interfaces in multidomain proteins. Here, we have compared following features: solvent accessible surface area, residue propensity and hydrophobicity, prevalence of hydrogen bond, disulfide bonds and secondary structure content of domain interfaces.

Before characterizing different interface properties, we analyzed contribution of DDIs from multidomain proteins. The statistics is summarized in Figure 2.4. From CATH database, we considered protein structures having resolution ≤2.5Å and made the dataset non-redundant at 40% sequence identity. This set was used to represent number of interacting domain pairs out of total number of domains in multidomain proteins.

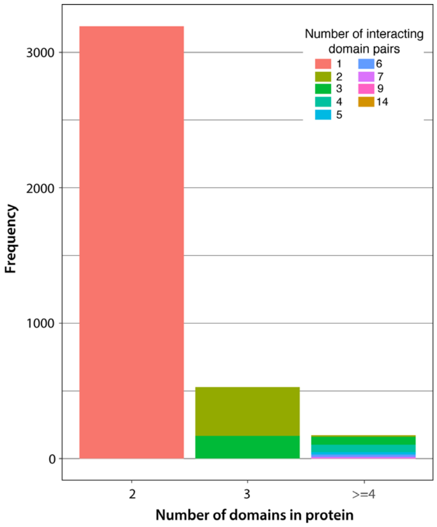


Figure 2.4 Distribution of interacting domains (CATH) in multidomain proteins. Figure showing the frequency of interacting domain pairs in multidomain proteins. The x-axis represents the total number of domains in a protein. The color represents the number of interacting domain pairs.

2.2.4.2.1 Analyzing physiochemical properties of DDI’s and PPI’s

We compared various physiochemical features of interacting domain-domain with protein-protein interfaces. First, we analyzed the sizes of DDI and PPI interface using buried surface area as a measure of interface size as well as the strength of binding (Jones &Thornton, 1996). We observed that mean (standard deviation (sd)) buried surface area of 2132 (1009) Å^2 and 3722 (2345) Å^2 for DDI and PPI respectively. This shows that protein-protein interfaces are larger in size compared to DDI as can also be seen in the distribution of BSA (Figure 2.5A). Next, we also compared the interface in terms of number of interfacial residues. This distribution is shown in Figure 2.5B. The observed mean (sd) number of interfacial residues in DDI and PPI are 50.8(23.8) and 86.4 (53.4) respectively. Since PPI interfaces can constitute more than one domain, it could be the reason of the large relative interface sizes. Another possibility is that domain linkers can potentially restrict the size of interface or contacts between two domains resulting in small interface sizes (Jones et al., 2000). Subsequently, we analyzed the hydrophobic/polar nature of interfaces. For this, we considered contribution of non-polar/polar surface area to interface region (buried surface area). The distributions of non-polar and polar buried surface area for DDI/PPI are shown in Figures 2.5C and 2.5D respectively. As can be seen both polar/non-polar buried surface area is less for DDI than PPI that could also be because of small interface area of DDI.

Since interface sizes are different in DDI and PPI, we compared the relative contribution of hydrophobic surface to interface (Figure 2.6). On an average, hydrophobic surface contributes 65.8% to DDI interface and 66.3% of the same is observed in PPI. This shows that interfaces are dominantly hydrophobic, which has been observed in PPI (Tsai et al., 1996). Importantly, nature of hydrophobicity of DDI is comparable to that of PPI.

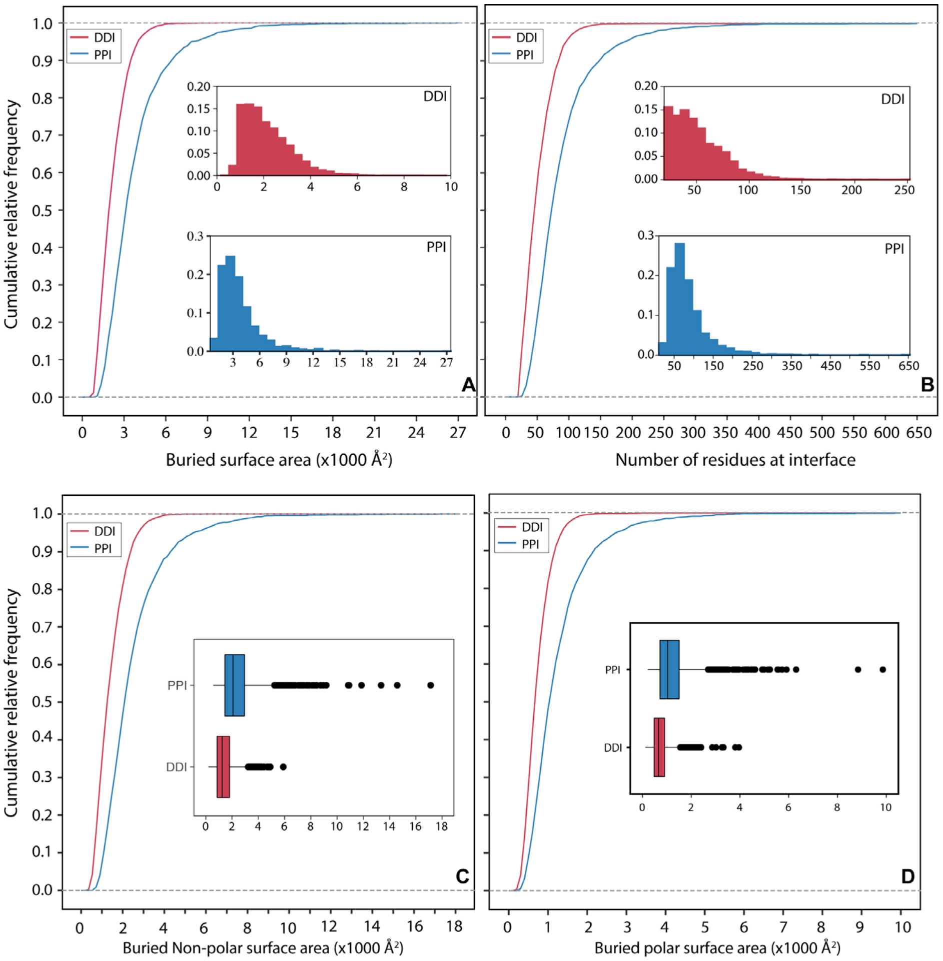


Figure 2.5 Comparison of various physiochemical features between intra-chain domain and protein-protein interfaces. Cumulative distribution of interface features for A) Buried surface area (BSA); B) the number of interfacial residues; Buried C) Non-polar ASA; D) Polar ASA.

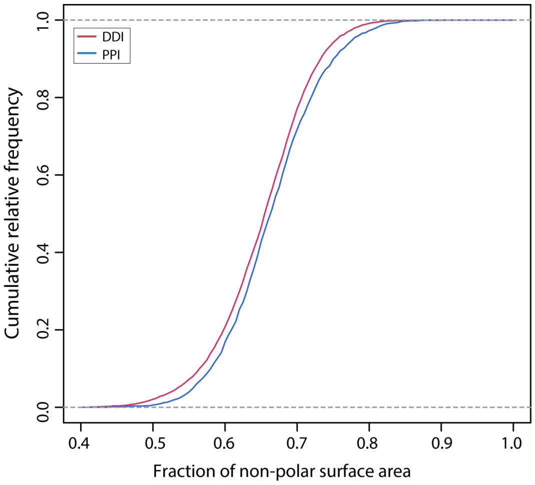


Figure 2.6 Comparison of relative hydrophobicity. Cumulative distribution of relative hydrophobic BSA of interfacial region of DDI and PPI.

Usually, a greater number of stabilizing interactions are observed in large interfaces that contribute to stability of interaction interfaces (Pace et al., 2014). We characterized these stabilizing interactions such as the number of hydrogen and disulphide bonds observed at DDI/PPI interface. As mentioned in methods, hydrogen bonds are identified using STRIDE program and a cut-off distance of 2.2Å between Sγ atoms of Cysteine residues is used to define a disulphide bond at the interface. On an average there are ~2 and ~4 hydrogen bonds per interface in DDI and PPI respectively. However, frequency of occurrence of these interactions is directly proportional to the size of interface or the area of contact. Hence, we normalize the number of hydrogen bonds with number of residues at the interface. The distribution is shown in Figure 2.7A and summary is shown in Table 2.1. Similarly, we calculated number of potential disulphide bonds at interfaces. The distribution and summary of disulphide bond is shown in Figure 2.7B and Table 2.1 respectively. These analyses showed that DDI has smaller interface compared to PPI. Despite this, the relative interaction features such as hydrophobic, hydrogen bonds, and disulphide bonds are comparable between PPI and DDI. Thus, suggesting physiochemical features of intra-chain domain interaction and PPI interfaces are similar despite differences in their interface sizes.

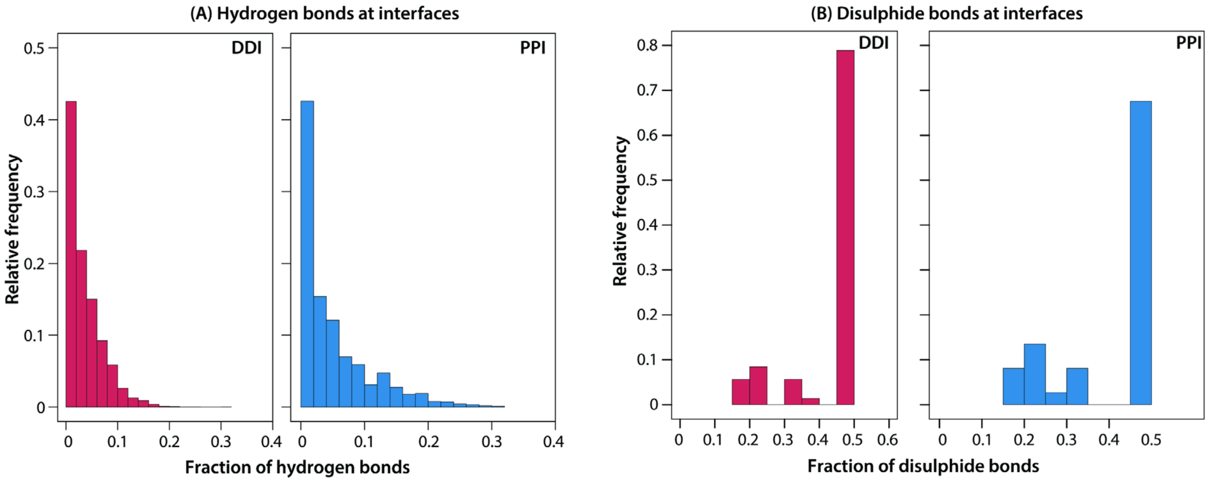


Figure 2.7 Comparison of relative abundance of hydrogen and disulphide bonds between intra-chain domain and protein-protein interfaces. Histogram showing relative frequency distribution of normalized: A) number of interface hydrogen bonds, and B) possible number of disulphide bonds.

Table 2.1 Summary of mean (SD) of various interface features compared between DDI and PPI interfaces

Interface parameters	DDI (5137) Mean (SD)	PPI (1517) Mean (SD)
Total buried ASA (Å^2)	2132.4 (1008.9)	3722.2 (2345.1)
Polar buried ASA (Å^2)	727.5 (341.5)	1250.5 (843)
Non-polar buried ASA (Å^2)	1405.0 (706.2)	2471.7 (1559.8)
Hydrogen bond	0.034 (0.036)	0.049 (0.059)
Disulphide bond	0.451 (0.10)	0.422 (0.118)

Next, we analyzed preferential occurrence of amino acids at interfaces by computing amino acids propensities at interfaces. In amino acid propensity calculation, we have normalized the frequency of amino acid occurrence at the interface to the same on protein surface (see methods). The surface exposed residues are defined based on relative ASA of a residue in the protein structure. As can be observed in Figure 2.8, hydrophobic amino acids (TRP, ILE, LEU, MET, PHE, and VAL) are prevalent at domain interfaces. This is consistent with the observation of buried surface area being dominated by hydrophobic surfaces. Apart from this, polar residue TYR and CYS are also found to be dominantly present at interfaces. Importantly, the propensities observed at DDI are similar to the ones at PPI. ARG has been frequently observed to be located in hot spots regions of protein-protein interfaces in binding energy experiments due to its positively charged side-chain interactions with TYR, TRP and PHE (Bogan &Thorn, 1998; Glaser et al., 2001). In our results, we observed that propensity of ARG is nearly same and approximately approaches propensity value of one in both DDI and PPI. HIS also appeared to be prevalent at the interfaces as it is involved in making Π-Π stacking interactions with TRP, TYR and PHE (Liao et al., 2013).

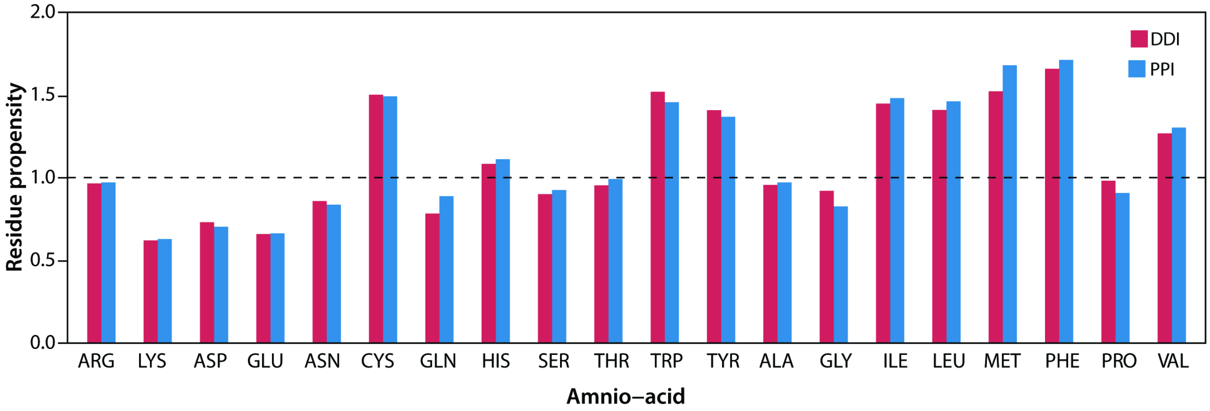


Figure 2.8 Amino acid propensities of DDI and PPI interfaces. Bar plot showing propensities of amino acids observed at intra-chain domain and PPI interfaces.

Next, we compared the secondary structural content of the interfaces in DDI and PPI. In our analysis, we considered the following secondary structural motifs namely, helices, strands, turns, coils (Kabsch &Sander, 1983). As can be seen in Figure 2.9, we found all major secondary structure types occur at interacting interfaces, with relatively more prevalence of regular secondary structure elements such as helices and sheets. There is no significant difference in the secondary structure content between DDI and PPI interfaces.

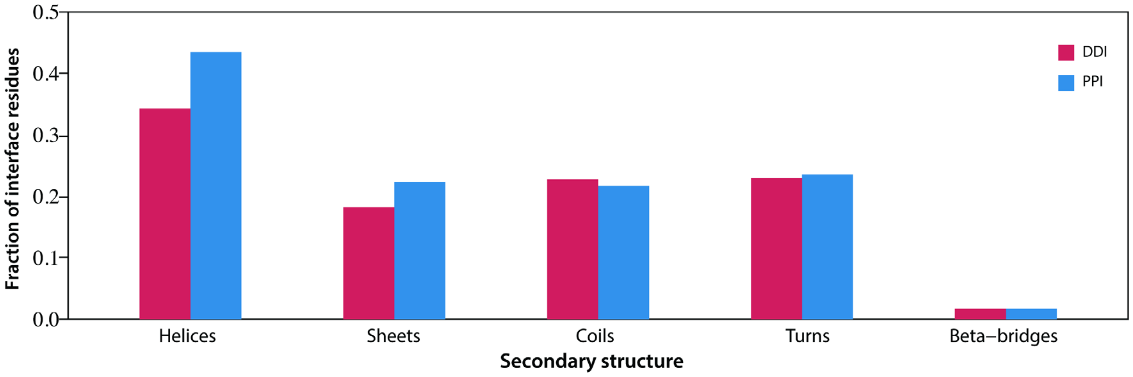


Figure 2.9 Distribution of secondary structural elements. Figure showing occurrence fraction of secondary structural elements at DDI and PPI interfaces

In this work, we have compared several physiochemical, sequence propensity and structural content between domain- domain interfaces and protein- protein interfaces. These have been studied in order to gain insight into interfaces either of PPI or among domains in multidomain proteins. The analysis on comprehensive dataset suggests that in general buried surface area of intra-chain domain interfaces is less than PPI. Despite this, most physiochemical features are comparable between DDI and PPI interfaces. Thus, suggesting interfaces formed either between two chains or within a chain share similar characteristics.

2.3 Conservation of intra-chain domain-domain interfaces (DDI) and their geometry in experimentally known tertiary structures of same protein

2.3.1 Background

Proteins even in their global lowest energy state can sample a wide array of conformations and many of these conformers are important for the function of protein (Lindorff-Larsen et al., 2005). However, a single model of the protein structure is thought to represent an incomplete content of information regarding what defines as a biological molecule and therefore, it alone cannot be used for precise functional annotation of protein and its mechanistic interpretation (Srivastava et al., 2012). But, still most of the structural studies of proteins use solid, rigid, crystal structures of protein solved by dominating experimental technique called X-ray crystallography (Burra et al., 2009). X-ray method produces only a single model from the protein ensemble and accounts for more than 90% of structures deposited in the PDB database (Berman et al., 2000; Krishnan &Rupp, 2012). In line with this fact, it has been observed that PDB consists of redundant entries of multiple different models of the same protein and on an average, each protein structure is represented more than 4 times (Burra et al., 2009). An analysis of these alternative conformers of identical proteins can provide significant insights into protein’s intrinsic conformational variability as well as we can study their response to various environmental changes. A study has shown that 2 or more models of the protein in PDB vary approximately up to 0.4 Å and these variations can be as large as tens of angstroms when the crystallization condition changes (Berman et al., 2000; Mowbray et al., 1999).

From earlier studies, it is known that residues at the interface can undergo relatively large conformational changes than rest of the protein structure (Betts &Sternberg, 1999; Chakravarty et al., 2015; Rajamani et al., 2004). These include changes in formation of specific interactions such as hydrogen bond or refinement of shape complementarityat interface of proteins (Janin &Chothia, 1990). There has been considerable work on documenting domain motions in proteins (Bennett et al., 1984; Kobayashi et al., 2015; Lee et al., 2003; Ravera et al., 2014). However, most of these concerns with changing the relative orientation of domain as well as characterizing the required bending, twist involved in resulting. But, the effect of these domain motions on the intra-domain interfaces has not been extensively studied. The insights into structural changes at interfaces are important as these can assist in accurate modeling of domain-domain interfaces in multidomain protein that plays essential role in multidomain protein function and stability.

In this section of the chapter, we have systematically performed an analysis to investigate the extent of conformational variability at domain-domain interfaces in proteins having multiple experimentally determined structures. Such studies can assist in modeling intra-chain domain interactions especially, when it is known that protein-ligand interactions can lead to conformation change in protein structures.

2.3.2 Methodology

The aim of this study was to analyze the extent of structural variations in protein structures experimentally determined for the same protein sequence. For this analysis, we relied on structural domain database CATH (v3.5.0) to extract multidomain proteins. Using this database, we constructed a non-redundant (at 70% sequence identity) dataset of multidomain proteins following the methodology schematically shown in Figure 2.10. Briefly, protein structures having at least two delineated and defined domains in CATH database were extracted and mapped to Uniprot unique identifier (“UniProt,” 2017). These mappings between PDB and Uniprot IDs were obtained from EBI-SIFTS database (Velankar et al., 2013). Subsequent to mapping of PDB and Uniprot IDs, we ensured that PDB sequence as given in ‘ATOM’ record is at least 70% of the length of sequence in Uniprot. This is to avoid alignment issues in the next stage. The steps described below are followed to maintain a consistent residue numbers across multiple structures of same sequence and facilitates comparison of protein structures. Since structures of same sequence can have different PDB residue numbers, we performed global sequence alignment of PDB and Uniprot sequences usinga locally implemented Needleman-Wunsch algorithm (Needleman &Wunsch, 1970) of global sequence alignment. Based on the global sequence alignment, PDB residues were renumbered to corresponding residue position from the sequence in Uniprot for PDB ids having sequence identity ≥ 95% to Uniprot sequence. Rest other PDB ids (sequence identity < 95%) were not considered for residue mapping. Thus, we obtained a list of Uniprot identifiers with mapped PDB ids along with chain ID (renumbered residue number). For each Uniprot id, if there are multiple PDB chains associated with it, then one of the longest proteins as a representative structure for the PDB id was selected. Following this, all possible combinations of structural domainsin the dataset were generated, and identified proteins having interacting domains relying on interatomic contact criteria of 20 residues at the interface. Finally, we have Uniprot mapped to PDB id (with chain id) and associated information whether it has intra-chain interacting domains. We removed protein sequences (Uniprot ids), which have only one PDB chain, or having no intra-domain interactions in any possible combination of domains. In case we have only one structure associated with Uniprot, attempts were made to include NMR structure to make the dataset comprehensive. Hence, we have composite dataset of both X-ray and NMR mapped structures. The Uniprot sequence was made non-redundant at 70% sequence identity using CD-Hit program (Velankar et al., 2013). Since CATH domain number is not in the order of N-to-C of protein sequence, we ensured that domains are linearly arranged properly. The C2-mult-ddi-data2 dataset consists of 1489 proteins. The dataset is available as excel sheet in the file Chap02-Dataset.xlsx available at GitHub repository, URL: https://github.com/riviverma/thesis-md-proteins/.

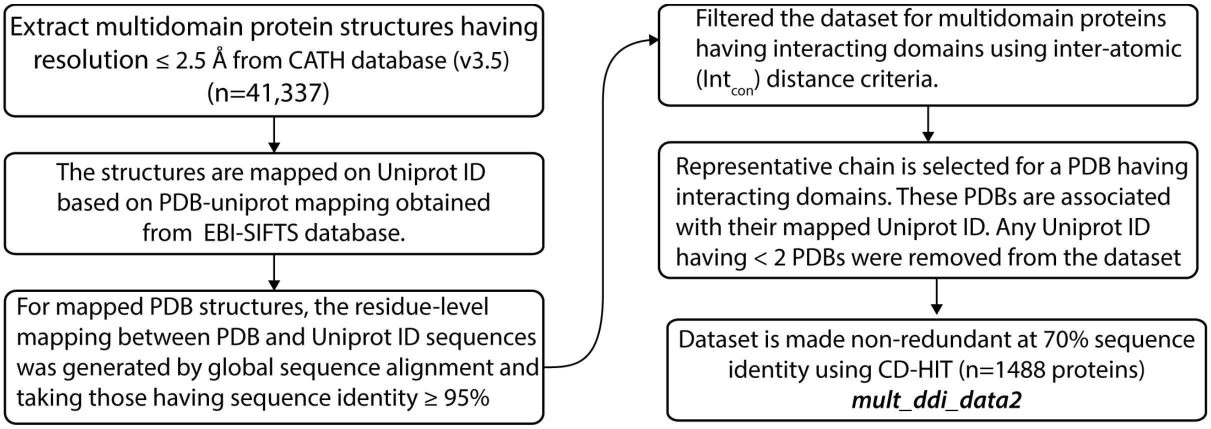

Extract multidomain protein structures having resolution ≤ 2.5 Å from CATH database (v3.5) (n=41,337)
The structures are mapped on Uniprot ID based on PDB-uniprot mapping obtained from EBI-SIFTS database.
For mapped PDB structures, the residue-level mapping between PDB and Uniprot ID sequences was generated by global sequence alignment and taking those having sequence identity ≥ 95%
Filtered the dataset for multidomain proteins having interacting domains using inter-atomic (Int$_{con}$) distance criteria.
Representative chain is selected for a PDB having interacting domains. These PDBs are associated with their mapped Uniprot ID. Any Uniprot ID having < 2 PDBs were removed from the dataset
Dataset is made non-redundant at 70% sequence identity using CD-HIT (n=1488 proteins)
mult_ddi_data2

Figure 2.10 Overview of the methodology used for generating C2-mult-ddi-data2. Figure showing important steps in construction of the C2-mult-ddi-data2dataset.

2.3.2.1 Identification of ligand bound to proteins

Since our dataset consists of both ligand unbound (apo) and ligand bound (holo) structures, we segregated structures bound to ligand/chemical compounds.Such identified ligand bound structures were used to probe whether there is a ligand induced conformational changes in domains. First, we categorized ligands into three subsets: DNA/RNA; small ligands, those with number of heavy atoms < 6, and rest all are put as large ligands. Then, we used NPIDB (Kirsanov et al., 2013; Zanegina et al., 2016) and LPC program (Sobolev et al., 1999) to identify which proteins are interacting with DNA/RNA molecules and ligands respectively. The interaction of ligand with domain is considered if there is at least on residue of domain is interacting with ligand. Thus, we identified following domains datasets: dom-DNA-RNA (bound to DNA/RNA); dom-apo (ligand unbound); dom-small-ligand and dom-large ligand are ligand bound structures with small and large ligand respectively.

The apo dataset was further divided into wild type (dom-apo-WT) and structures having mutation in domains (dom-apo-MT) based on the information in PDB file of the keyword “MUTATION”.

2.3.2.2 Measures of structural variation

In order to assess the structural variability at the domain-domain interfaces in multidomain proteins, we relied on following measures of assessment:

1. Interface Root Mean Square Deviation (RMSD): We calculated Cα RMSD of interfacial residues and used as a metric for structural variation of domain interfaces. The individual domain structures were optimally superposed using Kabash algorithm (Kabsch, 1976) as implemented in TM-score program (Zhang &Skolnick, 2004). Subsequent to this, superposed Cα interfacial residues were extracted for RMSD calculations. As had been mentioned before, we have used interactomic contact criteria to extract interfacial residues. Since the interfacial residues between two domains in multiple structures of a protein could vary from same set of interacting residues to no overlapping residues, it poses a problem in comparing domain-domain interfaces across multiple multidomain protein structures using RMSD. Note such cases occur because domain interactions are observed in other structures of the same sequence. To address this, we have defined two sets of interfacial residues: a) The union of all interface residues identified for a given interacting domain present across structures of the given sequence is referred to as Int-union; and b) similarly, the set of common (intersection) interface residues is categorized as Int-intersection. The RMSD for Int-union and Int-intersection residues are referred as RMSD-union and RMSD-intersection respectively. This is shown schematically in Figure 2.11.

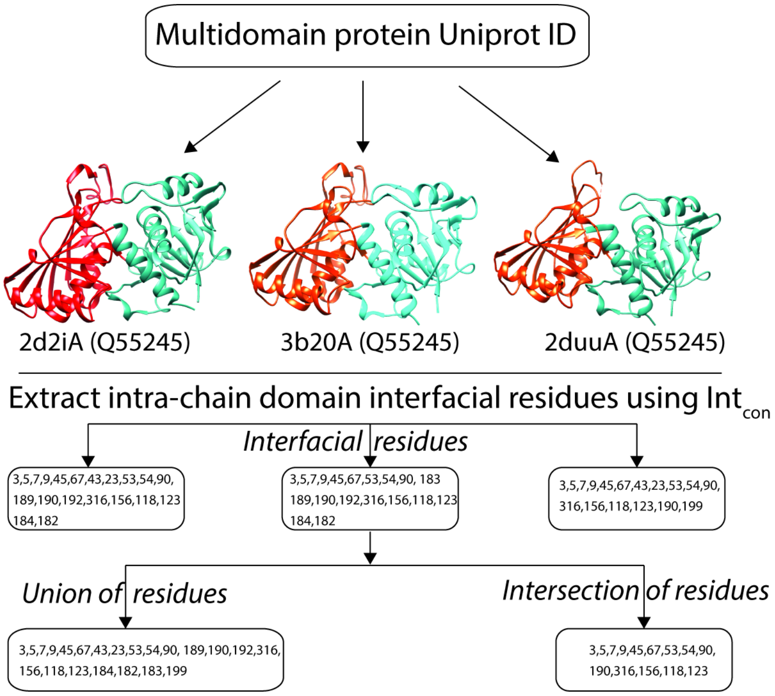


Figure 2.11 Schematic shows definition of Int-union and Int-intersection interfacial residues. Figure shows steps for generating union/intersection of interfacial residues.

2. Inter-domain geometry: We followed the method described by Han et al., (Han et al., 2006) to calculate inter-domain geometry or their relative orientation. In this procedure, first we find the best superposed domains as assessed by TM-score (Zhang &Skolnick, 2004) and then optimally superpose second domain. The translation and rotation required to superpose is calculated as metric to define domain geometry. The geometry between two domains is called as conserved if the translation less than 5Å and rotation less than 20°.
3. Interface similarity score (IS-score): The interface similarity score is output of structural alignment of interfaces by iAlign (Gao &Skolnick, 2010a). This interface alignment program was developed for aligning protein-protein interfaces. It essentially performs structural alignment of residues at the interfaces to detect their geometrical similarity. Since iAlign does not align individual proteins involved in PPI to detect similar interfaces, it can find structurally similar interfaces among all PPIs. We used iAlign version 1.0b7 for the structural alignment of domain-domain interfaces assuming each domain is equivalent to a protein in PPI. The similarity between interfaces is quantified using IS-score. Here, IS-score includes both geometric match score and conservation of the contact pattern between interfaces. IS-score is given by the equation:

IS-score = $(S + s_0) / (1 + s_0)$, where

$$S = \frac{1}{L_Q} \max \left[\sum_{i=1}^{Na} f_i / (1 + \frac{d_i^2}{d_0^2}) \right],$$

where, L_Q is the length of query interface; N_a is alignment length between query and template; d_i is distance (in Å) between Cα residue of aligned pairs; f_i is contact overlap defined by $f_i \equiv (c_i/a_i + c_i/b_i)/2$; where a_i and b_i are number of interfacial contacts of template and query interfaces at i^{th} position in the alignment respectively, and c_i is number of overlapping interfacial contacts at the same i^{th} position; d_0 is given by

$$d_0 \equiv \begin{cases} 1.24(L_Q - 15)^{1/3} - 1.8 \; For \; sequential \; alignment \\ 0.7(L_Q - 15)^{1/3} - 0.1 \; For \; non - sequential \; alignment \end{cases}$$

The length independent score S is obtained by normalizing it by s_0, which is given by $s_0 \equiv 0.18 - 0.35/{L_Q}^{0.3}$. The normalized S score is referred to as IS-score, which has the maximum score of one alignment between two identical structure (Gao &Skolnick, 2010a). Based on random protein-protein complexes, the p-value is calculated for IS-score. The statistically significantly similar interfaces are those with IS-score having p-value <0.05 and suggests that two aligned interfaces are similar and has some biological relevance.

2.3.3 Results and Discussion

In this study, we have systematically investigated the extent of intra-chain interface conformational variations in proteins having multiple experimentally determined structures. This will assist accurate modeling of intra-chain domain interfaces by recognizing proteins, which can undergo large interface variations. As has been mentioned before, first we generated all possible domain combinations for a protein in the dataset, and then filtered any domain interfaces having no interactions in multiple structures of a protein. Subsequent to this, we determined interfacial residues (Int-union and Int-intersection) and structural superposition of a domain-domain pair to study structural variations. This was measured using interfacial RMSD, relative domain geometry and IS-score.

We analyzed interface structural changes in non-redundant (at 70% sequence identity) dataset of 1489 multidomain proteins having more than one protein structure and these have at least one domain-domain interface. The superposition of domains resulted in mean (SD) RMSD of 1.3(2.8) and 0.7(1.9) for Int-union and Int-intersection interface residues respectively. The distribution of RMSD-union and RMSD-intersection is shown in Figures 2.12A and 2.12B respectively. As can be observed, most of domain pairs (~81%) have RMSD-union ≤ 1Å and ~90% of these have RMSD-union ≤ 5Å. Thus, suggesting that interface does not show much structural change as measured by RMSD. However, there are domain interfaces, which show large RMSD of 22 Å. We also examined the inter-domain geometry as measured by translation and rotation (see methods). Considering, 5Å and 20º change in angle as no change, ~95% of domain pairs conserved inter-domain geometry (Figure 2.13).

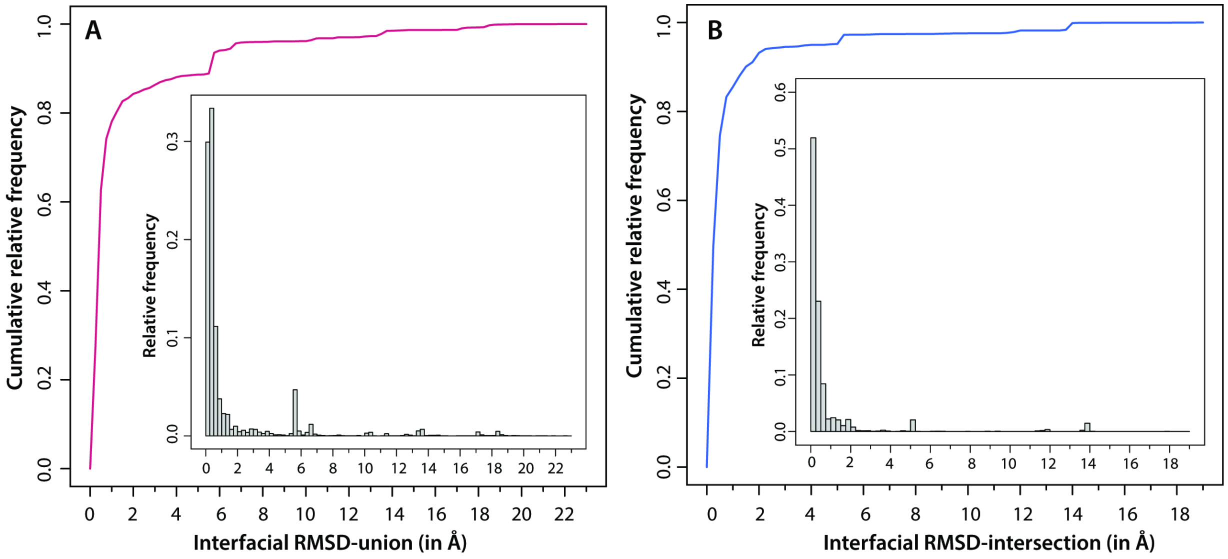


Figure 2.12 Cumulative distribution of interface RMSD. Figure showing cumulative distribution of: A) RMSD-union and B) RMSD-intersection of interfacial residues. Respective histogram of interfacial RMSD is shown in the inset.

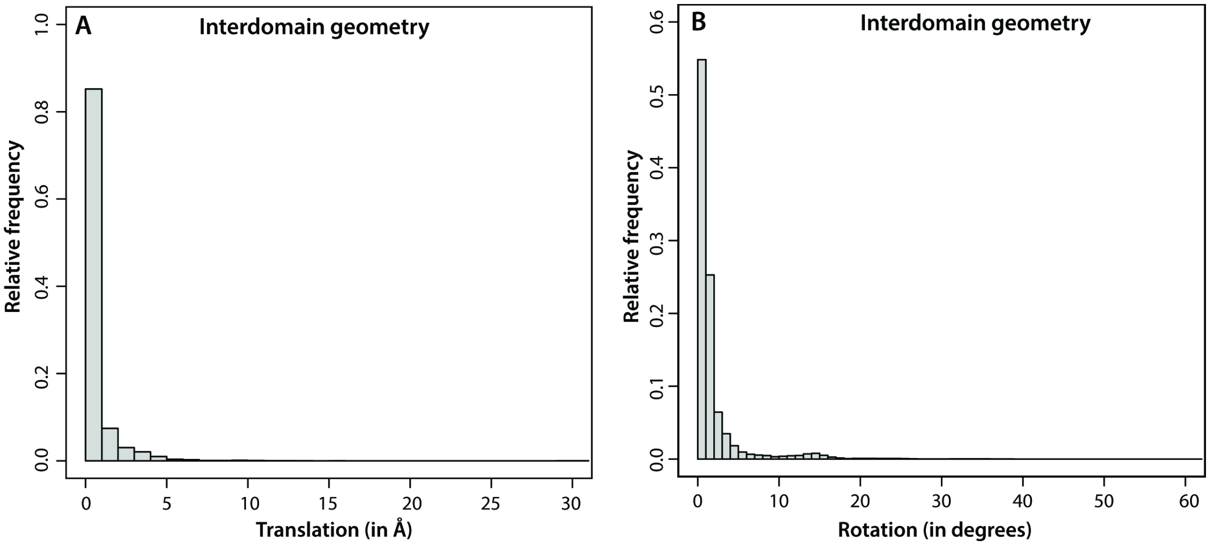


Figure 2.13 Inter-domain geometries as measured by translation/rotation. Figure shows histograms of translation and rotation metrics to assess inter-domain geometry of domain pairs. The extreme rotation/translation values are not shown in the figure.

