Gene2Phen – A web-based tool to build, visualize and compare phenotype specific subnetworks

by

Bilgesu ERDOĞAN

A Dissertation Submitted to the

Graduate School of Sciences and Engineering

in Partial Fulfillment of the Requirements for

the Degree of

Masters of Science

in

Biomedical Sciences and Engineering

KOÇ
UNIVERSITY

February 2, 2017

To my family…

ABSTRACT

Diseases are commonly the result of dysregulated complex interactions involving large sets of genes and proteins as products of these genes, and their cooperation with other cellular components. Interpreting protein-protein interactions at both network and molecular interaction levels with mutation knowledge requires a comprehensive research process that is fed from different sources. In this thesis, we developed a web-based tool, Gene2Phen, by integrating large-scale protein-protein interaction network, 3D protein structure information and interface mutation knowledge to aid researchers in exploring and comparing the molecular mechanism of different phenotypes. Gene2Phen works as an automatized pipeline tool to build, visualize and compare phenotype specific subnetworks, to examine protein- protein interactions associated with their structure and mutation data. Gene2Phen web tool prioritizes the human protein-protein network based on seed genes specific to a phenotype. From the prioritized-PPI network, users can generate a phenotype specific subnetwork. The phenotype-specific subnetworks can be visualized and compared interactively. Genome annotations and topological properties of each protein are shown in this interactive network representation. A unique feature of Gene2Phen is its ability to display 3D structural models of protein-protein interactions and their predicted protein-protein interfaces. Users can see the list of mutations which are mapped on predicted protein-protein interfaces. This allows users to study mutations altering protein-protein interfaces and their role in the phenotype-specific subnetworks. Gene2Phen, by automating the integration of protein-protein networks, protein structure, and disease - related mutations at large scale, will not only boost the productivity and efficiency, but it may be the leveraging step to the novel solutions/studies.

ÖZET

Hastalıklar genellikle çok sayıda gen ve bu genlerin ürünü olan proteinlerin diğer hücresel bileşenlerle işbirliğiyle oluşturduğu karmaşık etkileşim mekanizmalarında meydana gelen bozulmalar sonucu gelişir. Protein – protein etkileşimlerini hem etkileşim ağı seviyesinde hem de mutasyon bilgisi ile birlikte moleküler seviyede incelemek ve yorumlamak, farklı kaynaklardan beslenen kapsamlı bir araştırma süreci gerektirir. Bu tezde, geniş ölçekli protein-protein etkileşim ağını, üç boyutlu protein yapısı bilgilerini ve etkileşim ara yüzlerinde görülen mutasyon bilgisini entegre ederek, araştırmacılara farklı fenotiplerin moleküler mekanizmalarını keşfetme ve diğer fenotiplerle karşılaştırmalarında yardımcı olacak web tabanlı bir araç olan Gene2Phen'i geliştirdik. Gene2Phen, protein-protein etkileşimlerini yapılarına ve mutasyon verilerine bağlı olarak inceleyebilmek için geliştirilmiş olup fenotipe özgü alt ağların oluşturulması, görselleştirilmesi ve karşılaştırılması için otomatikleştirilmiş bir iletişim hattı işlevi görür. Gene2Phen web aracı, bir fenotipe özgü tohum genlerine dayalı olarak insan protein-protein etkileşim ağını önceliklendirir. Kullanıcılar, önceliklendirilmiş etkileşim ağından fenotipe özgü bir alt ağ oluşturabilirler. Fenotipe özgü alt ağlar görselleştirilebilir ve interaktif olarak karşılaştırılabilir. Bu interaktif ağ gösteriminde her protein, genom ek açıklamaları ve topolojik özellikleri ile birlikte gösterilir. Gene2Phen'i eşsiz kılan bir özelliği, protein-protein etkileşimlerinin üç boyutlu yapısal modellerini ve öngörülen protein-protein ara yüzlerini görüntüleme yeteneğidir. Kullanıcılar, tahmin edilen protein - protein ara yüzleri üzerine eşlenmiş olan mutasyonların listesini görebilirler. Bu özellik kullanıcıların protein - protein ara yüzlerini değiştiren mutasyonları ve bu mutasyonların fenotipe özgü alt ağlardaki yerlerini öğrenebilmelerini sağlar. Gene2Phen, protein-protein ağlarının, protein yapısının ve hastalıkla ilgili mutasyonların büyük ölçekte entegrasyonunu otomatikleştirerek yalnızca verimliliği ve etkililiği artırmakla kalmayıp yeni çözümler ve araştırmalar için manivela gücü sağlayabilir.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my advisors Prof. Attila Gürsoy and Prof. Özlem Keskin for their guidance, patience and motivation. Their continuous support helped me in all the time of research and writing of this thesis. One simply could not wish for friendlier supervisors. I will always be grateful for being in their group.

I also want to thank to my committee members, Assoc. Prof. Engin Erzin, Assoc. Prof. Nurcan Tunçbağ and Assist. Prof. Tuğba Bağcı Önder for their valuable time and feedback.

I am proud to be a member of Koç University. I am grateful for the financial support and countless facilities that Koç University provided to me. I would also like to thank to the Scientific and Technological Research Council of Turkey (TUBITAK) for partially supporting this work (Project No: 113E164).

I could not image a better office environment, every result described in this thesis was accomplished with the help and support of fellow officemates. Thus, I would like to thank to former and current COSBI lab members: Emine Güven-Maiorov, Emel Şen Kılıç, Deniz Demircioğlu, Alper Başpınar, Nilay Karahan, Doğuş Doğru, Cemal Yamak, Serena Muratcıoğlu, Efe Elbeyli, A. Derya Cavga, Farideh Halakou, E. Sıla Özdemir, Mahan Rajaei Golsefidi, A. Tuğrul Balcı, and Ayşe S. Yazgılı for their help and all the fun we have had.

I am deeply grateful to my housemates Serena Muratcıoglu, Yasemin Vardar and Gizem Tabak for their endless support and all the amazing time we shared.

I would like to thank to all members of Solidarity Between Offices (OFAD): Bakış Doğru, Houman Bahmani Jalali, Rustamzhon Melikov, Nazlı Ataç, Anıl Fettahoğlu, Erkut Demirhan, Ömer Kırnap, Aysima Hacısüleyman and Özgün Can Önder for making

my days brighter in Koc University. I will always remember our best summer in 2016.

I am so thankful to have very good friends in my life. I would particularly like to thank Efe Elbeyli, Serena Muratcıoğlu, Bakış and Doğuş Doğru and Laviş (the cat) were being there whenever I need. I can easily say that I have yet another family from now on. I also would like to thank my lifelong friends from IFL and Bilkent for continuing to share all the great times wherever we are, and the new pals Maxim Core-club for welcoming to the new working environment at the end of my master journey.

Finally, I owe my deepest gratitude to my family. Without their love, emotional support, understanding and endless encouragement I would not be able to pursue this research work. It is a bit sad that such an enormous parenting capacity has been dedicated to just one child. I have always felt so lucky and proud that I had such wonderful and loving parents. Thank you for always being there and for making me who I am today.

TABLE OF CONTENT

ÖZET i

ACKNOWLEDGEMENTS i

Chapter 1 1

INTRODUCTION 1

Chapter 2 4

LITERATURE REVIEW 4

2.1 Protein-protein interaction networks 4

2.2 Genotype-phenotype relationship using networks 5

2.3 Phenotype-specific subnetworks with network prioritization algorithms 7

2.4 Network topology and PPI networks 8

2.5 Integrating structural information to PPI networks 11

2.6 Related Software 13

Chapter 3 17

Gene2Phen – A Web-based tool to build, visualize and compare phenotype specific subnetworks with their structure and mutation data 17

3.1 Main Functionalities of Gene2Phen Web Tool 17

3.2 Workflow of Methodology 19

3.3 Inputs 21

3.3.1 Seed Genes 21

3.3.2 Cutoff Score 22

3.4 Processing Steps 23

3.4.1 Ranking Human PPI for Phenotype 1 and 2 23

3.4.2 Filtering PPIs According to the Cut-off Value 26

3.4.3 Generating Phenotype Specific Subnetworks 27

3.4.4 Generating Interactive Network Visualization 28

3.5 Output Features 29

3.5.1 Topological Scores 29

3.5.2 Phenotype scores 30

3.5.3 Gene Annotations 30

3.5.4 Populating edges of PPI network with structural models 30
3.5.5 3D View of Complex Structures..31
3.5.6 Mapped Mutations on Interfaces 32
3.6 Software Architecture 34
3.7 Database 35
3.7.1 The human protein - protein interaction network 35
3.7.2 gene2pdb table 37
3.7.3 PRISM Results Table..38
3.7.4 Mutation Files 39
3.8 Implementation with Cytoscape.js library 39
Style and Layout 40
Chapter 4..42
RESULTS AND CASE STUDY 42
4.1 Gene2Phen Web Tool 42
4.2 Gene2Phen Web Tool Usage 42
4.2.1 Main Page 43
4.2.2 Selection and Submission Page 44
4.2.3 Results Page 46
4.2.3.1 Network Visualization Panel 46
4.2.3.2 Gene Annotation Panel 48
4.2.3.3 PDB-PDB Interaction Panel 49
4.3 Case Study 51
4.3.1 ELANE – VCAM1 & ELANE - CXCR4 53
4.3.2 ELANE – VCAM1 & ELANE – CSF3 56
Chapter 5..61
CONCLUSION...61
REFERENCES 62

List of Figures

Figure 2.1 Representation of nodes (proteins) and an edge (interaction) in a PPI network 5
Figure 2.2 A summary of recent computational methods for prioritization of candidate genes, and their input/output requirements [19]. 7
Figure 2.3 Schematic showing a bottleneck and the four categories of nodes in a network 10
Figure 2.4 Schematic representation of template-based prediction [38] 13
Figure 3.1 Workflow of the methodology followed in Gene2Phen.............................. 21
Figure 3.2 Basic I/O representation of GUILD.. 24
Figure 3.3 Filtration phase of human PPI network according to the phenotype scores. 26
Figure 3.4 JSON content for nodes 27
Figure 3.5 JSON format for edges 28
Figure 3.6 Illustration of populating edges of the subnetworks with structural data..... 31
Figure 3.7 A schematic diagram of mutation mapping... 33
Figure 3.8 The structure of Gene2Phen tool 34
Figure 4.1 Overview of Gene2Phen main page 43
Figure 4.2 Overview of Selection and Submission Page 45
Figure 4.3 Overview of Results Page... 46
Figure 4.4 Network representation at the visualization panel...................................... 47
Figure 4.5 An edge tooltip example from the Network Visualization Panel................ 47
Figure 4.6 A node tooltip example from the Network Visualization Panel.................. 48
Figure 4.7 An example from the Gene Annotation Panel 49
Figure 4.8 PRISM predictions for all PDB-PDB interactions of ELANE and CSF3 genes and the mutations data on predicted interfaces. 50
Figure 4.9 A sample JSmol view of PRISM prediction data of a PDB-PDB interaction. 50
Figure 4.10 Mutations on Interface Residues of 1CD9_A and 4NZL_A 51
Figure 4.11 Brain and lung metastasis subnetwork with 0.5 cutoff threshold............... 53
Figure 4.12: ELANE entry in KEGG Pathway database 54
Figure 4.13: KEGG pathways that ELANE is participated 55
Figure 4.14 ELANE-CXCR4, ELANE-VCAM1 and ELANE-CSF interactions 56
Figure 4.15 PRISM predictions for the protein products of ELANE and VCAM1....... 57
Figure 4.16 All interface mutations on PDB structures of ELANE interacts with VCAM1.. 58
Figure 4.17 PRISM predictions for the protein products of ELANE and CSF3 59
Figure 4.18 All interface mutations on PDB structures of ELANE interacts with CSF3 59

List of Tables

Table 2.1 Currently available software that is related to our work. 14
Table 3.1 BIANA integrated database release from March 2013 36
Table 3.2 Description of interactome table. 37
Table 3.3 Description of gene2pdb table 37
Table 3.4 All tables in prismDatabase 38
Table 3.5 Description of results table... 38
Table 4.1: Breast cancer lung and brain metastasis associated gene sets 51
Table 4.2 ELANE Mutations on ELANE-VCAM and ELANE-CSF3 interactions 60

Chapter 1

INTRODUCTION

Proteins have an important role in all phases of biological processes including metabolism, immunity, signaling etc. During these processes, organization of proteins constitutes complex forms which perform diverse functions acting as molecular machines inside the body [1]. To improve our comprehension of how gene functions and organizations are constituted at the level of an organism needs to investigate molecular interactions between all biological elements in cells.

Since protein interactions perform important tasks in life-sustaining functions, occurrence of an abnormal condition which causes a disorder in protein interactions results in disease formation. Therefore, protein interactions have charge of healthy and diseased states in organisms since protein-protein interactions are central for both function and control [2]. A disease is rarely a result of an abnormality on a single gene, it is commonly the result of complex interactions involving large sets of genes and their cooperation with other cellular components. This basis leads to the advancement of the network-based approaches to comprehend human disease. Utilization of protein interaction networks during the investigation into molecular mechanism of a disease, helps us to understand how a specific phenotypic profile occurs [3]. This insight can increase the prevention, diagnosis, and treatment opportunities.

