Analysis of Human Protein-Protein Interaction Network and Cancer Proteins Using Structural Information

by

Gözde Kar

A Thesis Submitted to the
Graduate School of Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science
in
Computational Science and Engineering

Koc University

September 2008

Koc University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Gözde Kar

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:

_____

Assoc. Prof. Özlem Keskin

_____

Assoc. Prof. Attila Gürsoy

_____

Asst. Prof. Mehmet Sayar

Date: _____

ABSTRACT

Protein-protein interaction networks reveal that some proteins are highly connected to others (acting as hub proteins), whereas some others have a few interactions. The same or overlapping binding sites should be repeatedly used in hub proteins (single interface hub proteins) making them promiscuous. Alternatively, multi-interface hub proteins make use of several distinct binding sites to bind to different partners. Understanding the interactions with respect to their physical and chemical properties requires the atomistic details of the proteins, namely the three-dimensional structures. Then again, cancer-related proteins are more likely to act as hubs in interaction networks. In this thesis, we investigate "what features of cancer-related protein interfaces make them act as hubs" and "how it is possible for them to bind to many different proteins with varying affinity". We provide a detailed analysis of human protein-protein interaction network including cancer-related interactions. First we analyze the global behavior of cancer-related proteins, second we hold a structural perspective to elucidate how these proteins interact and figure out which interactions can occur simultaneously and which ones exclude each other. The results reveal that cancer-related proteins tend to interact with their partners through distinct interfaces, thus corresponding mostly to multi-interface hubs (56% of cancer-related proteins are multi-interface) and constituting the nodes with higher essentiality in the network (76% of them are essential). In addition, they have smaller, more planar, more charged and less hydrophobic binding sites compared to non-cancer ones which may indicate low affinity and high specificity of the cancer-related interactions. These findings might be important in obtaining new targets in cancer as well as finding the details of specific binding regions of putative drug candidates in cancer.

ÖZET

Protein-protein etkileşim ağları ortaya çıkarıyor ki bazı proteinler diğerleriyle oldukça iletişim halindeyken (merkez proteinler olarak görev almaktadırlar), bazılarının birkaç tane iletişimi vardır. Aynı veya üst üste gelen bağlanma yerleri merkez proteinlerde (tek ara yüzeyli merkez proteinler) ardı sıra kullanılmalıdır ki bu onları farksız kılar. Alternatif olarak, çoklu-ara yüzeyli merkez proteinler birkaç farklı bağlanma yüzeyini kullanarak farklı partnerlere bağlanırlar. Etkileşimleri fiziksel ve kimyasal özelliklerine göre anlamak proteinlerin atomik detaylarını; yani üç boyutlu yapıları gerektirir. Ve yine, kanserle ilgili proteinler etkileşim ağlarında merkez proteinler olarak görev yapmaya daha eğilimlidirler. Bu tezde, "kanserle ilgili protein ara yüzeylerinin hangi özellikleri onların merkez proteini olarak görev yapmasını sağlıyor" ve "bunların birçok farklı proteine değişen eğilimlerle bağlanması nasıl mümkün oluyor" araştırıyoruz. Kanserle ilgili etkileşimleri içeren insan protein-protein etkileşim ağının detaylı bir analizini sağlıyoruz. İlk olarak, kanserle ilgili proteinlerin genel davranışlarını inceliyor, ikinci olarak yapısal bir perspektiften bakarak bu proteinler nasıl etkileşime giriyor ve hangi etkileşimler aynı anda oluyor ve hangileri aynı anda olmuyor, aydınlatıyoruz. Sonuçlar gösteriyor ki, kanserle ilgili proteinler partnerleriyle farklı ara yüzeylerden etkileşime giriyor, böylece daha çok çoklu-ara yüzeyli merkez proteinlere denk geliyor (kanser proteinlerin %56 sı çoklu-ara yüzeyli) ve ağdaki gerekli düğümleri oluşturuyor. (bunların %76 sı gereklidir). Buna ek olarak, kanserli olmayanlarla karşılaştırıldığında, kanserle ilgili etkileşimlerin daha küçük, daha düzlemsel, daha yüklü, az hidrofobik bağlanma yüzeyleri vardır. Bu bulgular kanserde varsayılan ilaç adaylarının spesifik bağlanma bölgelerinin detaylarını bulmak açısından önemli olduğu kadar kanserde yeni hedefler bulmak açısından da önemlidir.

ACKNOWLEDGEMENTS

I would like to thank Assoc. Prof. Özlem Keskin (my advisor) and Assoc. Prof. Attila Gürsoy for their helpful discussions and their support during my M.S. study. Also, I would like to thank my thesis committee members for their critical reading.

I would like to thank my friends at Koc University; my former and recent office mates (Ekin Tüzün, Emre Güney, M. Cengiz Ulubaş, Nurcan Tunçbağ), my friends in ENG 110 (Sefer Baday, Özge Engin, Bahar Öndül, E. Besray Ünal, Osman N. Yoğurtçu) for their support and good times during these two years.

And finally, I would like to thank my family for their encouragement and continuous support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Protein-protein interactions play important roles in most cellular and molecular processes. Since proteins interact with other proteins, and with DNA, RNA and small molecules to become functional, observing these interactions all together in a network representation is crucial. Thereby, analysis of protein-protein interaction networks provides valuable information about protein function and complex cellular processes [1].

With the increase of protein interaction measurements, protein-protein interaction networks serve as important tools to understand the molecular mechanisms of disease. The studies, which have focused on analyzing the network properties of disease genes, have concluded that network topology of disease genes are quite different from those not involved in disease [2-4]. Thus, based on network topology, or functional relatedness of genes and known interactions, potential disease genes are discovered [5, 6].

Recently, the studies, which analyze structural interaction networks, have gained importance. The underlying reason is that protein interaction networks do not reflect the chemical and structural properties of interactions, and a better understanding of how molecules interact can be obtained only from three-dimensional (3D) structures [7]. However, determining 3D structures has been a hard work; thus, studies have presented methods to model interactions on complexes of known 3D structures [7-9]. In addition, structural modeling has been combined by network analysis, which has revealed relationships between network topology, genomic features and structure [10].

In this thesis study, differently from previous structural studies, we integrate protein-protein interfaces into structural network, focus on cancer-related proteins and investigate interface properties of cancer/noncancer protein interactions to shed light on interaction details. We provide a detailed analysis and comparison of six interaction networks: 1) human protein-protein interaction network, (PIN), 2) human cancer-related protein-protein interaction network, cPIN, a sub-network of the first. Then, we characterize the interactions in these networks by combining three-dimensional protein structures. Thus, we have: 3) the network constructed by selecting genes for which three-dimensional protein data is available, SPIN, a sub-network of the first, 4) human cancer-related structural protein-protein interaction network, cSPIN, a sub-network of SPIN. We map the known structural data into these networks whenever complex structure is available. For the rest, we predict the complex structures of the interactions through structural templates and hot spots using PRISM [11, 12]. So the last two resulting networks are "structural interface" networks: 5) human structural protein interface network (iSPIN) and 6) structural cancer-related protein interface network (ciSPIN). These six networks are analyzed and compared to highlight the advantages of using structures. The results reveal that cancer-related proteins tend to interact with their partners through distinct interfaces, thus corresponding mostly to multi-interface hubs and constituting the nodes with higher essentiality in the network. In addition, they have smaller, more planar and more hydrophilic binding sites compared to non-cancer ones which may indicate low affinity and high specificity of the cancer-related interactions.

The outline of the thesis is arranged as follows; Chapter 2 includes the summary of the related work in the literature. Studies about the analysis of protein-protein interaction networks, applications of protein networks to disease and structural interaction networks are presented.

In Chapter 3, the methodology of this thesis study is illustrated in detail. The approaches to map interactions to known three-dimensional structures, to define network topology and genomic properties, to predict interaction interfaces and verify

them, to classify network hub proteins with respect to number of interfaces and analysis of interface properties are explained.

Chapter 4 is the results part and includes the analysis of human protein-protein interaction network and its corresponding structural interaction network.

As presented in Chapter 5, this thesis study is concluded with discussion of the results and future directions.

Chapter 2

LITERATURE REVIEW

In this chapter, the summary of the related work in the literature is presented. Firstly, the applications of protein networks to disease are described. Then, the studies analyzing protein-protein interaction networks in terms of topological properties, protein function, expression dynamics and other genomic features are explained. This chapter continues with works about structural networks which combine structural modeling with network analysis.

2.1. Protein-protein interaction networks

Proteins do not function in isolation, but rather they interact with other proteins and with DNA, RNA and small molecules to accomplish their function. Thus, observing these interactions all together in a network representation seems to be crucial. Protein-protein interaction networks provide a valuable framework to annotate protein function and understand complex cellular processes [1]. Moreover, they serve as an important tool to understand the molecular mechanisms of disease.

2.1.1. Protein networks in disease

This section reviews the protein network analysis of human disease: the works, which aim to identify new candidate disease genes and investigate their network properties, are summarized. With the increase of human interactome data, research groups have focused on studying the networks underlying human disease [13].

Wachi et al. (2005), for the first time, have studied the predicted human interaction map for the analysis of cancer. They have analyzed the topological features (degree distribution and centrality) of differentially expressed genes in lung squamous cancer tissues. They have found that squamous cell lung cancer genes are highly connected and central; hence display the same topological properties of essential genes. These findings are explained by the idea that squamous cell lung cancer genes are essential for survival and proliferation of the cancerous tissues [4].

In a comprehensive study including all known human cancer genes, Jonsson and Bates (2006) have illustrated that network topology of cancer proteins is different from non-cancer proteins in the human interactome. Cancer proteins, on average, have twice as many interaction partners as non-cancer proteins and belong to larger communities [3].

Goh et al. (2007) have constructed a diseasome bipartite network which shows all disorders and the genes associated with the disorders. To infer human genetic disorders, they have used OMIM (Online Mendelian Inheritance in Man) [14] database. They have observed that disease genes tend to be coexpressed in specific tissues and interact mostly with other disease genes. Also, they have investigated the essentiality of genes in the human disease network; essential genes tend to encode highly connected proteins. However, in contrast to the previous studies, most of the disease genes are found to be nonessential and have no tendency to encode highly connected proteins. But, they have excluded the disease genes with somatic mutations from this finding and argued that somatic mutations leading to severe disease phenotypes should affect the functional center. When they have studied the properties of somatic cancer genes (Cancer Genome Census; www.sanger.ac.uk/genetics/CGP/Census they have observed the functional and topological centrality of the somatic cancer genes; they are more likely to encode hubs and be more essential [2].

In a recent study, Rambaldi et al. (2008) have determined genomic and network properties of ~600 human genes related with different cancer types. They have found that cancer related genes duplicate less than other human genes, independently of their molecular function and type of cancer they belong to. They have stated that genes mutated in cancer encode proteins which position in the centre of highly interconnected clusters of the human protein interaction network. Moreover, they have suggested that cancer genes are more fragile than other human genes toward perturbations (gene dosage modifications). Dosage effects on the cell cycle and sensitivity of cancer genes have been also experimentally confirmed [15].

In addition to the network topology analysis of disease, research groups have also aimed to predict new disease-associated genes. Human genetic diseases are mostly caused by multiple genes, which are likely to be functionally related. One approach for finding novel disease genes is to investigate the functional relatedness of genes [16]. Oti et al. (2006) have showed that the use of protein-protein interaction data have improved the prediction of candidate disease genes. A gene has been predicted as disease-related if it has a protein interaction with a known disease gene and is found within chromosomal locations associated with the given disease [5]. In a similar study, Franke et al. (2006) have constructed a functional human gene network based on known interactions and functional relationships. They have related this network to candidate genes from different disease loci. Around each disease gene, susceptibility loci containing 100 genes are constructed. Their algorithm ranks candidate genes in the susceptibility loci and is able to detect disease-causing genes with a 2.8-fold increase over random selection [17].

It has been observed that disease genes preferentially interact with other disease-causing genes. Thus, heritable disease genes might share some topological features in the protein network [6]. Based on this approach, Xu et al. (2006) have performed a study in which several topological features of protein network are compared between known disease

genes and genes known not to be involved in disease. They have trained their classifier on a set of known disease genes and non-disease genes and applied the classifier to over 5000 human genes. They have found novel 178 putative disease genes and some of them are validated by biological experiments [6].

2.1.2. Linking network topology to protein function and genomic features

In network studies, finding relations between network topology and functional, genomic features has been an important goal. Such a study has been accomplished by Han et al (2004) in which they have linked network topology to expression dynamics. They have classified highly connected proteins ("hubs") in the yeast interactome network into date and party hubs based on their partners' expression profiles. Calculating the average Pearson correlation coefficients between hubs and their partners for mRNA expression, they have uncovered two kinds of hubs; the ones with relatively high average Pearson correlation coefficients are defined as "party" and the other with relatively low average Pearson correlation coefficients are defined as "date". Party hubs presumably interact with most of their partners simultaneously, whereas date hubs interact with their partners at different times [18].