We analyzed domain pairs having high interfacial RMSD that showed following reasons for large variation at interfaces: a. Comparison of interacting and non-interacting domain interfaces; and b. comparison of apo and holo (ligand/RNA/DNA) bound domains. Below we discuss some of the examples of domain pairs with drastic change in interfaces. The diphtheria toxin protein (UniProt id: P00588) has two experimental structures known *viz.* 1toxA and 1mdtA and it has three CATH domains (A01-catalytic domain(C); A02-translocation domain (T); and A03-receptor-binding domain(R)) corresponding to receptor binding, transmembrane and catalytic domains (Bell &Eisenberg, 1996). Of these the interface of two domains (A01 and A03) are found to have interfacial RMSD of 3.65Å between these domains from 1mdt and 1tox (Figure 2.14). Moreover, the rotational angle and translation is 179.59° and 30.84Å respectively. In its monomeric form (1mdtA), these domains (A01-A03) are interacting, however, in dimeric form (1toxA) the domain is swapped to interact with domain from other chain. Hence, the intra-chain domain interaction is absent. This involves significant structural transition. It has been shown that hinge region connecting T and R domain spans 8 residues (379-386) which show largest RMS deviation between closed dimeric and open monomeric structures (Bennett &Eisenberg, 1994).

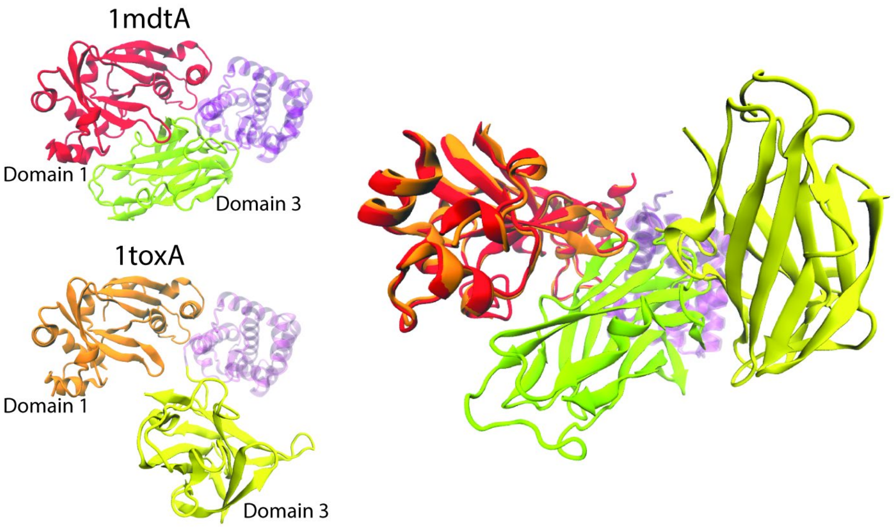


Figure 2.14 Example of distinct intra-chain domain interfaces in two functional states of protein. Figure showing motion of domains affecting domain-domain interface between two functional states of diphtheria toxin proteins as observed in experimentally determined structures (1mdtA and 1toxA). Both structures are shown on the left and their superposed domains are shown on the right. The interacting domains 1 and 3 of 1mdtA are shown solid new cartoon representation colored red and green colors respectively. Similarly, domains 1 and 3 of 1toxA are shown in orange and yellow colors respectively. The second domain is shown in transparent new cartoon representation in purple color.

Previous studies on protein structures have provided evidence that protein undergoes conformational changes could be induced by binding of the ligand, however, it is not necessary that ligand binding always leads to large conformation change (Brylinski &Skolnick, 2007; Cooper, 1976; Ha &Loh, 2012; Morange, 2006).

As has been mentioned before, in many cases large structural interface variations in identical proteins were found to be from DNA/RNA bound proteins either compared with each other or with other apo forms of the protein. One such protein is T7 RNA polymerase (UniProt id P00573) protein, which has 5 CATH domains. The comparison of domain interface of 1st and 5th domain between structures 1qlnA (A01-A05) and 1mswD (D01-D05) resulted in RMSD of 22.07Å (Figure 2.15). The initiation complex (1qlnA) structure upon interaction with DNA undergoes structural transition to form elongation complex (1mswD) (Cheetham &Steitz, 1999). It has been shown that N-terminal domain is unique to T7 RNA polymerase, which undergoes large conformational change in transition from initiation to elongation phase of transcription.

Moreover, secondary structural elements exhibit translations of 68.49Å and rotation of ~135° with some melting of helices and other changes in secondary structures (Yin &Steitz, 2002).

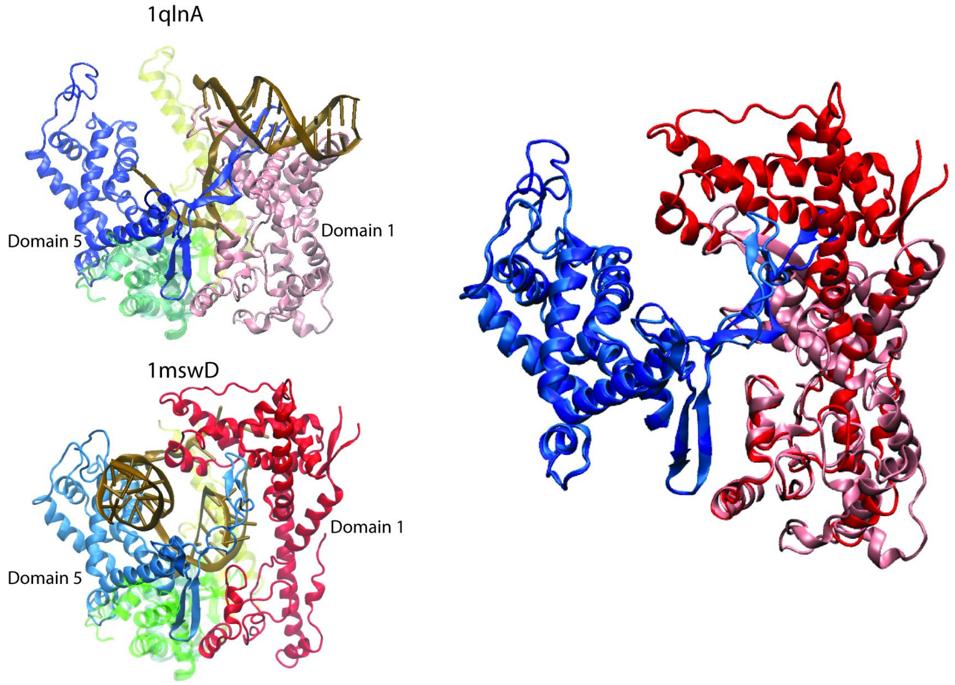


Figure 2.15 Example shows intra-chain domain motion upon DNA binding. Figure showing changes in domain-domain interface between two experimentally determined structures (1qlnA and 1mswD) of T7 RNA polymerase. Two individual structures are shown on the left and their superposed domains are shown on the right. The domains 1 and 5 of 1qlnA are shown solid in new cartoon representation and are in pink and blue colors respectively. Similarly, domains 1 and 5 of 1mswD are shown in red and sky blue colors respectively. Rest other domains are shown in different colors and are kept as transparent. The DNA is shown in ochre color.

The domain motions has been known in multidomain proteins due to binding of ligand that lead to changes involved in domain interfaces (Gerstein et al., 1994). One such example of multidomain protein bound to ligand with large change in RMSD is discussed below. We found interfacial RMSD of 21.9Å when domain interface formed by A01 and A02 domains from two structures (3fktA and 1sgzA) of the protein beta-secretase I (UniProt id P56817) were compared with each other. This is an aspartyl protease and a promising drug target for Alzheimer's disease (Barrow et al., 2008). The structure 3fktA is bound to a ligand (spiropiperidineiminohydantoin inhibitor) is in closed conformation of the protein, whereas the structure (1sgzA) is an apo form and is in the open conformation. The closing and opening of substrate binding site is controlled by a small set of residues and is called flap region (Hong &Tang, 2004). On binding the substrate, the flap moves from open to closed conformation by breaking several hydrogen bonds between flap residues (Hong et al., 2000, 2002). The destabilizing changes are compensated by the interactions of the substrate and protein structure (Hong &Tang, 2004) (Figure 2.16).

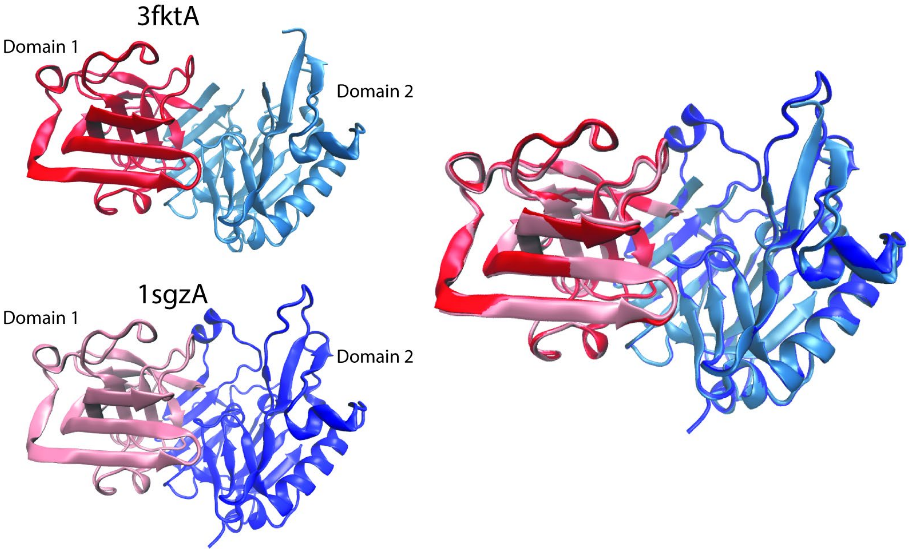


Figure 2.16: Example of change in intra-chain domain interface on ligand binding. Figure shows beta-secretase-I protein, which undergoes conformational change involving domain-domain interface on binding spiropiperidineiminohydantoin inhibitor. 3fktA is an apo form and 1sgzA is a holo form of the protein. Individual structures are shown on the left and their superposed domains are shown on the right. The solid new cartoon representation domains 1 and 2 of 3fktA are shown in red and sky blue colors respectively. Similarly, domains 1 and 2 of 1sgzA are shown in light red and dark blue colors respectively.

2.3.3.1 Contribution of conformational differences in each category of protein environment

To investigate the possibility of different experimental conditions contributing towards the conformational variations observed in same proteins, we categorically divided the main dataset based on the bound state of both the domain pairs in comparison. To define a domain pair belonging to a particular subset, both domains forming interface are either unbound (apo) or bound to ligands/DNA/RNA. We calculated interface RMSD for following datasets dom-apo, dom-DNA-RNA, dom-small-ligand, and dom-large-ligand (see methods). This will provide structural changes within bound form of structures i.e. ligand binding leads to similar ensemble of structures. The distribution of RMSD (Figure 2.17) and summary statistics is given in Table 2.2.

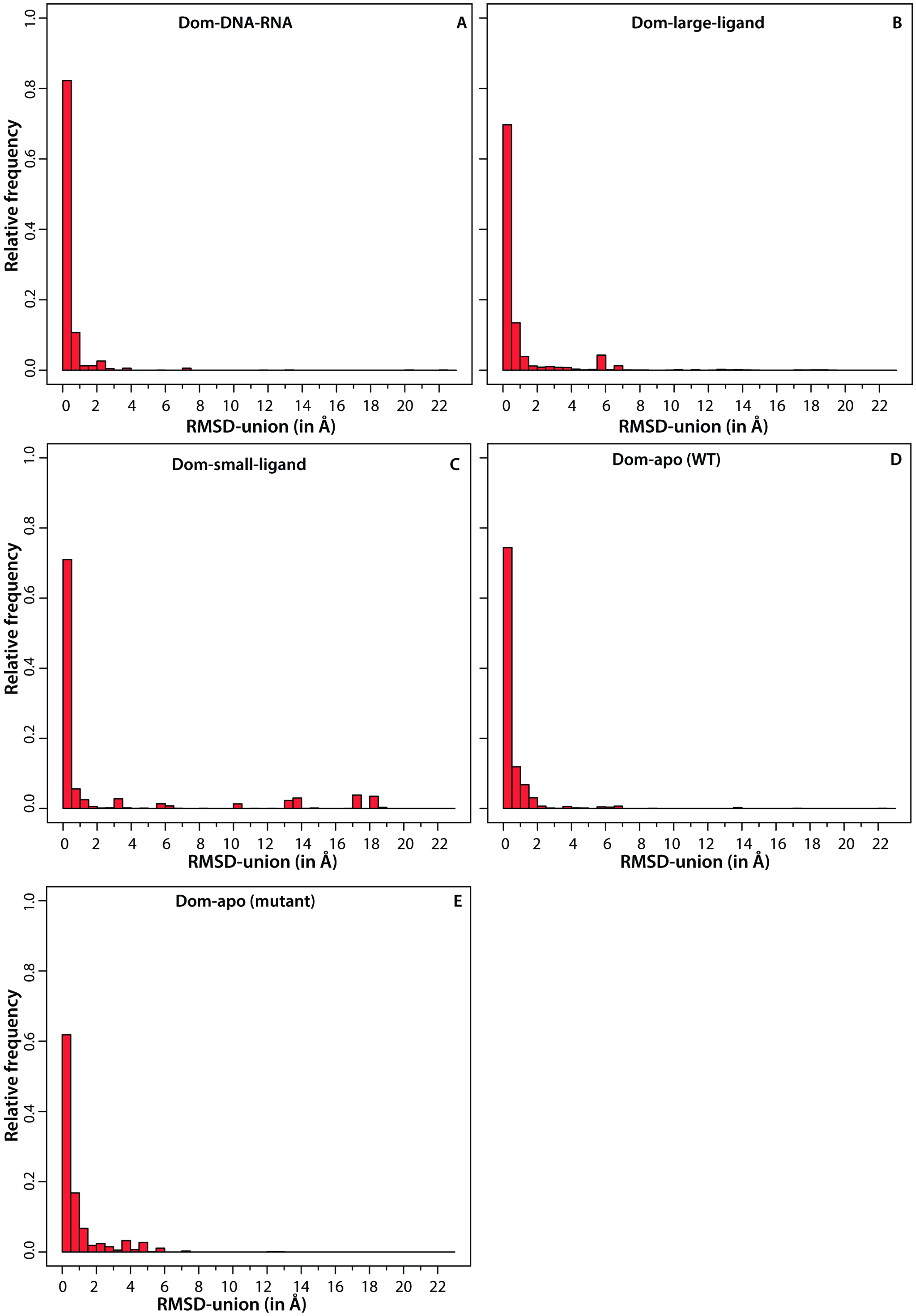


Figure 2.17 Distribution of RMSD-union of domain pairs. Histogram showing RMSD-union distribution for domain pairs A) bound to DNA/RNA, B) bound to large ligands, C) bound to small ligands, D) unbound in wild type form and E) unbound in mutant type.

Table 2.2: Union RMSD for bound and unbound domain pairs

Dataset	Number of domain pairs	RMSD Mean(SD)	RMSD Maximum
Dom-DNA-RNA	4450	0.47(1.11)	22.07
Dom-large-ligand	112377	1.03(2.24)	22.55
Dom-small-ligand	8048	2.72(5.45)	18.86
Dom-apo (wild type)	1477	0.62(1.36)	22.25
Dom-apo (mutant type)	744	0.89(1.37)	12.54

Finally, we aligned interfaces using iAlign, which structurally aligns interface and does not rely on sequence derived structural equivalences, which were used in previous RMSD analyses. Here only interfaces having at least 20 residues were aligned and the distribution of this is shown in Figure 2.18. Most of domain pairs (92%) have IS-score > 0.8. The domain interface pairs having low IS-scores are in cases where there is significant change in interface of one or both structures.

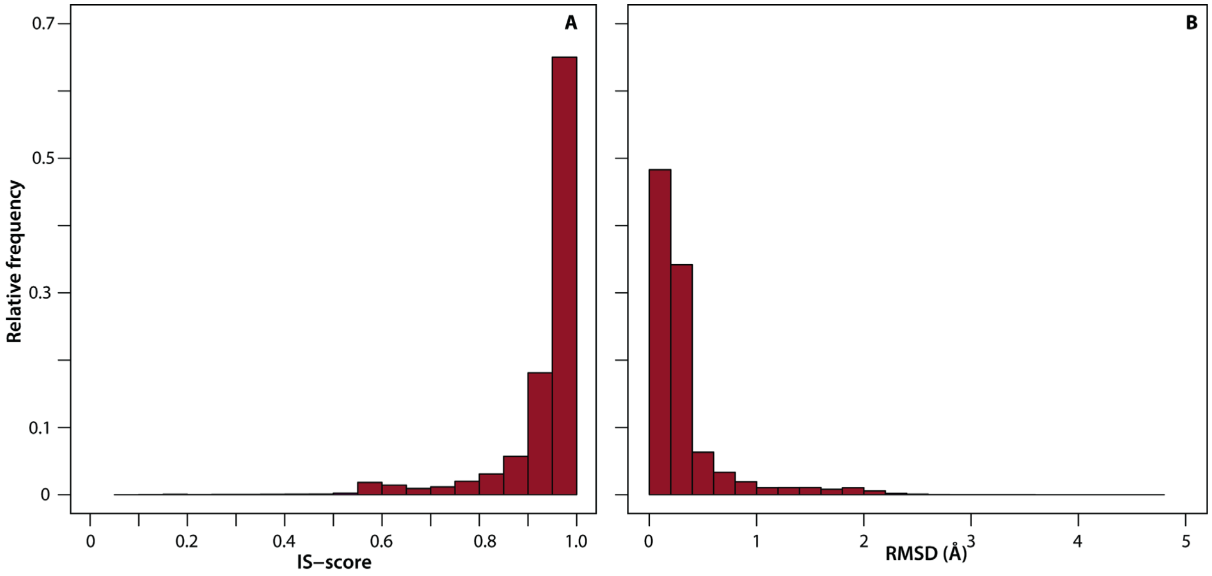


Figure 2.18 Distribution of scores obtained from structural alignment of interfaces. Histogram shows distributions of (A) IS-score and (B) interfacial RMSD.

2.3.4 Conclusions

In this chapter, we compared ways of defining intra-chain domain interfaces that showed interatomic contact criteria to define interfaces (at least 20 residues) results in definition similar to biologically relevant interfaces as known from protein-protein interactions. Using interatomic definition of domain-domain interfaces, we analyzed physiochemical and other properties between intra-chain domain interfaces and protein-protein interaction interfaces. This showed that domain interfaces are smaller in size with respect to PPI interfaces. Interestingly, both DDI and PPI interfaces show similarity in terms of hydrophobicity, hydrogen bonds, and secondary structures despite having differences in the size of interfaces. Thus, suggesting interfaces formed either between two chains or within a chain share similar physiochemical characteristics.

Further, we assessed structural variation of interfaces across multiple structures of a multidomain protein. We performed this study with an aim to understand the structural changes at interfaces, which can be considered in modeling of multidomain proteins. The comparison of intra-chain domain interfaces using RMSD showed that in general interfaces do not exhibit large variation as mean RMSD is ~1.3 Å with most (~90%) domain pairs having RMSD < 5Å. However, we have observed that some domain interfaces have large RMSD, which is mostly due to ligand binding or other factors based on their functions. Thus, suggesting intra-chain domain interfaces do not vary much under varying condition, however, a careful inspection is required when protein is known to interact with chemical compounds/DNA/RNA.

Chapter 3

Structural conservation of domain-domain interfaces (DDI) and their geometry in multidomain proteins having certain level of structural similarity

3.1 Introduction

The knowledge of three-dimensional (3-D) structure of protein can provide detailed insight into the molecular function of proteins. This propelled efforts for high throughput experimental determination of protein tertiary structures in large collaborations such as in structural genomics (Terwilliger et al., 2009). Despite this, there is ever increasing gap between known protein sequences and experimental protein tertiary structures. The protein structure prediction can potentially fill this gap by providing reliable protein models. In the past decade, significant progress has been made in the area of protein tertiary structure prediction, which has convincingly shown that reliable models of proteins can be exploited to decipher molecular details of protein function. It has been known that proteins having high sequence identity (>30%) usually share same fold, however, the reverse is not true as proteins belonging to same fold may show sequence identity as low as 8-10% (Balaji & Srinivasan, 2007; Illergård et al., 2009; Rost, 1997; Chothia & Lesk, 1986; Flores et al., 1993; Hubbard & Blundell, 1987; Russell & Barton, 1994). Thus, it has been known that 3-D structures of proteins in a fold are conserved well than their amino acid sequences. This has led to development of methods to reliably identify fold of a protein sequence (fold recognition) as a step towards modelling of protein 3D structure. Since *ab initio* methods of structure prediction is both time consuming and results in low prediction accuracy, the alternate method of template based modelling approaches have dominated the area of structure prediction (Fiser, 2004). The template based modelling (TBM) approaches involve identifying a reliable template/s of a given query sequence for its tertiary structure prediction. Hence, a successfully TBM method requires template to be present in PDB as well as ability to establish relationship between template to protein sequence. Recently, is has been shown that PDB is at least complete for single domain protein (Zhang et al., 2006). However, it is not known whether PDB is likely complete for multidomain proteins. Thus, modelling multidomain proteins using TBM approach could be used for prediction of individual domains. However, it remains as a challenge to assemble two or more domains. Thus, methods developed for modelling multidomain proteins, usually, model individual domains followed by assembly of domains. There are two main approaches known to predict structure of a multidomain protein from individual modeled domain structures: a) rigid body docking of individual domains, b) sampling the degrees of freedom of inter-domain linker (Wollacott et al., 2007). Utilizing these approaches, following programs have been developed to model multidomain proteins: MDA (Hertig et al., 2015), AIDA (Xu et al., 2014), ROSETTA (Rohl et al., 2004), DEMO (Zhou et al., 2019). These methods are able to assemble the domains, however, does not exclusively address the reliable modelling of domain interfaces, which are known to play role in function of multidomain proteins.

Several studies have been performed to understand the nature of interactions between domains in homologous multidomain proteins. One such study has examined the geometry of domain combinations for classical Rossmann superfamily combinations with 8 catalytic superfamilies. This showed that within a superfamily-superfamily pairs, relative orientation of domains and interfaces are conserved. However, the same is not conserved between two superfamilies (Bashton &Chothia, 2002). Furthermore, geometrical relationship between domains is not conserved when domains sequential order is reversed. Subsequent studies on the extent of conservation of domain-domain geometry and molecular structure of interface among homologous two-domain proteins have shown that ~60% of pairs conserve their geometry and interface and ~38% of pairs have variable geometries and interface. Interestingly, variable geometry and interface can be found even in homologous structures (Han et al., 2006). Another study has noted that usually the relative positioning of two superfamily-related domains in unrelated proteins are not similar (Apic et al., 2001a). These studies suggested that spatial domain orientations might be mostly affected by functional restraints. In a separate study on a small dataset of remotely related multidomain proteins, it was observed that relative positions of domains are conserved (Rekha et al., 2005). In later studies, it was shown using pseudo-torsion angle based on 2 domains center of mass and 2 Cα of domain boundary residues to measure mutual domain orientation that inter-domain geometry (IDG) is evolutionarily conserved (Bhaskara et al., 2013). Moreover, IDG seems to be affected primarily by change in interacting surfaces and/or inter-domain linkers (IDLs). IDG shows correlation with structural interface similarity IS-score (Gao &Skolnick, 2010a). These suggested that preservation of interaction constraints IDG. Further, it was shown that IDLs modulate domain interaction by varying its length, conformations and local structure.

Despite significant progress in understanding the extent of conservation of domain-domain interaction interfaces, these studies have not explored the possibility of structural relatedness between interfaces formed of homologous domains. In this study, we have comprehensively analyzed structural similarity of intra-chain domain-domain interfaces using a CATH (Orengo et al., 1997) hierarchy of structural similarity between corresponding domain pairs. Next, we explored whether there are evolutionary or functional constraints on conserving the structural interfaces among multidomain enzymes.

Previous studies on the conservation of domain geometry and interface structure considered only homologous two domain protein chains. This provides comparison of domain interfaces in proteins having two domains. However, it does not address the extent of conservation of intra-chain domain interfaces when homologous relationship is considered only at the level of domains involved in DDI of multidomain proteins having more than 2 domains. We investigated whether two interacting domains and their corresponding homologous domains form similar interface in multidomain proteins? This could be addressed by deriving structurally equivalent residues by aligning corresponding homologous domains of two DDIs. However, this may not be a suitable approach for this study because of difficulty in structurally aligning distantly related domains and in many instances equivalent surface of domain might not form part of interface region. Thus, we have resorted to use structural alignment of only interface regions between interacting homologous domain-domain pair. For this, we have relied on alignment of interfaces obtained from iAlign (Gao &Skolnick, 2010a).

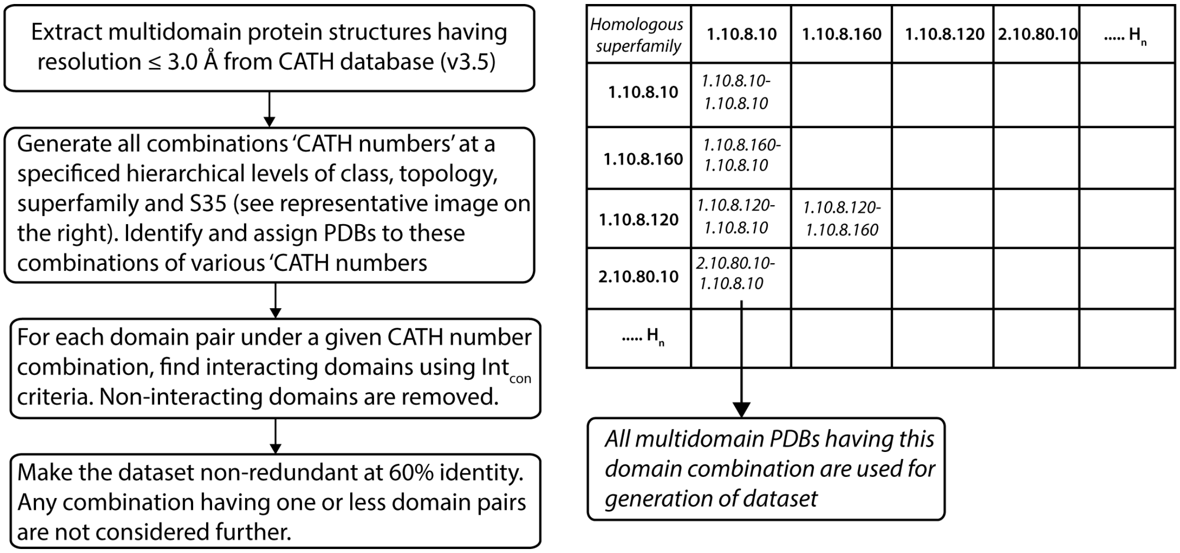
3.2 Materials and methods

We have investigated whether structurally related domains, where relatedness is defined at a given hierarchical level from structural database CATH, form structurally similar interfaces. This will provide the extent of structural conservation of interfaces between DDIs where domains are closely or distantly related. The interfaces were aligned using iAlign and assessed using interface similarity score (IS-score).

3.2.1 Construction of intra-chain domain-domain dataset at a given hierarchical level of CATH

In order to understand the relationship between structural divergence of domains and their interface conservation in multidomain proteins, we relied on structural domain database CATH, which classifies domains hierarchically based on their structural relatedness as established by secondary structures, and their topology and/or connectivity (Orengo et al., 1997). We constructed datasets having pairs of domain-domain interface, which are related at a specified CATH level. The schematic describing the construction of dataset is shown in Figure 3.1. Briefly, from the CATH database (v 3.5.0) we extracted all protein structures with resolution 3.0 Å or better with having at least two CATH classified domains. In CATH, it is possible to find a delineated domain without associated with classification. Subsequently, we extracted classification of domains in these multidomain proteins and prepared a list of 'CATH numbers' and associated domains at hierarchical classification levels of class, topology, homologous superfamily and 'S' (S35) which we considered as equivalent to family level relationship. At each of these classification levels, we generated all-against-all combinations of 'CATH numbers' that represents combination of domains in a protein (Figure 3.1). These were considered as order independent *i.e.* A-B and B-A is same. Further, for each pair of 'CATH number', we find all proteins consisting of these domains and filter for those proteins having the interacting domain pair. The interacting domains are defined as those having at least 20 interface residues using $\mathrm{Int}_{\mathrm{con}}$ criteria (as defined in Chapter 2). Next, from the set of interacting domain pairs we selected non-redundant proteins at 60% sequence identity. The domain combinations in previous step having one or no representative proteins are excluded from further analysis. For instance, given interacting domains D1 and D2 in a protein, we strive to find another non-redundant protein having interacting domains D1' and D2' such that both D1-D1'and D2-D2' are related only till the specified level of CATH *i.e.* if we are extracting superfamily level pairs, then D1-D1' both belongs to same superfamily but are from different family (S35). This ensures that both domains are diverged to the same extent. Following this procedure, it resulted in 1320, 44285, 49301, 1287353 domain-domain interface pairs at level of family, homologous superfamily, topology (fold) and class respectively. The resulting dataset *C3-mult-ddi-data1* is available as excel sheet in Chap03-Dataset.xlsx available at https://github.com/riviverma/thesis-md-proteins/.

The structural conservation of DDI was evaluated using Interface Similarity score (IS-score) obtained by aligning interfaces using iAlign. The domain orientation was evaluated using translation and rotation required for one domain to optimally align unto its corresponding domain partner in another protein (see Chapter 2 methods section 2.3.2.2).


Extract multidomain protein structures having resolution ≤ 3.0 Å from CATH database (v3.5)
Generate all combinations 'CATH numbers' at a specificed hierarchical levels of class, topology, superfamily and S35 (see representative image on the right). Identify and assign PDBs to these combinations of various 'CATH numbers
For each domain pair under a given CATH number combination, find interacting domains using Int$_{con}$ criteria. Non-interacting domains are removed.
Make the dataset non-redundant at 60% identity. Any combination having one or less domain pairs are not considered further.
All multidomain PDBs having this domain combination are used for generation of dataset

Homologous superfamily	1.10.8.10	1.10.8.160	1.10.8.120	2.10.80.10	..... H_n
1.10.8.10	*1.10.8.10-1.10.8.10*				
1.10.8.160	*1.10.8.160-1.10.8.10*				
1.10.8.120	*1.10.8.120-1.10.8.10*	*1.10.8.120-1.10.8.160*			
2.10.80.10	*2.10.80.10-1.10.8.10*				
..... H_n					

Figure 3.1 Overview of intra-chain domain interaction dataset construction at specified CATH classification level. Left panel shows flowchart of crucial steps in generating dataset. The right panel shows a representative of 'CATH number' combination (involved in one of the dataset construction steps) at the level of 'homologous superfamily'.

While constructing previous datasets, we imposed a strict condition on relatedness of domains in the aligned pair of DDIs *i.e.* domains should be related only till the specified level (CATH number). However, this eliminates many domain combinations from the dataset that may frequently occur in nature because only a limited repertoire of these combinations is found in genomes. Hence, we extended our dataset at the level of superfamily and S35/family to include such cases for at least one domain. For instance, one domain is related at superfamily level in a pair of DDI, but the other domain may belong to same family. Moreover, this analysis would be useful in including native interfaces, which could aid in preparing a compendium of representative template interfaces library. The resulting dataset *C3-mult-ddi-data2* is excel sheet in Chap03-Dataset.xlsx available at GitHub repository (https://github.com/riviverma/thesis-md-proteins).

While constructing the datasets, we considered only pairs of interacting intra-chain domains, where domains are related at specified structural relatedness. However, we did not consider linear arrangement of domains in two proteins such as whether domains are separated by one or more domains, and reversed order of domains. To understand whether interface similarity depends on the linear arrangement of domains, we subdivided our C3_mult_ddi_data1 based on linear separation of domains, and its linear order, into following different subcategories (see Figure 3.2). Various categories are:

a. Consecutive domains pairs: In a pair of DDIs, domains are contiguous in sequence; i.e. *a single linker region separates domains*.
b. Consecutive and reversed consecutive pairs: These consist pairs of DDIs, where one contiguous domain-domain interface is aligned with other domain-domain in reverse order.
c. Non-consecutive domain pairs: This has both DDIs having interacting domains separated by one or more domains.
d. Non-consecutive and reversed non-consecutive pairs: It has both DDIs formed of non-consecutive domains, however, in one of the DDIs the order of domain is reversed with respect to the other.
e. Consecutive/non-consecutive domain pairs: This is one DDI as consecutive and other DDI consists of non-consecutive domains.
f. Consecutive/non-consecutive and reversed consecutive/non-consecutive domain pairs: This is same as case (e) given above, however, one of the DDIs has reversed order of domains.

The list of pdb domain-domain interfaces pairs C3-mult-ddi-data3 are provided in the excel sheet in the Chap03-Dataset.xlsx available at (https://github.com/riviverma/thesis-md-proteins/).

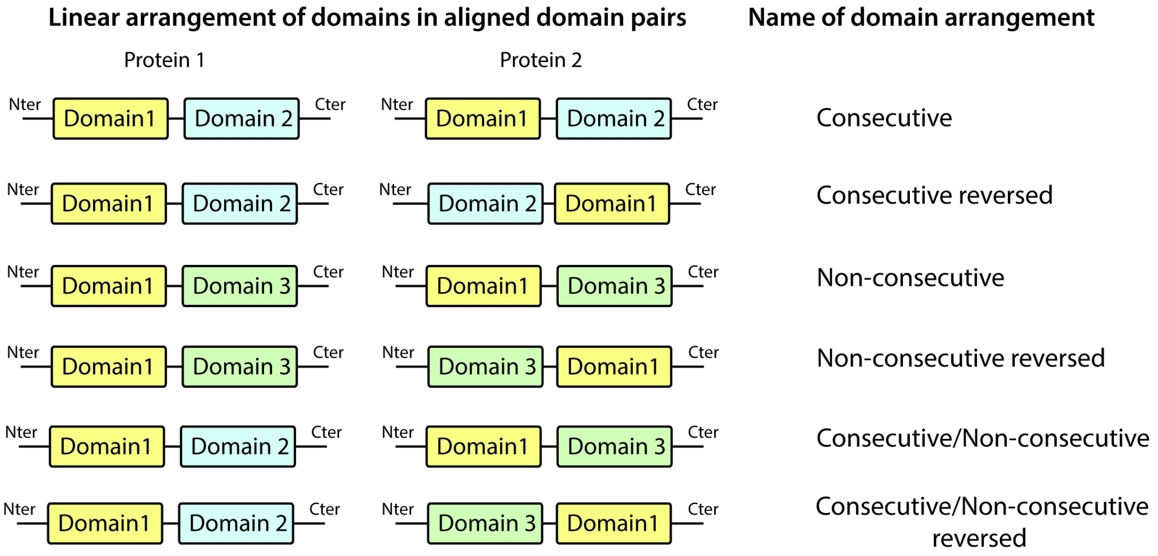

Linear arrangement of domains in aligned domain pairs
Name of domain arrangement
Protein 1
Protein 2
Nter
Domain1
Domain 2
Domain 3
Cter
Consecutive
Consecutive reversed
Non-consecutive
Non-consecutive reversed
Consecutive/Non-consecutive
Consecutive/Non-consecutive reversed

Figure 3.2 Categories of linear arrangement of domains. The figure shows various linear arrangements of domains along with classification category of domain order.

3.2.2 Dataset to study functional constraints on domain-domain interfaces

To explore whether there are functional constraints to conserve intra-chain domain interfaces of multidomain proteins during evolution, we have relied on well-characterized functional class of proteins *i.e.* enzymes. Moreover, enzymes are assigned Enzyme Commission (EC) numbers based on their reaction/substrate catalyzed (see section 1.3.1) that can be used for quantifying the functional similarities between two multidomain enzymes. To investigate this, we constructed datasets of multidomain enzymes extracted from CATH database (v3.5.0). Following the procedure described in methods (section 3.2.1.1), the dataset of pairs of DDIs related at superfamily (N=28409) and family (S35) (N=812) were generated. The dataset *C3-mult-enz-data* is given as excel sheet in the Chap03-Dataset.xlsx at https://github.com/riviverma/thesis-md-proteins/.