Combining structural information to PPI networks is essential to have an understanding on the mechanism of interactions. This approach helps us to identify which interaction partners are competing with each other to bind the same region on a particular protein [4]. In addition diseases are commonly caused by mutations in these binding interfaces [5]. In this way, investigation of pathological variations as a subsidiary part of the human PPI network can help to find out the genotype to phenotype relationships behind human diseases.

Interpreting protein-protein interactions at both broadly network level and at individually structural level with mutation data requires a comprehensive study process that is fed from different sources for researchers. A tool that can automate this process would make their work easier. In this thesis, we developed a web-based tool by integrating large-scale protein-protein interaction network data, 3D protein structure information and mutation knowledge to aid researchers in exploring and comparing the molecular mechanism of different phenotypes. Our software works as an automatized pipeline tool to build, visualize and compare phenotype specific subnetworks, to predict and examine protein-protein interactions associated with their structure and mutation data. Gene2Phen (Genotype to Phenotype Sub-Networks) web-tool receives two sets of genes which acts as seed genes for prioritizing PPI network to generate two phenotype-specific subnetworks. For network prioritization the tool uses guilt-by-association methods of GUILD software [6]. After resulting of the query, users will find an interactive representation of two phenotype-specific subnetworks. Users can display details about proteins, interactions and mutations and have a look at the available structural data by clicking on the elements of the networks. Edges are colored based on the availability of structural information about the corresponding protein-protein interaction. To implement graph analysis and visualization functions on displayed subnetworks, we used Cytoscape js which is an open source JavaScript network library [7]. Then, these two subnetworks are enriched with structural information of 3D structural models of known protein-complexes and predicted its protein-protein interfaces. For predicting the interface structures of interacting couples, we used PRISM which is a template-based PPI prediction method [8]. When there are structures available for the selected PPI, they are shown in a JSmol view. Lastly, we mapped mutations on protein structures and on interaction interfaces. Our system obtains mutation knowledge from cBioPortal [9].

Although there are several tools for mapping and visualizing mutations on 3D protein structures in a network representation such as dSysMap [10], underlying reason to this work is the lack of tools which enable easily visualize, analyze and compare different phenotype-specific subnetworks on the same network representation. Our software peaks generic PPI network visualization with its enriched interactive graph representation, gene prioritization, graph analysis options, 3D structural knowledge and mutation mapping. This study presents a better alternative when we need to analyze the common protein-protein interfaces among different networks or investigate mutations happening on the same or different proteins of the networks.

The outline of the thesis is arranged as follows: Chapter 2 covers an extensive literature review on related work. This chapter starts with general information about protein-protein interaction networks and continues by explanation of genotype – phenotype relationship and PPI networks in human disease. Then it explains how topological properties and structural knowledge are used to make inference from biological networks. Finally, it provides a comprehensive summary about previous studies related to visualization, analysis and comparison of PPI networks. Chapter 3 contains materials and methods to develop Gene2Phen tool, describes main functionalities provided to the users, and explain design and implementation details. Chapter 4 starts with introduction of the tool as a result of the work done with case studies. As the final chapter, Chapter 5 concludes this thesis study with discussions of the tool and the future works.

Chapter 2

LITERATURE REVIEW

This chapter covers an extensive literature review on related work. We start with general information about protein-protein interaction networks and continue by explanation of genotype – phenotype relationship and PPI networks in human disease. Then we explain how topological properties and structural knowledge are used to make inference from PPI networks. Finally, we provide a comprehensive summary about previous studies related to visualization, analysis and comparison of PPI networks.

2.1 Protein-protein interaction networks

Proteins are biological macromolecules which are fundamental components for organisms and perform in almost every process within cells such as sensing the environment, regulating the activity of metabolic and signaling enzymes and DNA replication [11]. Proteins mainly vary from each other according to their amino acid sequences, which are defined by the nucleotide sequence of their genes. Then they usually end up some distinct three-dimensional structures which specify their activity.

Initial studies on how genes function at the molecular level examined and identified protein interactions individually, however today we know that this paradigm is not adequate for explaining biological processes within organisms and the subject is more complicated. Genes, proteins and other biological entities usually accomplish their functions within an intricate network of interactions, therefore, a single biomolecule has an impact on a broad range of other cell components. Accordingly, analyzing PPIs on the network basis become crucial to understand various biological functions and processes. A comprehensive, integrative approach which includes PPI networks has been established throughout the years [12].

Generic representation of PPI networks is node-edge diagrams. In a protein-protein interaction network, nodes represent proteins and edges represent physical interactions between proteins. Figure 2.1 shows the interaction between HRAS and RAF1 and illustrates the corresponding representation in a PPI network.

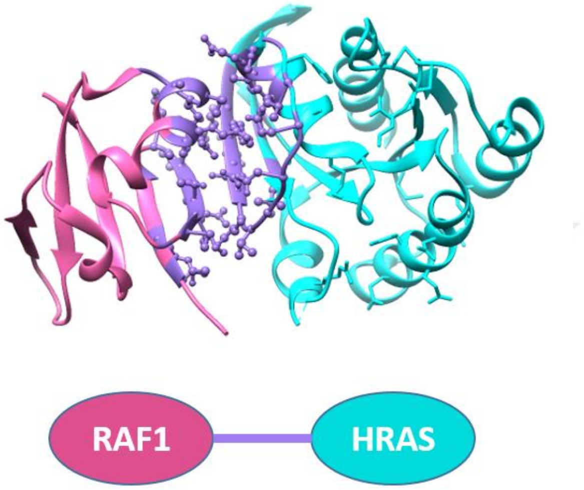


Figure 2.1 Representation of nodes (proteins) and an edge (interaction) in a PPI network

Protein-protein interaction networks have been utilized more and more to acknowledge the molecular basis of diseases. Current and promising PPI network applications on diseases can be categorized into four according to their objectives, which are, recognition of new disease genes, definition of their network features, identification of disease associated subnetworks and network-based disease classification [13]. This thesis concentrates on identifying disease-related subnetworks and recognition of new disease genes.

2.2 Genotype-phenotype relationship using networks

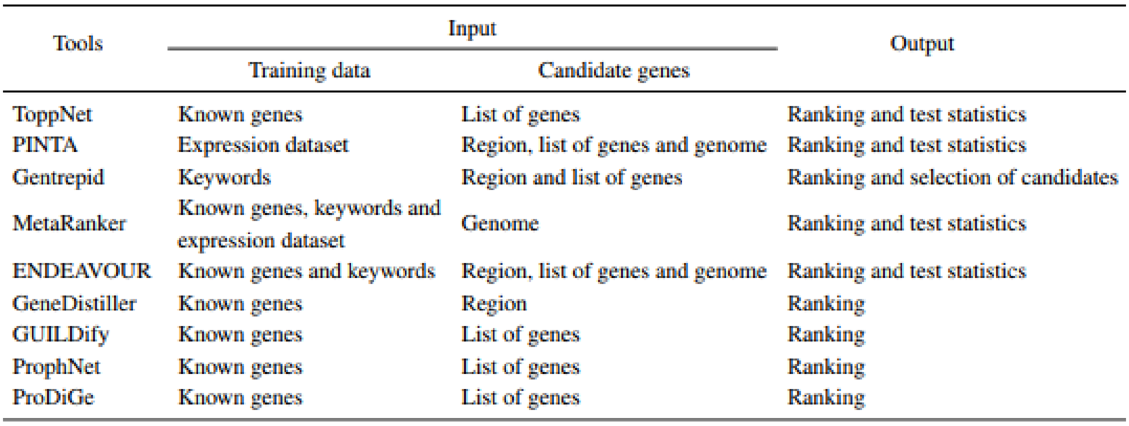
Before focusing on genotype-phenotype relationship, we need to define genotype and phenotype concepts individually. Genotype basically means the complete genetic characteristics of an organism which can be revealed by genome sequencing. However, this concept can also refer to a specific gene or gene sets related to a disease. On the other hand, phenotype implies our actual physical features which are generally visible. Similarly, phenotype concept can also refer to specific observations such as healthy or different diseased states of an organism. Thus, phenotype can be considered as the outcome of genotype with environmental factors. Sickle cell disease is a common way to explain genotype to phenotype relation. A red blood cell is normally in a round shape which makes it carry large amount of oxygen molecules. A mutation on the gene responsible for haemoglobin changes the shape of red blood cells into a crescent. That causes low levels of oxygen transport leading to sickle cell anaemia.

Genotype – phenotype relationships have been investigated with different approaches. “One gene – one enzyme – one function” concept was suggested by Beadle and Tatum, later Hartwell et al. introduced a new idea in 1999 [14]. Their theory suggests that rather than the individual properties of one protein, the whole complex generally determines the phenotype. Subsequent studies on large-scale protein-protein interactions largely confirmed Harwell’s modularity approach, while Tatum’s linear-relationship hypothesis is turning into an exception case for explaining genotype-phenotype relationships [15]. Generally, genotype–phenotype relationships are complex, this fact can be illustrated with examples of coronary heart disease and diabetes which are the most common polygenic human diseases.

Screening phenotypes throughout the genome alongside sequencing information obtained from patients generate high scaled, complex genotype and phenotype data [16]. New advances which consider the modularity of biological systems are needed to model and predict phenotypes, especially complex phenotypes from genotype information. The main goal is to provide predictive models of the disease mechanisms of complex diseases as well as define healthy states. The reason is that we hope accurate models will likely drive novel approaches to personalized medicine, where we better understand the genetic background of patients and can predict their proteome response to the environment in the context of their interacting gene products [17]. Still, many significant challenges remain to developing a network-based understanding of biology. In this thesis, two of these challenges we concentrate on is how the candidate network markers are identified and how the phenotype-specific subnetworks are selected from this candidate set.

2.3 Phenotype-specific subnetworks with network prioritization algorithms

The concept of gene prioritization become more and more popular with the states of both the progress made in computational biology and the large amount of genomic data publicly available. This concept was introduced in 2002 [18], and the first approach to solve this problem was described. Then, many different algorithm have been introduced (Figure 2.2) [19]. The principle was almost common among all strategies, which is called ‘guilt-by association’: the most favourable candidates will be the ones that are similar to the genes already known to be associated to desired biological interest.



Tools	Input		Output
	Training data	Candidate genes	
ToppNet	Known genes	List of genes	Ranking and test statistics
PINTA	Expression dataset	Region, list of genes and genome	Ranking and test statistics
Gentrepid	Keywords	Region and list of genes	Ranking and selection of candidates
MetaRanker	Known genes, keywords and expression dataset	Genome	Ranking and test statistics
ENDEAVOUR	Known genes and keywords	Region, list of genes and genome	Ranking and test statistics
GeneDistiller	Known genes	Region	Ranking
GUILDify	Known genes	List of genes	Ranking
ProphNet	Known genes	List of genes	Ranking
ProDiGe	Known genes	List of genes	Ranking

Figure 2.2 A summary of recent computational methods for prioritization of candidate genes, and their input/output requirements [19].

Functional annotations are generally the base of current disease candidate gene prioritization methods however they are limited. Therefore, when network-based candidate gene prioritization methods which are integrated with functional annotation performs better than all other gene features or annotations [20].

Phenotype-specific subnetwork prediction is another important aspect as much as prioritization of candidate gene. Similar to gene prioritization methods, subnetwork prediction has various computational methods, however, scoring the network is the common point of all strategies in order to obtain a candidate disease related subnetwork. Scoring a subnetwork could be performed according to to gene expression profiles, co-expression or set cover approach etc.

In the previous study of our work, Engin et al. created two different metastasis PPI networks for breast cancer, and involve protein structures to explain the genotype-phenotype relationships [21]. In that study, they generated PPI subnetworks by using initially known phenotype-related seed genes for both brain and lung metastasis, and they scored each interaction with following guilt-by-association principle [6]. After defining a cut-off as a threshold, they obtained brain metastasis subnetwork and lung metastasis subnetwork.

2.4 Network topology and PPI networks

The proteins and the protein-protein interactions create a network with several topological properties. Besides the knowledge on proteins and their interactions, knowledge of topological properties of the PPI networks can be used to create accurate models for predicting unknown protein-protein interactions and their biological significance on disease phenotypes [22]. PPI networks are need to be examined by their topological properties in order to distinguish the candidate disease genes because the genes which are related to same disease phenotype may be in the same functional pathways or probably have interactions with each other. Topological analysis of PPI networks also helps us to compare and characterize networks [23]. Several topological measures such as node degree, betweenness, closeness, clustering coefficient may contribute to the prediction of phenotype-gene association.

On earlier studies, topological centrality was considered as one of the key measures [24]. Topological centrality generally consists of node degree and betweenness centrality. Node degree of a gene indicates the number of interactions the gene has with its neighboring genes. Betweenness centrality of a gene corresponds to the number of shortest paths passing through that node. Hub nodes are selected as proteins with high node degree scores. Nodes with high betweenness centrality score are called bottleneck nodes representing the critical points of the network since they control most of the information flow in the network.

In earlier studies, Barabasi and Albert used growth and preferential connectivity principles in order to define the scale-free topology. According to these principles, scale-free networks grow as new nodes interact with the already-exist nodes in the network, and the new nodes favor to be attached the nodes with more connections. This mechanism creates the term hubs and presence of hubs is the key feature of scale-free networks [25]. Later, the evolutionary origin of scale-free topology is explained with biological evidences and its characteristic features such as topological robustness against accidental failures, relationship between being in the hub status and controlling viability or cell growth on complex biological networks are investigated [26].

