Enriching Protein-protein Interaction Networks with Alternative Conformations and Accelerated Filtering of Dissimilar Interfaces

by

Farideh Halakou

A Dissertation Submitted to the

Graduate School of Sciences and Engineering

In Partial Fulfillment of the Requirements for

the Degree of

Doctor of Philosophy

in

Computer Science and Engineering

January 8, 2020

Enriching Protein-protein Interaction Networks with Alternative Conformations and Accelerated Filtering of Dissimilar Interfaces

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a doctoral dissertation by

Farideh Halakou

and have found that it is complete and satisfactory in all respects,

and that any and all revisions required by the final

examining committee have been made.

Committee Members:

_____

Prof. Attila Gürsoy (Advisor)

_____

Prof. Özlem Keskin Özkaya (Co-Advisor)

_____

Prof. Öznur Özkasap

_____

Assoc. Prof. Mehmet Sayar

_____

Prof. Türkan Haliloğlu

_____

Asst. Prof. Sefer Baday

Date: _____

I would like to dedicate this thesis to my family for nursing me with love and affections, and their dedicated partnership for success in my life.

Abstract

The common practice in structural protein-protein interaction (PPI) networks is to investigate just one specific conformation for each protein. Yet it is not a comprehensive representation as it neglects the conformational changes of proteins which may lead to different protein interactions, functions, and downstream signaling. In this dissertation, a new representation is proposed for structural PPI networks which inspects the alternative conformations of proteins. This representation uses a method to get all available structures of proteins from protein data bank (PDB) and clusters them based on their sequence and structural similarities. Then, it investigates the alternative conformations of each protein.

A large-scale study is done by creating breast cancer lung and brain metastasis sub-networks and equipping them with alternative conformations of the proteins. PPI network analyses showed novel genes and cellular pathways which play important roles in each sub-network. By examining alternative conformations of proteins, the docking results coverage increased from 54% to 76%. In addition, the effects of conformational changes on specific interactions are shown. The conformational changes of KPNB1 directs it to bind to SNAI1 and SNUPN in the open and close conformations, respectively. Exploring the alternative conformations of CXCL12 shows that a point mutation on the binding surface can inhibit CXCL12 homodimerization, as a complex structure, and alter its functions.

PRISM is used for protein docking purposes in the sub-networks. The primary limitations of PRISM are its long running time and restricted interface database. These limitations derive mainly from the all-to-all protein surface and interface structure comparison process which is very time-consuming. So, to speed up this process, a new filtering method is proposed which creates 1D descriptors of protein structures to rapidly filter the dissimilar interface. This method uses a library of non-redundant small protein fragments and Geometric hashing technique to create the protein descriptor vectors. Based on the results, thousands of interface comparisons could be done in seconds using these descriptor vectors. 70%-80% of grossly dissimilar interfaces could be filtered rapidly and the remaining candidate similar interfaces would be compared by an alignment method.

Özetçe

Yapısal protein-protein etkileşimi ağlarındaki yaygın uygulama, her protein için sadece bir spesifik konformasyonu araştırmaktır. Ancak, bu yöntem farklı protein etkileşimlerine, fonksiyonlarına ve aşağı akım sinyallerine yol açabilecek yapısal değişiklikleri ihmal ettiği için kapsamlı bir temsil değildir. Bu tezde, proteinlerin alternatif konformasyonlarını inceleyen yapısal protein etkileşim ağları için yeni bir gösterim önerilmiştir. Bu gösterim, proteinlerin tüm kullanılabilir yapılarını protein veri bankasından elde etmek için bir yöntem kullanır ve bunları sekanslarına ve yapısal benzerliklerine göre kümeler. Daha sonra, her proteinin alternatif konformasyonlarını araştırır.

Meme kanseri, akciğer ve beyin metastazı alt ağları oluşturularak ve proteinlerin alternatif konformasyonları ile donatılarak büyük ölçekli bir çalışma yapıldı. Protein-protein etkileşim ağı analizleri, her alt ağda önemli roller oynayan yeni genler ve hücresel yolaklar gösterdi. Proteinlerin alternatif konformasyonlarını inceleyerek kenetlenme sonuçları kapsamı %54'ten %76'ya yükseldi. Ek olarak, konformasyonel değişikliklerin spesifik etkileşimler üzerindeki etkileri gösterildi. KPNB1'in yapısal değişiklikleri, onu açık ve kapalı konformasyonlarda sırasıyla SNAI1 ve SNUPN'ye bağlanmaya yönlendirir. CXCL12'nin alternatif konformasyonlarının araştırılması, bağlanma yüzeyi üzerindeki bir nokta mutasyonunun karmaşık bir yapı olarak CXCL12 homodimerizasyonunu inhibe edebileceğini ve işlevlerini değiştirebileceğini göstermektedir.

PRISM, alt ağlarda protein yerleştirme amacıyla kullanıldı. PRISM'in temel kısıtlamaları uzun çalışma süresi ve sınırlı arayüz veritabanıdır. Bu sınırlamalar esas olarak çok zaman alan protein yapısı karşılaştırma işleminden kaynaklanmaktadır. Bu nedenle, bu süreci hızlandırmak için, protein yüzeyi ve arayüz yapılarının 1D tanımlayıcılarını oluşturmak için yeni bir yöntem önerilmektedir. Bu yöntem, protein tanımlayıcı vektörleri oluşturmak için fazlalık içermeyen küçük protein parçalarının bir kütüphanesini ve Geometrik karma tekniğini kullanır. Sonuçlara dayanarak, bu tanımlayıcı vektörler kullanılarak binlerce arayüz karşılaştırması saniyeler içinde yapılabilir. Çok benzer olmayan arayüzlerin %70-%80'i hızla filtrelenebilir ve diğer aday benzer arayüzler bir hizalama yöntemi ile karşılaştırılabilir.

Acknowledgments

In the first place, I offer my sincerest gratitude to my supervisors, Prof. Attila gursoy and Prof. Ozlem Keskin for the chance of being a member in their research group, their friendship and support during my PhD.

I would like to thank my thesis committee, professors Turkan Haliloglu, Mehmet Sayar, Oznur Ozkasap, and Sefer Baday for their precious time.

I thank my officemates for their friendship and good memories we had together.

I also thank my parents for their continuous support during every step of my education.

Finally, I thank my beloved husband Arman for his patience, support and encouraging me throughout this journey.

Table of Contents

List of Tables

List of Figures

Nomenclature

PPI	Protein-Protein Interaction
PDB	Protein Data Bank
RMSD	Root Mean Square Deviation
PRISM	Protein Interactions by Structural Matching

Chapter 1

INTRODUCTION

Traditional Protein-Protein Interaction (PPI) networks, which use a node and edge representation, lack some valuable information about the mechanistic details of biological processes. Mapping protein structures to these PPI networks not only provides structural details of each interaction but also helps us to find the mutual exclusive interactions. Yet it is not a comprehensive representation as it neglects the conformational changes of proteins which may lead to different interactions, functions, and downstream signaling. In this thesis study, a new representation for structural PPI networks is proposed which inspects the alternative conformations of proteins.

The study started by creating breast cancer lung and brain metastases sub-networks as the interested phenotype. Topological and functional analyses are done on these sub-networks. Then, the combined network is equipped with different conformers of proteins. The results showed a considerable increase in the protein docking predictions by using alternative conformations. The protein-protein interaction predictions increased from 54% to 76% using the alternative conformations. The benefits of investigating structural data and alternative conformations of proteins are shown through three case studies.

The major limitation in the structural PPI creation is the long running time of the docking methods which show if two/more proteins can bind to each other. The time-consuming step of docking methods is the structural comparison of proteins. To tolerate this problem, a new compact representation is proposed to quickly filter out grossly dissimilar protein structures. This representation converts the 3D structure of proteins to a 1D descriptor vector using a library of small protein fragments and Geometric hashing technique. The descriptor vectors show the frequency of small protein fragments in the library. Using this method, PRISM can remove ~70%

of the dissimilar interfaces for a target protein in less than 1 minute. It can also be used for fast interface comparisons and clustering. So, by comparing PRISM interfaces with newly deployed interface structures in PDB, the PRISM interface database would be enlarged.

This thesis study is arranged in six chapters. Chapter 2 is a comprehensive review of the well-known studies done in PPI networks and protein structural comparisons. Chapter 3 describes the breast cancer lung and brain metastases sub-network creation and the related topological and functional analyses. Then, the new representation embedding the alternative conformations of proteins in PPI networks is explained in chapter 4 including the results and case studies. Chapter 5 designates the new representation for protein structures and shows the results of protein interface/surface structures. Finally, chapter 6 gives a conclusion of the whole thesis study.

Chapter 2

LITERATURE REVIEW

In this chapter, the studies related to protein-protein interaction networks representation are reviewed, and their advantages and limitations are mentioned. Then, the recent online resources in the field of protein conformational changes are reviewed. These online databases provide a valuable repertoire of data about protein shape dynamics. Finally, a comprehensive review of the state of the art methods to compare protein structures are given.

2.1 Protein-protein Interaction Networks

Protein-protein interaction network is a useful tool to analyze the protein interactions in system level. The most common representation of PPI networks is a graph demonstration. In these PPI graphs, nodes represent the proteins and edges represent their interactions. This abstract representation provides a global picture of biological processes and protein function and helps us to simplify complex cellular systems [1].

2.1.1 Structural PPI Networks

To deeply understand functional roles and binding mechanisms of proteins, we need to include an extra piece of information in the PPI networks which comes from structural data. Addition of the structural information to the traditional PPI networks enables us to answer some essential questions in systems biology.

A) The first question would be, how is it possible for some proteins to have tens and even hundreds of interactions in PPI networks? Since proteins have a limited surface area, a single protein cannot interact with such a large number of partners at the same time. Tsai et al. [2] show

that the problem arises from the traditional representation of PPI networks. In these networks, all the protein products of a single gene are mapped into a single node. So, in protein interaction networks, each node represents a collection of proteins, each with a distinct conformation or spliced form. Even though the conformations are not totally different, the small differences suffice to bring new interactions. Nevertheless, the number of interactions can vary from one protein to another. Structural analysis suggests that the essential proteins in the PPI networks have more binding sites than other proteins [3]. On the other hand, it has been shown that the larger the number of partners a hub protein has, the higher the probability of the hub essentiality is [4].

B) The next question would be, which interactions can occur simultaneously and which are mutually excluded? To answer this question, we need to investigate a new dimension in PPI networks which is time [5]. One of the first pioneering works in this field was done by Kim et al. [6]. They mapped the protein interactions to known interface structures by using their sequence similarity. They identified mutually exclusive interactions by inspecting the usage of the same binding interface on their structural yeast interaction network. Similar to this study, Tuncbag et al. [5] used p53 pathway as a case study and predicted the simultaneous and mutually exclusive interactions in their network. Kuzu et al. [7] used structural interaction networks to find simultaneous and mutually exclusive interactions on ERK and MAPK pathways.

C) Another question would be, how can mutations, located on interfaces, surfaces, or buried in the core regions of proteins, affect the PPI networks? It has been shown that mutations located on interaction interfaces of a protein are more likely causing distinct interruptions in the overall interactome [8]. Therefore, they can result in different biological phenotypes. Mosca et al. [9] created a resource named dSysMap which systematically maps disease-related missense mutations on the structurally annotated human interactome. Their study shows that if pairs of mutations are located in different interfaces of the same protein, they usually cause different phenotypes. However, if they are located in the same interface of interacting proteins, they are most likely producing the same phenotype. To predict the effects of mutations on PPI networks, Moretti et al. [10] performed a community-wide assessment of available methods. They computationally designed influenza binders HB36.4 and HB80.3, and then created single point mutant variants corresponding to all 20 amino acids at each position of the binders. They showed that mutations can influence binding if they disrupt the folded state.

2.1.2 Proteins Conformational Changes

The common practice in structural PPI networks is to inspect just one protein structure for each node. However, proteins are flexible [11, 12] and their conformations change based on the factors like post-translational modifications, pH level, and environment. Thus, it is needed to consider an ensemble of protein conformers for each node in PPI networks. The conformational changes dictate the protein function. Therefore, characterization of this association between protein conformation and its function, helps us to understand how to alter and regulate protein activity. The importance of investigating multiple conformers of the proteins in PPI networks lies in the fact that protein conformational changes cause activation of specific pathways [13-15]. A proper example could be RAS-family proteins. RAS proteins switch between GDP-bound inactive form to the GTP-bound active form, as two conformational states, in response to receptor-mediated extracellular signals [16]. They regulate proliferation, differentiation, motility and cytoskeletal re-organization [17, 18]. The conversion of GTP-bound to GDP-bound form is regulated by guanine nucleotide exchange factors (GEFs), and the conversion of GDP-bound to GTP-bound form is mediated by GTPase activating proteins (GAPs) [19]. Active RAS interacts with its effector proteins and activates downstream effectors like Raf kinase, phosphatidylinositol 3-kinase and Ral guanine nucleotide-dissociation stimulator [20]. Schematic representation of RAS conformational states and their corresponding interactions are shown in Figure 2.1.

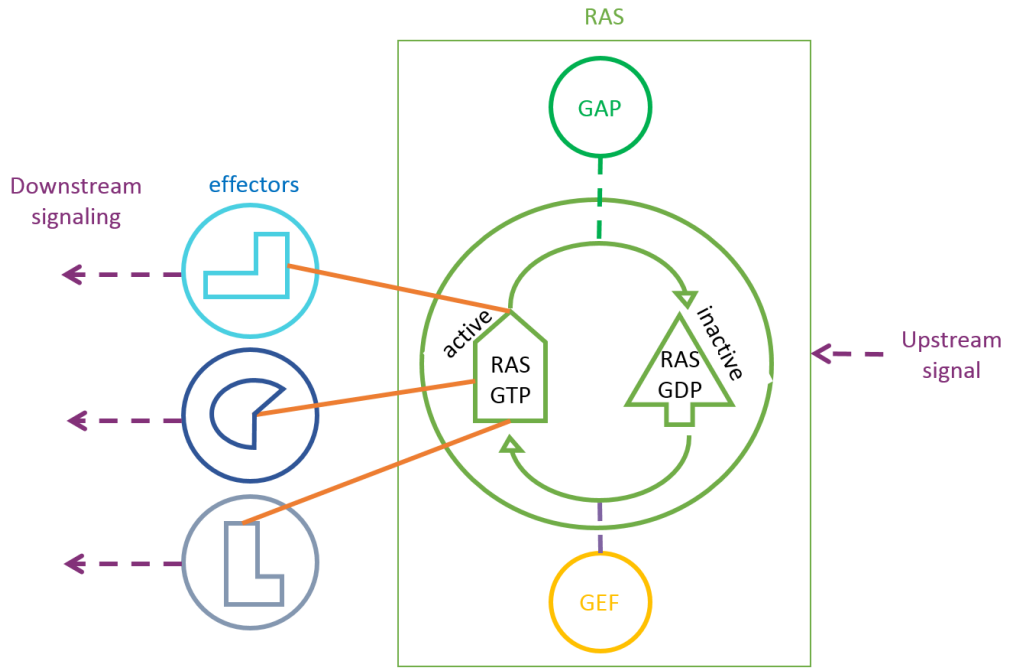

Figure 2.1 Schematic representation of RAS protein conformational states. In the active form, RAS can bind to its effectors and activate different signaling pathways.

RAS example intensifies the investigation of different conformers of proteins in structural PPI networks. Protein Data Bank (PDB) [21] provides protein structures which include alternative conformations of the same protein coming from different experiments. Kuzu et al. [22] used an ensemble of protein conformations to dock the difficult cases of a docking benchmark data set. Applying the alternative conformations of proteins significantly increased the probability of successful protein interaction prediction in their study. They created the network of ERK interactions as a case study. Also, ensemble of protein conformations are used in docking to model small-molecule-protein or peptide-protein interactions [23, 24]. However, ensemble docking studies are neither at network level, nor for protein interactions, they usually make use of computationally generated conformations rather than multiple conformations obtained from PDB.

There are some recent databases which explore the conformational changes and flexibility of protein structures deposited in PDB database such as CCProf[25], PDBFlex[26], and CoDNaS 2.0 [27]. PDBFlex database provides the analysis of structural variations between the different depositions for the same protein in the PDB database. It clusters different structures of proteins based on their sequence similarity and animates its dynamics. CCProf database also offers

conformational changes of protein structures plus some other beneficial biological features like potential binding target site, secondary structure, conservation, phosphorylation site and catalytic site. CoDNaS database covers a large proportion of available protein structures derived from X-ray crystallography and NMR experiments in the PDB database. It clusters the protein structures using their sequence identity too. There are ~15 conformers in average for each protein chain in this database.

2.2 Protein Structure Comparison

Many applications in proteomics, like protein docking methods or function prediction tools, involve comparing the protein structures, whether protein chain, surface or interface. There are different methods for protein structure comparison which can be classified into two categories: a) alignment-based and b) feature-based methods [28, 29]. Following sections describe some of the recent works done in these fields.

2.2.1 Alignment-based Protein Structure Comparison Methods

Alignment-based methods can provide accurate correspondence between the protein structures. These methods, which are usually computationally expensive, can be sequence order-dependent or sequence order-independent which are suitable for protein backbone comparison and protein surface comparison respectively. Pang et al. [30] proposed a sequence-independent method named PBSalign for protein surface comparison. PBSalign gets a pair of protein-protein interfaces, each consists of two binding sites, as input. It generates a triangulated mesh for each binding site of an interface using MSMS program [31]. So, the binding site surfaces are represented as connected non-directed graphs in which nodes characterize a surface point and edges show adjacency of the nodes on the surface. After the graph creation, each node is associated with various surface properties like electrostatic potential and hydrophobicity. Possible correspondences between the two binding site surfaces can be found by using the similarity of these properties. Finally, PBSalign uses MaxSurf to find the maximal overlapping surface of two binding sites while minimizing the RMSD of alignment. The results show that PBSalign accurately

finds the similar binding sites and performs alignments better than the compared protein-protein interface comparison tools.

Konc and Janezi [32] proposed a local structure alignment method, named ProBiS, to detect structurally similar regions on protein surfaces. ProBis can compare complete protein surfaces by using a maximum clique algorithm and dividing graphs into subgraphs. It represents the protein surfaces as graphs in which the residues are shown as vertices labeled with their physicochemical properties. These graphs are then divided into n overlapping subgraphs, where n is the number of vertices. The similarity of subgraphs is defined based on their distance matrices subtraction [33]. A product graph is created for each pair of similar query and database protein's subgraphs to make sure they have sufficiently similar geometrical measures of vertices and physicochemical properties. Then maximum clique is applied to detect the largest common vertex substructure between the two protein graphs. ProBis is used successfully to detect protein–protein, protein–small ligand and protein–DNA binding sites.

MultiProt [34] is another alignment-based protein structure comparison method. The first version of MultiProt was written in 1991 [35] to compare protein structures automatically. Multiprot is not dependent to the protein sequence, and it is not sensitive to insertions or deletions. Instead it uses protein structural information i.e. atomic coordinates and types. To align the protein structures, it uses Geometric Hashing technique [36]. It selects an ordered triplet of noncollinear points to make a plane and uses this plane to compute the coordinates of all other atoms in the protein. It repeats this process for all noncollinear triplets in the protein structure. Each selected triplet defines a rotation model of the protein structure. Two proteins align on top of each other if their atomic coordinates and types match in at least one of their rotation models. Multiprot is applied on protein structures and outperformed other methods, which use additional information for comparison, in finding the alignments.

2.2.2 Feature-based Protein Structure Comparison Methods

Feature-based protein comparison methods exploit shape descriptors for fast comparison of protein structures without explicit alignment of residues. Feature-based methods gain their speed

by representing the protein structures abstractly, typically as vectors, and comparing these representations quickly.

There are studies in pattern recognition field, which represent the protein 3D structures by feature vectors. In these methods, similarity of proteins is measured by comparing the feature vectors. For instance, using the α carbon distance matrix as a feature to represent the tertiary structure of a protein chain is widely used [37-39]. The distance matrix of a protein structure shows the intra-molecular distances of α carbons in a protein chain. As the secondary structure elements are conserved in the distance matrix, it can be used to retrieve the protein tertiary structure. The major benefit of using the distance matrix is that it is translation, rotation and scale invariance because it transforms the protein 3D structure to a 2D matrix.

Karim et al. [39] used distance matrices for fast and accurate retrieval of similar proteins from a large database. They introduced two new features i.e. the co-occurrence matrix of oriented gradient and pyramid histogram of oriented gradient to increase the speed and accuracy of the comparisons. They analyzed the tertiary structure and α carbon distance matrices by representing them as gray-scale images and observed that not all data in the matrix are equally important. The co-occurrence matrix of the oriented gradient of the distance matrix is introduced as the vital feature to compare the protein structures. The method has linear running time and in terms of accuracy it was almost similar to MASASW [40]. There is a tool, called MOMA, for fast and flexible comparison of protein structure chains [41]. This tool relies on the 2D matrices containing angles and distances between the secondary structure elements of the protein. MOMA uses dynamic programming to match a group of substructures to compare the protein structures. Besides the satisfactory speed and accuracy of the method, it has some weaknesses. MOMA is topology-dependent protein alignment method i.e. it depends on the connectivity order of secondary structure elements. In addition, it does not provide the structural alignments at the residue-level as it aligns the secondary structure elements.

There are also some dimension reduction methods which applied to protein surface shapes to find their similarities. Yang et al. [42] applied various dimension reduction methods to map the 3-D protein surfaces onto 2-D maps. They enriched the 2-D surfaces with two biochemical features: electrostatic potential and hydrophobicity. Then to find the similarities of the protein surfaces, they used image registration methods. As proteins can be found in different spatial orientations, an

image registration method needs to be performed to reorient and superpose them before similarities can be found. Based on their results, the functional site of chymotrypsin is depicted by using the 2-D representation. Also, it has been shown to detect structural similarities in heat shock proteins. However, as protein surfaces are more complicated than basic 3D objects, the running time of the most dimension reduction methods increase significantly.

2D methods are aside, there are methods to describe proteins as vectors and compare their 3D structures by comparing their descriptive vectors. An example can be 3D-Surfer tool [38] which compares protein surface shape similarities. It uses a global surface shape representation by using three-dimensional Zernike descriptors, which shows a protein structure as a vector using a series of 3D functions. It gets a protein surface as input and computes a vector of 121 numbers called invariants. This compact representation can be used to search among several thousand protein structures in a short amount of time. Besides, 3D Zernike descriptors are rotation invariant, so protein structures need not be aligned for comparison. FragBag [29] is another method which uses a library of small protein backbone fragments to convert a protein 3D structure to a 1D vector. It describes a protein structure by frequencies of its consisting fixed-length backbone fragments. Fragbag uses RMSD measure to find the similarities between the protein fragments and library fragments. Based on Fragbag results, fragment length 11 could reach better results than shorter fragment lengths.

Pang et al. [28] used a similar idea to compare protein binding sites using a compact representation called visual words representation (PBSword). PBSword is consisted of four steps. First, it selects feature points of binding site surfaces and extract corresponding geometric features. Geodesic distances and angles between the feature points are used as geometric features. Second, a visual vocabulary is built by clustering the feature point descriptors collected from the binding sites. Third, the feature points are associated with their nearest visual words in the visual vocabulary. This allows each binding site to be represented as a vector in which the frequencies of visual words are shown. Finally, the binding sites are compared by a scoring function which calculates the similarity between the two visual word vectors. The experimental results show that PBSword can achieve comparable performance to an alignment-based method i.e. iAlign [43] and it filters out ~80% of the dissimilar binding sites.

Using distances and angles between residues in the protein backbone is a common idea to compare the protein structures. Guyon and Tuffery [44] proposed a shape similarity score, based on Binet–Cauchy kernel named BC score, which is rotation invariant. The BC score is based on the correlation score between protein fragment shapes having the same size. It is normalized between -1 and +1 indicating perfect shape anti-similarity and perfect similarity. The BC score is fast and simple to compute. Based on their results, performing a search by using BC score is >50 times faster than the using RMSD.

2.2.3 Hybrid Protein Structure Comparison Methods

To benefit from the strengths of both alignment-based and feature-based methods, which are precision and efficiency, they can be used together to form an efficient robust protein comparison method. Feature-based methods could be used in a first stage for fast database search and then, only for positive matches, a residue-level structure alignment method could be executed in a sequential manner.

Binkowski and Joachimiak [45] proposed a sequence order independent hybrid method to measure protein surface similarities. They used a three-dimensional object recognition algorithm to assess the surface shape similarities. Then alignment of conserved residues is applied to measure the physicochemical texture similarities. They proposed a metric, named SurfaceShapeSignature (SSS), to describe the global structure of protein surfaces which can rapidly filter grossly dissimilar surfaces. Using SSS reduced the library size by 86% without eliminating any true positives. For the alignment step, they utilized two variants of the RMSD i.e. the coordinate root mean square distance (cRMSD), and the orientation root mean square distance (oRMSD) [46]. They assessed their method to find functionally similar surfaces which bind to different ligands, surfaces with different functions which bind the same ligand, and binding surfaces of ubiquitous and conformationally flexible ligands.

Wang et al. [47] proposes a new protein surface descriptor, named local average distance (LAD), which works based on either the geodesic distance (GD) or Euclidean distance (ED) for pairwise flexible protein structure comparison. At first, LAD profile of each protein structure is formed by averaging the distance values inside a sliding window which scans the protein structures

from N- to C-terminus. Then, local alignments of the profiles are used to find the similarity of protein structures. LAD descriptors have shown to be effective in finding deformed structures and insensitive to large bending motions. The method was compared with seven structural alignment methods and seven shape descriptors. Based on the results, it outperformed all other methods in finding the similar structures.

Chapter 3

ANALYSES OF PROTEIN-PROTEIN INTERACTION SUB-NETWORKS OF BREAST CANCER LUNG AND BRAIN METASTASES

Breast cancer is the second leading cause of cancer death in women and metastasis has been known to cause majority of the breast cancer deaths [48, 49]. Metastasis is a multistage process which occurs by the spread of malignant cells from primary tumor to distant organs [50]. Breast tumors are most commonly metastasized to the lung, liver, bone and brain [51]. The distribution of metastases is not a random process. Each tumor type has apparent characteristics of metastasis through specific organs, which is not exactly revealed [52]. The metastasis process does not occur because of a change in expression, mutation or loss of function of only one gene/protein. It occurs because of intracellular complex relationships of many proteins/genes which are biologically related to each other [53]. Network analysis of organ-specific breast cancer metastasis is one of the promising methods to identify novel gene markers and cellular pathways [54]. Engin et al. [55] generated protein interaction sub-networks for brain and lung metastasis of breast cancer. They created protein interaction sub-networks by using seed genes for corresponding metastasis profiles and ranked each interaction based on their relation to the seed genes by using a guilt-by-association approach. Here, the brain and lung metastasis sub-networks of breast cancer are generated by using the same seed genes as in [55]. However, the sub-networks are filtered to remove the functional interactions to get a network significantly enriched in direct protein-protein interactions. The topological and functional analyses are done by using significantly bigger sub-networks to get more reliable results.

Figure 3.1 shows the steps followed in this study to generate the metastases sub-networks of breast cancer to do different analyses on them.

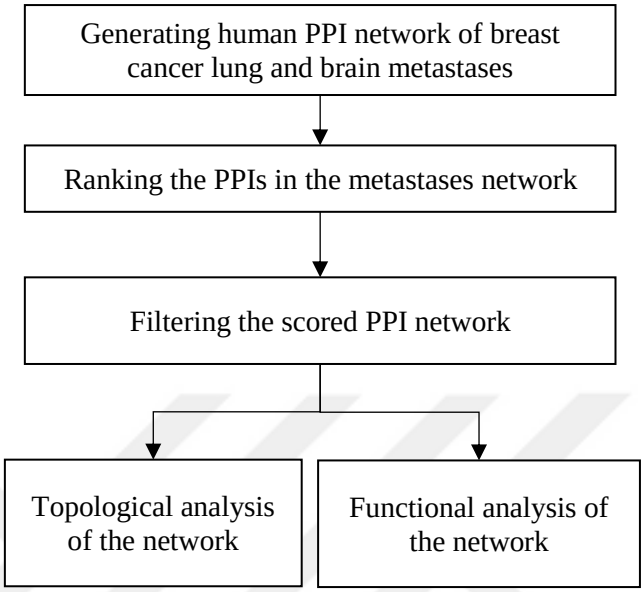

Figure 3.1 Flowchart of the process followed for generating the metastases network and applying the analyses

3.1 Generating Metastasis PPI Sub-networks

Advances in systems biology over the past years have provided a large amount of experimental high-throughput data on protein interactions. This information is spread across multiple databases with some overlap. In addition, given the variations in the experimental techniques used, each database has a different level of reliability for its data. Therefore, a tool is required to unify these data independently of the identifiers used in each database. The PPI network for breast cancer lung and brain metastasis is constructed by using GUILDify webserver [56]. This webserver uses BIANA [57] to construct the network and scores the nodes by using GUILD [58], which is based on network-based prioritization algorithms. GUILDify webserver first maps the seed genes, which are previously known to be associated with a specific phenotype, into a genome wide human interaction network provided by BIANA and scores other genes based on their association with the seed genes by using a combination of network-based parameters i.e. NetScore, NetZscore and NetShort [57].

Genes identified by Minn et al. [59] and Bos et al. [60] are used as seed genes, which were also used previously by Engin et al. [55] as well, to generate lung and brain metastasis sub-networks of breast cancer. They identified 18 genes mediating breast cancer to lung metastasis and 17 genes mediating lung cancer to brain metastasis. Table 3.1 shows the seed genes used.

Table 3.1 The genes mediating breast cancer to lung and brain metastases based on [30, 31].

Genes associated with lung metastasis	**Genes associated with brain metastasis**
MMP1	MMP1
RARRES3	RARRES3
FSCN1	FSCN1
ANGPTL4	ANGPTL4
LTBP1	LTBP1
PTGS2	PTGS2
KYNU	SEPP1
TNC	LAMA4
C10orf116	PLOD2
CXCL1	COL13A1
CXCR4	SCNN1A
KRTHB1	RGC32
VCAM1	PELI1
LY6E	TNFSF10
EREG	B4GALT6
NEDD9	HBEGF
MAN1A1	CSF3
ID1	

3.2 Ranking the PPIs in the metastases network

GUILDify scores genes based on their associations with brain and lung metastasis phenotypes. Thus, each node (gene) has two separate scores which lead to two different sub-networks namely BMSN (Brain Metastasis Sub-Network) and LMSN (Lung Metastasis Sub-Network). GUILDify scores were presented only for the genes but not for their interactions (edges). So, the score of the interactions is calculated by averaging the scores of their associated genes. The score of each interaction shows its relevance to the seed genes. GUILDify score varies between 0 and 1, so as the interaction scores. The score close to 1 indicates stronger relevance to a specific phenotype.

The two sub-networks i.e. BMSN and LMSN, produced by GUILDify webserver are merged to have a single network. This network consisted of 12,172 genes and 324,280 interactions coming from various databases including UniProt [61], TREMBL [62], GO [63], OMIM [64], Reactome [65], INTACT [66], MINT [67], DIP [68], and HPRD [69].

3.3 Filtering the Metastasis PPI Sub-networks

The examination of the interacting protein pairs and their affinities from experimental methods includes both direct and indirect interactions. To use only physical protein interactions, the data are filtered by using experimental protein interactions from STRING database [70] with medium confidence score to eliminate bulk network clusters and further computational load in 3D protein interaction analysis. STRING database collects available experimental information coming from primary interaction databases and assigns confidence score by considering both direct and indirect interactions.

Using the STRING database[70], human interaction network, consisted of 165,184 protein interactions, is downloaded. From this network, only protein interactions coming from experimental data with the confidence score ≥ 0.4 (Table 3.2) are considered.

Table 3.2 Human interaction network extracted from STRING Database.

STRING database protein interactions	**Number of interactions**	**Number of proteins**
Whole human protein interactions	4,850,628	20,770
Experimentally proved human protein interactions	618,574	16,571
Experimentally proved medium confidence score human protein interactions	165,184	12,260

After creation of the STRING human interaction network, it is compared with the merged network and the interactions presented in both networks are selected. After this process, the

network diminished to 43,903 interactions and 9,544 genes. Before filtering the network, there were many proteins with extremely high degrees. As an example, FSCN1 had 804 neighbors and it had a mesh-like cluster. However, after filtering the network its degree decreased to only 18 nodes.

To proceed with the topological and functional analysis, the first 1000 interactions having the highest GUILDify scores are selected from each sub-network. GULDify scores of edges range between 0.08-0.55 and 0.04- 0.58 in BMSN and LMSN respectively. There were 95 nodes and 73 edges in common between BMSN and LMSN (Figure 3.2). The merged network contained 1,927 interactions where 1,240 of them have experimental evidence for binding, 497 of these 1,240 interactions are physical interactions and 107 of them come from APMS. By investigating breast tissue gene expressions of the genes presented in the sub-networks, it has been observed that 99% and 87% of genes in BMSN and LMSN are expressed in the breast tissue, respectively.

Figure 3.2 The union of BMSN and LMSN top 1000 edges network. Purple, green and black edges indicate BMSN, LMSN, and common edges between two sub-networks respectively.

3.4 Topological Analysis of Sub-networks

It has been shown that biological networks are scale-free [71] i.e. the degree distribution in these networks follows a power law, $P(k) \sim k^{-\gamma}$. In other words, the majority of nodes in scale-free networks have a few connections to other nodes, while a few nodes are connected to many other nodes. The degree distributions for BMSN and LMSN are shown in Figure 3.3 and Figure 3.4 respectively.

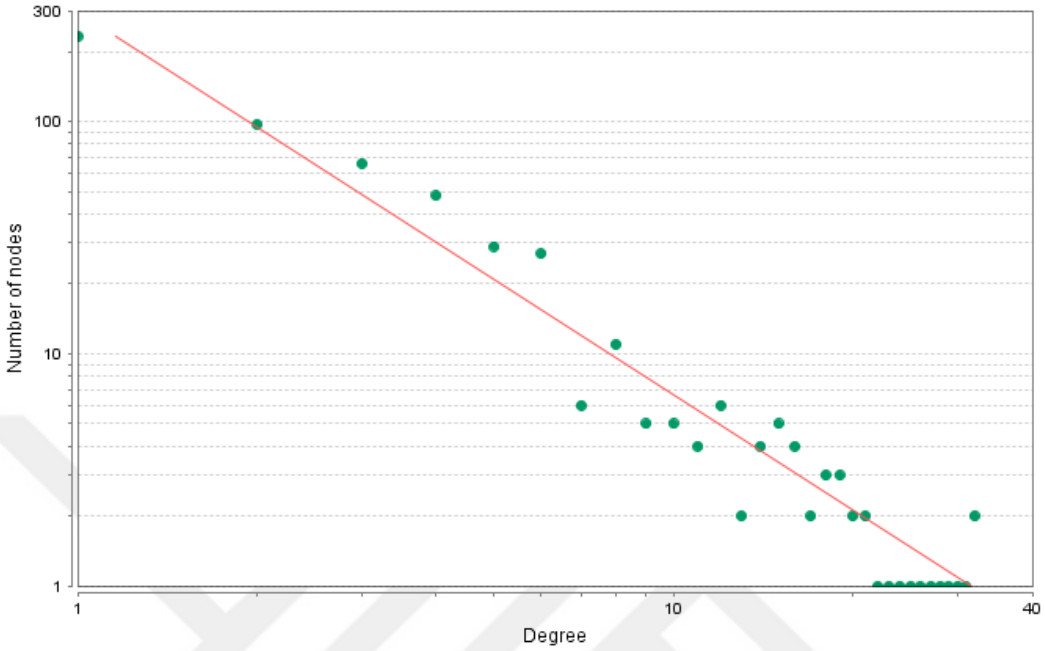

Figure 3.3 Degree distribution of BMSN. The red line shows the fitted power law of the form $y=ax^b$ with a=297.11 and b=-1.646.