The quantitative assessment of functional similarity was based on EC number assignment to the multidomain enzymes and their associated Gene Ontology (GO) terms (Ashburner et al., 2000). We used mapping of multidomain PDB structure to EC numbers available at EC-PDB database (https://www.ebi.ac.uk/thornton-srv/databases/enzymes/). We extracted GO numbers associated with Uniprot IDs, which were obtained by its mapping to PDB structure from EBI-SIFTS database (Velankar et al., 2013). The functional similarity is assessed by:

1 EC number overlaps between two enzymes: It is ordered EC numbers common between two enzymes.

2 GO term overlap: This is computed as fraction of overlapping molecular function GO terms similarity between two enzymes given by equation:

$$Fraction\ of\ GO\ term\ overlap = \frac{number\ (N_{go1}\ \cap N_{go2})}{number\ (N_{go1}\ \cup N_{go2})}$$

where, N_{go1} and N_{go2} are GO molecular function terms in protein 1 and 2 respectively.

3.2.3 Analysis of intra-chain domain interfaces in multidomain proteins using SCOP database

Previous comparative studies between CATH and SCOP structural databases have emphasized that there is no significant difference in domain assignments between CATH and SCOP databases (Murzin et al., 1995), however in a non-redundant set, 23.6% of CATH interfaces had no SCOP equivalent and 37.3% of SCOP had no CATH equivalent (Jefferson et al., 2008). We have primarily using CATH database for domain definitions; however, we extended this study to domains derived from SCOP. This was performed with an objective to study whether domain definitions can affect the results of interface similarity between domain pairs related at specified classification levels. For this study, we extracted domain assignments from SCOPe (Fox et al., 2014) database (v2.05). Briefly, we took multidomain PDB structures from SCOP having resolution equal to or better than 3Å. From these, we extracted domain sequences and clustered them within each family at 60% sequence identity using CD-Hit program (Li & Godzik, 2006). Subsequently, each domain was assigned a cluster number according to the cluster it belongs. This provided additional classification level to SCOP to make non-redundant entries in final dataset. Following procedure outlined in method (section 1.2.1.1) we constructed pairs of DDIs at family, superfamily, fold and class levels of SCOP having 190504, 207996, 47640, 5422808 number of domain pairs respectively. We considered four major structural classes of SCOP: Allα, Allβ, α AND β (α+β), α OR β (α/β) in this study. The dataset 'C3-scop-mult-ddi' of is given as excel sheets in Chap03-Dataset.xlsx in GitHub repository at https://github.com/riviverma/thesis-md-proteins/.

3.3 Results and Discussion

3.3.1 Structural conservation of domain-domain interfaces and their geometries in multidomain proteins

In this work, we have systematically and comprehensively addressed the question: whether two domain-domain interfaces consisting of structurally related domains between them would form similar interfaces (Figure 3.3). To investigate this, we have systematically constructed a large dataset of intra-chain domain-domain interface pairs such that domains show a specified level of structural relatedness. This analysis will provide the extent of conservation of domain interfaces in the realm of structural space as well as provide an understanding of how structural relatedness between domains shapes the evolution of intra-chain domain interfaces.

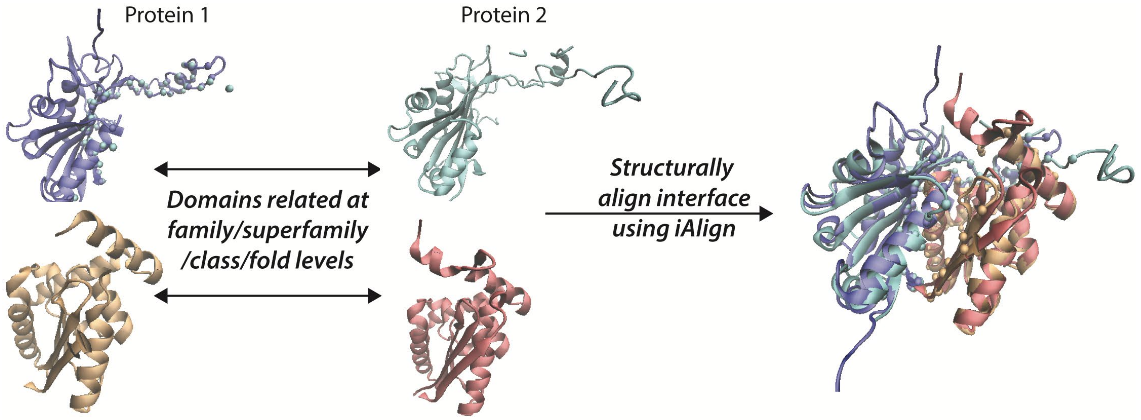

Protein 1
Protein 2
Domains related at family/superfamily /class/fold levels
Structurally align interface using iAlign

Figure 3.3 Schematics of domain interface comparison. Figure showing the comparison of interfaces between two intra-chain domains related at a specified structural relatedness.

3.3.1.1 Structural domains related at family or homologous superfamily in CATH

As described in methods, the dataset *C3-mult-ddi-data1* of pairs of domain-domain interfaces related at specified level were compared. First, we analyzed structural similarity of DDIs where domains in aligned pairs are family (‘S’ level of CATH) related using IS-score as obtained from iAlign. The distribution of IS-score for 1320 pairs of DDIs is shown in Figure 3.4A and summary statistics of the same is summarized in Table 3.1. The mean (SD) IS-score is 0.69 (0.15) for family related domain interfaces. The statistical significance of IS-score is quantified by p-value, which has been shown to be significant at value <0.05 (Gao &Skolnick, 2010b), which suggests statistically significant similarities. Based on this p-value cut-off, ~99% of pairs interface similarities is found to be statistically significant. The mean (SD) interfacial RMSD of these pairs is 1.03 (0.45) Å. These suggest family related domain-domain pairs usually have similar interfaces.

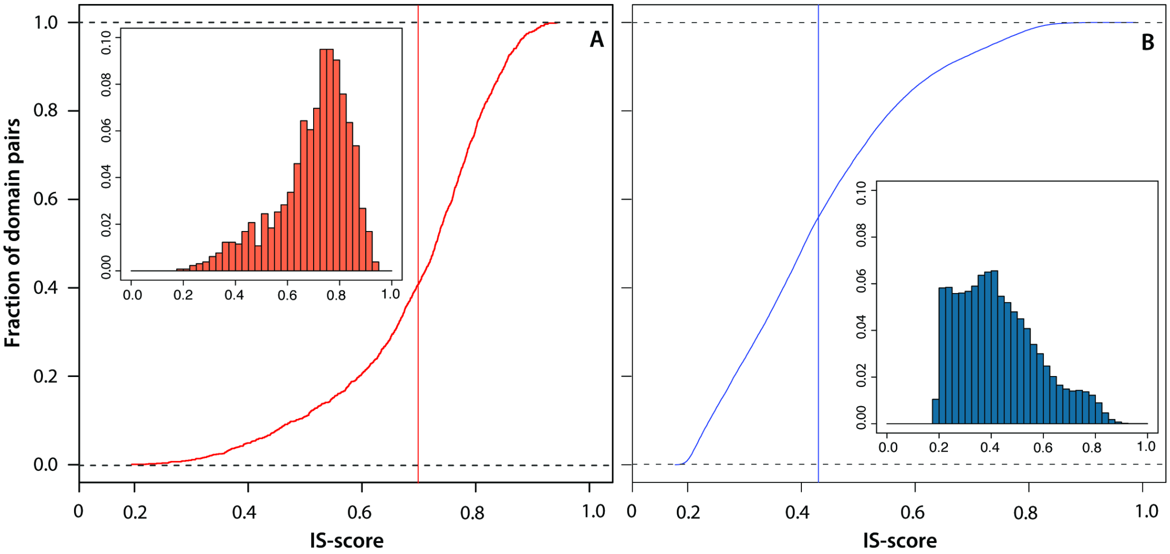


Figure 3.4 Distribution of IS-score of domain interface similarities. Figure showing cumulative distribution of IS-score of interfaces when two aligned domain interfaces have domains related at A) family (S35) or B) superfamily hierarchical levels. The inset of each panel has histogram of IS-score distribution.

Table 3.1: Summary of comparison of interfaces of domains pairs related at a specified hierarchical level in CATH.

CATH level	Total number of domain pairs	Average (SD) significant IS-score, % significant cases
Family	1320	0.700 (0.14), 99
Homology	44285	0.425 (0.15), 73
Topology	49301	0.217 (0.02), 3.4
Class	1287353	0.214 (0.02), 1.7

We analyzed pairs of DDIs with low IS-score and found that some of these have either utilized different interface region than topologically equivalent region or one of the domains has relatively rotated with respect to other. Below, we discuss some of these examples. The interface alignment of human apolactoferrin (PDB id: 1cb6) first and second domains to corresponding interface formed by homologous domains of porcine serum transferrin (PDB id: 1h76) results in IS-score of 0.26 (Figure 3.5A). Importantly, the pig transferrin is bound to two Fe^{3+}, of these the site of Fe binding lies in the domain1 and domain 2 interfaces (Figure 3.5A). The unbound form (human lactoferrin protein) does not have residues spatially oriented to facilitate binding of Fe ion, suggesting the domain motion is required for binding iron. Thus, we observed that change in binding interface accompanied by ~62° change in relative rotation of domains.

The interface alignment of DDI formed by first two N-terminal domains of *Escherichia coli* Elongation factor (EF-Ts) (PDB id: 1efuB) (Kawashima et al., 1996), and corresponding domains of *Thermus thermophilus* EF-Ts (PDB id: 1aipC) viz. domain 1 and 3 (Wang et al., 1997) results in IS-score of 0.32 (Figure 3.5B). The length of EF-Ts in *E. coli* is longer, where additional C-terminal sequence forms a subunit mimicking the N-terminal region (Wang et al., 1997). This leads to significant change in oligomerization state, which is hetero-tetrameric complex structure with EF-Tu protein in *T. thermophilus*, whereas same in *E. coli* is a simple hetero-dimeric structure (Wang et al., 1997). It seems that domain-domain interface of N-terminal domains in EF-Ts of *E. coli* have adjusted in order to accommodate these large scale changes in oligomerization because these domains form the part of inter-protomer interface.

In some cases, interface similarity is not significant despite domains belonging to same family. For instance, a non-significant match is found for domain-domain interfaces formed by Ras-domain (domain 1) and EF-Tu domain 3 in proteins: elongation factor (eEF1A) of yeast (PDB id: 1f60A) and corresponding homologue from archaea (*Sulfolobus solfataricus*) (PDB id: 1jnyA). The IS-score of 0.14 is found between domain interfaces (domains 1 and 3) from these two proteins. The analysis showed that there is significant change in interface region of both domains (Figure 3.5C) as revealed by superposition of full-length chains. This is because 1jny is bound to GDP, whereas 1f60 is bound to Guanine exchange factor (GEF), which facilitates exchange of GDP to GTP (Andersen et al., 2000; Vitagliano, 2001). Moreover this change is accompanied by ~ 74° relative rotation between these domains (Figure 3.5C). Thus, the interfaces are not found to be statistically significantly similar.

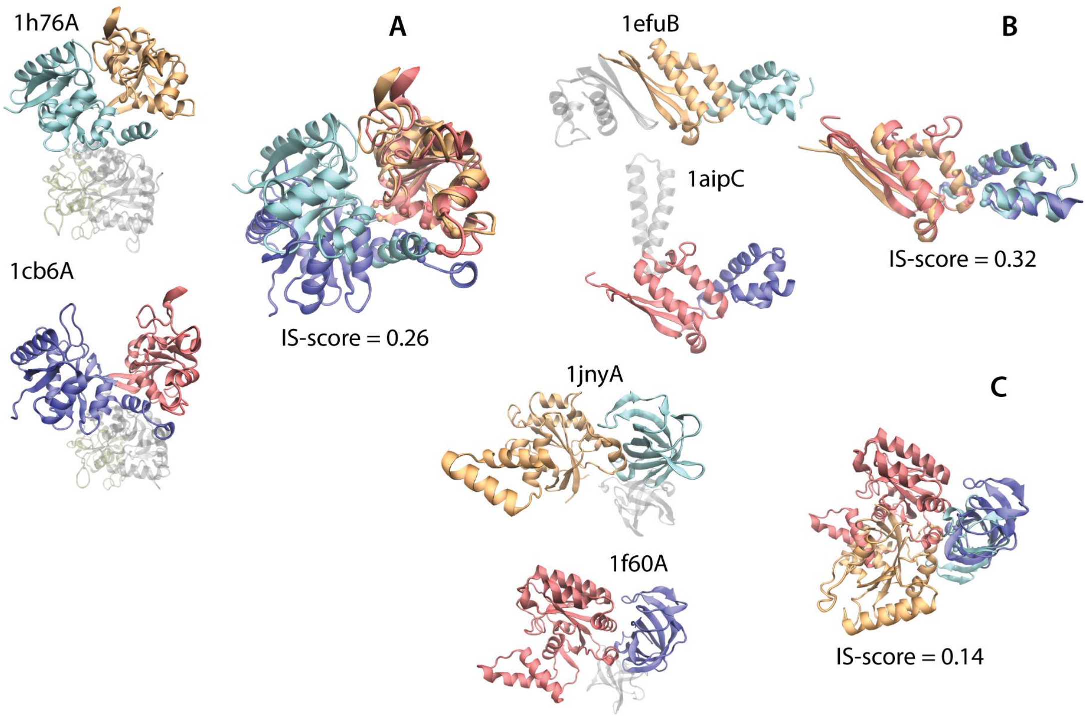


Figure 3.5 Examples of domain interface alignment of family related domains. Figure shows examples of low IS-score from family related domains in aligned domain pairs. In all panels (A, B and C), the arrangements of domains in individual proteins are shown on left and interface alignment only of aligned domain pairs are shown on the right. All structures are shown in new cartoon representation. A) 1h76A domains 1 and 2 are colored in cyan and orange respectively; domains 1 and 2 of 1cb6A are shown in blue and red colors respectively. Rest other domains are in transparent silver color. B) 1efuB domain 1 and 3 are colored in cyan and orange respectively; domains 1 and 3 of 1aipC are shown in blue and red colors respectively. Rest other domains are in transparent silver color. C) Domains 1 and 2 of 1jnyA are shown in cyan and orange respectively; Domains 1 and 2 of 1f60A are shown in blue and red colors respectively.

Next, we analyzed structural alignment of 44,285 pair of domain-domain interfaces having homologous superfamily relatedness. Using p-value of 0.05 as statistical significance of IS-score, as before, we found ~73% of alignments scores are significant. The distribution of these scores is shown in Figure 3.4 B and summary statistics tabulated in Table 3.1. As can be seen, IS-score has wide distribution ranging from IS-score value of 0.2 till 0.9. The mean (SD) IS-score and interfacial RMSD are 0.43(0.15) and 2.2(0.8) Å respectively. These indicate as domain-domain pair diverges the interfaces are not strictly conserved and variations among interfaces are observed. We analyzed low IS-score cases, and found most of these have slightly twisted relative domain interfaces. For example, domain interface in a two-domain protein putrescine transport system of *E. coli* (PDB id: 1a99B) has IS-score of 0.21 with interface of molybdate-binding proteins in *Azotobacter vinelandii* (PDB id: 1atgA). Both proteins belong to a diverse class of periplasmic receptors, which has variable ligand specificity. Interestingly, ligand binding site is at the cleft between domains and it lies at domain-domain interfaces in almost all periplasmic receptors (Lawson et al., 1998; Vassylyev et al., 1998). It has been known that ligand binding leads to conformational change involving hinge-bending. We found that 1a99B and 1atgA are bound to putrescine and sulfate anion (same binding site is involved in binding molybdate) respectively. Upon closer inspection of interface, we found that domain interface is oriented in a way to accommodate the appropriate sized ligand. From domain perspective, it is relatively rotated by ~24° in one of the protein (Figure 3.6A).

In another example, the structural alignment of domain interface of first and second domains of protein copper-containing amine oxidase from bovine serum (PDB id: 1tu5B) with interface of corresponding domains (Domain 3 and 2) of lysyl oxidase *Pichiapastoris* (Duff et al., 2006; Lunelli et al., 2005) (PDB id: 1w7cA) resulted in 0.29 IS-score with RMSD of 2.6 Å. Since the linear order of domain in these proteins is reversed in the way domain pair was chosen, there is rotation of 150° is observed in one domain with respect to the other. This could possibly the reason of low IS-score. However, the IS-score is statistically significant. Thus suggesting modeling of such domain interfaces cannot be derived from one multidomain template.

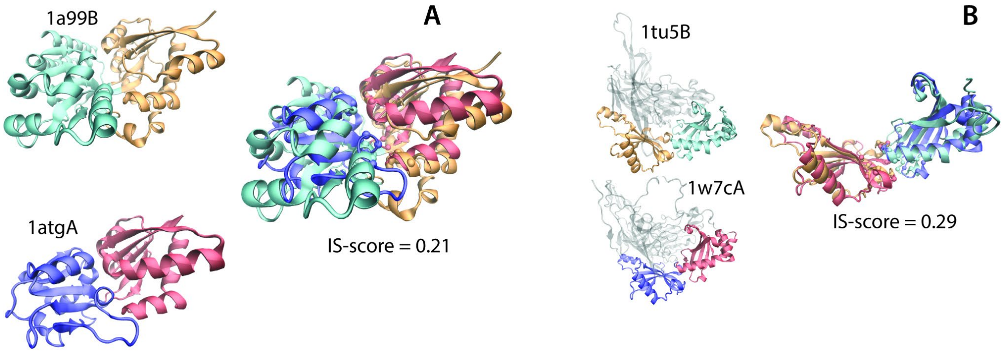


Figure 3.6 Example of domain interface alignment of superfamily related domains. Figure shows examples of low IS-score from superfamily related domains in alignments of domain pairs. In all panels (A and B), the arrangements of domains in individual proteins are shown on left and interface alignment only of aligned domain pairs are shown on the right. All structures are shown in new cartoon representation. A) 1a99B domain 1 and 2 are colored in cyan and orange respectively; domains 1 and 2 of 1atgA are shown in blue and red colors respectively. B) 1tu5B domains, *viz.*1 and 2 are shown in orange and cyan respectively; domains 2 and 3 of 1w7cA are shown in blue and red colors respectively. Rest other domains are in transparent silver color.

3.3.1.2 Structural domains related at fold or class in CATH

Having observed high conservation of domain interfaces in domains related at family level or to a lesser extent in homologous families (superfamily), we studied DDIs involving domains related at topology (Fold) or class levels. As expected, only 3.4% of statistically significant interface similarity could be observed when domains are related at fold level. Similarly, at class related domains, we observed ~2% of statistically significant DDIs. The distribution of both fold and class data is shown in Figure 3.7. This shows that detecting a template related to a domain in multidomain protein may not be sufficient to accurately model the domain-domain interface.

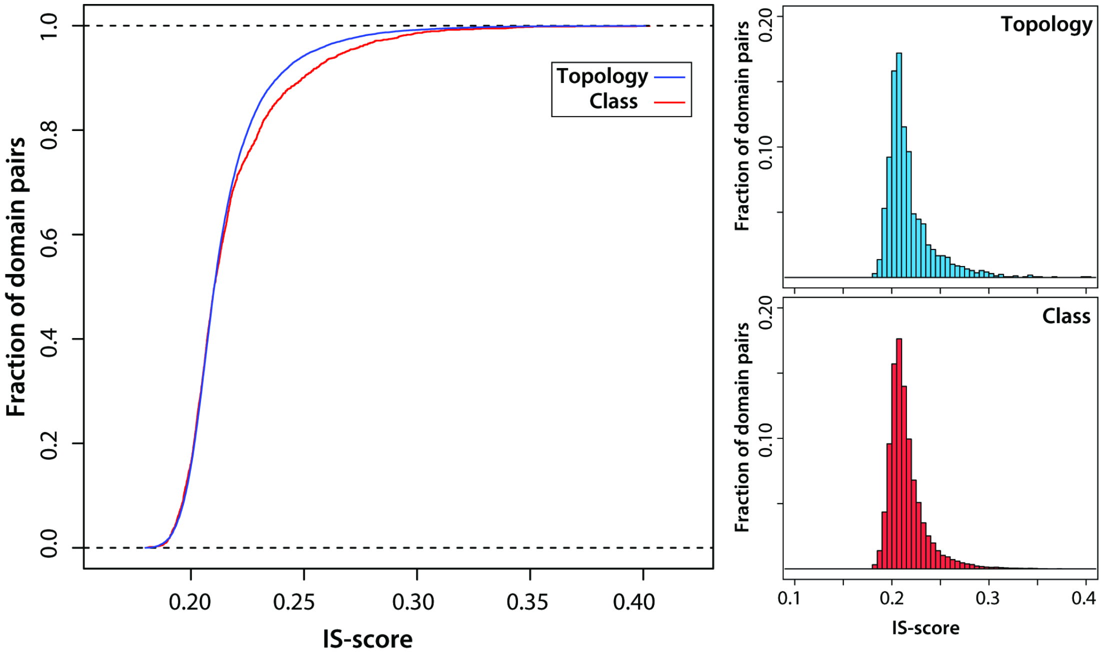


Figure 3.7 IS-score distribution of domains related at the levels of topology/class. Figure showing cumulative distribution of IS-score of interfaces when two aligned domain interfaces have domains related at topology (fold) or class hierarchical levels. The panel on the right shows histograms of the IS-score distribution.

One such example is the comparison of first and second domains in 5-keto-4-deoxyuronate isomerase (KduI) protein from *E. coli* (PDB id: 1xruA) and corresponding domains in proteins of Cupin superfamily (PDB id: 2vqaB). The IS-score between interfaces was 0.351 and RMSD was 2.72Å. Interestingly, KduI belongs to the class of Tim-barrels (5.3.1.17) and cupins have beta-barrels fold. Both class of families bind to metals, which aids in regulating the location of protein folding. It has been shown that although KduI is a member of different family than cupins but it is more structurally homologous to Cupins family than with the members of its own family because of the same structural fold in two different protein families (Figure 3.8). This suggest that interfaces could be conserved with no or low relatedness between domains.

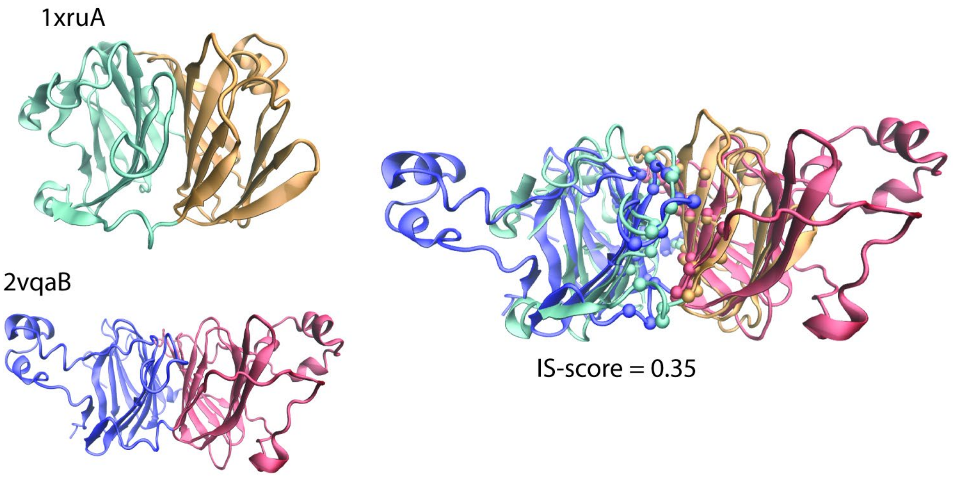


Figure 3.8 Example of domain interface alignment of topology related domains. Figure showing example of interface similarity between domains related at topology (fold) level. In the left panel, the arrangements of domains in individual proteins are shown and interface alignment only of aligned domain pairs are shown on the right. All structures are shown in new cartoon representation. 1xruA domain 1 and 2 are colored in orange and cyan respectively; domains 1 and 2 of 2vqaB are shown in blue and red colors respectively.

3.3.1.3 Analysis of inter-domain geometry of domains related at various CATH levels

Further, we performed analysis on the extent of divergence of intra-chain domain geometry as measured by rotation and translation required to superpose the second domain after optimal superposition of the best possible aligned domain. This analysis would be helpful in understanding whether domain pairs conserve their mutual orientation in order to conserve intra-chain domain interfaces. According to the definition used to analyze geometry (*see Chapter2 methods section 2.3.2.2),* a translation ≤ 5Å and rotation ≤ 20° between two domain pairs is considered to be conserved geometry (Han et al., 2006). Since previous analysis showed conservation of interfaces at the level of family or limited conservation at homology, we have restricted our analysis to family and homology related domain pairs. The distribution of translation and rotation is shown in Figures 3.9 A-D. In general, mean translation and rotation for family related domains are 1.8Å and 5.9° respectively. The same mean translation and rotation for homology related domains are 7.3Å and 26.7° respectively. Using the criteria of (Han et al., 2006), we observed that ~97% of pairs of DDIs having family related domains conserve their geometry. The same in homology related domains is ~73% of pairs of DDIs have conserved geometry.

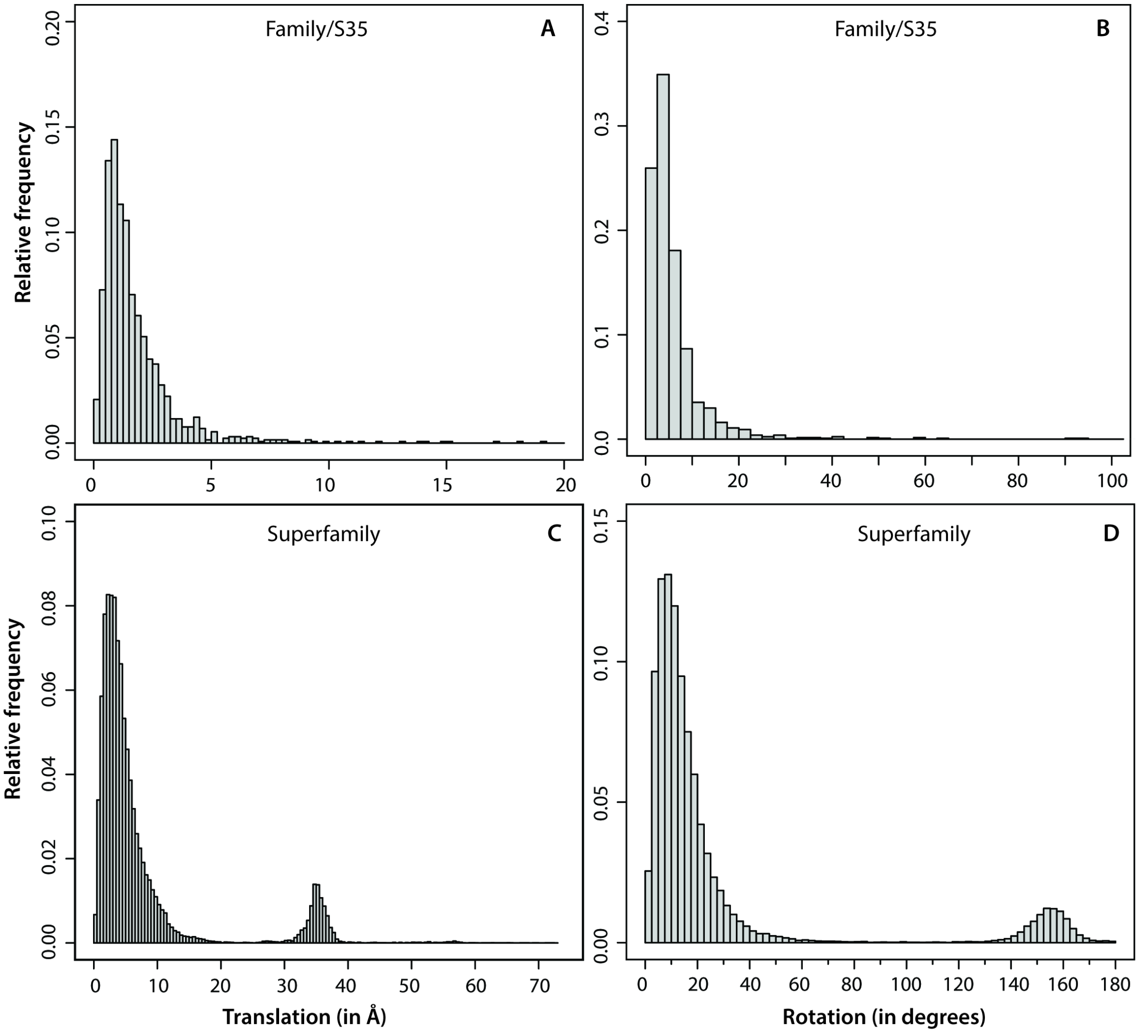


Figure 3.9 Comparison of inter-domain geometry among domain pairs related at family/homologous superfamily levels of CATH. Scatter plot showing the relationship between IS-score to the relative domain orientations as measured by translation and rotation for family (A and C) and superfamily related domains (B and D).

Next, we examined whether IS-score is correlated with translation and/or rotation. The plots for these are shown in Figures 3.10 A-D. We could not observe any apparent correlation between IS-score and rotation and/or translation. Interestingly, we observed that many pairs have large rotation in homology related pairs. This is observed because the domain arrangement being compared has reversed linear domain in one of the aligned domain pairs. One such example is β-trefoil lectin HA33/A (PDB: 1ybi) and β-trefoil lectin HA33/C (3aj6) protein from Clostridium botulinum toxin type A neurotoxin, which has translation of 14.3Å and rotation of 63.7° with IS-score of 0.11862 (Figure 3.11). H33/A binds to glycolipids and glycoproteins containing galactose and H33/C recognize sialic acid-containing glycolipids and glycoproteins. The dissimilarity in domain orientation is because these are two different serotypes. HA33/C has a longer N-terminus located at the interface of the domains that does not undergo the post-translational cleavage but a rotation of 60˚ is observed between C- terminal domains (Arndt et al., 2005; Nakamura et al., 2011).

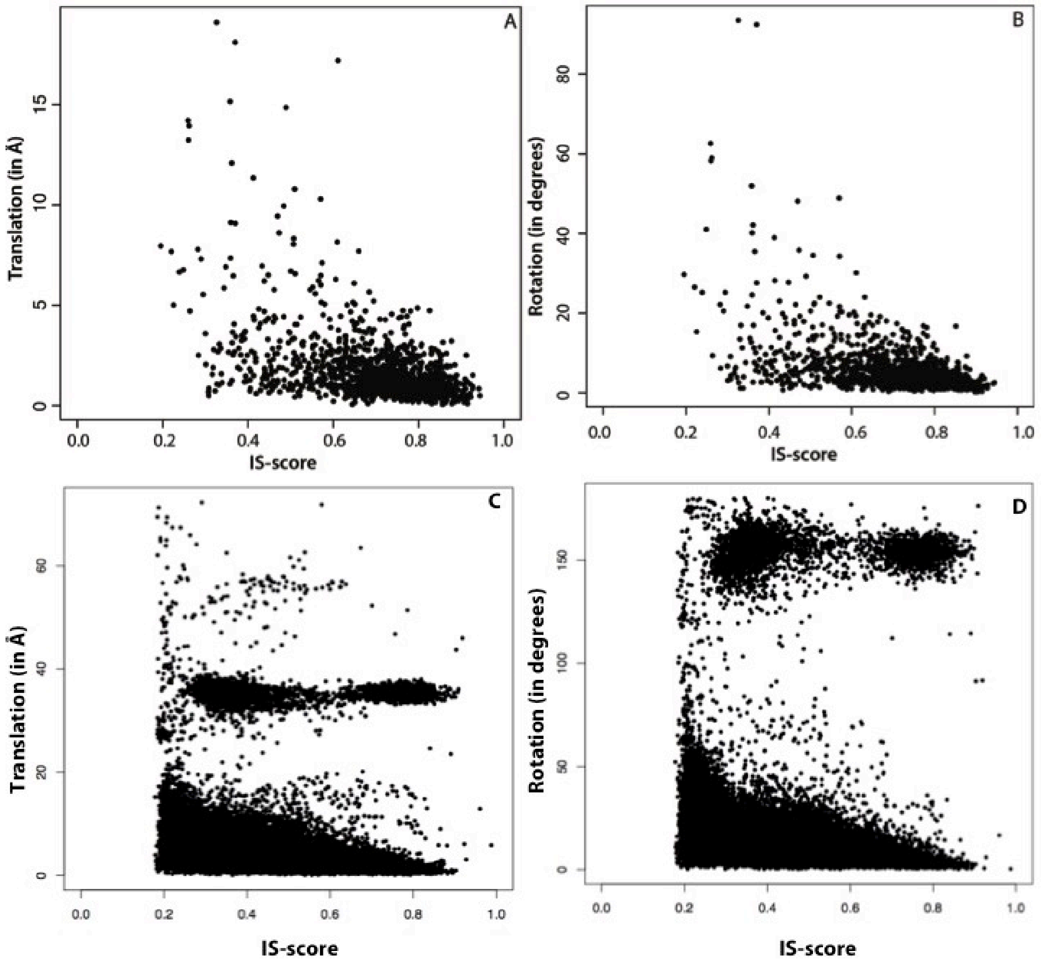


Figure 3.10: Scatter plots showing relation between IS-score and inter-domain geometry. Scatter plots show the relation between IS-score and translation and rotation for family /superfamily related domains. Plots A) and C) are for translation for family/superfamily related domains respectively. Plots B) and D) are for rotation for family/superfamily related domains respectively.

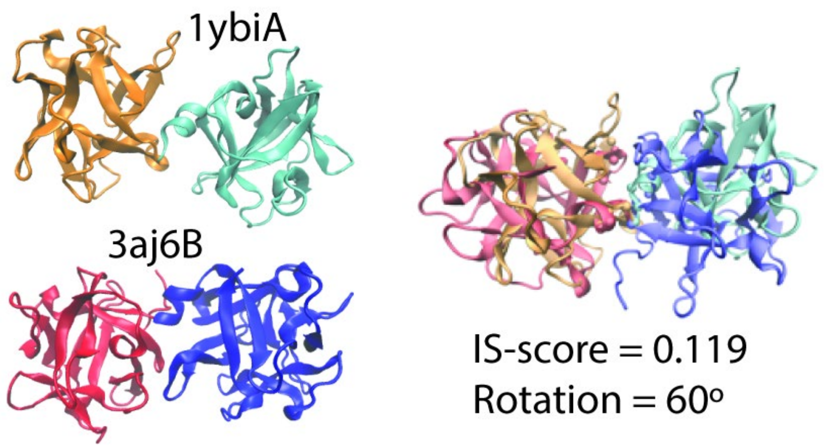


Figure 3.11 Example showing large rotation in domain interface of homology related domains. Figure shows an example of interface alignment of homology related domains pairs where one domain pair has relative large rotation. The structures are shown in new cartoon representation. Domains of 1ybiA (domain 1 and 2) are colored in cyan and orange respectively; similarly, domains 1 and 2 of 3aj6B are shown in red and blue colors respectively. The aligned domain pairs are shown on the right.

We also considered various possibilities in linear domain arrangement (Figure 3.12) and we did not find any pattern in these and IS-score of domain pairs either at homology or family level. The results are shown in Figure 3.12.

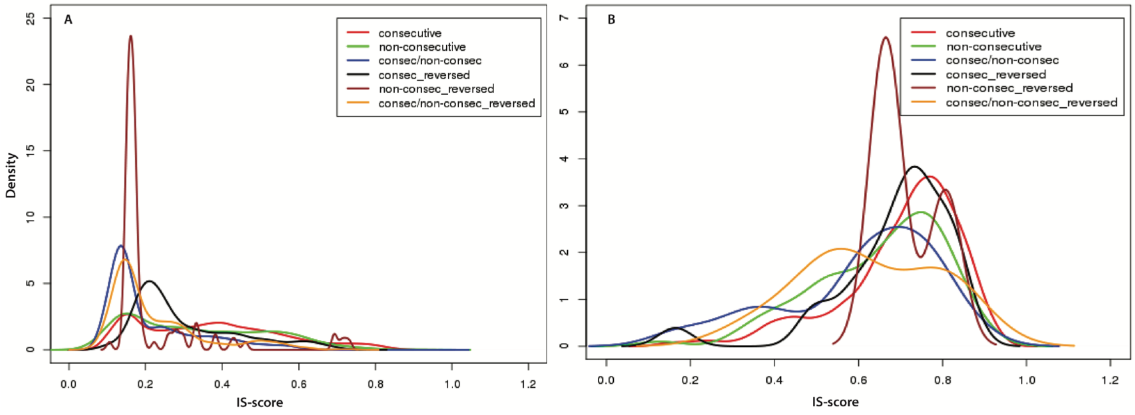


Figure 3.12 IS-score distribution of aligned domain pairs having varied domain arrangements. Figure showing IS-score distribution of different sub classifications of the linear arrangement of the domains at (A) homology and (B) family levels.