At first, network studies focused highly on hub proteins. Among the studies investigating lethality and centrality in protein networks, one suggests that likelihood that removal of a protein will prove lethal correlates with the node degree of the protein. The showed in their results that, even though non-hub proteins constitute about 93% of the total number of proteins, only about 21% of them are essential. On the other hand, Hub proteins are covered only the 0.7% of all proteins, but single deletion of 62% or so of these proves lethal. This implies that hub proteins are three times more likely to be essential than non-hub proteins. [27] However, then it is proposed that the correlation between hubness and lethality is not depend on the structure of network. The reason is that hub proteins possess a vast number of interactions, which is why they are more likely to get involved in essential protein-protein interactions. These findings bring new aspects on the topological robustness [28].

Han et al. suggested that in protein-interaction network of yeast, there are two categories of protein hubs: (i) party hubs interacting with most of their partners concurrently, (ii) date hubs interacting with their partners at different times or in different locations. Even the party and date hubs have similar essentiality [29], their removal will effect on the network connectivity differently. The removal of date hubs falls the all network apart, while the removal of party hubs does not have an effect on the connectivity of global network.

Yu et al is one of the first studies emphasizing the importance of betweenness. In their study, they categorized all proteins as hub-bottlenecks, non-hub-bottlenecks, hub-non-bottlenecks and non-hub-bottlenecks. Figure 2.3 shows a schematic diagram of bottleneck and the four categories of nodes in a network. Yu et al stated that date hubs correspond mostly to hub-bottlenecks, while the party-hubs correspond mostly to hub–non-bottlenecks. From point of view of gene essentiality, they concluded that in regulatory networks, betweenness centrality is better determinant, while node degree is much more effective in PPI networks [30].

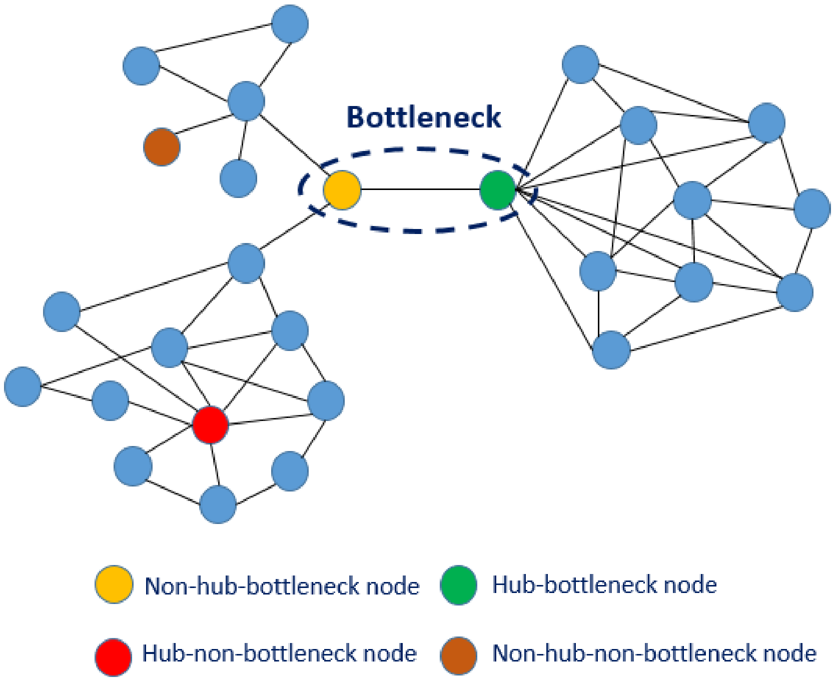


Figure 2.3 Schematic showing a bottleneck and the four categories of nodes in a network

2.5 Integrating structural information to PPI networks

A PPI network representation provides so much information about the sets of interacting proteins like whether two proteins are binding or do not, or the number of interactions a protein can have. Addition of structural knowledge to the PPI networks creates an extra dimension of data to the representation. Knowledge on structural details of two protein's interactions allows us to detect many other proteins trying to bind the same region on the surface of other proteins. On the other hand, this extra knowledge may help us to realize which interactions are unlikely to happen concurrently. Besides, there may be protein pairs using the similar interface architectures in order to interact with each other. The drug targeting on anyone of these PPIs will has a high probability of targeting the others as well, since the tendency of ligands to bind to similar binding sites. Moreover, determining the interface region of a PPI helps us to locate whether the mutations occur in the interface or not [21].

Previously known complex 3D structures are one of the best sources of protein interaction data in order to identify putative interactions of other proteins [31]. The mechanism of interactions has been understood better recently, after the incorporating 3D protein structure information to PPI networks. In 2006, Kim et al. integrated structural information with yeast PPINs [32]. They compared and structurally mapped edges by using sequence similarity to known complexes. They analyze the interfaces of each interaction. If a common partner protein uses the same interface while interacting with other the partners, these interactions are categorized as mutually exclusive. If the partners use different interfaces, those interactions are categorized as simultaneously possible.

However, there are limitations on known structural data and also there are limitations on the classical network node-and-edge representation which shows only which proteins interact; not how they interact [33]. Structural networks provide this information. Protein-protein interface structures can indicate which binding partners can interact simultaneously and which are competitive and can help forecasting potentially harmful drug side effects. From this standpoint, structural PPI networks enable to figure out the disease phenotypes. Kar et al.[34]'s analysis on protein interfaces of cancer related structural networks shows that the strength and specificity of the interactions of hub proteins/cancer proteins are different than the interactions of non-hub/non-cancer proteins, respectively.

Currently, large-scale cancer genome sequencing projects generates massive amounts of somatic mutation data, however, how to identify driver mutations and significantly mutated genes remains a great challenge [35]. By constructing a 3D structural human PPI network, Wang et al. systematically examine relationships between genes, mutations and associated disorders [36]. They find that in-frame mutations are mostly occur on the interfaces regions of binary interactions, and disease specificity is affected by different mutations located on different locations of an interface. They also predict 292 candidate genes for 694 unknown disease-to-gene associations. In another study, Engin et al. [21] mapped over 1.2 million non-synonymous somatic cancer mutations onto 4,896 experimentally determined protein structures and analyzed their spatial distribution. Then they use all available human protein complexes on PDB to construct a bipartite structural PPI network. Analysis of frequently mutated cancer genes within this network revealed that tumor-suppressors, but not oncogenes, are significantly enriched with functional mutations in homo-oligomerization regions. These studies show combination of point mutation data of 3D structures and PPI network help us to understand human disease mechanisms.

To integrate structural information to PPI networks we surely need to resolved 3D structure data. Although the experimental methods have provided thousands of interactions, this amount just corresponds quite small part of the whole genome size and also experimental methods are time consuming and expensive. Also, a large number of PPIs for non-modelled species are still uncertain. The demand for additional PPIs has led to the development of the computational prediction approaches of PPIs over the past decade. Some examples of these approaches are Prism, Haddock, Rosetta, and ZDOCK.

Template-based protein–protein interaction prediction tools are widely used computational approaches. These computational techniques also provide valuable insights for protein engineering and drug discovery. Hence, more efficient and less error prone computational techniques for protein–protein interaction prediction and structural modeling are of paramount importance in the biological sciences [8].

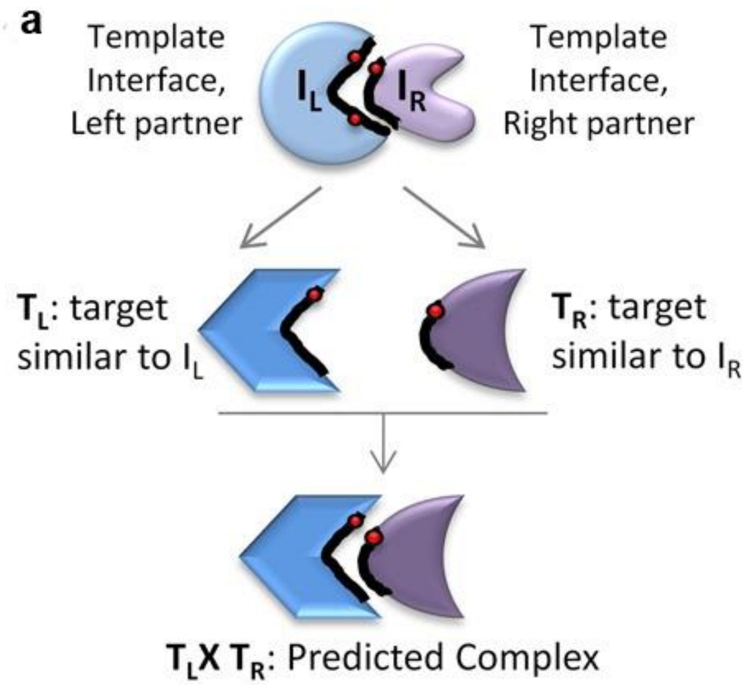

a
Template Interface, Left partner
I_L
I_R
Template Interface, Right partner
T_L: target similar to I_L
T_R: target similar to I_R
T_LX T_R: Predicted Complex

Figure 2.4 Schematic representation of template-based prediction [8]

2.6 Related Software

dSysMap [10] is a web-source used for mapping the disease-related missense mutations on the protein-protein interaction network which makes it one of the most important software related to our work. This genotype to phenotype relationship is used for developing deeper explanations for human diseases.

dSysMap has its own database for wide variety of human diseases. Users could visualize the constructed networks and build their own networks by giving list of proteins of interest. Users can navigate through the network, viewing the positions of disease-related mutations on the nodes and edges as well as on the high-resolution structures of proteins and complexes. However dSysMap is limited to experimentally known interactions and it does not have a network prioritization feature. Table 2.1 lists the other related software to this thesis with brief summaries.

Table 2.1 Currently available software that is related to our work.

Software Name	Description
Structure-PPi [37]	- Allow to analyze the effect of Single-Nucleotide Variants (SNVs) over functional sites, P-P interfaces, other annotated mutations on 3D structure of protein complex
The MI Bundle [38]	- Building disease networks - Identify variations in molecular complexes that may affect molecular interactions - Integrated analyses at the genomics, molecular, network and structural biology levels
PinSnps [39]	- Performs data analyses of PPI networks by using genetic and functional information mapped onto protein structures like genetic variants, SNPs, mutants' impact on the protein
ELASPIC [40]	- Evaluate the effect of mutations protein structures in the UniProt database - Allows predicted modeled wild-type and mutated structures, to be managed and viewed online and downloaded
G23D [41]	- Allows mapping of genomic positions/variants on 3D protein structures

	- Provides structural modeling of mutations - Allows analysis of intramolecular contacts of protein, functional predictions and predictions of thermo-stability changes
StructMAn [42]	- Analyzes the possible location of the amino acid residue corresponding to non-synonymous single nucleotide variants (nsSNVs) in the 3D protein structure relative to other proteins, nucleic acids and low molecular-weight ligands
MutaBind [43]	- Performs Mutation Mapping - Calculate change in binding affinity - No network visualization and analysis

Although having similar functionalities, the point that distinguishes Gene2Phen from the other related software is bringing all the following range of capabilities which are network prioritization, generating phenotype-specific subnetworks, interactive network visualization, displaying 3D structural models of PPIs and their predicted protein-protein interfaces, mutation mapping on interface structure and simultaneous phenotype-specific sub-network comparison together. Especially populating the PPI network with structural models and mutation mapping along with sub-network comparison constitutes the main contribution and novel part of this thesis work.

Gene2Phen tool provides various distinguishing biological uses, such as identification of new disease candidate genes or potential drug targets, or to predict novel disease associations. As an example, structural analysis on phenotype specific subnetworks may show that some PPI interfaces in two compared subnetworks have the same evolutionary origin which may be involved in similar molecular pathways that are shared by these two phenotypes. In addition, mapping the mutations on the interface regions of the proteins in two compared sub-networks may reveal that some mutations to be involved in the mechanisms differentiating between two phenotypes.

Another distinctive usage of Gene2Phen tool can be detection of proteins trying to bind the same region on the surface of other proteins, so that users can identify mutually exclusive interactions which are unlikely to happen concurrently. Besides, there may be protein pairs using the similar interface structure in order to interact with each other, this deduction helps drug targeting since these PPIs will has a high probability of targeting the others as well.

Chapter 3

Gene2Phen – A Web-based tool to build, visualize and compare phenotype specific subnetworks with their structure and mutation data

In this chapter, the materials and methods for design and implementation of Gene2Phen (Genotype to Phenotype) Web Tool are presented. The section starts with the preview and a brief commentary on main functionalities that user can find at Gene2Phen, and this is followed by step-by-step explanation of the inputs, processing steps and output features in the methodology workflow. This is followed by software architecture of the system, database management information, and web-based properties related to interactive network visualization.