In another study, de Lichtenberg et al (2005) have mapped microarray expression data onto yeast protein-protein interaction network. Thus, they have analyzed the dynamics of protein complexes during the yeast cell cycle. Their time-dependent interaction network infers functional linkages between proteins and they have revealed previously unknown components and modules [19].

Network studies have usually focused on protein hubs. In the study of Yu et al (2007), another topological feature has been given importance; betweenness centrality, which measures the total number of nonredundant shortest paths going through a certain protein [20]. Proteins with a high betweenness centrality have been defined as bottlenecks. Yu et al

(2007) have analyzed the expression dynamics and gene essentiality of hubs and bottlenecks in protein-protein interaction network and regulatory network in yeast. In terms of expression correlation, they have stated that hub-nonbottlenecks (hubs with low betweenness centrality) are relatively well coexpressed with their neighbors, whereas most of the hub-bottlenecks (hubs with high betweenness centrality) are not very well coexpressed. Thus, they have concluded that hub-bottlenecks tend to be date-hubs and hub-nonbottlenecks tend to be party-hubs. When they have considered the gene essentiality, they have found that a protein's degree (number of interaction partners) is a stronger determinant of essentiality in the protein interaction network, whereas in the regulatory network, betweenness is a stronger determinant of essentiality. These findings have indicated the relationships between dynamics and topological features in the interaction network [20].

2.2. Structural networks

Protein-protein interaction networks provide a basis to understand biological processes and protein function. However, they provide a rather abstract network representation of proteins since they do not reflect the chemical and structural aspects of interactions. At this point, introducing structural networks is necessary to reveal biological reality of the interactions. Structural network studies aim to find answers to such questions; which interactions can occur simultaneously and which ones exclude each other, do the interacting proteins form transient or stable complexes, what are the affinities and kinetic constants [21].

The pioneering work of Aloy & Russell (2002) has presented a method to model interactions on complexes of known 3D structures and illustrated how 3D structures can be used to infer molecular details of the interactions [7]. Their following studies have also focused on finding appropriate 3D structures to model binary interactions [8, 9]. They have

modeled complexes on a large scale in yeast using homology modeling, electron microscopy and affinity purification [8].

In an outstanding study, Kim et al (2006) have combined structural modeling with network analysis. They have mapped the interactions to known structures of interfaces and distinguished the interfaces of each interaction. For two or more proteins interacting with a common protein, if they use the same binding interface, the interactions have been categorized as mutually exclusive. If the proteins use different interfaces, interactions are simultaneously possible. They further have classified the network hubs as single-interface and multi-interface hubs. The former has at most two distinct binding interfaces and are enriched in signaling proteins, whereas the latter has more than two binding interfaces and tend to be members of large and stable complexes. Regarding the expression dynamics, they have found that multi-interface hubs are relatively well coexpressed with their neighbors, whereas single-interface hubs are not very well coexpressed. Thereby, they have stated that single-interface hubs seem to correspond to date-hubs and multi-interface hubs correspond to party-hubs [10]. In a recent study, they have used the same structural interaction network to explore the role of intrinsic disorder in this network. Disordered regions are the flexible and unordered segments of proteins [22]. They have found that single-interface hubs tend to be more disordered than the proteome average. On the contrary, the disorder level of multi-interface hubs cannot be distinguished from other proteins. The higher level of disorder of single-interface hubs is rationalized by their tendency to bind their partners in a cascade [22].

Regarding the pathways, including known or modeled structures into pathways provides a much better understanding of them. The order of interactions occurring through a pathway can be clarified by showing which interactions cannot occur simultaneously owing to a same binding interface. In addition, structured pathways can highlight the molecular mechanisms for a particular disease and provide a rational basis for designing

drug targets. In a recent study, Schroeder and his group (2007) have constructed a cell map related to pancreatic cancer by linking pathway approach, known interactions and structure-based interaction predictions. They have predicted 40 novel interactions that are specific for this disease [23].

Chapter 3

METHODOLOGY

This chapter contains the methodology followed in this study to generate a structured network and analyze the protein-protein interfaces, linking structure to network topology. First, the method for mapping the interactions onto known structures is presented. Then, interface prediction in the structured network is explained. Finally, the approach to define the topological properties and gene essentiality and analysis of interface properties for our networks are explained.

3.1 Human protein-protein interaction and cancer-associated protein interaction datasets

We have studied the human interactome constructed by Jonsson & Bates (2006) [3] and referred to this network as 'PIN'. They have used an orthology-based method in which BLAST [24] searches were run for the human genome against all proteins in the DIP [25] and MIPS Mammalian Protein-Protein Interaction databases [26]. They have analyzed their putative interactions giving confidence scores based on the level of homology to proteins found experimentally to interact and the amount of experimental data available. After the ROC curve analysis, with a sensitivity of 85 % and specificity of 82 %, the human interactome consisted of 108113 binary gene-gene interactions and 13584 genes. From these interactions, the redundant ones, i.e. the interactions for which the RefSeq ID corresponding to the same genes, are omitted. Thereby, the network (PIN) consists of

85083 interactions. The list of cancer genes are taken from the comprehensive census of human cancer genes provided by Futreal et al (2004) [27]. 10724 interactions are cancer-related in this interactome. In addition, we collected a set of known cancer genes from the Memorial Sloan Kettering computational biology website CancerGenes (http://www.mskcc.org/cancergenes) by the query of "tumor suppressor", "oncogene" and "stability" genes. We combined that list with the known cancer genes of Futreal et al. [27]. Thus, cancer related interactions number increases to 27413.

3.2 Mapping interactions to known 3D structures

We used Swiss-Prot Knowledgebase [28] to map the binary interactions to known structures. The human genes, for which 3D structures are known, are compiled from Swiss-Prot Knowledgebase. For each gene-gene interaction in the human interactome, a known complex structure is searched. If a known structure is not available for the interaction, we search for the structures of each gene and map each gene to the corresponding structure (as a single chain). If any of the genes in the binary interaction does not have a structural representation, then this interaction is omitted. The method can be further clarified by an example. In the human interactome, one of the binary interactions is TP53-MDM2 interaction. The interaction is represented by a known complex structure in PDB [29] as 1ycr. However, for the TP53-MDM4 interaction, there occurs no known complex structure. In this case, TP53 was represented by one of its corresponding structures; basically by the one with the highest resolution for which the PDB ID is 1aie_chain A. Similarly, for MDM4, the structure is 2cr8_chainA. In total, 206 interactions were mapped to known complexes. The summary of the mapping procedure is illustrated in **Figure 3.1.**

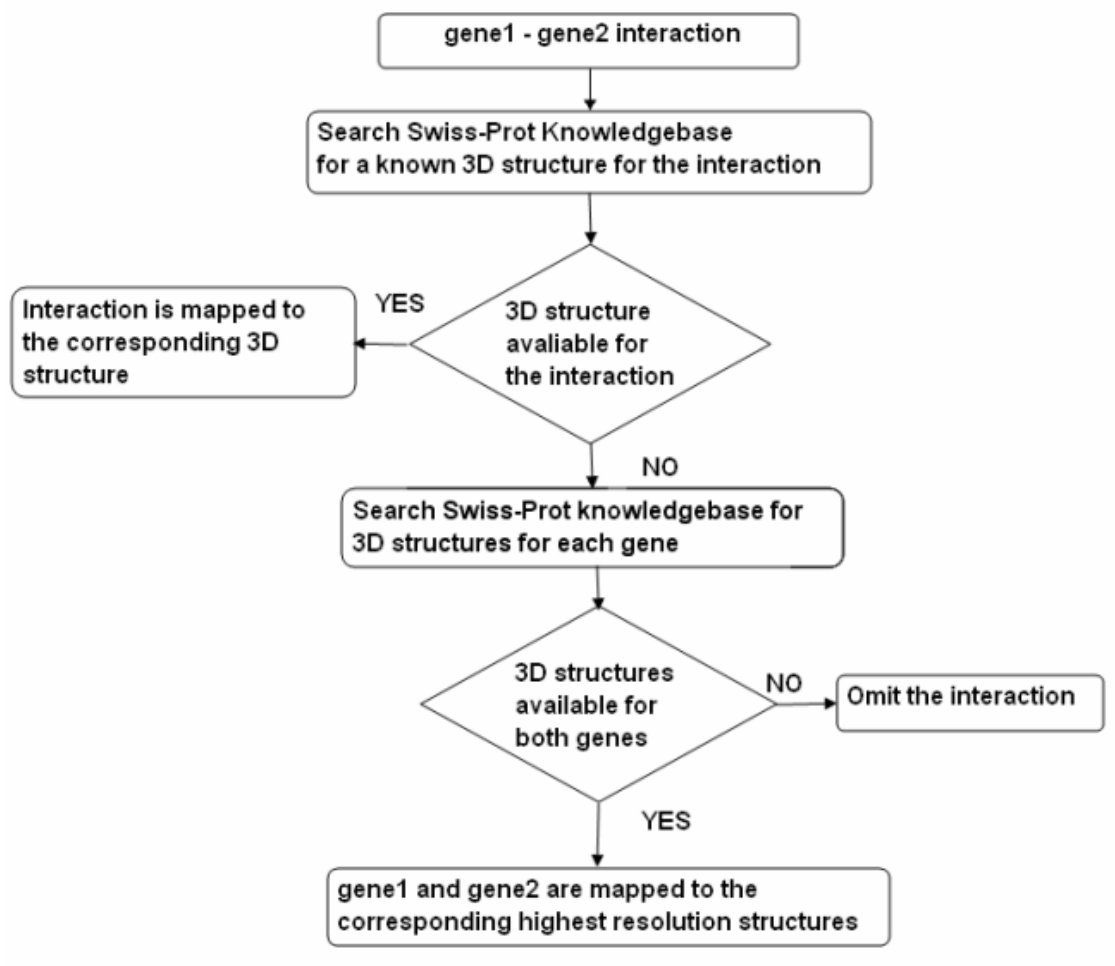

Figure 3.1. Flowchart shows the method of mapping interactions to 3D structures. The method is applied for all the interactions in the human interactome (PIN).

The mapped protein-protein interaction network called 'structural protein interaction network' (SPIN) which consists of 1702 nodes (proteins) and 5312 edges (interactions). From 5312 interactions, 206 interactions are mapped to known 3D structures. Thereby, the interfaces of these 206 interactions are known. On the other hand, the interfaces of the remaining 5106 interactions are left for further prediction.

When the list of cancer-related proteins are searched through 1702 proteins, 466 of them are found to be encoded by cancer-related genes (cancer gene information from Futreal et al. [27] and the Memorial Sloan Kettering computational biology website CancerGenes http://www.mskcc.org/cancergenes), the rest (1236 of them) is taken as encoded by noncancer genes. As a result, we define the 'cancer structural subnetwork' ('cSPIN'), as the one consisting of cancer-cancer and cancer-noncancer gene interactions. Our cSPIN contains 1303 proteins and 3221 interactions. The total number of proteins and interactions for each network is summarized in section 4, **Table 4.1**.

3.3 Definition of Hubs and Bottlenecks

Degree represents the number of interaction partners of a protein. Betweenness is a measure of the total number of shortest paths going through a certain node or edge in the network [30]. We defined hubs as the proteins that are in the top 20% of the degree distribution in PIN and SPIN. That corresponds to proteins with ≥9 interactions. Accordingly, we defined bottlenecks as the top 20% proteins with the highest betweenness values. To calculate betweenness within the network, we used NetworkX (NX) (https://networkx.lanl.gov/wiki) which is a Python package. Hubs were classified as hub-bottlenecks and hub-nonbottlenecks according to high betweenness or low betweenness, respectively.

3.4 Determination of essential human genes

Goh et al (2007) [2] predicted the essentiality of a human gene using phenotype information of the corresponding mouse orthologs. A human gene was defined as "essential" if a knock-out of its mouse ortholog results in lethality. Here; embryonic/prenatal lethality and postnatal lethality are considered as lethal phenotypes, and the rest of phenotypes as non-lethal ones. We obtain the human-mouse orthology and

mouse phenotype data from Mouse Genome Informatics (http://www.informatics.jax.org) on May 10, 2008. Of 1702 proteins in our SPIN, 1536 have mouse orthologs and phenotype information. According to our classification, we find 497 genes to be essential and the rest is non-essential.