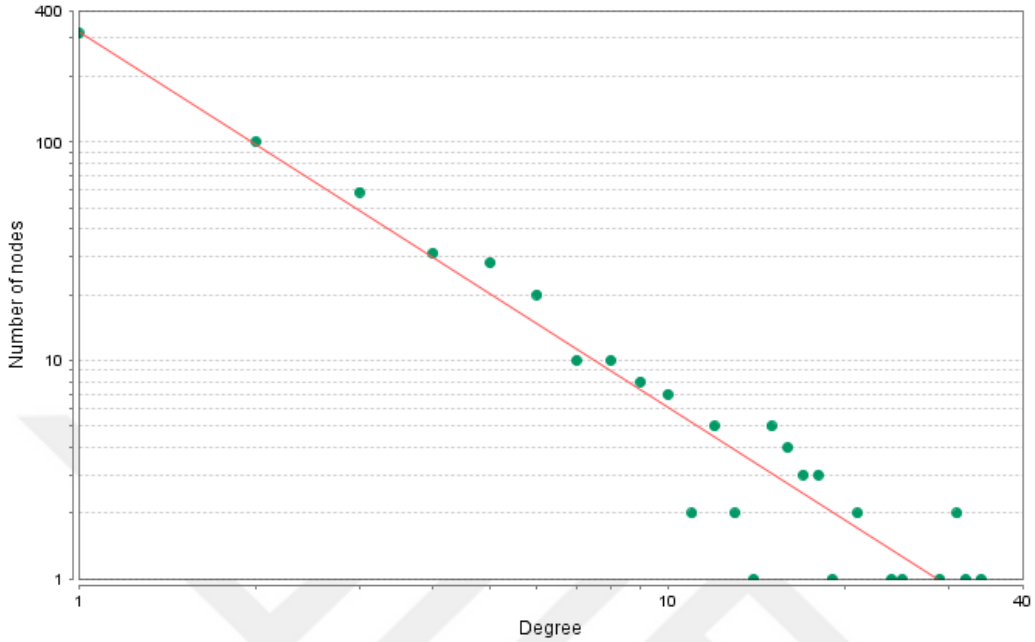

Figure 3.4 Degree distribution of LMSN. The red line shows the fitted power law of the form $y=ax^b$ with a=317.46 and b=-1.716.

The graphs indicate that the sub-networks follow the scale-free networks pattern. Further network parameters for each generated sub-network are shown in Table A.5.

As hub-bottleneck proteins are suggested to be the most important proteins for network cross-talk and biological essentiality [72, 73], the hub and bottleneck nodes are identified in the largest connected component of BMSN and LMSN.

The proteins in sub-networks are classified into three categories: hub-bottleneck, hub-nonbottleneck and nonhub-bottleneck. The average node degree was 3.73 and 3.32 for BMSN and LMSN, respectively. The proteins having more than 19 interactions and betweenness centrality score higher than 0.07 in each sub-network are classified as hubs and bottlenecks, respectively. These thresholds are selected based on the degree and betweenness centrality distributions in the sub-networks (Figure 3.3-6).

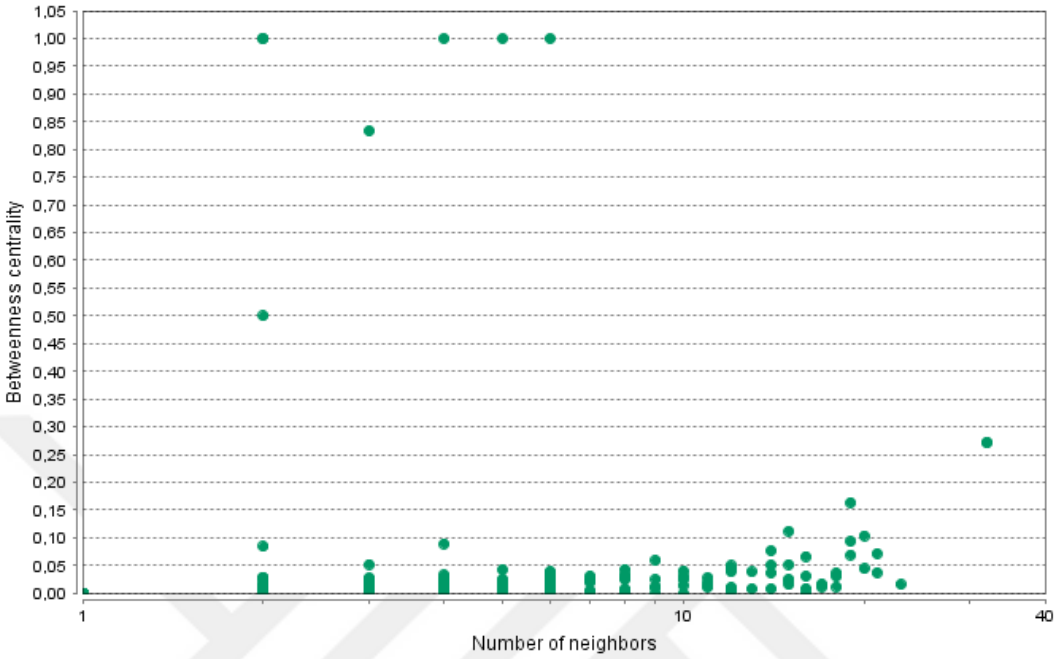

Figure 3.5 Betweenness centrality distribution of BMSN.

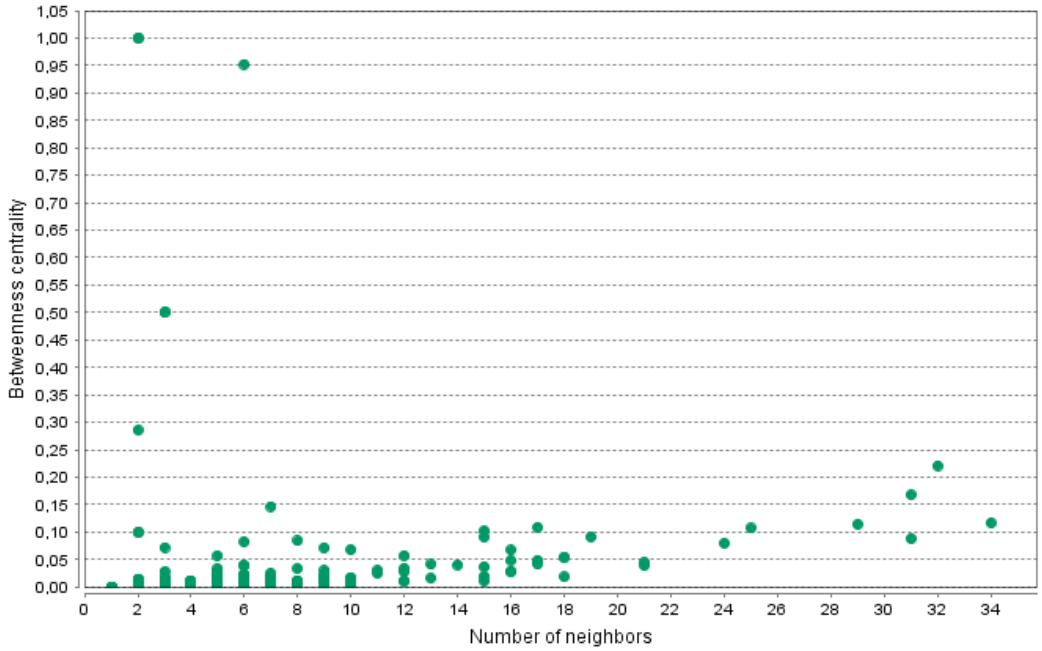

Figure 3.6 Betweenness centrality distribution of LMSN.

In BMSN, there were six hub-bottleneck proteins, four hub-nonbottleneck proteins, and four nonhub-bottleneck proteins as shown in Table 3.3. In this sub-network, proteins encoded by seed genes were not classified as hub-bottleneck proteins. Only two of the proteins encoded by the seed genes were nonhub-bottleneck proteins. The same observation was made for LMSN. In LMSN, seven hub-bottleneck proteins, three hub-nonbottleneck proteins, and seven nonhub-bottleneck proteins are identified as shown in Table 3.4. Only two of the hub-bottleneck proteins and two of the nonhub-bottleneck proteins were also products of the seed genes in LMSN.

Table 3.3 Proteins classified as hub-bottleneck, hub-nonbottleneck and nonhub-bottleneck in BMSN. * represents proteins which are products of the seed genes.

Hub-Bottleneck Proteins	**Hub-Nonbottleneck Proteins**	**Nonhub-Bottleneck Proteins**
CSNK2A1	RPL35	MMP1*
PRKCA	PSMA2	CMA1
RPL5	PRPF19	SCNN1A*
ITGB1	NCBP1	PFN1
ACTA1		
PA2G4		

Table 3.4 Proteins classified as hub-bottleneck, hub-nonbottleneck and nonhub-bottleneck in LMSN. * represents proteins which are products of the seed genes.

Hub-Bottleneck Proteins	**Hub-Nonbottleneck Proteins**	**Nonhub-Bottleneck Proteins**
CXCR4*	CXCR2	CAV1
ELANE	CCR1	ID1*
CCR5	VCAN	IGFBP3
NEDD9*		LTBP1*
DPP4		MYO10
MMP9		ITGB5
MMP2		CXCL12

It is observed that 77% of the proteins identified as hub or bottleneck in LMSN and 16% in BMSN were previously known as specific metastasis-related proteins. The difference between

these two ratios may be derived from the fact that there are plethora of proteins and pathways known to be associated with lung metastasis profile of breast cancer in literature, whereas brain metastasis has many aspects that are still waiting to be revealed [60].

3.4.1 Association of Hub and Bottleneck Proteins to Metastases Subtypes

According to the topological analysis, CSNK2A1, PRKCA, RPL5, ITGB1, PA2G4 and ACTA1 proteins are identified as hub-bottleneck proteins in BMSN. These proteins are already known to have roles in breast cancer proliferation [74-79]. However, only RPL5 (60s ribosomal protein L5) and ITGB1 (Integrin β1) proteins are known to be associated with breast cancer brain metastasis profile [76, 77]. RPL5 protein is required for rRNA maturation and formation of 60S ribosomal subunits. Increased expression of ribosomal proteins causes proliferation and growth of tumors and enhance metastasis capabilities [80]. Also, high expression of ITGB1 increased cell adhesion and metastatic diffusion of tumor cells. PA2G4, Proliferation associated protein 2G4, is an RNA binding protein and involved in growth regulation [81]. This protein is known to be associated with cell growth inhibition and differentiation and known to be related to breast cancer progression [82]. CSNK2A1, casein kinase 2 alpha1 protein, is serine/threonine protein kinase which is highly oncogenic pleiotropic protein that have many different interaction partners [83]. Similar to casein kinase 2, protein kinase C alpha (PRKCA) is also serine/threonine protein kinase that has a diverse role on cellular signaling pathways and highly associated with many tumorigenic and metastatic profiles [84]. The results suggest that both kinases are important for brain metastasis phenotype of breast cancer. Actin alpha 1 (ACTA1) protein also seems to be important for brain metastasis profile. ACTA1 belongs to actin protein families which play a role in cell motility, structure and integrity. It is one of the vital proteins in metastasis since actin cytoskeleton forms major structures in invasion mechanism [85].

In LMSN, CXCR4, ELANE, CCR5, NEDD9, DPP4, MMP9 and MMP2 proteins are identified as hub-bottleneck proteins. CXCR4 and NEDD9 were already associated with breast cancer lung metastasis profile since they were products of the seed genes. Apart from these, Elastase neutrophil expressed protein (ELANE), and matrix metallopeptidase 9 (MMP9), were identified as novel candidates that might be associated with breast cancer lung metastasis. Elastase plays an important

role in tumor microenvironment as it contributes to cytokine and chemokine production, which have vital roles in both inflammation and cancer [86]. Akizuki et al. [87] found that breast cancer patients with poor survival rate have high concentration of ELANE. It seems that this enzyme play an active role on tumor progression that leads to metastasis in human breast cancer. CCR5, Chemokine receptor 5 is a protein involved in immune system and act as receptor to chemokines. It plays an important role in breast cancer cell invasiveness and breast cancer lung metastasis [88]. DPP4, dipeptidyl peptidase-4 is associated with immune regulation and apoptosis [89]. This protein aids lung colonization of breast cancer cells via fibronectin mediated cell adhesion [90]. Both MMP2 and MMP9 belong to zinc-metalloproteinases class of enzymes. They are involved in extracellular matrix degradation. Suppression of Mmp2 in mice models decreases the lung metastasis cases [91]. Also, in breast cancer lung metastasis gene expression profiles of Minn et al. [59], MMP2 was highly expressed. MMP9 was previously shown to be associated with tumor migration of breast cancer cells [92]. Although, the specific association of MMP9 with lung metastasis was not reported. The results suggest that MMP9 might also be related with lung metastasis phenotype of breast cancer.

Nonhub-Bottleneck Proteins also have important regulatory roles in metabolic pathways and have a connector role in PPI networks. In this category, MMP1, CMA1, SCNN1A and PFN1 proteins are identified in BMSN. MMP1 and SCNN1A proteins were already associated with breast cancer brain metastasis profile since they were products of the seed genes. However, CMA1 and PFN1 proteins were not previously identified to be associated with brain metastasis of breast cancer. Chymase (CMA1) is chymotryptic serine proteinase which contributes to extracellular matrix degradation [93]. This enzyme acts indirectly by activating matrix metalloproteinases such as MMP1 which are highly associated with breast cancer metastasis profiles. The results suggest that this enzyme plays an important role in brain metastatic profile. While overexpression of CMA1 might accelerate the metastasis process, overexpression of profiling 1 (PFN1) which is an actin monomer binding protein, might be used to decelerate the metastasis process. This protein was previously identified as tumorigenicity suppressor of breast cancer cells [94]. Thus, PFN1 can be a good candidate to further analysis in breast cancer brain metastasis.

In LMSN, there are also nonhub-bottleneck proteins i.e. CAV1, ID1, IGFBP3, LTBP1, MYO10, ITGA5 and CXCL12. ID1, DNA-binding protein inhibitor, and LTBP1, latent TGF-beta binding protein, are previously identified to be associated with lung metastasis of breast cancer and used as seed genes to generate the sub-networks. All other nonhub-bottleneck proteins, except ITGA5, were also identified to be associated with lung metastasis [59, 95-97]. CAV1, Caveolin 1, contributes to RAS-ERK metabolic pathway and high expression of this gene have tumor suppressor effects in breast cancer and metastasis [98]. Also, in the study of Minn et al. [59], it was identified to have low expression profile in lung metastasis of breast cancer. In LMSN, it was one of the most significant proteins as it binds to many proteins which are products of the seed genes. IGFBP3, insulin like growth factor binding protein 3 has also an important role in lung metastasis. Blouin et al. [95] recently identified that IGFBP3 knock out on mice experiments accelerate growth of the lung metastasis nodules of breast cancer. MYO10 is an actin-based motor protein which is also known as Myosin X. MYO10 is highly expressed in aggressive breast cancer subtypes and has a role in cancer invasion and metastasis. In vivo experiments show that high expression of MYO10 is necessary for extravasation of breast cancer cells into lung [96]. CXCL12, C-X-C motif chemokine 12, is a ligand of CXCR4. Thus, it also plays a significant role in breast cancer metastasis progression [97]. ITGA5, integrin alpha-5 belongs to integrin alpha chain family, which is known to be associated with cell adhesion and cell-surface mediated signaling [99]. This protein is one of the candidates which may be further analyzed to find the exact role in in lung metastasis.

In hub-nonbottleneck protein category, there are RPL35, PSMA2, PRPF19 and NCBP1 proteins in BMSN. Except NCBP1, these proteins were previously associated or predicted to be associated with breast cancer but not exactly with brain metastasis [54, 100, 101]. RPL35, 60S ribosomal protein L35, is a candidate brain metastasis causing protein in BMSN. It is highly expressed in glioblastoma multiform brain tumors [102]. Similar to RPL5, this protein also contributes to proliferation and tumor outgrowth [103]. High expression of RPL35 was observed in male breast cancer patients and it is associated with protein synthesis upregulation [100]. PSMA2, proteasome subunit alpha type 2, is a peptidase that has a role in misfolded protein degradation. Downregulation of PSMA2 was observed in MCF-7 breast cancer cells [101]. Since it is related to degradation of proteins, the increase in concentration of unwanted proteins might

help the metastasis progress in brain. PRPF19, pre-mRNA processing factor 19, suggested to be downregulated in breast cancer cells and brain specific metastasis [54]. PRPF19 plays a role in DNA repair and mRNA splicing. NCBP1 protein, nuclear cap-binding protein 1, is also one of the candidates that might be important in brain metastasis as its association with breast cancer or brain metastasis profile is not known yet.

Similar to BMSN, hub-nonbottleneck proteins are identified in LMSN too. CXCR2, CCR1 and VCAN proteins are recognized as hub-nonbottleneck proteins in the lung metastasis sub-network. The association of CXCR2 and VCAN with lung metastasis is already known [104, 105]. Here the results suggest that CCR1 might also serve as a lung metastasis affecting protein. CXCR2, also known as interleukin 8 receptor beta, is a chemokine receptor associated with lung metastasis like CXCR4. Knockdown experiments identified that this chemokine have a role in breast cancer tumor growth and lung metastasis [104]. VCAN, Versican protein, is a large extracellular matrix proteoglycan. High expression of VCAN causes invasive and aggressive tumor phenotype [106]. Also, expression of Versican was also leading to lung metastasis in mice models [105]. CCR1, Chemokine receptor type 1, was also one of the chemokine family proteins present in the sub-network. Inhibition of CCR1 prevents tumor growth of the breast cancer [107].

Table 3.5 and Table 3.6 list all the candidate genes/proteins which are not previously known to be associated with breast cancer brain and lung metastasis profiles. The results suggest that these genes and their protein products might be important and need to be further investigated by using experimental approaches.

Table 3.5 Candidate genes/proteins related to lung metastasis profile of breast cancer according to topological analyses and literature search

Official symbol	**Functional role**	**Neighborhood ranking**	**Breast cancer gene expression levels**	**Topological importance**
ELANE	Serine proteinase. Modifies functions of natural killer cells. Cytokine/chemokine production affector [86]	1st neighbor of seed gene CXCR4 and VCAM1	Upregulated	Hub-Bottleneck
MMP9	Zinc metalloproteinase. Associates with tumor migration [92]	1st neighbor of seed gene CXCL1	Upregulated	Hub-Bottleneck
CCR1	Chemokine receptor. Proliferation and tumor outgrowth [108]	2nd neighbor of seed gene MMP1	Upregulated	Hub-nonbottleneck
ITGA5	Cell adhesion and cell-surface mediated signaling [99]	2nd neighbor of seed gene MMP1 and CXCR4	Upregulated	Nonhub-bottleneck

Table 3.6 Candidate genes/proteins related to brain metastasis profile of breast cancer according to topological analyses and literature search

Official symbol	**Functional role**	**Neighborhood Ranking**	**Breast cancer gene expression levels**	**Topological Importance**
CSNK2A1	Serine/threonine kinase. Highly associated with tumorigenic metastatic profiles [83]	1st neighbor of seed gene SCNN1A	Upregulated	Hub-Bottleneck
PRKCA	Serine/threonine kinase. Highly associated with tumorigenic metastatic profiles [84]	1st neighbor of seed gene FSCN1	Upregulated	Hub-Bottleneck
ACTA1	Belongs to Actin protein family. Plays role on cell motility and in invasion mechanism [85]	1st neighbor of seed gene FSCN1	Upregulated	Hub-Bottleneck
PA2G4	Associated with proliferation. Increase growth regulatory signals [81]	2nd neighbor of seed gene FSCN1	Upregulated	Hub-Bottleneck
RPL35	60S ribosomal protein. Protein synthesis upregulation and tumor outgrowth [103]	3rd neighbor of seed gene SCNN1A	Upregulated	Hub-Nonbottleneck
PSMA2	Functions as a peptidase. Have a role on misfolded protein degredation [101]	3rd neighbor of seed gene SCNN1A	Downregulated	Hub-Nonbottleneck

NCBP1	Have a role in Pre-mRNA splicing, translation regulation [109]	2nd neighbor of seed gene FSCN1	No Info	Hub-Nonbottleneck
CMA1	Serine Proteinase. Important function on extracellular matrix degredation [93]	1st neighbor of seed gene MMP1	Upregulated	Nonhun-Bottleneck
PFN1	Actin monomer binding protein. Tumorigenicity suppressor [94]	2nd neighbor of seed gene FSCN1	Upregulated	Nonhun-Bottleneck

3.5 Functional Analysis of the Sub-networks

Functions of the proteins in the sub-networks are analyzed to address particular signaling pathways, predicting phenotypes and finding target points for repressing or provoking certain response mechanisms in each metastasis sub-network. Engin et al. [55] observed that infectious diseases and immune system was more important in breast cancer lung metastasis, while in brain metastasis, transport and catabolism was more vital.

As shown in Figure 3.7, the functional analysis revealed that especially chemokine signaling pathway and cytokine-cytokine signaling interactions are significant for lung metastasis. It was previously observed that chemokines play important role in metastasis and tumor progression [110]. Moreover, CCR5 (C-C chemokine receptor type 5), CXCR2 (C-X-C chemokine receptor type 2), and CCR1 (C-C chemokine receptor type 1) were also topologically important genes in the sub-networks which are related to chemokine signaling pathway. Chemokines play a key role in regulating the immune response [108]. Therefore, the potential relationship between immune response and lung metastasis can be interpreted.

Regulation of ribosome synthesis is an important factor in cell growth regulation [111]. It was observed that pathways related to genetic information processing such as translation and transcription are essential for the breast cancer brain metastasis. It means that growth control might be the main target to form brain metastasis. Based on the results shown in Table 3.3, especially

RPL5 (Ribosomal protein L5), RPL35 (Ribosomal protein L35), and PSMA2 (proteasome subunit alpha type-2) are topologically significant for BMSN, and they are classified in pathways related to the genetic information processing.

The percentages of the observed KEGG classes in the sub-networks are shown in Figure 3.8 and Figure 3.9. According to the functional analysis, a major relationship between the lung metastasis and pathways related to the infectious diseases is detected. Also, such a relationship is observed between the brain metastasis and pathways related to the genetic information processing i.e. translation.

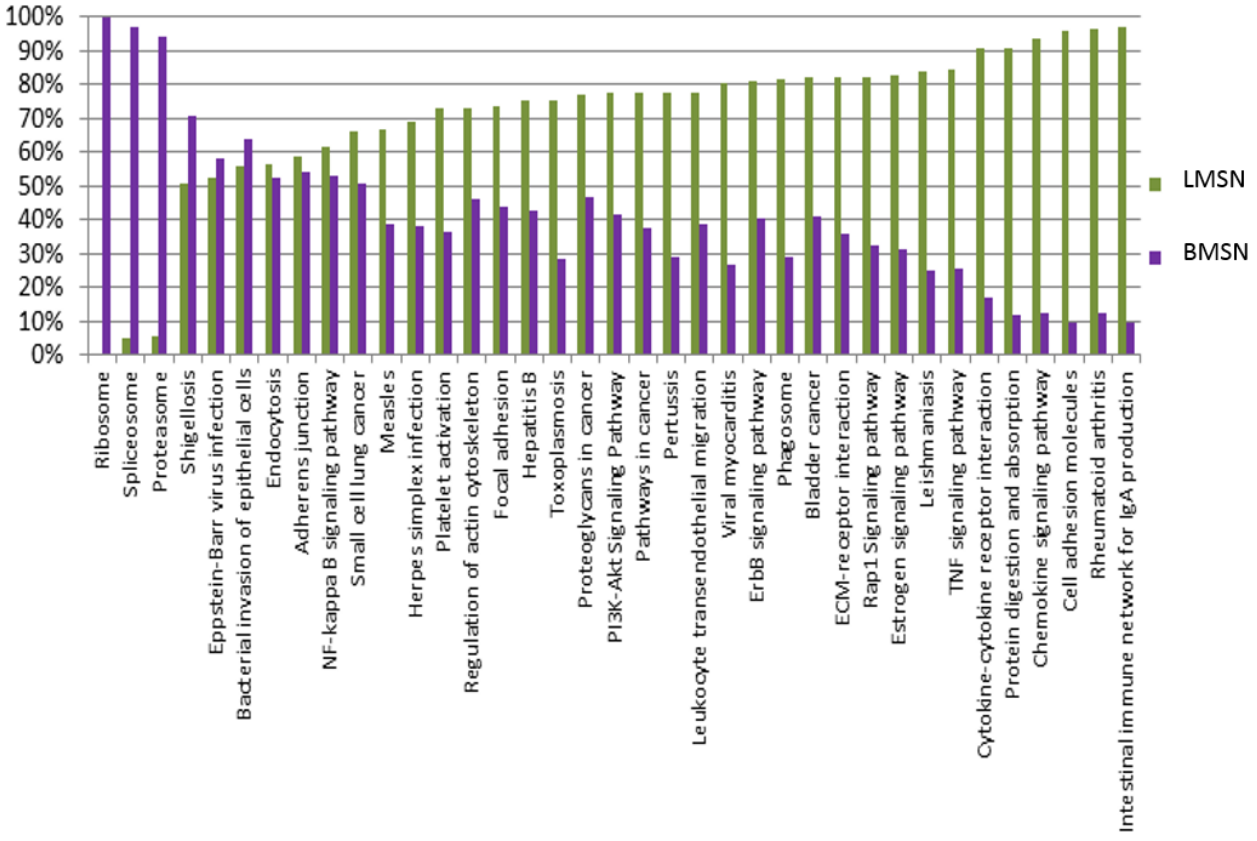

Figure 3.7 The most significant KEGG pathways for LMSN and BMSN sub-networks (pV<=0.05)

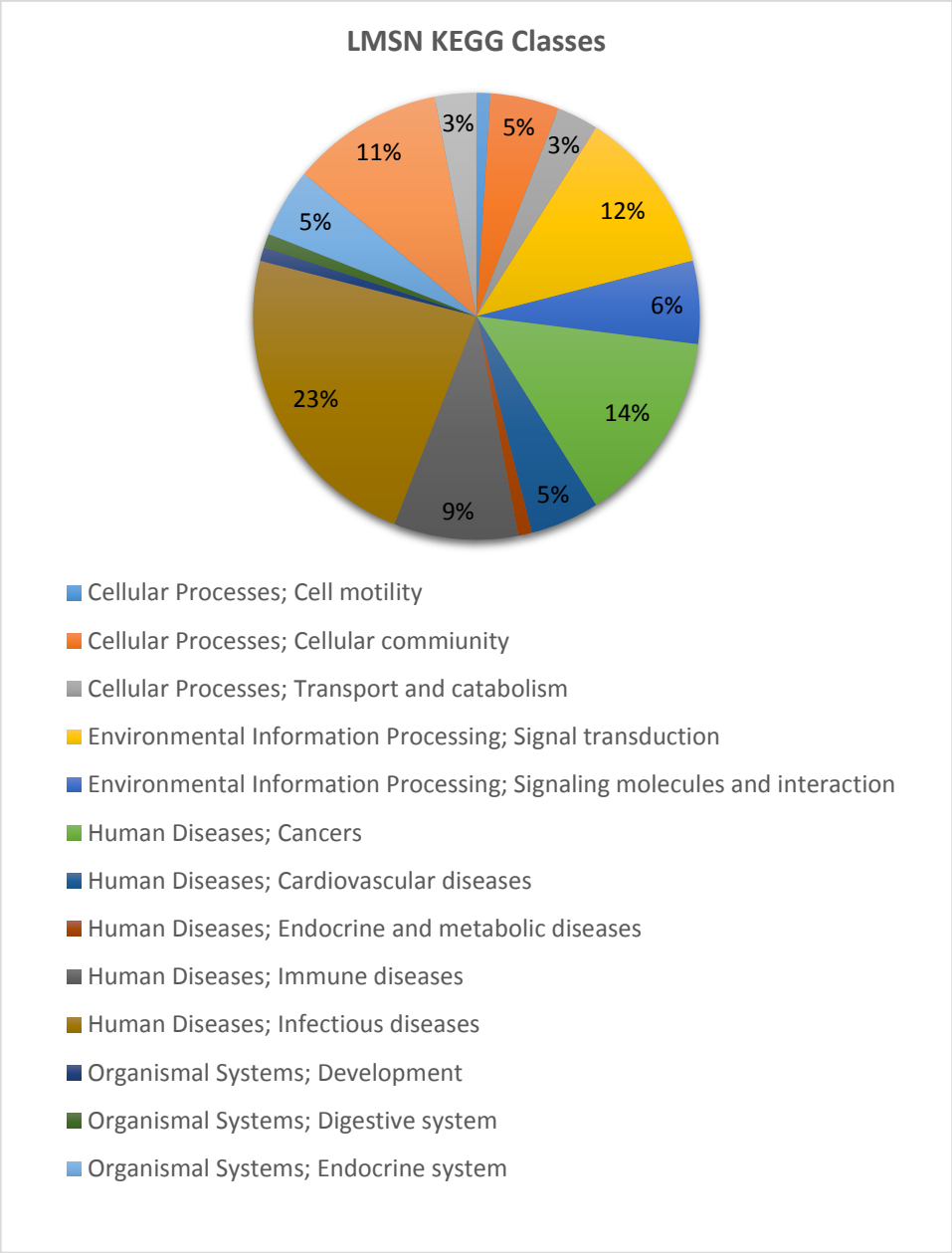

Figure 3.8 The percentages of the observed KEGG classes in LMSN.

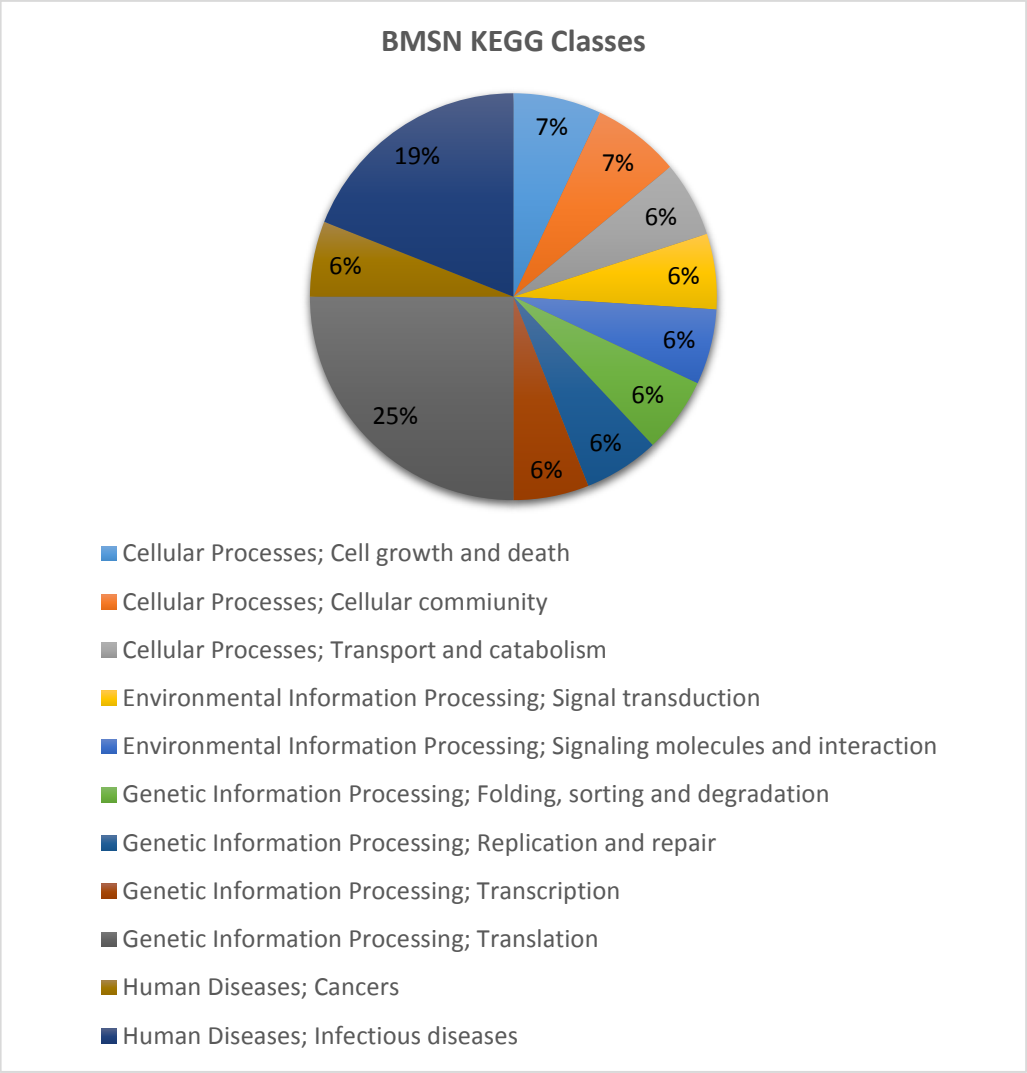

Figure 3.9 The percentages of the observed KEGG classes in BMSN.

3.6 Concluding Remarks

Breast cancer is the second most common cause of death among women. However, it is not deadly if the cancerous cells remain in the breast. The life threat starts when cancerous cells travel to other parts of body like lung, liver, bone and brain. So, most breast cancer deaths derive from metastasis to other organs. In this study, novel proteins and cellular pathways that play important roles in brain and lung metastases of breast cancer are introduced using PPI networks.

The topological analysis identified genes such as RPL5, MMP2 and DPP4 which are already known to be associated with lung or brain metastasis. Additionally, four and nine novel candidate genes that are specific to lung and brain metastases are found, respectively. The functional analysis showed that KEGG pathways associated with the immune system and infectious diseases, particularly the chemokine signaling pathway, are important for lung metastasis. On the other hand, pathways related to genetic information processing were more involved in brain metastasis.

Chapter 4

ENRICHING TRADITIONAL PROTEIN-PROTEIN INTERACTION NETWORKS WITH ALTERNATIVE CONFORMATIONS OF PROTEINS

In this chapter, the proposed new PPI representation is explained to have a more realistic view of PPI networks (Figure 4.1).

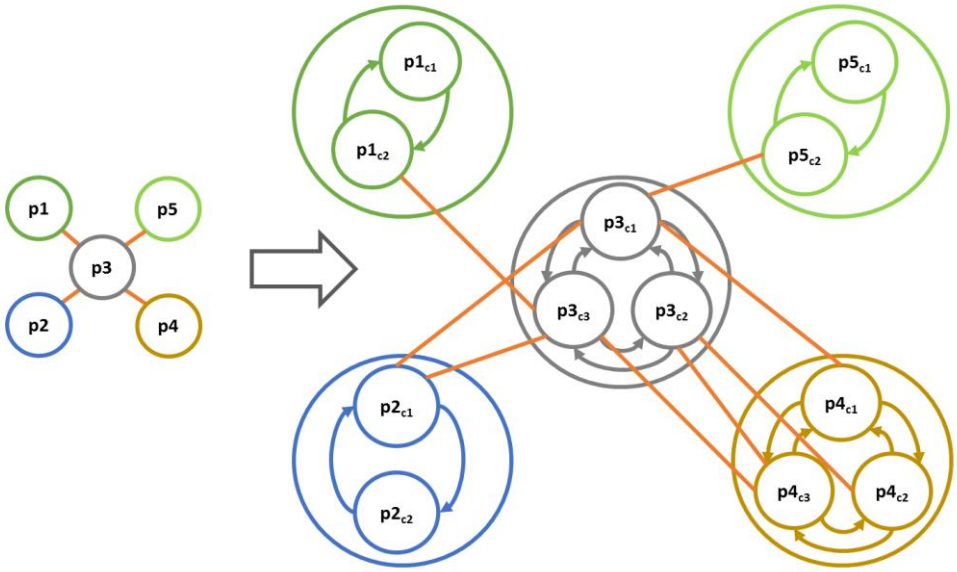

Figure 4.1 Structural PPI network in traditional representation on the left, and the new representation investigating proteins conformational changes on the right. Arrows show inter-conversions between alternative conformations of the same protein.