3.1 Main Functionalities of Gene2Phen Web Tool

The main functions that users can take advantage of by using Gene2Phen tool can be listed and summarized as follows:

- *Users can prioritize the human protein-protein interaction network according to desired set of genes that user aim to investigate.* The tool ranks all interactions in the human PPI network according to the level of being associated with desired set of genes. In the generated scored-PPI network, higher scores imply higher associability to those genes and therefore higher probability of being related to investigated phenotype.
- *Users can generate a phenotype specific subnetwork from the scored-PPI network by selecting a cut-off score.* The tool eliminates the interactions which have a lower association score than the selected cut off score, thus generates a desired phenotype-specific subnetwork. Via cut-off score selection users can control the level of specificity since choosing higher threshold score causes narrowing the subnetwork down further.

- *Users can visualize generated phenotype-specific subnetworks as a generic network representation.* In this representation, nodes indicate proteins and edges indicate physical interactions between proteins. The displayed phenotype specific subnetwork provides a bird’s eye view of all interactions to the user and a general framework that the user can easily get into the details.
- *Users can enrich their research by examining the topological significance of proteins in the phenotype-specific subnetworks.* The tool displays the phenotype-specific subnetwork with numerical degree centrality, closeness centrality and betweenness centrality values for each protein when the user hovers the mouse on them. These centrality scores help users to identify whether the protein is a hub or bottleneck in the network or not.
- *Users can retrieve genome annotations for each protein in the phenotype-specific subnetwork.* To investigate protein-protein interactions associated with a phenotype or disease, researchers mostly need to access some biological information about the genes which are expressed and used in synthesis of functional gene products, generally proteins. When the user clicks on a gene the tool provides short and detailed description, aliases, chromosome location, UNIPROT ID and GENE ID for each gene in the network.
- *Users can go into the details on predicted possible interface structures of interacting protein-protein pairs in the phenotype-specific subnetworks.* The tool provides the template interface which used to predict the PDB-PDB interaction, along with energy scores and structure information of predicted interfaces. Users may need to identify competitive interaction partners which can bind the same interface region on a specific protein, so that they can detect which interactions can happen at the same time and which cannot. From another perspective, in the case of ligands which are prone to attach to similar binding region, a drug which targets any of these structurally similar protein-protein interactions will be likely to target the others as well.

- *Users can view 3D structural models of known protein-complexes and their predicted protein-protein interfaces*. For selected PDB-PDB interaction, user can visualize the complex structure, can identify the non-interface and interface residues of both two proteins, can download contacts of interface residues and the PDB file of the complex structure. The downloaded PDB file can be viewed with other visualization tools afterwards.
- *Users can see the list of mutations which are mapped on predicted protein-protein interfaces.* For each mutation on predicted interface structure, the tool provides information about mutation status (e.g. somatic), mutation type (e.g. missense mutation), amino acid change (e.g. P99S), functional impact score, chromosome number, reference allele and variant allele.
- *Users can use all features of the tool for analyzing and comparing two phenotypes at the same time.* Our tool is designed as all output features and results can be displayed in same frame in order to make easier to investigate two different phenotypes concurrently.

3.2 Workflow of Methodology

Our methodology stages have been tailored to the needs of researchers working on genotype to phenotype relationships. With the escalation in large scale methods to map functional connections between genes, many researchers are now examining data sets as networks. Main reason of this tendency is that genes usually perform their functions by interacting with other genes and in overall picture they construct biological networks which serve as molecular machines or dynamic biological pathways [26]. Due to the functional cooperation of genes and in PPI networks, researchers have shaped their studies towards a reverse approach that diseases or other phenotypic variations are derived from some perturbations of this molecular networks by genetic and environmental effects [44].

In this direction, interpreting protein-protein interactions from broadly network level to individually structural level with mutation data requires a comprehensive study process that is fed from different sources for researches on genotype-phenotype relationship. We have developed a web-based software works as an automatized pipeline tool to build, visualize, analyze and compare phenotype specific subnetworks, to predict protein-protein interactions associated with their structure and mutation data. The tool only demands sets of seed genes from users according to which phenotypes are investigated, then it automatizes the research methodology below (Figure 3.1) and brings into use.

First, the system ranks human protein – protein interaction network for both two different phenotypes according to seed gene sets coming from user’s selection. The ranking logic is basically based on being associated with specific seed genes which are known to be associated with a phenotype, makes this gene probable to be related to same phenotype. Then the scored human PPI network is filtered by a cut-off value and phenotype specific subnetworks are generated. After that, interactive visualization part is coming to play and topological scores for nodes, phenotype scores for edges and gene annotations enhances the generic network representation. Then, these two subnetworks are enriched with structural information of predicted protein-complexes and its protein-protein interfaces. If there are available structures of the predicted PPI, user can visualize it. Lastly, mutations are mapped on interaction interfaces of complex structures.

The subsequent sections follow the workflow of the methodology in Figure 3.1, and explains the input data which comes from user, the processing steps of Gene2Phen, and the output features provided in detail.

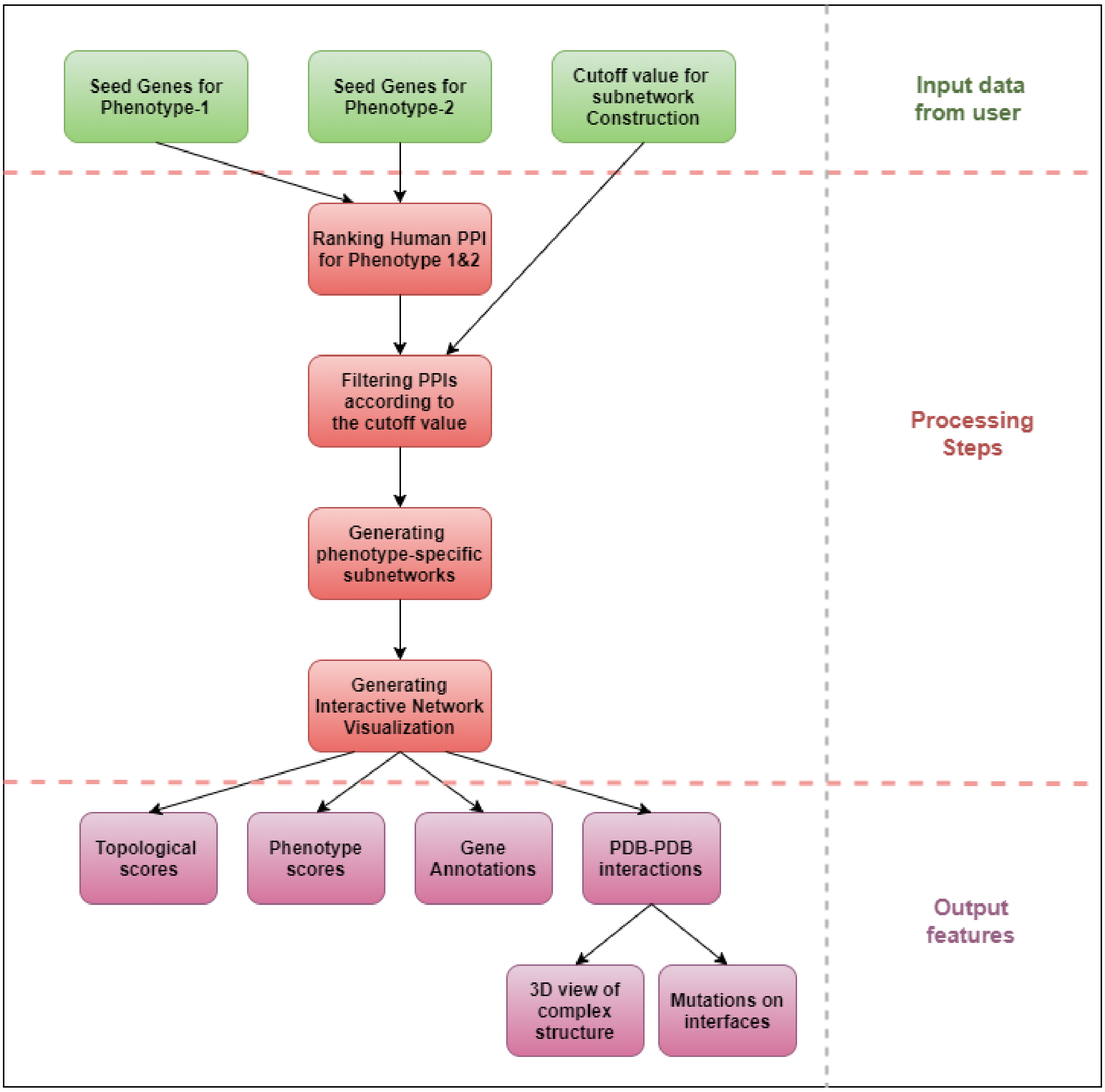

Seed Genes for Phenotype-1
Seed Genes for Phenotype-2
Cutoff value for subnetwork Construction
Input data from user
Ranking Human PPI for Phenotype 1&2
Filtering PPIs according to the cutoff value
Generating phenotype-specific subnetworks
Generating Interactive Network Visualization
Processing Steps
Topological scores
Phenotype scores
Gene Annotations
PDB-PDB interactions
3D view of complex structure
Mutations on interfaces
Output features

Figure 3.1 Workflow of the methodology followed in Gene2Phen

3.3 Inputs

3.3.1 Seed Genes

To detect disease gene candidates for investigated phenotypes, Gene2Phen takes advantage of guilt-by-association methods. The logic behind these algorithms can be illustrated as follows: Let us consider that a company needs to recruit new software engineers for their open positions and the headhunters of the company are seeking talented candidates who have expertise in software development. In case of a large number of applications, both narrowing down the candidate pool with prescreening interviews and the one-by-one interviews with all candidates would be time consuming and mostly inaccurate. To have a shorter hiring time and higher rate of acceptance, they may consider using an alternative recruitment method called employee referral which means to search potential candidates from their existing employees' networks. In this logic, starting from their employees as known to be good engineers and motivating them to refer a candidate would be an effective strategy.

In a similar manner, when researchers need to identify new candidate genes that responsible for investigated phenotype, guilt-by-association methods which ranks all genes based on their proximity to known disease genes (seeds) and narrows down the number of candidates therefore these methods provide user for focusing on the highly associated genes and increase the accuracy. In this project, Gene2Phen tool requires to get a set of seed genes from user for both two investigated phenotypes for network prioritization process. User can pick their seed genes by searching the literature on genes which are identified as linked to a phenotype in experimental studies such as RNA-sequencing, microarray analysis etc. The input format for seed genes is the official gene symbol approved by the Human Gene Nomenclature, which is a short-abbreviated version of the gene name [A3].

3.3.2 Cutoff Score

The human PPI network that is used in this project includes 12,834 genes and 339,811 gene-gene interactions. After the step of ranking human PPI is done, cutoff score parameter is needed for filtering PPIs of the whole network to make easier to study. The cutoff score can have any value between zero and one. If the user sets the cutoff score to the minimum (zero), since the tool would not be able to eliminate any interaction even if it is completely unrelated to the phenotype, the output would be the whole human PPI network to the user. On the contrary, if the user sets the cutoff score to the maximum (one), the tool eliminates all interactions except the few possible seed gene – seed gene interactions which are already known to be related to the phenotype by user, therefore the output network would be extremely specific that is not able to give any insight for new candidate genes. Between zero and one, users can try various values to generate phenotype specific subnetworks in different size and to control the level of specificity since choosing higher threshold score makes the subnetwork smaller.

3.4 Processing Steps

3.4.1 Ranking Human PPI for Phenotype 1 and 2

For ranking of human PPI network, we used a software named GUILD, a network-based disease candidate gene prioritization framework. GUILD (Genes Underlying Inheritance Linked Disorders) includes several algorithms of “guilt-by-association” to prioritize a list of candidate genes associated with a phenotype. Guilt-by-association approaches are based on a set of genes associated with a phenotype, named seeds, and the tendency that other genes associated with the same phenotype will interact with the seeds as previously explained in Chapter 3.3.1.

GUILD consists of implementations of 8 algorithms: NetScore, NetZcore, NetShort, NetCombo, fFlow, NetRank, NetWalk and NetProp [6]. We employed the NetCombo algorithm in GUILD using the default parameters as in to rank all the proteins of the human PPI network. This algorithm combines the algorithms of NetScore, NetZcore and NetShort.

NetScore adopts a message-passing scheme such that each node sends the information associated with it as a message to all its neighbors and the neighbors convey these messages to their neighbors. NetScore takes into consideration alternative shortest paths within the distance of at most number of iterations links at each so-called repetition-cycle. At the end of the repetition cycle, the node scores are updated according to messages received so far and the message passing is restarted.

NetZcore assigns a normalized score using the distribution of the scores of neighboring nodes. The normalization uses a random model of networks and it is calculated with the Z-score formula: z=(x-m)/s, where m is the average of scores of neighboring nodes with similar distribution in the random network and s is the standard deviation. The distribution is obtained with hundred network-replicates obtained by randomly shuffling the scores among nodes with similar degree.

NetShort accumulates the weighted shortest path lengths between a node and the rest of nodes in the network, where each edge-weight is inversely proportional to the average of the scores of the two nodes connected by the edge (i.e. edges connecting high scoring nodes are shorter).

NetCombo combines the output scores from NetScore, NetZcore and NetShort in a consensus scheme by averaging normalized scores (z-scores) of a node in these methods. It requires the output files of NetScore, NetZcore and NetShort.