3.3.2 Structural conservation of domain-domain interfaces in multidomain proteins having certain level of structural similarity as documented in CATH database: an extension

Next, we extended our interface similarity comparison at family and superfamily levels of structural similarity by incorporating domain pairs having at least on domain having specified relationship at the specified level. This is to address the question whether interface similarity is better when pairs of DDIs share high similarity. In this part of the work, one of the domain pair has to be related at a particular level of CATH but the other domain pair can be similar at any lower (detailed) level. This analysis would be more beneficial in detecting templates, which are more realistic to exist in nature to accurately model target interfaces. As seen in Figure 3.13, the peak of statistically significant IS-scores of domain pairs related at homology level is now shifted towards the right with high IS-score (Mean (SD) 0.62(0.16)) values (Figure 3.13A). The mean of IS-score at family level remains same around 0.7(0.15). This observation clearly signifies that extent of structural interface conservation depends on structural similarity between those domain pairs. Interestingly, the inter-domain geometry of domain pairs related at homology level also observed to be conserved for most of the cases (Figure 3.13 B and C).

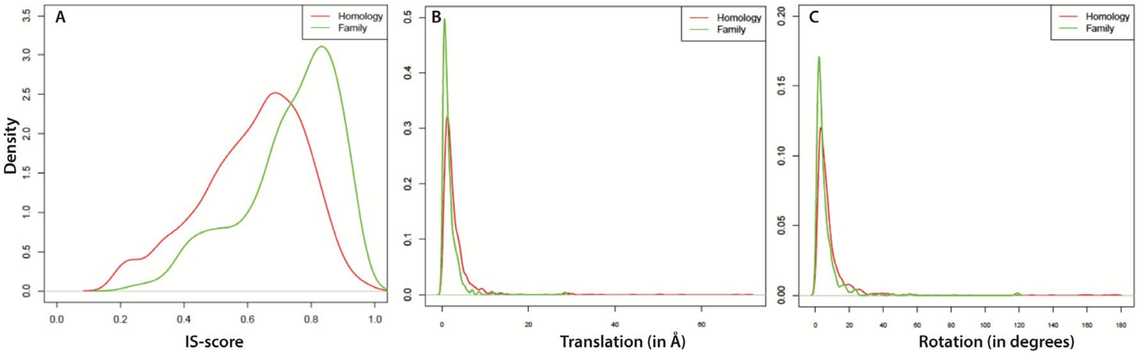


Figure 3.13 Density plots showing various measures of interface similarities of domain pairs related at family/homology levels. Figure shows density distribution of interface alignment metrics: (A) Statistically significant IS-score; and inter-domain geometry as measured by (B) translation and (C) rotation of family/homology related domains in domain pairs.

3.3.3 Role of functional constraints in maintaining the similar interface

It is known that domain interfaces could harbors the functional sites in multidomain proteins (Hirako & Shionyu, 2012). In order to explore whether in evolution the domain interfaces formed of structurally related domains are conserved to facilitate similar function, we analyzed interface conservation among family/superfamily related domain-domain pairs to their function conservation in enzymes. For this, we constructed a multidomain enzyme structures dataset *C3-mult-enz-data* (as discussed in methods section) using the mapping between PDB ids and EC (Enzyme Commission) numbers. The pairs of domains were selected at structural similarity level of homology and family in a similar manner as described in the first dataset (methods section 3.2.1.1). The interface comparison for these pairs was done using iAlign program. In order to understand whether interface conservation is functionally constrained, we calculated two parameters: *EC_overlap* which is simply counting the number of EC terms shared between two proteins and *fraction_GO_overlap*, which is defined as fraction of common GO terms (molecular function) between two proteins. The results in Figure 3.14 (A and B) showed the correlation between interface similarity score and fraction of GO terms overlap at homology (Pearson correlation coefficient=0.301) and family (Pearson correlation coefficient=0.065) level. It can be seen that interface conservation of domain pairs related at superfamily does not necessarily lead to function conservation as measured by GO overlap. The EC number overlap and corresponding IS-score is shown in Table 3.2. It can be seen that even if all the four EC number overlaps at homology level, the average IS-score is not more than 0.4, whereas at family level, domains do share conservation of interface and function. This signifies that interfaces between domain pairs related at homology level are not under functional constraints to maintain a similar interface. Additionally, we did observe that at homology level, interface conservation span a wide range of IS-scores. They show high similarity between their interfaces along with other proteins, which share no or low similarity of interfaces. The proteins showing low similarity of interfaces might serve an explanation for proteins, which do not conserve their function. Moreover, it has been demonstrated in some studies that protein families within superfamilies show diversity in their structures as well as functions (Das et al., 2015; Ga et al., 2006).

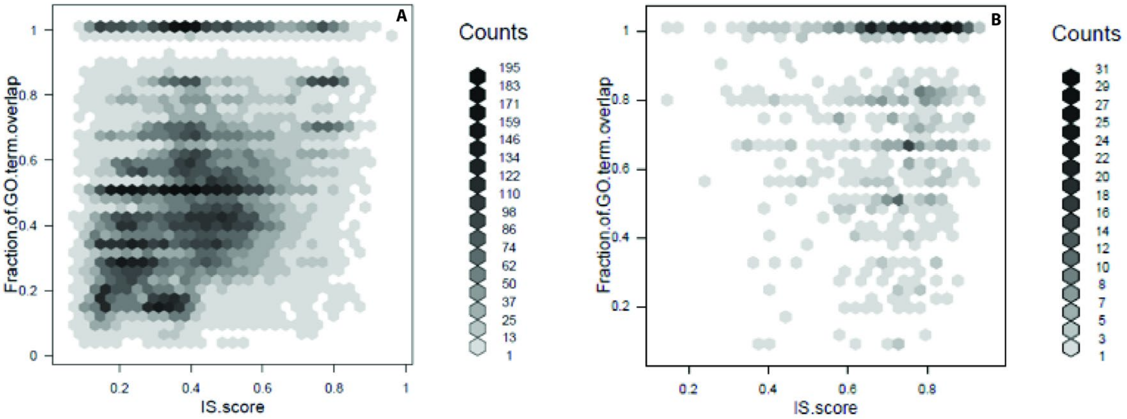


Figure 3.14 Hexagonal binning plots showing relationship between GO term overlap and IS-score. Panel (A) is plot of domains pairs related at superfamily and (B) is for family related domain pairs.

From Table 3.2, we observed that at family level, there are eight proteins with no EC number overlap. One such example is *trans*-sialidases (PDB id: 2jkbA), which bind to two different substrates and therefore their mode of carrying out the chemical reaction is different (Figure 3.15A). The *trans*-sialidase binds to α2-3-linked substrates (EC number: 3.2.1.18) and intra-molecular *trans*-sialidase (PDB id: 3sliA) binds to α2-3 linked sialic acids and produce 2,7-anhydro-Neu5Ac (EC number: 4.2.2.15). The other example includes the Keto-pantoatereductase (KPR) of *E.coli* (PDB id: 1ks9A) and ovine 6-phosphogluconate dehydrogenase (6PGDH) of *Ovisaries* (PDB id: 2pgdA). Both the proteins undergo large conformational change on binding the redox cofactor producing a closed conformation of the active site (Tchigvintsev et al., 2012). Additionally, KPR exists as a monomer and has an asymmetric unit, whereas 6PGDH exists as homo-2-mer and has a cyclic symmetry (Figure 3.15B).

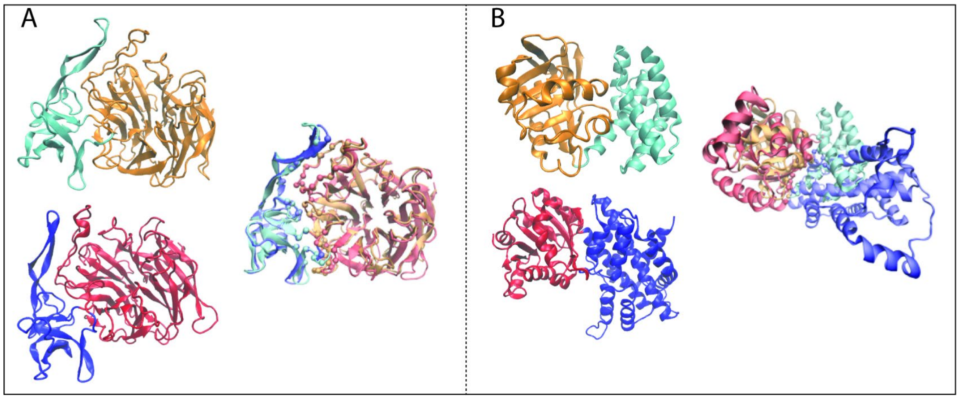


Figure 3.15 Examples of domain interface alignment of family related domains with no conserved function. Figure showing examples of no EC number overlap from family related domains in an alignment of domain pairs. Both structures are shown in new cartoon

representation. A) 2jkbA domain 3 and 2 are colored in cyan and orange respectively; domains 3 and 2 of 3sliA are shown in red and blue respectively. B) 1ks9A domains, *viz.*1 and 2 are shown in orange and cyan respectively; domains 1 and 2 of 2pgdA are shown in blue and red colors respectively. The alignments of interfaces are also shown on the right in each panel.

Table 3.2: Summary of EC number overlap and their corresponding average IS-score at homology and family level.

Number of pairs	EC number overlap	Mean IS-score	
1677	0	0.271	**Homology**
1607	1	0.237	
3065	2	0.450	
5284	3	0.452	
2284	4	0.441	
Number of pairs	**EC number overlap**	**Mean IS-score**	**Family**
8	0	0.66	
19	1	0.76	
11	2	0.67	
68	3	0.74	
485	4	0.72	

3.3.4 Analysis of structural conservation of intra domain interfaces in multidomain proteins having certain level of structural similarity as documented in SCOP database

Both SCOP (Murzin et al., 1995) and CATH (Orengo et al., 1997) are two standard structural domain databases. The domain definitions in both databases largely overlap, however, in classification there are difference between CATH and SCOP (Csaba et al., 2009). Therefore, to explore whether observations of CATH dataset can be extended to SCOP domain definitions, we carried out similar analysis of domain-domain interface similarity of SCOP domains at various hierarchical levels. To analyze the similarities in interfaces at different levels of structural similarity, we generated a multidomain dataset from SCOPe database in a similar manner as that for CATH database. The domain interfaces from different levels were compared using iAlign program and IS-scores were calculated to quantify the structural conservation of the interfaces.

The distribution of IS-scores at different levels is plotted in Figure 3.16 and a detailed statistics is summarized in Table 3.3. The class and fold showed very restrictive distribution similar to as observed in CATH database. The domains at these levels of structural similarity do not share any evolutionary relationship among each other and therefore, low conservation of interfaces is expected. However, we found an interesting observation that interfaces of domain pairs related at superfamily or family level of SCOP do not share strong similarities as seen in CATH. The mean of significant IS-score (p-value < 0.05) at superfamily and family level was 0.32 and 0.47 respectively.

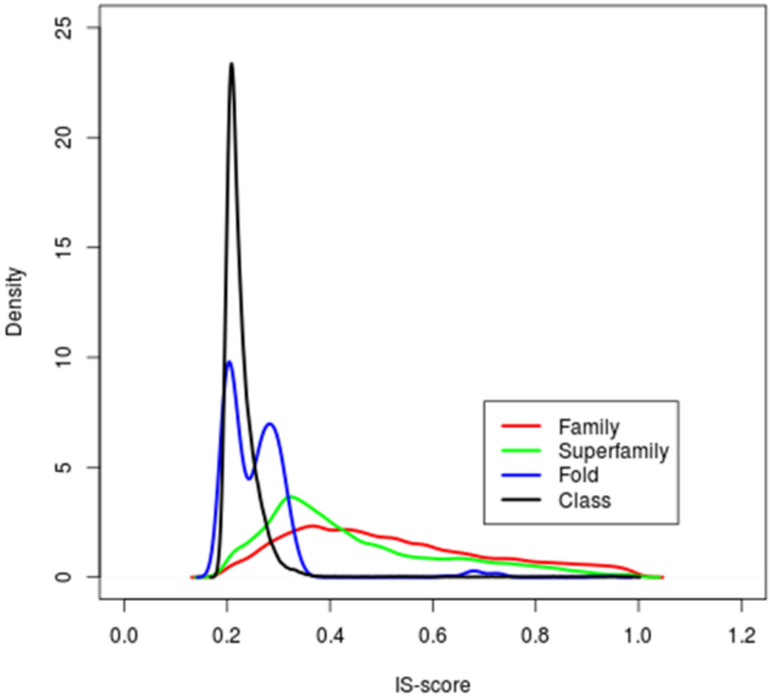


Figure 3.16 IS-score distribution of domains related at various levels in aligned pairs (SCOP). Density distribution plot of IS-score of related domains in aligned domain pairs at different hierarchical levels (class/fold/superfamily/family) of SCOP.

Table 3.3: Summary statistics of interface similarity among multidomain proteins in SCOPe database

Different levels of structural similarity	**SCOPe** **Mean (SD) of significant IS-score**
Class	0.306(0.056)
Fold/Topology	0.323(0.076)
Superfamily/Homologous family	0.469(0.16)
Family	0.535(0.18)

3.4 Conclusion

In this present work, we have comprehensively compared pairs of intra-chain domain interfaces, where domains between them are structurally closely, remotely or completely unrelated. The results showed that closely related domains form similar interfaces, whereas interfaces of distantly related domains have significant relatively large variation in interface similarities. Thus, some homology related pairs of DDIs conserve domain interfaces but many others show poorly conserved interfaces. Further, we examined the functional constraints to conserve interfaces on enzymes and found that functional constraints do not seemingly contribute towards conserving structural similarity of interfaces in homologous enzymes. Thus, interfaces can be conserved without any relatedness in the function of proteins even though domains either belong to same family or homologous superfamily.

Overall, this study investigates the scope of the structural similarity that facilitates the detection of a broad range of templates significantly divergent from the targets. Additionally, these analyses also provide important details in designing a methodology to model multidomain proteins by exploiting conserved interface residues among structural neighbors.

Chapter 4

Understanding structural relatedness of domain-domain interfaces among sequence/structurally unrelated domains

4.1 Introduction

In the previous chapter, similar domain-domain interfaces (DDIs) are observed despite there is no sequence/structural similarity between aligned domain pairs. Thus, indicating interfaces of completely unrelated domain pairs can be similar. The same property has been previously observed among protein-protein interfaces (Gao &Skolnick, 2010b). This prompted us to investigate the redundancy of domain interfaces as well as understand the structural space of DDIs.

The protein structural space is viewed as collection of known tertiary structures and it has been of great interest to understand the nature of structural space in terms of geometrical/structural relatedness. This can enhance knowledge about protein structure evolution and assist in developing approaches for prediction or designing tertiary structures. The studies over years have found that there are a finite number of secondary structure spatial arrangements with similar topological connections (folds), which can be assumed to show a general hierarchical organization and the fold space is continuous (Chothia, 1992; Finkelstein &Ptitsyn, 1987; Sadreyev et al., 2009; Valas et al., 2009). The partial similarity observed between structural fold is also referred to as ‘gregariousness’, which measures the number of folds having significant structural overlap with a given fold (Harrison et al., 2002). The study using single domain library of proteins structures has shown that the structural space is likely complete largely due to packing of compact, hydrogen-bonded secondary structural elements (Zhang et al., 2006). The completeness of the structural space indicates the continuous nature of fold space, which is quite unlikely to be caused by evolutionary divergence, instead directed by the protein folding rules (Skolnick et al., 2009). It was found that even two random protein structures could be connected through no more than seven steps using 0.4 TM-score cut-off to establish non-random structural relationships. TM-score of 0.4 indicates partial alignment with 40% coverage between structural cores of different proteins (Zhang et al., 2005; Skolnick et al., 2009). The completeness of single domain structural space was further extended to analyze nature of protein quaternary structures or protein-protein interface structural space (Gao &Skolnick, 2010b; Kim et al., 2006; Zhang et al., 2006). The study of Gao and Skolnick showed that structural space of protein interfaces is indeed degenerate mainly because of relatively flat interacting surfaces generated as a consequence of packing compact, hydrogen-bonded secondary structural elements. It was shown further that structural space of interaction interface is close to complete in terms of geometric similarity (Gao &Skolnick, 2010b) and can be utilized for prediction of protein-protein interfaces (Gao &Skolnick, 2011).

In this chapter, we have investigated the structural degeneracy of domain interfaces among domain pairs, which are unrelated by sequence and structure. Further, we also explored the connectivity of interface structural space. In general such knowledge of structurally similar interfaces among unrelated domains can facilitate modeling of domain interactions in multidomain proteins by identifying native-like interfaces. Unlike protein-protein interaction interfaces intra-chain domain interfaces are constrained in their interaction due to linker region separating interacting domains. For instance, interface accessible for interactions between domains sequentially next to each other will be restricted by allowed conformations of inter-domain linkers. However, two domains separated by one or more domains could be less constrained in their mode and available accessible surface for interaction. Considering these constraints, we analyzed degeneracy of domain-domain interfaces for following cases: a) continuous domains separated by inter-domain linkers (chained domain) constrained system, b) continuous domains separated by one or more domains (unchained domains) less constrained system, and c) discontinuous chained domains two linker constrained system and d) inter-chain domain interfaces (no constraint system) Furthermore, we compared inter-domain interfaces extracted from protein-protein interactions with all intra-chain domain interfaces to find structurally similar interfaces with a view to evaluate possibility of using either dataset to construct interface template library. We used non-sequential mode of iAlign to align domain interfaces and assessed similarity based on IS-score. Thus, this study provides geometrical similarity across the structural space of domain-domain interactions and its comparison with protein-protein interfaces.

4.2 Methodology

4.2.1 Domain-domain dataset

To incorporate constraints, as described before, between domains we constructed four separate domain-domain datasets *viz.* domain-CC-2 (continuous and chained domain of only 2 domain proteins), domain-CC-M (continuous chained domain of proteins having 3 or more domains), domain-CU-M (continuous chained and unchained domains of proteins having 3 or more domains), domain-DC-2 (discontinuous chain domains of only 2 domain proteins). The detailed steps of dataset generation are described in Figure 4.1.

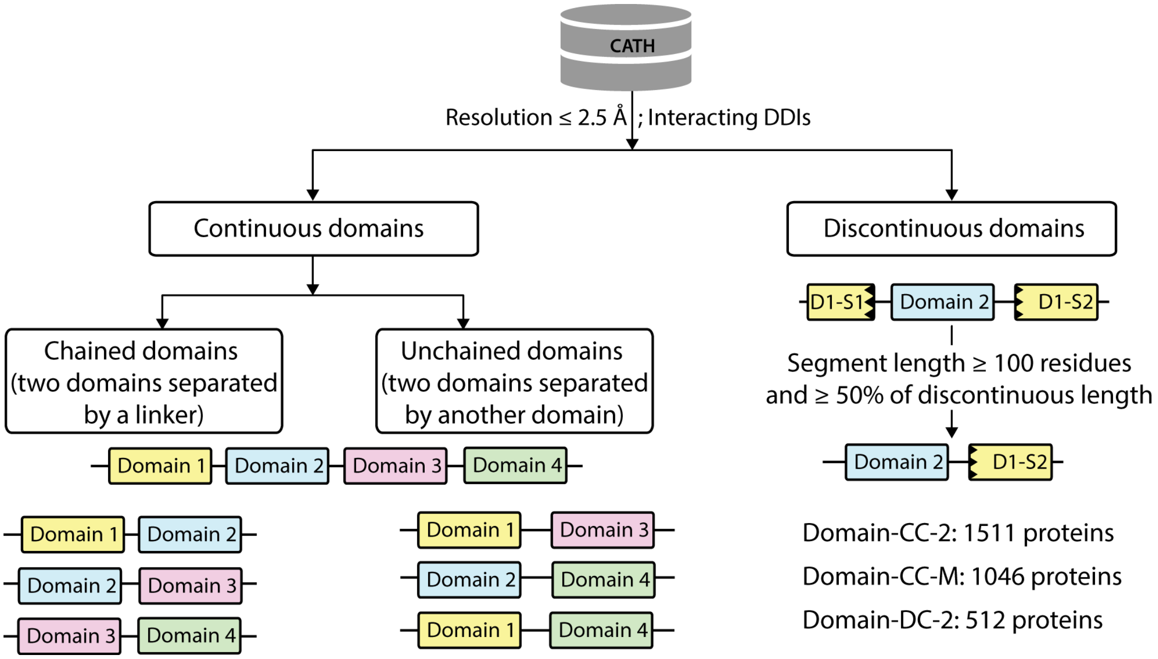

CATH
Resolution ≤ 2.5 Å ; Interacting DDIs
Continuous domains
Discontinuous domains
Chained domains (two domains separated by a linker)
Unchained domains (two domains separated by another domain)
Domain 1
Domain 2
Domain 3
Domain 4
D1-S1
D1-S2
Segment length ≥ 100 residues and ≥ 50% of discontinuous length
Domain-CC-2: 1511 proteins
Domain-CC-M: 1046 proteins
Domain-DC-2: 512 proteins

Figure 4.1 Overview of dataset. Schematic representation of methodology employed in the construction of intra-chain domain- domain dataset.

In this study, we have used CATH domain definitions and classification from database version v4.1.0. From CATH database, we took X-ray crystal structures (PDBID with chain identifier) having resolution ≤ 2.5Å with at least two classified domains. This is because CATH database defines a domain without classifying at various hierarchical levels such as class, architecture, topology and homology. Thus, obtained PDB IDs were mapped to UniProt identifiers using the mapping available, documented in EBI-SIFTS database. For each pdb entry, we used iAlign (v1.0b7) to extract interacting domains using the definition: two domains are said to be interacting if at least 20 residues are involved in interatomic contacts between interfaces of these domains. As has been mentioned in previous chapters, domain-domain interfacial residue is defined as those residues having at least one heavy atom within 4.5Å of a heavy atom in another domain. All interfacial residues together constitute the domain-domain interface for a domain pair. Following this, we generated a list of PDB entry having interacting domains for each UniProt id. Further, we selected a representative PDB structure for each UniProt entry having longest length and the best possible resolution. This dataset is divided into 2 dataset a) structures with only 2 domains (set A) and b) rest all structures (set B).

From set A (2 domain proteins), we selected pdb entries with continuous domains *i.e.* domain consisting of only one single segment. These are referred to as continuous chained domains. For pdb entries with missing residues between two domains, we imposed a criterion on the length of inter-domain linkers to consider them as continuous domains. Based on the observed minimum CATH domain length, we consider two domains as continuous having a maximum of 13 residues inter-domain linker length. Though it is a strict criterion, this would ensure that at least there is no intervening CATH domain in pdb entries having missing residues between domains. This filtering resulted in a redundant dataset of 2411 continuous domain proteins. This was used to prepare a list of 1511 non-redundant (40% sequence identity at domain level) consecutive continuous domain dataset (domain-CC-2) using the approach described in Chapter 2 section 2.2.1.2. This approach ensures that at least one domain is non-redundant at 40% sequence identity.

We used set A to extract PDB entries having either one or both discontinuous domains *i.e.* more than one segment defines a domain. Since, a domain has linear disconnected sequences; it poses a serious problem in sequence alignment to create non-redundant dataset. To address this issue, we examined the coverage and length distribution of segments in all CATH discontinuous domains. This analysis showed that ~63% of discontinuous domains have a major segment, which contributes ≥ 70% to the domain length. Based on this, we decided to represent discontinuous domain with one major segment with empirical condition on the length and its contribution to domain length as well as to domain interface. A major segment represents a domain if it is the longest segment, with a minimum length of 100 residues, and it contributes ≥ 50% to the total discontinuous domain length. Additionally, this major segment should consist at least 20 interfacial residues, thus ensuring that the selected segment is the representative of an interacting domain. We obtained 851 redundant PDB entries employing the above criteria that were used to construct 512 non-redundant dataset (domain-DC-2) at 40% sequence identity using the method as described in Chapter 2.

To examine the effect of one or more domain lying between two interacting domains, we used multidomain proteins with at least three domains (setB). Using set B (given above), we generated two separate subsets: a) chained continuous domain pairs and b) all combinations of consecutive and non-consecutive continuous domain pairs. Here, non-consecutive domains have at least one intervening domain between two interacting domains. Following the procedure to construct domain-CC-2 dataset as described before, resulted in list of 1113 domain pairs. For the second set, we made all possible combinations of interacting domains for a given pdb entry that lead to a set of 1553 domain pairs. These redundant datasets were made non-redundant at 40% sequence identity using the approach described before in Chapter 2. Thus, final dataset has 759 entries in consecutive continuous domain dataset (domain-CC-M) and a list of 1046 domain pairs in consecutive/non-consecutive continuous dataset (domain-CU-M). All datasets are given in Chap04-Dataset.xlsx available at https://github.com/riviverma/thesis-md-proteins/.

4.2.2 Protein-protein interaction dataset

To prepare protein-protein interaction dataset or inter-chain domain interactions, we culled the 17659 heteromers protein-protein interaction dataset obtained from previous work (Maheshwari &Brylinski, 2015) to extract only inter-chain domain interactions. From this dataset, we considered interacting proteins with only unique chain order and discarded discontinuos domains. Next, this dataset was curated based on the criteria given below: First, we selected protein pairs having structures with resolution ≤ 2.5Å and CATH domains defined for both proteins (1366). Next, if a multidomain protein is involved in PPI, then we identify interacting domain pairs (these are interacting inter-chain domains) between two proteins using iAlign criteria, as has been given in the previous section and only consider protein pairs in PPIs having valid interacting domains between two proteins (Figure 4.2). We performed this additional step because we need to compare fold of respective domains in the process to find best match of the query domain-domain interface. These steps led to a total of 1464 interacting inter-chain domain pairs in 1233 dimer entries.

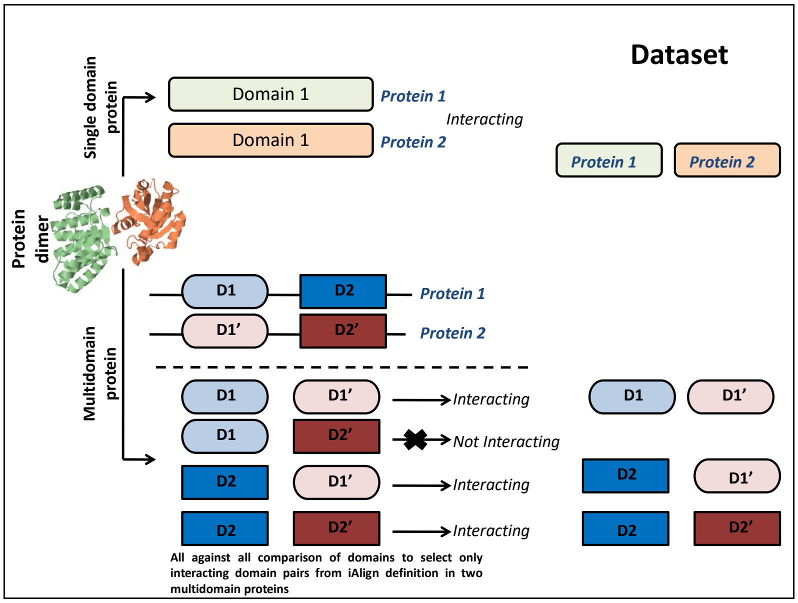


Figure 4.2 Overview of generation of intra-and inter-chain dataset. Schematic shows the generation of inter-chain and intra-chain domain-domain dataset.

4.3 Results and Discussion

Previous study on PPI have shown that interaction interfaces are degenerate mostly due to functional restraints and the packing of compact, hydrogen-bonded secondary structures, which generates flat interacting surfaces having common geometrical shapes. Unlike PPI, domain interactions could be constrained by the length of inter-domain linkers (even encompassing one or more domains), which might restrict the accessible surface available for domain interactions. To investigate structural degeneracy among domain interfaces, we have structurally aligned interacting interfaces of non-redundant consecutive/non-consecutive domains, which are sequence/structurally unrelated, formed of continuous/discontinuous segments. Furthermore, we have compared these interfaces with protein-protein interaction interfaces to examine degeneracy between intra-and inter-chain interfaces. From the aligned interfaces, the best structural match for a given interface is the one with the highest IS-score. Below we describe results of similarity searches in DDIs and comparison with inter-chain interfaces.

4.3.1 Similarity among consecutive continuous domain-domain interfaces

As described before, we have constructed two non-redundant (pairwise sequence identity < 40%) consecutive and continuous domains datasets *viz.* domain-CC-2 (1511 structures) and domain CC-M (759 structures). The motive of using latter dataset is to investigate whether consecutive domain interfaces are affected in presence of additional domains.

As mentioned before, domain-CC-2 dataset consists of DDIs separated by one inter-domain linker. Since our objective is to detect similar interfaces formed by domain pairs without any significant structural or sequence relationship, we generated a list of structurally dissimilar proteins for each member of domain-CC-2, by searching against dataset DDIs (1511 structures). The list of dissimilar domain pairs were defined on following conditions: a) No two domains between domain pairs lie within same CATH topology (fold), b) has no significant structural similarity as assessed by TM-score, *i.e.* for all combinations of domains in domain pairs the TM-score<0.4, and c) no domain sequence have significant sequence similarity as assessed by PSI-BLAST with E-value > 1 (this ensures no domain sequence relatedness). The conditions (a) and (b) removes structurally related domains in aligned pairs. Thus, we obtained a list of structurally dissimilar proteins for each 1511 proteins (domain-CC-2), which served as template library to search query DDI for identifying similar interfaces using iAlign. The IS-score was used as the metric to measure interface similarity and it length of query DDI was used for normalization factor in IS-score (see methods).

The results of the closest interfacial match for each 1511 DDI as assessed by the best IS-score are shown in Figure 4.3 and relevant statistics are summarized in Table 4.1. As is shown the mean (standard deviation (SD)) IS-score of the best interfacial similarity is 0.307 (0.026). Importantly, this is higher than the mean (SD) IS-score of 0.207 (0.036) obtained for the best matches among random PPI interfaces (Gao &Skolnick, 2010b). Thus, showing that domain interfaces of sequence/structurally unrelated domains are not random. Further, ~88% of these DDIs have the best structurally similar interface with a significant IS-score (p-value <0.05). The mean (SD) RMSD of interface is 3.3 (0.5) Å, a mean (SD) residue coverage f_{res} of 86% (9%), and a mean (SD) contact coverage f_{con} of 55% (9%). Both average residue and contact coverage were calculated with respect to the query DDI. This shows that overall significant region of query DDI was part of aligned region and contacts of aligned region were similar between two domain pairs. This analysis finds that for most intra-chain domain interfaces one could find structurally similar interface even though interfaces are formed of sequence or structurally unrelated domains. This redundancy of interfaces has also been previously observed among interfaces formed in PPIs.

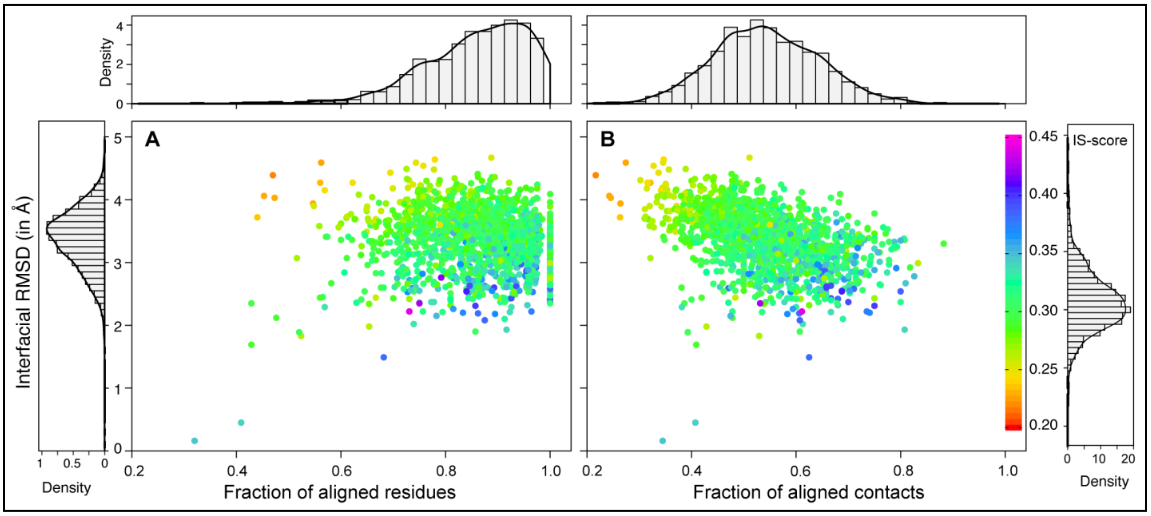


Figure 4.3 Plot of the best interface matches for Domain-CC-2 dataset. Scatter plot of the interfacial RMSD versus (A) fraction of aligned residues (f_{res}) and (B) fraction of aligned contacts (f_{con}) for the closest match of 1511 domain-domain interfaces extracted from proteins having only two CATH classified structural domains. Each point is represented using color gradient based on IS-score. Histogram and density plots of RMSD, f_{res}, f_{con} and IS-score are shown surrounding main scatter plot.

Following the criteria described above to define sequence/structural unrelated domain pairs, we searched for similar interfaces for each of 759 intra-chain domain interfaces (domain-CC-M) derived from proteins having more than 2 domains. The search statistics of the best structural interface match is summarized in Table 4.1 and shown in Figure 4.4. The mean (SD) IS-score of the best matched interface is 0.298 (0.029) and ~74% of these have significant IS-score. The other parameters of fraction aligned residues and contacts are similar as other datasets (Table 4.1). This shows that similar interfaces could be found among structurally unrelated domain pairs in consecutive domains in proteins having more than 2 domains. Importantly, these are observed for interfaces formed by consecutive domains, which can be constrained in their intra-chain domain interactions due to inter-domain linkers.

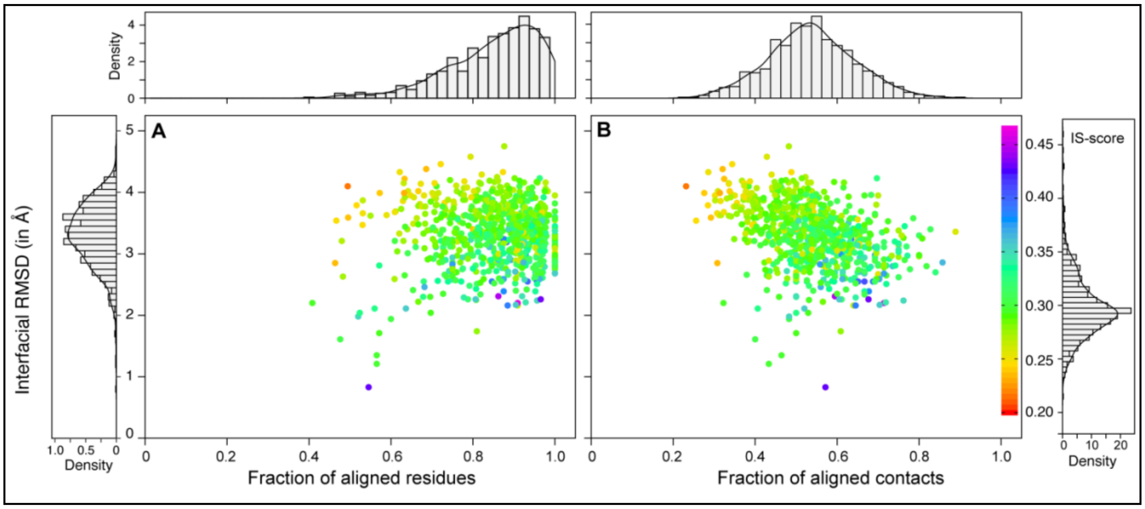


Figure 4.4 Scatter plots of the best interface matches for Domain-CC-M dataset. Scatter plot of interfacial RMSD versus (A) fraction of aligned residues (f_{res}) and (B) fraction of aligned contacts (f_{con}) for the closest match of 759 consecutive continuous domain-domain interfaces extracted from proteins with >2 CATH structural domains. Each point is colored based on IS-score. Histogram of IS-score is shown.