3.5 Extracting the Protein-Protein Interfaces in the SPIN using PRISM

PRISM (protein interactions by structural matching) [11, 12] is a web server to predict protein-protein interactions and protein interfaces. The prediction algorithm uses structural and evolutionary similarities to find possible binary interactions between proteins (targets) through similar known interfaces (templates). Here, target proteins are the proteins in our SPIN dataset for which we want to predict the interaction interfaces. As template interfaces, we used the representative interfaces which are generated from the nonredundant data set of protein-protein interfaces [31] available at http://prism.ccbb.ku.edu.tr/interface , for which the interactions are biological according to NOXclass [32] outputs. There are 1478 template interfaces.

PRISM prediction algorithm starts with extracting the surfaces of target proteins by invoking NACCESS [33]. Template interfaces are split into their complementary partner chains and these partners are structurally aligned with the surfaces of the target proteins. Similarity between the target surface and one partner of the template interface is measured using a scoring function based on two factors. First one is the structural similarity, in which RMSD and residue match ratio between target protein and the template interface is scored. The other factor considers evolutionary similarity in which hotspot match ratio is scored. (PRISM obtains the information on hotspots from Hotsprint [34], a web server for predicting hotspots at protein interfaces.) Then, combining these scores, PRISM predicts the most possible interactions occurring between the target proteins.

3.6 Elimination of crystal packing interfaces and interactions

After we obtain the interfaces of the proteins in our network using PRISM, non-biological interfaces, if any, should be eliminated. The interfaces having a biological score greater than 60% according to the NOXclass [32] outputs are accepted as biologically relevant. Thus, 357 interaction interfaces are predicted and most of them (80%) have biological score greater than 80%. Also, including the known interfaces coming from 3D structures, the resulting network which includes interface information is called 'iSPIN'. It consists of 534 proteins and 563 interactions. The subnetwork of cancer-related interactions (ciSPIN) includes 381 proteins and 375 interactions. The protein and interaction numbers are given in **Table 3.1.**

3.7 Hub classification: Single-interface and multi-interface hub

Kim et al. (2006) [10] has classified protein hubs as singlish-interface and multi-interface hubs. The former has at most two distinct binding interfaces, whereas the latter has more than two binding interfaces. In this study, we also classified the hubs in iSPIN according to the number of distinct binding interfaces; we defined single-interface hubs as the protein hubs with only one distinct binding interface and multi-interface hubs as the ones with more than one distinct binding interface. To distinguish overlapping interfaces from non-overlapping interfaces, we looked at the shared residue percentage of the interfaces of hub proteins. We defined shared residue percentage as the ratio of number of shared residues to the number of total interface residues. If the interface residues are shared at a percentage greater than 20%, then the corresponding interface is an overlapping one and interactions occurring through this interface are mutually exclusive, i.e. the interactions cannot occur at the same time. On the other hand, if the interface is not shared at all, meaning that shared residue percentage is less than 20%, then this is a non-overlapping

interface and the interaction through this interface is simultaneously possible, independent of each other.

3.8 Interface property analysis

For interface analysis, we used PROTORP [35] which invokes NACCESS [33], SURFNET [36] and PRINCIP (SURFNET) [36] for interface accessible surface area and gap volume and planarity calculation, respectively. PROTORP calculates the amino acid composition of residues defined in the interface as a percentage value of those classified as polar, non-polar and charged as described previously by Jones and Thornton [37]. The amino acid compositions are weighted and then normalized by the complex ASA values which are calculated using NACCESS.

3.9 Statistical tests

To determine the significance of the difference between interface properties of cancer and noncancer genes, statistical tests (nonparametric, one-tailed) are performed using GraphPad InStat software.

3.10 Network topology analysis

All the parameters describing the network topology are calculated using NetworkAnalyzer, which is a Java plugin for Cytoscape [38].

Chapter 4

PROTEIN-PROTEIN INTERACTION NETWORK TO STRUCTURAL INTERACTION NETWORK

4.1 Structural Protein Interface Network (iSPIN)

We illustrate how to obtain a structure-integrated network from PIN. The seed network is the human protein-protein interaction network (PIN) where the nodes are proteins and the edges are interactions. We search which proteins in this networks have structural information in PDB and constructed a subnetwork with the extracted structures called as SPIN (see Methods for the details). To further integrate protein interfaces into SPIN, we map the known structural data of complexes into SPIN whenever complex structure is available. If a known structure is not available for an interaction, we predict the complex structures of the two interacting proteins by using structural templates and hot spots through PRISM [11, 12]. The resulting network which contains known complexes (from Protein Data Bank (PDB) [29]) and predicted complexes (from PRISM) contains interface knowledge, and is called iSPIN. The subsets of PIN, SPIN and iSPIN, which contains cancer-related interactions, are called cPIN, cSPIN and ciSPIN, respectively (See Methods section for further information). **Table 4.1** lists the protein and interaction numbers in each of the interaction networks. Three networks; from PIN to iSPIN are represented in **Figure 4.1**.We should note that there is a dramatic decrease in the number of proteins when going from PIN to SPIN. As seen in **Figure 4.1**, while PIN contains the information of gene interactions, SPIN only contains those with PDB IDs. And finally iSPIN contains the

information at the atomic level; protein interfaces. Although we provide a topological analysis of the networks, the main concern of this study is to present interface analysis of cancer-related proteins and in addition predict which interactions can and cannot occur simultaneously and ultimately, to emphasize the importance of using structures in network studies.

Table 4.1. The number of proteins and interactions in each network.

Network name	Protein	Interaction	Known complex in PDB
PIN	13584	85083	NA
cPIN	8990	27413	NA
SPIN	1702	5312	206
cSPIN	1303	3221	149
iSPIN	534	563	206
ciSPIN	381	375	149

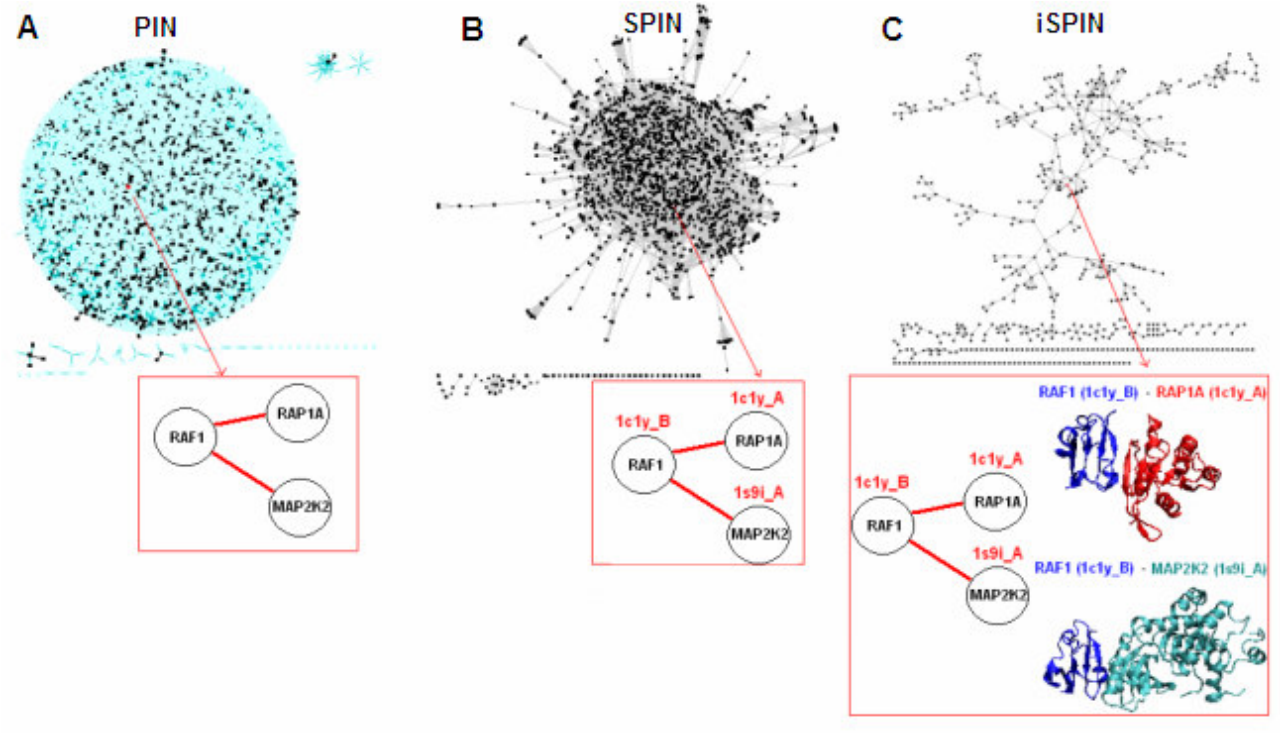

Figure 4.1 Representation of PIN,SPIN and iSPIN. In A) proteins in PIN are represented; the ones colored black have PDB IDs and the ones colored blue do not have PDB IDs. In B) The proteins with PDB ID and interactions among them constitutes SPIN. In C) The proteins with PDB ID and protein interface information and their interactions constitutes iSPIN. The zoomed representations give idea about what type of information each network contains; PIN is an abstract representation of interactions, SPIN is a subset of PIN with information of PDB IDs, and iSPIN contains the most detailed information including protein interfaces into network. All the networks are visualized using Cytoscape [38].

4.2 Analysis of the interface properties in iSPIN

We present the interface properties of interactions in iSPIN (both predicted and known PDB interfaces) such as the interface accessible surface area (ASA), planarity, gap volume index and residue composition at the interfaces.

4.2.1 Cancer proteins have smaller, more planar, less tightly packed and less hydrophobic binding sites compared to non-cancer ones

Physical properties of interfaces are computed for the interactions in iSPIN. We classify the interactions into two groups; 'cancer-related interactions' are those in which at least one partner is a cancer-related protein in a binary interaction and 'noncancer interactions' are the ones in which none of the proteins are known to be involved in cancer. In this way, there are 363 and 186 cancer-related and non-cancer interactions, respectively. For interface ASA calculation, PROTORP [35] invokes NACCESS [33] and difference in ASA between monomers and the complex is given as the interface ASA. Interface residues are defined as those residues that have an ASA that decreases by >1.0 Ångström^2 on complexation. Regarding the interface ASA (ΔASA), cancer proteins on average are observed to have a smaller interface ASA (1009.1 Å ^{2}) than that of noncancer proteins (1242.9 Å^2) (standard deviations and p-values are summarized in **Table 4.2**). These results indicate that the complex interfaces which are formed through the interactions of cancer proteins are less buried, or likewise, the monomeric surfaces of cancer proteins are less exposed. When the ASA of the complex structures are considered, it is found that ASA of cancer proteins (2210.9 Å ^{2}) are statistically significantly smaller than that of noncancer proteins (2628.1 Å ^{2}). It is known that transient complexes have smaller interface areas [39]. Our results show that cancer proteins use a smaller surface area while interacting and we know that they have many interaction partners [3], thus it may be hypothesized that they are more likely to be involved in transient interactions.

We also investigate the complementarity of the interfaces. Gap volume gives a measure of complementarity and closeness of packing of the interface between the two interacting proteins, by measuring the volume of empty space between them. Gap volume index is the ratio of gap volume to the interface area, therefore it estimates the volume enclosed between any two molecules, delimiting the boundary by defining a maximum allowed distance from both interfaces [40]. PROTORP uses SURFNET [36] for gap volume calculation. For the cancer related interactions, the average gap volume (5076.8 $Å^3$) is found to be significantly smaller than the average gap volume of noncancer interactions (5574.5 $Å^3$) (p-value = 0.0194 at $\alpha = 0.05$). This is an artificial result or outcome of the smaller interfaces of the cancer proteins since volume is proportional to the surface area. On the other hand, the average gap volume indeces for these two categories are 2.76A˚ and 2.54 A˚, respectively. The difference is significant; meaning cancer related interactions are less optimized in terms of complementarity. This indicates that, in fact, the complementarity and packing of two types (cancer/noncancer) are distinguishable from each other.

To analyze the shapes of the interfaces, the planarity indices are used. PROTORP uses PRINCIP program which is a part of SURFNET package [36]. PRINCIP models a best-fit plane through the atoms defined in the interaction site using principal component analysis. For cancer-related interactions, the average planarity index (2.84) is smaller than that of non-cancer ones (3.06) with p-value 0.02 indicating that cancer-related interfaces are more planar. It is known that there is a high correlation between the planarity of the interfaces and the ASAs of the interfaces [41]. As the ASAs of the interfaces increases, the planarity index also increases, the interfaces become less planar, deviating from their principal axes. It is also known that transient complexes usually have more planar interfaces [39]. Here, consistent with previous findings, we observe that cancer proteins use more planar binding sites in their complexes. The results are summarized in **Table 4.2**.