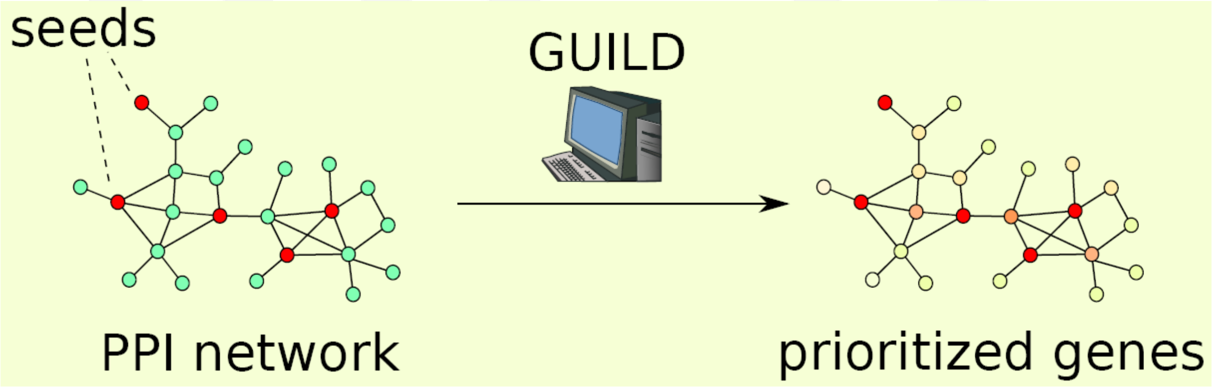


Figure 3.2 Basic I/O representation of GUILD

Since network prioritization process is performed under the guidance of two different set of seed genes for each phenotype, there will be calculated two phenotype scores for all edges in the human PPI network. For network prioritization process, Gene2Phen tool generates input node file and input edge file to make GUILD algorithms work on.

<u>Input node file:</u>

Input node scores file containing node (e.g. gene) identifier followed by its phenotypic relevance score (e.g. association with the disease phenotype for that protein/gene) on each line. The values need to be separated by whitespace(s). That is;

<node_id> <node_score>

At this stage, Gene2Phen only has two sets of seed genes about the phenotypic relevance of nodes in the network. Therefore, to prioritize them, it sets the node scores of seed genes to 1, while the rest of the nodes is set to 0.001.

<u>Input edge file:</u>

Input edge scores file containing node identifier followed by score of the edge its phenotypic relevance score (e.g. association with the disease phenotype for the proteins/genes it is connecting) and node identifier (the interaction partner) on each line. The values are separated by whitespace(s). Thus, a line in this file looks like;

<node_id> <edge_score> <node_id>

Before the network prioritization process, our human PPI network is not weighted, therefore the input edge scores file always has the same content, first node ID, second node ID and a symbolic value of the edge score between them (one) in order to register that they are interacted.

<u>Output node file:</u>

Output node scores file containing node identifier followed by its “calculated” phenotypic relevance score (e.g. association with the disease phenotype for that protein/gene) on each line. The values are separated by whitespace(s). The format of a line would be;

<node_id> <node_score>

Gene2Phen uses the output node file to generate output edge file. Two output node files are generated for two phenotypes.

Output edge file:

Output edge scores file containing two node identifiers followed by their "calculated" phenotypic relevance scores (e.g. association with the disease phenotype for that protein/gene) for two phenotypes on each line. The values are separated by whitespace(s). The format of a line would be;

<node_id1> <node_id2> <edge_score1><edge_score2>

As an input for generating output edge scores file, Gene2Phen use the input edge file and output node files of two phenotypes. For each interaction on each line in the input edge file, it retrieves the calculated phenotype-1 scores of interacting nodes, takes the average and write to the <edge_score1>, then repeats the same process for phenotype-2.

3.4.2 Filtering PPIs According to the Cut-off Value

Once Gene2Phen completes the stage of ranking human PPI network, each interaction in the output edge-scores file passes through the filtering phase. If an interaction does not have a phenotype score above the threshold for neither phenotype-1 nor phenotype-2, it is filtered by the tool. This phase is the key for generating phenotype-specific subnetworks.

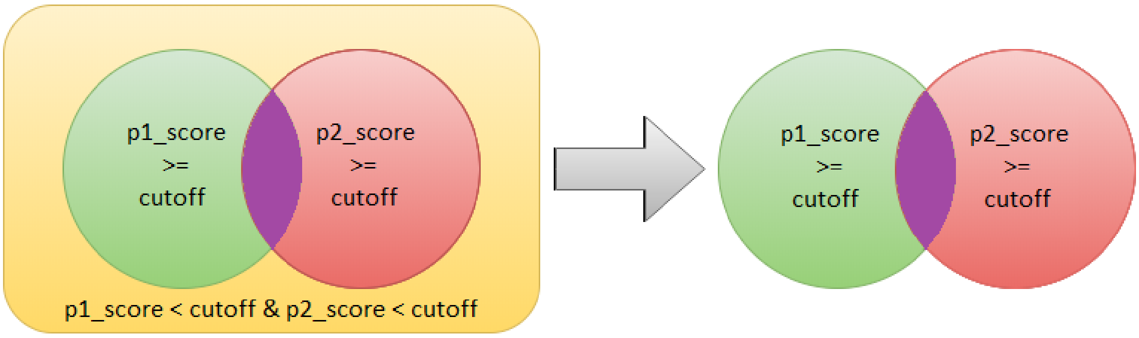
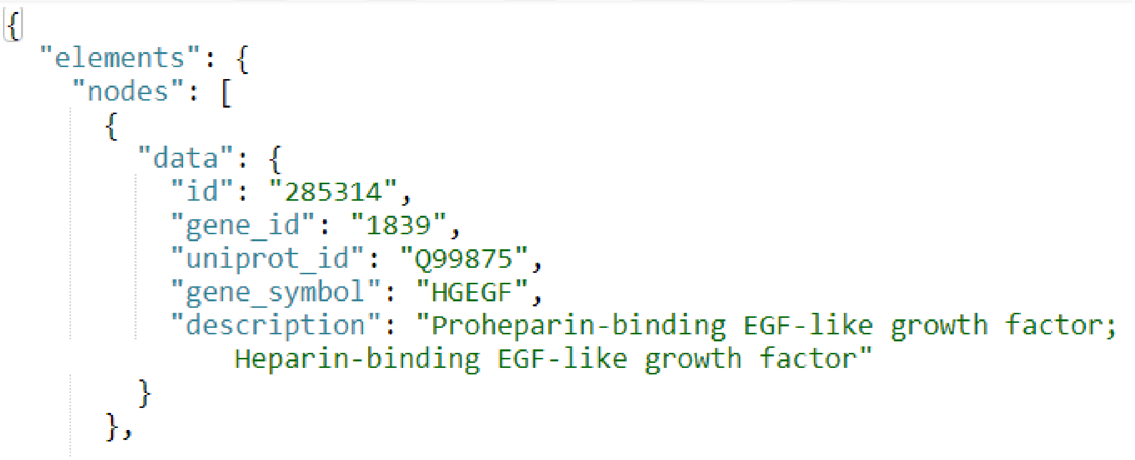


Figure 3.3 Filtration phase of human PPI network according to the phenotype scores.

3.4.3 Generating Phenotype Specific Subnetworks

At this intermediary step, output edge scores file - which only includes two gene IDs and two phenotype scores for each interaction - is evolved to a file contains a collection of two types of objects (i.e. nodes and edges) with comprehensive data coming from the database tables, in JSON (JavaScript Object Notation) format. JSON is a lightweight data-interchange format. It is easy for people to read and write. It is also easy for machines to parse and generate.

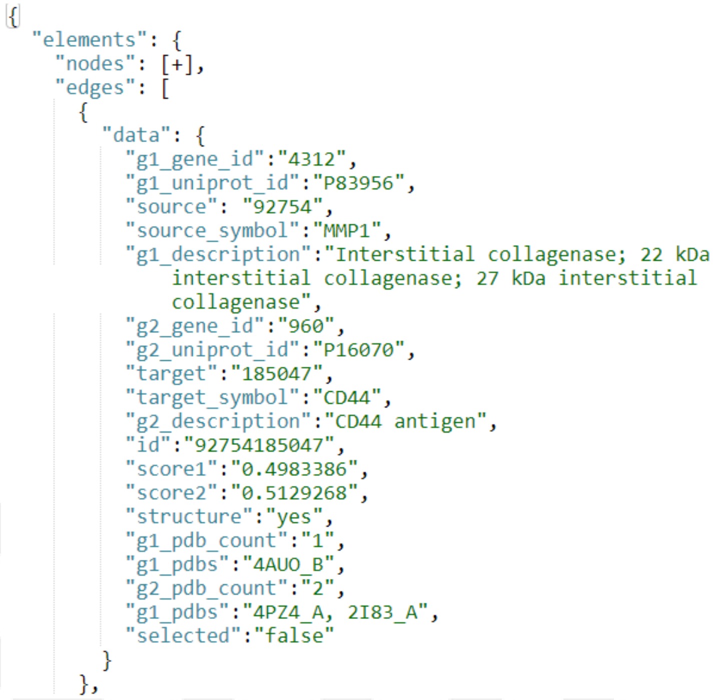
Even the BIANA ID is the key value for each interaction, users will need to see other details and identifiers for the elements in phenotype specific subnetworks, also for proceeding the next steps which generates the output features of the tool will use these data. Therefore, Entrez Gene ID, UniProt ID, HUGO gene symbol and a short description is added to the data object of each node (Figure 3.4).



```
{
  "elements": {
    "nodes": [
      {
        "data": {
          "id": "285314",
          "gene_id": "1839",
          "uniprot_id": "Q99875",
          "gene_symbol": "HGEGF",
          "description": "Proheparin-binding EGF-like growth factor;
              Heparin-binding EGF-like growth factor"
        }
      },
```

Figure 3.4 JSON content for nodes

Alongside the data pairs explained above for each node, phenotype-1 and phenotype-2 scores, an interaction ID, availability status of structure for the interaction, PDB counts and PDB sets for Gene-1 and Gene-2 is added to the edge elements.



```
{
  "elements": {
    "nodes": [+],
    "edges": [
      {
        "data": {
          "g1_gene_id":"4312",
          "g1_uniprot_id":"P83956",
          "source": "92754",
          "source_symbol":"MMP1",
          "g1_description":"Interstitial collagenase; 22 kDa
              interstitial collagenase; 27 kDa interstitial
              collagenase",
          "g2_gene_id":"960",
          "g2_uniprot_id":"P16070",
          "target":"185047",
          "target_symbol":"CD44",
          "g2_description":"CD44 antigen",
          "id":"92754185047",
          "score1":"0.4983386",
          "score2":"0.5129268",
          "structure":"yes",
          "g1_pdb_count":"1",
          "g1_pdbs":"4AUO_B",
          "g2_pdb_count":"2",
          "g1_pdbs":"4PZ4_A, 2I83_A",
          "selected":"false"
        }
      },
```

Figure 3.5 JSON format for edges

Generated JSON file of phenotype specific subnetworks will be used as an input for generating interactive network visualization, and the output features which will be explained following sections.

3.4.4 Generating Interactive Network Visualization

At the interactive network visualization step, Cytoscape.js, an open source network library called is highly utilized. Using this library allows to easily display and manipulate enhanced, interactive network representation for phenotype specific subnetworks.

The main feature of the library is to provide an opportunity to display and manage graphs interactively. In Gene2Phen, when user clicks on a node, annotation panel displays the details about selected gene. Clicking on an edge brings a table of predicted PDB-PDB interactions for selected edge. Mouseover events of nodes and makes the tooltips pop up, node tooltips display topological scores of the selected node while edge tooltips display phenotype-1 and phenotype-2 scores. The library also allows selection, pinch-to-zoom, and panning for both touch and non-touch operated devices. Cytoscape.js depends on event-driven model with a core API [9].

During initialization phase, Cytoscape.js receives the graph elements and their style properties, layout options, and any other graph properties as JSON objects which is generated in the previous step. Style properties are compatible with CSS standards. Gene2Phen tool also exploits several layout graphs. Users can choose a suitable layout algorithm according to the context and size of their graph.

3.5 Output Features

3.5.1 Topological Scores

If users want to see topological scores of a gene, they can hover with the mouse over the node and a tooltip will pop out. Node tooltips display the degree centrality, normalized degree centrality, closeness centrality, normalized closeness centrality and betweenness centrality. To generate the tooltips that displays topological scores, Gene2Phen tool utilizes the built-in functions of Cytoscape.js library for each algorithm.

- *Degree centrality* is defined as the number of links occurrence upon a node (the number of edges that a node has). Nodes that have higher degree centrality are considered as more important since they are hubs in a network, they can take charge in several pathways, can spread information to many nodes or can prevent the network from breaking up.
- *Normalized degree centrality* is normalization of degree centrality score of a node.
- *Closeness centrality* is defined as the average proximity to other nodes. Nodes which have higher proximity average means that those nodes are closer to the other nodes which are assumed to be more important.
- *Normalized closeness centrality* is normalization of closeness centrality score.
- *Betweenness centrality* is defined as fraction of shortest paths that pass through the node. Higher betweenness centrality means that node tends to connect other nodes which makes them critical (bottleneck) in the networks.