Table 4.1: Summary statistics of the best similar interfaces for datasets

Dataset	Mean (SD) of				Significant matches
	IS-score	RMSD	Residue coverage	Contact coverage	
Domain-CC-2 (1511)	0.307 (0.026)	3.3 (0.5)	86% (9)	55% (9)	88%
Domain-CC-M (759)	0.298 (0.029)	3.2 (0.5)	86% (10)	55% (10)	74%
Domain-CU-M (1046)	0.30 (0.027)	3.3 (0.5)	86% (10)	55% (10)	78%
Domain-DC-2 (512)	0.286 (0.02)	3.5 (0.3)	83% (10)	51% (8)	64%

In previous work on PPI interfaces (Gao &Skolnick, 2010b), it was shown that that protein-protein interfaces are observed to be structurally degenerate mostly due to functional constraints, physical constraints as a result of packing of compact hydrogen bonded secondary structure elements, and almost flat interfaces in which case geometrical similarity can be established. Using the same reasoning, we performed detailed analysis to understand basis of similar interfaces and observed that except functional constraints other two features can explain the existence of similar interfaces. As has been observed before in PPIs, similar interfaces are formed due to limited ways of packing of secondary structure elements despite dissimilar folds between domain pairs. Moreover, multidomain proteins are mostly globular like single domain proteins and packing of secondary structures may contribute to protein stability. The examples of packing of α-helices and/or β-strands are shown in Figures 4.5A-C. The DNA polymerase III beta sliding clamp protein consists of 3 topologically equivalent domains of α/β class that has anti-parallel helices bracketing the four stranded anti-parallel β-sheet (Wolff et al., 2014). Two such anti-parallel β-sheets of consecutive domains interact to form an extended β-sheet and two parallelly oriented β-strands mainly constitute the domain interface (Figure 4.5C). The pullulanase enzyme (2fh8A) has four domains. Of these, first two and last domain belongs to Immunoglobulin-like fold (all β-class) and rest one domain adopts TIM barrel fold (α/β class) (Mikami et al., 2006). The β-sheets from first two domains come together in a parallel β-strand orientation to form interface of these domains (Figure 4.5C). As shown in Figure 4.5C, the first two consecutive domain interfaces of DNA polymerase III beta sliding clamp and pullulanase are similar due to packing of parallel β-strands, even though domains belonging to different CATH class.

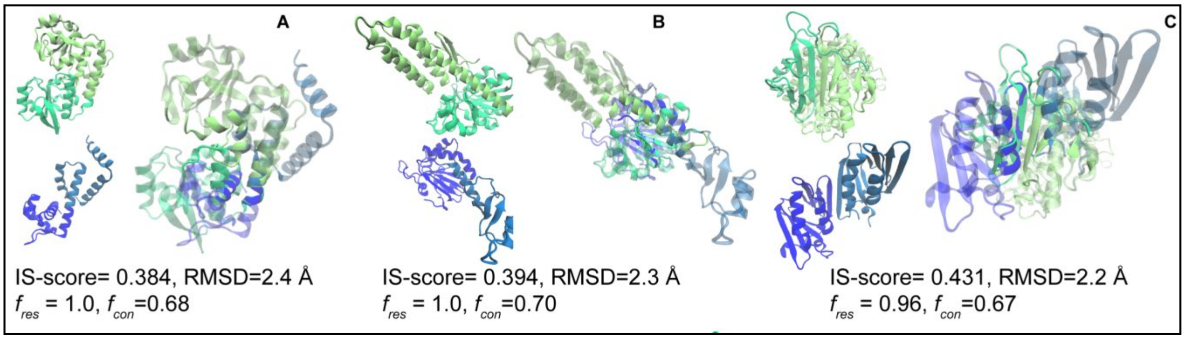

A
B
C
IS-score= 0.384, RMSD=2.4 Å
f_{res} = 1.0, f_{con}=0.68
IS-score= 0.394, RMSD=2.3 Å
f_{res} = 1.0, f_{con}=0.70
IS-score= 0.431, RMSD=2.2 Å
f_{res} = 0.96, f_{con}=0.67

Figure 4.5 Examples of similar domain-domain interface pairs from multidomain proteins. Two domains of template protein are shown in green and lime colors, while target protein domains are shown in blue and sky blue colors. A) Periplasmic receptor CeuE (domains 1 and 2 of 4inoA) and manganese transport regulator (MNTR) protein (domains 1 and 2 of 2f5fB), PDB identifier is followed by chain identifier. B) RhoA-dependent invasion protein (domains 1 and 2 of 4ldrB) and Peroxiredox in protein (domains 1 and 2 of 2v2gA). C) Pullulanase enzyme (domains 3 and 4 of 2fh8A) and DNA polymerase sliding clamp (domains 1 and 2 of 4tr8B).

The flat interfaces have been observed in PPI that can show geometrical similarity with ease and more so in non-sequential alignment of interfaces. The investigation of domain interface alignments showed that these are also rather flat (Figure 4.6A) and observed similar interfaces between domains pairs having different secondary structure elements at the interfaces. The interface of domains 1 and 3 of hyaluronate lyase enzyme (1n7oA) that belongs to mainly beta class aligns with mainly helical domains (5 and 6) of serum albumin (4f5uA) as shown in Figure 4.6B. Most of the non-significant cases are due to one of the domain enveloping other domain and in some cases, the interaction interface is very small constituting of loops.

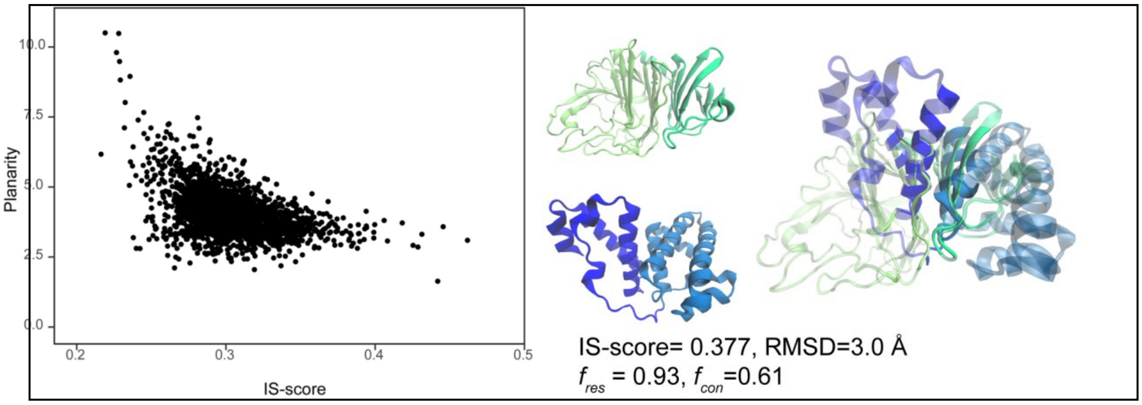


Figure 4.6 Domain- domain interfaces are flat. A) Scatter plot showing relationship between planarity of domain-domain interface to the best IS-score of interface obtained for each representative 2270 consecutive domains. Planarity is measured using PRINCIP program in SURFNET (Laskowski, 1995) suit of programs that is a root-mean square deviation between interface Cα-atoms and the best fit of plane through the interface Cα-atoms. B) Hyaluronate lyase enzyme (domains 1 and 3 of 1n7oA) and serum albumin (domains 5 and 6 of 4f5uA).

4.3.2 Similarity among consecutive/non-consecutive domain-domain interfaces

Having shown degeneracy for consecutive domains in two-domain proteins, we extended the same to non-consecutive domains, which will have little or no direct restrictions by inter-domain linkers as consecutive domains. In this analysis, we constructed a list of 282 non-redundant domain pairs, which have non-consecutive domains. These were combined with previous set of multidomain consecutive domains (domain-CC-M) to generate a combined set of 1046 (domain-CU-M) domain pairs. Using similar criteria described before, search for the best domain interface match resulted in a mean (SD) IS-score of 0.30 (0.027). Of these, ~78% of proteins have statistically significant interface (P-value<0.5). The distribution of various parameters for the best match of interfaces is shown in Figure 4.7 and summarized in Table 4.1. The best matched domain alignment have an interface average rmsd (SD) of 3.3 (0.5) Å, a mean (SD) residue coverage f_{res} of 86% (10%), and a mean (SD) contact coverage f_{con} of 55% (10%), respectively. These analyses found that interfaces are degenerate even when one or more domains separate aligned domain pair.

Of 282 non-consecutive domain pairs, 76% of domains have significantly similar interfaces (P-value <0.05) having mean (SD) IS-score 0.30 (0.026). The visual inspection of non-significant alignments showed that some of these non-consecutive domains are involved in interactions with other domains. Moreover, these interfaces have one of the interfaces is enveloping surface of other domain.

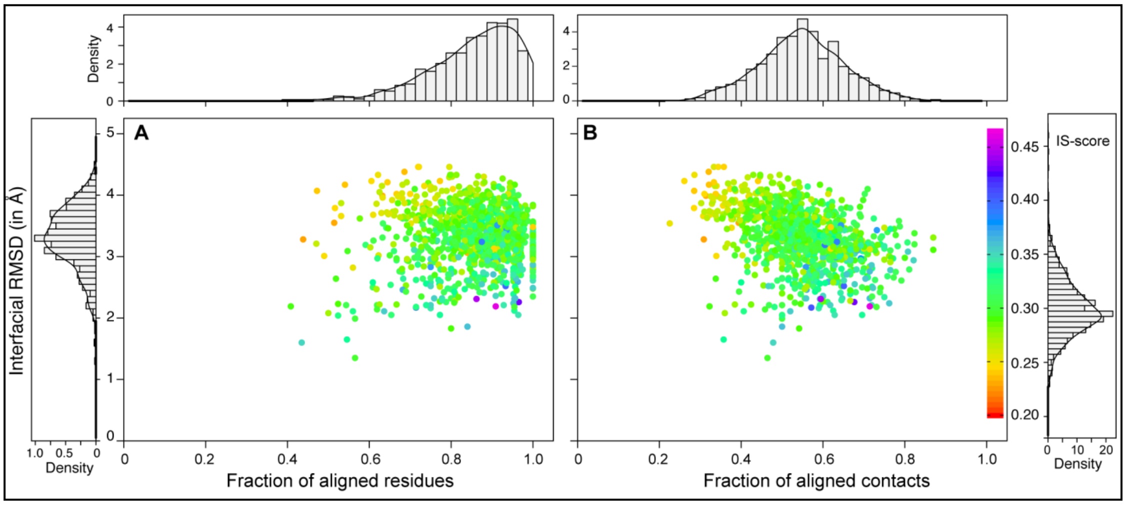


Figure 4.7 Scatter plots of the best interface matches for domain-CU-M dataset. Scatter plot of interfacial RMSD versus (A) fraction of aligned residues (f_{res}) and (B) fraction of aligned contacts (f_{con}) for the closest match of 1046 domain-domain interfaces extracted from proteins having > 2 CATH structural domains. Distribution of IS-score is shown as histogram.

4.3.3 Similarity among discontinuous domain-domain interfaces

Since structural domains can consist more than one segment (discontinuous), we performed similar analysis of finding the best structural match for interfaces involving one or both discontinuous domains. Such discontinuous domains involve two or more inter-domain linkers. Through this analysis, we investigated whether structural degeneracy is observed even in interfaces of discontinuous domains. Since generating non-redundant dataset with multiple segments of sequence is not appropriate, we considered only longest segment with empirical criteria for representing discontinuous domain and prepared the dataset (domain-DC-2) of 512 domain pairs (see methods for detail). Following the same approach to find structural matched interfaces, the results of the best structural matches for discontinuous domain interfaces are shown in Figure 4.8. The mean (SD) IS-score of the best interface match for discontinuous domains is 0.28 (0.02). Of these, 64% of domain pairs have statistically significant interface similarity. These domain pairs have interface average rmsd (SD) of 3.5(0.3) Å, a mean (SD) residue coverage f_{res} of 83% (10%), and a mean (SD) contact coverage f_{con} of 51% (8%), respectively. This shows interfaces of discontinuous domains are degenerate and suggesting this is a general feature observed among DDIs.

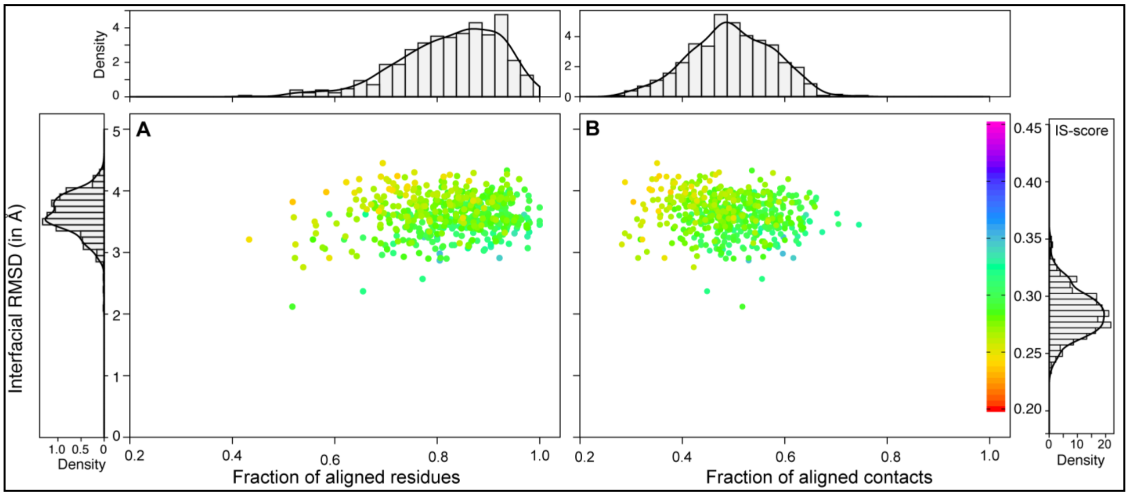


Figure 4.8 Distribution of scores for the best structural interface matches of domain-DC-2. Scatter plots of interfacial RMSD versus (A) fraction of aligned residues (f_{res}) and (B) fraction of aligned contacts (f_{con}) for the closest match of 512 consecutive and non-continuous domains extracted from proteins with only two CATH structural domains. Each point is colored based on IS-score.

4.3.4 Similarity between domain-domain and protein-protein interfaces

As domain-domain interface of consecutive/non-consecutive continuous or discontinuous domains are shown to be degenerate, it can be suggested that these might share similarity to inter-chain domain interfaces in PPI under no sequence or structural similarity at the level of domains. Moreover, DDIs and PPIs have been shown to share similar physiochemical properties. Such comparison would also facilitate including inter-chain domain interfaces in template library of domain interfaces to improve modeling of interfaces in multidomain proteins by identifying native-like interfaces.

For this analysis, we constructed protein-protein interaction dataset and pruned to remove structures without any CATH assigned domains. Further, PPIs having valid inter-chain domains (from two monomers) were considerd to construct non-redundant inter-chain domain interfaces (*see* Chapter 2). Following the procedure described above to find the best structural match for a query interface, we searched for the best structural match of an inter-chain domain interface (target-PPI) in the template library of the DDIs (consisting of datasets: domain-CC-2, domain-CC-M, and domain-CU-M). The highest IS-score found among all 3 template library dataset is taken as the best matched domain interface for each PPI interface. Figure 4.9 shows the distribution of parameters of the best structural match to inter-chain interfaces and other statistics is summarized in Table 4.1. The mean (SD) IS-score of the closest match of inter-chain interface to intra-chain domain interface is 0.311 (0.031). Among these, significant matches are found for ~86% of protein-protein interfaces. These significant domain pairs have interface average RMSD (SD) of 3.2 Å (0.45). These have mean (SD) residue coverage f_{res} of 88% (10%), and mean (SD) contact coverage f_{con} of 58% (10%). This shows inter-chain domain and intra-chain domain interfaces are similar despite unrelated domains.

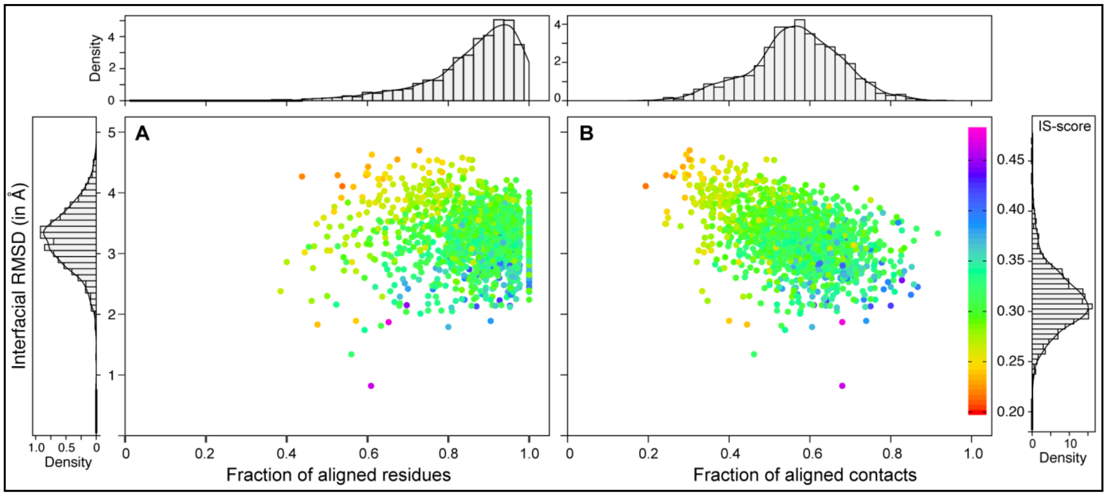


Figure 4.9 Structural comparisons of inter-chain domain and intra-chain domain interfaces. Scatter plots of interfacial RMSD versus (A) fraction of aligned residues (f_{res}) and (B) fraction of aligned contacts (f_{con}) for the closest match of 1464 protein-protein interfaces with intra-chain domain interfaces. Distribution of IS-score is shown as histogram.

Further, we found that similar packing of secondary structures as well as flat interfaces are mostly responsible for observed similarities between intra-chain and inter-chain domain interfaces. This is illustrated in representative examples shown in Figure 4.10. The first example shows a noticeable overlap between the PPI of 4pjeC/E [formed by major histocompatibility complex class I protein (domain C02) and domain 1 of T-cell receptor] and the DDI of 1ospO [outer surface protein A with two domains having antiparallel β-sheet topology], which consists of similar antiparallel β-strands, that is detected in the interface alignment (Figure 4.10A). Another similar structural interfaces, between packed anti-parallel β-sheets, which belong to the β-sandwich scaffold, have also been demonstrated between intra-chain domains of 2o62A and inter-chain domains of 3qnzB and 3qnzA (Figure 4.10B).

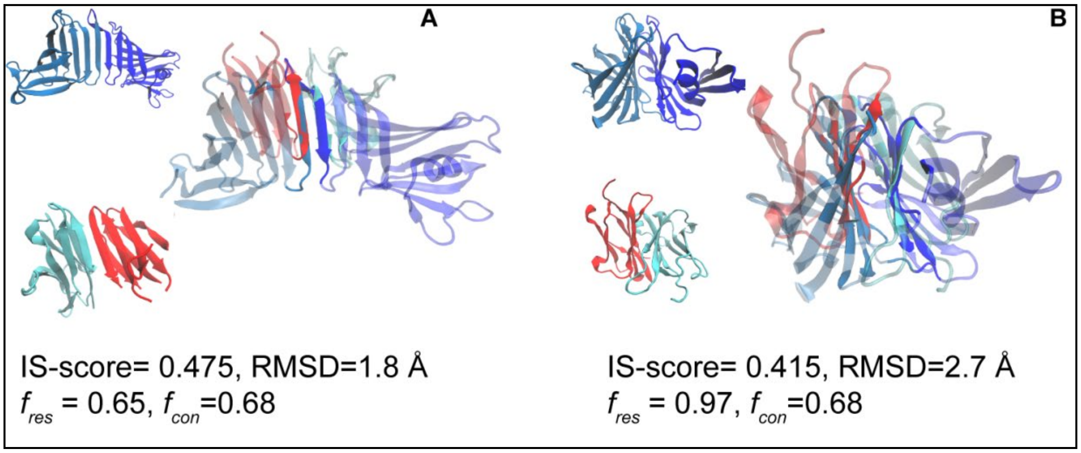

A
B
IS-score= 0.475, RMSD=1.8 Å
f_{res} = 0.65, f_{con}=0.68
IS-score= 0.415, RMSD=2.7 Å
f_{res} = 0.97, f_{con}=0.68

Figure 4.10 Examples show similarity between domain-domain and protein-protein interfaces. Two domains of template (domain-domain interface) protein are shown in blue and sky blue colors, while two interacting proteins are shown in red and cyan colors. A) Two domains of outer surface protein A (domains 1 and 2 of 1ospO) is aligned with human major histocompatability complex with T-cell receptor (4pjeC (domain C02/ 4pjeE (domain E01). B) Structural interface alignment of protein of unknown function (domains 1 and 2 of 2o62A) is complexed with antibody fragments (3qnzB (domain B02)/3qnzA (domain A02).

Since intra-and inter-chain domain interfaces are degenerate, we investigated whether combining inter-chain with intra-chain domain interfaces could increase the number of significant structural matches for DDIs. For this, we searched for the closest match for consecutive continuous domain interfaces (1511) against template domain interface library consisting of a) 1464 inter-chain, b) 1511 intra-chain, and c) combined set of both inter and intra-chain domains. The closest match of 1511 interface against 1464 PPI resulted in mean (SD) IS-score of 0.304 (0.027) having ~85% of significant matches. The best structural neighbors of

1511 interfaces in a combined dataset of both intra-chain and inter-chain domain interfaces achieve average (SD) IS-score of 0.314 (0.027). This is statistically significantly (p-value << 0.001 in the paired t-test) different than the mean IS-score of 0.304 and 0.307 obtained from searching against inter-chain and intra-chain domain respectively. Interestingly, the number of significant IS-score matched interfaces also increased from ~85 (88) for inter (intra)-chain to 93%. Thus, suggesting that including inter-domain interfaces can enrich the template interface library constructed from intra-chain domain interfaces.

4.3.5 Connectivity of DDIs interface structural space

To understand the continuity and connectivity of structural space for intra-chain domain interfaces, we analyzed this using a directed graph. A directed graph consists of domain interfaces as vertices, which are joined by a directed edge that points from template to query (target) interfaces drawn based on a predefined threshold of IS-score and edge weight is one. We considered a directed edge because IS-score is not transitive and it is not same for two interfaces when target interface is changed, i.e., IS-score for A-B is not same as B-A, where B and A are target interfaces respectively. An interface I_A is said to be k^{th} neighbor of I_B, if the minimum path length from node I_A to I_B is $\leq k$. Since domain-CC-2 dataset is the largest among DDIs, we performed network analysis only for this dataset. The fraction of all possible directed pairs at a given k^{th} neighbor for varying IS-score is shown in Figure 4.11. We observed that at a significant IS-score threshold of 0.26 about ~84% of all directed interface pairs are at most separated by the eighth neighbor. The largest strongly connected component (LSCC), where all nodes are connected bidirectionally to at most k^{th} neighbor consists of ~83% of interfaces at a threshold of 0.26 and k=8 (Figure 4.11B). The related size of LSCC drops drastically to approximately 3% at IS-score of 0.30, which probably is the critical threshold below which nodes are densely connected and structural space is continuous.

As we have found that including inter-chain interfaces of PPIs improve overall closely related matches for intra-chain domain interfaces, we examined whether connectivity of structural space can be improved by including inter-chain domain interfaces. For this, we utilized search results of domain-CC-2 against PPI and vice-versa to include only edges between interfaces (nodes) from PPI and DDI. The summary of all possible directed pairs at given *k* as a function of IS-score and LSCC at given IS-score as a function of k are shown in Figures 4.11C and 4.11D respectively. At IS-score threshold of 0.26, roughly 90% of all directed pairs are at the most eighth neighbor and LSCC consists of ~89% of interfaces at k=8. The LSCC increases by ~6% in comparison to graph without inter-chain interface connectivity. The LSCC for IS-score threshold of 0.30 is also increased to ~6%. This shows that structural space of domain-domain interface is continuous and connected, which improves by including inter-chain interfaces.

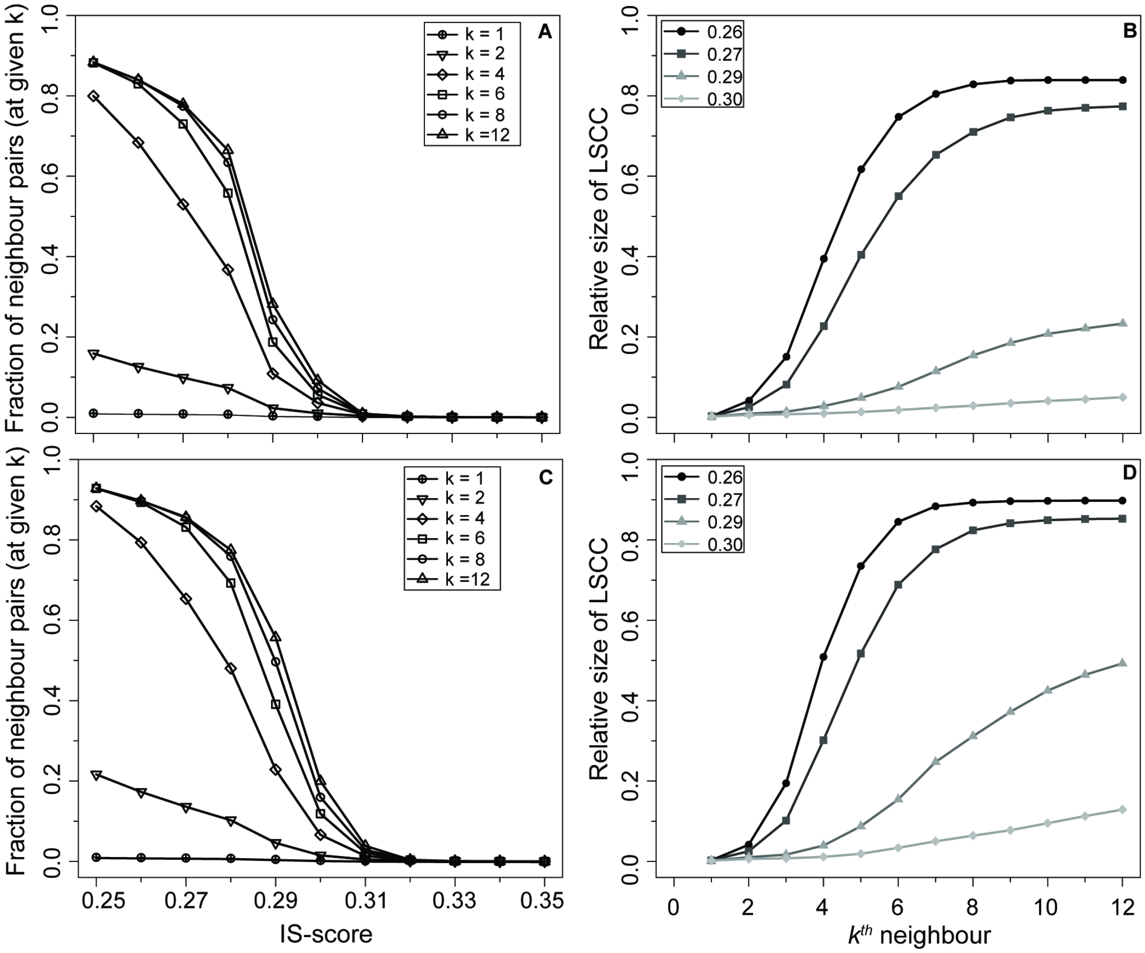


Figure 4.11 Domain interface graph connectivity. The fraction of directed pairs of nodes (interfaces), which are connected with at most k-th neighbor plotted as a function of IS-score (A) and (C). Fraction is $n_k/(Nx(N-1)$, $n_k$is the number of kth neighbor pairs and N is total number of interfaces in a graph. The relative size of LSCC at different k for graphs generated at a given IS-score thresholds (B) and (D). The graph for intra-chain domain interfaces is shown in A and B, whereas combined intra-and inter-chain interface network is shown in C and D panels.

4.4 Conclusions

In the present work, we have investigated DDI interface structural degeneracy among structurally unrelated consecutive/non-consecutive domains, which consist of continuous or discontinuous segments. The results showed that intra-chain domain interfaces are also degenerate as has been observed for interfaces of PPI. Moreover, inter-domain linker constraints does not affect the general features of interfaces and it is likely that interfaces are maintained by allowing linker length to accommodate appropriate interactions among domains to either facilitate their function and/or stability.When we analyzed interfaces to understand reasons for similar interfaces, we observed that similar interfaces are because limited ways of packing of secondary structural elements and flat nature domain-domain interfaces. This flat nature allows geometrical match even between different interfacial secondary structural elements. Overall, it indicates a possibility that domain interactions probably evolve from non-specific interface to specific interactions during evolution based on its functional or structural constraints. Further searching for similar inter-chain interfaces within intra-chain domain interfaces showed degeneracy among them, even, when domains are from the same or different protein. The graph analysis of relatedness among interfaces showed that domain-domain interface space is highly connected and continuous, which increase on considering inter-chain domain interfaces. This information could be exploited to construct template interfaces, which can assist in interface modeling in multidomain proteins. Moreover, this study suggests towards a possibility that interaction domain interfaces evolve from a non-specific to specific interaction depending on the functional/structural significance of interfaces.

Chapter 5

A method to improve ranking of docked domain structures using interface constraints

5.1 Introduction

The average number of domains in prokaryotic and eukaryotic proteins is found to be 1.5 and 2.1 respectively (Brocchieri &Karlin, 2005; Apic et al., 2003; Zmasek &Godzik, 2012; Ekman et al., 2005). Each domain within a multidomain protein usually has a 3D compact shape associated with independent functions, which are usually conserved among homologous proteins. Domains can be found in various domain architectures with some present as single domain as well. It has been argued that function of protein can be elucidated by associating function to their constituent domains. Since tertiary structure of proteins can aid in deciphering their function, there has been concerted effort to determine structures of proteins. Presently, single domain proteins dominate PDB database and there is a greater need to obtain tertiary structure of multidomain proteins either experimentally or computational prediction methods. This can involve methods to assemble domains to correctly predict orientation of domains (Ben-Zeev et al., 2005). The computational methods known for structure prediction include: homology modeling, threading and *ab initio* approaches. In case of multidomain proteins, homology modeling can assist in determining orientation of domains when a reliable template is found for the target protein. However, in many instances it is likely that no templates are found for certain domains or despite having templates for every domain of a multidomain protein the interactions among domains are not available. This is possible when domains of target protein identifies template from different protein structures. Moreover, even if there exists a homologous template, the domains may not interact in a similar way as it could have been in the target protein (Aloy & Russell, 2002). As a result, the focus on *ab initio* approaches has been increasing in addition to homology modeling. Among multiple approaches, docking offers a promising tool for modeling of multidomain proteins through *ab initio* approach of assembling domains. In this approach the homology modeling is used for modeling individual domains but domains are assembled *ab initio* by relying on docking of domains. Moreover, Critical Assessment of Prediction of Interactions (CAPRI) experiments on blind tests of prediction of interactions have shown that docking methods have been successful in predicting the structure of the protein complexes.

5.1.1 Protein docking

In computational scheme, docking tries to find the best match between two molecules namely receptor and ligand. The problem of molecular docking can be defined as follows: predict the "correct" bound association given the atomic coordinates of two molecules. Generally, no additional data is provided to perform docking. However, additional biochemical information, especially knowledge of binding sites, may be given which can greatly facilitates the problem of docking. Nevertheless, it should be noted that there are multiple binding sites present on the surface of the protein but while docking, it is assumed that the primary site of binding usually participate in bound conformations (Halperin et al., 2002). Usually, docking is performed between two protein chains, where whole proteins are "docked" to generate models of the bound protein complex. Similarly, multidomain protein can be modeled by docking of modeled domains to find appropriate interacting pose between domains in multidomain protein. Modeling the structure of the multidomain protein has implications in the field of protein complex modeling, as it is possible that any of the components of the complex is made up of multiple domains. The problem of multidomain modeling can be tackled using divide and conquer approach instead of performing docking directly where firstly; domain orientations are modeled followed by the assembly of domains through docking (Cheng et al., 2008).

In general, docking involves two steps. In the first step, considering the two individual proteins as rigid bodies and using their atomic coordinates, a large number of candidate alternative conformations are generated. This step is called exploration step in which only a small portion of the conformational space is searched while keeping the balance between amount of search space examined and computational expense (Hernández-Santoyo et al., 2013).

Secondly, scoring functions are used to rank these potential solutions, called as refinement step. The scoring functions consist of mathematical functions, which predict the strength of binding affinity between components forming the complex. These functions generally include geometric complementarity, electrostatic interactions, buried surface area and other energy potential functions. Electrostatic interactions are included through coulomb potential scoring functions in FFT based methods such as ZDOCK (Chen &Weng, 2002). In the last recent years, a number of different algorithms and scoring functions have been developed (Eisenstein & Katchalski-Katzir, 2004; Halperin et al., 2002; Smith & Sternberg, 2002; Vajda & Camacho, 2004) with different accuracies and computational efficiencies. Once the near- native solutions are known, they are subject to further refinement.

5.1.2 Fast Fourier Transform (FFT)

The first step of the docking involves the efficient representation of the protein structures that need to be docked. FFT method is employed for the same and it was first proposed by Katchalski-Katzir (Katchalski-Katzir et al., 1992). In this method, the protein structure is projected on a cubic grid of size N^3. To mark the relative position of a molecule, each grid point is given some weight. For all translations of one protein relative to the other, the correlation function of the weights associated with the two proteins is calculated by FFT and the calculation is repeated for all orientations. The Figure 5.1 illustrates the procedure of FFT.

5.1.3 Assessment of docking results

The success of docking programs in predicting a correct docked pose in CAPRI experiments is usually evaluated using two parameters (i) fraction of native contacts and (ii) RMSD. A native contact between two residues is defined when the distance between two heavy atoms of both the residues is less than 5Å (Lensink &Wodak, 2013) in interacting protein structures determined experimentally. Therefore, having greater fraction of native contacts recalled in a model structures suggest a better docked pose. The second assessment criteria is interface root-mean-square deviation (iRMSD) which consider residues lying at the interface and measures the distance between the experimentally known positions of backbone atoms in the reference structure and the equivalent residues in the predicted one after superimposing the two structures. However, the major challenge in searching the correct poses lies in the flexibility of the interacting protein chains. The search efficiency is determined by number of degrees of freedom included in the conformational search (Sousa et al., 2006). An iRMSD of less than 2Å is usually considered as a good performance.

In this chapter, we describe an approach to model multidomain protein using domain-domain rigid body docking, where the primary focus is to detect native like interface structures. As a preliminary analysis, we evaluated whether rigid body docking can generate docked conformations near native to domain-domain interfaces and such docked orientations can be identified using structural alignment to DDI template library.

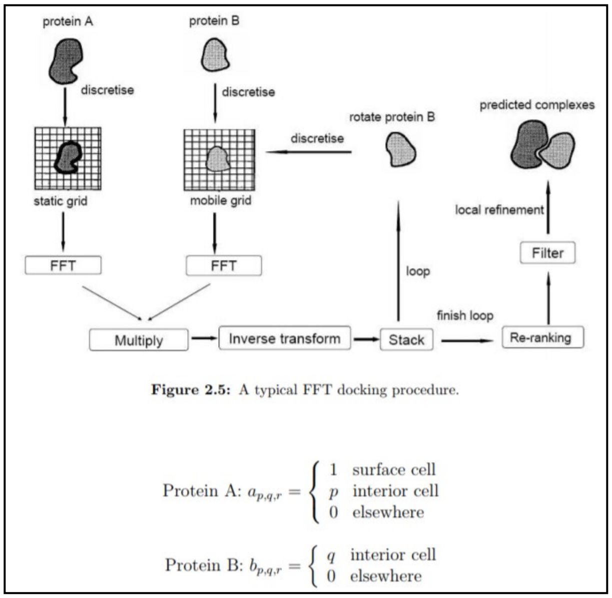


Figure 2.5: A typical FFT docking procedure.

$$\text{Protein A: } a_{p,q,r} = \begin{cases} 1 & \text{surface cell} \\ p & \text{interior cell} \\ 0 & \text{elsewhere} \end{cases}$$

$$\text{Protein B: } b_{p,q,r} = \begin{cases} q & \text{interior cell} \\ 0 & \text{elsewhere} \end{cases}$$

Figure 5.1 Schematics of FFT docking procedure. Figure showing steps of FFT docking (as adapted from *http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.891.6031&rep=rep1&=type=pdf*)

5.2 Methodology

5.2.1 Construction of benchmark dataset of docking of interacting domains

We compiled a benchmark dataset of non-redundant two domain proteins to evaluate the reliability of docking to find near-native domain-domain interfaces. For this, we considered all two-domain proteins having continuous domains (single segment domain) as defined in CATH (v4.1.0) and resolution ≤ 2.5Å. Of these, we took proteins having intra-chain domain interactions as defined using interatomic distance criteria. Subsequently, these filtered pdb entries were made non-redundant at 60% sequence identity using CD-HIT. This resulted in a set of 1375 non-redundant interacting intra-chain domain-domain proteins (Figure 5.2). Due to limitations of computing resources, we performed benchmarking on a subset of 1107 domain-domain pair. This set of interacting pair of domains is referred to as native structure. The dataset (Chap05-Dataset.xlsx) is available at GitHub repository (https://github.com/riviverma/thesis-md-proteins/).