This representation covers proteins' alternative conformations and shows their specific interactions with their binding partners. In this representation, each protein p in the network is an ensemble of that proteins states/conformations $\{p_{c1}, p_{c2}, \dots p_{cn}\}$ which can inter-convert to each other. If p1 and p2 have N and M alternative conformations respectively, the edge {p1, p2} shows that at least one $\{p1_{ci}, p2_{cj}\}$ $i \leq N$, $j \leq M$ is physicochemically possible. This new representation reveals two important facts about PPI networks. First, each specific protein conformer can contribute to the number of potential interactions the protein can have. In Figure 4.1 for example,

if p3 is in conformation 2 ($p3_{c2}$), it can only bind to p4 if p4 is in its c2 or c3 conformations. However, if p3 is in conformation c3, it may bind to p1, p2, or p3. Second, each edge in the PPI network has a strength based on the present conformations of the proteins in the network at any time point. Namely, some edges can be formed more probably as the corresponding interacting proteins can bind to each other using different conformations. For example, edge {p3, p5} can be speculated to be less probable than {p3, p4}, as the latter can happen regardless of which conformation p3 is in.

To create the multi-conformational PPI network, the steps shown in Figure 4.2 is followed.

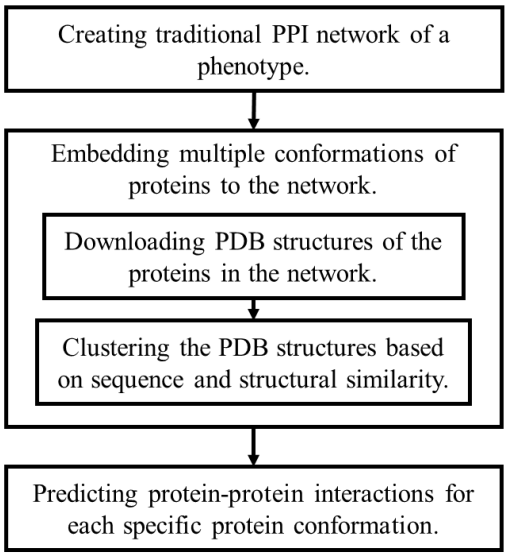

Figure 4.2 Flowchart of creating multi-conformational PPI network.

The process starts with creating a traditional PPI network for the phenotype of interest. In this network, nodes show the proteins and edges show the protein-protein interactions. The merged network of breast cancer lung and brain metastases is used in this study. Afterwards, available PDB structures of each protein in the network are mapped to it. Protein structures are all stored in PDB database [21]. As it is a redundant database, mapping all structures without any curation can cause unnecessary repetition in the 3D protein-protein interaction predictions. Thus, similar PDB structures which correspond to the same protein should be eliminated. For each protein, its alternative conformations are clustered according to their sequence and structural similarities. Then, the possible protein-protein interactions using each specific protein conformation are investigated using a docking method. PRISM is used in this study for interaction prediction because unlike many docking methods designed just for a binary interaction, PRISM can get a PPI

network as the input which makes it suitable for large scale studies. The graphical illustration of the multi-conformational PPI network creation procedure is illustrated in Figure 4.3.

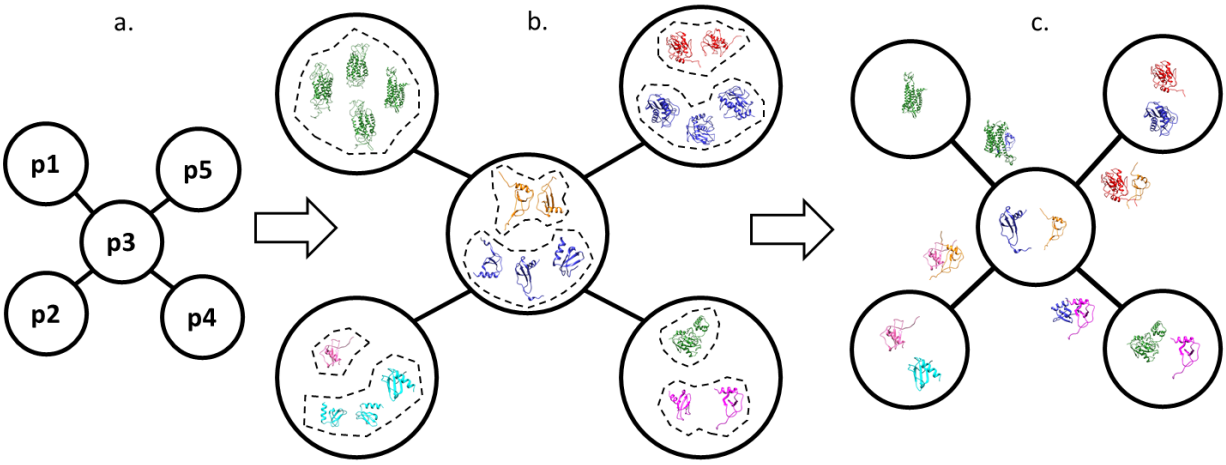

Figure 4.3 Graphical illustration of creating multi-conformational PPI network. a. Traditional PPI network. Each node represents a protein. b. Available PDB structures for each protein. Similar conformations for each protein, indicated with the same color, are clustered together. c. From each cluster, just one PDB structure is selected as its representative. Two proteins may bind to each other by using specific conformations.

After clustering the protein structures according to their similarities, one or more structures remain as their clusters' representatives for each protein. To predict if two connected proteins in the network can bind to each other or not, all permutations of their representative structures are explored. As explained in the following sections, some interactions in the network are feasible only by using specific conformations of the interacting proteins. This fact may change the topology of the network at each specific time point as each protein can form different interactions based on its conformation in the network. This is one of the major points that traditional PPI networks are not able to show.

4.1 Clustering PDB Structures to Form Alternative Conformations

Protein structures are clustered using their sequence identity and structural similarity. Then, a representative structure for each cluster is chosen showing an alternative conformation of the same protein in the network. To map proteins to their known 3D structures, UniProt Knowledgebase [61], which provides a list of corresponding available structures in PDB, is used. In this step, protein structures which have less than 30 residues are eliminated. Afterwards, all available

structures for each UniProt ID are structurally aligned using TM-align tool [112] on top of each other to quantify their structural differences. In order to find the unique protein structures, agglomerative complete linkage clustering method is applied based on sequence similarity and RMSD value of the protein structures. If two protein structures, corresponding to the same UniProt ID, have more than 95% sequence identity and smaller than 2 Å RMSD value, they will be put together in a cluster. This procedure is applied until there is not any protein structures left without an assigned cluster information. Representative structures for each cluster are selected based on the best intra-cluster similarity and RMSD value. These representative structures are the alternative structures of nodes corresponding to proteins in the merged structural network. It should be noted that it is possible that some clusters represent different parts of the same protein because PDB structures do not necessarily cover the complete protein sequence. Since they all correspond to different parts of the protein, they still will contribute to some specific interactions in the PPI network. Therefore, all representatives are kept in the network.

4.2 Mapping Protein Conformations to Network

UniProt Knowledgebase[61] is used to map the proteins to their known 3D structures. The protein structures corresponding to the same UniProt ID are clustered according to their sequence identity and structural alignment using the clustering algorithm described in section 4.1. Figure 4.4 shows the clustering statistics of the proteins in the sub-networks.

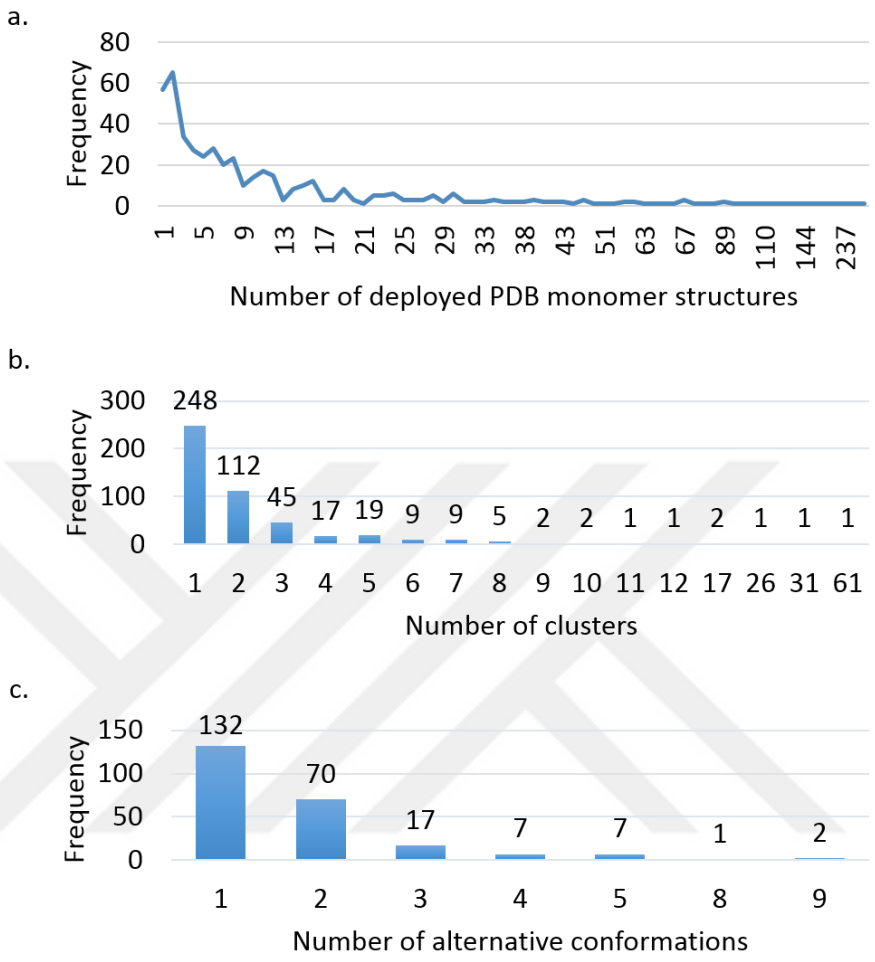

Figure 4.4 PDB database and clustering statistics. a. Number of deployed monomer structures for each protein in the PDB database. b. Number of different clusters for each protein based on the clustering algorithm. These clusters may represent different parts of the same protein or alternative conformations of it. c. Number of the alternative conformations each protein has in the network.

Figure 4.4.a represents the number of monomer structures deployed for each protein in the PDB database. As can be seen from the figure, ~90% of the network's proteins have at least two available monomer structures in the PDB database. HBA1 (hemoglobin subunit alpha 1) is the protein with maximum number of available structures in the network with 439 PDB monomer structures. Figure 4.4.b shows the number of clusters for each protein based on the clustering algorithm. As can be seen from the figure, ~50% of the proteins has just one cluster meaning that all their PDB structures shows the same conformation of the protein. After clustering, HBA1's PDB structures grouped in just 3 different clusters. In this step, APP (amyloid beta precursor protein), with 237 PDB monomer structures, had the maximum number of clusters with 61 clusters. As mentioned earlier, PDB structures of a protein may cover different parts of the protein. So, if a

protein has 3 clusters based on the clustering algorithm, it does not necessarily mean that it has three alternative conformations in the PDB database as those clusters may represent different overlapping/non-overlapping parts of the protein. Two clusters are considered overlapping if they share two or more residues. Therefore, to discover the exact number of alternative conformations each protein has, the protein clusters are mined to find the proteins having at least two clusters, to find the different PDB structures covering the same part of the protein. Figure 4.4.c shows that ~22% of the proteins in the network has at least two alternative conformations saved in the PDB database. So, the remaining 28% of proteins having at least two clusters, in fact have different regions of proteins saved in PDB database instead of alternative conformations. APP (amyloid precursor protein) and CALM1 (calmodulin 1) have the maximum number of conformations with nine alternative conformations in the network. CALM has 163 PDB monomer structures grouped in 31 clusters.

Figure 4.5 shows the merged structural network in which node size represents the node degree in the network and the color of nodes correspond to alternative conformations, i.e. the darker the color is, the more alternative conformation it has. As can be seen from the figure, high degree nodes in the network are all light pink except CXCR4 (C-X-C chemokine receptor type 4) which is a small structure consists of a random coil. Although this observation may seem counterintuitive, it can be explained as follows: the definition of alternative conformation in this study considers large conformational changes for the structures belonging to the same part of a protein, however small local conformational changes in proteins' binding sites may disrupt a specific interaction or promotes another interaction [113].

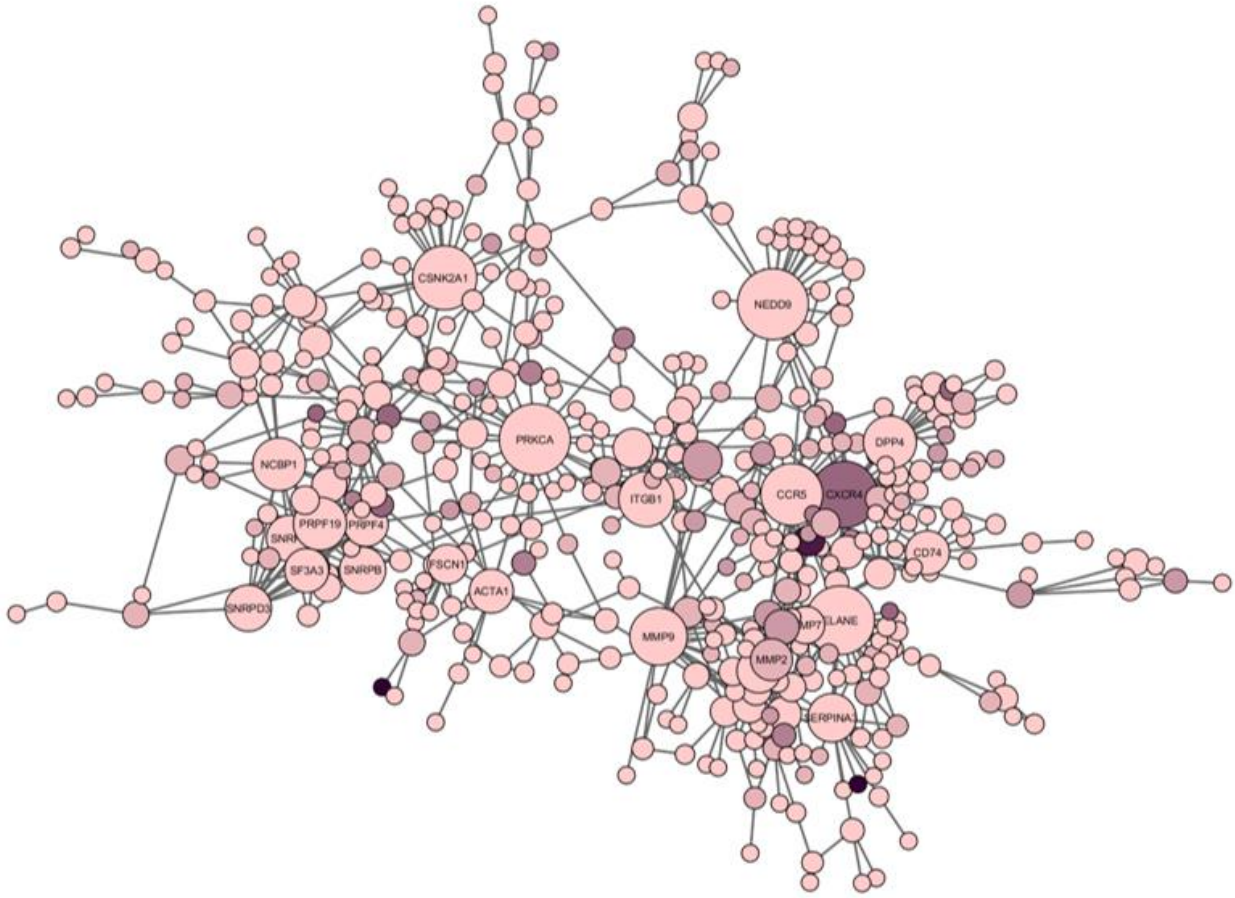

Figure 4.5 Merged structural network. Node size shows the node degree. Node color corresponds to the number of alternative conformations, i.e. the darker the node color is, the more alternative conformation it has.

4.3 Modelling 3D Structures of the Binary Protein Interactions

PRISM webserver [114-116] is used for modelling the binary protein interactions in the structural BMSN and LMSN. PRISM uses a template interface set to predict binding sites of two interacting proteins. Its prediction algorithm investigates the surface regions of two proteins and if they are similar to two sides of a crystal or NMR interface structure, it models their complex structure using that template interface. The rationale of using interfaces instead of whole protein structure for docking purposes in PRISM is that similar binding sites can be found even in globally different protein structures [117]. PRISM algorithm consisted of four steps. In the first step, it extracts the surfaces of target proteins. Then, it splits the template interfaces into two sides and using MultiProt [34] it calculates the similarity of each template interface side to the target protein

surfaces in the second step. Multiprot can find the structural similarities of proteins in an order-independent manner which is necessary for surface/interface comparisons. After finding two target protein surfaces which are similar to two sides of a template interface, and making sure that at least one hotspot at each side is conserved, PRISM transforms the surfaces to the template interface to form a protein complex in the third step. Finally, in the last step, FiberDock [118, 119] is used for flexible refinement of the complexes and ranking them based on the binding energy scores. Fiberdock models both backbone and side-chain flexibility. The current PRISM webserver provides 22,604 template interfaces which were extracted from binary protein interactions stored in the PDB database [120].

Not all the proteins in the sub-networks have available 3D structures in the PDB database, so their interactions could not be modelled using the PRISM webserver. As Table 4.1 shows, only ~44% of the proteins in the network have PDB structures.

Table 4.1 Coverages of PDB and PRISM in the sub-networks.

Network	**Original Network**		**Covered by PDB**		**Covered by PRISM**	
	#of nodes	**#of edges**	**#of nodes**	**#of edges[1]**	**#of edges[2] using single conformation**	**#of edges using multiple conformations**
BMSN	567	1000	262	376	186	254
LMSN	622	1000	280	387	224	322
Merged Network	1094	1927	484	720	390	547

[1]If proteins at both ends of an edge have PDB structures. [2]If PRISM has a prediction for that edge.

After clustering the protein structures, all alternative conformations of the interacting proteins are investigated for each interaction/edge in the network to predict if they can bind to each other and how they bind. To do so, more than 4,500 PPIs are submitted to PRISM. PRISM webserver modelled ~76% of the structural interactions submitted to it (Table 4.1). To show the effect of using multiple conformations of proteins in PPI predictions instead of a single conformation, one conformation is selected randomly for each of the proteins in the network and their PPI predictions are checked. As shown in Table 4.1, investigating just one conformation for each protein in the

network diminished the PRISM predictions by ~22%, so the PRISM coverage of the network edges reduced to just ~54%. Therefore, using one conformation for protein docking is not enough and may bring incomplete results.

Equipped with protein structures and their conformations in the network, the importance of investigating protein structures, their conformational changes, and effects of mutations on the protein-protein interactions are presented through the following three case studies.

4.4 Conformational Selection Mechanism of KPNB1

KPNB1 (Importin subunit beta-1) is a member of Karyopherin family which transports proteins into and out of the nucleus. It has been known that unlike importin α, members of importin β family can bind to proteins directly and import them independently [121]. KPNB1 has 62 interactions in the network in which 41 of its neighbors has 3D structures available in the PDB database as shown in Figure 4.6.

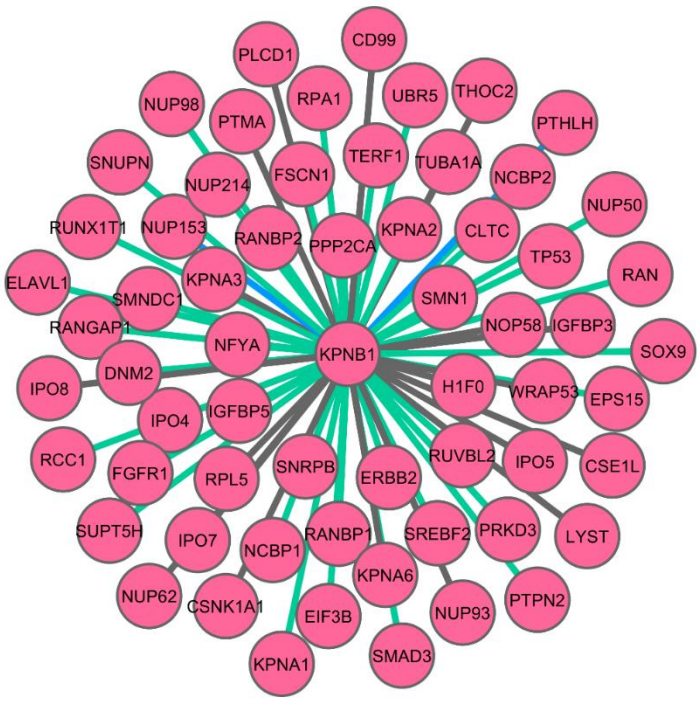

Figure 4.6 KPNB1 cluster in the network. Color of the edges corresponds to PDB and PRISM coverages. Blue edges are covered by PDB, and green edges are covered by PRISM.

KPNB1 is a protein of length 876 and there is its almost complete structure in the PDB database. There are 19 PDB structures for KPNB1 which are clustered to 7 different clusters based on the algorithm i.e. 3w5kA, 1qgrA, 3lwwC, 1qgkA, 1ibrB, 2q5dB, and 2qnaA. Figure 4.7 shows a coarse-grained representation of KPNB1 and its 7 different cluster representatives based on the algorithm.

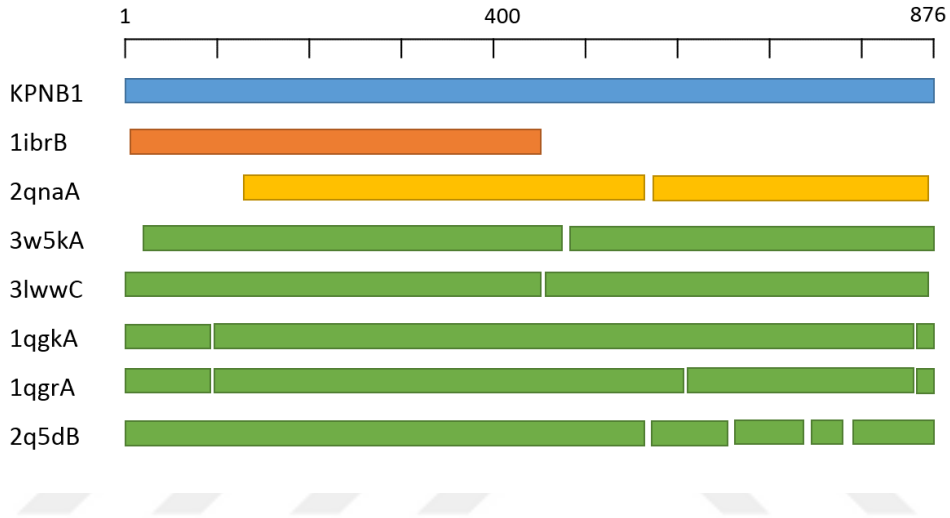

Figure 4.7 KPNB1 protein and its alternative conformations representative structures after clustering all PDB structures of it.

As can be clearly seen from the figure, PDB structures 1ibrB and 2qnaA cover different parts of KPNB1 they are not used as alternative conformations of KPNB1 in the analysis. 3D structures of the five alternative conformations of KPNB1, i.e. 3w5kA, 3lwwC, 1qgkA, 1qgrA, and 2q5dB, are shown in Figure 4.8 and the RMSD values of their aligned structures are shown in Table 4.2. These structures show the open and close conformations of KPNB1.

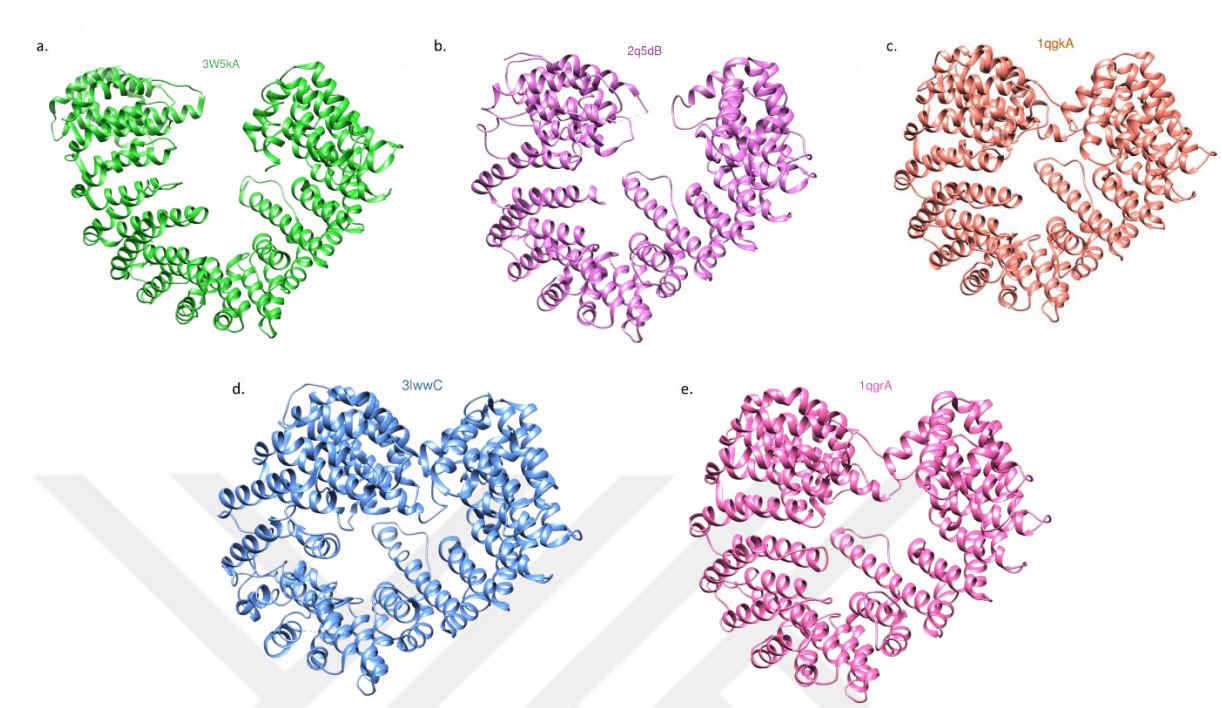

Figure 4.8 Alternative conformations of KPNB1 protein. a) 3w5kA, b) 2q5dB, c) 1qgkA, d) 3lwwC, and e) 1qgrA.

Table 4.2 RMSD values of aligned structures of KPNB1's alternative conformations.

	2q5dB	**1qgkA**	**3lwwC**	**1qgrA**
3w5kA	3.65 Å	3.69 Å	4.33 Å	4.39 Å
2q5dB		3.09 Å	5.05 Å	4.12 Å
1qgkA			4.34 Å	2.87 Å
3lwwC				3.22 Å

All possible interactions of KPNB1 with its neighbors using their alternative conformations are submitted to PRISM. For representation purposes, only 6 different interactions of KPNB1 in the network is investigated. Based on the PRISM predictions, the structural network shown in Figure 4.9 is reached.

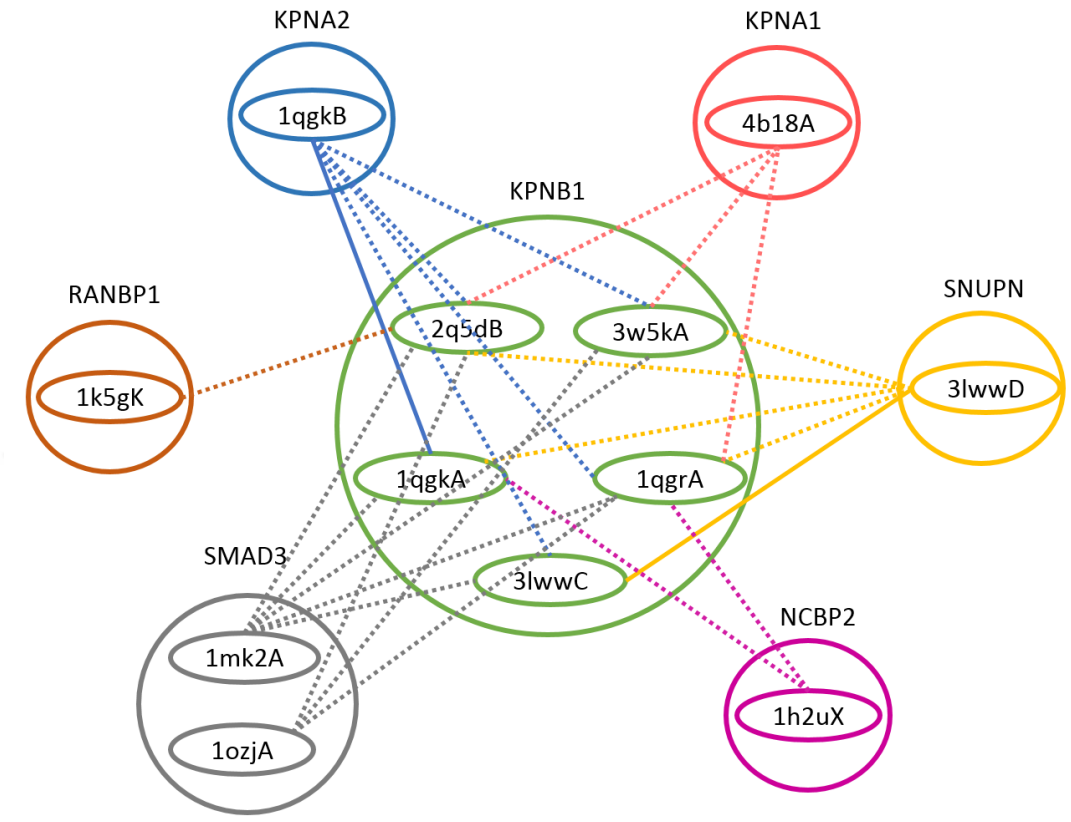

Figure 4.9 KPNB1's existing and predicted interactions by PRISM. Interactions having the complex structures in PDB database are mentioned with solid lines. Dotted lines indicate the PRISM predictions.

The complex structures of interactions KPNB1-SNUPN and KPNB1-KPNA2 are already available in the PDB database namely PDB ids 3lww and 1qgk, respectively. The close conformation of KPNB1 i.e. 3lwwC is bound to SNUPN (Snurportin-1). KPNB1 uses SNUPN adapter to bind cargo molecules to import m_3G-capped U snRNPs [122]. SNUPN has an IBB (importin-β-binding) domain in its N-terminal which is used to bind to KPNB1 and then transport to the nucleus. It has been shown that after each round of import, SNUPN uses CRM1 to return back to the cytoplasm [123]. KPNB1 close conformation 3lwwC bound to SNUPN is represented in Figure 4.10.a. This complex protein structure has the global binding energy of -49.82 based on the Fiberdock webserver [118, 119]. To detect the binding residues in this complex structure, HotRegion webserver [124] is used. The binding residues of this complex structure are listed in Table A1, and they are indicated with opaque colors in the figure. UCSF Chimera package [125] is used for molecular graphics and analyses.

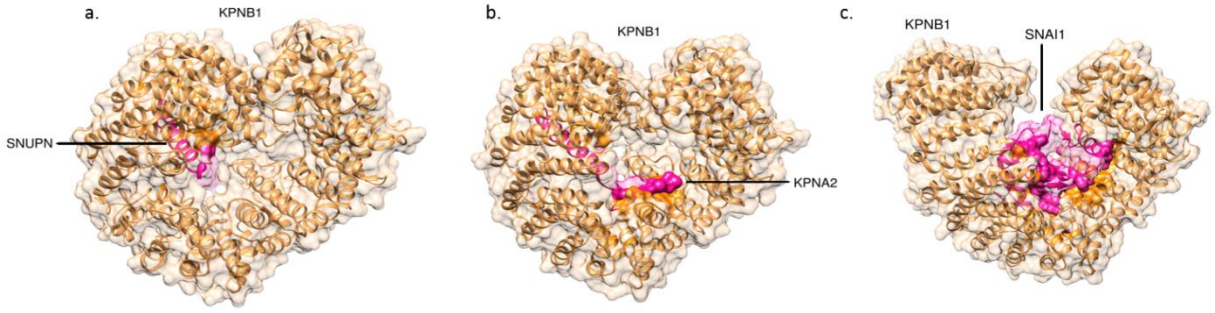

Figure 4.10 KPNB1 complex structures with a. SNUPN (PDB ID 3lwwCD) b. KPNA2 (1qgkAB), and c. SNAI1 (3w5kAB). The binding residues on the surface are indicated as opaque.

Notwithstanding the high RMSD values between open and close conformations of KPNB1, Bhardwaj and Cingolani [126] provided structural evidence that open and close conformations of KPNB1 bind to SNUPN IBB domain. Same results are observed in PRISM predictions as shown in Figure 4.9. It should be noted that, though SNUPN binds to all conformers of KPNB1, it has a different affinity, in terms of binding energy, towards each conformer. For example, the predicted binding energy scores for the interaction of SNUPN with 2q5dB and 1qgkA conformers are -67.38 and -53.32 respectively.

Another existent interaction of KPNB1 is with KPNA2 (PDB id 1qgk). KPNA2 (Importin subunit alpha-1) is an adapter protein for KPNB1 which binds to substrates containing a nuclear localization signal (NLS) motif. Importin β binds to the IBB domain of importin α and forms a heterodimer. Cytosolic proteins having NLS-containing motifs form a complex with this heterodimer and transported into the nucleus by using nuclear pore complexes (NPCs) which are embedded on the nuclear membrane [127, 128]. The available PDB complex structure for KPNB1-KPNA2 uses a KPNB1 close conformer as shown in Figure 4.10.b. The binding residues of this complex structure are listed in Table A2, and they are indicated with opaque colors in the figure. There are 17 residues in common between KPNB1 interacting surface with SNUPN and KPNA2 based on HotRegion webserver [124] shown with boldface in Table A1 and Table A2 which indicates that KPNB1 uses almost the same region to bind to these proteins. This complex protein structure has the global binding energy of -138.40 based on the Fiberdock webserver. However, based on the PRISM results, KPNA2 can bind to other conformers of KPNB1 though with

significantly fluctuating binding energies e.g. -78.25 and -22.32 to bind to 1qgrA and 3lwwC respectively.

Liang et al. [129] showed that KPNA2/KPNB1 canonical pathway is responsible for p65 (RELA proto-oncogene, NF-kB subunit) import to the nucleus. They tested 6 different members of the importin α family and found that KPNA2 plays the major role in p65 import. p65 uses different members of importin β and α families for its import but it has a high affinity towards KPNB1 and KPNA2. Based on their results, p65 hetero trimer is imported to the nucleus through Nuclear Pore Complexes (NPCs). The proteins involved in this pathway are connected to each other in the network, in the order RELA-KPNA2-KPNB1-NUP153. Figure 4.11 shows this pathway in which the interaction KPNB1-KPNA2 is represented with its structural details. Despite the large conformational changes of KPNB1, its binding site to KPNA2 remains almost intact enabling it to bind to KPNA2 in different conformations. It is important to know that using which alternative conformations, the proteins involved in a pathway can bind to each other and pass the signal.

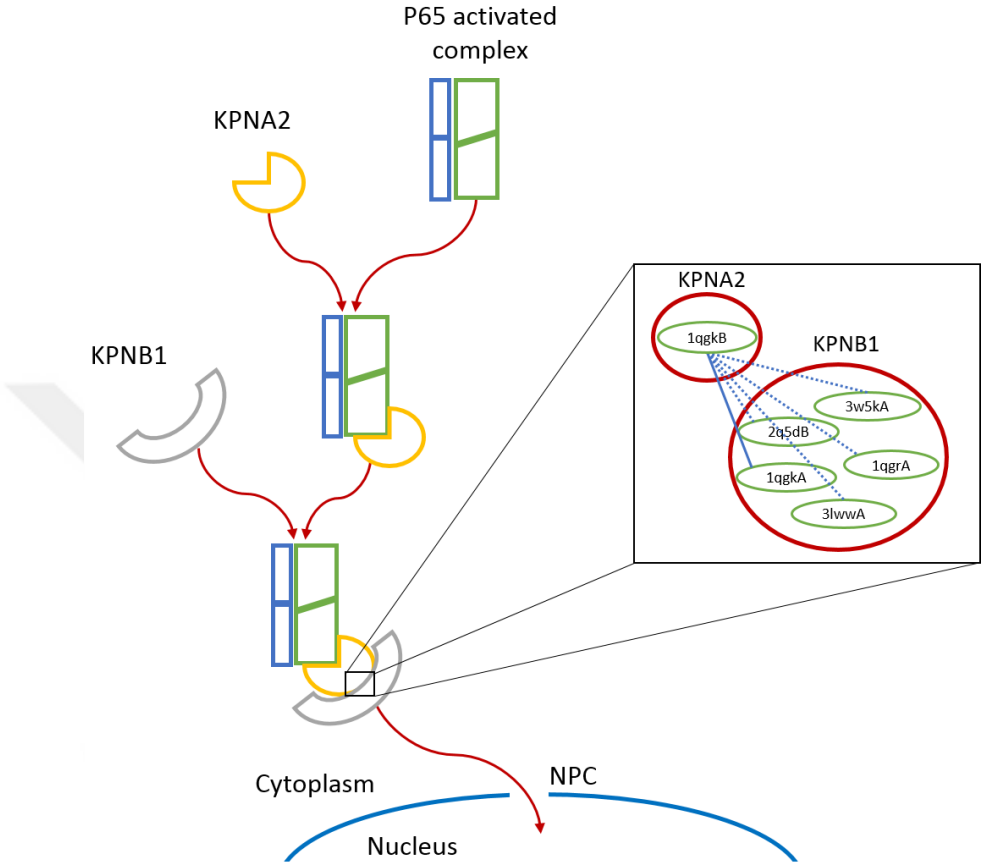

Figure 4.11 Pathway for p65 import based on Liang et al.[129].