3.5.2 Phenotype scores

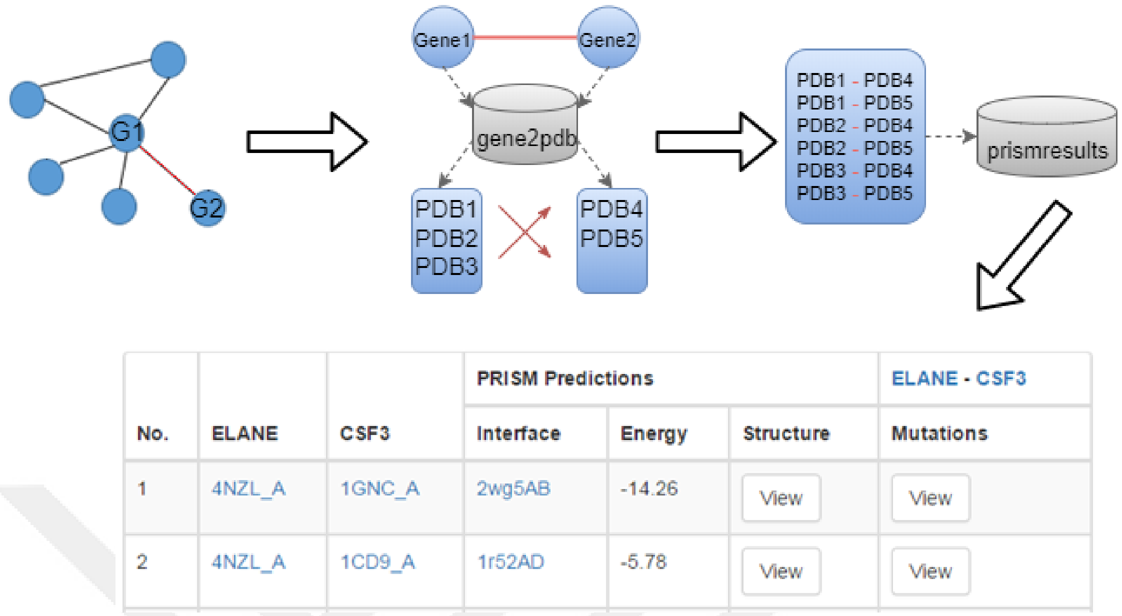
At the network prioritization step, the algorithms of GUILD software score the genes, not interactions between genes. Therefore, Gene2Phen tool calculates the edge phenotype scores by getting the average of phenotype scores of interacting genes calculated by GUILD. When a user hovers with the mouse over an edge, phenotype-1 and phenotype-2 scores are displayed. Phenotype scores get a value between zero and one, and higher values indicates higher associability to that phenotype.

3.5.3 Gene Annotations

For all nodes in the phenotype specific subnetworks, user can find annotation data which includes description, aliases, short description, chromosome location, UniProt ID and Gene ID of selected gene. Annotation data is obtained from BioGene. BioGene is a simple web service where scientists can query a gene and retrieve information about its functions and references. It primarily uses Entrez Gene, a gene database provided by NCBI [45]. Query results of the tool are exhaustive; therefore, the tool extract only some fundamental elements from annotation data for ease of use.

3.5.4 Populating edges of PPI network with structural models

Up to this stage, Gene2Phen provides a common representation style for phenotype specific subnetworks which is a graph demonstration. In these graphs, nodes represent the genes and edges represent gene interactions. This concept provides a global picture of biological processes and protein function. However, without having molecular and structural-level data about which proteins can interact with, or which interactions can occur at the same time, or which residues would be on the interface; it is not possible to deeply understand functional roles and binding mechanisms of proteins in phenotypic variation. Therefore, Gene2Phen adds new perspective to generic network representation by mapping all structurally-known protein products of a gene to each node, and all predicted protein-protein interactions of two genes to each edge.



			PRISM Predictions			ELANE - CSF3
No.	ELANE	CSF3	Interface	Energy	Structure	Mutations
1	4NZL_A	1GNC_A	2wg5AB	-14.26	View	View
2	4NZL_A	1CD9_A	1r52AD	-5.78	View	View

Figure 3.6 Illustration of populating edges of the subnetworks with structural data

Gene2Phen obtains all data on structural prediction of protein-protein interactions from its database which consist of PRISM results. PRISM (PRotein Interactions by Structural Matching) is a template-based protein interaction prediction tool and needs the 3D structures of queried genes stored in PDB database. Since it takes PDB IDs as an input, our tool firstly fetches the PDB sets of two interacting genes from gene2pdb table in database, then lists all possible combinations for PDB-PDB interactions for the selected edge. Then it connects to the PRISM results table and fetches the results for each PDB pairs (Figure 3.6). Prediction results of PRISM consists of 3 items which are interface, binding energy score and 3D structure data. Binding energy scores of predicted complexes are calculated by using Fiberdock [46].

3.5.5 3D View of Complex Structures

For each predicted PDB-PDB interaction, Gene2Phen demonstrates the 3D view of complex structure via JSMol molecule viewer applet. JSMol JMolApplet is a web browser JavaScript application that can be integrated easily into web pages as Gene2Phen

[47]. Besides viewing the structure, users also can download the complex structure in as a pdb file, and contacts of interface residues as a txt file.

3.5.6 Mapped Mutations on Interfaces

For each predicted PDB-PDB interaction, Gene2Phen maps mutations on interface region and displays annotation data of those mutations in a list view. A schematic diagram of mutation mapping steps can be seen in Figure 3.7.

Gene2Phen obtains the mutation data from CBioPortal, is a web application that allows users to explore, visualize, and analyze cancer genomics data [9]. CBioPortal Web API provides direct programmatic access to all genomic data stored within the server. Their web service is REST-based, this means that users can make a query by concatenating input entries to the base URL and receive the output data as a TXT or an XML response. At the first step of mutation mapping process, Gene2Phen requests the full set of annotated extended mutation data for Gene1 and Gene2. This output data also gives the residue indexes of mutations on the gene, however, a residue-level mapping of UniProtKB entries to PDB entries is required to be able to identify whether those mutations are on one of two interacted PDBs or not.

While many links are provided to Protein Data Bank (PDB) files, getting a regularly updated mapping between UniProtKB entries and PDB entries at the chain or residue level is not straightforward. At the second step, we have utilized PDBSWS, a automatically maintained database which performs a residue-level mapping by aligning the sequences from PDB and UniProtKB [48]. PDBSWS is also REST-based, when Gene2Phen tool makes queries for interacting PDBs, PDBSWS returns residue-level mapping of given PDB and related genes. From the first step of mutation mapping, the tool has the residues of all mutations on the interacting genes, by combining those two outputs gives all mutations on predicted complex. However, at this stage we still have no clue about the mutations on interface residues since the tool does not have the interface residues of given PDB-PDB interactions yet.

At the last step, Gene2Phen makes another RESTful query to obtain the interface residues of selected PDB-PDB interaction. After having the interface residues and combining with the previously obtained all mutation data on predicted complex, by checking whether mutations are on the interface or not, Gene2Phen accomplishes the mutation mapping on interface region of selected PDB-PDB interaction.

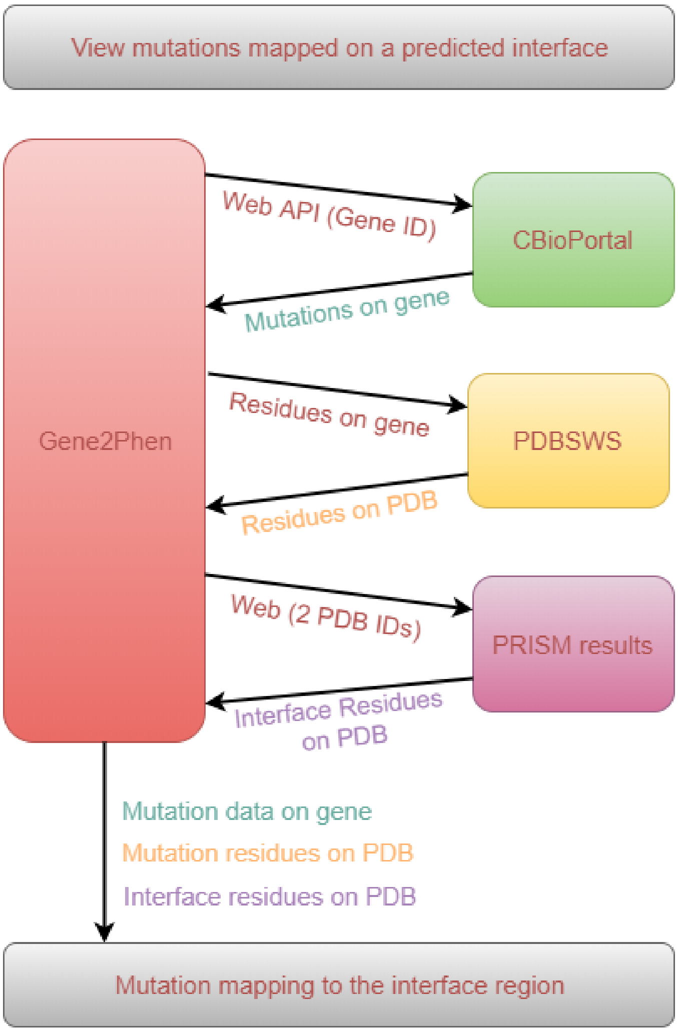


Figure 3.7 A schematic diagram of mutation mapping

3.6 Software Architecture

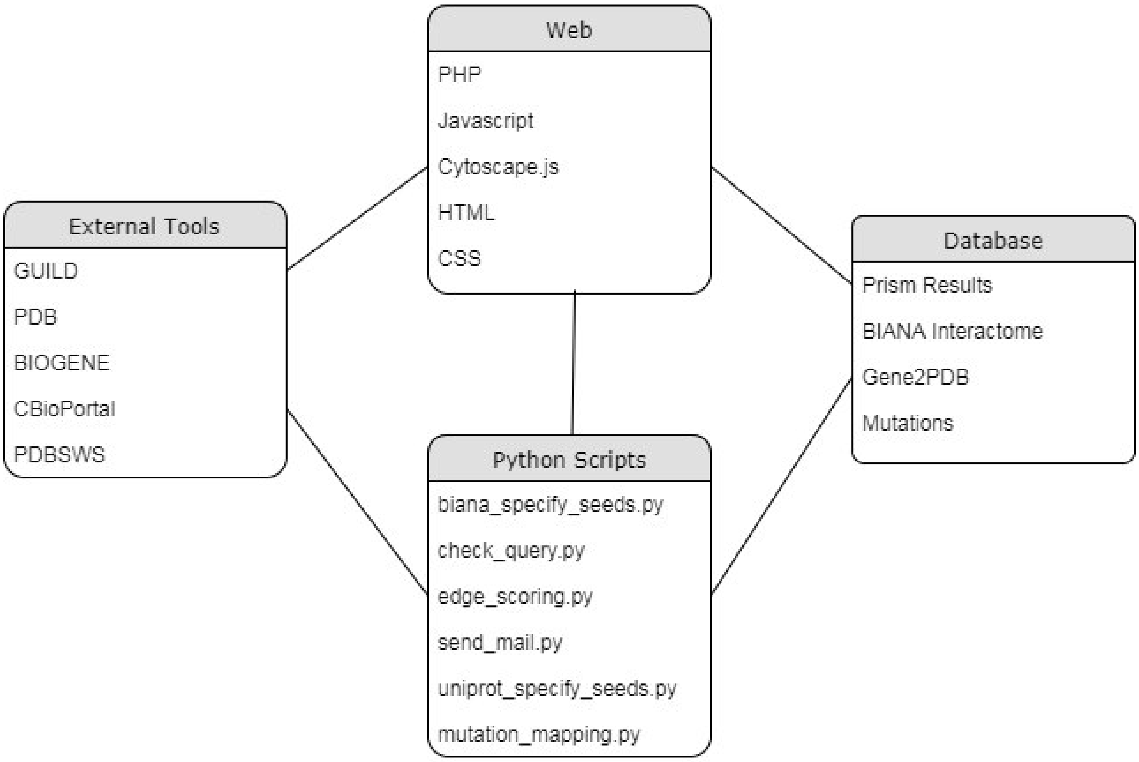


Figure 3.8 The structure of Gene2Phen tool

In this project PHP is employed as server-side scripting language. It is currently the most popular server-side scripting language. PHP supports all major operating systems and wide range of databases including MySQL.

In the database part of Gene2Phen web tool MySQL is utilized. MySQL is the most popular relational database management system. MySQL can provide concurrent multi user access to its databases. SQL statements to select, insert and update the fields of tables are embedded into PHP codes.

JavaScript is one of the three main languages for web development part of the project along with HTML and CSS. HTML is used to define the content of the pages of Gene2Phen, CSS is used to specify the appearance of visualization and layout details, while JavaScript programs the behaviour of the output features of the tool. Especially Cytoscape.js, an open network library written in JavaScript for analysis and visualization is highly utilized in generating interactive network visualization step. The implementation details will be explained on the following pages.

External sources of the Gene2Phen tool are mostly have Web APIs, excluding GUILD software which its standalone version is used. Yet these sources are mostly employed at the generation steps of output features, excluding GUILD which is employed at the network prioritization phase.