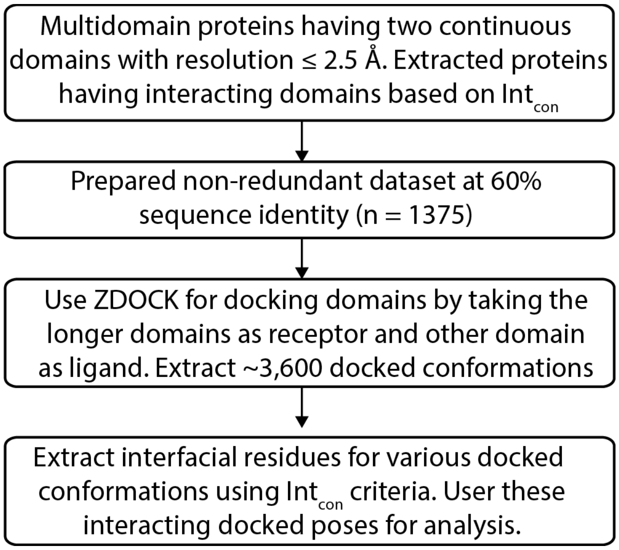

Multidomain proteins having two continuous domains with resolution ≤ 2.5 Å. Extracted proteins having interacting domains based on Int$_{con}$
Prepared non-redundant dataset at 60% sequence identity (n = 1375)
Use ZDOCK for docking domains by taking the longer domains as receptor and other domain as ligand. Extract ~3,600 docked conformations
Extract interfacial residues for various docked conformations using Int$_{con}$ criteria. User these interacting docked poses for analysis.

Figure 5.2 Overview of dataset construction. Flowchart showing steps in the construction of dataset and of docking protocol.

In order to dock domains, we used FFT method as implemented in ZDOCK program (Chen &Weng, 2002). Following the approach of ZDOCK, we took the longest domain as static receptor and the other domain as ligand, which will span all possible docked orientations. By default, ~2000 docked poses are given by ZDOCK. However, to span all possible docked poses from FFT we modified the criteria and obtained the maximum number of possible docked poses, which could be to a maximum of 3600 conformations. These poses were also scored based on shape complementarity, electrostatics, and statistical potential terms defined by ZDOCK.

5.2.2 Measures for assessment of domain-domain docking

The performance assessment of docking was based on scoring of docked poses with respect to native structure. For this, we used the criteria (section 2.2.2(d)) to define domain-domain interface and docked pose not satisfying this is not considered for any further assessment. Subsequently, we used following measures to assess interacting domain-domain docked poses:

a. Global RMSD: This is RMSD obtained after optimal superposition of each docked structure to the native structure as a single domain. This does not involve superposition of individual domains, which will provide no information.
b. Interface RMSD: The interfacial residues of interacting domains are extracted from the native structures as has been discussed before. For this, individual docked domains are first superposed on native structure and RMSD only for interfacial residues are calculated as interfacial RMSD.
c. Interface similarity (IS-score): The interacting domain-domain docked poses were aligned to native interface using iAlign.

Importantly, while calculating these parameters, the order of domains was kept intact in each docked complex as that was in its corresponding native structures. Since we are considering all possible solutions, the best docked poses were considered within top N docked ranked poses, where N varies from 1 to 20.

5.2.3 Construction of DDI template library

We used pre-compiled dataset of non-redundant interacting domain pairs, which share a structural similarity at the level of homology as defined by CATH (section 3.2). From this pair of interacting domain-domain interfaces, we selected any one as representative and performed all-against-all alignment of representative domain interfaces using iAlign. Thus, obtained IS-scores were used for clustering DDIs using program APCLUSTER (Frey &Dueck 2007; Bodenhofer et. al., 2011), which relies on similarity as a metric for clustering any set of objects. This results in 389 clusters of interfaces. We randomly selected any one member from a cluster to construct DDI template library. Thus, we have an interface template consisting of 389 domain-domain interfaces. The docked structures were aligned to the template interface library using iAlign and ranked using IS-score in order to identify near native docked poses.

5.2.4 Method to re-rank docked poses using interface alignment

In our proposed preliminary approach, we rely on interface similarity of docked poses to a representative interface template library for identifying native like interfaces. This approach involved performing interface alignment of all docked poses with 389 template library interface followed by re-ranking of docked poses using IS-score obtained from iAlign. The detailed procedure is shown in Figure 5.3. Briefly, we find the best structural matching interface for each docked pose by aligning it with all interfaces in template library followed by ranking interfaces using IS-score, which is normalized by length of docked pose interface. Thus, obtained list of the best IS-score for each docked pose is again ranked by IS-score to identify the docked conformation having the highest IS-score, which is considered as the near native structure from docking.

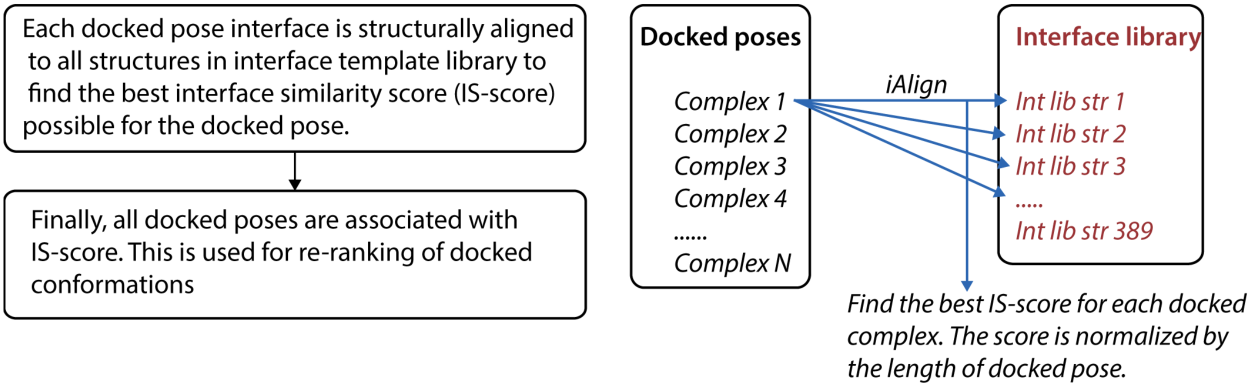

Each docked pose interface is structurally aligned to all structures in interface template library to find the best interface similarity score (IS-score) possible for the docked pose.
Finally, all docked poses are associated with IS-score. This is used for re-ranking of docked conformations
Docked poses
Complex 1
Complex 2
Complex 3
Complex 4
......
Complex N
iAlign
Interface library
Int lib str 1
Int lib str 2
Int lib str 3
.....
Int lib str 389
Find the best IS-score for each docked complex. The score is normalized by the length of docked pose.

Figure 5.3 Protocol describing re-ranking of docked complexes using interface constraints. The figure shows the schematics used for detecting native like interfaces by aligning docked interfaces to template library.

This procedure of interface structural alignment is found to be extremely computationally time consuming. To alleviate this problem, we examined docked poses from various structures and observed that many docked poses are not physically feasible because of constraints of linker region between domains. Unlike protein-protein docking where two structures are independent entity, here two domain structures being docked are physically linked as they are from same protein structure, thus many docking solutions are not possible. In order to restrict the docking solution space obtained by ZDOCK, we filtered docked poses based on: a. distance between last Cα and the first Cα of two connecting consecutive domains, and b. there will be limited surface area accessible for interactions between two connected globular objects. These are discussed in detail below.

5.2.5 Using C-alpha distance criteria to limit the number of docked poses

As mentioned before the structural alignment is time consuming task, we used distance cut-off between the last Cα of the first domain and the first Cα of the following connected (next) domain to filter docked solutions. The criteria used are as follows:

1. We imposed a distance cut-off of 6Å between last Cα of 1st domain and first Cα of 2nd domain for protein without linker region.
2. For proteins with inter-domain linker, a cut-off of 4(n+1) Å was applied; where n is the number of residues in a linker is used. The distance 4Å is used because mean distance between 2 Cα is 3.8Å. We relaxed this distance cut-off with an assumption that in subsequent modeling of full-length multidomain protein one can connect them and general force-field will be able to get to desired optimal geometry.

5.2.6 Using domains as interacting spheres to limit docking solution space

Another criterion we used to restrict docked solution space is by assuming interaction between connected domains as two linked rigid spheres of same size connected by a non-extendable linker. With connecting sphere assumption by a short inter-domain linker, the possible interacting space between spheres becomes a limiting solution (Figure 5.4). This is assuming that domain interaction site (presume it as interaction point) is within quarter of sphere surface (semi-hemisphere) with the domain linker site on the sphere. With these assumptions, it can be found that maximum Euclidean distance between interface interaction site and the point of linker region will be less than $\sqrt{2}$R, where R is sphere radius. Here, we have also considered interaction between identical spheres (Figure 5.4). In order to follow this geometrical sphere approximation of proteins, we used a well-known criterion to identify globular proteins, which involves relating number of residues to radius of gyration given by:

$R_g(P)=2.2x(N)^{0.38}$, here Rg (P) is predicted radius of gyration and N is the length of protein.

We assumed a domain to be a spherical object if, the calculated radius of gyration (R_g) of a domain is less than $R_g(P)$. Then, the point of interaction on sphere is represented by the centroid of interfacial residues. Thus, R_g is the radius of domain (with sphere approximation) and centroid of interfacial residues is the point of interaction (on the sphere). Thus, we followed following approach to find feasible space:

a. Each domain is assumed to be spherical, if $R_g<R_g(P)$, where R_g is calculated for each domain.
b. For each docked pose, we calculated interface centroids on both domains. Next, we calculated Euclidean distance between the C-terminal (Cα) of the first domain and centroid of interface on this domain. The same is repeated in the other domain with distance being calculated between N-terminal of this domain and its interface centroid.
c. We used relaxed criteria and defined any docked pose as a feasible solution, if the distance between centroid and connecting point is less than equal to $\sqrt{2}R_{g.}$

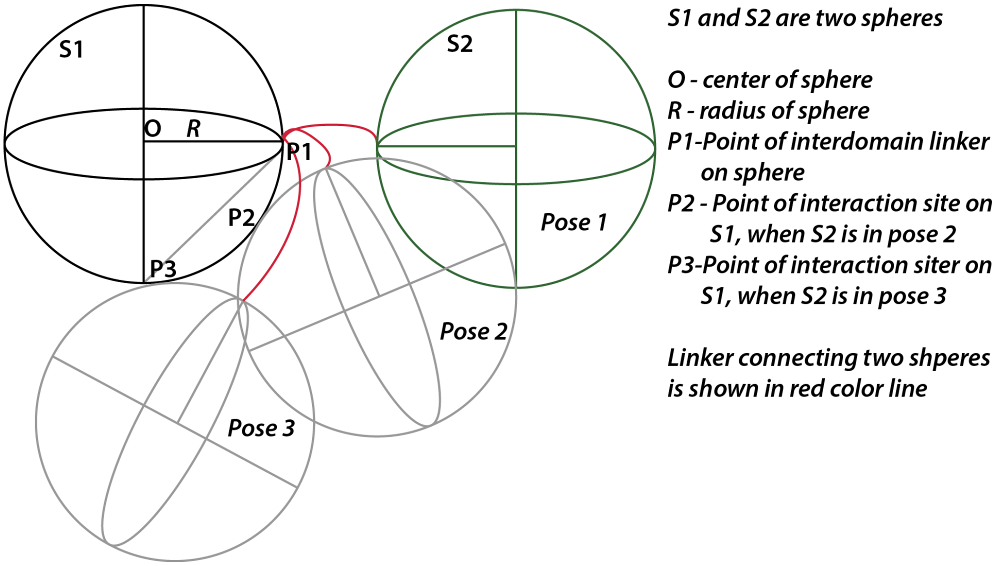

S1
S2
O
R
P1
P2
P3
Pose 1
Pose 2
Pose 3
S1 and S2 are two spheres
O - center of sphere
R - radius of sphere
P1-Point of interdomain linker on sphere
P2 - Point of interaction site on S1, when S2 is in pose 2
P3-Point of interaction siter on S1, when S2 is in pose 3
Linker connecting two shperes is shown in red color line

Figure 5.4 Illustration of sphere approximation to limit docking solution space. Figure shows if two interacting spheres are allowed to interact on surfaces lying within a quarter sphere from a point (linker site) on protein sphere, then it is feasible to define the maximum Euclidean distance between linker site (P1) and possible points of interaction (such as P2, and P3).

Even though this is rudimentary approach, with this gross sphere approximation we could reduce the docked solution space and decrease the computing time to find reasonable docked poses.

5.3 Results

5.3.1 Benchmark results

In order to reliably model intra-domain interface, as a preliminary approach we have applied the similarity of domain-domain interaction interface to filter the docked poses to identify the pose close to native structure. For the initial assessment, we have docked domains extracted from experimentally determined two domain protein structures followed by identifying the best-docked pose using global RMSD as a metric. We have generated a maximum of 3600 docked poses using ZDOCK. This program utilizes a combination of shape complementarity, electrostatics, and statistical potential terms to rank docked solutions. Since we are using native structure for docking, it is worthy to note that ZDOCK will provide reasonably ranked solutions with rank one as the most likely pose close to native structure. However, the same would be challenge to obtain for modeled domain structures.

For calculating global RMSD, the native structure was aligned with all ZDOCK docked conformations and these were re-ranked based on RMSD. The distribution of global RMSD is shown in Figure 5.5A. It is apparent from the figure that for most (~94%) of the benchmark proteins, global RMSD is < 2 Å. Using global RMSD as a metric to rank docked poses, we observed that the best docked pose in 67% of cases is ranked 1 in ZDOCK.

5.3.2 Using interface similarity score and interface RMSD to rank the docked poses

Next, we analyzed the distributions of interface similarity score (IS-score) and interface RMSD of native domain-domain interface to the interface between domains obtained from docking. This will allow us to examine whether docking is able to generate native-like interfaces. The IS-score was computed by performing structural alignment of docked interfaces with native domain-domain interface. As is seen in Figure 5.5B, most (85%) docked poses have IS-score ≥ 0.7 suggesting that docking is able to recapitulate native-like interfaces. The re-ranking of docking solutions using IS-score, showed that ~67% of proteins with the best IS-score is ranked one by ZDOCK.

Further, we employed an interfacial RMSD as a metric to rank docked poses according to the most similar interface with the native interface. Since our dataset has multidomain proteins having domain separated by linker region (140 proteins), we assessed whether proteins having linker regions have any difference in docking with respect to proteins without linker region. The distribution of interfacial RMSD is shown in Figure 5.5C, which shows multidomain proteins with or without linker regions results in interfacial RMSD for most domain-domain docked solutions within 2 Å of native interface. Thus, suggesting docking results in poses, which have native like interfaces. The re-ranking of docked poses using interfacial RMSD results in similar results of rank 1 from ZDOCK ranked positions in 67% of cases as has been observed in previous metrics.

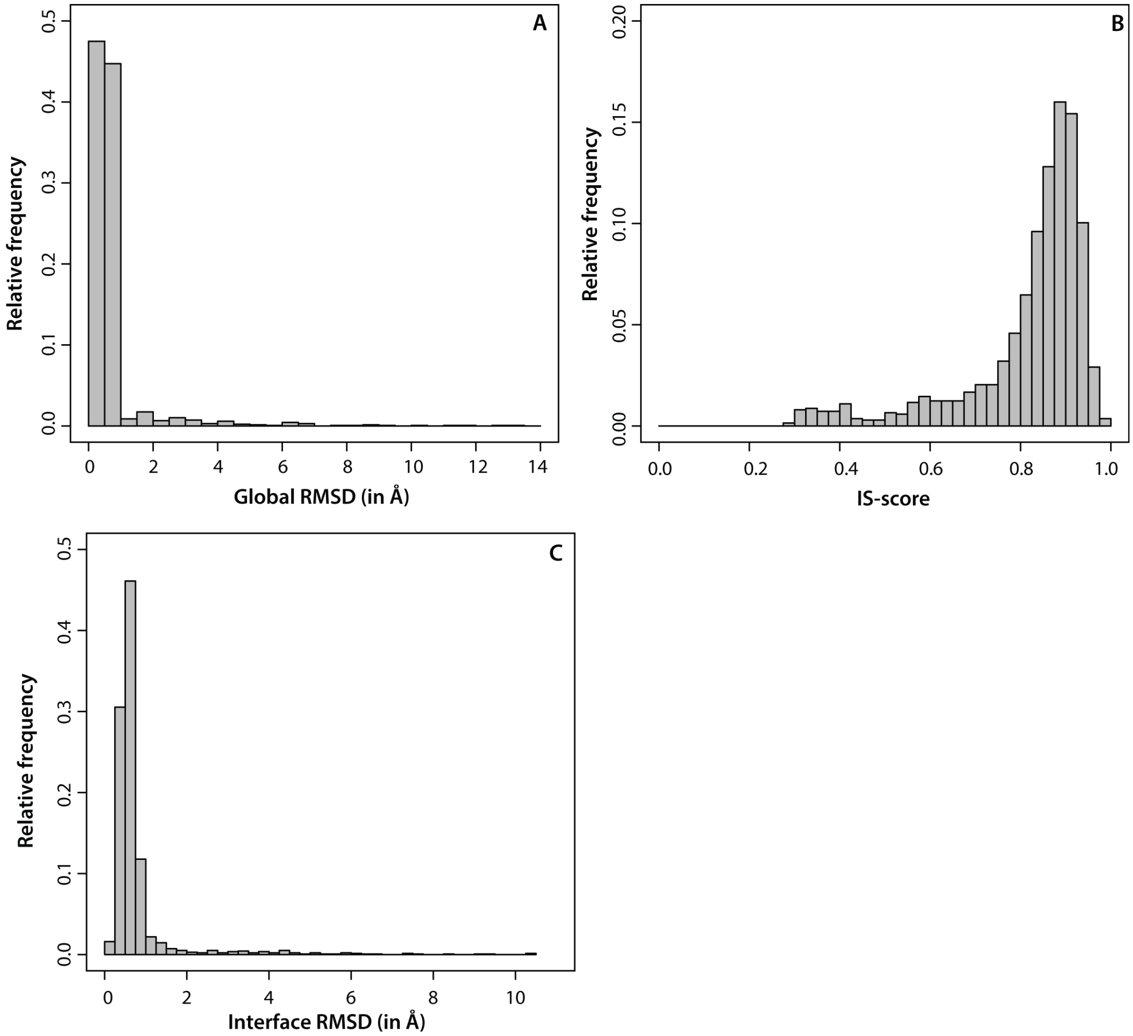


Figure 5.5 Histogram of the best scores between native and docked complexes. The best solution from docking is identified using global RMSD, IS-score, and interface RMSD, which are obtained by optimally superposing all docked complexes to native structures. Subsequently, the best score of metric used for evaluation is plotted as histogram. These are shown for the best A) Global RMSD, B) IS-score and C) interface RMSD among all docked complexes.

5.3.3 Finding docking solution in interface template library

Having shown that docking results in native-like interface, the challenge is to identify such interfaces in the absence of native structure. For this, we proposed a preliminary solution that is to align all docked poses on template library of intra-chain domain-domain interfaces to find the native like interface. As described in methods section, we obtained 389 template library of interfaces using clustering based on IS-score. Each docked conformation subsequent to alignment with all template interface, we identify the best possible template interface as assessed by IS-score and assign this score to the docked conformation. Essentially, we are searching for the best template (IS-score) for each docked pose and all docked poses will be re-ranked based on this template-pose IS-score (Figure 5.3). This will possibly identify interfaces, which will resemble and match to interface in native structure.

Subsequent to alignment with template library interface and re-ranking the docked solutions based on the template-pose IS-score, we considered the top 5/10/15/20 docked complexes and find the best ranked position from the ranked list based on IS-score computed between interfaces of native structure and docked complexes. Thus, providing the best possible rank of complex identified in our approach in the ‘true’ ranked list of docked complexes based on interface similarities. This will assist in evaluating the performance of our approach for its ability to identify and rank near-native pose. The Figure 5.6A shows the cumulative fraction of proteins for which a near native pose can be identified within given rank. For instance, our approach for a query protein identifies a pose within top 5, which has the best rank in native-docked pose as 200; however, if we consider top 20 poses in our approach the best rank can either remain same or improve to a lower rank. This is evident in Figure 5.6A, as it can be seen that the best ranked docked complex is greatly improved from taking top 5 to top 20 poses. In this case, we have not considered the quality of docked solution in terms of IS-score. It is possible that the best docked solution may have insignificant similarity to native. Since our approach is time consuming, we applied two additional filters to improve the speed as well as reduce number of non-feasible solutions.

5.3.4 Using domain contact points as a measure to reduce docking space

As many docked solution are not feasible because the domain are tethered to each other, we applied the filter criteria of distance between the last Cα of first domain and first Cα of second domain arranged in sequential order (see methods). The proteins having no inter-domain linker between domains, a distance between last Cα of domain1 and first Cα of domain2 should be less than 6Å. Whereas, a distance cut-off of 4 * (n+1) Å was used for proteins having two domains separated by a linker region, where n is the length of the linker. The docked complexes, which followed this cut-off, were retained and others were excluded from the analysis. This eliminated 9 proteins from the dataset, for which we could not find any complex, which satisfy the given criteria. Following the approach described before, we considered top 5/10/15/20 docked poses from our protocol to identify native-interfaces solutions. These were subjected to similar analysis as before and results are shown in Figure 5.6B. As can be seen in the results, the percentage of finding a near-native solution among top 5 best docked poses was increased from 44% to 70%.

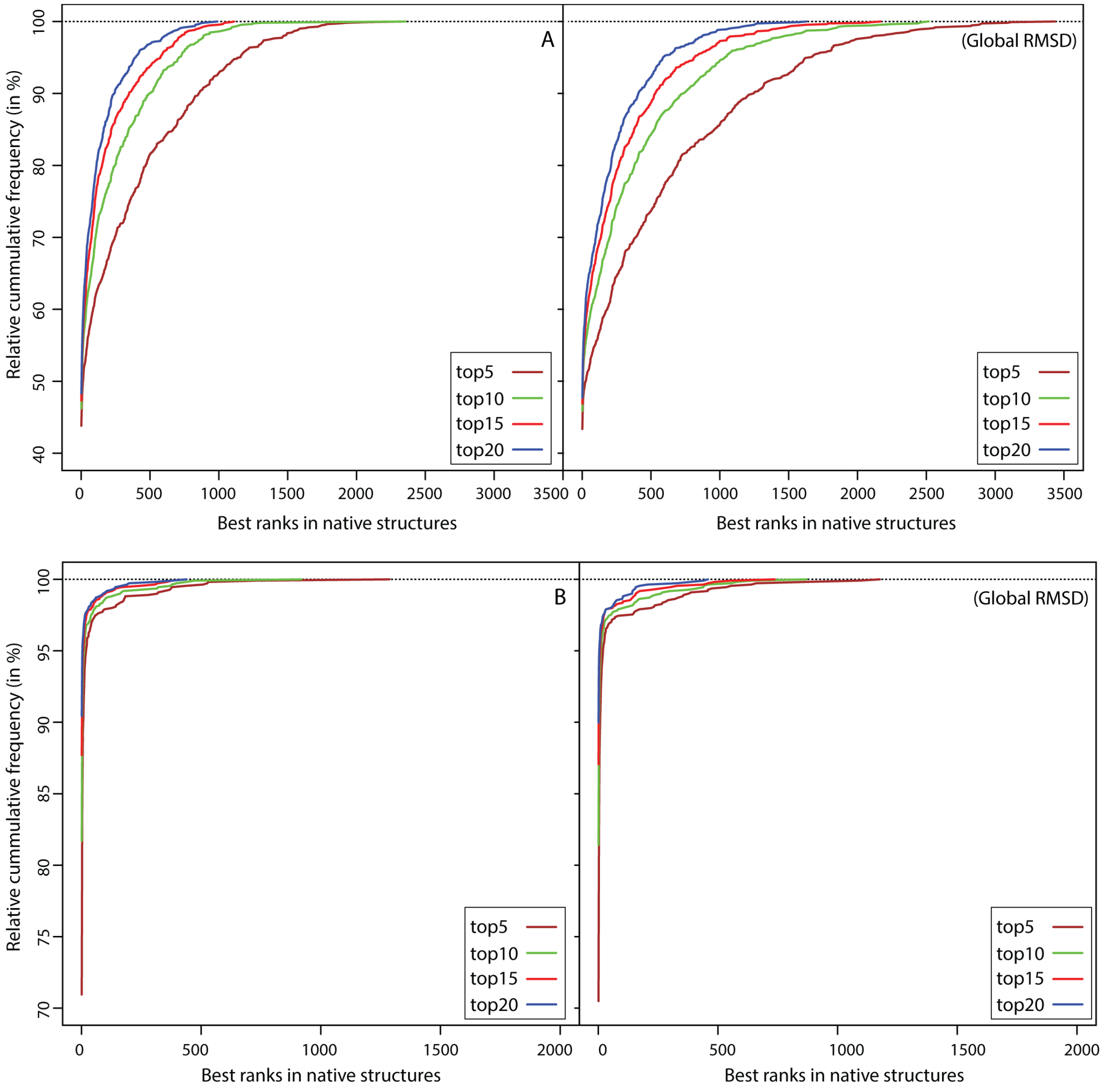


Figure 5.6 Enrichment analyses by reducing docking solution space. Figure showing the best ranks from top 5/10/15/20 of docked complexes (obtained by aligning to template library) plotted with respect to ranks of the same in IS-score based ranking of docked poses to native structure. The panel (A) is without any filter and (B) is using distance between ends of domains to filter solution of docked complexes.

5.3.5 Globularity of proteins as a measure to reduce docking solution space

The second filter criterion is to use protein globularity to eliminate non-likely solutions. In this respect, we used a protein globularity measure to filter the possible docked solutions. We assumed protein structural domains as spheres with their interfaces being represented as centroids and used Cα distances between domains as contact points. Applying this approach, we found only 223 proteins, which satisfy this globularity criterion for both the domains. From these 223 two-domain proteins, we filtered out the docked complexes and reduced the feasible docked complexes to a median of 670 from ~3000 complexes. Subsequently, we find the best docked structure in this smaller docking solution space in terms of IS-score. Out of 223 proteins, 203 proteins, approximately 92% of cases consisted of native-like interfaces. This implies that a simple sphere representation of domains could be used to reduce docked complexes.

5.4 Conclusion

In the present work, we utilized the knowledge of interface alignment in re-ranking of docking solutions and analyzed if this method can reliably identify a near-native docked pose. For this analysis, we constructed a benchmark dataset of continuous 2-domain proteins from CATH_v4.1.0 and each protein was subjected to a rigid body docking. In order to identify a correct docked conformation in a decoy of structures for each protein, we used measures such as interfacial RMSD (iRMSD), global RMSD (gRMSD) and interface similarity score (IS-score). Considering that structure with rank1 from ZDOCK is the most similar structure to the native, we found that in 67% of the cases, docked complexes have native interface. However, this would not be possible for docked complexes of modeled domains. Thus, we have proposed IS-score based ranking of docked complex by searching in template library of interfaces. Since, the comparison of interface of each docked complex with the templates could become a computationally extensive comparison given a large docking space; we employed two filters of distance between ends of domains and protein shape. The globularity measure and inter-domain distance cut-off filters not only resulted in reducing the number of docked complexes but also showed an enrichment of identifying a near-native docked structure. The percentage of finding proteins with the correct interface in top20 best docked poses increased to 90% in the interface template library. These results suggest that incorporating interface information in the docking studies can result in improving the docking predictions and can provide more accurate models for structure prediction. This is preliminary analysis into using interface similarity score, which can be improved by subsequently subjecting docked poses for refinement in modeling multidomain proteins.

All datasets used in Chapters 2-5 are deposited in github under the repository:

https://github.com/riviverma/thesis-md-proteins/

References

Acuner Ozbabacan, S. E., Engin, H. B., Gursoy, A., & Keskin, O. (2011). Transient protein-protein interactions. *Protein Engineering, Design and Selection*, 24(9), 635-648.

Adzhubei, A. A., & Sternberg, M. J. (1994). Conservation of polyproline II helices in homologous proteins: Implications for structure prediction by model building. *Protein Science : A Publication of the Protein Society*, *3*(12), 2395–2410.

Alberts, B. (1998). The cell as a collection of protein machines: Preparing the next generation of molecular biologists. *Cell*, *92*(3), 291–294.

Aloy, P., Ceulemans, H., Stark, A., & Russell, R. B. (2003). The Relationship Between Sequence and Interaction Divergence in Proteins. *Journal of Molecular Biology*, *332*(5), 989–998.

Aloy, P., & Russell, R. B. (2002). Interrogating protein interaction networks through structural biology. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(9), 5896–5901.

Andersen, G. R., Pedersen, L., Valente, L., Chatterjee, I., Kinzy, T. G., Kjeldgaard, M., & Nyborg, J. (2000). Structural Basis for Nucleotide Exchange and Competition with tRNA in the Yeast Elongation Factor Complex eEF1A:eEF1Bα. *Molecular Cell*, *6*(5), 1261–1266.

Apic, G., Gough, J., & Teichmann, S. A. (2001a). An insight into domain combinations. *Bioinformatics (Oxford, England)*, *17 Suppl 1*, S83-89.

Apic, G., Gough, J., & Teichmann, S. A. (2001b). Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *Journal of Molecular Biology*, 310(2), 311-325.

Apic, G, & Russell, R. (2010). Domain recombination: A workhorse for evolutionary innovation. *Science Signaling*, *3*(139), pe30–pe30.

Apic, G., Huber, W., & Teichmann, S. A. (2003). Multi-domain protein families and domain pairs: Comparison with known structures and a random model of domain recombination. *Journal of Structural and Functional Genomics*, *4*(2–3), 67–78.

Argos, P. (1988). An investigation of protein subunit and domain interfaces. *Protein Engineering, Design and Selection*, *2*(2), 101–113.

Arkin, M. R., & Wells, J. A. (2004). Small-molecule inhibitors of protein-protein interactions: Progressing towards the dream. *Nature Reviews. Drug Discovery*, *3*(4), 301–317.

Arkin, M. R., & Whitty, A. (2009). The road less traveled: Modulating signal transduction enzymes by inhibiting their protein–protein interactions. *Current Opinion in Chemical Biology*, *13*(3), 284–290.

Arndt, J. W., Gu, J., Jaroszewski, L., Schwarzenbacher, R., Hanson, M. A., Lebeda, F. J., & Stevens, R. C. (2005). The structure of the neurotoxin-associated protein HA33/A from Clostridium botulinum suggests a reoccurring beta-trefoil fold in the progenitor toxin complex. *Journal of Molecular Biology*, *346*(4), 1083–1093.

Arviv, O., & Levy, Y. (2012). Folding of multidomain proteins: Biophysical consequences of tethering even in apparently independent folding. *Proteins*, *80*(12), 2780–2798.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene Ontology: Tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29.

Auton, M., & Bolen, D. W. (2005). Predicting the energetics of osmolyte-induced protein folding/unfolding. *Proceedings of the National Academy of Sciences*, *102*(42), 15065–15068.

Aytuna, A. S., Gursoy, A., & Keskin, O. (2005). Prediction of protein–protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics*, *21*(12), 2850–2855.

Bae, K., Mallick, B. K., & Elsik, C. G. (2005a). Prediction of protein interdomain linker regions by a hidden Markov model. *Bioinformatics (Oxford, England)*, *21*(10), 2264–2270.

Bahadur, R. P., & Zacharias, M. (2008). The interface of protein-protein complexes: Analysis of contacts and prediction of interactions. *Cellular and Molecular Life Sciences: CMLS*, *65*(7–8), 1059–1072.

Bahadur, Ranjit Prasad, Chakrabarti, P., Rodier, F., & Janin, J. (2003). Dissecting subunit interfaces in homodimeric proteins. *Proteins*, *53*(3), 708–719.

Balaji, S., & Srinivasan, N. (2007). Comparison of sequence-based and structure-based phylogenetic trees of homologous proteins: Inferences on protein evolution. *Journal of Biosciences*, *32*(1), 83–96.

Barabási, A.-L., & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, *286*(5439), 509–512.

Barrow, J. C., Stauffer, S. R., Rittle, K. E., Ngo, P. L., Yang, Z., Selnick, H. G., Graham, S. L., Munshi, S., McGaughey, G. B., Holloway, M. K., Simon, A. J., Price, E. A., Sankaranarayanan, S., Colussi, D., Tugusheva, K., Lai, M.-T., Espeseth, A. S., Xu, M., Huang, Q., … Vacca, J. P. (2008). Discovery and X-ray crystallographic analysis of a spiropiperidine iminohydantoin inhibitor of beta-secretase. *Journal of Medicinal Chemistry*, *51*(20), 6259–6262.

Bashton, M., & Chothia, C. (2002). The geometry of domain combination in proteins. *Journal of Molecular Biology*, *315*(4), 927–939.

Bashton, M., & Chothia, C. (2007). The Generation of New Protein Functions by the Combination of Domains. *Structure*, *15*(1), 85–99.

Basu, M. K., Carmel, L., Rogozin, I. B., & Koonin, E. V. (2008). Evolution of protein domain promiscuity in eukaryotes. *Genome Research*, *18*(3), 449–461.

Basu, M. K., Poliakov, E., & Rogozin, I. B. (2009). Domain mobility in proteins: Functional and evolutionary implications. *Briefings in Bioinformatics*, *10*(3), 205–216.

Batey, S., Randles, L. G., Steward, A., & Clarke, J. (2005). Cooperative Folding in a Multi-domain Protein. *Journal of Molecular Biology*, *349*(5), 1045–1059.

Batey, S., Scott, K. A., & Clarke, J. (2006). Complex Folding Kinetics of a Multidomain Protein. *Biophysical Journal*, *90*(6), 2120–2130.

Bell, C. E., & Eisenberg, D. (1996). Crystal Structure of Diphtheria Toxin Bound to Nicotinamide Adenine Dinucleotide. *Biochemistry*, *35*(4), 1137–1149.

Ben-Naim, A. (2006). On the driving forces for protein-protein association. *The Journal of Chemical Physics*, *125*(2), 024901.

Bennett, M. J., & Eisenberg, D. (1994). Refined structure of monomeric diphtheria toxin at 2.3 A resolution. *Protein Science : A Publication of the Protein Society*, *3*(9), 1464–1475.

Bennett, Melanie J., Schlunegger, M. P., & Eisenberg, D. (1995). 3D domain swapping: A mechanism for oligomer assembly. *Protein Science*, *4*(12), 2455–2468.

Bennett, W. S., Huber, R., & Engel, Jür. (1984). Structural and Functional Aspects of Domain Motions in Proteins. *Critical Reviews in Biochemistry*, *15*(4), 291–384.

Ben-Zeev, E., Kowalsman, N., Ben-Shimon, A., Segal, D., Atarot, T., Noivirt, O., Shay, T., & Eisenstein, M. (2005). Docking to single-domain and multiple-domain proteins: Old and new challenges. *Proteins: Structure, Function, and Bioinformatics*, *60*(2), 195–201.

Berggård, T., Linse, S., & James, P. (2007). Methods for the detection and analysis of protein–protein interactions. *PROTEOMICS*, *7*(16), 2833–2842.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977). The Protein Data Bank: A computer-based archival file for macromolecular structures. *Journal of Molecular Biology*, *112*(3), 535–542.

Betel, D., Breitkreuz, K. E., Isserlin, R., Dewar-Darch, D., Tyers, M., & Hogue, C. W. V. (2007). Structure-Templated Predictions of Novel Protein Interactions from Sequence Information. *PLOS Computational Biology*, *3*(9), e182.

Betts, M. J., & Sternberg, M. J. E. (1999). An analysis of conformational changes on protein–protein association: Implications for predictive docking. *Protein Engineering, Design and Selection*, *12*(4), 271–283.

Bhaskara, R. M., de Brevern, A. G., & Srinivasan, N. (2013). Understanding the role of domain-domain linkers in the spatial orientation of domains in multi-domain proteins. *Journal of Biomolecular Structure & Dynamics*, *31*(12), 1467–1480.