Looking at the PDB database, other complexes for KPNB1 are found which complex with RAN, SNAI1, and PTHLH. Among all available structures of KPNB1, only 3w5kA is bound to the zinc finger protein SNAI1. SNAI1 is a transcription factor which has been found in a number of carcinomas and melanomas [130]. Its expression in breast carcinoma is associated with tumor growth and metastasis [131-133]. Besides, in vivo experiments showed that silencing of SNAI1 significantly diminishes tumor occurrence and growth [134].

To see if other conformers of KPNB1 can also bind to SNAI1, all other alternative conformations of KPNB1 with SNAI1 are submitted to the PRISM webserver. Interestingly, PRISM can only find the complex structure using 2q5dB conformer of KPNB1 for these submissions. In this complex, SNAI1 binds to 2q5dB from the same binding site it uses to interact

with 3w5kA. The binding energy score for this complex is -139.17. Therefore, KPNB1 conformers 3w5kA and 2q5dB can bind to SNAI1 and import it into the nucleus. KPNB1 open conformation 3w5kA bound to SNAI1 is shown in Figure 4.10. The binding residues of this complex are listed in Table A3, and they are indicated with opaque colors in the figure. There are 19 residues in common between KPNB1 interacting surface with SNAI1 and KPNA2 based on Hotregion webserver [124] shown with italic typeface in Table A2 and Table A3 which indicates that KPNB1 uses almost the same region to bind to these proteins.

These evidences lead us to infer that some protein conformers are more limited in terms of their binding partners' quantities e.g. KPNB1 can bind to SNUPN or SNAI1 in open conformation though it can't bind to SNAI1 in close conformation. Therefore, each specific protein conformer put a limitation on the diversity of possible binding partners.

4.5 Conformational Changes of CXCL12 Leading to Different Downstream Signaling

CXCL12 (C-X-C motif chemokine 12) is the ligand of CXCR4 (C-X-C chemokine receptor type 4) which is a seed gene in LMSN. This interaction has GUILDify score of ~0.5 which places it in the top 5% interactions in LMSN. It is known that in-vivo inhibition of these two genes reduce breast cancer metastasis progress of lung [135]. Chemokines are small proinflammatory chemoattractant cytokines which bind to specific G-protein-coupled receptors. CXCL12 is expressed in several organs including lung, liver, brain, skeletal muscle, kidney, heart, skin, and bone marrow. The binding of CXCL12 to CXCR4 is known to induce intracellular signaling through several different pathways initiating signals related to chemotaxis, cell survival and/or proliferation [136].

The CXCL12-CXCR4 interaction is involved in tumor progression, angiogenesis, metastasis, and survival. There are efforts to block metastatic dissemination by inhibiting CXCR4 activation [137] to inhibit cancer malignancy [138, 139]. It has been shown that binding to CXCR4 N-terminus (CXCR4 1–38) promotes CXCL12 dimerization [140]. Drury et al. [141] found that oligomeric changes of CXCL12 inducing cellular migration with monomer but not dimer [142]. Interestingly, they have also observed that dimeric CXCL12 exhibited receptor interactions and downstream signaling different from the monomeric chemokine. Their results show that

monomeric CXCL12 activates β-arrestin-2 recruitment and filamentous-actin accumulation. On the other hand, dimeric CXCL12 weakly recruits β-arrestin-2 and diminishes actin polymerization compared to monomeric CXCL12 [143]. This indicates that monomeric and dimeric CXCL12 can have different outcomes in metastasis phenotypes upon CXCR4 binding.

Different signaling outcomes might be related to alternative conformations of CXCL12 upon CXCR4 binding. There are 20 available structures in PDB database for CXCL12 which are divided to 11 clusters based on the clustering method. It is known that conformational changes can be related to point mutations on the protein interfaces which may change their dimerization profile and structural arrangement. There are several mutations known on the CXCL12 protein. One important mutation is H25R substitution at the binding interface which discourages CXCL12 dimerization, thus it is preferentially monomeric [141, 144]. However, some studies have also reported that binding of the CXCR4 N-terminus ($CXCR4_{NT}$) induces H25R mutated CXCL12 ($CXCL12_{H25R}$) dimerization [145]. Here, the homodimerization of $CXCL12_{WT}$ and $CXCL12_{H25R}$, and the interaction of CXCL12 with $CXCR4_{N\text{-terminus}}$ are investigated by using PRISM to better understand the conformational differences and their downstream effects. Figure 4.12.a shows the pairwise superposition of CXCL12WT (PDB ID: 2j7zA) conformation and its H25R mutated conformation (PDB ID: 2kolA). As can be seen from the figure they don't align perfectly on top of each other, especially in N- and C-terminal parts. Besides, there is a significant difference in their surfaces at residue 25 as shown in Figure 4.12.a Based on the clustering method, these two conformations belong to different clusters.

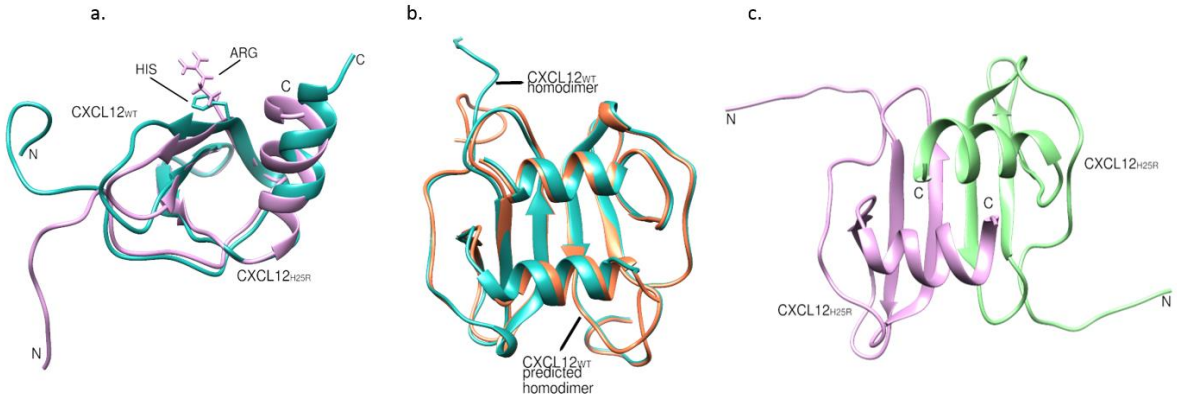

Figure 4.12 CXCL12 monomer and dimer structures. a. Superposition of $CXCL12_{WT}$ (PDB ID: 2j7zA) and $CXL12_{H25R}$ (PDB ID: 2kolA) conformations. The RMSD value of aligned structures is 2.10 Å. b. Superposition of the PRISM prediction for $CXCL12_{WT}$ homodimer (orange) and the available structure in PDB i.e. 2j7z (blue). c. PRISM prediction for $CXCL12_{H25R}$ homodimer.

PRISM webserver predicted the homodimer formation for both $CXCL12_{WT}$, and $CXCL12_{H25R}$. However, in case of the $CXCL12_{H25R}$ the binding energy diminished significantly as expected. The binding energy for $CXCL12_{WT}$ homodimer is -101.41, while it is -17.24 for $CXCL12_{H25R}$ homodimer. The binding residues of these homodimer structures are listed in Table A4. Figure 4.12.b shows the superposition of the PRISM prediction for CXCL12WT homodimer and the available structure in PDB database. There is no structure in PDB database for CXCL12H25R homodimer to compare with the PRISM prediction shown in Figure 4.12.c.

Afterwards the binding of monomeric CXCL12 variants to $CXCR4_{NT}$ is examined. For both interactions, $CXCL12_{WT}$-$CXCR4_{NT}$ and $CXCL12_{H25R}$-$CXCR4_{NT}$, the PRISM has predictions with acceptable binding energies of -87.25 and -52.48 respectively. Tyrosine sulfation is known to promote the protein−protein interactions. There are three potential sulfation sites on $CXCR4_{NT}$ at positions 7, 12, and 21 [145] which PRISM used two of them in its predictions (Table A5). Figure 4.13.a and Figure 4.13.b show the PRISM predictions for interactions $CXCL12_{H25R}$-$CXCR4_{NT}$ and $CXCL12_{WT}$-$CXCR4_{NT}$ respectively, in which CXCR4 surrounded CXCL12.

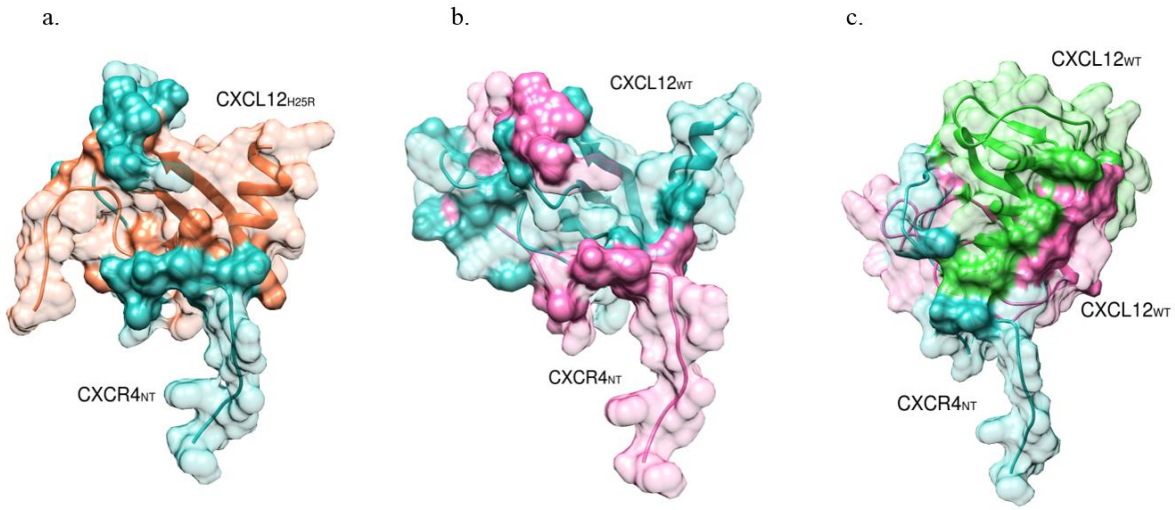

Figure 4.13 PRISM predictions for a. $CXCL12_{H25R}$ (orange)-$CXCR4_{NT}$ (blue) interaction, b. $CXCL12_{WT}$ (pink)-$CXCR4_{NT}$ (blue) interaction, and c. $CXCL12_{WT}$:$CXCR4_{NT}$(pink:blue)-$CXCL12_{WT}$ (green) trimer structure. Binding residues are indicated as opaque.

Based on these results, it can be concluded that CXCL12 binds to $CXCR4_{NT}$ in wildtype and mutated forms. As monomeric CXCL12–$CXCR4_{NT}$ interaction is said to stimulate the dimerization of CXCL12, the interaction of $CXCL12_{WT}$:$CXCR4_{NT}$ heterodimer to $CXCL12_{WT}$, and $CXCL12_{H25R}$:$CXCR4_{NT}$ heterodimer to $CXCL12_{H25R}$ are inspected. Interestingly, the PRISM could only find interaction for $CXCL12_{WT}$:$CXCR4_{NT}$ heterodimer to $CXCL12_{WT}$ with the binding energy of -85.31 and could not predict a binding for $CXCL12_{H25R}$:$CXCR4_{NT}$ heterodimer to $CXCL12_{H25R}$. This shows that binding of $CXCR4_{NT}$ to $CXCL12_{H25R}$ cannot induce homo-dimerization of $CXCL12_{H25R}$ as it may partly occupy the dimerization interface of $CXCL12_{H25R}$. Therefore, a single point mutation can inhibit CXCL12 dimerization which can contribute to metastatic profile and tumor dissemination. Figure 4.13.c shows the PRISM prediction for the trimer structure of $CXCL12_{WT}$:$CXCR4_{NT}$-$CXCL12_{WT}$. Figure 4.14 and Table A6 show the superposition of this predicted complex with available dimer structure of $CXCL12_{WT}$, and list of the binding residues of this complex structure respectively.

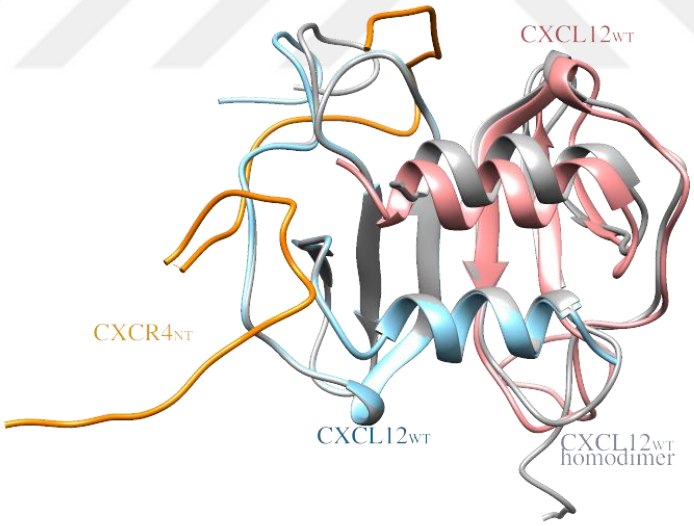

Figure 4.14 Superposition of PRISM prediction for $CXCL12_{WT}$:$CXCR4_{NT}$ (blue:orange)-$CXCL12_{WT}$ (pink) trimer structure and the available $CXCL12_{WT}$ homodimer (grey) structure in PDB i.e. 2j7zAB.

Consequently, as proteins are dynamic and changing conformations based on environmental factors, they favor different binding partners at each time point. Therefore, investigating alternative conformations for each protein is vital in structural PPI networks to have a precise understanding of protein interactions and signaling pathways.

All the interactions for these two case studies are submitted to Zdock [146], GRAMM-X [147], and PatchDock [148, 149] webservers too. The common residues between their predictions and PRISM predictions are listed in Table A7, and Table A8.

4.6 Mutual Exclusive Interactions of NEDD9 with SMAD3 and NCK1

NEDD9 (neural precursor cell expressed, developmentally down regulated 9), SMAD3 (SMAD family member 3), and NCK1 (NCK adaptor protein 1) complex is an example of mutually exclusive interactions in LMSN. NEDD9 is a scaffolding protein, and it was previously identified as the lung metastasis causing gene in breast cancer [60]. It was used as a seed gene in LMSN creation and previously identified as metastatic hub protein in cancer signaling [150]. Figure 4.15 shows NEDD9 cluster i.e. its first neighbors in LMSN. NEDD9 is associated with many cellular processes such as cell adhesion, migration, proliferation, apoptosis and homeostasis [151]. The major role of NEDD9 is to promote assembly of proteins in order to maintain these cellular processes through cell signaling.

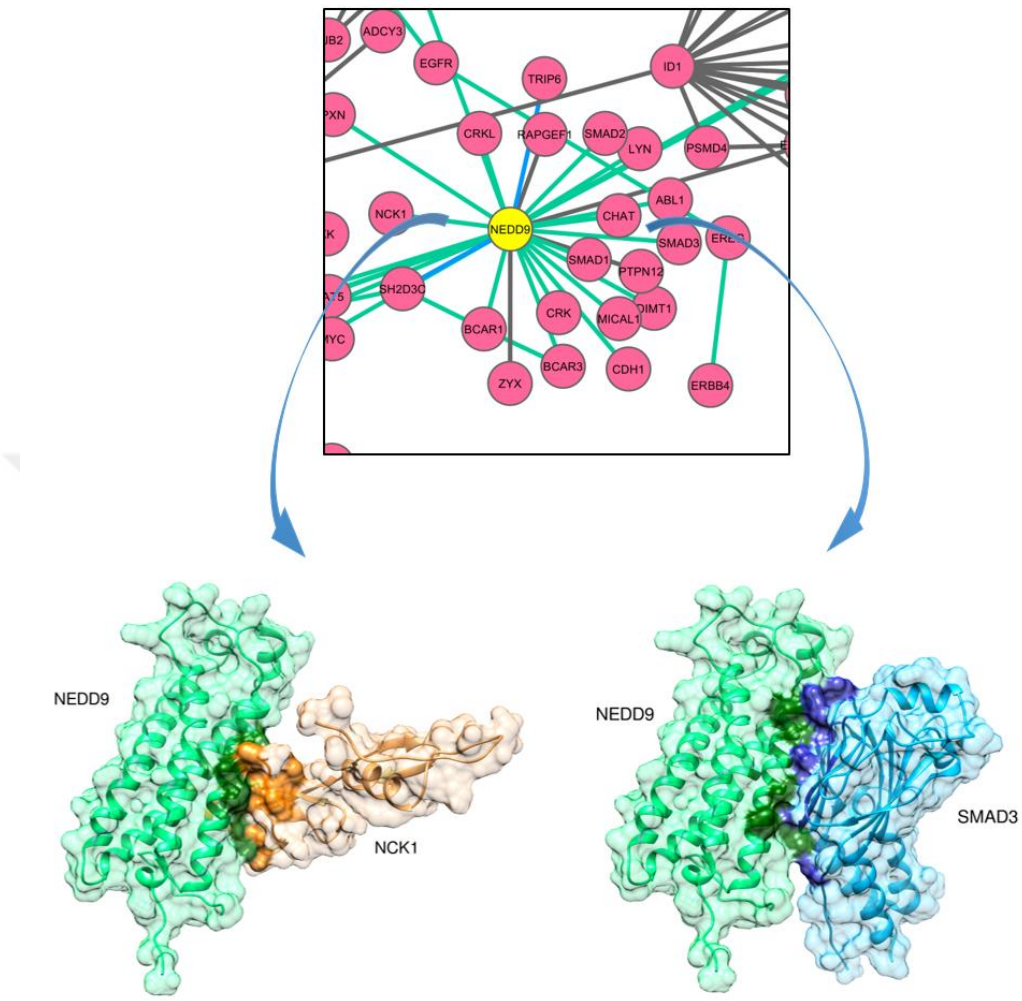

Figure 4.15 NEDD9 cluster in LMSN and PRISM predictions for its interactions with NCK1 and SMAD3. Top) Color of the edges corresponds to PDB and PRISM coverages. Blue edges are covered by PDB, and green edges are covered by PRISM. Bottom) Binding residues are indicated with opaque colors in each complex protein structure.

Both N-terminal and C-terminal sequences of NEDD9 are known to bind to SMAD3 causing rapid turnover of NEDD9 [152, 153]. It has been found that the residues 301-330 of SMAD3 (PDB ID: 1MK2), which forms the MH2 domain, are more important for binding to NEDD9 [152]. NEDD9 has only one structure available in PDB which covers residues 399-563 (PDB ID: 2L81) presenting a serine-rich four helix bundle domain. As shown in Figure 4.15, the PRISM webserver predicted that NEDD9 binds to SMAD3 MH2 domain with binding energy score of -34.29.

On the other hand, PRISM predictions suggest that NEDD9 binds to NCK1 from the same region it binds to SMAD3 (Figure 4.15). The binding energy score of the prediction is -22.46.

NCK1 is cytoplasmic non-catalytic region of tyrosine kinase adaptor protein which facilitates the construction of larger protein assemblies. NCK1 has both SH2 and SH3 domains in its structure [154]. Previously, it was identified that NCK1 binds to NEDD9 by using its SH2 domain [155]. The same result is observed in the PRISM predictions. The predicted interface of NEDD9-NCK1 interaction consists of 26 residues and NEDD9-SMAD3 interaction interface consists of 33 residues. There are 7 residues in common between these interfaces (Table A9).

Based on these results, NEDD9-SMAD3 and NEDD9-NCK1 interactions are mutually exclusive so these two edges cannot be present at the same time in the PPI network. Thus, to achieve the actual topology of a PPI network, inspecting the structural aspects of proteins are necessary. By considering the mutual exclusive interactions, the degree of NEDD9 in LMSN at any specific time point may decrease significantly.

4.7 Concluding Remarks

The main objective of this study was to emphasize the importance of investigating alternative conformations of proteins in structural PPI networks to analyze the protein interactions precisely. This information enable researchers to observe the effects of conformational changes, mutations, and mutual exclusiveness on specific protein interactions and downstream signaling.

To this end, an extensive protein-protein interaction network for breast cancer lung and brain metastasis is built by combining various databases. To map the protein structures into the network, a clustering algorithm is proposed based on sequence and structural similarity of protein structures to prevent repetition in the analyses. PRISM is used to model the 3D structures of the PPIs in the network. However, predicting the whole interactions in the metastasis sub-networks was not possible since not all of the proteins have structures stored in the PDB database.

By using alternative conformations of proteins in the sub-networks, the effects of protein conformational changes on protein interactions and downstream signaling are observed. The conformational changes of KPNB1 which directs it to bind to SNAI1 in the open conformation and bind to SNUPN in the open/close conformation are inspected. It shows the limitations posed by the conformational changes of proteins which fluctuates their ability to stablish various interactions. This is an example of conformational change effects exist in the PDB database. As

another case study, the interactions $CXCL12_{WT}$-$CXCR4_{NT}$ and $CXCL12_{H25R}$-$CXCR4_{NT}$ are examined. Based on the results, $CXCL12_{H25R}$ and $CXCL12_{WT}$ can bind to $CXCR4_{NT}$, and this interaction has been shown to direct the migration of metastatic breast cancer cells to specific tissues. However, binding to $CXCR4_{NT}$ inhibits $CXCL12_{H25R}$ dimerization which is not the case for $CXCL12_{WT}$. Dimerization of CXCL12 as the ligand of CXCR4 is important as it stimulates intracellular calcium flux but fails to activate F-actin polymerization or β-arrestin 2 recruitment which has been shown to cause cellular idling that blocks the metastatic tumor formation.

Investigating the structural models, the mutual exclusive interactions are detected. As a case study, NEDD9 mutually exclusive interaction partners are examined. PRISM predicted that NEDD9-SMAD3 and NEDD9-NCK1 interactions are mutually exclusive. This difference can cause activation of different pathways which lead to different metastasis phenotypes. It is known that NEDD9-SMAD3 interaction causes proteolytic cleavage of NEDD9 which induces cell detachment, apoptosis and individual motility of metastasis cells. However, exact role of NCK1-NEDD9 binding was not previously examined in detail.

Chapter 5

ACCELERATED FILTERING OF DISSIMILAR INTERFACES

As explained in the previous chapter, PRISM is used for structural network creation. PRISM has some limitations which restricted the study. The major limitation of PRISM is its long running time which forced the study to focus on a small portion of the sub-networks for structural analysis. This problem derives mainly from the interface search step of PRISM which tries to find the similarity between target protein surfaces and the template interfaces. Another limitation of PRISM is its template interface dataset, which includes interfaces coming from PDB complexes till 2012. Since PRISM predictions are mainly based on these template interfaces, even a new interface structure may change the docking results for many PPIs. So, as new protein complexes are deployed to PDB, there is a need to combine these newly-found interaction interfaces to the PRISM database. They may/may not be similar to the existent PRISM template interfaces. The major problem in updating the PRISM template interface library, in a frequent manner, is again the structural comparison of newly found interfaces with the existent ones which is very time-consuming. The PRISM algorithm is shown in Figure 5.1.

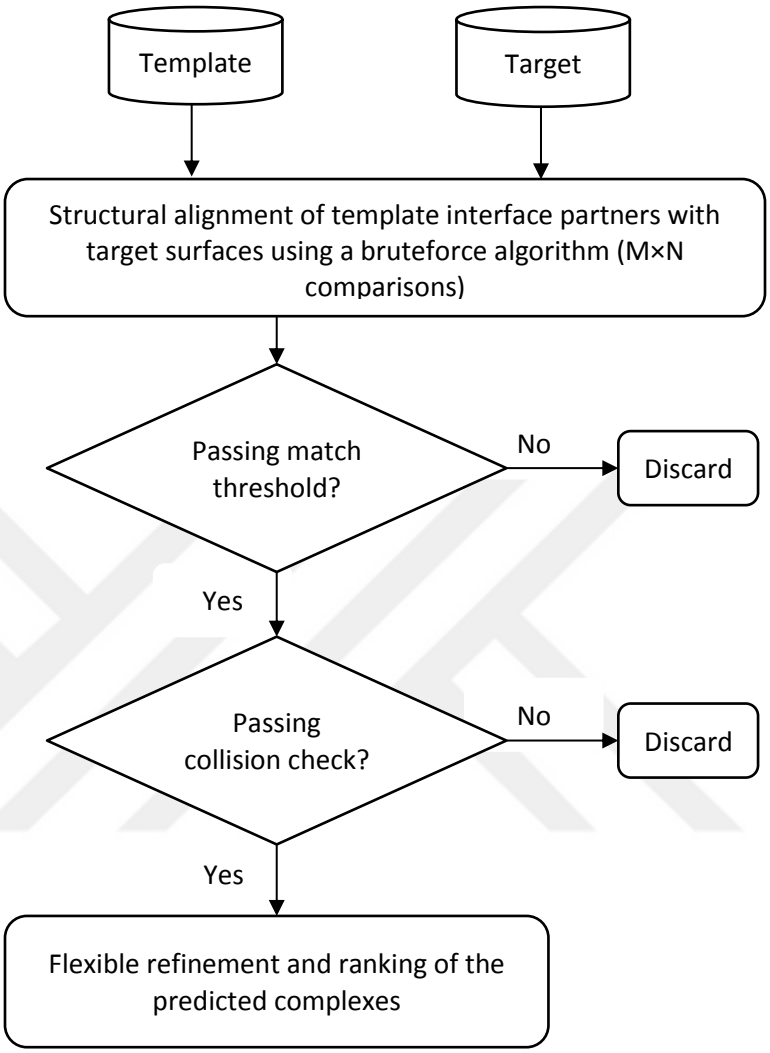

Figure 5.1: Flowchart of the PRISM algorithm.

One of the computationally expensive steps of PRISM is the structural alignment of template interface sides with target protein surfaces. This step splits each template interface into two partners and aligns the structure of each partner with the target surfaces using the MultiProt engine [34]. As Figure 5.2 shows, it is a many-to-many comparison problem which uses structural alignment for each comparison in PRISM.

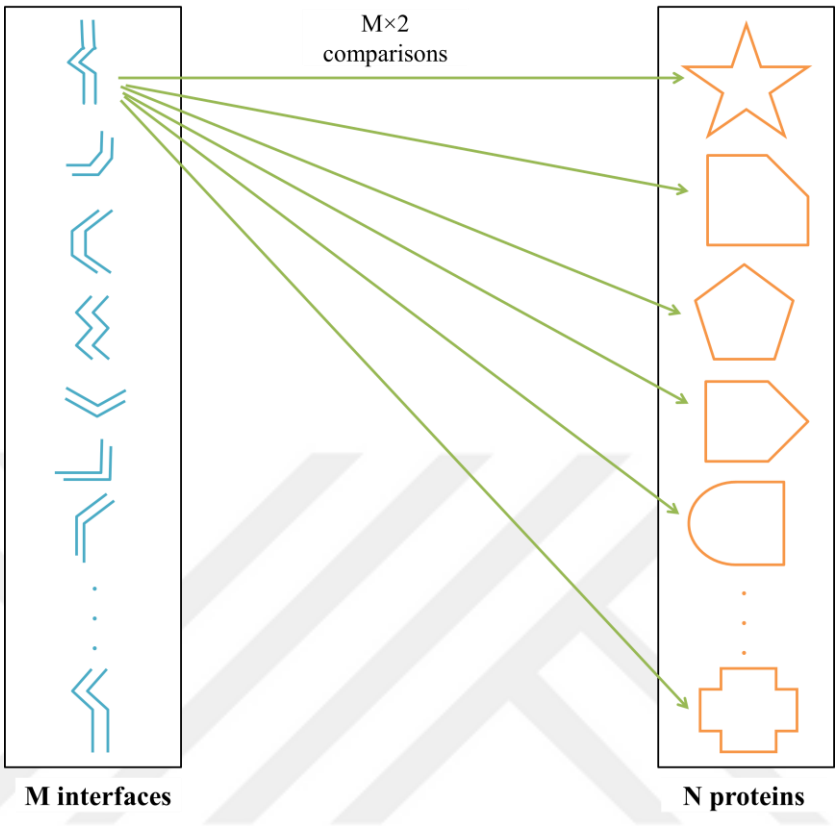

Figure 5.2: Schematic representation of the structural alignment of template interface partners with target surfaces in PRISM.

PRISM needs one second, on average, for structural matching of each query protein with a template interface side. For a target data set composed of N proteins and a template set composed of M interfaces, structural matching of all targets with all template interfaces takes M×2×N×1 seconds running time. As the current version of PRISM consisted of 22,604 interfaces in its library, structural alignment step takes more than 12 hours for one query protein. For the structural network of breast cancer lung and brain metastasis, which had 484 proteins with available structures, ~254 days was needed to run PRISM if just one structure for each protein investigated. However, there were proteins with multiple conformations in the network. The server version of PRISM is used in this study which is faster than the stand-alone version as it uses MySQL database and parallelization. However, it lasted several weeks to finish all the runs for the network.

So, to be able to enlarge the PRISM interface dataset and speeding up the structural matching step, a new representation for protein surface and interface structures is proposed in this study. This representation provides coarse-grained descriptors of protein structures, either interfaces or

surfaces, which can be used to quickly compare the protein structures. It converts 3D structures of proteins to 1D descriptors using a library of small protein fragments. This non-redundant library covers all 3D shapes of protein structures. The descriptors show the count of each library fragment found in the protein structures. Figure 5.3 shows a graphical demonstration of the descriptor creation for a protein structure.

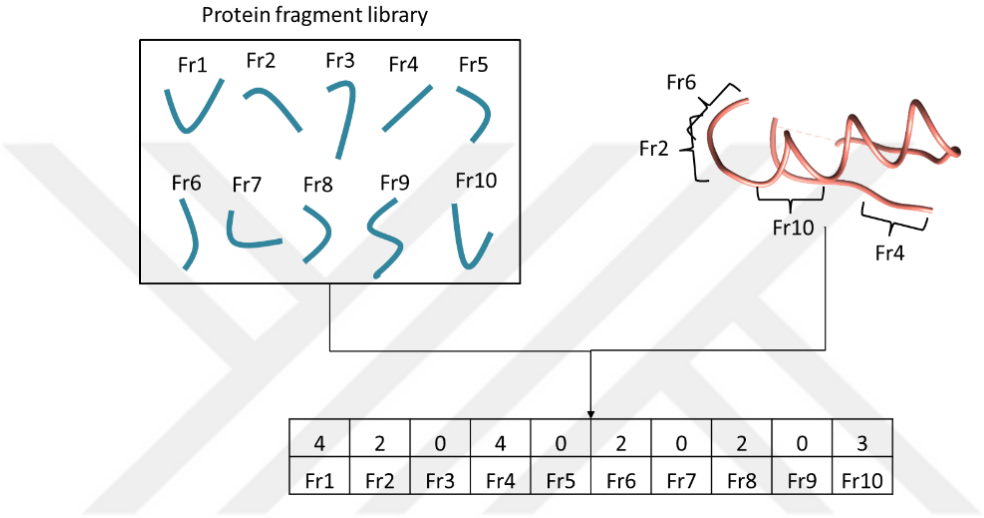

Figure 5.3: Graphical demonstration of descriptor creation for a protein structure.

The proposed representation is inspired by FragBag [29] algorithm and Geometric hashing technique [36]. Figure 5.4 shows the steps of the proposed method. By applying this feature-based algorithm at first step of PRISM, a small subset of potential similar interfaces will remain to be compared with the query protein surfaces through MultiProt. This representation can be used for interface comparisons to update the PRISM interface library.

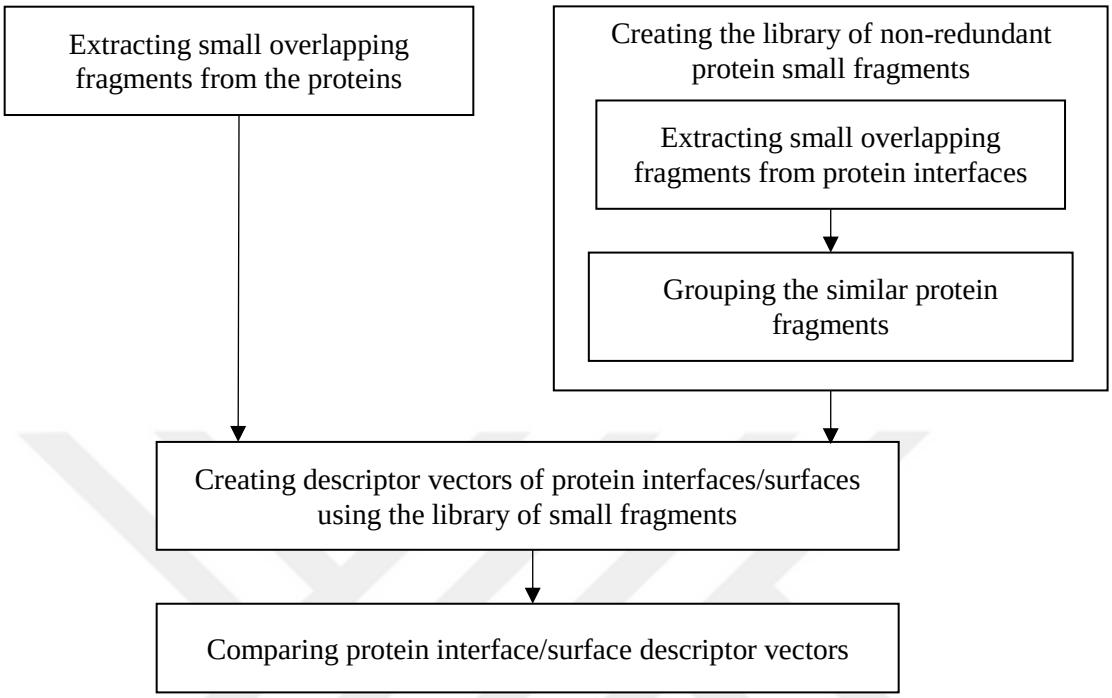

Figure 5.4: Flowchart of the proposed feature-based method for protein surface/interface comparison.

The process starts with creating a library of non-redundant small protein fragments. This library is created based on the PRISM interface dataset. The process continues with grouping the similar protein fragments to have a non-redundant library of protein interface building blocks. Using this library, descriptor vectors for the protein interface and surface structures can be created. The algorithm extracts the small protein fragments from the interface and surface structures and finds which library fragment is similar to each of these extracted fragments. Finally, comparing protein surfaces and interfaces are possible by using these descriptor vectors.

This chapter continues with step by step explanation of the proposed method and corresponding results. Then, PRISM template interface dataset and the related analyses are shown.

5.1 Extracting Small Structural Fragments from PRISM Interfaces

The process started by extracting small overlapping fragments from PRISM interface dataset. For each interface side, starting from the first residue on the protein chain, the continuous fragments of length k, k =3..10 in this study, are extracted. To cover all possible protein interface fragment shapes, the overlap size is set to maximum i.e. two consecutive continuous fragments are different in just one residue. Figure 5.5 and Figure A1 show a schematic representation of a protein backbone residues and a real PRISM interface side file from them the fragments of size 4 are extracted with overlap size 3, respectively.