Previously, smaller interfaces have been shown to display a reduced hydrophobic effect [42]. We investigate if this finding agrees with our results. Residue composition in the interface (polar, non-polar or charged) is analyzed using PROTORP. The residue compositions are normalized by the ASA in the complex structures (see Methods). The results reveal that cancer-related interactions show a reduction in hydrophobicity and an increase in charged interactions; thus having more hydrophilic interfaces compared to non-cancer interactions. Although, in general, it is agreed that protein-protein interfaces are highly hydrophobic and hydrophobicity is a dominant force in protein-protein interactions [43], there are also studies indicating the importance of hydrophilic interface regions. In the study of Tormo et al. (1999), the interactions of NK (natural killer) receptors (which regulates NK cell function) have been studied and the interface of C-type-lectin-like receptor family (Ly49 A) has been detected to be highly hydrophilic and dominated by charged interactions [44]. For our iSPIN interfaces, charged interactions appear to play important roles as well, implying the significance of electrostatics in binding. A recent study indicated that favorable electrostatic interactions were not prerequisite for stable complex formation between proteins whereas hydrophobic effects were found to be favorable in native complexes [45]. Here, we also observe that cancer related proteins, which are intrinsically more disordered and transient [46], have less hydrophobic interactions as opposed to other proteins.

Table 4.2. Average interface properties of cancer and non-cancer interactions.

Interface property	**Cancer-related interactions**	**Non-cancer interactions**	**p-value (at α=0.05)**
ΔASA ($A^{\circ 2}$)	1009.1(±611)	1242.9(±942)	3.1e-005
ASA of the complex ($A^{\circ 2}$)	2210.9 (±1475)	2628.1(±1947)	3.2e-004
Planarity (A°)	2.84(±1.28)	3.06(±1.23)	0.02

Gap Volume Index	2.76(±1.48)	2.54(±1.27)	0.035
% Polar residues in interface	29.7 (±14.8)	30.7 (±13.5)	Not significant (0.07)
% Non-polar residues in interface	27.1 (±13.6)	28.8 (±12.9)	0.0035
% Charged residues in interface	43.2 (±16.6)	40.5 (±15.4)	0.003

4.2.2 Hub proteins have smaller, more planar, less tightly packed binding sites compared to non-hub ones

We also classified the interactions as "hub-involved" and "non-hub-involved" interactions. For the hub-involved interactions, at least one protein of the binary interaction is a hub protein, whereas for the latter, none of the proteins correspond to a hub. There are 455 and 94 hub-involved and non-hub-involved interactions, respectively. As hub proteins, we considered the hubs of SPIN. We find that, on average, hub proteins tend to form smaller, more planar interfaces with their partners. In contrast to previous studies [47, 48], we find no significant difference on the residue composition of the interfaces (including charged residue content) of hub proteins. In terms of complementarity of the interfaces, hub proteins form looser complexes (gap volume index of 2.72 versus 2.49). The results are summarized in **Table 4.3**.

Table 4.3 Average interface properties of hub and nonhub interactions

Interface property	**Hub-involved interactions**	**Nonhub-involved interactions**	**p-value (at α=0.05)**
	Single-interface hubs		
ΔASA ($A^{\circ 2}$)	1011.0 (±434)	1459.9 (±1484)	1.8e-004
	1022.1(374)		0.0050
ASA of the complex ($A^{\circ 2}$)	2230.0 (±1326)	2943.9 (±2691)	0.0015
	2228.1(1178)		0.0011
Planarity (A°)	2.82 (±1.13)	3.34 (±1.72)	0.0043
	2.97(1.13)		0.0370
Gap Volume Index	2.72 (±1.40)	2.49 (±1.48)	0.0350
	2.53(1.06)		Not significant (0.17)
% Polar residues in	30.5 (±14.5)	29.9 (±13.3)	Not significant (0.40)
interface	32.5 (±14.6)		0.016
% Non-polar residues in	28.0 (±13.3)	28.1 (±13.1)	Not significant (0.38)
interface	28.6(±12.2)		Not significant (0.29)
% Charged residues in	41.5 (±16.4)	42.0 (±15.4)	Not significant (0.38)
interface	38.8(±0.16)		0.011

4.2.3 Multi-interface and single-interface hubs have different interface characteristics

Some hubs are single-interface (communicating with their partners by using the same interface) whereas others are multi-interface. We mapped hub proteins of SPIN onto iSPIN. The proteins which correspond to hubs of SPIN and have interaction number≥3 are considered as multi-interface or single-interface hubs. In this way, there are 79 hub proteins from which 42 of them are multi-interface and 37 are single-interface hubs. Interestingly

when we compared the interface characteristics of these two types of hubs, we observe that they have different characteristics. Usually multi-interface hubs are similar to non-hub interfaces. On the other hand, single-interface hubs are more polar and less charged than the multi-interface hubs and non-hub proteins (See the second lines in each row of **Table 4.3**).

4.3 Network Topology of the networks and relationship with essentiality

To address the topological properties of each network, we calculated the degree distribution of proteins which is a measure of the number of interaction partners of the proteins. In **Figure 4.2**, the topological properties are visualized for SPIN. It is observed that the degree distribution of the proteins decreases slowly following a power-law ($P(k) \sim k^{\gamma}$,where k is the number of partner proteins) for each network. (For example, in **Figure 4.2**A, $R^2 = 0.914$ for power law fit for SPIN) This implies the fact that the networks have scale-free properties [49]. On average proteins in SPIN have 6.24 interaction partners. In **Figure 4.2**B, average clustering coefficient which is a measure of proteins to form clusters in the network [49] is shown. Average clustering coefficient decreases as the number of interactions of the proteins increases since sparsely connected proteins are neighbors of highly connected proteins (hub proteins). For the hub proteins, the number of neighbors increase, but the number of connected pairs does not increase that much which causes average clustering coefficient to decrease. This behavior indicates the hierarchical organization in the protein interaction network [49]. In **Figure 4.2**C, the topological coefficient which is a relative measure for the extent to which a protein shares neighbors with other proteins [50] is displayed. The decreasing behavior of the topological coefficient as the number of interactions of a protein increases confirms the modular network organization, hub proteins do not share more neighbors with other proteins than the sparsely connected proteins. **Figure 4.2**D shows the shortest length distribution and

indicates that proteins are closely linked. The topological properties of other networks (PIN, cPIN, cSPIN, iSPIN, ciSPIN) show similar trends to those of SPIN explained above. When the cancer subnetworks are compared with the original ones (cPIN with PIN, cSPIN with SPIN and ciSPIN with iSPIN), the average clustering coefficient values are lower; i.e. the proteins have a lower tendency to form clusters. This is reasonable since cancer proteins are the key nodes that link different pathways and they spread throughout the network to function in these pathways. For example, the Cancer Cell Map (http://cancer.cellmap.org/cellmap/), which is a collection of human-focused cellular pathways implicated in cancer, contains ten pathways each having around 100-400 interactions and usually the cancer genes function in more than one pathway. The network parameters calculated for each network are displayed in **Table 4.4**.

Table 4.4 Network parameters calculated for each network

Parameter	**Network type**					
	PIN	**cPIN**	**SPIN**	**cSPIN**	**iSPIN**	**ciSPIN**
Number of nodes	13584	8990	1702	1303	534	381
Number of edges	85083	27413	5312	3221	563	375
Clustering coefficient	0.109	0.080	0.143	0.113	0.089	0.051
Characteristic path length	4.086	4.589	4.661	5.064	9.533	8.221
Network diameter	11	11	11	11	23	20
Network density	0.001	0.001	0.004	0.004	0.004	0.005
Avg. number of neighbors	11.27	5.45	6.24	4.94	2.11	1.97

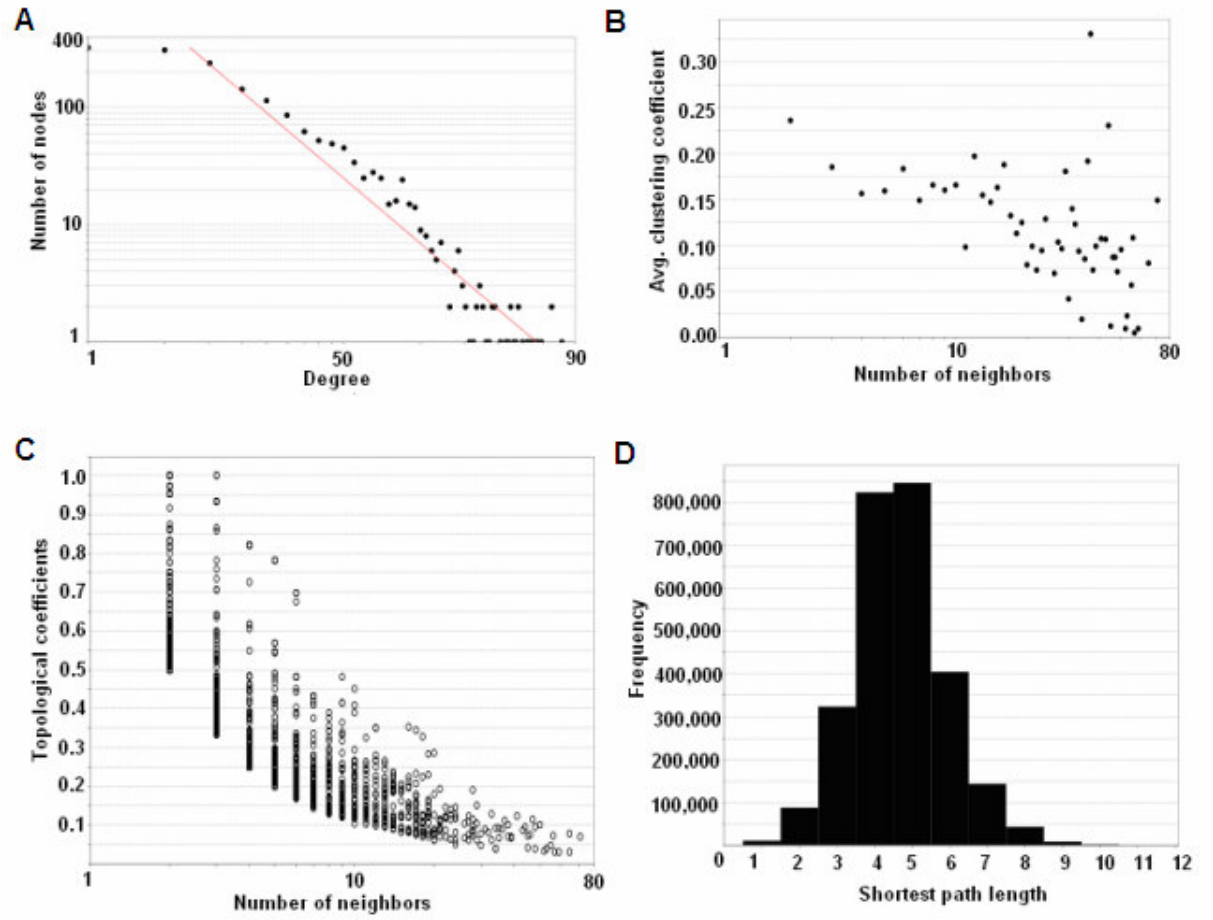

Figure 4.2. Topological properties of SPIN. (A) Degree distribution of proteins, $R^2 = 0.914$ for power law fit (B) Average clustering coefficient (C) Topological coefficients (D) Shortest path length distribution

4.3.1 Hubs are more important to characterize essential genes rather than bottlenecks

Recently, Yu et al (2007) [20] have analyzed the significance of hubs (proteins with high degree distribution) and bottlenecks (proteins with high betweenness) in the yeast protein-protein interaction network and regulatory networks. They have investigated which quantity (degree distribution or betweenness) is a better predictor of protein essentiality. It has been found that in directed networks, for example in regulatory network, betweenness is a more important feature in terms of essentiality. In yeast regulatory networks, they have observed that bottlenecks (both hub-bottlenecks and nonhub-bottlenecks) are generally products of essential genes, whereas hub-nonbottlenecks are not essential at all. When they have analyzed the protein-protein interaction network in yeast (undirected network), they found that degree is a much better predictor of essentiality since hub-nonbottlenecks are much more essential than nonhub-bottlenecks.

We have also investigated how degree and betweenness correlate with essentiality in protein-protein interaction network in human. We have classified all proteins into four categories; hub-bottleneck, hub-nonbottleneck, nonhub-bottleneck and nonhub-nonbottleneck. **Figure 4.3 (A**, **B)** show the essentiality of different categories of proteins, in PIN and in SPIN. In addition to these networks, a random network is generated from PIN which is the same size as SPIN and has the same average degree distribution. In **Figure 4.3**, the hub-bottlenecks are the most essential category compared to others in both networks. It is observed that hub-nonbottlenecks are more essential than nonhub-bottlenecks; i.e. degree is a more important parameter in terms of essentiality all in PIN, SPIN and the random network. This finding confirms the hypothesis stated by Yu et al (2007) [20].