Python scripts are more readable and easy to understand in comparison with PHP, therefore processing steps are implemented in Python. These scripts are employed by the client side and connect to database tables or external sources when needed.

Database tables of Gene2Phen tool are explained in detail at the following sections.

3.7 Database

3.7.1 The human protein - protein interaction network

There are numerous databases consisting of experimental data on protein interactions. To work with a comprehensive protein-protein interaction network, all data needs to be queried accurately and overlapping data should be combined. Because when such an overlap occurs, it affects reliability since the methods to get data may be different, the source organism may be different, or same data may be organized in different ways. To avoid these possible issues, we used a human PPI network generated via BIANA (Biological Integration and Network Analysis) bioinformatics tool. BIANA brings numerous databases together and handles integration of data which has different identifiers. When two proteins have the same sequence data, the same UNIPROT accession number or the same Entrez Gene ID, BIANA assumes that those two proteins are the same.

Our tool uses the same database version with the current version of GUILDIFY (web version of GUILD) uses, which is BIANA integrated database release from March 2013 (includes UniProt, GO, OMIM, Drugbank, HPRD, IntAct, DIP, BioGrid, MPACT and Reactome databases) (Table 3.1).

Table 3.1 BIANA integrated database release from March 2013

Database Name	Database Version
SWISSPROT	Mar 2013
TREMBL	Mar 2013
OMIM	Jul 2013
GO	Jun 2013
DRUGBANK	Jul 2013
BIOGRID	Jan 2013
DIP	Jan 2013
HPRD	Apr 2010
INTACT	Jan 2013
MINT	Dec 2011
MPACT	Oct 2008
REACTOME	Dec 2012

The human PPI network taken from BIANA integrated database release is located to interactome table in our database. The detailed description of interactome table is given at Table 3.2. The meaning of each column is explained below.

Field: The name of the corresponding column in the database table.

Type: The type of individual values the column can have. Numbers in parenthesis indicate the size allocated in bytes for each value.

Key: Shows whether the column act as part of a primary key (PK) for that table.

Default: The default value assigned to that column if one is not set explicitly.

Extra: Miscellaneous information.

Table 3.2 Description of interactome table.

Field	**Type**	**Null**	**Key**	**Default**	**Extra**
biana_id	varchar(10)	NO	PK	NULL	
gene_symbol	varchar(20)	NO		NULL	
gene_id	varchar(10)	NO		NULL	
uniprot_id	varchar(20)	NO		NULL	
description	varchar(200)	NO		NULL	

Interactome table has 5 columns which are biana_id, gene_symbol, gene_id, uniport_id, description. Biana_id is the primary key of the interactome table. Since BIANA ID numbers are specific to GUILD software, we added gene_id, uniport_id and gene_symbol to the table so that users can distinguish genes with global identifiers. This table is employed at generating phenotype specific subnetworks (Chapter 3.4.3) step when the filtered output edge-score network is enhanced for generating json files of the subnetworks.

3.7.2 gene2pdb table

Gene2pdb table has three columns which are gene_id, pdb_count, and pdbs. Gene ID is the primary key of gene2pdb table. Pdb_count stores the number of protein structures available in database for each gene and pdbs column stores the PDB IDs of those structures' as a concatenated string.

Table 3.3 Description of gene2pdb table

Field	**Type**	**Null**	**Key**	**Default**	**Extra**
gene_id	varchar(10)	NO	PK	NULL	
pdb_count	int(11)	NO		NULL	
pdbs	varchar(2000)	NO		NULL	

Gene2pdb table is employed at populating PPI edges with structural models. When an interaction is selected, the tool retrieves the PDB counts of interacting genes, and takes the Cartesian products of two sets of PDBs to predict possible PDB-PDB interactions.

3.7.3 PRISM Results Table

PRISM database keeps all user information data available, target structures, template structures, prediction results Multiprot results etc. PRISM database consists of 9 tables (Table 3.4). Gene2Phen tool only utilizes the results table.

Table 3.4 All tables in prismDatabase

Tables_in_prismDatabase
ip_addr
Job
Jobs
Multiprot
Passed
Results
Targets
Templates
templatesN

Results table (Table 3.5) contains final prediction results from PRISM runs. It includes target protein information, which interface is used as a template, energy value of the interaction, location of the generated complex protein and date value that indicates prediction time of the job. Gene2Phen utilizes the results table at “populating edges of PPI network with structural models” phase (Chapter 3.5.4).

Table 3.5 Description of results table

Field	**Type**	**Null**	**Key**	**Default**	**Extra**
target1	varchar(30)	NO	PK	NULL	
target2	varchar(30)	NO	PK	NULL	
interface	varchar(30)	NO	PK	NULL	
energy	double	NO		NULL	
structure	varchar(200)	NO	PK	NULL	
date_column	timestamp	NO		CURRENT_TIMESTAMP	

When an interaction is selected, after the tool takes the Cartesian products of two sets of PDBs to predict possible PDB-PDB interactions, Gene2Phen queries the result table in condition that target1 and target2 is equal to each interaction partners in the list of possible PDB-PDB interactions. Result queries return interface, energy and structure columns.

3.7.4 Mutation Files

Mutation files for each gene is retrieved from CBioPortal through the Web API and downloaded. Gene2Phen tool requests the full set of annotated extended mutation data and downloads in tsv format it is not previously downloaded. The tab-delimited file contains following columns: Entrez gene ID, HUGO gene symbol, case ID, sequencing center responsible for identifying the mutation, somatic or germline mutation status, mutation type (such as nonsense, missense or frameshift), validation status, amino acid change, predicted functional impact score by Mutation Assesor, links to the various views from Mutation Assessor, chromosome where the mutation occurs, start and end position of mutation and mutation profile id. Gene2Phen tool utilizes to the downloaded mutation files at the mutation mapping stage on predicted interfaces. Since these files contains all mutations are observed on a gene, the tool eliminates those are not on the contact region of interacting PDBs. Then it displays interface mutations with the data coming from selected columns of the mutation files.

3.8 Implementation with Cytoscape.js library

The architecture of Cytoscape.js consists of two main parts, the graph instance named “core” and the collection. The core is the main entry point into the library. From the core, layouts can be run, viewport can be altered, and other operations are performed on the graph as a whole. The core provides several functions to access elements in the graph. Each of these functions returns a collection, a set of elements in the graph. Functions are available on collections that allow us to get data about elements in the collection, filter the collection, perform operations on the collection, and traverse the graph about the collection, and so on.

Gestures

The visualization panel supports several gestures: Grab and drag background to pan, grab and drag nodes, mouse wheel to zoom, tap to select, tap background to unselect, tap hold background to unselect, etc.

Style and Layout

For visualization, the container, elements, style, and layout options usually should be set.

Style: Style is used for modifying the visual details in the network, such as differentiating edges which are more associated to phenotype-1, phenotype-2 and the edges which do not have structural data by using eles.addClass(), eles.removeClass(), etc..

Layout: In graph visualization, layout is used for algorithmically positioning the nodes in the graph. Which layout is initially run is specified by the name field. In the visualization panel, Springy, which is a force directed layout algorithm in JS is used. Springy simulates the real world physics to make the network looks attractive with realistic motion. In visualization panel we prefer to limit the running time of Springy layout with 4000 ms in order to keep the balance between attractive design and usable design since Springy layout does not allow to zoom in/out.

In case of visualizing very large networks, springy layout cannot working effectively since the calculations required for simulation becomes costly. In such cases our tool uses plainer layout named “cose”.

3.8.3 Events

Mouse over node event: Displays a tooltip which returns topological scores of the node by using eles.degreeCentrality(),eles.betweennessCentrality(),eles.closenessCentrality(), eles.degreeCentralityNormalized(), eles.closenessCentralityNormalized(),

Mouse over edge event: Displays a tooltip which returns phenotype scores of the edge by using e.data.score1 and e.data.score2 from json file of the generated phenotype specific subnetworks.

Click node event: Displays Gene Annotation table of selected gene.

Click edge event: Displays the PDB-PDB Interaction table of selected interaction.

Checkbox filter event: Hides the edges which have no structure by using cy.getElementById(i).addClass('filtered')

Chapter 4

RESULTS AND CASE STUDY

4.1 Gene2Phen Web Tool

The Gene2Phen web tool serves as an automatized pipeline to build, visualize and compare phenotype specific subnetworks, to predict and examine protein-protein interactions associated with their structure and mutation data.

Gene2Phen web-tool receives two sets of genes which acts as seed genes for prioritizing PPI network to generate two phenotype-specific subnetworks. For network prioritization the tool uses guilt-by-association methods of GUILD software. [6] After resulting of the query, users will find an interactive representation of two phenotype-specific subnetworks. Users can display details about proteins, interactions and mutations and have a look at the available structural data by clicking on the elements of the networks. Edges are colored based on the availability of structural information about the corresponding protein-protein interaction. To implement graph analysis and visualization functions on displayed subnetworks, we used Cytoscape js which is an open source network library written in JS. [7] Then, these two subnetworks are enriched with structural information of 3D structural models of known protein-complexes and predicted its protein-protein interfaces. For predicting the interface structures of interacting couples, we used PRISM which is a template-based PPI prediction method. [8] When there are structures available for the selected PPI, they are shown in a JSmol view. Lastly, we mapped mutations on protein structures and on interaction interfaces. Our system obtains mutation knowledge from cBioPortal. [9]

4.2 Gene2Phen Web Tool Usage

Users can access the Gene2Phen features through three pages:

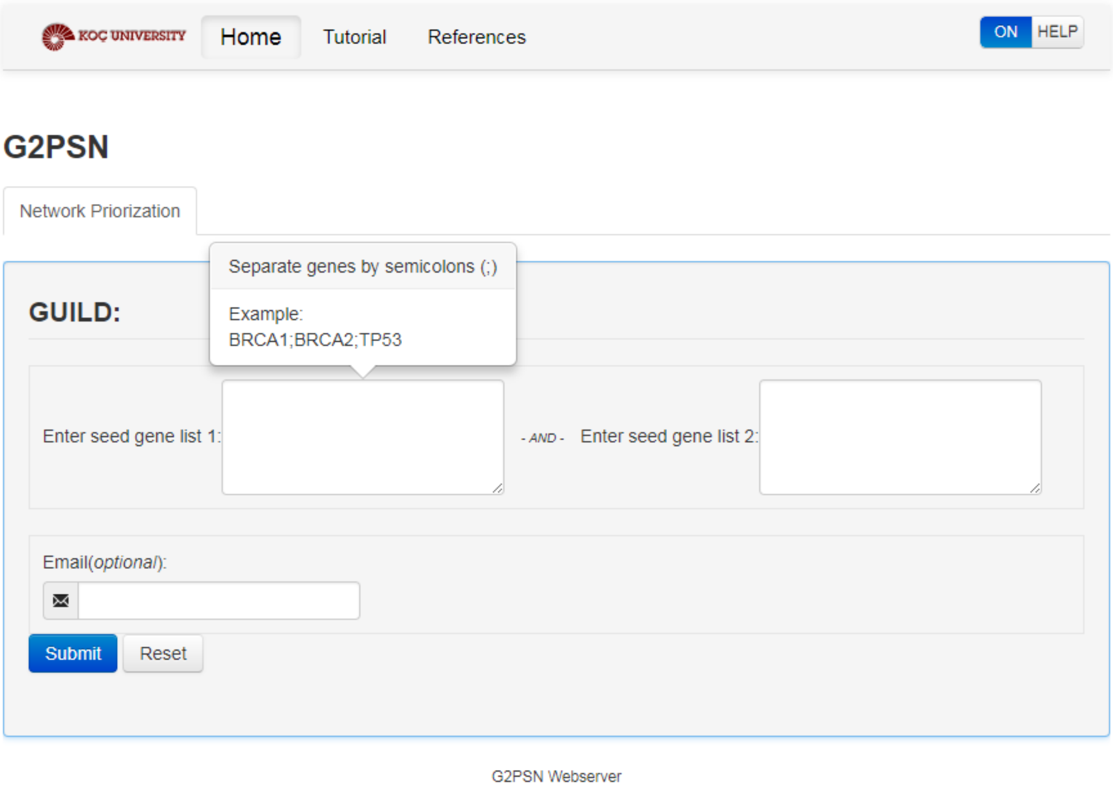
1) The Gene2Phen main page for entering the two sets of genes for query,

2) The Selection and Submission page to select or eliminate genes for submission,

3) The Results page to display the phenotype-specific subnetworks with all visualization, comparison and analysis functionalities.

4.2.1 Main Page

The Main page is used to initiate the first step of generating phenotype-specific subnetworks, which is network prioritization. For this step our tool uses network-topology based prioritization algorithms in GUILD to score relevance of gene products with respect to given gene symbols.