Bhaskara, R. M., Padhi, A., & Srinivasan, N. (2014). Accurate prediction of interfacial residues in two-domain proteins using evolutionary information: Implications for three-dimensional modeling. *Proteins: Structure, Function, and Bioinformatics*, *82*(7), 1219–1234.

Bhaskara, R. M., & Srinivasan, N. (2011). Stability of domain structures in multi-domain proteins. *Scientific Reports*, *1*, 40.

Bhattacharyya, R., Pal, D., & Chakrabarti, P. (2004). Disulfide bonds, their stereospecific environment and conservation in protein structures. *Protein Engineering, Design and Selection*, *17*(11), 795–808.

Björkholm, P., & Sonnhammer, E. L. L. (2009). Comparative analysis and unification of domain–domain interaction networks. *Bioinformatics*, *25*(22), 3020–3025.

Bjorklund, A. K., Ekman, D., Light, S., Frey-Skott, J., & Elofsson, A. (2005). Domain rearrangements in protein evolution. *Journal of Molecular Biology*, 353(4), 911-923.

Blake, C. C., Mair, G. A., North, A. C., Phillips, D. C., & Sarma, V. R. (1967). On the conformation of the hen egg-white lysozyme molecule. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, *167*(1009), 365–377.

Bodenhofer, U., Kothmeier, A., & Hochreiter, S. (2011). APCluster: An R package for affinity propagation clustering. *Bioinformatics (Oxford, England)*, *27*(17), 2463–2464.

Bogan, A. A., & Thorn, K. S. (1998). Anatomy of hot spots in protein interfaces. *Journal of Molecular Biology*, *280*(1), 1–9.

Bolser, D. M., & Park, J. H. (2003). Biological Network Evolution Hypothesis Applied to Protein Structural Interactome. *Genomics & Informatics*, *1*(1), 7–19.

Bonetta, L. (2010). Protein-protein interactions: Interactome under construction. *Nature*, *468*(7325), 851–854.

Bonvin, A. M. (2006). Flexible protein–protein docking. *Current Opinion in Structural Biology*, *16*(2), 194–200.

Bornberg-Bauer, E., Beaussart, F., Kummerfeld, S. K., Teichmann, S. A., & Weiner, J., 3rd. (2005). The evolution of domain arrangements in proteins and interaction networks. *Cellular and Molecular Life Sciences CMLS*, 62(4), 435-445.

Bourne, P. E., & Weissig, H. (2003). *Structural Bioinformatics* (Vol. 44).

Branden, C., & Tooze, J. (1999). *Introduction to Protein Structure* (2nd ed.). Garland Publishing.

Brender, J. R., & Zhang, Y. (2015). Predicting the Effect of Mutations on Protein-Protein Binding Interactions through Structure-Based Interface Profiles. *PLOS Computational Biology*, *11*(10), e1004494.

Brenner, S. E., Koehl, P., & Levitt, M. (2000). The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Research*, *28*(1), 254–256.

Briggs, S., & Smithgall, T. (1999). SH2-kinase linker mutations release Hck tyrosine kinase and transforming activities in Rat-2 fibroblasts. *The Journal of Biological Chemistry*, *274*(37), 26579–26583.

Brocchieri, L., & Karlin, S. (2005). Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Research*, *33*(10), 3390–3400.

Bronowska, A. K. (2011). Thermodynamics of Ligand-Protein Interactions: Implications for Molecular Design. *Thermodynamics - Interaction Studies - Solids, Liquids and Gases*.

Brylinski, M., & Skolnick, J. (2007). What is the relationship between the global structures of apo and holo proteins? *Proteins: Structure, Function, and Bioinformatics*, *70*(2), 363–377.

Buljan, M., Frankish, A., & Bateman, A. (2010). Quantifying the mechanisms of domain gain in animal proteins. *Genome Biology*, *11*(7), R74.

Burley, S. K., & Bonanno, J. B. (2002). Structuring the universe of proteins. *Annual Review of Genomics and Human Genetics*, 3, 243-262.

Burra, P. V., Zhang, Y., Godzik, A., & Stec, B. (2009). Global distribution of conformational states derived from redundant models in the PDB points to non-uniqueness of the protein structure. *Proceedings of the National Academy of Sciences*, *106*(26), 10505–10510.

Caffrey, D. R., Somaroo, S., Hughes, J. D., Mintseris, J., & Huang, E. S. (2004a). Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Science: A Publication of the Protein Society*, *13*(1), 190–202.

Campbell, I. D., & Baron, M. (1991). The structure and function of protein modules. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *332*(1263), 165–170.

Carugo, O., & Argos, P. (1997). Protein-protein crystal-packing contacts. *Protein Science : A Publication of the Protein Society*, *6*(10), 2261–2263.

Carugo, O. (2006). *Rapid Methods for Comparing Protein Structures and Scanning Structure Databases* (Vol. 1). Bentham Science Publishers.

Cazals, F., Proust, F., Bahadur, R. P., & Janin, J. (2006). Revisiting the Voronoi description of protein-protein interfaces. *Protein Science*, 15(9), 2082-2092.

Chacón, P., & Wriggers, W. (2002). Multi-resolution contour-based fitting of macromolecular structures. *Journal of Molecular Biology*, *317*(3), 375–384.

Chakrabarti, P., & Janin, J. (2002). Dissecting protein-protein recognition sites. *Proteins*, *47*(3), 334–343.

Chakravarty, D., Janin, J., Robert, C. H., & Chakrabarti, P. (2015). Changes in protein structure at the interface accompanying complex formation. *IUCrJ*, *2*(6), 643–652.

Chakravarty, D., Guharoy, M., Robert, C. H., Chakrabarti, P., & Janin, J. (2013). Reassessing buried surface areas in protein–protein complexes. *Protein Science : A Publication of the Protein Society*, *22*(10), 1453–1457.

Chandonia, J.-M., Hon, G., Walker, N. S., Lo Conte, L., Koehl, P., Levitt, M., & Brenner, S. E. (2004). The ASTRAL Compendium in 2004. *Nucleic Acids Research*, *32*(Database issue), D189-192.

Cheetham, G. M. T., & Steitz, and T. A. (1999). Structure of a Transcribing T7 RNA Polymerase Initiation Complex. *Science*, *286*(5448), 2305–2309.

Chen, R., & Weng, Z. (2002). Docking unbound proteins using shape complementarity, desolvation, and electrostatics. *Proteins: Structure, Function, and Bioinformatics*, *47*(3), 281–294.

Cheng, A. C., Coleman, R. G., Smyth, K. T., Cao, Q., Soulard, P., Caffrey, D. R., Salzberg, A. C., & Huang, E. S. (2007). Structure-based maximal affinity model predicts small-molecule druggability. *Nature Biotechnology*, *25*(1), 71–75.

Cheng, T. M., Blundell, T. L., & Fernandez-Recio, J. (2008). Structural assembly of two-domain proteins by rigid-body docking. *BMC Bioinformatics*, *9*(1), 441.

Choi, I.-G., & Kim, S.-H. (2006). Evolution of protein structural classes and protein sequence families. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(38), 14056–14061.

Chothia, C., & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, *5*(4), 823–826.

Chothia, C. (1992). One thousand families for the molecular biologist. *Nature*, *357*(6379), 543–544.

Chothia, C., & Gough, J. (2009). Genomic and structural aspects of protein evolution. *The Biochemical Journal*, *419*(1), 15–28.

Chothia, C., & Janin, J. (1975). Principles of protein–protein recognition. *Nature*, *256*(5520), 705–708.

Chou, P. Y., & Fasman, G. D. (1974). Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins. *Biochemistry*, *13*(2), 211–222.

Chung, S. Y., & Subbiah, S. (1996). A structural explanation for the twilight zone of protein sequence homology. *Structure*, *4*(10), 1123–1127.

Clore, G. M., & Gronenborn, A. M. (1998). NMR structure determination of proteins and protein complexes larger than 20 kDa. *Current Opinion in Chemical Biology*, *2*(5), 564–570.

Cohen, F. E., & Prusiner, S. B. (1998). Pathologic Conformations of Prion Proteins. *Annual Review of Biochemistry*, *67*(1), 793–819.

Connolly, M. L. (1986). Shape complementarity at the hemoglobin alpha 1 beta 1 subunit interface. *Biopolymers*, *25*(7), 1229–1247.

Cooper, A. (1976). Thermodynamic fluctuations in protein molecules. *Proceedings of the National Academy of Sciences*, *73*(8), 2740–2741.

Csaba, G., Birzele, F., & Zimmer, R. (2009). Systematic comparison of SCOP and CATH: A new gold standard for protein structure analysis. *BMC Structural Biology*, *9*(1), 23.

Cunningham, B. A., Gottlieb, P. D., Pflumm, M. N., & Edelman, G. M. (1971). Immunoglobulin Structure: Diversity, Gene Duplication, and Domains**Supported by grant GB 8371 from the National Science Foundation and by grants AM 04256 and AI 09273 from the National Institutes of Health. In B. Amos (Ed.), *Progress in Immunology* (pp. 3–24). Academic Press.

Das, S., Dawson, N. L., & Orengo, C. A. (2015). Diversity in protein domain superfamilies. *Current Opinion in Genetics & Development*, *35*, 40–49.

Das, S., & Smith, T. F. (2000). Identifying nature's protein lego set. In *Advances in Protein Chemistry* (Vol. 54, pp. 159–183). Academic Press.

Dasgupta, S., Iyer, G. H., Bryant, S. H., Lawrence, C. E., & Bell, J. A. (1997). Extent and nature of contacts between protein molecules in crystal lattices and between subunits of protein oligomers. *Proteins*, *28*(4), 494–514.

De Las Rivas, J., & Fontanillo, C. (2010). Protein–Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks. *PLoS Computational Biology*, *6*(6).

De Las Rivas, J., & Fontanillo, C. (2012). Protein–protein interaction networks: Unraveling the wiring of molecular machines within the cell. *Briefings in Functional Genomics*, *11*(6), 489–496.

de Vries, S. J., van Dijk, M., & Bonvin, A. M. J. J. (2010). The HADDOCK web server for data-driven biomolecular docking. *Nature Protocols*, *5*(5), 883–897.

Decanniere, K., Transue, T. R., Desmyter, A., Maes, D., Muyldermans, S., & Wyns, L. (2001). Degenerate interfaces in antigen-antibody complexes11Edited by J. Thornton. *Journal of Molecular Biology*, *313*(3), 473–478.

Deremble, C., & Lavery, R. (2005). Macromolecular recognition. *Current Opinion in Structural Biology*, *15*(2), 171–175.

Díez, P., Dasilva, N., González-González, M., Matarraz, S., Casado-Vela, J., Orfao, A., & Fuentes, M. (2012). Data Analysis Strategies for Protein Microarrays. *Microarrays*, *1*(2), 64–83.

Dominguez, C., Boelens, R., & Bonvin, A. M. J. J. (2003). HADDOCK: A Protein−Protein Docking Approach Based on Biochemical or Biophysical Information. *Journal of the American Chemical Society*, *125*(7), 1731–1737.

Doolittle, R. F. (1995). The multiplicity of domains in proteins. *Annual Review of Biochemistry*, *64*, 287–314.

Duan, Y., Reddy, B. V. B., & Kaznessis, Y. N. (2005). Physicochemical and residue conservation calculations to improve the ranking of protein–protein docking solutions. *Protein Science : A Publication of the Protein Society*, *14*(2), 316–328.

Duff, A. P., Cohen, A. E., Ellis, P. J., Hilmer, K., Langley, D. B., Dooley, D. M., Freeman, H. C., & Guss, J. M. (2006). The 1.23 Å structure of *Pichia pastoris* lysyl oxidase reveals a lysine–lysine cross-link. *Acta Crystallographica Section D Biological Crystallography*, *62*(9), 1073–1084.

Edwards, H., & Deane, C. M. (2015). Structural Bridges through Fold Space. *PLOS Computational Biology*, *11*(9), e1004466.

Eisenberg, D., Marcotte, E. M., Xenarios, I., & Yeates, T. O. (2000). Protein function in the post-genomic era. *Nature*, *405*(6788), 823–826.

Eisenstein, M., & Katchalski-Katzir, E. (2004). On proteins, grids, correlations, and docking. *Comptes Rendus Biologies*, *327*(5), 409–420.

Ekman, D., Björklund, A. K., Frey-Skött, J., & Elofsson, A. (2005). Multi-domain proteins in the three kingdoms of life: Orphan domains and other unassigned regions. *Journal of Molecular Biology*, *348*(1), 231–243.

Elez, K., Bonvin, A. M. J. J., & Vangone, A. (2018). Distinguishing crystallographic from biological interfaces in protein complexes: Role of intermolecular contacts and energetics for classification. *BMC Bioinformatics*, *19*(Suppl 15), 438.

Enright, A. J., Iliopoulos, I., Kyrpides, N. C., & Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature*, *402*(6757), 86–90.

Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M., Pieper, U., & Sali, A. (2006). Comparative Protein Structure Modeling Using Modeller. *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]*, *0 5*, Unit-5.6.

Eyster, K. (1998). Introduction to signal transduction: A primer for untangling the web of intracellular messengers. *Biochemical Pharmacology*, *55*(12), 1927–1938.

Fabiola, F., & Chapman, M. S. (2005). Fitting of high-resolution structures into electron microscopy reconstruction images. *Structure (London, England: 1993)*, *13*(3), 389–400.

Fersht, A. R. (1987). The hydrogen bond in molecular recognition. *Trends in Biochemical Sciences*, *12*, 301–304.

Finkelstein, A. V., & Ptitsyn, O. B. (1987). Why do globular proteins fit the limited set of folding patterns? *Progress in Biophysics and Molecular Biology*, *50*(3), 171–190.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., . . . Punta, M. (2014). Pfam: the protein families database. *Nucleic Acids Research*, 42(Database issue), D222-230.

Fischer, T., Paczkowski, M., Zettel, M., & Tsai, J. (2007). *A Guide to Protein Interaction Databases* (pp. 753–799).

Fiser, A. (2004). Protein structure modeling in the proteomics era. *Expert Review of Proteomics*, *1*(1), 97–110.

Fiser, A. (2010). Template-Based Protein Structure Modeling. *Methods in Molecular Biology (Clifton, N.J.)*, *673*, 73–94.

Flaugh, S. L., Kosinski-Collins, M. S., & King, J. (2005). Contributions of hydrophobic domain interface interactions to the folding and stability of human gammaD-crystallin. *Protein Science: A Publication of the Protein Society*, *14*(3), 569–581.

Flores, T. P., Orengo, C. A., Moss, D. S., & Thornton, J. M. (1993). Comparison of conformational characteristics in structurally similar protein pairs. *Protein Science : A Publication of the Protein Society*, *2*(11), 1811–1826.

Fong, J. H., Geer, L. Y., Panchenko, A. R., & Bryant, S. H. (2007). Modeling the Evolution of Protein Domain Architectures Using Maximum Parsimony. *Journal of Molecular Biology*, *366*(1), 307–315.

Forslund, K., Henricson, A., Hollich, V., & Sonnhammer, E. L. L. (2008). Domain Tree-Based Analysis of Protein Architecture Evolution. *Molecular Biology and Evolution*, *25*(2), 254–264.

Fox, N. K., Brenner, S. E., & Chandonia, J.-M. (2014). SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Research*, *42*(Database issue), D304-309.

Frey, B. J., & Dueck, D. (2007). Clustering by Passing Messages Between Data Points. *Science*, *315*(5814), 972–976.

Furukawa, Y., Ishimori, K., & Morishima, I. (2002). Oxidation-state-dependent protein docking between cytochrome c and cytochrome b5: High-pressure laser flash photolysis study. *Biochemistry*, *41*(31), 9824–9832.

Fuxreiter, M. (2018). Towards a Stochastic Paradigm: From Fuzzy Ensembles to Cellular Functions. *Molecules*, 23(11).

Reeves, G.A., Dallman, T.J., Redfern, O.C., Akpor, A., & Orengo, C.A. (2006). Structural diversity of domain superfamilies in the CATH database. *Journal of Molecular Biology*, *360*(3), 725–741.

Gabb, H. A., Jackson, R. M., & Sternberg, M. J. (1997). Modelling protein docking using shape complementarity, electrostatics and biochemical information. *Journal of Molecular Biology*, *272*(1), 106–120.

Gaines, J. C., Acebes, S., Virrueta, A., Butler, M., Regan, L., & O'Hern, C. S. (2018). Comparing side chain packing in soluble proteins, protein-protein interfaces, and transmembrane proteins. *Proteins: Structure, Function, and Bioinformatics*, *86*(5), 581–591.

Gao, M., & Skolnick, J. (2010a). iAlign: A method for the structural comparison of protein-protein interfaces. *Bioinformatics (Oxford, England)*, *26*(18), 2259–2265.

Gao, M., & Skolnick, J. (2010b). Structural space of protein–protein interfaces is degenerate, close to complete, and highly connected. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(52), 22517–22522.

Gao, M., & Skolnick, J. (2011). New benchmark metrics for protein-protein docking methods. *Proteins: Structure, Function, and Bioinformatics*, *79*(5), 1623–1634.

Garbuzynskiy, S. O., Ivankov, D. N., Bogatyreva, N. S., & Finkelstein, A. V. (2013). Golden triangle for folding rates of globular proteins. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(1), 147–150.

Garma, L., Mukherjee, S., Mitra, P., & Zhang, Y. (2012). How Many Protein-Protein Interactions Types Exist in Nature? *PLOS ONE*, *7*(6), e38913.

George, R. A., & Heringa, J. (2002a). Protein domain identification and improved sequence similarity searching using PSI-BLAST. *Proteins: Structure, Function, and Bioinformatics*, *48*(4), 672–681.

George, R. A., & Heringa, J. (2002b). SnapDRAGON: A method to delineate protein structural domains from sequence data. *Journal of Molecular Biology*, *316*(3), 839–851.

George, R. A., & Heringa, J. (2002c). An analysis of protein domain linkers: Their classification and role in protein folding. *Protein Engineering, Design and Selection*, *15*(11), 871–879.

Gerstein, M., & Levitt, M. (1998). Comprehensive assessment of automatic structural alignment against a manual standard, the scop classification of proteins. *Protein Science: A Publication of the Protein Society*, *7*(2), 445–456.

Gerstein, M., Lesk, A. M., & Chothia, C. (1994). Structural Mechanisms for Domain Movements in Proteins. *Biochemistry*, *33*(22), 6739–6749.

Gerstein, M., Schulz, G., & Chothia, C. (1993). Domain Closure in Adenylate Kinase: Joints on Either Side of Two Helices Close Like Neighboring Fingers. *Journal of Molecular Biology*, *229*(2), 494–501.

Giver, L., Gershenson, A., Freskgard, P.-O., & Arnold, F. H. (1998). Directed evolution of a thermostable esterase. *Proceedings of the National Academy of Sciences*, *95*(22), 12809–12813.

Glaser, F., Steinberg, D. M., Vakser, I. A., & Ben-Tal, N. (2001). Residue frequencies and pairing preferences at protein–protein interfaces. *Proteins: Structure, Function, and Bioinformatics*, *43*(2), 89–102.

Gokhale, R. S., Tsuji, S. Y., Cane, D. E., & Khosla, C. (1999). Dissecting and Exploiting Intermodular Communication in Polyketide Synthases. *Science*, *284*(5413), 482–485.

Gough, J. (2005). Convergent evolution of domain architectures (is rare). *Bioinformatics*, *21*(8), 1464–1471.

Gough, J., Karplus, K., Hughey, R., & Chothia, C. (2001). Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure11Edited by G. Von Heijne. *Journal of Molecular Biology*, *313*(4), 903–919.

Gouzy, J., Eugéne, P., Greene, E. A., Kahn, D., & Corpet, F. (1997). XDOM, a graphical tool to analyse domain arrangements in any set of protein sequences. *Computer Applications in the Biosciences: CABIOS*, *13*(6), 601–608.

Grabowski, M., Niedzialkowska, E., Zimmerman, M. D., & Minor, W. (2016). The impact of structural genomics: the first quindecennial. *Journal of Structural and Functional Genomics*, 17(1), 1-16.

Gracy, J., & Argos, P. (1998). Automated protein sequence database classification. II. Delineation Of domain boundaries from sequence similarities. *Bioinformatics (Oxford, England)*, *14*(2), 174–187.

Guinn, E. J., Kontur, W. S., Tsodikov, O. V., Shkel, I., & Record, M. T. (2013). Probing the protein-folding mechanism using denaturant and temperature effects on rate constants. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(42), 16784–16789.

Gurung, A. B., Bhattacharjee, A., Ajmal Ali, M., Al-Hemaid, F., & Lee, J. (2017). Binding of small molecules at interface of protein–protein complex – A newer approach to rational drug design. *Saudi Journal of Biological Sciences*, *24*(2), 379–388.

Ha, J.-H., & Loh, S. N. (2012). Protein Conformational Switches: From Nature to Design. *Chemistry (Weinheim an Der Bergstrasse, Germany)*, *18*(26), 7984–7999.

Halperin, I., Ma, B., Wolfson, H., & Nussinov, R. (2002). Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins: Structure, Function, and Bioinformatics*, *47*(4), 409–443.

Hamon, V., & Morelli, X. (2013). Druggability of protein?protein interactions. In *Understanding and Exploiting Protein?Protein Interactions as Drug Targets* (Vol. 1–0, pp. 18–31). Future Science Ltd.

Han, J.-H., Batey, S., Nickson, A. A., Teichmann, S. A., & Clarke, J. (2007). The folding and evolution of multidomain proteins. *Nature Reviews. Molecular Cell Biology*, *8*(4), 319–330.

Han, J.-H., Kerrison, N., Chothia, C., & Teichmann, S. A. (2006). Divergence of Interdomain Geometry in Two-Domain Proteins. *Structure*, *14*(5), 935–945.

Harder, R. L., & Desmarais, R. N. (1972). Interpolation using surface splines. *Journal of Aircraft*, *9*(2), 189–191.

Harrison, A., Pearl, F., Mott, R., Thornton, J., & Orengo, C. (2002). Quantifying the Similarities within Fold Space. *Journal of Molecular Biology*, *323*(5), 909–926.

Heinig, M., & Frishman, D. (2004). STRIDE: A web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Research*, *32*(Web Server issue), W500–W502.

Henrick, K., & Thornton, J. M. (1998). PQS: A protein quaternary structure file server. *Trends in Biochemical Sciences*, *23*(9), 358–361.

Henry, E. R., Best, R. B., & Eaton, W. A. (2013). Comparing a simple theoretical model for protein folding with all-atom molecular dynamics simulations. *Proceedings of the National Academy of Sciences*, *110*(44), 17880–17885.

Hernández-Santoyo, A., Tenorio-Barajas, A. Y., Altuzar, V., Vivanco-Cid, H., & Mendoza-Barrera, C. (2013). Protein-Protein and Protein-Ligand Docking. *Protein Engineering - Technology and Application*.

Hertig, S., Goddard, T. D., Johnson, G. T., & Ferrin, T. E. (2015). Multidomain Assembler (MDA) Generates Models of Large Multidomain Proteins. *Biophysical Journal*, *108*(9), 2097–2102.

Hirako, S., & Shionyu, M. (2012). DINE: A Novel Score Function for Modeling Multidomain Protein Structures with Domain Linker and Interface Restraints. *IPSJ Transactions on Bioinformatics*, *5*, 18–26.

Holm, L., & Sander, C. (1996). Mapping the protein universe. *Science (New York, N.Y.)*, *273*(5275), 595–603.

Holm, L., & Sander, C. (1994). Parser for protein folding units. *Proteins: Structure, Function, and Bioinformatics*, *19*(3), 256–268.

Holm, L., & Sander, C. (1997). Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Research*, *25*(1), 231–234.

Hong, L., Koelsch, G., Lin, X., Wu, S., Terzyan, S., Ghosh, A. K., Zhang, X. C., & Tang, J. (2000). Structure of the protease domain of memapsin 2 (beta-secretase) complexed with inhibitor. *Science (New York, N.Y.)*, *290*(5489), 150–153.

Hong, L., & Tang, J. (2004). Flap Position of Free Memapsin 2 (β-Secretase), a Model for Flap Opening in Aspartic Protease Catalysis,. *Biochemistry*, *43*(16), 4689–4695.

Hong, L., Turner, R. T., Koelsch, G., Shin, D., Ghosh, A. K., & Tang, J. (2002). Crystal Structure of Memapsin 2 (β-Secretase) in Complex with an Inhibitor OM00-3. *Biochemistry*, *41*(36), 10963–10967.

Hou, J., Jun, S.-R., Zhang, C., & Kim, S.-H. (2005). Global mapping of the protein structure space and application in structure-based inference of protein function. *Proceedings of the National Academy of Sciences*, *102*(10), 3651–3656.

Hou, J., Sims, G. E., Zhang, C., & Kim, S.-H. (2003). A global representation of the protein fold space. *Proceedings of the National Academy of Sciences*, *100*(5), 2386–2390.

Hou, Q., De Geest, P. F. G., Vranken, W. F., Heringa, J., & Feenstra, K. A. (2017). Seeing the trees through the forest: Sequence-based homo- and heteromeric protein-protein interaction sites prediction using random forest. *Bioinformatics (Oxford, England)*, *33*(10), 1479–1487.

Hu, Z., Ma, B., Wolfson, H., & Nussinov, R. (2000). Conservation of polar residues as hot spots at protein interfaces. *Proteins: Structure, Function, and Bioinformatics*, *39*(4), 331–342.

Huang, N., & Jacobson, M. P. (2010). Binding-Site Assessment by Virtual Fragment Screening. *PLOS ONE*, *5*(4), e10109.

Huang, S.-Y. (2015). Exploring the potential of global protein-protein docking: An overview and critical assessment of current programs for automatic ab initio docking. *Drug Discovery Today*, *20*(8), 969–977.

Hubbard, R. E., & Haider, M. K. (2010). Hydrogen Bonds in Proteins: Role and Strength. In *ELS*. American Cancer Society.

Hubbard, S., & Thornton, J. (1993). *Naccess*.

Hubbard, T. J., & Blundell, T. L. (1987). Comparison of solvent-inaccessible cores of homologous proteins: Definitions useful for protein modelling. *Protein Engineering*, *1*(3), 159–171.

Ikebe, M., Kambara, T., Stafford, W. F., Sata, M., Katayama, E., & Ikebe, R. (1998). A Hinge at the Central Helix of the Regulatory Light Chain of Myosin Is Critical for Phosphorylation-dependent Regulation of Smooth Muscle Myosin Motor Activity. *Journal of Biological Chemistry*, *273*(28), 17702–17707.

Illergård, K., Ardell, D. H., & Elofsson, A. (2009). Structure is three to ten times more conserved than sequence—A study of structural response in protein cores. *Proteins*, *77*(3), 499–508.

Inanami, T., Terada, T. P., & Sasai, M. (2014). Folding pathway of a multidomain protein depends on its topology of domain connectivity. *Proceedings of the National Academy of Sciences*, *111*(45), 15969–15974.

Inbar, Y., Benyamini, H., Nussinov, R., & Wolfson, H. J. (2005). Combinatorial docking approach for structure prediction of large proteins and multi-molecular assemblies. *Physical Biology*, *2*(4), S156-165.

Ingólfsson, H., & Yona, G. (2008). Protein Domain Prediction. *Methods in Molecular Biology (Clifton, N.J.)*, *426*, 117–143.

Islam, S. A., Luo, J., & Sternberg, M. J. E. (1995). Identification and analysis of domains in proteins. *Protein Engineering, Design and Selection*, *8*(6), 513–526.

Ito, K., Uyeda, T. Q. P., Suzuki, Y., Sutoh, K., & Yamamoto, K. (2003). Requirement of domain-domain interaction for conformational change and functional ATP hydrolysis in myosin. *The Journal of Biological Chemistry*, *278*(33), 31049–31057.

Iyer, L. M., Leipe, D. D., Koonin, E. V., & Aravind, L. (2004). Evolutionary history and higher order classification of AAA+ ATPases. *Journal of Structural Biology*, *146*(1–2), 11–31.

Jaenicke, R. (1987). Folding and association of proteins. *Progress in Biophysics and Molecular Biology*, *49*(2–3), 117–237.

Janin, J. (1997). Specific versus non-specific contacts in protein crystals. *Nature Structural Biology*, *4*(12), 973–974.

Janin, J., & Chothia, C. (1990). The structure of protein-protein recognition sites. *Journal of Biological Chemistry*, *265*(27), 16027–16030.

Janin, J., & Rodier, F. (1995). Protein-protein interaction at crystal contacts. *Proteins*, *23*(4), 580–587.

Janin, J., & Wodak, S. J. (1983). Structural domains in proteins and their role in the dynamics of protein function. *Progress in Biophysics and Molecular Biology*, *42*(1), 21–78.

Janin, J., Bahadur, R. P., & Chakrabarti, P. (2008). Protein-protein interaction and quaternary structure. *Quarterly Reviews of Biophysics*, 41(2), 133-180.

Jefferson, E. R., Walsh, T. P., & Barton, G. J. (2008). A comparison of SCOP and CATH with respect to domain-domain interactions. *Proteins*, *70*(1), 54–62.

Jeong, H., Mason, S. P., Barabasi, A. L., & Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, 411(6833), 41-42.

Jin, L., Wang, W., & Fang, G. (2014). Targeting Protein-Protein Interaction by Small Molecules. *Annual Review of Pharmacology and Toxicology*, *54*(1), 435–456.

Johnson, J. L., Cusack, B., Hughes, T. F., McCullough, E. H., Fauq, A., Romanovskis, P., Spatola, A. F., & Rosenberry, T. L. (2003). Inhibitors tethered near the acetylcholinesterase active site serve as molecular rulers of the peripheral and acylation sites. *The Journal of Biological Chemistry*, *278*(40), 38948–38955.

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, *292*(2), 195–202.

Jones, S., & Thornton, J. M. (1996). Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(1), 13–20.

Jones, S., & Thornton, J. M. (1997). Analysis of protein-protein interaction sites using surface patches. *Journal of Molecular Biology*, *272*(1), 121–132.

Jones, Susan, Marin, A., & M.Thornton, J. (2000). Protein domain interfaces: Characterization and comparison with oligomeric protein interfaces. *Protein Engineering, Design and Selection*, *13*(2), 77–82.

Jung, W., Jeong, H.-H., & Lee, K. (2012). Protein Interactome and Its Application to Protein Function Prediction. *Protein-Protein Interactions - Computational and Experimental Tools*.

Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, *32*(5), 922–923.

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, *22*(12), 2577–2637.

Kannan, L., & Wheeler, W. C. (2012). Maximum Parsimony on Phylogenetic networks. *Algorithms for Molecular Biology*, *7*(1), 9.

Kastritis, P. L., & Bonvin, A. M. J. J. (2010). Are scoring functions in protein-protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark. *Journal of Proteome Research*, *9*(5), 2216–2225.

Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A. A., Aflalo, C., & Vakser, I. A. (1992). Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(6), 2195–2199.

Kawashima, T., Berthet-Colominas, C., Wulff, M., Cusack, S., & Leberman, R. (1996). The structure of the Escherichia coli EF-Tu· EF-Ts complex at 2.5 Å resolution. *Nature*, *379*(6565), 511–518.

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H., & Phillips, D. C. (1958). A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature*, *181*(4610), 662–666.

Keskin, O., Gursoy, A., Ma, B., & Nussinov, R. (2008). Principles of protein-protein interactions: What are the preferred ways for proteins to interact? *Chemical Reviews*, *108*(4), 1225–1244.

Keskin, O., Ma, B., & Nussinov, R. (2005). Hot regions in protein--protein interactions: The organization and contribution of structurally conserved hot spot residues. *Journal of Molecular Biology*, *345*(5), 1281–1294.

Keskin, O., Tuncbag, N., & Gursoy, A. (2016). Predicting Protein–Protein Interactions from the Molecular to the Proteome Level. *Chemical Reviews*, *116*(8), 4884–4909.

Kim, W. K., Bolser, D. M., & Park, J. H. (2004). Large-scale co-evolution analysis of protein structural interlogues using the global protein structural interactome map (PSIMAP). *Bioinformatics*, *20*(7), 1138–1150.

Kim, W. K., Henschel, A., Winter, C., & Schroeder, M. (2006). The Many Faces of Protein–Protein Interactions: A Compendium of Interface Geometry. *PLOS Computational Biology*, *2*(9), e124.

Kirsanov, D. D., Zanegina, O. N., Aksianov, E. A., Spirin, S. A., Karyagina, A. S., & Alexeevski, A. V. (2013). NPIDB: Nucleic acid-protein interaction database. *Nucleic Acids Research*, *41*(D1), D517–D523.

Klemm, J. D., Schreiber, S. L., & Crabtree, G. R. (1998). Dimerization as a Regulatory Mechanism in Signal Transduction. *Annual Review of Immunology*, *16*(1), 569–592.

Kobayashi, C., Matsunaga, Y., Koike, R., Ota, M., & Sugita, Y. (2015). Domain Motion Enhanced (DoME) Model for Efficient Conformational Sampling of Multidomain Proteins. *The Journal of Physical Chemistry B*, *119*(46), 14584–14593.

Kolodny, R., Petrey, D., & Honig, B. (2006). Protein structure comparison: Implications for the nature of “fold space”, and structure and function prediction. *Current Opinion in Structural Biology*, *16*(3), 393–398.

Koonin, E. V., Aravind, L., & Kondrashov, A. S. (2000). The impact of comparative genomics on our understanding of evolution. *Cell*, *101*(6), 573–576.

Koonin, E. V., & Galperin, M. Y. (2003). In *Sequence - Evolution - Function: Computational Approaches in Comparative Genomics*. Boston.

Kovermann, M., Rogne, P., & Wolf-Watz, M. (2016). Protein dynamics and function from solution state NMR spectroscopy. *Quarterly Reviews of Biophysics*, *49*.

Krishnan, V., & Rupp, B. (2012). Macromolecular Structure Determination: Comparison of X-ray Crystallography and NMR Spectroscopy. *ELS*.

Kryshtafovych, A., Fidelis, K., & Moult, J. (2011). CASP9 results compared to those of previous CASP experiments. *Proteins*, *79 Suppl 10*, 196–207.

Kumar, S., Ma, B., Tsai, C.-J., & Nussinov, R. (2000). Electrostatic strengths of salt bridges in thermophilic and mesophilic glutamate dehydrogenase monomers. *Proteins: Structure, Function, and Bioinformatics*, *38*(4), 368–383.

Kummerfeld, S. K., & Teichmann, S. A. (2005). Relative rates of gene fusion and fission in multi-domain proteins. *Trends in Genetics: TIG*, *21*(1), 25–30.

Kummerfeld, S. K., & Teichmann, S. A. (2009). Protein domain organisation: Adding order. *BMC Bioinformatics*, *10*(1), 39.

Laskowski, R. A. (1995). SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions. *Journal of Molecular Graphics*, *13*(5), 323–330, 307–308.

Laskowski, R. A. (2003). Structural quality assurance. *Methods of Biochemical Analysis*, *44*, 273–303.

Lawrence, M. C., & Colman, P. M. (1993). Shape complementarity at protein/protein interfaces. *Journal of Molecular Biology*, *234*(4), 946–950.

Lawrence, M. C., & Colman, P. M. (1993). Shape Complementarity at Protein/Protein Interfaces. *Journal of Molecular Biology*, *234*(4), 946–950.

Lawson, D. M., Williams, C. E., Mitchenall, L. A., & Pau, R. N. (1998). Ligand size is a major determinant of specificity in periplasmic oxyanion-binding proteins: The 1.2 å resolution crystal structure of Azotobacter vinelandii ModA. *Structure*, *6*(12), 1529–1539.

Lee, B., & Richards, F. M. (1971). The interpretation of protein structures: Estimation of static accessibility. *Journal of Molecular Biology*, *55*(3), 379–400.

Lee, I., Blom, U. M., Wang, P. I., Shim, J. E., & Marcotte, E. M. (2011). Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Research*, *21*(7), 1109–1121.

Lee, I., Lehner, B., Vavouri, T., Shin, J., Fraser, A. G., & Marcotte, E. M. (2010). Predicting genetic modifier loci using functional gene networks. *Genome Research*, *20*(8), 1143–1153.

Lee, R. A., Razaz, M., & Hayward, S. (2003). The DynDom database of protein domain motions. *Bioinformatics (Oxford, England)*, *19*(10), 1290–1291.

Lehne, B., & Schlitt, T. (2009). Protein-protein interaction databases: Keeping up with growing interactomes. *Human Genomics*, *3*(3), 291.

Lensink, M. F., & Wodak, S. J. (2013). Docking, scoring, and affinity prediction in CAPRI. *Proteins: Structure, Function, and Bioinformatics*, *81*(12), 2082–2095.

Levitt, M. (2009). Nature of the protein universe. *Proceedings of the National Academy of Sciences*.