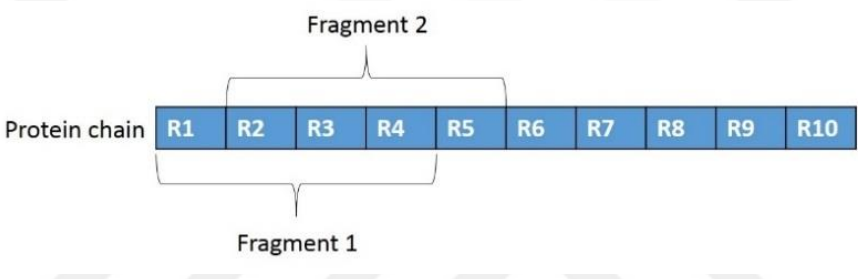

Figure 5.5 Schematic representation of a protein backbone residues from which the overlapping fragments are extracted.

R stands for residue number in protein backbone structure. Figure 5.6 shows the number of extracted continuous fragments from the PRISM interface dataset.

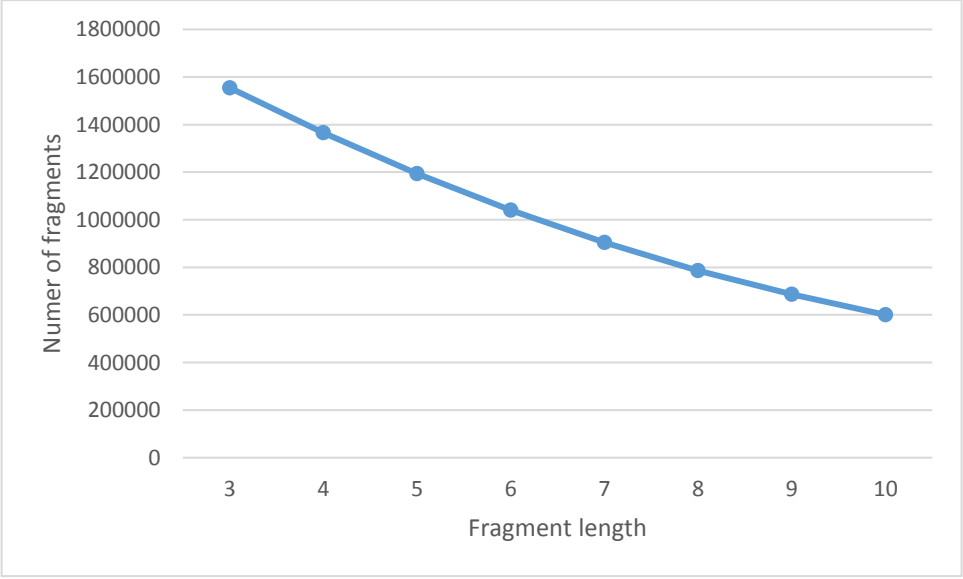

Figure 5.6 Number of extracted overlapping fragments from PRISM interface dataset.

As can be seen from the figure, the number of extracted fragments diminishes eye-catchingly by increasing the fragment length i.e. from 1,554,663 fragments of length 3 to just 600,462 fragments of length 10. The reason is that, as mentioned earlier, the protein backbone is not continuous on the protein interface, so as the fragment length rises, some of the small disconnected regions will be lost on the interface sides.

To decide about the fragment length for the upcoming analyses, the frequency of protein fragments of length k in each interface side is considered. Figure 5.7 shows the ratios of PRISM interface sides having different fragment lengths.

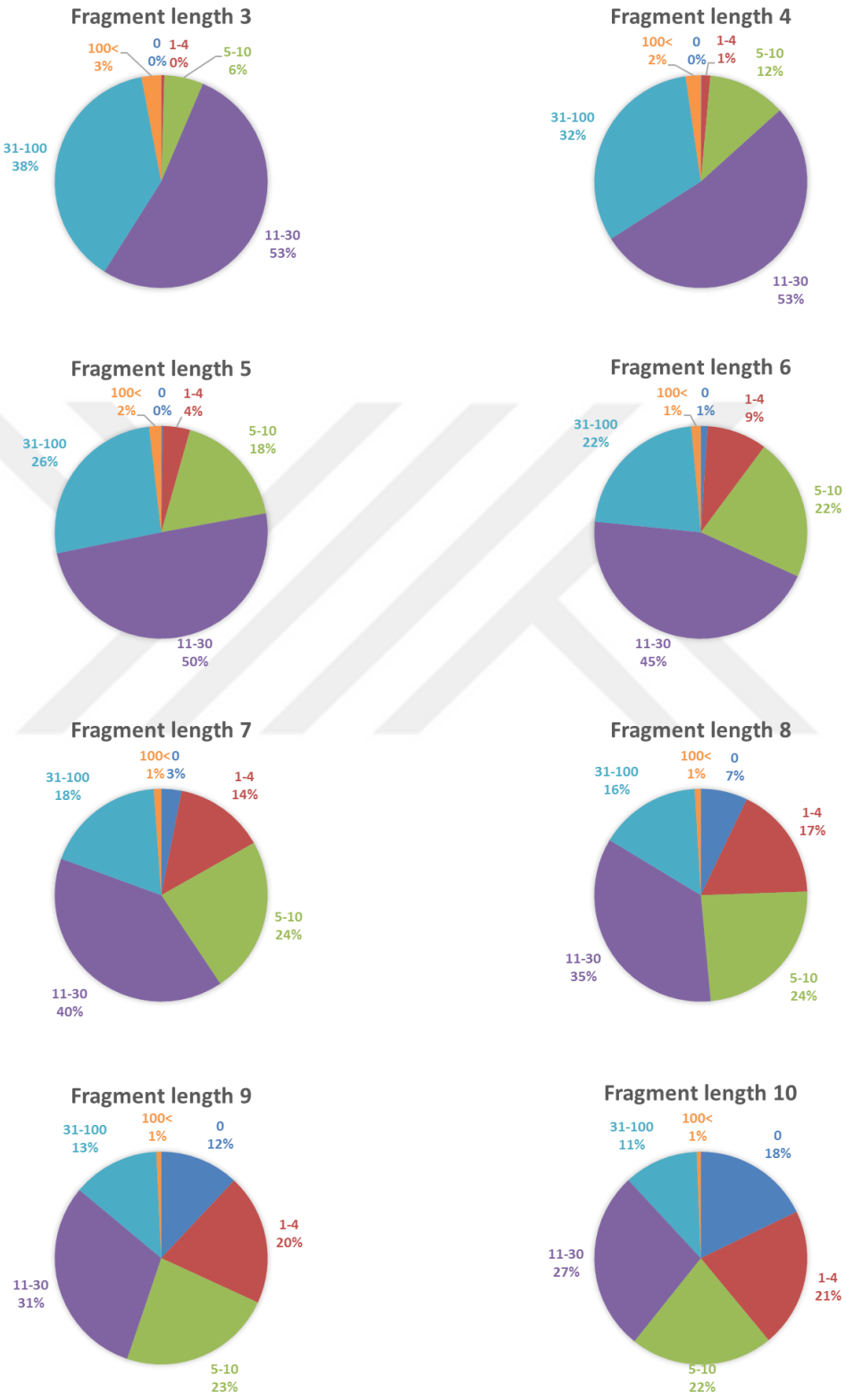

Figure 5.7 Ratios of PRISM interface sides for different fragment lengths. The numbers show the number of the fragments in each interface side.

As can be seen from the figure, 94% of the interface sides have at least 11 fragments of length 3. However, this ratio falls sharply to just 39% for fragment length 10. Also, if the fragment length is set to 10, 18% of the interface sides will be lost and even for 43% of the interface sides, the descriptor vectors will not be differentiative as they consist of only 1-10 fragments. Based on these graphs, fragment lengths 3-6 are selected as they seem to be more promising because of less information loss for the interface sides.

5.2 Grouping Similar Protein Fragments

To create the library of non-redundant protein structural fragments, the extracted protein fragments need to be grouped based on their shape similarities. This library would be representing the interface structure building blocks. In this study, Geometric hashing [36] is used to represent the protein fragments and grouping them i.e. instead of showing the protein fragments as PDB files, they are represented as their geometric hash tables. Using Geometric hashing makes the structure comparisons much faster as its time complexity is polynomial while finding the alignment between the protein structures, as PDB files, needs exponential time.

A geometric hash table of a protein fragment comprises all rotation models of that fragment using all permutations of the triplets selected as Geometric hashing Reference Set (RS). Pseudocode of creating the non-redundant protein fragment library is shown in Figure 5.8. For each interface side, starting from the first residue on the protein chain, the continuous fragments of length k, k =3..10 in this study, are extracted. To cover all possible protein interface fragment shapes, the overlap size is set to maximum i.e. two consecutive continuous fragments are different in just one residue. Figure 5.5 and Figure A1 show a schematic representation of a protein backbone residues and a real PRISM interface side file from them the fragments of size 4 are extracted with overlap size 3, respectively.

R stands for residue number in protein backbone structure. Figure 5.6 shows the number of extracted continuous fragments from the PRISM interface dataset.

.

```
Input: fragmentDeck
Output: seedFragment hash tables, seedFragment groups
Initialize: Set values for fragmentLength, binSize, and minClusterMembers variables
While (There is an ungroupedFragment in fragmentDeck)
    seedFragment ← Select randomly from fragmentDeck
    Create a fragment group consisted of only seedFragment
    Remove seedFragment from fragmentDeck
    Create hash table of seedFragment including its all rotation models
    For each ungroupedFragment in fragmentDeck
        Create hash table of ungroupedFragment including just one rotation model
        If ungroupedFragment hash table matches seedFragment hash table
            Add ungroupedFragment to seedFragment group
            Remove ungroupedFragment from fragmentDeck
    If (seedFragment group size > minClusterMembers)
        Write seedFragment hash table to file
```

Figure 5.8 Pseudocode of creating the library of protein fragments.

To group the protein fragments, one fragment is selected randomly to be the seed fragment to form a new fragment group. Hash table of the seed fragment is created consisting all rotation models of it. Each rotation model derives from a specific permutation of three points selected from the seed fragment. The running time of seed fragment hash table creation step is $L \times L\text{-}1 \times L\text{-}2 \times L\text{-}3 = O(L^4)$, where L is the fragment length.

After seed fragment hash table creation, it will be compared with all ungrouped fragment hash tables. To create the ungrouped fragment hash tables, just one rotation model is investigated because the seed fragment hash table includes all rotation models, if ungrouped fragment does not match the seed fragment using only one rotation model, it will not match using other rotation models too. So, the running time of hash table creation and then comparison in this step is L and S, respectively, where S is the size of seed fragment hash table. It should be noted that, if finding

the substructure matchings between the fragments was needed, it was necessary to investigate all rotation models of the ungrouped fragments because in each rotation model, the number of matched points may change. As the fragment length is set to small values, just the complete matches between the fragments were considered in this study. Based on the pseudocode given in Figure 5.8, the running time of grouping algorithm is $n \times L^4 \times n\text{-}1 \times L \times S \sim O(n^2 \times L^5 \times S)$ where n is the number of ungrouped fragments in fragmentDeck. Based on the values used in this study, L is much smaller than n. After tracing all ungrouped fragments for a seed fragment, if the number of fragments in seed fragment's group was more than a predefined threshold, then the seed fragment hash table was stored in the output file. The minimum number of group members is set to 5 in this study.

BinSize is a feature used for grouping the extracted small protein fragments using Geometric hashing. In Geometric hashing method, when rotating a shape to find its best alignment to another shape, instead of looking at exact point coordinates, they can be divided into binSize so the small movements in each point of the shapes would not make them different. The smaller the binSize, the stricter the grouping algorithm, e.g. if the binSize is set to 1, the algorithm can only find the protein fragments which have exactly the same shape. Figure 5.9 shows a graphical representation of how binSize affects the grouping. Figure A3 shows a seed fragment and its geometric hash table.

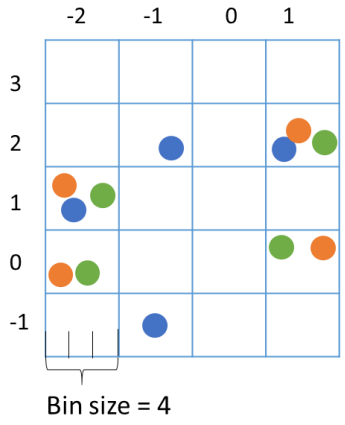

Figure 5.9 Graphical representation of three shapes i.e. green, blue, and orange in a geometric hash table. Each shape is in a specific rotation model. By setting binSize to 4, orange and green shapes will have the same coordinates and be grouped together.

Results of grouping the extracted continuous protein fragments of length 3-6 are shown in Table 5.1.

Table 5.1 Results of grouping the protein fragments of lengths 3-6.

Fragment length	Bin size	#of groups
3	2	29
4	2	135
5	2	619
6	2	2885
3	3	18
4	3	21
5	3	25
6	3	42
3	4	14
4	4	16
5	4	17
6	4	18

As can be seen from the results, in bin size two, the number of groups rise eye-catchingly as the fragment length increases. However, this trend slows down for bin size three and four as the grouping becomes smoother. As two consecutive residues in protein chain have the distance ~3.8 [156], there were bins inside which two residues were found in case of bin size four. This causes losing the information about some parts of the protein fragments which decreases the accuracy of the algorithm in comparing the protein fragments. So, bin size two and three seemed to be more promising for later analyses.

As mentioned earlier, only the complete matches are considered in this study as the fragment lengths are small enough. This makes the hash table comparisons to be very fast; because investigating only one rotation model of the ungrouped fragment is enough to perceive if it is similar to the seed fragment or not. Here, there are examples of running time differences between two grouping scenarios in which the first scenario tries to find the similar fragments based on complete matches and the second scenario tries to group the fragments by substructure matching. Grouping algorithm in the second scenario needs to investigate all rotation models of the

ungrouped fragments to see if they sub-structurally match the seed fragment or not. The results are shown in Table 5.2. As can be clearly seen, in the sub-structural matching scenario, the running time of the algorithm grows eye-catchingly because of the many rotation models investigated.

Table 5.2 Run time of fragment grouping algorithm in complete match and sub-structure match scenarios.

Fragment length	**Bin size**	**Full match**		**Sub-structural match**	
		Match size	**Run time (s)**	**Match size**	**Run time (s)**
4	2	4	102	3	430
5	2	5	729	4	2391
6	3	6	112	5	2292

The vital feature of the fragment libraries is how much they cover the protein structural space i.e. if they have enough diversity of protein fragment shapes to describe the protein structures. To check if the fragment libraries are sufficient to describe unseen protein structures, 10,000 protein interfaces are selected randomly among the PIFACE [120] cluster members. PIFACE is a database of 130,209 protein-protein interfaces which are divided into 22,611 clusters based on structural similarities. For each cluster in PIFACE, there is a representative protein-protein interface which has RMSD $\leq$ 2Å with each cluster member. These representative interface structures form the PRISM interface dataset.

The 1000 interface structures are selected such that there is at least one continuous fragment of length six in each interface side, as the maximum fragment length to be investigated is six in this study. This limitation is applied to make sure the results are not affected by short protein interfaces. The results are shown in Table 5.3. As the table represents, the fragment libraries are quite sufficient to create the descriptor vectors of the protein structures as the average ratios of missed fragments in interface structures are less than 8% for all fragment libraries. This ratio increases as the fragment length rises because the longer fragments can gain more diverse shapes in 3D space than smaller ones.

Bin size has a minor effect on the ratio as it is a factor forming the groups with more/less diverse members, and in the case of less diverse groups, the number of groups increases to cover the different shapes of fragments. However, in bin size three the missed fragments ratio is less than bin size two, as it accepts more diverse shapes as similar. It should be noted that, in the grouping process, groups having less than 5 members are disregarded and the missed fragments in these unseen interface structures may be among those skipped fragment shapes.

Table 5.3 Average ratios of missed fragments using different fragment libraries for unseen interface structures.

Fragment length	Bin size	Number of extracted fragments	Average ratio of fragments missed in all extracted fragments	Average ratio of fragments missed in each interface structure
3	2	928,791	0.02	0.02
4	2	847,951	0.03	0.04
5	2	771,321	0.04	0.05
6	2	699,768	0.05	0.08
3	3	928,791	0.02	0.02
4	3	847,951	0.02	0.03
5	3	771,321	0.03	0.04
6	3	699,768	0.04	0.05

In this study, the continuous small fragments are extracted from the protein structures and clustered. Then, the descriptor vectors are created based on these fragment clusters. However, as protein interfaces and surfaces are not flat and have many holes, they consisted of several disconnected regions. As continuous fragments are investigated in this study, by increasing the fragment length, many small fragments are lost because they are not connected to other parts on the protein structure.

To test how the results change by inspecting the disconnected regions in the fragment library, the fragments are extracted without considering if they are continuous or not. Even if those parts

are not in the protein surface/interface structure, they affect the protein surface/interface global shape. Number of extracted continuous vs discontinuous fragments are shown in Figure 5.10.

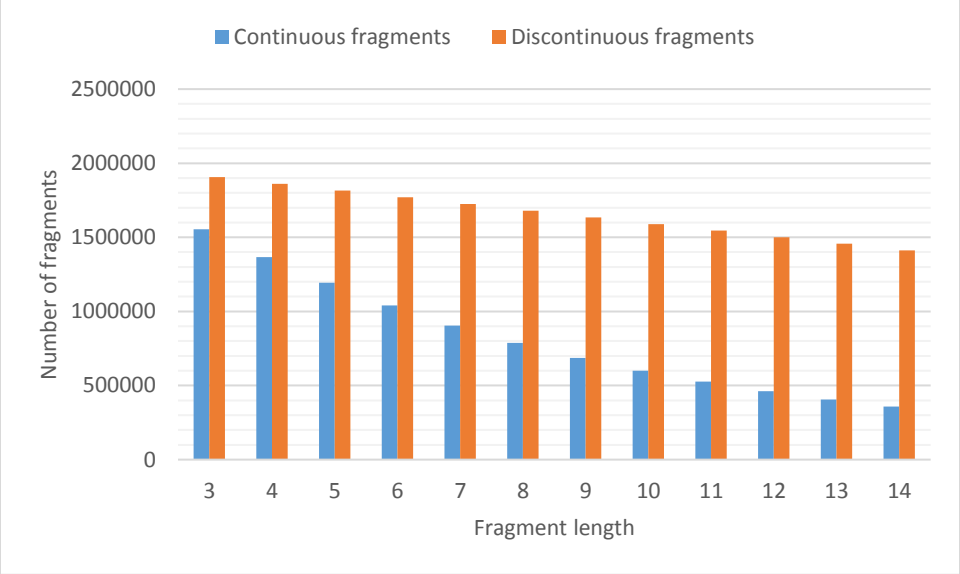

Figure 5.10: Number of extracted continuous vs discontinuous fragments

As can be clearly seen from the graph, in case of continuous fragments, when the fragment length increases, the number of fragments reduces sharply and many smaller protein fragments are lost.

The discontinuous fragments of size 5 are grouped with bin size set to 2 and match size set to 4. Surprisingly, it returned 4,527 different fragment groups, though in the continuous case it returned only 619 groups. Fragments of length 5 are small portions and this many groups derived from the fact that when the disconnected parts are included in the fragments, the next point in the fragment can be anywhere in the 3D space. By increasing the fragment length, the number of groups would be many more than this. In case of the continuous fragments, there is a threshold on the distances between continuous points on the protein structure because the distance between two consecutive alpha carbon in the protein backbone is ~3.5 Å. Therefore, the idea of including discontinuous fragments in the algorithm did not seem to work.

5.3 Creating Protein Interface Descriptor Vectors

Using the library of non-redundant protein fragments, protein structure descriptor vectors can be created. The descriptor vectors show each library fragment's count found in the protein structure. It is a list of size N, which is the number of protein fragments in the library. An example protein descriptor vector can be as Figure 5.11.

Fr_1	Fr_2	Fr_3	Fr_4	Fr_5	Fr_6	Fr_7	….	Fr_n
20	5	0	0	4	0	7	….	2

Figure 5.11: An example of protein structure descriptor vector.

Fr stands for fragment and it represents a fragment in the fragments library. The numbers below the fragments show the fragment count in the protein structure. This descriptor vector is a compact representation for a protein structure which can be used for fast protein structure comparison. The advantage of this representation is that it converts a protein structure from 3D space to a 1D vector.

Pseudocode of the algorithm to create the descriptor vectors for a list of proteins is shown in Figure 5.12. The algorithm extracts the small protein fragments from the protein structures based on the predefined fragment length. Then, it creates the protein fragments geometric hash tables based on the predefined bin size. In this step, the geometric hash table of a fragment includes just one rotation model as it is enough to find the similarity between the new fragment and the library fragment because the library fragment hash tables include all rotation models. The algorithm proceeds with comparing the hash tables to find which library fragment is similar to the new protein fragment. This process repeats for all fragments extracted from a protein structure to form its descriptor vector.

```
Input: proteinList, fragmentLibrary
Output: proteinDescriptorVectors
Initialize: Set values for fragmentLength and binSize.
For each protein in proteinList
    Set protein descriptor to a vector of zeros of size N
    For each fragment in protein structure
        Create hash table of fragment including just one rotation model
        For each libraryFragment in fragmentLibrary
            If fragment hash table matches libraryFragment hash table
                Increase corresponding index in protein descriptor vector by one
                Break
    Add protein descriptor vector to proteinDescriptorVectors
    Write proteinDescriptorVectors to file
```

Figure 5.12: Pseudocode of the protein descriptor vector creation algorithm.

The running time of this algorithm is $n \times L \times m \times S$, worst case, in which n is the number of fragments extracted from the protein list, L is the fragment length, m is the number of fragments in fragment library, and S is the size of fragment library hash table. Figure A2 shows an example of protein descriptor vectors. As the protein descriptor vectors are sparse, the zero indices are removed for memory usage.

To apply the algorithm on protein interface structures, 1000 PIFACE cluster representatives are selected randomly. These representative interface structures are selected such that they have at least 10 cluster members, and one continuous fragment of size 6 in each interface side, as the maximum fragment length investigated was 6. The selected interfaces have 108 cluster members in average, minimum 10 and maximum 491.

After creating the interface descriptor vectors, a metric is needed to express the similarity of descriptor vectors. One of the metrics used successfully in datamining and web page text mining applications to show the descriptors similarities is cosine similarity [29, 157, 158]. Cosine similarity formula is:

$$CosineSimilarity[X,Y] = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n}(x_i)^2}\sqrt{\sum_{i=1}^{n}(y_i)^2}}$$

Mean cosine similarities between descriptor vectors of PIFACE cluster representatives are shown in Figure 5.13.

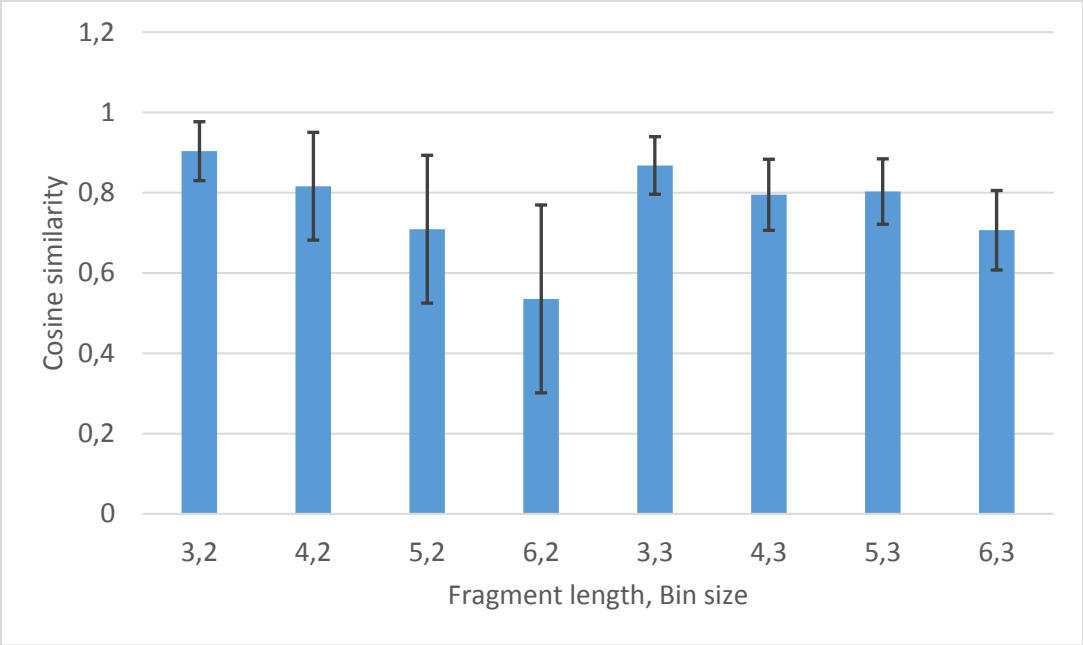

Figure 5.13 Mean cosine similarities between PIFACE cluster representatives.

As can be seen from the figure, cosine similarity decreases as the fragment length increases. This was expected intuitively because longer fragments can show the differences better than smaller fragments. The decrease trend is sharper for bin size two than three as smaller bin size is more sensitive to shape differences. Generally, the results for bin size two seems better than bin size three, however the standard deviation enlarges for bin size two as the fragment length increases which shows the instability of descriptor vectors in showing the shape differences between PIFACE cluster representatives monotonically. Figure 5.14 shows two examples of

PIFACE cluster representatives having different cosine similarities. The descriptor vectors of the interfaces shown are derived from fragment length six with bin size three. In Figure 5.14.a descriptor vectors of the interfaces 1cov24 and 3gj3AB have cosine similarity 0.96 because even the global shapes are different, the small fragments consisting the interfaces are similar. Cosine similarity between descriptor vectors of 1cov24 and 2j5iBD is 0.12 as there is not any local similarities between these two interface structures.

Figure 5.14.b shows that even there is a significant size difference between interface structures, they can have high cosine similarity because of their building blocks similarities. As the descriptor vectors of the interfaces 2xscED and 1a6dAB have cosine similarity 0.99. To prevent these comparisons, which may mislead the results, the protein interfaces having different sizes should be filtered out. So, before comparing two interface descriptor vectors, if their size difference is greater than 50% of the smaller interface, their cosine similarity is set to zero. This limitation, reduces the number of comparisons, though it does not affect the mean cosine similarity between the PIFACE cluster representatives as shown in Figure 5.15. It has been already shown that native protein interface structures have a high similarity score with other random interfaces with respect to the backbone geometry and binding pattern [159].

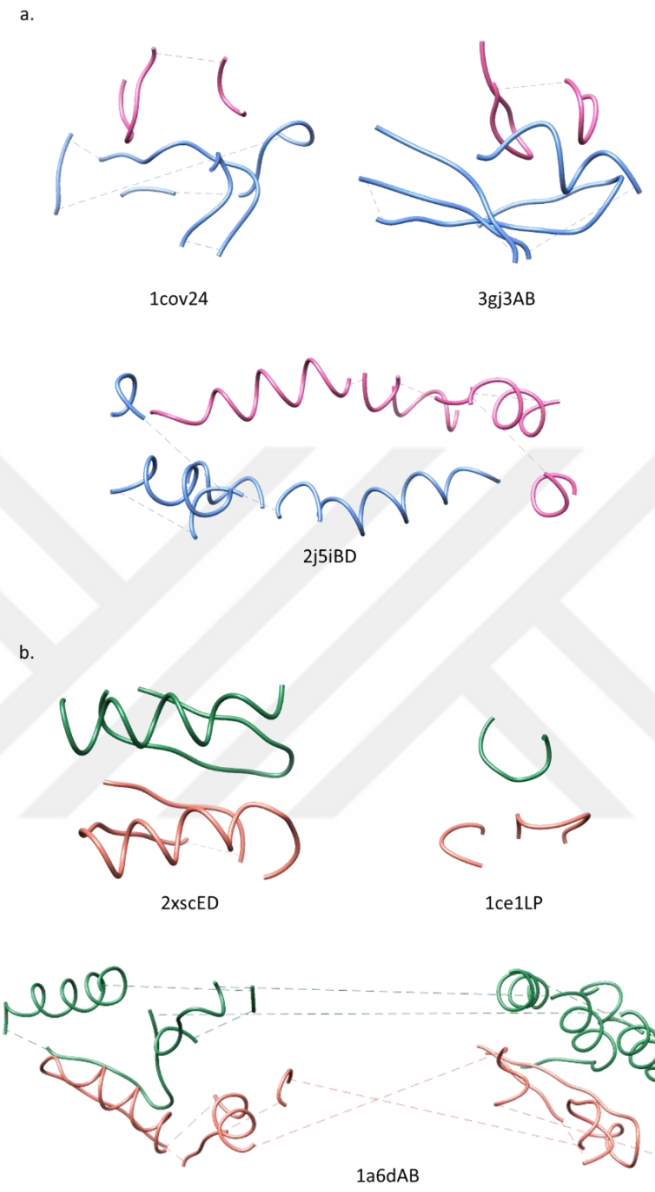

Figure 5.14 Examples of PIFACE cluster representatives having different cosine similarities. Interface sides colored differently. a) The descriptor vector of interface 1cov24 has cosine similarities of 0.96 and 0.12 with descriptor vectors of 3gj3AB and 2j5iBD respectively. b) The descriptor vector of interface 2xscED has cosine similarities of 0.99 and 0.25 with descriptor vectors of 1a6dAB and 1ce1LP respectively.

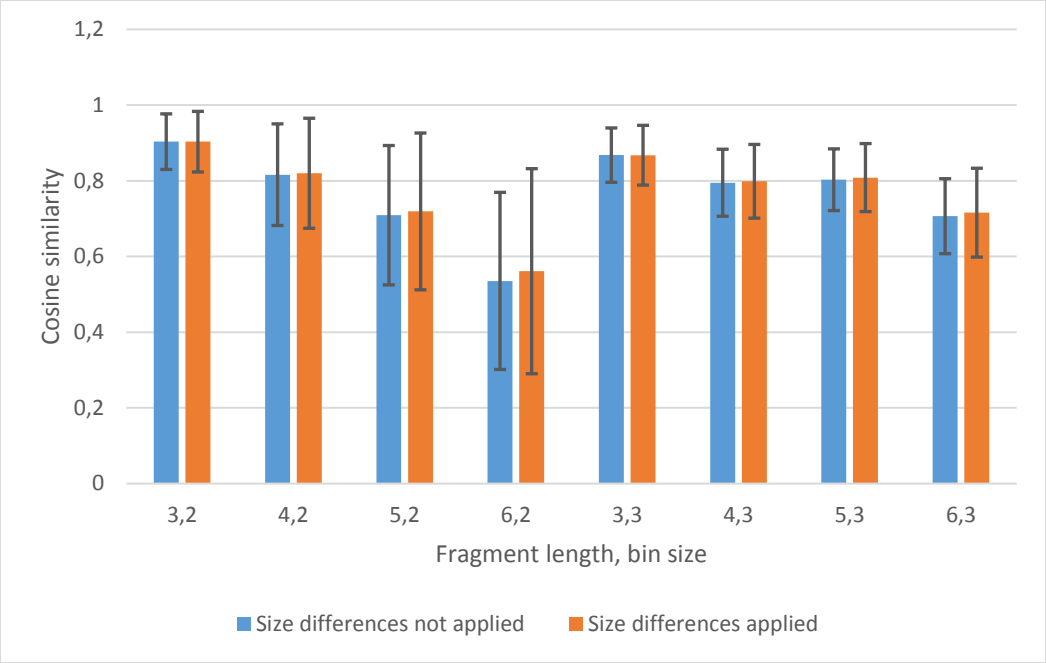

Figure 5.15 Mean cosine similarities between PIFACE cluster representatives in two cases: with and without inspecting the size differences.

To see the differences between descriptor vectors of the PIFACE cluster representatives and cluster members, the cosine similarities between members of each cluster are calculated for the selected 1000 clusters. Mean cosine similarities between descriptor vectors of PIFACE cluster representatives and cluster members are shown in Figure 5.16. As can be seen from the graph, there is a noteworthy difference in cosine similarity values for cluster representatives and cluster members. Mean cosine similarity for cluster members is around 1 in fragment length three. This value decreases for cluster members as the fragment length increases, though with lighter slope than for the cluster representatives, showing its stable behavior in detecting the similarities between truly similar interfaces which are grouped together in PIFACE clusters. Intuitively, bin size two is expected to show the differences between descriptor vectors of cluster representatives and cluster members better than bin size three, however the results are vice versa. The reason behind it can be the parameter values used in PIFACE to cluster the interface structures. PIFACE uses 3 Å as the threshold for RMSD values in structural comparisons [120].

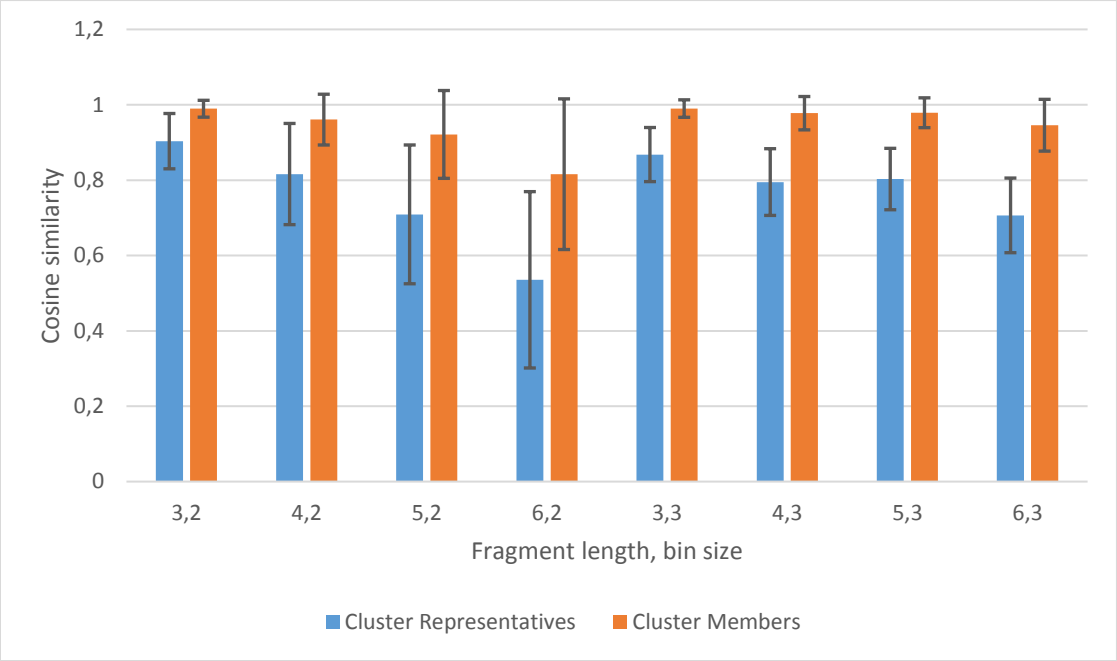

Figure 5.16 Mean cosine similarity between PIFACE cluster representatives and cluster members.

5.3.1 Protein Interface Comparison

Based on the results shown in Figure 5.16, there is a considerable difference in Cosine similarity values between PIFACE cluster members and cluster representatives. To see how helpful the descriptor vectors are in filtering the dissimilar interface structures, 1000 PIFACE interfaces are selected randomly. The descriptor vectors of these 1000 interfaces are compared with all other interfaces, if their sizes are not very different, and the Cosine similarities of them are calculated. Interfaces having the cosine similarity value greater than or equal to the predefined threshold are remained as potentially similar interfaces and the others are filtered out. The objective in this step is to see by inspecting the descriptor vectors, how many cluster members will be found and how many will be filtered. Table 5.4 shows for how many of the interfaces, the algorithm could find their corresponding PIFACE cluster members (at least 90% of them).

Table 5.4 Results of comparing 1000 interface descriptor vectors with PIFACE cluster members and representatives

Fragment Length, Bin size	Cosine similarity threshold	Number of interfaces for which at least 90% of cluster members are found	Ratios of filtered interfaces
3,3	0,9	922	80%
4,3	0,9	864	85%
5,3	0,9	793	86%
6,3	0,85	689	89%

Based on the results shown, using the descriptor vectors the algorithm could find the similar interface structures for majority of the target interfaces. The accuracy decreases as the fragment length rises. The reason behind it can be the fact that as the fragment length increases, the descriptor vectors get shorter, so they may not be describing the interface structure as good as the smaller fragments.