Although the relationship trend for the fraction of essential genes in different categories of proteins is quite the same among PIN and SPIN, the essentiality fraction values in SPIN are much higher than the ones in PIN. (see the y-axes of **Figure 4.3**A and **Figure 4.3**B)

The reason for higher fraction of essential genes in SPIN may stem from the possible bias towards proteins which have structural information since they are well studied proteins. Another reason would be the physical bias due to the fact that PIN is a large-scale data. To investigate the reason for this bias, we generated a random network from PIN which is the same size as SPIN and has the same average degree distribution. **Figure 4.3C** displays the fraction of essential genes in random network. We observe that the fraction of essentiality is higher for random network than PIN. However, compared to SPIN, the values are still much smaller. Thus, we conclude that the reason for higher essentiality in SPIN probably arises from a bias towards well studied proteins rather than a physical bias.

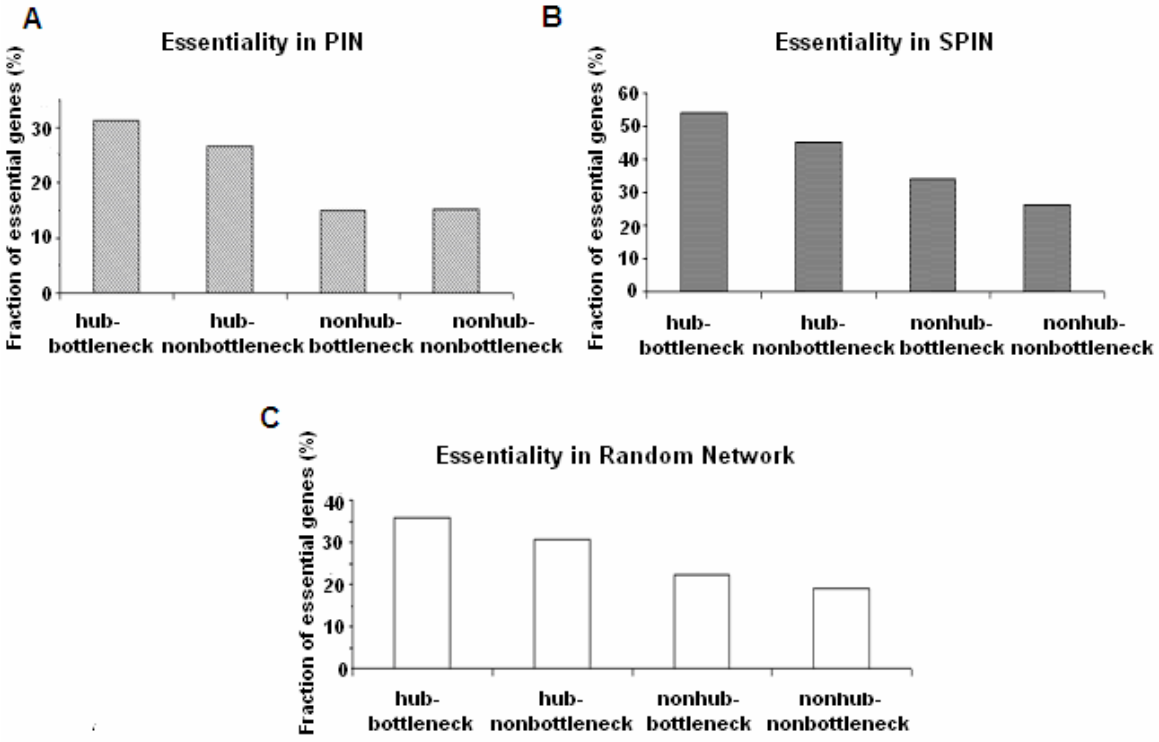

Figure 4.3 Essentiality of different categories of proteins in A) PIN, B) SPIN and C) random network.

4.3.2 Essentiality of cancer hubs is significantly higher than non-cancer hubs

It is known that hub proteins are more likely encoded by essential genes [18, 51]. In addition, somatic cancer genes are more likely to encode hub proteins [2]. From these, we can hypothesize that essential cancer genes are more likely to encode hub proteins than non-essential cancer genes. Thus, we classified all of the cancer genes in the networks as hub and non-hub, and observed that cancer-hubs are more essential than cancer-nonhubs, which confirms our above hypothesis; essential cancer genes are more likely to encode hub proteins than non-essential cancer genes. The essentiality percentage in each category; hubs and non-hubs are 50% (total 532) and 37% (total 650) for PIN, 66% (total 158) and 44% (total 286) for SPIN, 47% (total 85) and 37% (total 140) for random network, respectively. The essentiality percentage values are visualized in **Figure 4.4A**.

From a similar point of view, another question to be addressed is that; if we classify the hubs as cancer and non-cancer, which category is more essential. We found that when we classify the hub proteins as cancer-hubs and non-cancer-hubs, there is a significant difference in essentiality. In SPIN, there are 158 cancer hubs and 66% of them are essential while the value decreases to 38% when we consider the non-cancer-hub (197 non-cancer hubs) essentiality percentage. In PIN and the random network, similarly, cancer hubs (fraction of essential genes: 50% (532 in total) and 47% (85 in total), respectively) are much more essential than non-cancer hub proteins (fraction of essential genes: 24% (1801 in total) and 30% (246 in total), respectively). The fraction of essential genes in cancer hubs and non-cancer hubs for each network are shown in **Figure 4.4B**.

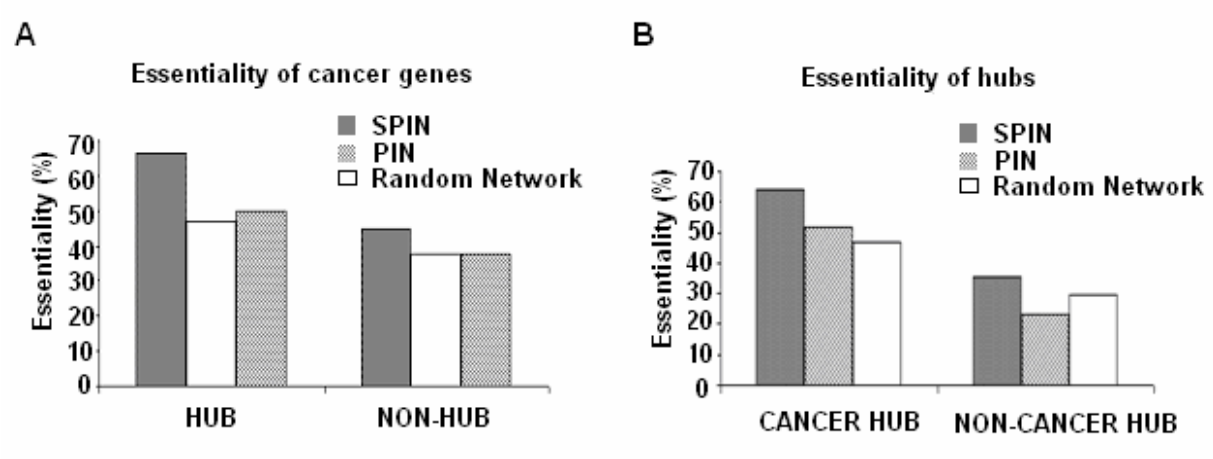

Figure 4.4 A) Essentiality of cancer proteins classified as hub & non-hub B) Essentiality of hub proteins classified as cancer and noncancer in each network.

4.3.3 Multi-interface and single-interface proteins: Correspondence with degree, betweenness and essentiality

Some hubs, as discussed above, are single-interface (communicating with their partners by using the same interface) whereas others are multi-interface. We investigated to which category (hub-bottleneck or hub-nonbottleneck) multi-interface and single-interface proteins mostly belong to. We observe that multi-interface proteins mostly correspond to hub-bottleneck proteins rather than hub-nonbottlenecks (71 % of multi-interface proteins are hub-bottlenecks.) When the single-interface proteins are considered, the percentage of hub-bottleneck correspondence decreases to 59%. To put it in other words, 58% of hub-bottleneck proteins are multi-interface and 42% are single-interface. Previously, we have shown that hub-bottlenecks are the most essential category of proteins in SPIN and in PIN. Here, in structural interface network, we find that the essentiality of multi-interface hubs (68% is essential) is higher than that of single-interface (52% is essential). This result agrees with a previous finding [10] indicating that it is the number of interfaces that leads

to higher essentiality. In addition, Aragues et al. (2007) found that yeast hubs with multiple interacting motifs are more likely to be essential than hubs with one or two interacting motifs [52]. Being more essential and corresponding mostly to hub-bottlenecks, multi-interface hubs are the key points in the protein-protein interaction network.

In terms of cancer/noncancer comparison, cancer proteins are more enriched in multi-interface proteins. (56% of cancer proteins are multi-interface, 44% being single-interface) This is reasonable since it is known that on average cancer proteins are longer with larger surface areas. To cope with many interactions at the same time, they tend to be multi-interface hubs, with distinct interfaces interacting with different proteins. Although cancer proteins tend to have more than one distinct interface, we found that on average their interfaces are smaller, which can indicate their binding behavior acting as hub proteins. In addition, average number of interfaces of cancer multi-interface hubs and noncancer multi-interface hubs are 2.5 and 2.3, respectively, being greater for the cancer multi-interface hubs. The correspondence of hub-bottlenecks, hub-nonbottlenecks to multi/single interface proteins and essentiality percentage in cancer/noncancer & multi/single interface proteins are displayed in **Table 4.5**.

Table 4.5. Correspondence of HB, H-NB to Multi/single interface proteins and Essentiality % in cancer/noncancer & Multi/single interface proteins

	HB	**H-NB**	**Total**
Multi-interface #	30	12	42
Single-interface #	22	15	37
		Essentiality percentage (%)	
Multi-interface		68	
Single-interface		52	

Cancer	76
Non-cancer	42

4.4 Case Studies

The interface information is an asset to predict which interactions can and cannot co-exist. In other words, it will help to deduce which interactions can occur simultaneously and which are mutually excluded. Addressing this question, may add a fourth dimension into interaction maps: that of sequence of processes. Including the *sequence* dimension in structural networks is an immense asset; transforming network node-and-edge maps into cellular processes, and assisting in the comprehension of cellular pathways and their regulation. Here, we present two case studies; a multi-interface cancer protein and a single-interface cancer protein in iSPIN and an inhibitor targeting the network for characterization of the interactions and inferring the order of processes. In **Figure 4.5**, a visualization of iSPIN is displayed together with multi-interface and single-interface proteins.

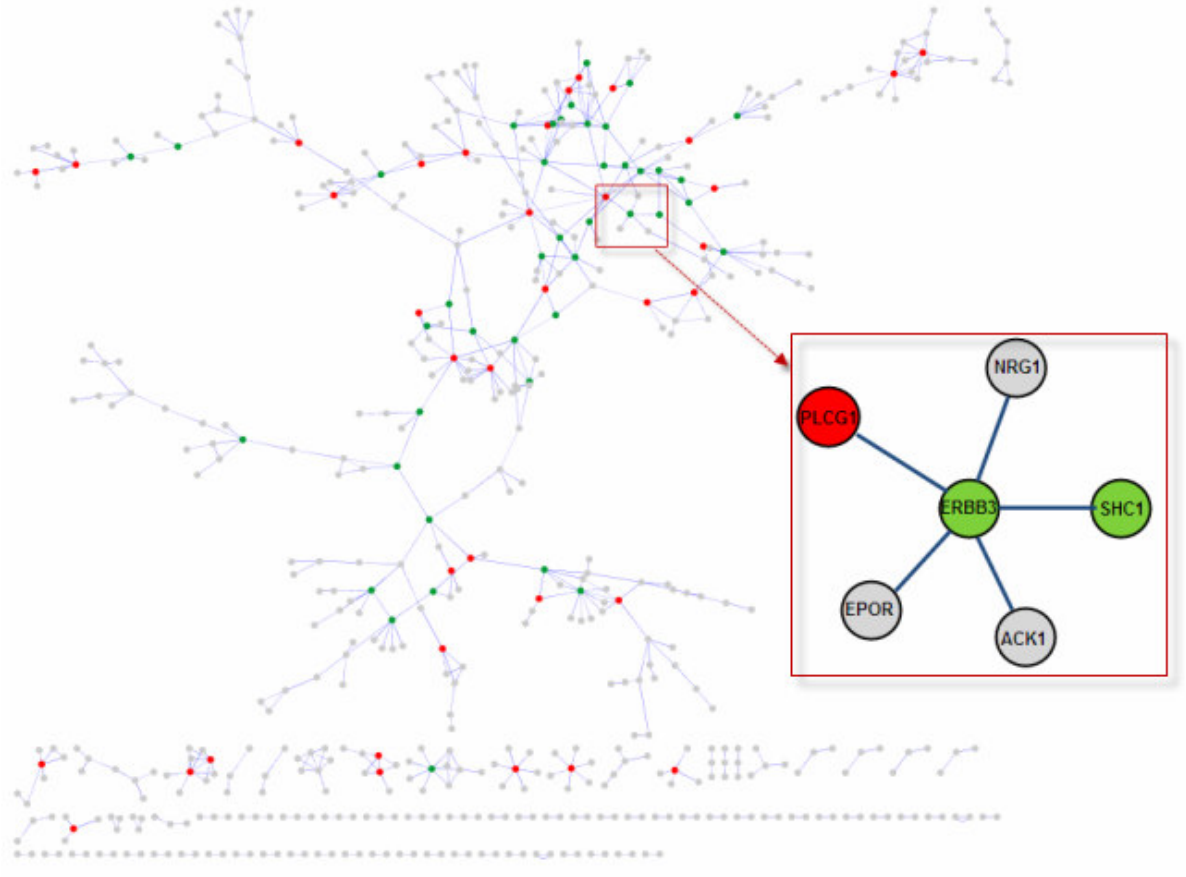

Figure 4.5 iSPIN representation. The nodes colored in green and red are multi-interface hubs and single-interface hubs, respectively. In the zoomed representation, the interactions of a multi-interface hub; ERBB3 is displayed.