Levitt, M., & Chothia, C. (1976). Structural patterns in globular proteins. *Nature*, *261*(5561), 552–558.

Levy, Y. (2017). Protein Assembly and Building Blocks: Beyond the Limits of the LEGO Brick Metaphor. *Biochemistry*, *56*(38), 5040–5048.

Li, B., & Kihara, D. (2012). Protein docking prediction using predicted protein-protein interface. *BMC Bioinformatics*, *13*(1), 7.

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics (Oxford, England)*, *22*(13), 1658–1659.

Liao, S.-M., Du, Q.-S., Meng, J.-Z., Pang, Z.-W., & Huang, R.-B. (2013). The multiple roles of histidine in protein interactions. *Chemistry Central Journal*, *7*, 44.

Lindorff-Larsen, K., Best, R. B., DePristo, M. A., Dobson, C. M., & Vendruscolo, M. (2005). Simultaneous determination of protein structure and dynamics. *Nature*, *433*(7022), 128–132.

Lionta, E., Spyrou, G., Vassilatis, D. K., & Cournia, Z. (2014). Structure-Based Virtual Screening for Drug Discovery: Principles, Applications and Recent Advances. *Current Topics in Medicinal Chemistry*, *14*(16), 1923–1938.

Lise, S., Walker-Taylor, A., & Jones, D. T. (2006). Docking protein domains in contact space. *BMC Bioinformatics*, *7*(1), 310.

Littler, S. J., & Hubbard, S. J. (2005). Conservation of Orientation and Sequence in Protein Domain–Domain Interactions. *Journal of Molecular Biology*, *345*(5), 1265–1279.

Liu, J., & Rost, B. (2004a). Sequence-based prediction of protein domains. *Nucleic Acids Research*, *32*(12), 3522–3530.

Liu, J., & Rost, B. (2004b). CHOP: Parsing proteins into structural domains. *Nucleic Acids Research*, *32*(Web Server issue), W569–W571.

Lo Conte, L., Chothia, C., & Janin, J. (1999). The atomic structure of protein-protein recognition sites. *Journal of Molecular Biology*, *285*(5), 2177–2198.

London, N., Raveh, B., Movshovitz-Attias, D., & Schueler-Furman, O. (2010). Can Self-Inhibitory Peptides be Derived from the Interfaces of Globular Protein-Protein Interactions? *Proteins*, *78*(15), 3140–3149.

London, N., Raveh, B., & Schueler-Furman, O. (2013). Druggable protein–protein interactions – from hot spots to hot segments. *Current Opinion in Chemical Biology*, *17*(6), 952–959.

Loregian, A., Marsden, H. S., & Palù, G. (2002). Protein–protein interactions as targets for antiviral chemotherapy. *Reviews in Medical Virology*, *12*(4), 239–262.

Lunelli, M., Di Paolo, M. L., Biadene, M., Calderone, V., Battistutta, R., Scarpa, M., Rigo, A., & Zanotti, G. (2005). Crystal Structure of Amine Oxidase from Bovine Serum. *Journal of Molecular Biology*, *346*(4), 991–1004.

Lupas, A., & Koretke, K. (2008). *Evolution of Protein Folds* (pp. 131–151).

Ikebe, M., Kambara, T., Stafford, W.F., Sata, M., Katayama, E., & Ikebe, R. (1998). A hinge at the central helix of the regulatory light chain of myosin is critical for phosphorylation-dependent regulation of smooth muscle myosin motor activity. *The Journal of Biological Chemistry*, *273*(28), 17702–17707. doi: 10.1074/jbc.273.28.17702

Macalino, S. J. Y., Basith, S., Clavio, N. A. B., Chang, H., Kang, S., & Choi, S. (2018). Evolution of In Silico Strategies for Protein-Protein Interaction Drug Discovery. *Molecules : A Journal of Synthetic Chemistry and Natural Product Chemistry*, *23*(8).

Maheshwari, S., & Brylinski, M. (2015). Predicted binding site information improves model ranking in protein docking using experimental and computer-generated target structures. *BMC Structural Biology*, *15*.

Majumdar, I., Kinch, L. N., & Grishin, N. V. (2009). A Database of Domain Definitions for Proteins with Complex Interdomain Geometry. *PLOS ONE*, *4*(4), e5084.

Makley, L., & Gestwicki, J. (2013). Expanding the Number of 'Druggable' Targets: Non-Enzymes and Protein-Protein Interactions. *Chemical Biology & Drug Design*, *81*, 22–32.

Mandal, S., Moudgil, M., & Mandal, S. K. (2009). Rational drug design. *European Journal of Pharmacology*, *625*(1–3), 90–100.

Marcotte, E. M., Pellegrini, M., Ng, H.-L., Rice, D. W., Yeates, T. O., & Eisenberg, D. (1999). Detecting Protein Function and Protein-Protein Interactions from Genome Sequences. *Science*, *285*(5428), 751–753.

Marion, D. (2013). An Introduction to Biological NMR Spectroscopy. *Molecular & Cellular Proteomics : MCP*, *12*(11), 3006–3025.

Markus, M., & Benezra, R. (1999). Two Isoforms of Protein Disulfide Isomerase Alter the Dimerization Status of E2A Proteins by a Redox Mechanism. *Journal of Biological Chemistry*, *274*(2), 1040–1049.

Marsden, R. L., McGuffin, L. J., & Jones, D. T. (2002). Rapid protein domain assignment from amino acid sequence using predicted secondary structure. *Protein Science : A Publication of the Protein Society*, *11*(12), 2814–2824.

Marsh, J. A., & Teichmann, S. A. (2010). How do proteins gain new domains? *Genome Biology*, *11*(7), 126.

McCoy, A. J., Chandana Epa, V., & Colman, P. M. (1997). Electrostatic complementarity at protein/protein interfaces11Edited by B. Honig. *Journal of Molecular Biology*, *268*(2), 570–584.

McGuire, A. L., Gabriel, S., Tishkoff, S. A., Wonkam, A., Chakravarti, A., Furlong, E. E. M., . . . Kim, J. S. (2020). The road ahead in genetics and genomics. *Nature Reviews. Genetics*, 21(10), 581-596.

Meinguet, J. (1979). *Multivariate interpolation at arbitrary points made simple*.

Mikami, B., Iwamoto, H., Malle, D., Yoon, H.-J., Demirkan-Sarikaya, E., Mezaki, Y., & Katsuya, Y. (2006). Crystal structure of pullulanase: Evidence for parallel binding of oligosaccharides in the active site. *Journal of Molecular Biology*, *359*(3), 690–707.

Miller, S., Janin, J., Lesk, A. M., & Chothia, C. (1987). Interior and surface of monomeric proteins. *Journal of Molecular Biology*, *196*(3), 641–656.

Minor, D. L. (2007). The Neurobiologist's Guide to Structural Biology: A Primer on Why Macromolecular Structure Matters and How to Evaluate Structural Data. *Neuron*, *54*(4), 511–533.

Mintseris, J., & Weng, Z. (2005). Structure, function, and evolution of transient and obligate protein–protein interactions. *Proceedings of the National Academy of Sciences*, *102*(31), 10930–10935.

Miyazaki, S., Kuroda, Y., & Yokoyama, S. (2002). Characterization and prediction of linker sequences of multi-domain proteins by a neural network. *Journal of Structural and Functional Genomics*, *2*(1), 37–51.

Miyazawa, S., & Jernigan, R. L. (1985). Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules*, *18*(3), 534–552.

Monod, J., Wyman, J., & Changeux, J. P. (1965). ON THE NATURE OF ALLOSTERIC TRANSITIONS: A PLAUSIBLE MODEL. *Journal of Molecular Biology*, *12*, 88–118.

Moore, A. D., Björklund, A. K., Ekman, D., Bornberg-Bauer, E., & Elofsson, A. (2008). Arrangements in the modular evolution of proteins. *Trends in Biochemical Sciences*, *33*(9), 444–451.

Morange, M. (2006). The Protein Side of the Central Dogma: Permanence and Change. *History and Philosophy of the Life Sciences*, *28*(4), 513–524. JSTOR.

Mowbray, S. L., Helgstrand, C., Sigrell, J. A., Cameron, A. D., & Jones, T. A. (1999). Errors and reproducibility in electron-density map interpretation. *Acta Crystallographica Section D*, *55*(7), 1309–1319.

Mullard, A. (2012). Protein-protein interaction inhibitors get into the groove. *Nature Reviews. Drug Discovery*, *11*(3), 173–175.

Munoz, V., & Eaton, W. A. (1999). A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proceedings of the National Academy of Sciences*, *96*(20), 11311–11316.

Murzin, A. G., Brenner, S. E., Hubbard, T., & Chothia, C. (1995). SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, *247*(4), 536–540.

Nakamura, T., Tonozuka, T., Ito, S., Takeda, Y., Sato, R., Matsuo, I., Ito, Y., Oguma, K., & Nishikawa, A. (2011). Molecular diversity of the two sugar-binding sites of the β-trefoil lectin HA33/C (HA1) from Clostridium botulinum type C neurotoxin. *Archives of Biochemistry and Biophysics*, *512*(1), 69–77.

Narayanan, C., Bafna, K., Roux, L. D., Agarwal, P. K., & Doucet, N. (2017). Applications of NMR and computational methodologies to study protein dynamics. *Archives of Biochemistry and Biophysics*, *628*, 71–80.

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*(3), 443–453.

Nelson, D. L., Lehninger, A. L., & Cox, M. M. (2008). *Lehninger principles of biochemistry* (5thed.): New York : W.H. Freeman.

Nooren, I. M. A., & Thornton, J. M. (2003). Diversity of protein-protein interactions. *The EMBO Journal*, *22*(14), 3486–3492.

Ofran, Y., & Rost, B. (2003). Analysing six types of protein-protein interfaces. *Journal of Molecular Biology*, *325*(2), 377–387.

Oliva, R., Vangone, A., & Cavallo, L. (2013). Ranking multiple docking solutions based on the conservation of inter-residue contacts. *Proteins: Structure, Function, and Bioinformatics*, *81*(9), 1571–1584.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., & Thornton, J. M. (1997). CATH--a hierarchic classification of protein domain structures. *Structure (London, England: 1993)*, *5*(8), 1093–1108.

Ortiz, M. A., Light, J., Maki, R. A., & Assa-Munt, N. (1999). Mutation analysis of the Pip interaction domain reveals critical residues for protein–protein interactions. *Proceedings of the National Academy of Sciences*, *96*(6), 2740–2745.

Otterbein, L. R., Cosio, C., Graceffa, P., & Dominguez, R. (2002). Crystal structures of the vitamin D-binding protein and its complex with actin: Structural basis of the actin-scavenger system. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(12), 8003–8008.

Ozdemir, E. S., Halakou, F., Nussinov, R., Gursoy, A., & Keskin, O. (2019). Methods for Discovering and Targeting Druggable Protein-Protein Interfaces and Their Application to Repurposing. *Methods in Molecular Biology (Clifton, N.J.)*, *1903*, 1–21.

Pace, C. N., Scholtz, J. M., & Grimsley, G. R. (2014). Forces Stabilizing Proteins. *FEBS Letters*, *588*(14), 2177–2184.

Pandit, S. B., & Skolnick, J. (2010). TASSER_low-zsc: an approach to improve structure prediction using low z-score-ranked templates. *Proteins*, 78(13), 2769-2780.

Park, J., & Bolser, D. (2001). Conservation of Protein Interaction Network in Evolution. *Genome Informatics. International Conference on Genome Informatics*, *12*, 135–140.

Park, J., Lappe, M., & Teichmann, S. A. (2001). Mapping protein family interactions: Intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast11Edited by J. Karn. *Journal of Molecular Biology*, *307*(3), 929–938.

Pascual-García, A., Abia, D., Méndez, R., Nido, G. S., & Bastolla, U. (2010). Quantifying the evolutionary divergence of protein structures: The role of function change and function conservation. *Proteins*, *78*(1), 181–196.

Pascual-García, A., Abia, D., Ortiz, Á. R., & Bastolla, U. (2009). Cross-Over between Discrete and Continuous Protein Structure Space: Insights into Automatic Classification and Networks of Protein Structures. *PLOS Computational Biology*, *5*(3), e1000331.

Patthy, L. (1996). Exon shuffling and other ways of module exchange. *Matrix Biology*, *15*(5), 301–310.

Patwardhan, A. (2017). Trends in the Electron Microscopy Data Bank (EMDB). *Acta Crystallographica. Section D, Structural Biology*, 73(Pt 6), 503-508.

Perkins, J. R., Diboun, I., Dessailly, B. H., Lees, J. G., & Orengo, C. (2010). Transient Protein-Protein Interactions: Structural, Functional, and Network Properties. *Structure*, *18*(10), 1233–1243.

Perutz, M. F., Rossmann, M. G., Cullis, A. F., Muirhead, H., Will, G., & North, A. C. (1960). Structure of haemoglobin: A three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis. *Nature*, *185*(4711), 416–422.

Phillips, D. C. (1966). The three-dimensional structure of an enzyme molecule. *Scientific American*, *215*(5), 78–90.

Ponstingl, H., Henrick, K., & Thornton, J. M. (2000). Discriminating between homodimeric and monomeric proteins in the crystalline state. *Proteins*, *41*(1), 47–57.

Ponting, C. P., & Russell, R. R. (2002). The Natural History of Protein Domains. *Annual Review of Biophysics and Biomolecular Structure*, *31*(1), 45–71.

Postic, G., Ghouzam, Y., Chebrek, R., & Gelly, J.-C. (2017). An ambiguity principle for assigning protein structural domains. *Science Advances*, *3*(1), e1600552.

Prasad Bahadur, R., Chakrabarti, P., Rodier, F., & Janin, J. (2004). A Dissection of Specific and Non-specific Protein–Protein Interfaces. *Journal of Molecular Biology*, *336*(4), 943–955.

Prieto, C., & Rivas, J. D. L. (2010). Structural domain–domain interactions: Assessment and comparison with protein–protein interaction data to improve the interactome. *Proteins: Structure, Function, and Bioinformatics*, *78*(1), 109–117.

Privalov, P. L., & Khechinashvili, N. N. (1974). A thermodynamic approach to the problem of stabilization of globular protein structure: A calorimetric study. *Journal of Molecular Biology*, *86*(3), 665–684.

Rajamani, D., Thiel, S., Vajda, S., & Camacho, C. J. (2004). Anchor residues in protein–protein interactions. *Proceedings of the National Academy of Sciences*, *101*(31), 11287–11292.

Rao, V. S., Srinivas, K., Sujini, G. N., & Kumar, G. N. S. (2014). *Protein-Protein Interaction Detection: Methods and Analysis* [Research article]. International Journal of Proteomics.

Ravera, E., Salmon, L., Fragai, M., Parigi, G., Al-Hashimi, H., & Luchinat, C. (2014). Insights into Domain–Domain Motions in Proteins and RNA from Solution NMR. *Accounts of Chemical Research*, *47*(10), 3118–3126.

Rekha, N., Machado, S. M., Narayanan, C., Krupa, A., & Srinivasan, N. (2005). Interaction interfaces of protein domains are not topologically equivalent across families within superfamilies: Implications for metabolic and signaling pathways. *Proteins*, *58*(2), 339–353.

Richardson, J. S. (1981). The anatomy and taxonomy of protein structure. *Advances in Protein Chemistry*, *34*, 167–339.

Robinson, C. R., & Sauer, R. T. (1998). Optimizing the stability of single-chain proteins by linker length and composition mutagenesis. *Proceedings of the National Academy of Sciences*, *95*(11), 5929–5934.

Rohl, C. A., Strauss, C. E. M., Misura, K. M. S., & Baker, D. (2004). Protein Structure Prediction Using Rosetta. In *Methods in Enzymology* (Vol. 383, pp. 66–93). Academic Press.

Rossmann, M. G., Morais, M. C., Leiman, P. G., & Zhang, W. (2005). Combining X-ray crystallography and electron microscopy. *Structure*, *13*(3), 355–362.

Rossmann, M. G., Moras, D., & Olsen, K. W. (1974). Chemical and biological evolution of a nucleotide-binding protein. *Nature*, *250*(5463), 194–199.

Rost, B. (1997). Protein structures sustain evolutionary drift. *Folding and Design*, *2*, S19–S24.

Roy, A., Kucukural, A., & Zhang, Y. (2010). I-TASSER: A unified platform for automated protein structure and function prediction. *Nature Protocols*, *5*(4), 725–738.

Rual, J. F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., . . . Vidal, M. (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062), 1173-1178.

Russell, R. B., & Barton, G. J. (1994). Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts secondary structure and accessibility. *Journal of Molecular Biology*, *244*(3), 332–350.

Ruvinov, S., Wang, L., Ruan, B., Almog, O., Gilliland, G. L., Eisenstein, E., & Bryan, P. N. (1997). Engineering the Independent Folding of the Subtilisin BPN‘ Prodomain: Analysis of Two-State Folding versus Protein Stability. *Biochemistry*, *36*(34), 10414–10421.

Sadowski, M., & Taylor, W. (2010). On the evolutionary origins of “Fold Space Continuity”: A study of topological convergence and divergence in mixed alpha-beta domains. *Journal of Structural Biology*, *172*(3), 244–252.

Sadreyev, R. I., Kim, B.-H., & Grishin, N. V. (2009). Discrete—Continuous Duality of Protein Structure Space. *Current Opinion in Structural Biology*, *19*(3), 321–328.

Samanta, U., Bahadur, R. P., & Chakrabarti, P. (2002). Quantifying the accessible surface area of protein residues in their local environment. *Protein Engineering, Design and Selection*, *15*(8), 659–667.

Sanchez, C., Lachaize, C., Janody, F., Bellon, B., Röder, L., Euzenat, J., Rechenmann, F., & Jacq, B. (1999). Grasping at molecular interactions and genetic networks in Drosophila melanogaster using FlyNets, an Internet database. *Nucleic Acids Research*, *27*(1), 89–94.

Savage, H. J., Elliott, C. J., Freeman, C. M., & Finney, J. L. (1993). Lost hydrogen bonds and buried surface area: Rationalising stability in globular proteins. *Journal of the Chemical Society, Faraday Transactions*, *89*(15), 2609–2617.

Savitsky, P., Bray, J., Cooper, C. D. O., Marsden, B. D., Mahajan, P., Burgess-Brown, N. A., & Gileadi, O. (2010). High-throughput production of human proteins for crystallization: The SGC experience. *Journal of Structural Biology*, *172*(1), 3–13.

Schlunegger, M. P., Bennett, M. J., & Eisenberg, D. (1997). Oligomer Formation By 3D Domain Swapping: A Model For Protein Assembly And Misassembly. In F. M. Richards, D. S. Eisenberg, & P. S. Kim (Eds.), *Advances in Protein Chemistry* (Vol. 50, pp. 61–122). Academic Press.

Scott, D. E., Bayly, A. R., Abell, C., & Skidmore, J. (2016). Small molecules, big targets: Drug discovery faces the protein–protein interaction challenge. *Nature Reviews Drug Discovery*, *15*(8), 533–550.

Scott, K. A., Steward, A., Fowler, S. B., & Clarke, J. (2002). Titin; a multidomain protein that behaves as the sum of its parts11Edited by J. Karn. *Journal of Molecular Biology*, *315*(4), 819–829.

Selkoe, D. J. (1998). The cell biology of beta-amyloid precursor protein and presenilin in Alzheimer's disease. *Trends in Cell Biology*, *8*(11), 447–453.

Selzer, T., & Schreiber, G. (1999). Predicting the rate enhancement of protein complex formation from the electrostatic energy of interaction11Edited by B. Honig. *Journal of Molecular Biology*, *287*(2), 409–419.

Shangary, S., & Wang, S. (2009). Small-molecule inhibitors of the MDM2-p53 protein-protein interaction to reactivate p53 function: A novel approach for cancer therapy. *Annual Review of Pharmacology and Toxicology*, *49*, 223–241.

Sharan, R., & Ideker, T. (2006). Modeling cellular machinery through biological network comparison. *Nature Biotechnology*, *24*(4), 427–433.

Sheinerman, F. B., Norel, R., & Honig, B. (2000). Electrostatic aspects of protein–protein interactions. *Current Opinion in Structural Biology*, *10*(2), 153–159.

Shindyalov, I. N., & Bourne, P. E. (2000). An alternative view of protein fold space. *Proteins: Structure, Function, and Bioinformatics*, *38*(3), 247–260.

Shoemaker, B. A., Panchenko, A. R., & Bryant, S. H. (2006). Finding biologically relevant protein domain interactions: Conserved binding mode analysis. *Protein Science : A Publication of the Protein Society*, *15*(2), 352–361.

Shoichet, B. K., & Kuntz, I. D. (1991). Protein docking and complementarity. *Journal of Molecular Biology*, *221*(1), 327–346.

Siddiqui, A. S., & Barton, G. J. (1995). Continuous and discontinuous domains: An algorithm for the automatic generation of reliable protein domain definitions. *Protein Science : A Publication of the Protein Society*, *4*(5), 872–884.

Sistla, R. K., K. V., B., & Vishveshwara, S. (2005). Identification of domains and domain interface residues in multidomain proteins from graph spectral method. *Proteins: Structure, Function, and Bioinformatics*, *59*(3), 616–626.

Skolnick, J., Arakaki, A. K., Lee, S. Y., & Brylinski, M. (2009). The continuity of protein structure space is an intrinsic property of proteins. *Proceedings of the National Academy of Sciences*, *106*(37), 15690–15695.

Skrabanek, L., Saini, H. K., Bader, G. D., & Enright, A. J. (2008). Computational prediction of protein-protein interactions. *Molecular Biotechnology*, *38*(1), 1–17.

Smith, G. R., & Sternberg, M. J. E. (2002). Prediction of protein-protein interactions by docking methods. *Current Opinion in Structural Biology*, *12*(1), 28–35.

Sobolev, V., Sorokine, A., Prilusky, J., Abola, E. E., & Edelman, M. (1999). Automated analysis of interatomic contacts in proteins. *Bioinformatics (Oxford, England)*, *15*(4), 327–332.

Sonnhammer, E. L., & Kahn, D. (1994). Modular arrangement of proteins as inferred from analysis of homology. *Protein Science : A Publication of the Protein Society*, *3*(3), 482–492.

Sousa, S. F., Fernandes, P. A., & Ramos, M. J. (2006). Protein-ligand docking: Current status and future challenges. *Proteins*, *65*(1), 15–26.

Srivastava, S. K., Gayathri, S., Manjasetty, B. A., & Gopal, B. (2012). Analysis of Conformational Variation in Macromolecular Structural Models. *PLOS ONE*, *7*(7), e39993.

Sriwastava, B. K., Basu, S., Maulik, U., & Plewczynski, D. (2013). PPIcons: Identification of protein-protein interaction sites in selected organisms. *Journal of Molecular Modeling*, *19*(9), 4059–4070.

Sugiki, T., Kobayashi, N., & Fujiwara, T. (2017). Modern Technologies of Solution Nuclear Magnetic Resonance Spectroscopy for Three-dimensional Structure Determination of Proteins Open Avenues for Life Scientists. *Computational and Structural Biotechnology Journal*, *15*, 328–339.

Sukhwal, A., & Sowdhamini, R. (2013). Oligomerisation status and evolutionary conservation of interfaces of protein structural domain superfamilies. *Molecular BioSystems*, *9*(7), 1652–1661.

Tanaka, T., Kuroda, Y., & Yokoyama, S. (2003). Characteristics and prediction of domain linker sequences in multi-domain proteins. *Journal of Structural and Functional Genomics*, *4*(2–3), 79–85.

Taylor, P., Blackburn, E., Sheng, Y. G., Harding, S., Hsin, K.-Y., Kan, D., Shave, S., & Walkinshaw, M. D. (2008). Ligand discovery and virtual screening using the program LIDAEUS. *British Journal of Pharmacology*, *153*(Suppl 1), S55–S67.

Taylor, W. R. (1999). Protein structural domain identification. *Protein Engineering*, *12*(3), 203–216.

Taylor, W. R. (2007). Evolutionary transitions in protein fold space. *Current Opinion in Structural Biology, 17*(3), 354-361.

Taylor, W. R. (2020). Exploring Protein Fold Space. *Biomolecules, 10*(2).

Tchigvintsev, A., Singer, A., Brown, G., Flick, R., Evdokimova, E., Tan, K., Gonzalez, C. F., Savchenko, A., & Yakunin, A. F. (2012). Biochemical and Structural Studies of Uncharacterized Protein PA0743 from Pseudomonas aeruginosa Revealed NAD+-dependent l-Serine Dehydrogenase. *Journal of Biological Chemistry*, *287*(3), 1874–1883.

Teichmann, S. A. (2002). Principles of protein-protein interactions. *Bioinformatics*, *18Suppl 2*, S249.

Terwilliger, T. C., Stuart, D., & Yokoyama, S. (2009). Lessons from Structural Genomics. *Annual Review of Biophysics*, *38*, 371–383.

Thiel, P., Kaiser, M., & Ottmann, C. (2012). Small-Molecule Stabilization of Protein–Protein Interactions: An Underestimated Concept in Drug Discovery? *Angewandte Chemie International Edition*, *51*(9), 2012–2018.

Topf, M., & Sali, A. (2005). Topf, M. & Sali, A. Combining electron microscopy and comparative protein structure modeling. Curr. Opin. Struct. Biol. 15, 578-585. *Current Opinion in Structural Biology*, *15*, 578–585.

Tsai, C. J., Lin, S. L., Wolfson, H. J., Nussinov, R., Alerting, E., Tsai, C., Lin, S. L., Wolfson, H. J., & Nussinov, R. (1996). *Studies of protein-protein interfaces: A statistical analysis of the hydrophobic effect*.

Turinsky, A. L., Razick, S., Turner, B., Donaldson, I. M., & Wodak, S. J. (2010). Literature curation of protein interactions: Measuring agreement across major public databases. *Database: The Journal of Biological Databases and Curation*, *2010*.

UniProt: The universal protein knowledgebase. (2017). *Nucleic Acids Research*, *45*(Database issue), D158–D169.

Vajda, S., & Camacho, C. J. (2004). Protein-protein docking: Is the glass half-full or half-empty? *Trends in Biotechnology*, *22*(3), 110–116.

Valas, R. E., Yang, S., & Bourne, P. E. (2009). Nothing about protein structure classification makes sense except in the light of evolution. *Current Opinion in Structural Biology*, *19*(3), 329–334.

van Leeuwen, H. C., Strating, M. J., Rensen, M., de Laat, W., & van der Vliet, P. C. (1997). Linker length and composition influence the flexibility of Oct-1 DNA binding. *The EMBO Journal*, *16*(8), 2043–2053.

van Zundert, G. C. P., Rodrigues, J. P. G. L. M., Trellet, M., Schmitz, C., Kastritis, P. L., Karaca, E., Melquiond, A. S. J., van Dijk, M., de Vries, S. J., & Bonvin, A. M. J. J. (2016). The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *Journal of Molecular Biology*, *428*(4), 720–725.

Vassylyev, D. G., Tomitori, H., Kashiwagi, K., Morikawa, K., & Igarashi, K. (1998). Crystal Structure and Mutational Analysis of the *Escherichia coli* Putrescine Receptor: STRUCTURAL BASIS FOR SUBSTRATE SPECIFICITY. *Journal of Biological Chemistry*, *273*(28), 17604–17609.

Velankar, S., Dana, J. M., Jacobsen, J., van Ginkel, G., Gane, P. J., Luo, J., Oldfield, T. J., O'Donovan, C., Martin, M.-J., & Kleywegt, G. J. (2013). SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Research*, *41*(D1), D483–D489.

Verboven, C., Bogaerts, I., Waelkens, E., Rabijns, A., Van Baelen, H., Bouillon, R., & De Ranter, C. (2003). Actin-DBP: The perfect structural fit? *Acta Crystallographica. Section D, Biological Crystallography*, *59*(Pt 2), 263–273.

Vitagliano, L., Masullo, M., Sica, F., Zagari, A., & Bocchini, V. (2001). The crystal structure of Sulfolobus solfataricus elongation factor 1α in complex with GDP reveals novel features in nucleotide binding and exchange. *The EMBO Journal*, *20*(19), 5305–5311.

Vogel, C., Berzuini, C., Bashton, M., Gough, J., & Teichmann, S. A. (2004). Supra-domains: Evolutionary Units Larger than Single Protein Domains. *Journal of Molecular Biology*, *336*(3), 809–823.

Vogel, C., Teichmann, S. A., & Pereira-Leal, J. (2005). The relationship between domain duplication and recombination. *Journal of Molecular Biology*, *346*(1), 355–365.

Vogl, T., Jatzke, C., Hinz, H.-J., Benz, J., & Huber, R. (1997). Thermodynamic Stability of Annexin V E17G: Equilibrium Parameters from an Irreversible Unfolding Reaction,. *Biochemistry*, *36*(7), 1657–1668.

Wang, Y., Jiang, Y., Meyering-Voss, M., Sprinzl, M., & Sigler, P. B. (1997). Crystal structure of the EF-Tu˙EF-Ts complex from Thermus thermophilus. *Nature Structural Biology*, *4*(8), 650–656.

Webb, E. C. (1993). Enzyme Nomenclature: A Personal Retrospective. *The FASEB Journal*, *7*(12), 1192–1194.

Wei, M., Ye, D., & Dunaway-Mariano, D. (2001). Investigation of the role of the domain linkers in separate site catalysis by Clostridium symbiosum pyruvate phosphate dikinase. *Biochemistry*, *40*(45), 13466–13473.

Weiner, J., Beaussart, F., & Bornberg-Bauer, E. (2006). Domain deletions and substitutions in the modular protein evolution. *The FEBS Journal*, *273*(9), 2037–2047.

Weiner, J., & Bornberg-Bauer, E. (2006). Evolution of Circular Permutations in Multidomain Proteins. *Molecular Biology and Evolution*, *23*(4), 734–743.

Weiser, J., Shenkin, P. S., & Still, W. C. (1999). Approximate atomic surfaces from linear combinations of pairwise overlaps (LCPO). *Journal of Computational Chemistry*, *20*(2), 217–230.

Wernisch, L., Hunting, M., & Wodak, S. J. (1999). Identification of structural domains in proteins by a graph heuristic. *Proteins: Structure, Function, and Bioinformatics*, *35*(3), 338–352.

Wetlaufer, D. B. (1973). Nucleation, Rapid Folding, and Globular Intrachain Regions in Proteins. *Proceedings of the National Academy of Sciences*, *70*(3), 697–701.

Wheelan, S. J., Marchler-Bauer, A., & Bryant, S. H. (2000). Domain size distributions can predict domain boundaries. *Bioinformatics (Oxford, England)*, *16*(7), 613–618.

Wlodawer, A., Minor, W., Dauter, Z., & Jaskolski, M. (2008). Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. *The FEBS journal, 275*(1), 1-21.

Wodak, S. J., & Janin, J. (1978). Computer analysis of protein-protein interaction. *Journal of Molecular Biology*, *124*(2), 323–342.

Wolff, P., Amal, I., Oliéric, V., Chaloin, O., Gygli, G., Ennifar, E., Lorber, B., Guichard, G., Wagner, J., Dejaegere, A., & Burnouf, D. Y. (2014). Differential modes of peptide binding onto replicative sliding clamps from various bacterial origins. *Journal of Medicinal Chemistry*, *57*(18), 7565–7576.

Wollacott, A. M., Zanghellini, A., Murphy, P., & Baker, D. (2007). Prediction of structures of multidomain proteins from structures of the individual domains. *Protein Science : A Publication of the Protein Society*, *16*(2), 165–175.

Wriggers, W., Chakravarty, S., & Jennings, P. A. (2005). Control of protein functional dynamics by peptide linkers. *Biopolymers*, *80*(6), 736–746.

Xu, D., Tsai, C. J., & Nussinov, R. (1997). Hydrogen bonds and salt bridges across protein-protein interfaces. *Protein Engineering*, *10*(9), 999–1012.

Xu, D., Jaroszewski, L., Li, Z., & Godzik, A. (2014). AIDA: Ab initio domain assembly server. *Nucleic Acids Research*, *42*(W1), W308–W313.

Xu, D., Jaroszewski, L., Li, Z., & Godzik, A. (2015). AIDA: Ab initio domain assembly for automated multi-domain protein structure prediction and domain–domain interaction prediction. *Bioinformatics*, *31*(13), 2098–2105.

Xu, D., & Zhang, Y. (2012). Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins*, *80*(7), 1715–1735.

Xu, J., & Zhang, Y. (2010). How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics (Oxford, England)*, *26*(7), 889–895.

Xue, L. C., Dobbs, D., Bonvin, A. M. J. J., & Honavar, V. (2015). Computational prediction of protein interfaces: A review of data driven methods. *FEBS Letters*, *589*(23), 3516–3526.

Yan, C., Wu, F., Jernigan, R. L., Dobbs, D., & Honavar, V. (2008). Characterization of protein-protein interfaces. *The Protein Journal*, *27*(1), 59–70.

Yan, Y., & Moult, J. (2005). Protein Family Clustering for Structural Genomics. *Journal of Molecular Biology*, 16.

Yang, A. S., & Honig, B. (2000). An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. *Journal of Molecular Biology*, *301*(3), 665–678.

Yang, H., Carney, P. J., Mishin, V. P., Guo, Z., Chang, J. C., Wentworth, D. E., Gubareva, L. V., & Stevens, J. (2016). Molecular Characterizations of Surface Proteins Hemagglutinin and Neuraminidase from Recent H5Nx Avian Influenza Viruses. *Journal of Virology*, *90*(12), 5770–5784.

Yellaboina, S., Tasneem, A., Zaykin, D. V., Raghavachari, B., & Jothi, R. (2011). DOMINE: A comprehensive collection of known and predicted domain-domain interactions. *Nucleic Acids Research*, *39*(Database issue), D730-735.

Yin, H., & Hamilton, A. D. (2005). Strategies for targeting protein-protein interactions with synthetic agents. *Angewandte Chemie - International Edition*, *44*(27), 4130–4163.

Yin, Y. W., & Steitz, T. A. (2002). Structural Basis for the Transition from Initiation to Elongation Transcription in T7 RNA Polymerase. *Science*, *298*(5597), 1387–1395.

Yu, L., Tanwar, D. K., Penha, E. D. S., Wolf, Y. I., Koonin, E. V., & Basu, M. K. (2019). Grammar of protein domain architectures. *Proceedings of the National Academy of Sciences*U S A, 116(9), 3636-3645.

Zacharias, M. (2010). Accounting for conformational changes during protein-protein docking. *Current Opinion in Structural Biology*, *20*(2), 180–186.

Zanegina, O., Kirsanov, D., Baulin, E., Karyagina, A., Alexeevski, A., & Spirin, S. (2016). An updated version of NPIDB includes new classifications of DNA–protein complexes and their families. *Nucleic Acids Research*, *44*(D1), D144–D153.

Zhang, Y., Arakaki, A. K., & Skolnick, J. (2005). TASSER: An automated method for the prediction of protein tertiary structures in CASP6. *Proteins: Structure, Function, and Bioinformatics*, *61*(S7), 91–98.

Zhang, Y., Hubner, I. A., Arakaki, A. K., Shakhnovich, E., & Skolnick, J. (2006). On the origin and highly likely completeness of single-domain protein structures. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(8), 2605–2610.

Zhang, Y., & Skolnick, J. (2004a). Automated structure prediction of weakly homologous proteins on a genomic scale. *Proceedings of the National Academy of Sciences*, *101*(20), 7594–7599.

Zhang, Y., & Skolnick, J. (2004b). Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, *57*(4), 702–710.

Zhang, Y., & Skolnick, J. (2005). TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, *33*(7), 2302–2309.

Zhao, X.-M., Wang, Y., Chen, L., & Aihara, K. (2008). Protein domain annotation with predicted domain-domain interaction networks. *Protein and Peptide Letters*, *15*(5), 456–462.

Zhou, X., Hu, J., Zhang, C., Zhang, G., & Zhang, Y. (2019). Assembling multidomain protein structures through analogous global structural alignments. *Proceedings of the National Academy of Sciences*, *116*(32), 15930–15938.

Zhu, H., Domingues, F. S., Sommer, lngolf, & Lengauer, T. (2006). NOXclass: Prediction of protein-protein interaction types. *BMC Bioinformatics*, *7*, 27.

Zinzalla, G., & Thurston, D. E. (2009). Targeting protein–protein interactions for therapeutic intervention: A challenge for the future. *Future Medicinal Chemistry*, *1*(1), 65–93.

Zmasek, C. M., & Godzik, A. (2012). This Déjà Vu Feeling—Analysis of Multidomain Protein Evolution in Eukaryotic Genomes. *PLOS Computational Biology*, *8*(11), e1002701.