It should be noted that many of the filtered cluster members are filtered because of their size differences with the target interfaces. There are clusters in PIFACE in which interfaces of different sizes are clustered together e.g. interfaces of size 80 and 170 residues. So, if the interfaces are not filtered based on their sizes, the number of interfaces for which at least 90% of cluster members are found increases from 922 to 973 for fragment length three. Although, the ratio of filtered interfaces decreases to just 41%.

The comparison algorithm takes ~0.2 seconds to compare a protein interface structure with all PRISM interfaces. So, it could be efficiently used for interface clustering and ascertaining if the new interface structures deployed in PDB are similar to the existent PRISM interfaces or not.

5.4 Comparing Protein and Interface Descriptor Vectors

In the previous sections, the interface descriptor vectors are shown to be useful in differentiating similar and dissimilar interface structures. Here, the descriptor vectors are used to find the interface sides on protein surfaces. Of course, there is a huge size difference between the interface sides and protein surfaces to be compared.

In this section, Euclidean distance and Histogram intersection similarity are also used for comparison besides Cosine similarity. Euclidean distance formula is:

$$Euclidean\ distance[X,Y] = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Histogram intersection formula is:

$$Histogram\ intersection\ similarity[X,Y] = \frac{\sum_{i=1}^{n} min(x_i, y_i)}{min(\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} y_i\ )}$$

x_i and y_i indicate item i in the descriptor vectors x and y, respectively.

To apply the algorithm on protein surface structures, 1000 PIFACE cluster members are selected randomly. These interface structures are selected such that they have at least one continuous fragment of size six in each interface side, as the maximum fragment length investigated in this step was six. Then, one side of each interface structure is selected randomly and the whole pdb structure of that interface side is downloaded. It should be noticed that on each pdb structure there can be several binding sites, so for each pdb in the list there can be several PRISM interface representatives. Naccess [160] is used to extract the surfaces of the downloaded pdb structures by calculating the accessible surface area. Surface depth is set to 6 Å as PRISM uses this value.

After extracting the surfaces of the 1000 pdb structures, their descriptor vectors are created using the code shown in Figure 5.12. Then, the protein descriptor vectors are compared with PRISM interface side descriptor vectors to see if the comparison algorithm can find the representative interface structures, for which this protein belong to their clusters. The results of comparing these descriptor vectors with PRISM interface side descriptor vectors are shown in Figure 5.17. Here, only the bin size three is investigated as in previous sections it has been shown to get better results in interface comparisons. The pdbs in the input list has 2 PRISM interface representatives in average (minimum 1 and maximum 10 representatives).

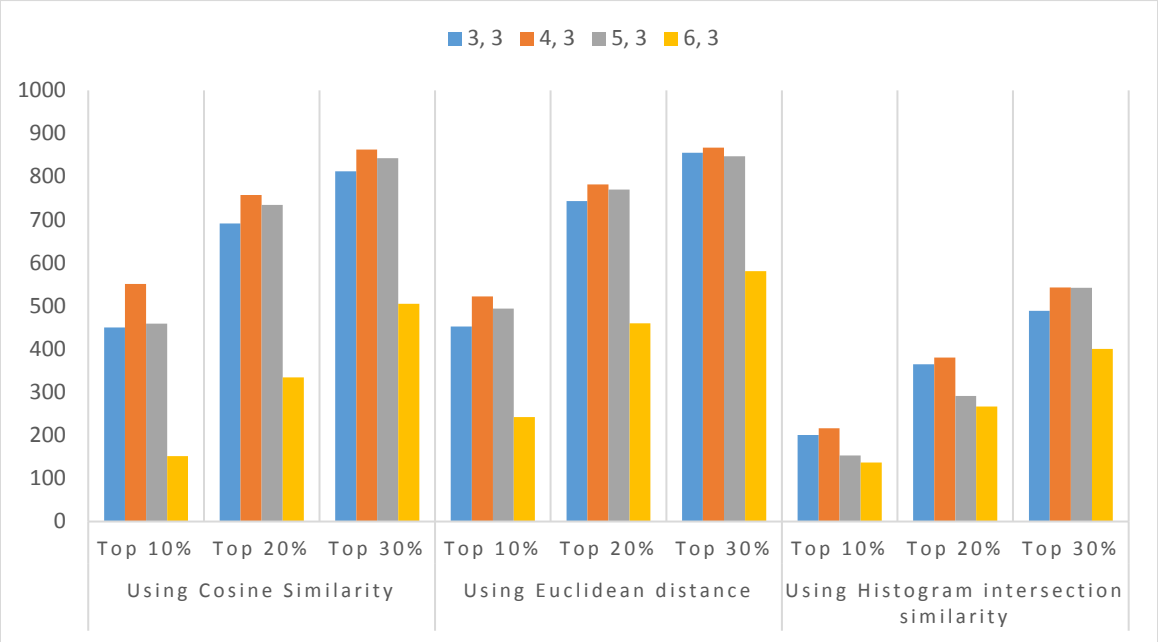

Figure 5.17 Comparison results of 1000 protein descriptor vectors with PRISM interface side descriptor vectors. It shows for how many proteins the algorithm could find the representative interfaces in top X% results. Each color shows the results for a specific fragment length, used for extracting the fragments, with bin size three.

As can be clearly seen from the figure, Histogram intersection could not find the correct results as good as the other two metrices. Histogram intersection measures the overlap between the two descriptor vectors, and as the protein interface sides are small structures, they can be perfectly overlapped with different parts, maybe other than the binding regions, of a protein surface. Euclidean distance and Cosine similarity results were roughly similar to each other.

In terms of fragment length, shorter fragments seem better than the longer ones as fragment length six has a considerable decrease in the number of correctly found representative interfaces. This can be derived from the fact that the larger the fragments are, the shorter the descriptor vectors become, and shorter descriptor vectors cannot describe the protein structures well. Fragment length 4 got slightly better results in comparison with other fragment lengths. The comparison algorithm takes 0.9 seconds to compare a protein surface structure with all PRISM interfaces.

5.4.1 Comparing Protein sub-surfaces with PRISM Interfaces

As the distances between fragments are not considered in the descriptor vectors, an interface side descriptor vector can match with a protein descriptor vector because it has similar fragments on different parts of its surface which might be far from each other. So, instead of comparing the whole protein surface's descriptor vector, several sub-surfaces can be extracted from the protein surface, and then descriptor vectors of them can be compared with interface side descriptor vectors. The extracted sub-surfaces should contain fragments which are near to each other on the protein surface in 3D space.

To do so, the sub-surfaces of the selected 1000 proteins are extracted and the descriptor vector of each sub-surface is created. To extract the sub-surfaces, the algorithm starts with the first fragment on protein surface chain. If the distance between centers of fragment i and j are smaller than a predefined threshold, then fragment j is near enough to be a part of the subsurface i. The process continues till the last fragment on protein surface chain. While comparing the sub-surfaces of a specific protein with PRISM interface sides, the maximum Cosine similarity, minimum Euclidean distance, and maximum Histogram intersection similarity of all protein sub-surfaces are considered as the similarity of the whole protein surface to an interface side. The threshold values 10, 15, and 20 Å are applied to extract the sub-surfaces. The results are shown in Figure 5.18.

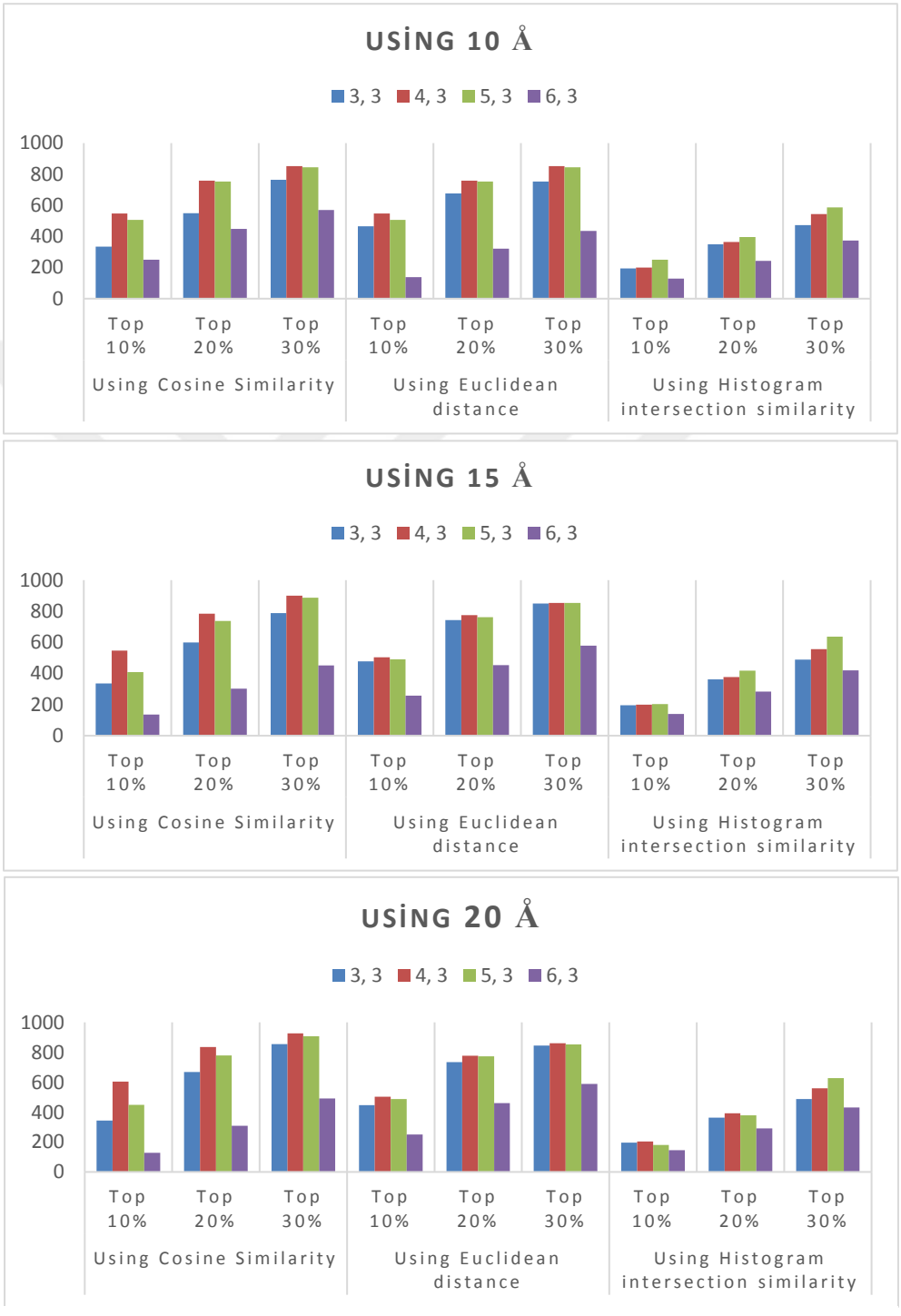

Figure 5.18 Comparison results of protein sub-surfaces descriptor vectors with PRISM interface side descriptor vectors. It shows for how many proteins the algorithm could find the representative interfaces in top X% results. Each color shows the results for a specific fragment length, used for extracting the fragments, with bin size three.

The results in this step is generally similar to the previous section. However, larger thresholds i.e. 15 and 20 Å got slightly better results as 10 Å was so small. For example, by using Cosine similarity in threshold 20 Å, the algorithm could find the representative interfaces for 927 proteins out of 1000.

In terms of running time, the proposed algorithm takes ~55 seconds to compare a protein surface structure with all PRISM interfaces. While, the same comparison takes 4,098 seconds in MultiProt. So, the proposed algorithm could be used to rapidly remove the grossly dissimilar interface structures in the preprocessing step of PRISM.

5.4.2 Comparing the Results with TM-align

TM-align [112] is an algorithm uses dynamic programming and TM-score [161, 162] rotation matrix to find the best structural match between two proteins. It returns the superposition of the aligned proteins as well as TM-score which shows their structural similarity. The value of TM-score lies in (0,1], where 1 indicates a perfect alignment between two protein structures. TM-score below 0.2 shows that there is no similarity between two structures while the score higher than 0.5 means the structures share the same fold.

To compare the TM-align with proposed algorithm, the data used in previous section and results taken by Cosine similarity for fragment length 4 are investigated. TM-align is applied to compare the 1000 protein surfaces with all PRISM interfaces. Based on TM-align results, there are 2139 similar interfaces for each protein surface on average (minimum 42 and maximum 4148 similar interfaces). The interfaces having TM-score higher than 0.5 got the Cosine similarity of ~0.7 on average in the proposed method. By setting the Cosine similarity to 0.7, 56% of the dissimilar interfaces would be filtered by the proposed algorithm and the remaining interfaces would be compared using an alignment-based method.

5.5 Prism Interface dataset

Protein interface is a part of protein surface which is used to bind to other proteins. Interface structures are usually consisted of several disconnected regions because of the holes exist on

protein surfaces. PRISM interface dataset includes 22,611 interfaces coming from protein complex structures in PDB. These unique interface structures are derived from PIFACE [120] database. For each interface in PRISM interface dataset, the coordinates of contacting and neighboring residues are stored, i.e. the residues directly involved in the protein-protein interaction and their neighbors within a radius of 6 Å.

Each interface consists of two interacting sides, so the PRISM interface dataset totally includes 45,222 interface sides. Figure 5.19 shows the frequency of protein interface side lengths. Most PRISM interface sides have 10-50 residues.

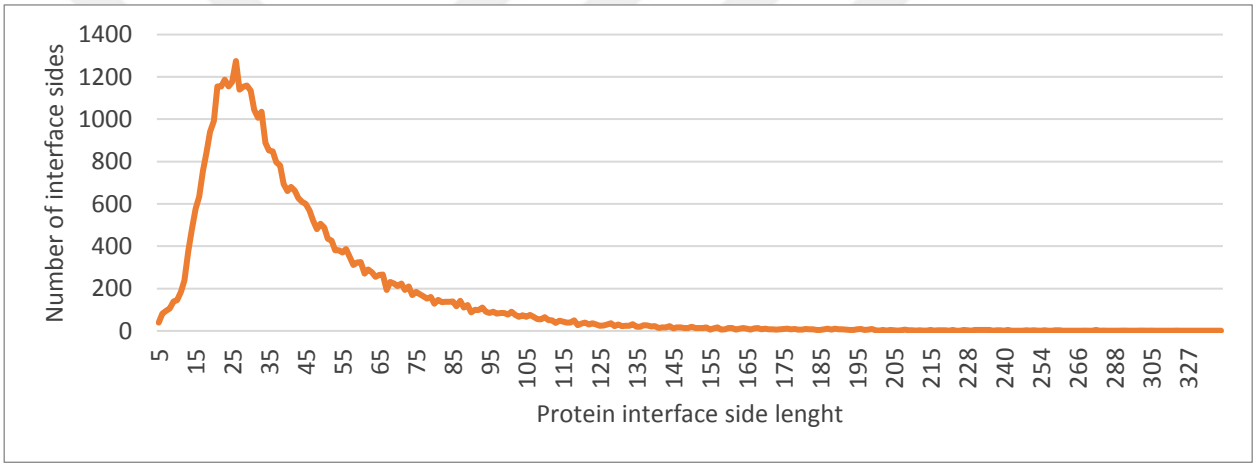

Figure 5.19 Frequency of PRISM interface side lengths.

As mentioned earlier, each side of a protein interface consists of several disconnected regions because of the holes existing on protein surfaces. Figure 5.20 lists the number of disconnected regions on PRISM interface sides. Majority of protein interface sides have 2-8 disconnected regions.

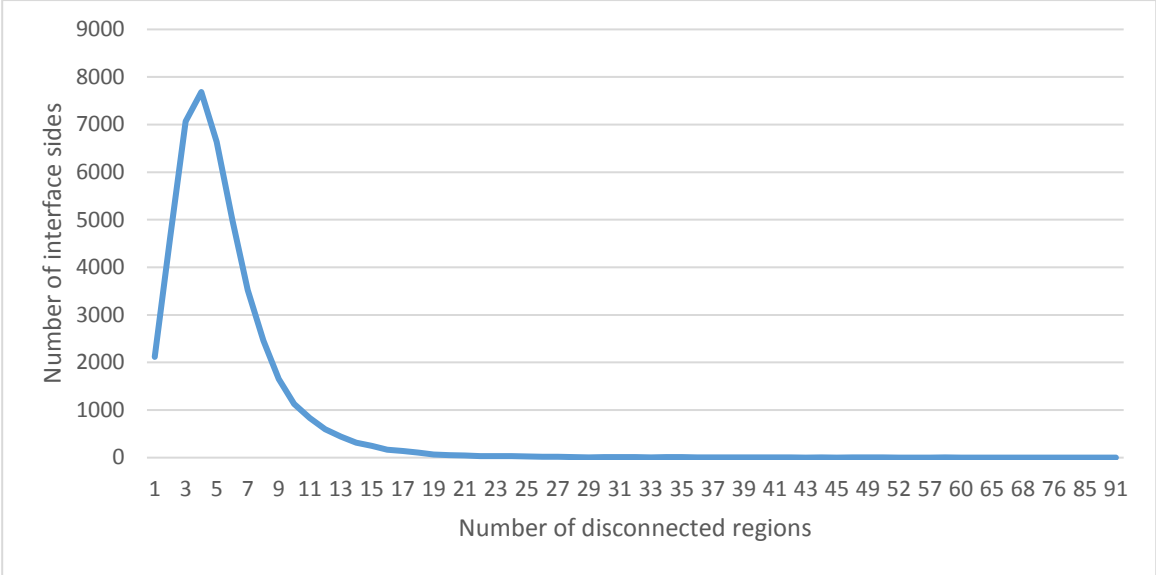

Figure 5.20 Frequency of PRISM interface sides consisted of several disconnected regions.

Figure 5.21 shows the lengths of the disconnected regions on protein interface sides. Disconnected regions of length 1-15 are the majority.

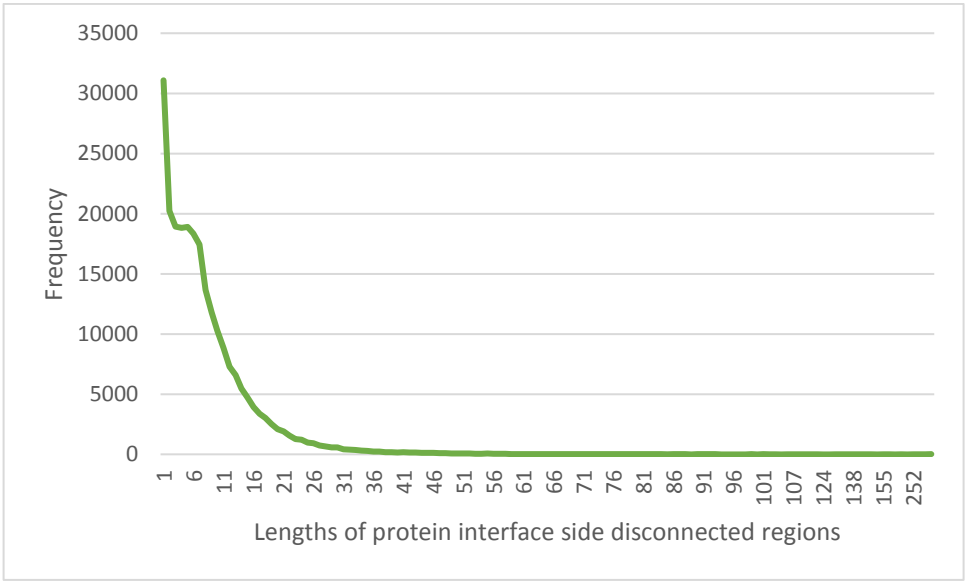

Figure 5.21 Number of protein interface side disconnected regions having different lengths.

Based on these graphs, it is decided to extract fragments of length 3-10 from the interface sides. To use the Geometric hashing technique, at least three residues are needed to create a plane to form a specific rotation of the protein fragment. Also, as the fragment length increases, the fragment

hash-table size and protein fragment library size will increase significantly because the protein fragment can form many different shapes in 3D space.

5.5.1 Labeling Fragments based on Protein Secondary Structures

To study the distributions of protein secondary structure elements in the interface regions, the extracted fragments are labelled based on the majority type of their residues secondary structures. For this analysis, fragment length six is selected as it is long enough, and it does not cause any significant fragments loss as shown in Figure 5.7. Fragments are labeled as: H (a fragment which is mostly helix), S (a fragment which is mostly strand), C (a fragment which is mostly random coil), HS (a fragment which is equally helix and strand), HC (a fragment which is equally helix and coil), SC (a fragment which is equally strand and coil), and HSC(a fragment which is equally helix, strand and coil). The distributions of these seven categories are shown in Figure 5.22.

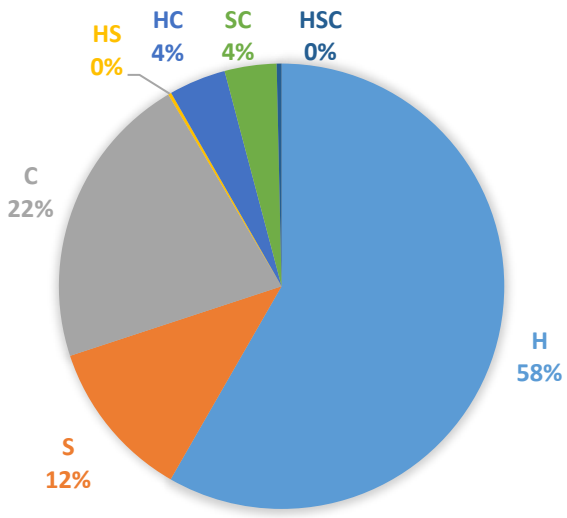

Figure 5.22 Secondary structure elements distribution in protein fragments extracted from protein interfaces.

Based on the graph, protein interfaces are mostly consisted of helices (~60%); and random coils are more favorable than strands to be used in protein-protein interactions. Similar results have been reported previously [163-165].

5.5.2 Distribution of Secondary Structure Elements in Fragment Groups

In section 5.5.1, the extracted fragments are labelled based on their residues' secondary structures. Based on Figure 5.22, the majority of fragments in protein interfaces are helices. Figure 5.23 shows the secondary structure elements distribution in the library of fragments i.e. after grouping the protein fragments. Fragment length six is selected for this analysis as section 5.5.1 with bin size three for grouping.

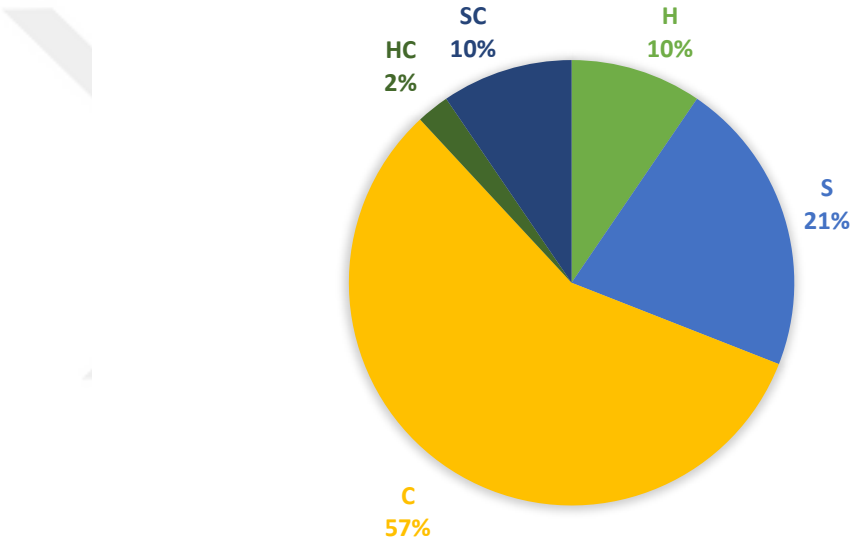

Figure 5.23 Secondary structure elements distribution in protein fragments library.

Even though coils just contributed to 22% of all extracted fragments from the protein interfaces, they have many different shapes in 3D space and form the majority of fragment groups as seen in the figure. On the other hand, all the extracted helices are grouped in a few groups as their shapes are similar to each other. This is intelligible intuitively because coils are freely moving and changing their shapes in 3D space than strands and helices.

To see if in the same geometry, protein fragments can be found having different secondary structures, the distribution of secondary structure elements inside each fragment group is investigated. Figure 5.24 shows the secondary elements distribution inside the fragment groups of length six which are grouped with bin size three. It should be noted that the number of fragments in each group varies significantly i.e. from 333,180 fragments in the most abundant group to only

5 fragments in the smallest groups. The groups in Figure 5.24 are ordered based on the number of fragments they consisted of.

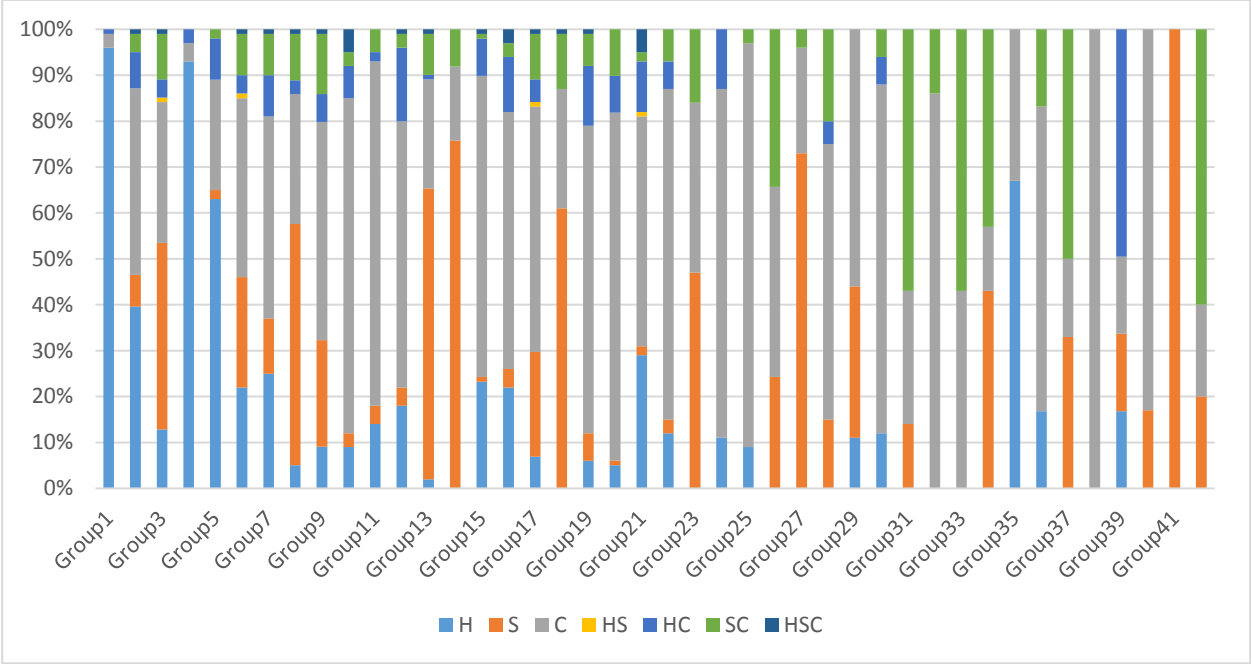

Figure 5.24 Secondary structure elements distribution inside protein fragment groups.

As can be seen from the figure, there are plenty of coils inside each group. As they are freely moving in 3D space and changing their conformation, they can form diverse geometric shapes.

To see the most abundantly used secondary structure elements in protein interface structures, the most crowded groups, having more than 10,000 members, are picked. Figure 5.25 shows the seed fragment structure for the most abundant fragment groups used in protein interfaces.

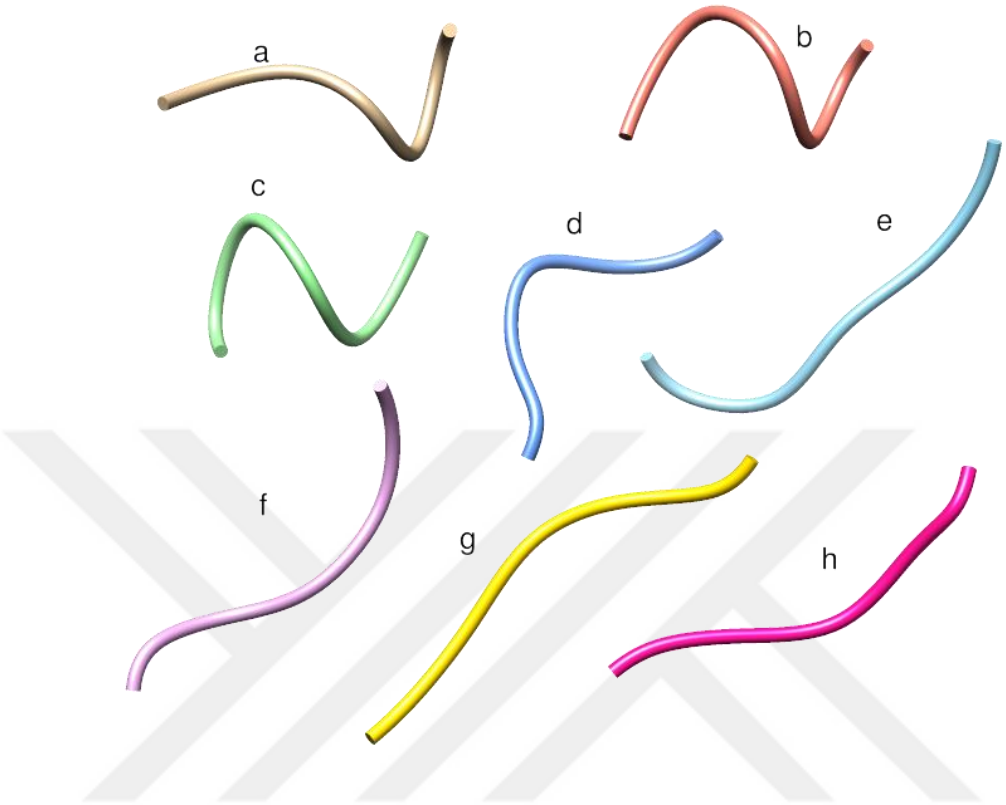

Figure 5.25 The most abundant fragment structures in protein interfaces

Fragments a, b, and c are helices, fragments d, e, and f are coils, and fragments g and h are strands.

5.6 Analyses of PRISM Interface Descriptor Vectors

In this section, PRISM interface library is analyzed based on the descriptor vectors of interfaces to have a deeper understanding of PRISM interfaces structures. Interface descriptors of fragment length six with bin size three is investigated.

Number of extracted fragments from interface sides are shown in Figure 5.26. Majority of the interface sides are consisted of 3-19 fragments.

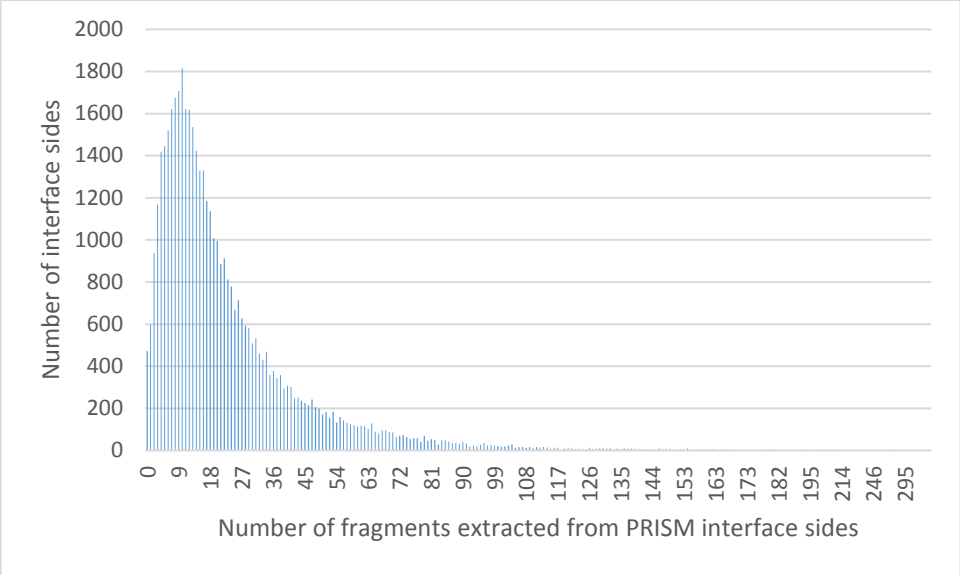

Figure 5.26 Ratios of PRISM interfaces side lengths in terms of the number of fragments extracted from them

5.6.1 Composition of PRISM Interfaces

As mentioned in Section 5.2, fragments of size six are clustered in 42 different clusters, i.e. there are 42 different fragments shapes of length six. Figure 5.27 shows that majority of PRISM interface sides are composed of just 3-6 different fragment shapes. So, inside an interface, several repeating fragments can be seen instead of many unique fragment shapes.

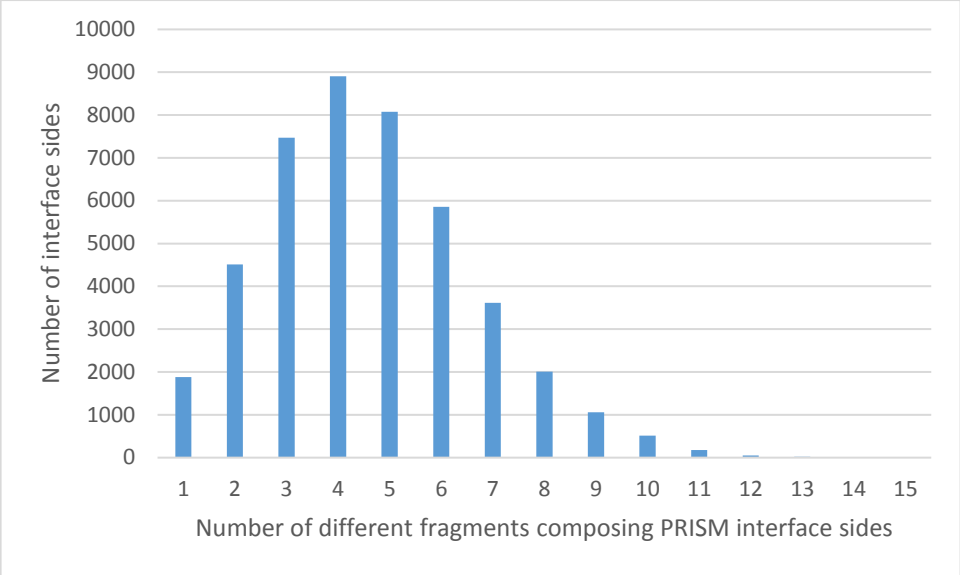

Figure 5.27 Ratios of PRISM interfaces sides composed of different fragments

Although the protein interface structures are mostly consisted of helices (section 5.5.1), inside each interface all secondary structures can be seen as shown in Figure 5.28.

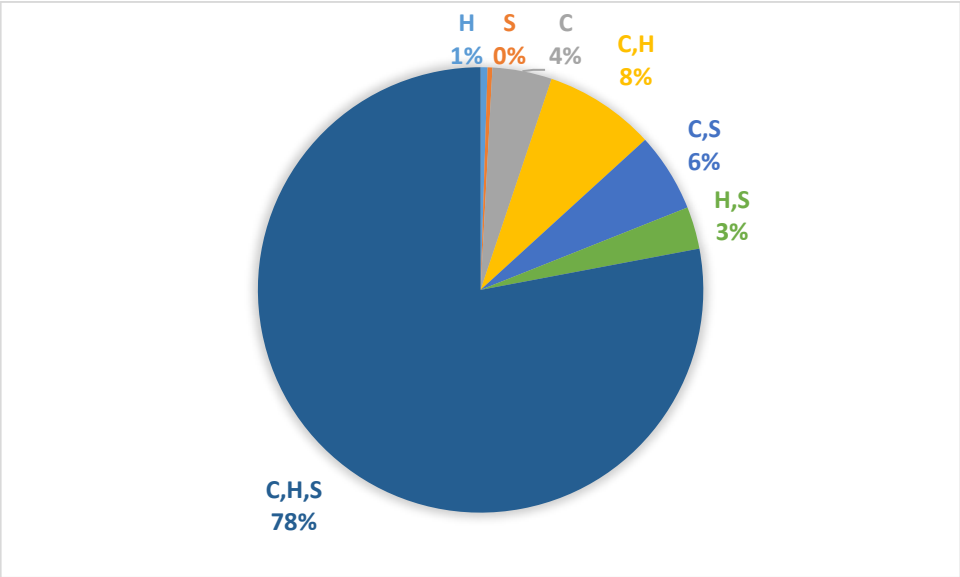

Figure 5.28 Secondary structure elements distribution in PRISM interface sides

5.6.2 Correlation between Protein Fragments

To see if there is a correlation between protein fragments consisting an interface descriptor vector, the Pearson correlation coefficient between fragments is calculated, for interface descriptors of fragment length six with bin size three, as follows:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

in which x and y are two sets of scores of the same size, and n is the number of pairs of scores.