4.4.1 A multi-interface hub: ErbB3 (Her3)

Here, we show how the interface information is used to deduce which interactions can and cannot co-exist. If each interaction partner of a hub protein use distinct interface on the hub while interacting, then these interactions are more likely to occur simultaneously. In addition, quaternary structure of the complex should well be considered to ensure that the interaction partners do not collide. To exemplify the idea, we present a so-called 'multi-

interface' hub protein; ERBB3 (or HER3), which is one of the hub proteins in SPIN. Receptor tyrosine-protein kinase erbB-3 precursor (ERBB3) belongs to EGF receptor subfamily and acts as a heregulin receptor and as an epidermal growth factor receptor. Amplification of this gene and/or overexpression of its protein have been reported in numerous cancers, including prostate, bladder, and breast tumors [53]. According to KEGG database [54], ERBB3 functions in ErbB signaling pathway and Calcium signaling pathway. In ErbB signaling pathway, NRG1 (neuregulin 1, heregulin), which is a direct ligand for ERBB3, binds and activates ERBB3. We have modeled this interaction using the PDB accession codes 1hae_A (NMR structure of heregulin) for NRG1; and 1m6b_A (crystal structure of ERBB3 taken from a homodimer structure) for ERBB3, respectively. PRISM results indicate that these two proteins (1hae_A and 1m6b_A) interact and using NOXclass [32], we find that the interaction is biologically relevant. Predicted binding sites on both proteins and interacting residues for NRG1-ERBB3 interaction are shown in **Figure 4.6A.** The interaction was experimentally studied in a previous study of Jones et al (1998) [55], where they mutated individual residues of the egf domain of heregulinβ (same as egf domain of heregulinα-NRG1- except four residues) to alanine in order to determine residues critical for binding receptors and initiating signal transduction. They have found that when His^{2}, Leu^{3}, Val^{4}, Phe^{13}, Val^{15}–Gly^{18}, Val^{23}, Arg^{31}, Lys^{35}, Gly^{42}–Gln^{46} residues were changed to alanine, binding affinity for ERBB3 was dramatically reduced. We have observed that most of these critical residues are included in our predicted binding site for NRG1. In **Figure 4.6A**, these residues are labeled.

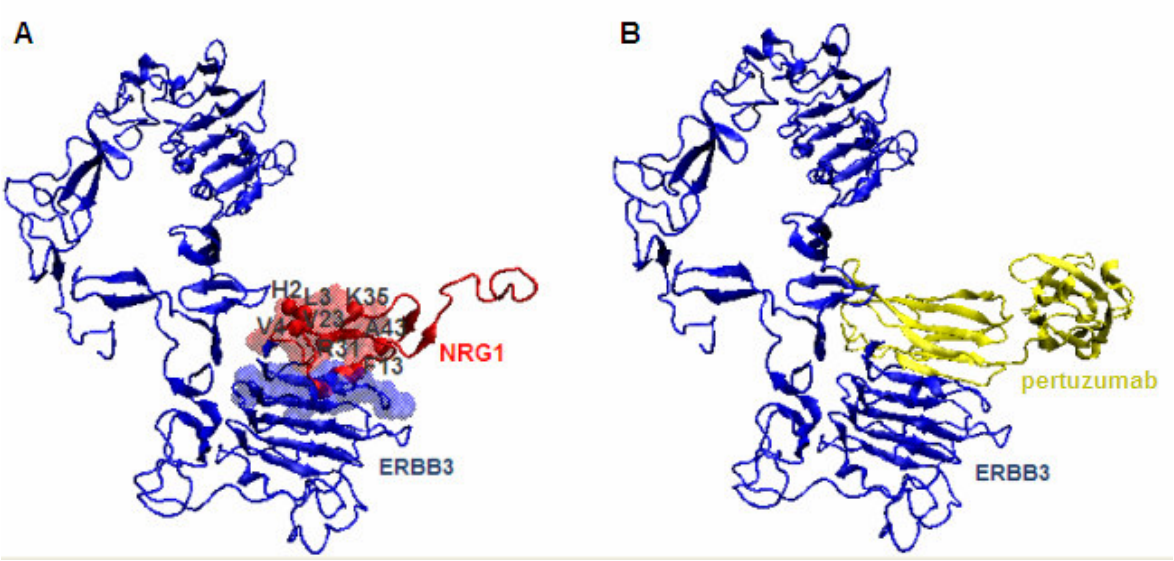

Figure 4.6 Representation of ERBB3-NRG1 interaction schematically using VMD [56] A) ERBB3 (1m6b_A) and NRG1 (1hae_A) are shown as newcartoon diagram in blue and red color, respectively. The transparent surface represents the interface region. The labeled residues (represented by their Cα atoms) of 1hae_A are reported to be critical for binding in a previous work [55]; i.e. when they are mutated to alanine, the binding affinity for ERBB3 was significantly reduced. B) HER3 (blue) – pertuzumab heavy chain (yellow) is shown, docked conformation using HEX [57]. Pertuzumab shares the same interface with NRG1 (see the next section).

In ErbB signaling pathway, NRG1 also binds to ERBB4, and the binding affinity is similar to that of ERBB3 [55]. According to our interface prediction, ERBB3 and ERBB4 binding interfaces on NRG1 are overlapping; i.e. the same binding site is used for the ERBB3 and ERBB4 interactions. Thereby, NRG1-ERBB3 and NRG1-ERBB4 interactions are mutually exclusive; they cannot occur at the same time.

According to the Calcium signaling pathway in KEGG [54], ERBB3 interacts with PLCG1. Although the interaction is not reported in public databases as in DIP [25], BIND [58], in a recent study, it has been observed on the protein microarrays [59]. PLCG1 (Phospholipase C-gamma-1) is a major substrate for heparin-binding growth factor 1 (acidic fibroblast growth factor)-activated tyrosine kinase. PDB structure of SH3 domain of PLCG1 is 1hsq. The predicted interface residues of ERBB3-PLCG1 (1m6b_A-1hsq_A) interaction are displayed in **Figure 4.7** labeled as A.

Two other possible interactions of ERBB3 occur with EPOR (Erythropoietin receptor) and ACK1 (Activated CDC42 kinase 1) according to the human interactome constructed by Jonsson and Bates. No experimental confirmation is available for these interactions yet, however they have high confidence scores to be interacting in Jonsson and Bates's network [3]. Since they are also predicted to be interacting via PRISM [11, 12], these interactions of ERBB3 are further investigated. Subcellular location for ERBB3, EPOR and ACK1 is the cell membrane. EPOR and ERBB3 function as single-pass type I membrane protein. The predicted interfaces for these interactions are illustrated in **Figure 4.7**, labeled as B and C.

Our results show that ERBB3 uses at least three different binding sites while interacting. Of these interactions, we propose that ERBB3 cannot interact with EPOR and ACK1 at the same time. Because if we model the quaternary structure of ERBB3-EPOR-ACK1 complex, the residues of EPOR and ACK1 will collide. Thus, they cannot bind simultaneously. But, we should keep in mind that proteins are dynamic, and hinge-like motion of the two domains of ERBB3 can eliminate the collision between EPOR and ACK1.

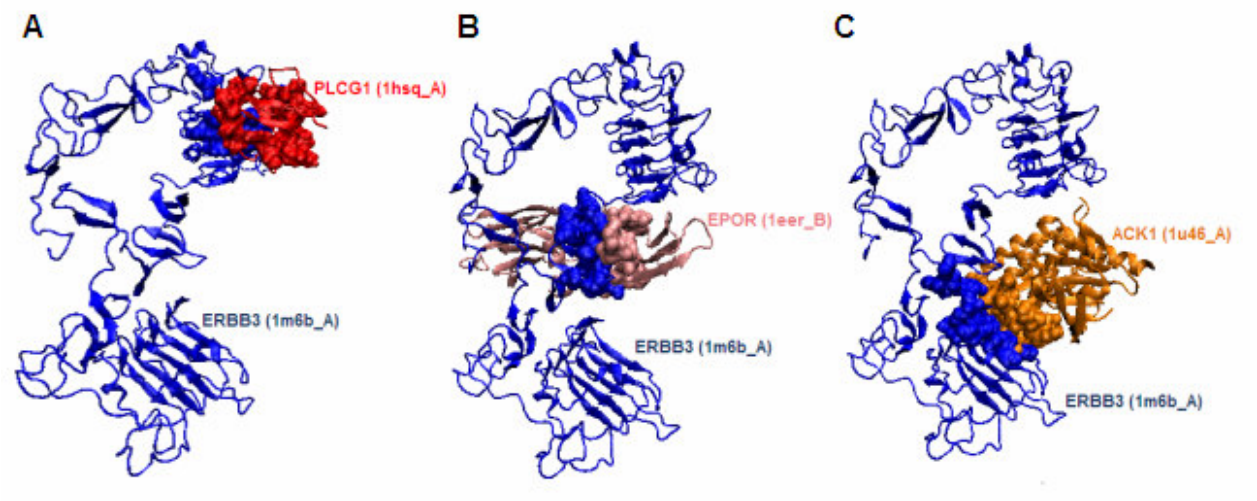

Figure 4.7 Ribbon diagram and interface representation of ERBB3 interactions with PLCG1, EPOR and ACK1 labeled as A, B, C respectively. ERBB3 (1m6b_A), PCLG1 (1hsq_A), EPOR (1eer_B) and ACK1 (1u46_A) are colored in blue, red, pink and orange respectively. Interface residues are shown as spheres.

4.4.2 An inhibitor affecting Erb signaling pathway: pertuzumab

To illustrate the importance of sequence of processes, we further focus on ERRB3 interactions and investigate how it functions if its partners use the same interface while interacting (in this case the interactions cannot occur at the same time). In general, HER/erbB family of proteins (EGFR (HER1), HER2, HER3, and HER4) activate intracellular signaling pathways in response to extracellular signals [53]. The signaling mechanism is as follows; first EGFR and HER3 are activated by ligand binding (ligands are EGF and NRG1 for EGFR and HER3, respectively), then EGFR or HER3 forms heterodimer with HER2 followed by the transphosphorylation of their C-terminal tails. Heterodimer formation of HER2 with EGFR and HER3 induces different pathways. For example, The PI3K/Akt pathway, which is critically important in tumorigenesis, is

activated by phosphorylated HER3. Deregulation of signaling functions of HER family of proteins causes cell transformation and tumorigenic growth [53]. In anti-cancer drug development, EGFR and HER2 proteins are the main targets. For example, pertuzumab, which targets HER2 dimerization region, tries to inhibit HER2-HER3 or HER2-EGFR interactions.

In a recent study [60] which investigates the effect of pertuzumab in lung cancer cells, it is found that pertuzumab blocked NRG1-stimulated phosphorylation of HER3 and in contrast, it failed to block epidermal growth factor (EGF)-stimulated phosphorylation of EGFR in human non-small cell lung cancer cell line 11_18. It is somewhat interesting since HER2 uses same binding region for dimerization with HER3 and EGFR and this region is assumed to be blocked by pertuzumab. However, at that point, it may be hypothesized that in addition to its inhibiting effect on dimerization region of HER2, pertuzumab should also affect the ligand binding region of HER3 and EGFR; namely HER3-NRG1 interaction and EGFR1-EGF interaction.

In order to investigate the effect of pertuzumab on HER3-NRG1 interaction, we used HEX [57], a protein docking program, to dock pertuzumab heavy chain (PDB ID 1s78) to HER3 (PDB ID 1m6b) and the docked conformation is visualized in **Figure 4.6B**. NOXclass results indicate that the docked conformation is biological (biological score is 70%). Although HER2 and HER3 are similar in structure, the interface region on HER2 and HER3 through which the interaction with pertuzumab occur, are not exactly the same in structure, but rather they use overlapping regions. We have observed that pertuzumab binding interferes with NRG1 binding region, which indicates that pertuzumab may also block ligand binding to HER3 and thus preventing HER3 activation. 36% of interface residues (8 out of 22) of HER3-NRG1 interface are also used by pertuzumab, which makes the interactions of HER3 with NRG1 and pertuzumab mutually exclusive. In **Figure 4.8**,

both interactions are visualized together and the black surface region shows the shared interface region.

Thus, our results indicate that pertuzumab may block NRG1 interaction region of HER3. Probably, pertuzumab would not affect the binding of EGF to EGFR and thus it is not effective against (EGF)-stimulated phosphorylation of EGFR in the aforementioned lung cancer cells.

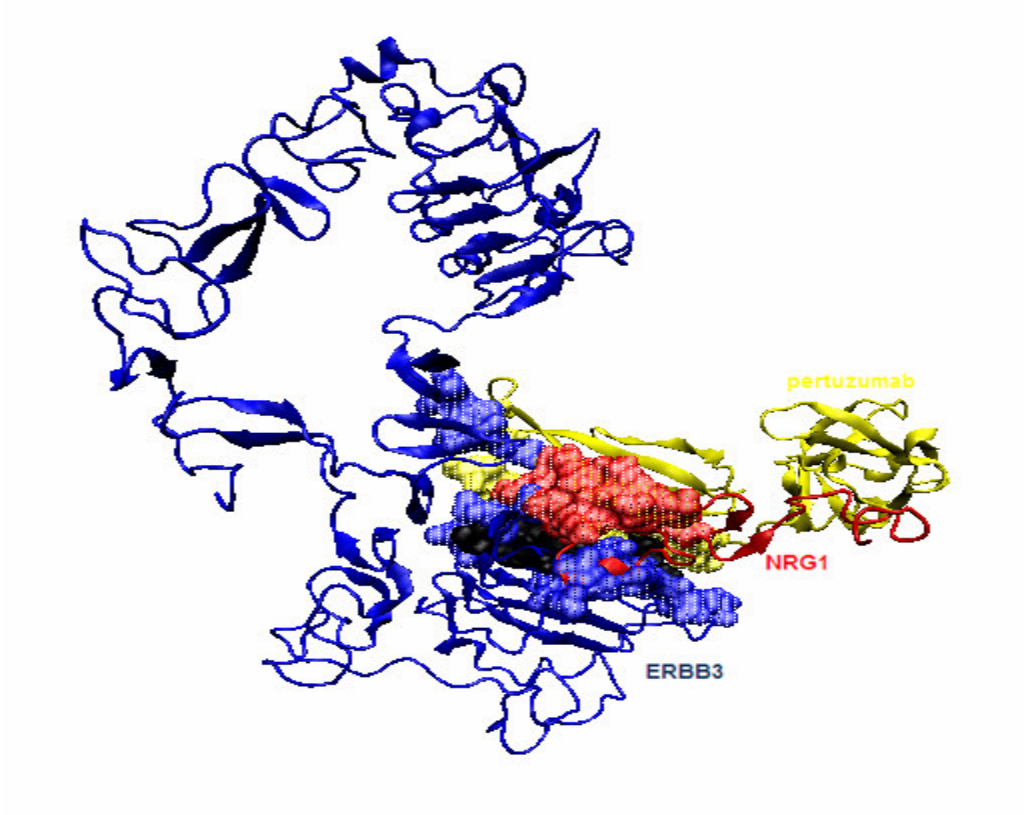

Figure 4.8 HER3 (blue) interactions with NRG1 (red) and pertuzumab (yellow). The interface regions for HER3-NRG1 and HER3-pertuzumab interactions are shown spheres in surface representation in blue-red and blue-yellow color, respectively. The black surface is the shared interface region which implies that these two interactions cannot occur at the same time.

4.4.3 A Single-interface Hub: RAF1

If the interaction partners of a hub protein use the same interface region, then these interactions are more likely to be mutually exclusive. For example, in iSPIN, RAF1 has 9 interactions partners which compete for binding. RAF proto-oncogene serine/threonine-protein kinase involves in the transduction of mitogenic signals from the cell membrane to the nucleus and protects cells from apoptosis mediated by STK3. Among its interaction partners, we were able to predict interaction interfaces for CDC25, YWHAZ and MAP2K2. Interaction with RAP1A is a known structure with PDB ID 1c1y. Also possible interactions of RAF1 in iSPIN are with RALA, DIRAS1, DIRAS2, CCNA2 and RRAD. Although the interface region is not completely the same for each interaction partner, most of the interface residues are shared (shared percentage > 20, which is the cutoff value for assigning the interface as distinct or same). Thus, these interactions cannot occur at the same time. Three predicted binding sites are illustrated in **Figure 4.9.** In **Figure 4.10**, RAF1 is displayed with its three binding partners; RAF1 (1c1y_B) is shown in blue color, the partners YWHAZ (1qja_A), MAP2K2 (1s9i_A) and CDC25A (1c25_A) are colored in red, cyan and purple respectively. Since the interface is highly shared, the interface region is assigned as the same (not distinct) for all of the interactions which hypothesize that RAF1 is a single-interface protein and involved in mutually exclusive interactions. RAF1 is a protein kinase and it is a signaling protein; thus it probably interacts transiently with most of its targets. A recent study confirms this interaction behavior of RAF1 by showing that the binding of Cdc25 and of Rad24 (14-3-3 homolog that is important in the DNA damage checkpoint) to Raf-1 is mutually exclusive [61].

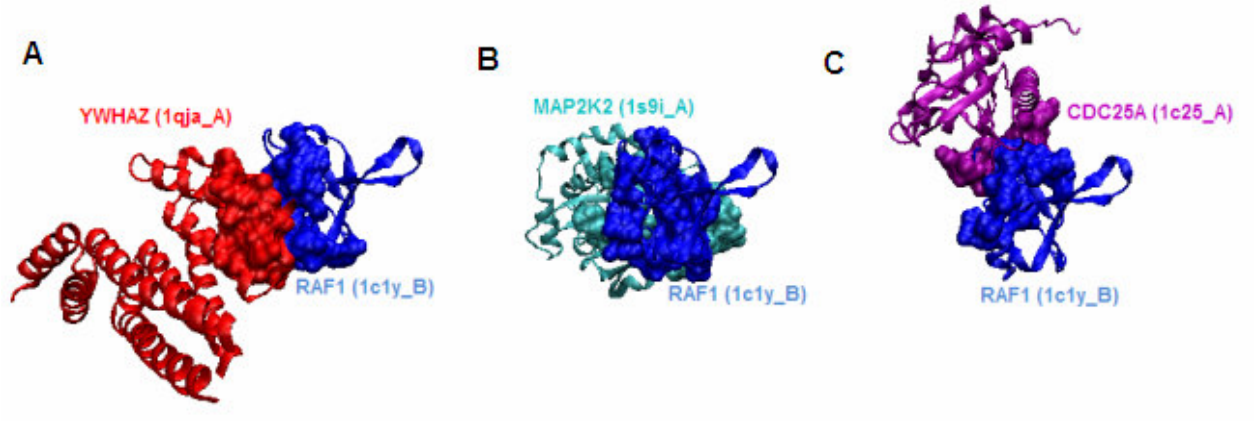

Figure 4.9 Ribbon diagram and interface representation of RAF1 interactions with YWHAZ, MAP2K2 and CDC25A labeled as A, B, C respectively. RAF1 (1c1y_B), YWHAZ (1qja_A), MAP2K2 (1s9i_A) and CDC25A (1c25_A) are colored in blue, red, cyan and purple respectively. Interaction interfaces of RAF1 through YWHAZ, MAP2K2 and CDC25A are highly overlapping; the interactions are mutually exclusive.

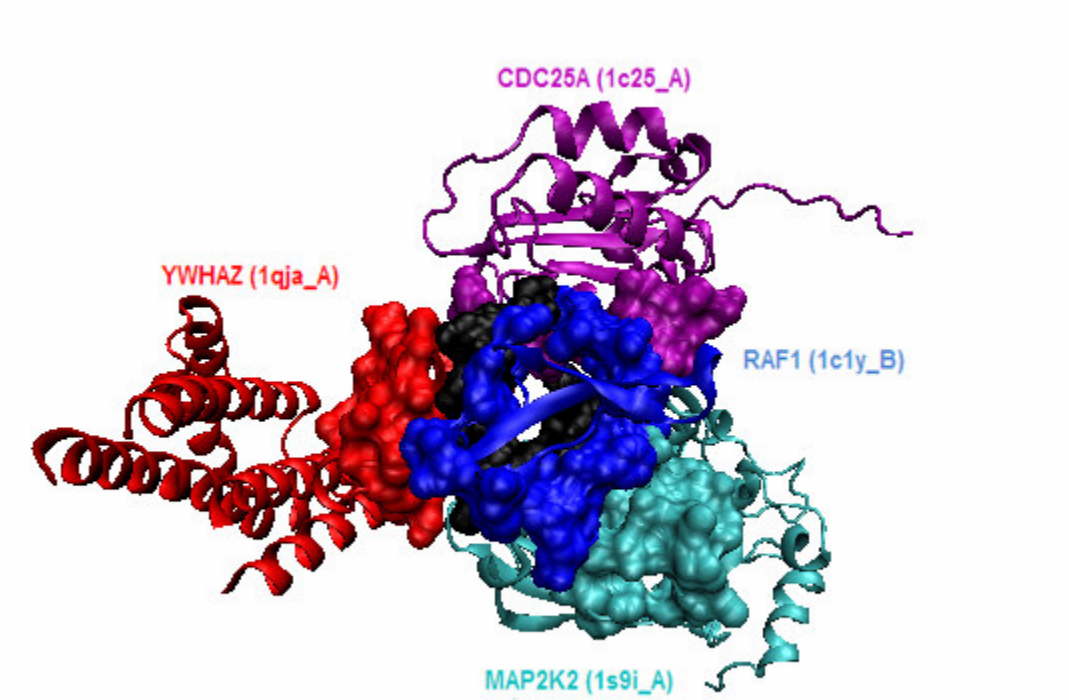

Figure 4.10 RAF1 is displayed with its three binding partners; RAF1 (1c1y_B) is shown in blue color, the partners YWHAZ (1qja_A), MAP2K2 (1s9i_A) and CDC25A (1c25_A) are colored in red, cyan and purple respectively. The interface regions are shown as surfaces; the surface colored in black is the shared interface residues between the interactions.

Chapter 5

CONCLUSION

The proposition that protein-protein interaction networks are scale-free [62] has directed the focus onto highly connected proteins, namely “hub proteins”, which are known for their critical significance. In addition, cancer genes are known to encode hub proteins. In this work, we analyzed hub proteins and cancer proteins in human protein-protein interaction network first by considering their global behavior in the network, second, from a structural perspective.

We know that hub proteins are considered to be essential proteins in the biological networks [51]. Upon investigating another topological property, betweenness, together with node degree to observe which one is a better determinant of essentiality, we found that hubs are more important to characterize essential genes rather than bottlenecks, similar to what is observed for protein interaction networks previously by Yu et al. [20]. Interestingly, this notion is found to be more pronounced for cancer-related hub proteins, being much more essential than non-cancer hubs in our network. Thus, the reason for cancer genes to act mostly as hub proteins stems from the fact that they correspond mostly to essential genes, which are known to encode hubs.

Although topological properties can provide an explanation of hub proteins in the global sense, we need more concrete reasons to clarify the ability of a hub to interact with multiple partners. One way to do that is using structural information. Integrating three-dimensional protein structures into our network revealed important aspects about hubs and

cancer-related proteins. Interface property analysis has served to identify structural tendencies in hubs that assist their binding to multiple proteins and observe the characteristics that differentiate hubs from non-hubs and cancer from non-cancer proteins. We find that hub proteins have smaller, less tightly packed and more planar interfaces compared to non-hub ones. The differences in interface properties are more pronounced when we consider the cancer/non-cancer case. The interfaces of cancer proteins, on average, are smaller in size, more planar, less tightly packed and more hydrophilic compared to non-cancer ones. Friedler et al. (2005) [63] observed a highly electrostatic binding site in a cancer protein, p53, interacting with Rad51 and other peptide sequences with different affinity. The results imply that cancer proteins and hubs interact with their partners with high specificity and low affinity. Therefore, it becomes possible for them to bind to many different proteins with varying affinity.