The results show among 42 fragments in fragment library, there are strong correlations between fragment 1 and 4, fragment 1 and 3, fragment 4 and 7, and fragment 1 and 7 with correlation coefficients of 0.81, 0.68, 0.63, and 0.59 respectively. Fragments 1 and 7 are strands and fragment 3 and 4 are helices. It shows that some strand and helix structures are coexist on the protein interfaces. However, coils don't have such preferences as they are free to move in space.

5.7 Concluding Remarks

This study focused on offering a compact representation of protein structures/sub-structures which can be used for fast comparison and clustering. Such representation can be used as a filtering method which would remove the grossly dissimilar proteins before applying a time-consuming structural comparison method. To this end, a new representation is proposed based on proteins consisting small fragments. This representation uses a library of non-redundant protein structural fragments to create a protein descriptor vector. PRISM interface dataset is used to create this library. Small fragments extracted from the PRISM interface library are clustered based on their similarity using Geometric Hashing. These small fragment libraries are shown to cover all possible fragment shapes in proteins.

Analyzing the structures of protein fragments showed that protein interfaces are mostly consisted of helices. However, they just consisted a small portion of fragment libraries as their 3D shapes are mostly similar. On the other hand, coils have been shown to take variety of 3D shapes and therefore consisting majority of fragment libraries.

Using the library of non-redundant protein fragments, the descriptor vectors of protein interface structures are created. Cosine similarity is used to compare the descriptor vectors. By setting a threshold on Cosine similarity and investigating size differences between interface structures, the interface descriptor vectors can be used for interface clustering. The interface descriptor vectors created using fragment length six with bin size three displayed the best results in this step. By analyzing the PRISM interface dataset, it has been shown that interfaces are consisted of several repeating fragments instead of many unique fragment shapes. Labeling the protein fragments showed that the protein interface structures are mostly consisted of helices, though inside each interface all secondary structures can be seen.

The proposed method is applied to compare the protein surface structures with PRISM interfaces using Euclidean distance, Histogram intersection similarity, and Cosine similarity. Based on the results, shorter fragments got better results than longer fragments. Also, Cosine similarity and Euclidean distance could find the similarities better than Histogram intersection. As there is a huge size difference between protein surfaces and interfaces, several sub-surfaces are extracted from the protein surface, and their descriptor vectors are compared with interface side descriptor vectors. The comparison results for sub-surfaces of length 20 Å were more promising than the previous results. By applying the TM-align on protein surfaces and interfaces, it has been found that comparisons having TM-score higher than 0.5 got the Cosine similarity of ~0.7 on average in the proposed method.

Chapter 6

CONCLUSION

This dissertation mainly focused on the creation and analysis of breast cancer lung and brain metastasis PPI sub-networks to find important proteins and cellular pathways using functional and topological analysis. It also explains the importance of enriching the structural PPI networks with protein conformational changes. Besides, a new representation for protein structures is proposed which can be used for fast protein comparisons.

The PPI sub-networks of breast cancer lung and brain metastases are created by merging the data coming from different databases and scoring the PPIs. Afterwards, the interactions are filtered based on STRING [70] database medium confidence score, and the top 1000 interactions for each sub-network is selected. The largest connected component of each sub-network is examined for topological analysis. Using the degree and betweenness centralities of nodes, four novel proteins i.e. ELANE, MMP9, CCR1, and ITGA5 are found to have important roles in lung metastasis. Proteins CSNK2A1, PRKCA, ACTA1, PA2G4, RPL35, PSMA2, NCBP1, CMA1, and PFN1 are also found to be vital in brain metastases. The functional analysis of lung metastasis sub-network showed the importance of KEGG pathways associated with the immune system and infectious diseases in this phenotype. The same observation is detected in brain metastasis sub-network for pathways related to genetic information processing.

Traditional protein-protein interaction (PPI) networks which use a graph representation provide a general view of proteins and their interactions. However, this commonly used representation cannot show which interactions can happen simultaneously or they are mutually exclusive. In addition, it is not easy to track the effects of mutations on the protein interactions in these networks. To cope with these problems, a new representation for structural PPI networks is proposed which investigates the alternative conformations of proteins. To get the alternative

conformations of proteins, all available PDB structures of the proteins in the sub-networks are downloaded, and then clustered based on the sequence and structural similarities. Next, PRISM is applied to predict the possible interactions of proteins using their alternative conformations.

This representation helps to precisely analyze the protein interactions in the sub-networks. The effects of protein conformational changes on protein interactions and downstream signaling are observed and explained. As a case study, open and close conformations of protein KPNB1 are examined which directs it to bind to SNAI1 and SNUPN respectively. It also shows the limitations posed by the conformational changes of proteins which fluctuates their ability to stablish various interactions.

Using PRISM, or any other docking software, for structural network creation needs a long running time, weeks or months, to finish. This limitation forces the structural studies to focus on a small portion of the PPI network. On the other hand, PRISM prediction results are dependent to its template interface library which includes interfaces coming from PDB complexes till 2012. Combining newly-found interaction interfaces to the PRISM database would be very time consuming as it needs thousands of 3D protein sub-structure comparisons. Therefore, a compact representation of protein structures/sub-structures is proposed to speed up the structural comparisons. As this representation creates a coarse-grained 1D descriptor for protein structures, either interfaces or surfaces, it can make the process of protein comparison and clustering much faster. It uses a library of non-redundant protein structural fragments, which are extracted from PRISM interface structures, to create the protein descriptor vector.

This representation has been shown to be applicable in protein interface clustering by setting a threshold on the similarity measure and investigating size differences between interface structures. It can be also used for filtering the grossly dissimilar protein structures in dataset searches. However, as there is a huge size difference between protein surface and interface structures, comparing the descriptor vectors of sub-surfaces with the interfaces was more promising. These sub-surfaces are selected by investigating the 3D distance between the fragments on protein surface.

Bibliography

1. Vidal, M., M.E. Cusick, and A.L. Barabasi, *Interactome networks and human disease.* Cell, 2011. **144**(6): p. 986-98.
2. Tsai, C.J., B. Ma, and R. Nussinov, *Protein-protein interaction networks: how can a hub protein bind so many different partners?* Trends Biochem Sci, 2009. **34**(12): p. 594-600.
3. Kim, P.M., et al., *Relating three-dimensional structures to protein networks provides evolutionary insights.* Science, 2006. **314**(5807): p. 1938-41.
4. He, X. and J. Zhang, *Why do hubs tend to be essential in protein networks?* PLoS Genet, 2006. **2**(6): p. e88.
5. Tuncbag, N., et al., *Towards inferring time dimensionality in protein-protein interaction networks by integrating structures: the p53 example.* Mol Biosyst, 2009. **5**(12): p. 1770-8.
6. Kim, Y.A., et al., *Module cover - a new approach to genotype-phenotype studies.* Pac Symp Biocomput, 2013: p. 135-46.
7. Kuzu, G., et al., *Constructing structural networks of signaling pathways on the proteome scale.* Curr Opin Struct Biol, 2012. **22**(3): p. 367-77.
8. Wang, X., et al., *Three-dimensional reconstruction of protein networks provides insight into human genetic disease.* Nat Biotechnol, 2012. **30**(2): p. 159-64.
9. Mosca, R., et al., *dSysMap: exploring the edgetic role of disease mutations.* Nat Methods, 2015. **12**(3): p. 167-8.
10. Moretti, R., et al., *Community-wide evaluation of methods for predicting the effect of mutations on protein-protein interactions.* Proteins, 2013. **81**(11): p. 1980-7.
11. Bennett, W.S. and R. Huber, *Structural and functional aspects of domain motions in proteins.* CRC Crit Rev Biochem, 1984. **15**(4): p. 291-384.
12. Keskin, O., et al., *Principles of protein-protein interactions: what are the preferred ways for proteins to interact?* Chem Rev, 2008. **108**(4): p. 1225-44.
13. Kahsai, A.W., et al., *Monitoring protein conformational changes and dynamics using stable-isotope labeling and mass spectrometry.* Nat Protoc, 2014. **9**(6): p. 1301-19.
14. Galandrin, S., G. Oligny-Longpre, and M. Bouvier, *The evasive nature of drug efficacy: implications for drug discovery.* Trends Pharmacol Sci, 2007. **28**(8): p. 423-30.
15. Goh, C.S., D. Milburn, and M. Gerstein, *Conformational changes associated with protein-protein interactions.* Curr Opin Struct Biol, 2004. **14**(1): p. 104-9.
16. Cebecauer, M., et al., *Signalling complexes and clusters: functional advantages and methodological hurdles.* J Cell Sci, 2010. **123**(Pt 3): p. 309-20.
17. Karnoub, A.E. and R.A. Weinberg, *Ras oncogenes: split personalities.* Nat Rev Mol Cell Biol, 2008. **9**(7): p. 517-31.
18. Schubbert, S., K. Shannon, and G. Bollag, *Hyperactive Ras in developmental disorders and cancer.* Nat Rev Cancer, 2007. **7**(4): p. 295-308.
19. Hall, B.E., D. Bar-Sagi, and N. Nassar, *The structural basis for the transition from Ras-GTP to Ras-GDP.* Proc Natl Acad Sci U S A, 2002. **99**(19): p. 12138-42.
20. Lu, S., et al., *Ras Conformational Ensembles, Allostery, and Signaling.* Chem Rev, 2016. **116**(11): p. 6607-65.
21. Sussman, J.L., et al., *Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules.* Acta Crystallogr D Biol Crystallogr, 1998. **54**(Pt 6 Pt 1): p. 1078-84.

22. Kuzu, G., et al., *Exploiting conformational ensembles in modeling protein-protein interactions on the proteome scale.* J Proteome Res, 2013. **12**(6): p. 2641-53.
23. Wong, C.F., *Flexible receptor docking for drug discovery.* Expert Opin Drug Discov, 2015. **10**(11): p. 1189-200.
24. Bolia, A., Z.N. Gerek, and S.B. Ozkan, *BP-Dock: a flexible docking scheme for exploring protein-ligand interactions based on unbound structures.* J Chem Inf Model, 2014. **54**(3): p. 913-25.
25. Chang, C.W., C.W. Chou, and D.T. Chang, *CCProf: exploring conformational change profile of proteins.* Database (Oxford), 2016. **2016**.
26. Hrabe, T., et al., *PDBFlex: exploring flexibility in protein structures.* Nucleic Acids Res, 2016. **44**(D1): p. D423-8.
27. Monzon, A.M., et al., *CoDNaS 2.0: a comprehensive database of protein conformational diversity in the native state.* Database (Oxford), 2016. **2016**.
28. Pang, B., et al., *Fast protein binding site comparisons using visual words representation.* Bioinformatics, 2012. **28**(10): p. 1345-52.
29. Budowski-Tal, I., Y. Nov, and R. Kolodny, *FragBag, an accurate representation of protein structure, retrieves structural neighbors from the entire PDB quickly and accurately.* Proc Natl Acad Sci U S A, 2010. **107**(8): p. 3481-6.
30. Pang, B., et al., *PBSalign: Sequence-Independent Local Alignment of Protein Binding Sites.* 2013 Ieee International Conference on Bioinformatics and Biomedicine (Bibm), 2013.
31. Sanner, M.F., A.J. Olson, and J.C. Spehner, *Reduced surface: an efficient way to compute molecular surfaces.* Biopolymers, 1996. **38**(3): p. 305-20.
32. Konc, J. and D. Janezic, *ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment.* Bioinformatics, 2010. **26**(9): p. 1160-8.
33. Konc, J. and D. Janezic, *Protein-protein binding-sites prediction by protein surface structure conservation.* J Chem Inf Model, 2007. **47**(3): p. 940-4.
34. Shatsky, M., R. Nussinov, and H.J. Wolfson, *A method for simultaneous alignment of multiple protein structures.* Proteins, 2004. **56**(1): p. 143-56.
35. Nussinov, R. and H.J. Wolfson, *Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques.* Proc Natl Acad Sci U S A, 1991. **88**(23): p. 10495-9.
36. Wolfson, H.J. and I. Rigoutsos, *Geometric hashing: An overview.* Ieee Computational Science & Engineering, 1997. **4**(4): p. 10-21.
37. Holm, L. and C. Sander, *Dali/FSSP classification of three-dimensional protein folds.* Nucleic Acids Res, 1997. **25**(1): p. 231-4.
38. Singh, A.P. and D.L. Brutlag, *Hierarchical protein structure superposition using both secondary structure and atomic representations.* Proc Int Conf Intell Syst Mol Biol, 1997. **5**: p. 284-93.
39. Karim, R., et al., *CoMOGrad and PHOG: From Computer Vision to Fast and Accurate Protein Tertiary Structure Retrieval.* Sci Rep, 2015. **5**: p. 13275.
40. Mirceva, G., et al., *Efficient approaches for retrieving protein tertiary structures.* IEEE/ACM Trans Comput Biol Bioinform, 2012. **9**(4): p. 1166-79.
41. Gutierrez, F.I., et al., *Efficient and automated large-scale detection of structural relationships in proteins with a flexible aligner.* BMC Bioinformatics, 2016. **17**: p. 20.

42. Yang, H., R. Qureshi, and A. Sacan, *Protein surface representation and analysis by dimension reduction.* Proteome Sci, 2012. **10 Suppl 1**: p. S1.
43. Gao, M. and J. Skolnick, *iAlign: a method for the structural comparison of protein-protein interfaces.* Bioinformatics, 2010. **26**(18): p. 2259-2265.
44. Guyon, F. and P. Tuffery, *Fast protein fragment similarity scoring using a Binet-Cauchy kernel.* Bioinformatics, 2014. **30**(6): p. 784-91.
45. Binkowski, T.A. and A. Joachimiak, *Protein functional surfaces: global shape matching and local spatial alignments of ligand binding sites.* BMC Struct Biol, 2008. **8**: p. 45.
46. Binkowski, T.A., L. Adamian, and J. Liang, *Inferring functional relationships of proteins from local sequence and spatial surface patterns.* Journal of Molecular Biology, 2003. **332**(2): p. 505-526.
47. Wang, H.W., et al., *A local average distance descriptor for flexible protein structure comparison.* Bmc Bioinformatics, 2014. **15**.
48. Weigelt, B., J.L. Peterse, and L.J. Van't Veer, *Breast cancer metastasis: markers and models.* Nature reviews cancer, 2005. **5**(8): p. 591-602.
49. Jemal, A., et al., *Global cancer statistics.* CA Cancer J Clin, 2011. **61**(2): p. 69-90.
50. Nguyen, D.X., et al., *WNT/TCF signaling through LEF1 and HOXB9 mediates lung adenocarcinoma metastasis.* Cell, 2009. **138**(1): p. 51-62.
51. Lu, X. and Y. Kang, *Organotropism of breast cancer metastasis.* J Mammary Gland Biol Neoplasia, 2007. **12**(2-3): p. 153-62.
52. Hess, K.R., et al., *Metastatic patterns in adenocarcinoma.* Cancer, 2006. **106**(7): p. 1624-33.
53. Yun, J., et al., *Signalling pathway for RKIP and Let-7 regulates and predicts metastatic breast cancer.* EMBO J, 2011. **30**(21): p. 4500-14.
54. Sanz-Pamplona, R., et al., *A taxonomy of organ-specific breast cancer metastases based on a protein-protein interaction network.* Mol Biosyst, 2012. **8**(8): p. 2085-96.
55. Engin, H.B., et al., *Integrating structure to protein-protein interaction networks that drive metastasis to brain and lung in breast cancer.* PLoS One, 2013. **8**(11): p. e81035.
56. Guney, E., J. Garcia-Garcia, and B. Oliva, *GUILDify: a web server for phenotypic characterization of genes through biological data integration and network-based prioritization algorithms.* Bioinformatics, 2014. **30**(12): p. 1789-90.
57. Garcia-Garcia, J., et al., *Biana: a software framework for compiling biological interactions and analyzing networks.* BMC Bioinformatics, 2010. **11**: p. 56.
58. Guney, E. and B. Oliva, *Exploiting protein-protein interaction networks for genome-wide disease-gene prioritization.* PLoS One, 2012. **7**(9): p. e43557.
59. Minn, A.J., et al., *Genes that mediate breast cancer metastasis to lung.* Nature, 2005. **436**(7050): p. 518-24.
60. Bos, P.D., et al., *Genes that mediate breast cancer metastasis to the brain.* Nature, 2009. **459**(7249): p. 1005-9.
61. UniProt, C., *The Universal Protein Resource (UniProt).* Nucleic Acids Res, 2007. **35**(Database issue): p. D193-7.
62. Boeckmann, B., et al., *The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.* Nucleic Acids Res, 2003. **31**(1): p. 365-70.
63. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.

64. McKusick, V.A., *Mendelian Inheritance in Man and its online version, OMIM.* Am J Hum Genet, 2007. **80**(4): p. 588-604.
65. D'Eustachio, P., *Reactome knowledgebase of human biological pathways and processes.* Methods Mol Biol, 2011. **694**: p. 49-61.
66. Kerrien, S., et al., *The IntAct molecular interaction database in 2012.* Nucleic Acids Res, 2012. **40**(Database issue): p. D841-6.
67. Licata, L., et al., *MINT, the molecular interaction database: 2012 update.* Nucleic Acids Res, 2012. **40**(Database issue): p. D857-61.
68. Salwinski, L., et al., *The Database of Interacting Proteins: 2004 update.* Nucleic Acids Res, 2004. **32**(Database issue): p. D449-51.
69. Keshava Prasad, T.S., et al., *Human Protein Reference Database--2009 update.* Nucleic Acids Res, 2009. **37**(Database issue): p. D767-72.
70. Franceschini, A., et al., *STRING v9.1: protein-protein interaction networks, with increased coverage and integration.* Nucleic Acids Res, 2013. **41**(Database issue): p. D808-15.
71. Barabasi, A.L. and Z.N. Oltvai, *Network biology: understanding the cell's functional organization.* Nat Rev Genet, 2004. **5**(2): p. 101-13.
72. Gursoy, A., O. Keskin, and R. Nussinov, *Topological properties of protein interaction networks from a structural perspective.* Biochem Soc Trans, 2008. **36**(Pt 6): p. 1398-403.
73. Yu, H., et al., *The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics.* PLoS Comput Biol, 2007. **3**(4): p. e59.
74. Romieu-Mourez, R., et al., *Protein kinase CK2 promotes aberrant activation of nuclear factor-kappaB, transformed phenotype, and survival of breast cancer cells.* Cancer Res, 2002. **62**(22): p. 6770-8.
75. Lonne, G.K., et al., *PKCalpha expression is a marker for breast cancer aggressiveness.* Mol Cancer, 2010. **9**: p. 76.
76. Sanz-Pamplona, R., et al., *Expression of endoplasmic reticulum stress proteins is a candidate marker of brain metastasis in both ErbB-2+ and ErbB-2- primary breast tumors.* Am J Pathol, 2011. **179**(2): p. 564-79.
77. Choi, Y.P., et al., *Cancer-associated fibroblast promote transmigration through endothelial brain cells in three-dimensional in vitro models.* Int J Cancer, 2014. **135**(9): p. 2024-33.
78. Drasin, D.J., T.P. Robin, and H.L. Ford, *Breast cancer epithelial-to-mesenchymal transition: examining the functional consequences of plasticity.* Breast Cancer Res, 2011. **13**(6): p. 226.
79. Ou, K., et al., *Quantitative profiling of drug-associated proteomic alterations by combined 2-nitrobenzenesulfenyl chloride (NBS) isotope labeling and 2DE/MS identification.* J Proteome Res, 2006. **5**(9): p. 2194-206.
80. Naora, H., *Involvement of ribosomal proteins in regulating cell growth and apoptosis: translational modulation or recruitment for extraribosomal activity?* Immunol Cell Biol, 1999. **77**(3): p. 197-205.
81. Figeac, N., et al., *ErbB3 binding protein-1 (Ebp1) controls proliferation and myogenic differentiation of muscle stem cells.* Developmental Biology, 2014. **386**(1): p. 135-151.
82. Hamburger, A.W., *The role of ErbB3 and its binding partners in breast cancer progression and resistance to hormone and tyrosine kinase directed therapies.* J Mammary Gland Biol Neoplasia, 2008. **13**(2): p. 225-33.

83. Ruzzene, M. and L.A. Pinna, *Addiction to protein kinase CK2: a common denominator of diverse cancer cells?* Biochim Biophys Acta, 2010. **1804**(3): p. 499-504.
84. Blobe, G.C., L.M. Obeid, and Y.A. Hannun, *Regulation of protein kinase C and role in cancer biology.* Cancer Metastasis Rev, 1994. **13**(3-4): p. 411-31.
85. Menhofer, M.H., et al., *The actin targeting compound Chondramide inhibits breast cancer metastasis via reduction of cellular contractility.* PLoS One, 2014. **9**(11): p. e112542.
86. Gregory, A.D. and A.M. Houghton, *Tumor-associated neutrophils: new targets for cancer therapy.* Cancer Res, 2011. **71**(7): p. 2411-6.
87. Akizuki, M., et al., *Prognostic significance of immunoreactive neutrophil elastase in human breast cancer: long-term follow-up results in 313 patients.* Neoplasia, 2007. **9**(3): p. 260-4.
88. Velasco-Velazquez, M., et al., *CCR5 antagonist blocks metastasis of basal breast cancer cells.* Cancer Res, 2012. **72**(15): p. 3839-50.
89. Karagiannis, T., P. Boura, and A. Tsapas, *Safety of dipeptidyl peptidase 4 inhibitors: a perspective review.* Ther Adv Drug Saf, 2014. **5**(3): p. 138-46.
90. Cheng, H.C., et al., *Lung endothelial dipeptidyl peptidase IV promotes adhesion and metastasis of rat breast cancer cells via tumor cell surface-associated fibronectin.* J Biol Chem, 1998. **273**(37): p. 24207-15.
91. Bates, A.L., et al., *Stromal matrix metalloproteinase 2 regulates collagen expression and promotes the outgrowth of experimental metastases.* J Pathol, 2015. **235**(5): p. 773-83.
92. Rolli, M., et al., *Activated integrin alphavbeta3 cooperates with metalloproteinase MMP-9 in regulating migration of metastatic breast cancer cells.* Proc Natl Acad Sci U S A, 2003. **100**(16): p. 9482-7.
93. Yahiro, E., et al., *Chymase inhibitors.* Curr Pharm Des, 2013. **19**(17): p. 3065-71.
94. Janke, J., et al., *Suppression of tumorigenicity in breast cancer cells by the microfilament protein profilin 1.* J Exp Med, 2000. **191**(10): p. 1675-86.
95. Blouin, M.J., et al., *Germ line knockout of IGFBP-3 reveals influences of the gene on mammary gland neoplasia.* Breast Cancer Res Treat, 2015. **149**(3): p. 577-85.
96. Arjonen, A., et al., *Mutant p53-associated myosin-X upregulation promotes breast cancer invasion and metastasis.* J Clin Invest, 2014. **124**(3): p. 1069-82.
97. Sun, Y., et al., *CXCL12-CXCR4 axis promotes the natural selection of breast cancer cell metastasis.* Tumour Biol, 2014. **35**(8): p. 7765-73.
98. Sloan, E.K., K.L. Stanley, and R.L. Anderson, *Caveolin-1 inhibits breast cancer growth and metastasis.* Oncogene, 2004. **23**(47): p. 7893-7.
99. Porter, J.C. and N. Hogg, *Integrins take partners: cross-talk between integrins and other membrane receptors.* Trends Cell Biol, 1998. **8**(10): p. 390-6.
100. Callari, M., et al., *Gene expression analysis reveals a different transcriptomic landscape in female and male breast cancer.* Breast Cancer Res Treat, 2011. **127**(3): p. 601-10.
101. Chen, S., et al., *Proteomic identification of differentially expressed proteins associated with the multiple drug resistance in methotrexate-resistant human breast cancer cells.* Int J Oncol, 2014. **45**(1): p. 448-58.
102. Kroes, R.A., et al., *The identification of novel therapeutic targets for the treatment of malignant brain tumors.* Cancer Lett, 2000. **156**(2): p. 191-8.
103. Neumann, F., et al., *Human ribosomal protein L7 inhibits cell-free translation in reticulocyte lysates and affects the expression of nuclear proteins upon stable transfection into Jurkat T-lymphoma cells.* Nucleic Acids Res, 1995. **23**(2): p. 195-202.

104. Sharma, B., et al., *Targeting CXCR2 enhances chemotherapeutic response, inhibits mammary tumor growth, angiogenesis, and lung metastasis.* Mol Cancer Ther, 2013. **12**(5): p. 799-808.
105. Yee, A.J., et al., *The effect of versican G3 domain on local breast cancer invasiveness and bony metastasis.* Breast Cancer Res, 2007. **9**(4): p. R47.
106. Wu, Y., et al., *Overlapping gene expression profiles of cell migration and tumor invasion in human bladder cancer identify metallothionein 1E and nicotinamide N-methyltransferase as novel regulators of cell migration.* Oncogene, 2008. **27**(52): p. 6679-89.
107. Robinson, S.C., et al., *A chemokine receptor antagonist inhibits experimental breast tumor growth.* Cancer Res, 2003. **63**(23): p. 8360-5.
108. Raman, D., T. Sobolik-Delmaire, and A. Richmond, *Chemokines in health and disease.* Exp Cell Res, 2011. **317**(5): p. 575-89.
109. Choe, J., et al., *Translation initiation on mRNAs bound by nuclear cap-binding protein complex CBP80/20 requires interaction between CBP80/20-dependent translation initiation factor and eukaryotic translation initiation factor 3g.* J Biol Chem, 2012. **287**(22): p. 18500-9.
110. Sarvaiya, P.J., et al., *Chemokines in tumor progression and metastasis.* Oncotarget, 2013. **4**(12): p. 2171-85.
111. Frank, D.J., B.A. Edgar, and M.B. Roth, *The Drosophila melanogaster gene brain tumor negatively regulates cell growth and ribosomal RNA synthesis.* Development, 2002. **129**(2): p. 399-407.
112. Zhang, Y. and J. Skolnick, *TM-align: a protein structure alignment algorithm based on the TM-score.* Nucleic Acids Res, 2005. **33**(7): p. 2302-9.
113. Bhardwaj, N., et al., *Integration of protein motions with molecular networks reveals different mechanisms for permanent and transient interactions.* Protein Sci, 2011. **20**(10): p. 1745-54.
114. Baspinar, A., et al., *PRISM: a web server and repository for prediction of protein-protein interactions and modeling their 3D complexes.* Nucleic Acids Res, 2014. **42**(Web Server issue): p. W285-9.
115. Ogmen, U., et al., *PRISM: protein interactions by structural matching.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W331-6.
116. Tuncbag, N., et al., *Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM.* Nat Protoc, 2011. **6**(9): p. 1341-54.
117. Muratcioglu, S., et al., *Advances in template-based protein docking by utilizing interfaces towards completing structural interactome.* Curr Opin Struct Biol, 2015. **35**: p. 87-92.
118. Mashiach, E., R. Nussinov, and H.J. Wolfson, *FiberDock: a web server for flexible induced-fit backbone refinement in molecular docking.* Nucleic Acids Res, 2010. **38**(Web Server issue): p. W457-61.
119. Mashiach, E., R. Nussinov, and H.J. Wolfson, *FiberDock: Flexible induced-fit backbone refinement in molecular docking.* Proteins, 2010. **78**(6): p. 1503-19.
120. Cukuroglu, E., et al., *Non-redundant unique interface structures as templates for modeling protein interactions.* PLoS One, 2014. **9**(1): p. e86738.
121. Jakel, S. and D. Gorlich, *Importin beta, transportin, RanBP5 and RanBP7 mediate nuclear import of ribosomal proteins in mammalian cells.* EMBO J, 1998. **17**(15): p. 4491-502.

122. Paraskeva, E., et al., *CRM1-mediated recycling of snurportin 1 to the cytoplasm.* J Cell Biol, 1999. **145**(2): p. 255-64.
123. Mitrousis, G., et al., *Molecular basis for the recognition of snurportin 1 by importin beta.* J Biol Chem, 2008. **283**(12): p. 7877-84.
124. Cukuroglu, E., A. Gursoy, and O. Keskin, *HotRegion: a database of predicted hot spot clusters.* Nucleic Acids Res, 2012. **40**(Database issue): p. D829-33.
125. Pettersen, E.F., et al., *UCSF Chimera--a visualization system for exploratory research and analysis.* J Comput Chem, 2004. **25**(13): p. 1605-12.
126. Bhardwaj, A. and G. Cingolani, *Conformational selection in the recognition of the snurportin importin beta binding domain by importin beta.* Biochemistry, 2010. **49**(24): p. 5042-7.
127. Nagoshi, E. and Y. Yoneda, *Dimerization of sterol regulatory element-binding protein 2 via the helix-loop-helix-leucine zipper domain is a prerequisite for its nuclear localization mediated by importin beta.* Mol Cell Biol, 2001. **21**(8): p. 2779-89.
128. Cingolani, G., et al., *Structure of importin-beta bound to the IBB domain of importin-alpha.* Nature, 1999. **399**(6733): p. 221-9.
129. Liang, P., et al., *KPNB1, XPO7 and IPO8 mediate the translocation ofNF-kappaB/p65 into the nucleus.* Traffic, 2013. **14**(11): p. 1132-43.
130. Peinado, H., D. Olmeda, and A. Cano, *Snail, Zeb and bHLH factors in tumour progression: an alliance against the epithelial phenotype?* Nat Rev Cancer, 2007. **7**(6): p. 415-28.
131. Martin, T.A., et al., *Expression of the transcription factors snail, slug, and twist and their clinical significance in human breast cancer.* Ann Surg Oncol, 2005. **12**(6): p. 488-96.
132. Come, C., et al., *Snail and slug play distinct roles during breast carcinoma progression.* Clin Cancer Res, 2006. **12**(18): p. 5395-402.
133. Blanco, M.J., et al., *Correlation of Snail expression with histological grade and lymph node status in breast carcinomas.* Oncogene, 2002. **21**(20): p. 3241-6.
134. Olmeda, D., et al., *SNAI1 is required for tumor growth and lymph node metastasis of human breast carcinoma MDA-MB-231 cells.* Cancer Res, 2007. **67**(24): p. 11721-31.
135. Muller, A., et al., *Involvement of chemokine receptors in breast cancer metastasis.* Nature, 2001. **410**(6824): p. 50-6.
136. Teicher, B.A. and S.P. Fricker, *CXCL12 (SDF-1)/CXCR4 pathway in cancer.* Clin Cancer Res, 2010. **16**(11): p. 2927-31.
137. Wu, X., et al., *Chemokine receptors as targets for cancer therapy.* Curr Pharm Des, 2009. **15**(7): p. 742-57.
138. Epstein, R.J., *The CXCL12-CXCR4 chemotactic pathway as a target of adjuvant breast cancer therapies.* Nat Rev Cancer, 2004. **4**(11): p. 901-9.
139. Horuk, R., *Chemokine receptor antagonists: overcoming developmental hurdles.* Nat Rev Drug Discov, 2009. **8**(1): p. 23-33.
140. Veldkamp, C.T., et al., *Recognition of a CXCR4 sulfotyrosine by the chemokine stromal cell-derived factor-1alpha (SDF-1alpha/CXCL12).* J Mol Biol, 2006. **359**(5): p. 1400-9.
141. Drury, L.J., et al., *Monomeric and dimeric CXCL12 inhibit metastasis through distinct CXCR4 interactions and signaling pathways.* Proc Natl Acad Sci U S A, 2011. **108**(43): p. 17655-60.
142. Veldkamp, C.T., et al., *Structural basis of CXCR4 sulfotyrosine recognition by the chemokine SDF-1/CXCL12.* Sci Signal, 2008. **1**(37): p. ra4.

143. Takekoshi, T., et al., *A locked, dimeric CXCL12 variant effectively inhibits pulmonary metastasis of CXCR4-expressing melanoma cells due to enhanced serum stability.* Mol Cancer Ther, 2012. **11**(11): p. 2516-25.
144. Veldkamp, C.T., et al., *The monomer-dimer equilibrium of stromal cell-derived factor-1 (CXCL 12) is altered by pH, phosphate, sulfate, and heparin.* Protein Sci, 2005. **14**(4): p. 1071-81.
145. Ziarek, J.J., et al., *Sulfopeptide probes of the CXCR4/CXCL12 interface reveal oligomer-specific contacts and chemokine allostery.* ACS Chem Biol, 2013. **8**(9): p. 1955-63.
146. Pierce, B.G., et al., *ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers.* Bioinformatics, 2014. **30**(12): p. 1771-3.
147. Tovchigrechko, A. and I.A. Vakser, *GRAMM-X public web server for protein-protein docking.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W310-4.
148. Duhovny, D., R. Nussinov, and H.J. Wolfson, *Efficient unbound docking of rigid molecules.* Algorithms in Bioinformatics, Proceedings, 2002. **2452**: p. 185-200.
149. Schneidman-Duhovny, D., et al., *PatchDock and SymmDock: servers for rigid and symmetric docking.* Nucleic Acids Research, 2005. **33**: p. W363-W367.
150. O'Neill, G.M., et al., *A new central scaffold for metastasis: parsing HEF1/Cas-L/NEDD9.* Cancer Res, 2007. **67**(19): p. 8975-9.
151. Guerrero, M.S., J.T. Parsons, and A.H. Bouton, *Cas and NEDD9 Contribute to Tumor Progression through Dynamic Regulation of the Cytoskeleton.* Genes Cancer, 2012. **3**(5-6): p. 371-81.
152. Nourry, C., et al., *Direct interaction between Smad3, APC10, CDH1 and HEF1 in proteasomal degradation of HEF1.* BMC Cell Biol, 2004. **5**: p. 20.
153. Liu, X., et al., *A novel ability of Smad3 to regulate proteasomal degradation of a Cas family member HEF1.* EMBO J, 2000. **19**(24): p. 6759-69.
154. Frese, S., et al., *The phosphotyrosine peptide binding specificity of Nck1 and Nck2 Src homology 2 domains.* J Biol Chem, 2006. **281**(26): p. 18236-45.
155. Minegishi, M., et al., *Structure and function of Cas-L, a 105-kD Crk-associated substrate-related protein that is involved in beta 1 integrin-mediated signaling in lymphocytes.* J Exp Med, 1996. **184**(4): p. 1365-75.
156. Chakraborty, S., et al., *Protein structure quality assessment based on the distance profiles of consecutive backbone Calpha atoms.* F1000Res, 2013. **2**: p. 211.
157. Mistry, M. and P. Pavlidis, *Gene Ontology term overlap as a measure of gene functional similarity.* Bmc Bioinformatics, 2008. **9**.
158. Boyack, K.W., et al., *Clustering more than two million biomedical publications: comparing the accuracies of nine text-based similarity approaches.* PLoS One, 2011. **6**(3): p. e18029.
159. Gao, M. and J. Skolnick, *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, 2010. **107**(52): p. 22517-22522.
160. Hubbard, S.J. and J.M. Thornton, *Naccess*, in *Department of Biochemistry and Molecular Biology*. 1993, University College: London.
161. Zhang, Y. and J. Skolnick, *Scoring function for automated assessment of protein structure template quality.* Proteins, 2004. **57**(4): p. 702-10.
162. Xu, J. and Y. Zhang, *How significant is a protein structure similarity with TM-score = 0.5?* Bioinformatics, 2010. **26**(7): p. 889-95.