To exemplify different kind of hubs with respect to the number of distinct binding sites, we present a multi-interface and a single-interface hub protein as case studies. Here, structural information provides answers regarding to which interactions are simultaneously possible and which ones exclude each other. Protein structures, whether known or modeled provide crucial information about binding behavior and interaction specificity. In this thesis study, we illustrated the advantages of including structures into protein interaction networks.

BIBLIOGRAPHY

[1] Hartwell LH, Hopfield JJ, Leibler S, Murray AW: **From molecular to modular cell biology**. *Nature* 1999, **402**(6761 Suppl):C47-52.

[2] Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL: **The human disease network**. *Proc Natl Acad Sci U S A* 2007, **104**(21):8685-8690.

[3] Jonsson PF, Bates PA: **Global topological features of cancer proteins in the human interactome**. *Bioinformatics* 2006, **22**(18):2291-2297.

[4] Wachi S, Yoneda K, Wu R: **Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues**. *Bioinformatics* 2005, **21**(23):4205-4208.

[5] Oti M, Snel B, Huynen MA, Brunner HG: **Predicting disease genes using protein-protein interactions**. *J Med Genet* 2006, **43**(8):691-698.

[6] Xu J, Li Y: **Discovering disease-genes by topological features in human protein-protein interaction network**. *Bioinformatics* 2006, **22**(22):2800-2805.

[7] Aloy P, Russell RB: **Interrogating protein interaction networks through structural biology**. *Proc Natl Acad Sci U S A* 2002, **99**(9):5896-5901.

[8] Aloy P, Bottcher B, Ceulemans H, Leutwein C, Mellwig C, Fischer S, Gavin AC, Bork P, Superti-Furga G, Serrano L *et al*: **Structure-based assembly of protein complexes in yeast**. *Science* 2004, **303**(5666):2026-2029.

[9] Aloy P, Pichaud M, Russell RB: **Protein complexes: structure prediction challenges for the 21st century**. *Curr Opin Struct Biol* 2005, **15**(1):15-22.

[10] Kim PM, Lu LJ, Xia Y, Gerstein MB: **Relating three-dimensional structures to protein networks provides evolutionary insights**. *Science* 2006, **314**(5807):1938-1941.

[11] Aytuna AS, Gursoy A, Keskin O: **Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces**. *Bioinformatics* 2005, **21**(12):2850-2855.

[12] Ogmen U, Keskin O, Aytuna AS, Nussinov R, Gursoy A: **PRISM: protein interactions by structural matching**. *Nucleic Acids Res* 2005, **33**(Web Server issue):W331-336.

[13] Kann MG: **Protein interactions and disease: computational approaches to uncover the etiology of diseases**. *Brief Bioinform* 2007, **8**(5):333-346.

[14] McKusick VA: **Mendelian Inheritance in Man and its online version, OMIM**. *Am J Hum Genet* 2007, **80**(4):588-604.

[15] Rambaldi D, Giorgi FM, Capuani F, Ciliberto A, Ciccarelli FD: **Low duplicability and network fragility of cancer genes**. *Trends Genet* 2008, **24**(9):427-430.

[16] Brunner HG, van Driel MA: **From syndrome families to functional genomics**. *Nat Rev Genet* 2004, **5**(7):545-551.
[17] Franke L, van Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C: **Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes**. *Am J Hum Genet* 2006, **78**(6):1011-1025.
[18] Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJ, Cusick ME, Roth FP *et al*: **Evidence for dynamically organized modularity in the yeast protein-protein interaction network**. *Nature* 2004, **430**(6995):88-93.
[19] de Lichtenberg U, Jensen LJ, Brunak S, Bork P: **Dynamic complex formation during the yeast cell cycle**. *Science* 2005, **307**(5710):724-727.
[20] Yu H, Kim PM, Sprecher E, Trifonov V, Gerstein M: **The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics**. *PLoS Comput Biol* 2007, **3**(4):e59.
[21] Kiel C, Beltrao P, Serrano L: **Analyzing protein interaction networks using structural information**. *Annu Rev Biochem* 2008, **77**:415-441.
[22] Kim PM, Sboner A, Xia Y, Gerstein M: **The role of disorder in interaction networks: a structural analysis**. *Mol Syst Biol* 2008, **4**:179.
[23] Dawelbait G, Winter C, Zhang Y, Pilarsky C, Grutzmann R, Heinrich JC, Schroeder M: **Structural templates predict novel protein interactions and targets from pancreas tumour gene expression data**. *Bioinformatics* 2007, **23**(13):i115-124.
[24] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.
[25] Xenarios I, Fernandez E, Salwinski L, Duan XJ, Thompson MJ, Marcotte EM, Eisenberg D: **DIP: The Database of Interacting Proteins: 2001 update**. *Nucleic Acids Res* 2001, **29**(1):239-241.
[26] Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Mark P, Stumpflen V, Mewes HW *et al*: **The MIPS mammalian protein-protein interaction database**. *Bioinformatics* 2005, **21**(6):832-834.
[27] Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR: **A census of human cancer genes**. *Nat Rev Cancer* 2004, **4**(3):177-183.
[28] Gasteiger E, Jung E, Bairoch A: **SWISS-PROT: connecting biomolecular knowledge via a protein database**. *Curr Issues Mol Biol* 2001, **3**(3):47-55.
[29] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank**. *Nucleic Acids Res* 2000, **28**(1):235-242.
[30] Girvan M, Newman ME: **Community structure in social and biological networks**. *Proc Natl Acad Sci U S A* 2002, **99**(12):7821-7826.

[31] Tuncbag N, Gursoy A, Guney E, Nussinov R, Keskin O: **Architectures and functional coverage of protein-protein interfaces**. *J Mol Biol* 2008, **381**(3):785-802.

[32] Zhu H, Domingues FS, Sommer I, Lengauer T: **NOXclass: prediction of protein-protein interaction types**. *BMC Bioinformatics* 2006, **7**:27.

[33] Hubbard SJ, & Thornton JM: **'NACCESS', Computer Program**. *Department of Biochemistry and Molecular Biology, University College London* (1993).

[34] Guney E, Tuncbag N, Keskin O, Gursoy A: **HotSprint: database of computational hot spots in protein interfaces**. *Nucleic Acids Res* 2008, **36**(Database issue):D662-666.

[35] Reynolds C, Damerell D, Jones S: **ProtorP: a protein-protein interaction analysis server**. *Bioinformatics* 2009, **25**(3):413-414.

[36] Laskowski RA: **SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions**. *J Mol Graph* 1995, **13**(5):323-330, 307-328.

[37] Jones S, Thornton JM: **Protein-protein interactions: a review of protein dimer structures**. *Prog Biophys Mol Biol* 1995, **63**(1):31-65.

[38] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks**. *Genome Res* 2003, **13**(11):2498-2504.

[39] Nooren IM, Thornton JM: **Structural characterisation and functional significance of transient protein-protein interactions**. *J Mol Biol* 2003, **325**(5):991-1018.

[40] Bahadur RP, Chakrabarti P, Rodier F, Janin J: **A dissection of specific and non-specific protein-protein interfaces**. *J Mol Biol* 2004, **336**(4):943-955.

[41] Keskin O, Nussinov R: **Similar binding sites and different partners: implications to shared proteins in cellular pathways**. *Structure* 2007, **15**(3):341-354.

[42] Tsai CJ, Lin SL, Wolfson HJ, Nussinov R: **Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect**. *Protein Sci* 1997, **6**(1):53-64.

[43] Hu Z, Ma B, Wolfson H, Nussinov R: **Conservation of polar residues as hot spots at protein interfaces**. *Proteins* 2000, **39**(4):331-342.

[44] Tormo J, Natarajan K, Margulies DH, Mariuzza RA: **Crystal structure of a lectin-like natural killer cell receptor bound to its MHC class I ligand**. *Nature* 1999, **402**(6762):623-631.

[45] Liang S, Li L, Hsu WL, Pilcher MN, Uversky V, Zhou Y, Dunker AK, Meroueh SO: **Exploring the molecular design of protein interaction sites with molecular**

dynamics simulations and free energy calculations. *Biochemistry* 2009, **48**(2):399-414.

[46] Iakoucheva LM, Brown CJ, Lawson JD, Obradovic Z, Dunker AK: **Intrinsic disorder in cell-signaling and cancer-associated proteins**. *J Mol Biol* 2002, **323**(3):573-584.

[47] Higurashi M, Ishida T, Kinoshita K: **Identification of transient hub proteins and the possible structural basis for their multiple interactions**. *Protein Sci* 2008, **17**(1):72-78.

[48] Patil A, Nakamura H: **Disordered domains and high surface charge confer hubs with the ability to interact with multiple proteins in interaction networks**. *FEBS Lett* 2006, **580**(8):2041-2045.

[49] Barabasi AL, Oltvai ZN: **Network biology: understanding the cell's functional organization**. *Nat Rev Genet* 2004, **5**(2):101-113.

[50] Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL: **Hierarchical organization of modularity in metabolic networks**. *Science* 2002, **297**(5586):1551-1555.

[51] Jeong H, Mason SP, Barabasi AL, Oltvai ZN: **Lethality and centrality in protein networks**. *Nature* 2001, **411**(6833):41-42.

[52] Aragues R, Sali A, Bonet J, Marti-Renom MA, Oliva B: **Characterization of protein hubs by inferring interacting motifs from protein interactions**. *PLoS Comput Biol* 2007, **3**(9):1761-1771.

[53] Maglott D, Ostell J, Pruitt KD, Tatusova T: **Entrez Gene: gene-centered information at NCBI**. *Nucleic Acids Res* 2005, **33**(Database issue):D54-58.

[54] Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes**. *Nucleic Acids Res* 2000, **28**(1):27-30.

[55] Jones JT, Ballinger MD, Pisacane PI, Lofgren JA, Fitzpatrick VD, Fairbrother WJ, Wells JA, Sliwkowski MX: **Binding interaction of the heregulinbeta egf domain with ErbB3 and ErbB4 receptors assessed by alanine scanning mutagenesis**. *J Biol Chem* 1998, **273**(19):11667-11674.

[56] Humphrey W, Dalke A, Schulten K: **VMD: visual molecular dynamics**. *J Mol Graph* 1996, **14**(1):33-38, 27-38.

[57] Ritchie DW: **Evaluation of protein docking predictions using Hex 3.1 in CAPRI rounds 1 and 2**. *Proteins* 2003, **52**(1):98-106.

[58] Bader GD, Donaldson I, Wolting C, Ouellette BF, Pawson T, Hogue CW: **BIND--The Biomolecular Interaction Network Database**. *Nucleic Acids Res* 2001, **29**(1):242-245.

[59] Jones RB, Gordus A, Krall JA, MacBeath G: **A quantitative protein interaction network for the ErbB receptors using protein microarrays**. *Nature* 2006, **439**(7073):168-174.

[60] Sakai K, Yokote H, Murakami-Murofushi K, Tamura T, Saijo N, Nishio K: **Pertuzumab, a novel HER dimerization inhibitor, inhibits the growth of human lung cancer cells mediated by the HER3 signaling pathway**. *Cancer Sci* 2007, **98**(9):1498-1503.

[61] Lee M, Yoo HS: **Human Raf-1 proteins associate with Rad24 and Cdc25 in cell-cycle checkpoint pathway of fission yeast, Schizosaccharomyces pombe**. *J Cell Biochem* 2007, **101**(2):488-497.

[62] Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL: **The large-scale organization of metabolic networks**. *Nature* 2000, **407**(6804):651-654.

[63] Friedler A, Veprintsev DB, Rutherford T, von Glos KI, Fersht AR: **Binding of Rad51 and other peptide sequences to a promiscuous, highly electrostatic binding site in p53**. *J Biol Chem* 2005, **280**(9):8051-8059.

VITA

Gözde Kar was born in Bursa, Turkey, on June 7, 1983. She received her B.Sc. Degree in Chemical Engineering from Boğaziçi University, Istanbul, in 2006. From September 2006 to September 2008 she worked as a research and teaching assistant at Koç University. She has worked on "Protein-protein interfaces" and "Rigidity of hot regions at protein interfaces". She has attended ISMB/ECCB'07 (Vienna, Austria) where has presented a poster about "Determination of the correspondence between rigidity/mobility and conservation of the interface residues".

She currently lives in Istanbul, Turkey, and will continue her education with Ph.D. in Chemical and Biological Engineering at Koç University, Turkey.