163. Guharoy, M. and P. Chakrabarti, *Secondary structure based analysis and classification of biological interfaces: identification of binding motifs in protein-protein interactions.* Bioinformatics, 2007. **23**(15): p. 1909-1918.
164. Keskin, O. and R. Nussinov, *Similar binding sites and different partners: implications to shared proteins in cellular pathways.* Structure, 2007. **15**(3): p. 341-54.
165. Yan, C., et al., *Characterization of protein-protein interfaces.* Protein J, 2008. **27**(1): p. 59-70.

Appendix 1

Supplementary Tables

Table A1 Interacting residues in protein complex 3lwwCD representing the interactions KPNB1-SNUPN.

Interface Name	Residue Number	Residue Type	Chain
3lwwCD	533	MET	C
3lwwCD	534	GLU	C
3lwwCD	583	LEU	C
3lwwCD	**586**	**ALA**	**C**
3lwwCD	**589**	**GLN**	**C**
3lwwCD	**626**	**GLU**	**C**
3lwwCD	**630**	**MET**	**C**
3lwwCD	637	GLU	C
3lwwCD	**665**	**GLN**	**C**
3lwwCD	**672**	**GLY**	**C**
3lwwCD	**676**	**ASP**	**C**
3lwwCD	**679**	**ARG**	**C**
3lwwCD	682	GLN	C
3lwwCD	712	GLN	C
3lwwCD	715	SER	C
3lwwCD	**719**	**ASP**	**C**
3lwwCD	**722**	**LEU**	**C**
3lwwCD	**763**	**GLU**	**C**

3lwwCD	**770**	**THR**	**C**
3lwwCD	**774**	**GLN**	**C**
3lwwCD	**820**	**GLY**	**C**
3lwwCD	**824**	**ASP**	**C**
3lwwCD	**861**	**LEU**	**C**
3lwwCD	**864**	**TRP**	**C**
3lwwCD	43	ARG	D
3lwwCD	44	ARG	D
3lwwCD	46	ARG	D
3lwwCD	47	LEU	D
3lwwCD	48	LEU	D
3lwwCD	50	LEU	D
3lwwCD	51	GLN	D
3lwwCD	52	LYS	D
3lwwCD	54	LYS	D
3lwwCD	55	ARG	D
3lwwCD	58	TYR	D
3lwwCD	59	VAL	D
3lwwCD	62	ALA	D
3lwwCD	63	ARG	D
3lwwCD	64	ARG	D

Bold residues show the common binding residues of KPNB1 to interact with SNUPN and KPNA2.

Table A2 Interacting residues in protein complexes 1qgkAB representing the interaction KPNB1-KPNA2.

Interface Name	Residue Number	Residue Type	Chain
1qgkAB	*281*	*GLU*	*A*
1qgkAB	*284*	*SER*	*A*
1qgkAB	*285*	*ASN*	*A*
1qgkAB	*288*	*ASP*	*A*
1qgkAB	339	ASP	A
1qgkAB	*340*	*ASP*	*A*
1qgkAB	*342*	*TRP*	*A*
1qgkAB	346	LYS	A
1qgkAB	*350*	*VAL*	*A*
1qgkAB	*353*	*MET*	*A*
1qgkAB	*354*	*LEU*	*A*
1qgkAB	*388*	*MET*	*A*
1qgkAB	*426*	*ASP*	*A*
1qgkAB	*427*	*THR*	*A*
1qgkAB	*430*	*TRP*	*A*
1qgkAB	469	ASN	A
1qgkAB	*472*	*TRP*	*A*
1qgkAB	*530*	*GLU*	*A*
1qgkAB	*582*	*SER*	*A*
1qgkAB	**586**	**ALA**	**A**
1qgkAB	**589**	**GLN**	**A**
1qgkAB	593	ARG	A
1qgkAB	*623*	*GLY*	*A*
1qgkAB	***626***	***GLU***	***A***
1qgkAB	*627*	*ASP*	*A*

1qgkAB	**630**	**MET**	**A**
1qgkAB	**665**	**GLN**	**A**
1qgkAB	**672**	**GLY**	**A**
1qgkAB	**676**	**ASP**	**A**
1qgkAB	**679**	**ARG**	**A**
1qgkAB	**719**	**ASP**	**A**
1qgkAB	**722**	**LEU**	**A**
1qgkAB	**763**	**GLU**	**A**
1qgkAB	767	GLU	A
1qgkAB	**770**	**THR**	**A**
1qgkAB	**774**	**GLN**	**A**
1qgkAB	**820**	**GLY**	**A**
1qgkAB	**824**	**ASP**	**A**
1qgkAB	860	THR	A
1qgkAB	**861**	**LEU**	**A**
1qgkAB	**864**	**TRP**	**A**
1qgkAB	11	ALA	B
1qgkAB	12	ALA	B
1qgkAB	13	ARG	B
1qgkAB	14	LEU	B
1qgkAB	16	ARG	B
1qgkAB	17	PHE	B
1qgkAB	18	LYS	B
1qgkAB	19	ASN	B
1qgkAB	20	LYS	B
1qgkAB	22	LYS	B
1qgkAB	27	MET	B
1qgkAB	28	ARG	B

1qgkAB	31	ARG	B
1qgkAB	32	ILE	B
1qgkAB	34	VAL	B
1qgkAB	35	ASN	B
1qgkAB	38	LEU	B
1qgkAB	39	ARG	B
1qgkAB	40	LYS	B
1qgkAB	43	LYS	B
1qgkAB	47	MET	B
1qgkAB	50	ARG	B
1qgkAB	51	ARG	B
1qgkAB	53	VAL	B
1qgkAB	54	SER	B

Bold and italic residues show the common binding residues of KPNB1 to interact with KPNA2 and SNUPN, and to interact with KPNA2 and SNAI1 respectively.

Table A3 Interacting residues in protein complexes 3w5kAB representing the interaction KPNB1-SNAI1.

Interface Name	Residue Number	Residue Type	Chain
3w5kAB	200	ASN	A
3w5kAB	242	VAL	A
3w5kAB	243	LYS	A
3w5kAB	246	SER	A
3w5kAB	247	LEU	A
3w5kAB	*281*	*GLU*	*A*
3w5kAB	*284*	*SER*	*A*
3w5kAB	*285*	*ASN*	*A*
3w5kAB	*288*	*ASP*	*A*
3w5kAB	289	GLU	A
3w5kAB	292	ASP	A
3w5kAB	340	ASP	A
3w5kAB	*342*	*TRP*	*A*
3w5kAB	*350*	*VAL*	*A*
3w5kAB	*353*	*MET*	*A*
3w5kAB	*354*	*LEU*	*A*
3w5kAB	*388*	*MET*	*A*
3w5kAB	395	GLU	A
3w5kAB	399	PRO	A
3w5kAB	*426*	*ASP*	*A*
3w5kAB	*427*	*THR*	*A*
3w5kAB	*430*	*TRP*	*A*
3w5kAB	437	GLU	A
3w5kAB	438	LEU	A
3w5kAB	440	PRO	A

3w5kAB	441	GLU	A
3w5kAB	*472*	*TRP*	*A*
3w5kAB	479	GLU	A
3w5kAB	480	ALA	A
3w5kAB	483	GLU	A
3w5kAB	491	GLN	A
3w5kAB	*530*	*GLU*	*A*
3w5kAB	572	SER	A
3w5kAB	575	ILE	A
3w5kAB	576	GLN	A
3w5kAB	578	ASN	A
3w5kAB	579	ASP	A
3w5kAB	*582*	*SER*	*A*
3w5kAB	583	LEU	A
3w5kAB	621	SER	A
3w5kAB	622	GLY	A
3w5kAB	623	GLY	A
3w5kAB	624	VAL	A
3w5kAB	*626*	*GLU*	*A*
3w5kAB	*627*	*ASP*	*A*
3w5kAB	157	LYS	B
3w5kAB	165	SER	B
3w5kAB	166	LEU	B
3w5kAB	167	GLY	B
3w5kAB	170	LYS	B
3w5kAB	173	ILE	B
3w5kAB	191	ARG	B
3w5kAB	192	PRO	B

3w5kAB	193	TRP	B
3w5kAB	196	GLN	B
3w5kAB	200	ARG	B
3w5kAB	220	ARG	B
3w5kAB	221	SER	B
3w5kAB	222	ASN	B
3w5kAB	224	ARG	B
3w5kAB	225	ALA	B
3w5kAB	228	GLN	B
3w5kAB	229	THR	B
3w5kAB	234	LYS	B
3w5kAB	237	GLN	B
3w5kAB	239	GLN	B
3w5kAB	240	ALA	B
3w5kAB	241	CYS	B
3w5kAB	242	ALA	B
3w5kAB	243	ARG	B
3w5kAB	246	SER	B
3w5kAB	247	ARG	B
3w5kAB	249	SER	B
3w5kAB	250	LEU	B
3w5kAB	253	LYS	B
3w5kAB	261	GLY	B
3w5kAB	262	CYS	B
3w5kAB	263	PRO	B
3w5kAB	264	ARG	B

Italic residues show the common binding residues of KPNB1 to interact with SNAI1 and KPNA2.

Table A4 Predicted binding residues of $CXCL12_{WT}$ (PDB ID: 2j7zA) homodimer and $CXCL12_{H25R}$ (PDB ID: 2kolA) homodimer.

2j7zA		**2j7zA**	
ILE	28	TYR	61
ILE	28	LEU	62
LEU	29	TYR	61
ILE	28	ALA	65
LYS	68	ALA	35
LYS	27	LEU	26
ASN	30	VAL	23
LYS	24	ASN	30
LEU	29	HIS	25
ILE	38	LEU	66
ILE	38	ALA	65
ALA	65	GLN	37
ASN	30	TYR	61
ILE	28	ILE	28
ILE	28	HIS	25
ILE	28	LEU	26
TYR	61	LEU	29
TYR	61	ILE	28
ALA	65	ILE	38
GLN	37	ALA	65
HIS	25	LYS	27
LEU	62	LEU	62
ALA	65	PRO	53
ASN	30	LYS	24
HIS	25	LEU	29
HIS	25	ILE	28

2kolA		**2kolA**	
ASN	67	GLU	63
LEU	29	ARG	25
LEU	26	ASN	30
LEU	29	LYS	27
LEU	29	LEU	26
ASN	30	LEU	62
ILE	28	ILE	28
ASN	33	ARG	25
ILE	28	LYS	27
ASN	30	LEU	26
ASN	30	ARG	25
ARG	25	LEU	29
THR	31	ARG	25
GLU	63	GLU	63
LYS	68	GLU	63
GLU	63	ASN	67
GLN	59	ASN	67
LYS	27	LYS	27
LEU	36	LEU	66
LYS	68	GLN	59
LEU	26	LEU	29
LYS	27	LEU	29
LYS	27	ILE	28
LEU	66	ASN	30
LEU	66	ALA	35
LEU	66	LEU	36

ASN	30	ARG	20
ARG	20	ASN	30
LEU	62	ILE	28
ILE	58	LEU	66
LYS	27	LYS	27
LYS	24	LEU	29
LYS	27	HIS	25
LEU	26	LEU	26
LEU	26	LYS	27
GLN	37	LYS	68
LEU	26	ILE	28
LYS	68	CYS	34
LEU	66	ILE	38
PRO	32	LYS	68
VAL	23	ASN	30
PRO	53	LEU	66
PRO	53	ALA	65
ALA	35	LYS	68
LEU	66	ILE	58
LEU	66	PRO	53
CYS	34	LYS	68
GLN	37	LYS	64
HIS	25	ASN	30
TYR	61	ASN	30

PRO	10	ARG	25
ASN	67	GLN	59
PRO	32	ARG	25
PRO	32	LYS	24
ARG	8	ARG	25
ARG	25	ASN	30
LEU	62	ASN	30

Table A5 Predicted binding residues of $CXCL12_{WT}$ (PDB ID: 2j7zA)-$CXCR4_{NT}$ (PDB ID: 2k03B), and $CXCL12_{H25R}$ (PDB ID: 2kolA)-$CXCR4_{NT}$ (2k03B).

2j7zA		**2k03B**	
GLN	**48**	**TYR**	**112**
LEU	5	THR	113
PHE	14	SER	118
PHE	14	GLY	119
GLN	59	PHE	129
LYS	56	PHE	129
LYS	54	PHE	129
ARG	8	GLU	115
LYS	54	GLU	126
THR	31	ILE	106
THR	**31**	**TYR**	**107**
LYS	1	GLU	115
LEU	**29**	**TYR**	**112**
VAL	49	SER	118
PHE	14	MET	124
PHE	14	SER	123
CYS	9	GLU	114
PHE	14	ASP	120
PRO	32	SER	105
PRO	32	ILE	104
PRO	32	ILE	106
LYS	54	PRO	127
GLU	15	SER	118
PRO	32	GLU	102
PRO	10	GLU	114
LYS	54	ALA	128

2kolA		**2k03B**	
ASP	52	LYS	125
ASP	52	GLU	126
GLN	48	SER	118
TYR	7	GLU	115
TYR	7	GLU	114
PHE	13	SER	123
LEU	**29**	**TYR**	**112**
VAL	49	SER	118
VAL	49	GLY	119
PHE	14	LYS	125
PHE	14	MET	124
PHE	14	SER	123
PRO	32	SER	105
PRO	32	ILE	104
PRO	**32**	**TYR**	**107**
PRO	32	ILE	106
LYS	54	PRO	127
PRO	32	GLY	103
PRO	32	GLU	102
LYS	54	PHE	129
LYS	54	ALA	128
ASN	33	GLY	103
ASN	33	ILE	106
ASN	33	ILE	104
CYS	50	SER	118
CYS	50	GLY	119

PRO	10	THR	113
ASN	33	GLY	103
LYS	**27**	**TYR**	**112**
ASN	33	ILE	106
ASN	33	ILE	104
SER	6	GLU	115
VAL	3	GLU	115
VAL	3	GLU	114
VAL	3	THR	113
VAL	**39**	**TYR**	**112**
ASN	**30**	**TYR**	**107**
ASN	30	ILE	106
PHE	13	GLY	119
PHE	13	SER	118
SER	4	THR	113

LYS	56	PHE	129
PHE	13	GLY	119

Bold rows represent the interactions of three sulfation sites on $CXCR4_{NT}$ at positions 7, 12, and 21.

Table A6 Predicted binding residues of $CXCL12_{WT}$:$CXCR4_{NT}$-$CXCL12_{WT}$ trimer structure.

$CXL12_{WT}$:$CXCR4_{NT}$		**CXCL12**	
ILE	28	TYR	61
ILE	28	LEU	62
LEU	29	TYR	61
LYS	68	ALA	35
MET	101	LYS	68
ASN	30	VAL	23
LYS	24	ASN	30
LEU	29	HIS	25
ILE	38	LEU	66
ILE	38	ALA	65
ALA	65	GLN	37
ASN	30	TYR	61
ILE	28	ILE	28
ILE	28	HIS	25
ASN	30	LYS	43
LEU	66	ILE	58
TYR	61	LEU	29
ILE	28	LEU	26
ALA	65	ILE	38
LEU	62	LEU	66
HIS	25	LYS	27
LEU	62	LEU	62
ALA	65	PRO	53
ASN	30	LYS	24
HIS	25	LEU	29

HIS	25	ILE	28
TYR	61	ILE	28
ARG	20	ASN	30
GLU	63	GLU	63
LEU	62	ILE	28
ILE	58	LEU	66
GLN	59	LEU	66
LYS	27	LEU	26
LEU	26	LEU	26
LEU	26	LYS	27
TYR	107	LYS	24
LEU	26	ILE	28
LEU	66	ILE	38
LYS	54	ASN	67
LYS	54	LEU	66
LYS	54	LYS	68
VAL	23	ASN	30
PRO	127	LYS	68
PRO	53	LEU	66
GLU	126	LYS	68
LEU	66	GLN	59
PRO	53	ALA	65
ASN	67	PRO	53
TYR	61	ASN	30
LEU	66	LYS	54
LEU	66	PRO	53
LYS	68	GLN	37
HIS	25	ASN	30

Table A7 Common interface residues in PRISM and other docking methods predictions for interactions of KPNB1 with SNUPN and KPNA2.

Docking Tools	Common Interface Residues in Interaction
	KPNB1-SNUPN (3w5kA-3lwwD)
PRISM and PatchDock	A:630,A:637,A:672,A:676,A:679,A:712,A:715,A:719,A:722,A:767,A:770,A:774,A:820,A:824,A:857,A:861,A:864,D:43,D:44,D:47,D:50,D:51,D:54,D:55,D:58,D:59,D:62,D:63
PRISM and GRAMM-X	A:665,D:50,D:56,D:63
PRISM and Zdock	A:626,A:630,A:637,A:665,A:672,A:676,A:679,A:712,A:715,A:719,A:722,A:767,A:770,A:774,A:820,A:824,A:857,A:861,A:864,D:44,D:47,D:48,D:50,D:51,D:54,D:55,D:56,D:58,D:59,D:62,D:63,D:64
	KPNB1-SNUPN (2q5dB-3lwwD)
PRISM and PatchDock	B:630,B:633,B:665,B:676,B:712,D:40,D:43,D:55,D:59,D:62,D:63,D:64
PRISM and GRAMM-X	B:857,D:40,D:43,D:44,D:62,D:63,D:64
PRISM and Zdock	B:589,B:593,B:679,D:43,D:44,D:47,D:51,D:55,D:59,D:62,D:63,D:64
	KPNB1-SNUPN (1qgkA-3lwwD)
PRISM and PatchDock	A:586,A:589,A:627,A:630,A:665,A:668,A:672,A:679,A:712,A:715,D:43,D:44,D:47,D:50,D:51,D:54,D:55,D:58,D:59,D:62,D:63
PRISM and GRAMM-X	A:679,A:682,A:719,A:722,A:723,A:770,A:774,A:824,D:51,D:55,D:58,D:59,D:62,D:63
PRISM and Zdock	A:593,A:630,A:633,A:637,A:665,A:672,A:676,A:679,A:712,D:43,D:44,D:47,D:50,D:54,D:55,D:58,D:59,D:62,D:63
	KPNB1-SNUPN (1qgrA-3lwwD)
PRISM and PatchDock	A:586,A:589,A:593,A:627,A:630,A:637,A:676,A:679,A:715,A:722,A:767,A:770,A:774,A:777,A:779,A:820,A:824,

	A:827,A:861,A:864,D:43,D:47,D:50,D:51,D:54,D:55,D:56,D:58,D:59,D:61,D:62,D:63,D:64
PRISM and GRAMM-X	A:679,A:682,A:719,A:722,A:767,A:770,A:774,A:779,A:820,A:824,A:861,A:864,D:51,D:52,D:55,D:56,D:58,D:59,D:61,D:62,D:63,D:64
PRISM and Zdock	A:626,A:630,A:637,A:665,A:672,A:676,A:679,A:682,A:715,A:719,A:722,A:763,A:767,A:770,A:774,A:779,A:824,A:861,A:864,D:43,D:44,D:47,D:51,D:52,D:54,D:55,D:56,D:58,D:59,D:62,D:63,D:64
	KPNB1-KPNA2 (3lwwC-1qgkB)
PRISM and PatchDock	B:11,B:12,B:13,B:16,B:28,B:36,B:39,B:40,B:42,B:44,B:51,B:53,B:54,C:340,C:589,C:630,C:679,C:719,C:767
PRISM and GRAMM-X	B:11,B:12,B:14,B:19,B:20,B:31,B:36,B:39,B:40,B:43,B:53,B:54,C:346,C:350,C:388,C:430
PRISM and Zdock	B:11,B:12,B:13,B:14,B:16,B:17,B:18,B:19,B:20,B:28,B:31,B:39,B:40,B:47,B:50,C:288,C:339,C:340,C:342,C:346,C:350,C:388,C:427,C:430,C:589,C:630,C:679,C:779,C:861
	KPNB1-KPNA2 (3w5kA-1qgkB)
PRISM and PatchDock	A:350,A:434,A:582,A:626,A:630,A:665,A:668,A:672,A:711,A:712,A:715,A:767,B:11,B:12,B:13,B:14,B:20,B:23,B:24,B:25,B:28,B:39,B:52
PRISM and GRAMM-X	B:11,B:12,B:13,B:14,B:23,B:25,B:51,B:52
PRISM and Zdock	A:284,A:288,A:340,A:342,A:350,A:354,A:388,A:430,A:469,A:575,A:576,A:579,A:855,B:11,B:12,B:13,B:14,B:17,B:18,B:19,B:24,B:25,B:28,B:31,B:50,B:51,B:52
	KPNB1-KPNA2 (2q5dB-1qgkB)
PRISM and PatchDock	A:593,A:679,A:722,B:12,B:13,B:14,B:16,B:17,B:18,B:19,B:20,B:22,B:23,B:24,B:27,B:28,B:31,B:35,B:38,B:42,B:46,B:48,B:50,B:51,B:53,B:54
PRISM and GRAMM-X	A:593,A:679,B:16,B:18,B:19,B:22,B:24,B:27,B:28,B:31,B:35,B:48,B:51,B:53,B:54
PRISM and Zdock	A:340,A:341,A:342,A:346,A:349,A:350,A:353,A:388,A:430,A:472,A:526,A:530,A:533,A:583,A:593,A:679,A:722,A:770,A:857,A:864,B:12,B:13,B:14,B:16,B:17,B:18,B:19,B:20,B:22,B:24,B:27,B:28,B:31,B:35,B:43,B:48,B:51,B:53,B:54
	KPNB1-KPNA2 (1qgrA-1qgkB)

PRISM and PatchDock	A:626,A:672,A:676,A:679,A:682,A:719,A:763,B:11,B:12,B:13,B:16,B:28,B:39,B:40,B:51,B:53
PRISM and GRAMM-X	B:28,B:39,B:42,B:46,B:52,B:53
PRISM and Zdock	A:626,A:672,A:676,A:679,A:682,A:719,A:763,B:11,B:12,B:13,B:16,B:28,B:39,B:40,B:51,B:53

Table A8 Common interface residues in PRISM and other docking methods predictions for $CXCL12_{WT}$ and $CXCL12_{H25R}$ homodimers and their interactions with $CXCR4_{NT}$.

Docking Tools	Common Interface Residues for interaction
	$CXCL12_{WT}$ homodimer (2j7zA-2j7zA)
PRISM and PatchDock	A:23,A:24,A:25,A:26,A:27,A:28,A:29,A:30,A:35,A:37,A:38,A:53,A:58,A:61,A:62,A:65,A:66,A:68,B:23,B:24,B:25,B:26,B:27,B:28,B:29,B:30,B:35,B:37,B:38,B:53,B:58,B:61,B:62,B:65,B:66,B:68
PRISM and GRAMM-X	A:23,A:24,A:25,A:26,A:27,A:28,A:29,A:30,A:35,A:37,A:38,A:53,A:58,A:61,A:62,A:65,A:66,A:68,B:23,B:24,B:25,B:26,B:27,B:28,B:29,B:30,B:35,B:37,B:38,B:53,B:58,B:61,B:62,B:65,B:66,B:68
PRISM and Zdock	A:34,A:35,A:53,B:32,B:34,B:35,B:53
	$CXCL12_{H25R}$ homodimer (2kolA-2kolA)
PRISM and PatchDock	A:25,A:27,A:29,A:30,B:10,B:25,B:27,B:29,B:30,B:31,B:32
PRISM and GRAMM-X	No common residues.
PRISM and Zdock	No common residues.
	$CXCL12_{WT}$-$CXCR4_{NT}$ (2j7zA-2k03B)
PRISM and PatchDock	A:1,A:4,A:5,A:6,A:8,A:13,A:14,A:32,A:54,A:56,A:59,B:106,B:107,B:112,B:113,B:114,B:118,B:120,B:123,B:124,B:126,B:127,B:128,B:129
PRISM and GRAMM-X	A:1,A:3,A:10,A:13,A:27,A:29,A:30,B:102,B:103,B:106,B:112,B:118,B:119,B:120,B:123,B:124,B:126,B:127,B:128,B:129
PRISM and Zdock	A:1,A:4,A:10,A:13,A:27,A:29,A:30,B:102,B:103,B:114,B:119,B:120,B:123,B:124,B:126,B:127
	$CXCL12_{H25R}$-$CXCR4_{NT}$ (2kolA-2k03B)
PRISM and PatchDock	A:7,A:13,A:32,A:33,A:49,A:54,A:56,B:102,B:103,B:104,B:105,B:106,B:107,B:112,B:114,B:118,B:119,B:124,B:125,B:126,B:128,B:129
PRISM and GRAMM-X	A:7,A:13,A:14,A:29,A:32,A:48,A:49,A:52,A:54,A:56,B:103,B:104,B:106,B:107,B:112,B:114,B:115,B:118,B:119,B:123,B:124,B:125,B:126,B:127,B:128,B:129
PRISM and Zdock	A:7,A:14,A:32,A:33,A:49,A:56,B:102,B:103,B:104,B:106,B:107,B:112,B:114,B:119,B:124,B:125,B:126,B:128

	$CXCL12_{wt}$:CXCR4NT- $CXCL12_{wt}$ (2j7zA:2k03B-2j7zA)
PRISM and PatchDock	A:24,A:25,A:27,A:29,A:30,C:24,C:25,C:27,C:29,C:30
PRISM and GRAMM-X	A:24,A:25,A:26,A:27,A:28,A:30,A:54,A:59,A:61,A:62,A:65,A:66,A:68,C:24,C:25, C:26,C:27,C:28,C:30,C:37,C:54,C:59,C:61,C:62,C:65,C:66,C:68
PRISM and Zdock	A:53,A:54,A:59,C:35,C:53,C:54,C:59

Table A9 Predicted binding residues of NEDD9 (PDB ID: 2l81A)-SMAD3 (PDB ID: 1mk2A), and NEDD9 (PDB ID: 2l81A)-NCK1 (2ci8A) interactions.

1mk2A		2l81A	
SER	253	**ALA**	**32**
SER	253	**GLN**	**31**
SER	253	**MET**	**35**
PRO	335	LEU	43
PRO	252	**ARG**	**28**
PRO	252	GLN	27
LEU	307	**SER**	**38**
TYR	296	**ALA**	**32**
TYR	297	**ARG**	**28**
CYS	337	GLU	60
PHE	303	**MET**	**35**
GLY	336	LEU	43
TYR	296	**MET**	**35**
TYR	296	**ARG**	**28**
GLY	336	ALA	64
ASP	309	ALA	42
LEU	307	ALA	42
ARG	291	ALA	42
PRO	335	HIS	57
ARG	294	**MET**	**35**
PRO	252	**GLN**	**31**
GLU	227	**ARG**	**28**
ASN	338	**LYS**	**67**
LEU	295	**MET**	**35**
GLU	305	**PHE**	**71**

2l81A		2ci8A	
ARG	**28**	TYR	284
ARG	**28**	VAL	287
LEU	70	HIS	301
LEU	70	GLY	300
LEU	70	ASP	304
HIS	77	ARG	299
GLU	74	GLY	300
PHE	**71**	ASP	304
PHE	**71**	PHE	305
LEU	29	ARG	299
MET	**35**	TRP	282
MET	**35**	PRO	281
PHE	78	ARG	299
GLU	34	PRO	281
GLN	**31**	TYR	284
GLN	**31**	TYR	283
ARG	28	GLN	291
MET	**35**	LEU	306
LYS	**67**	PHE	305
MET	35	ILE	307
MET	**35**	PHE	305
ALA	**32**	TYR	284
SER	**38**	PRO	281
SER	**38**	TRP	282
LEU	70	ARG	299

PRO	335	GLU	60
ARG	291	MET	41
ARG	294	**SER**	**38**
GLU	305	SER	39
GLU	305	GLY	36
LEU	307	SER	39
ILE	298	**ARG**	**28**
GLU	305	**MET**	**35**
GLY	336	SER	39

GLU	74	ARG	299

Bold residues show the common binding residues of NEDD9 to interact with SMAD3 and NCK1.

Appendix 2

Supplementary Figures

```
ATOM    146  CA  TYR A  18     -12.824  46.014  19.476  1.00 22.28           C
ATOM    158  CA  LYS A  19      -9.162  45.021  19.819  1.00 36.10           C
ATOM    167  CA  ASP A  20      -7.612  43.744  23.062  1.00 35.35           C
ATOM    175  CA  THR A  21      -4.270  45.016  24.342  1.00 75.22           C
ATOM    182  CA  GLU A  22      -2.685  42.544  21.927  1.00 62.89           C
ATOM    191  CA  GLY A  23      -4.590  43.844  19.023  1.00 41.88           C
ATOM    195  CA  TYR A  24      -6.950  40.881  19.030  1.00 27.87           C
ATOM    207  CA  TYR A  25     -10.626  40.951  18.324  1.00 23.58           C
ATOM    219  CA  THR A  26     -12.597  40.966  21.335  1.00 24.02           C
ATOM    268  CA  LEU A  33     -12.728  36.195  19.175  1.00 31.66           C
ATOM    276  CA  THR A  34      -9.665  34.753  17.515  1.00 33.57           C
ATOM    283  CA  LYS A  35      -5.948  35.244  17.302  1.00 31.07           C
ATOM    292  CA  SER A  36      -5.469  34.299  13.569  1.00 42.92           C
ATOM    298  CA  PRO A  37      -3.845  36.848  11.215  1.00 43.79           C
ATOM    305  CA  SER A  38      -6.021  35.657   8.563  1.00 41.39           C
ATOM    311  CA  LEU A  39      -9.078  37.570   9.943  1.00 40.77           C
```

```
ATOM    146  CA  TYR A  18     -12.824  46.014  19.476  1.00 22.28           C
ATOM    158  CA  LYS A  19      -9.162  45.021  19.819  1.00 36.10           C
ATOM    167  CA  ASP A  20      -7.612  43.744  23.062  1.00 35.35           C
ATOM    175  CA  THR A  21      -4.270  45.016  24.342  1.00 75.22           C
```

```
ATOM    158  CA  LYS A  19      -9.162  45.021  19.819  1.00 36.10           C
ATOM    167  CA  ASP A  20      -7.612  43.744  23.062  1.00 35.35           C
ATOM    175  CA  THR A  21      -4.270  45.016  24.342  1.00 75.22           C
ATOM    182  CA  GLU A  22      -2.685  42.544  21.927  1.00 62.89           C
```

```
ATOM    167  CA  ASP A  20      -7.612  43.744  23.062  1.00 35.35           C
ATOM    175  CA  THR A  21      -4.270  45.016  24.342  1.00 75.22           C
ATOM    182  CA  GLU A  22      -2.685  42.544  21.927  1.00 62.89           C
ATOM    191  CA  GLY A  23      -4.590  43.844  19.023  1.00 41.88           C
```

```
ATOM    175  CA  THR A  21      -4.270  45.016  24.342  1.00 75.22           C
ATOM    182  CA  GLU A  22      -2.685  42.544  21.927  1.00 62.89           C
ATOM    191  CA  GLY A  23      -4.590  43.844  19.023  1.00 41.88           C
ATOM    195  CA  TYR A  24      -6.950  40.881  19.030  1.00 27.87           C
```

```
ATOM    182  CA  GLU A  22      -2.685  42.544  21.927  1.00 62.89           C
ATOM    191  CA  GLY A  23      -4.590  43.844  19.023  1.00 41.88           C
ATOM    195  CA  TYR A  24      -6.950  40.881  19.030  1.00 27.87           C
ATOM    207  CA  TYR A  25     -10.626  40.951  18.324  1.00 23.58           C
```

```
ATOM    191  CA  GLY A  23      -4.590  43.844  19.023  1.00 41.88           C
ATOM    195  CA  TYR A  24      -6.950  40.881  19.030  1.00 27.87           C
ATOM    207  CA  TYR A  25     -10.626  40.951  18.324  1.00 23.58           C
ATOM    219  CA  THR A  26     -12.597  40.966  21.335  1.00 24.02           C
```

```
ATOM    268  CA  LEU A  33     -12.728  36.195  19.175  1.00 31.66           C
ATOM    276  CA  THR A  34      -9.665  34.753  17.515  1.00 33.57           C
ATOM    283  CA  LYS A  35      -5.948  35.244  17.302  1.00 31.07           C
ATOM    292  CA  SER A  36      -5.469  34.299  13.569  1.00 42.92           C
```

Figure A1 A sample PRISM interface side file and the first seven extracted overlapping fragments of it.

```
ATOM   2239  CA  GLY A 379     -32.838  87.937  85.802  1.00 49.92
ATOM   2243  CA  VAL A 380     -36.109  87.779  87.741  1.00 57.06
ATOM   2250  CA  GLY A 381     -36.180  84.443  89.625  1.00 60.48
ATOM   2254  CA  GLU A 382     -34.243  81.196  90.133  1.00 62.06
```

```
3   0   0   (1,4,2) (1,4,3) (4,1,2) (4,1,3)
0   2   0   (2,1,4)
0   1   0   (1,2,3) (2,1,3) (2,1,3) (2,3,1) (2,4,1) (3,1,4) (3,4,2)
0   -0  0   (1,3,4) (1,4,3) (2,1,4) (2,3,4) (2,4,1) (2,4,3) (3,1,2) (3,1,4)
-1  2   0   (3,4,1)
-1  1   0   (3,2,4) (3,4,2)
1   0   0   (1,2,3) (1,2,4) (1,2,4) (1,3,2) (1,4,2) (2,1,3) (2,1,4) (2,3,1)
1   1   0   (1,3,2) (1,4,2) (3,1,2) (4,1,3) (4,3,2)
1   2   0   (1,2,4)
2   2   0   (4,3,1)
2   1   0   (1,2,3) (1,3,4) (1,4,3) (2,3,4) (3,2,1) (4,1,2) (4,2,1) (4,3,2)
2   0   0   (1,3,2) (1,3,4) (2,4,1) (2,4,3) (3,1,2) (3,1,4) (4,2,1) (4,2,3)
```

Figure A3 A sample seed fragment and its geometric hash table. Some of the RSs are not shown in the figure because of width limitation.

```
1041AB_A.int    2:4,3:2,4:2,5:3,7:3,8:2,11:3,15:1,21:1,33:1,
1041AB_B.int    1:1,2:5,3:1,5:5,7:3,11:3,21:1,
11asAB_A.int    1:2,2:18,3:1,4:8,5:9,6:4,7:3,8:1,11:3,15:1,16:1,18:1,23:1,26:1,
11asAB_B.int    1:7,2:16,4:8,5:6,6:3,7:7,10:1,11:4,17:3,18:1,19:1,
1371AB_A.int    1:2,2:8,4:4,7:3,8:1,11:1,15:1,17:1,31:1,32:1,
1371AB_B.int    2:11,5:3,7:2,8:1,15:1,21:2,
13pkBC_B.int    2:32,3:1,5:3,6:4,7:4,10:1,11:2,12:1,16:1,25:1,
13pkBC_C.int    2:29,3:1,5:2,6:4,7:3,8:1,10:1,11:3,15:2,16:1,
1501AC_A.int    2:3,4:4,5:3,7:1,8:1,15:3,17:1,26:1,
1501AC_C.int    2:20,5:3,7:1,
1681AB_A.int    2:18,5:2,8:1,11:1,
1681AB_B.int    3:1,7:2,11:1,
1691CD_C.int    2:29,3:1,5:1,7:1,11:1,33:1,
1691CD_D.int    2:2,3:1,4:1,5:1,7:3,8:1,11:1,15:1,
1741AB_A.int    2:14,3:2,5:3,6:2,8:1,11:2,15:1,21:1,30:1,
1741AB_B.int    2:14,3:2,5:3,6:2,8:1,11:3,15:1,19:1,
```

Figure A2 An example of protein descriptor vectors. Each line has a protein name and its descriptor vector.

Vita

Farideh Halakou was born in Gonbad-e Kavous, Iran, on June 12, 1987. She received the BSc degree in Software Engineering from Mirdamad Higher Education Institute of Gorgan in 2008 and MSc degree in Information Technology Engineering from Kerman Graduate University of Technology in 2012. She received the PhD degree from Koc University in Computer Science and Engineering in 2019. From February 2014 to June 2018 she worked as teaching and research assistant at Koc University. Her research mainly focuses on embedding conformational changes of proteins into traditional protein-protein interaction networks.