G2PSN

GUILD:

Separate genes by semicolons (;)

Example:
BRCA1;BRCA2;TP53

Enter seed gene list 1:

- AND -

Enter seed gene list 2:

Email(*optional*):

G2PSN Webserver

Figure 4.1 Overview of Gene2Phen main page

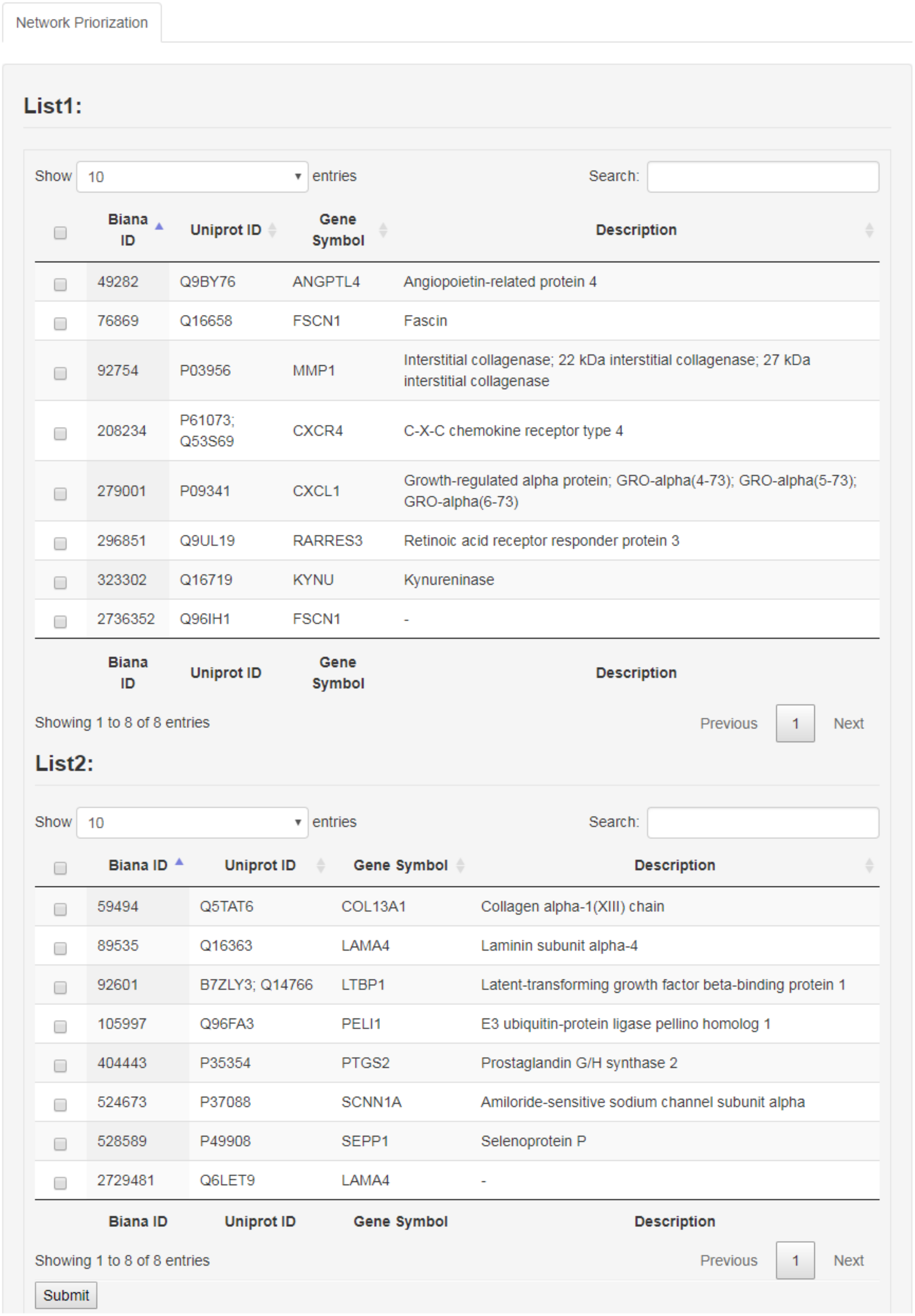
First, BIANA, a knowledge base containing data integrated from publicly available major data repositories, is queried with two sets of genes for two phenotypic profiles. Then these gene products are fed to a human interaction network (created using BIANA) as seed proteins. Finally, a score of relevance for each gene product in the network is calculated by the selected prioritization algorithm.

In order to enter a query which consist of a list of genes, user should separate the list by semicolons (e.g., BRCA1; BRCA2; TP53). Then, users can submit the prediction request (Figure 4.1). The tool use NetCombo as the default prioritization algorithm to execute the GUILD. The request will be put in a job queue to be executed in a cluster environment dedicated to the Gene2Phen server. Users will be given a link to follow the progress of their job status. Additionally, users can provide their e-mail addresses in the optional e-mail field to be notified when their jobs are submitted and their jobs are completed. After submission of the query, Selection and Submission page will be displayed.

4.2.2 Selection and Submission Page

First, for the user-provided gene symbols, BIANA-KB is queried and the products of the genes (e.g. proteins) associated with these gene symbols are listed. At this step, user may choose to use a subset of the listed genes or may choose all of them (Figure 4.2). Next, the products of these genes are used as seeds (initial gene-phenotype annotations) and NetCombo (default) method or any other selected method implemented in GUILD framework is run on a human protein-protein interaction network. The resulting scores are then listed along with the descriptive information of the gene products such as UniProt id, gene symbol, and description.

When the user selects the seed genes among the listed genes and presses the button for job submission, a link will be given for tracking the job submission process. When this process is completed, user will be able to access all visualization, comparison and analysis features of generated phenotype-specific subnetworks at the Results Page via this given link.



List1:

Show 10 entries Search:

□	Biana ID	Uniprot ID	Gene Symbol	Description
□	49282	Q9BY76	ANGPTL4	Angiopoietin-related protein 4
□	76869	Q16658	FSCN1	Fascin
□	92754	P03956	MMP1	Interstitial collagenase; 22 kDa interstitial collagenase; 27 kDa interstitial collagenase
□	208234	P61073; Q53S69	CXCR4	C-X-C chemokine receptor type 4
□	279001	P09341	CXCL1	Growth-regulated alpha protein; GRO-alpha(4-73); GRO-alpha(5-73); GRO-alpha(6-73)
□	296851	Q9UL19	RARRES3	Retinoic acid receptor responder protein 3
□	323302	Q16719	KYNU	Kynureninase
□	2736352	Q96IH1	FSCN1	-
	Biana ID	Uniprot ID	Gene Symbol	Description

Showing 1 to 8 of 8 entries

List2:

Show 10 entries Search:

□	Biana ID	Uniprot ID	Gene Symbol	Description
□	59494	Q5TAT6	COL13A1	Collagen alpha-1(XIII) chain
□	89535	Q16363	LAMA4	Laminin subunit alpha-4
□	92601	B7ZLY3; Q14766	LTBP1	Latent-transforming growth factor beta-binding protein 1
□	105997	Q96FA3	PELI1	E3 ubiquitin-protein ligase pellino homolog 1
□	404443	P35354	PTGS2	Prostaglandin G/H synthase 2
□	524673	P37088	SCNN1A	Amiloride-sensitive sodium channel subunit alpha
□	528589	P49908	SEPP1	Selenoprotein P
□	2729481	Q6LET9	LAMA4	-
	Biana ID	Uniprot ID	Gene Symbol	Description

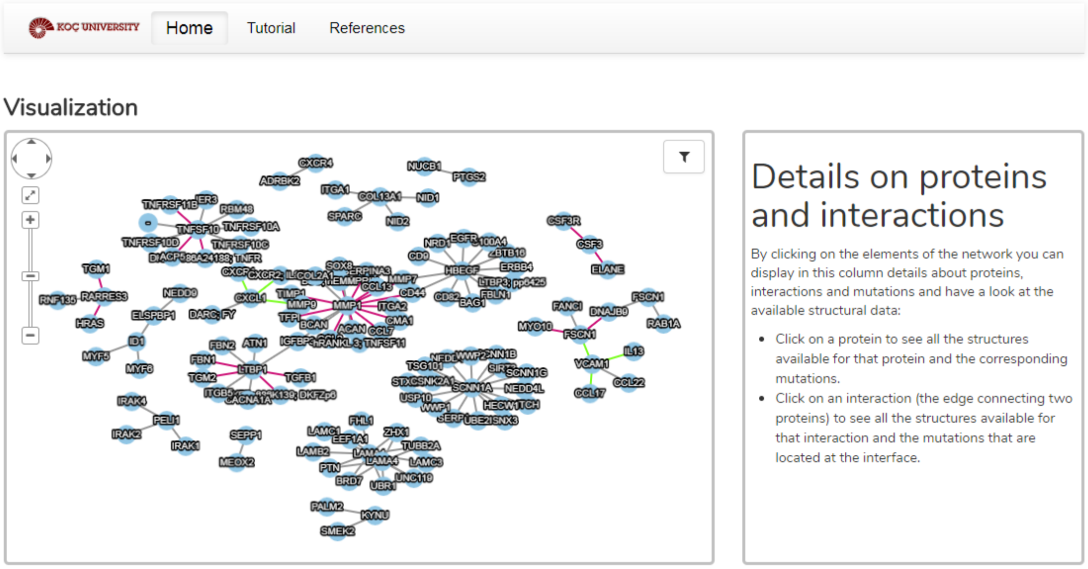
Showing 1 to 8 of 8 entries

Figure 4.2 Overview of Selection and Submission Page

4.2.3 Results Page

The previous page provides the links for the Result Page. This link is going to be available as soon as the scores are calculated by the server. "Access to results" link on this web page can be used go to the Result Page (when available).

In this page you can see two phenotype-specific subnetworks at Network Visualization Panel. (Figure 4.3) At the top left of the page user can find an interactive representation of the protein-protein interaction networks results from the query of the user. At the top right of the page Gene Annotation Panel is located. When user clicks on an edge, PDB-PDB Interactions Panel appears at the bottom.



Details on proteins and interactions

By clicking on the elements of the network you can display in this column details about proteins, interactions and mutations and have a look at the available structural data:

- Click on a protein to see all the structures available for that protein and the corresponding mutations.
- Click on an interaction (the edge connecting two proteins) to see all the structures available for that interaction and the mutations that are located at the interface.

Figure 4.3 Overview of Results Page

4.2.3.1 Network Visualization Panel

At this panel user will find an interactive representation of the protein-protein interaction networks (Figure 4.4). Nodes in this network represent proteins while edges represent protein-protein interactions (PPIs), as collected from public PPI databases.

The edges are colored based on the availability of structural information about the corresponding interaction:

🟩 Interactions which more associated with Phenotype-1

🟪 Interactions which more associated with Phenotype-2

⬜ Interaction without structural data

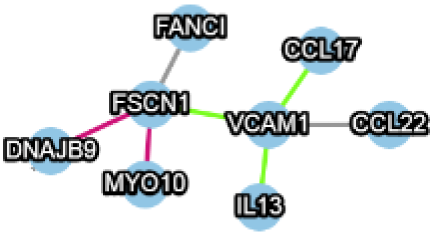
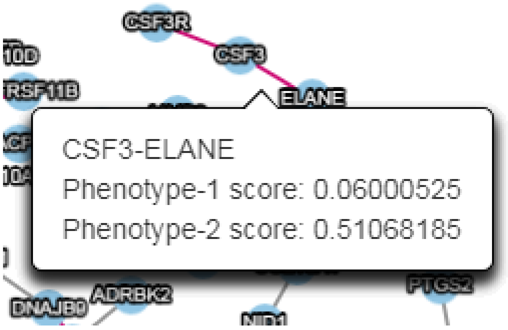


Figure 4.4 Network representation at the visualization panel

Interactive network representation means that you can move the nodes around, and you can select proteins (by clicking on the nodes) or interactions (by clicking on the edges). When user clicks on a protein or an interaction, our tool shows their corresponding details (including their structures and the mutations that can be mapped on them) in the Gene Annotation Panel and PDB-PDB Interactions Panel. Also, if users want to see details about one interaction, they can hover with the mouse over the edge and a tooltip will pop out (Figure 4.5).

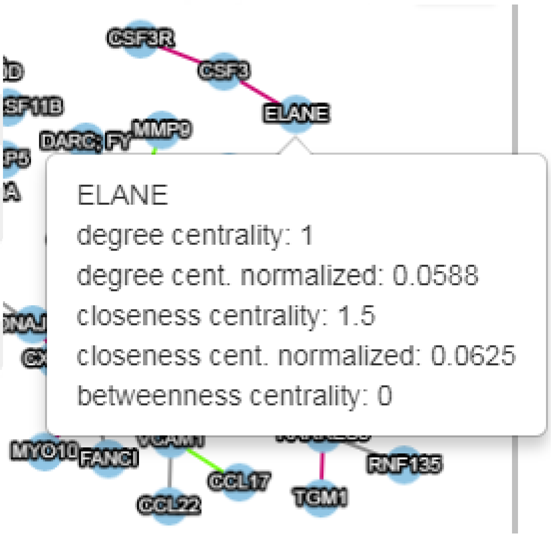


CSF3-ELANE
Phenotype-1 score: 0.06000525
Phenotype-2 score: 0.51068185

Figure 4.5 An edge tooltip example from the Network Visualization Panel

Edge tooltips show the relevance scores for two phenotypic profiles. In order to be shown an edge in this panel, at least one of two phenotype scores must be higher than the selected threshold by the user.

Likewise, if users want to see details about one protein, they can hover with the mouse over the node and a tooltip will pop out (Figure 4.6). Node tooltips show the topological properties of the protein which are degree centrality, normalized degree centrality, closeness centrality, normalized closeness centrality and betweenness centrality.

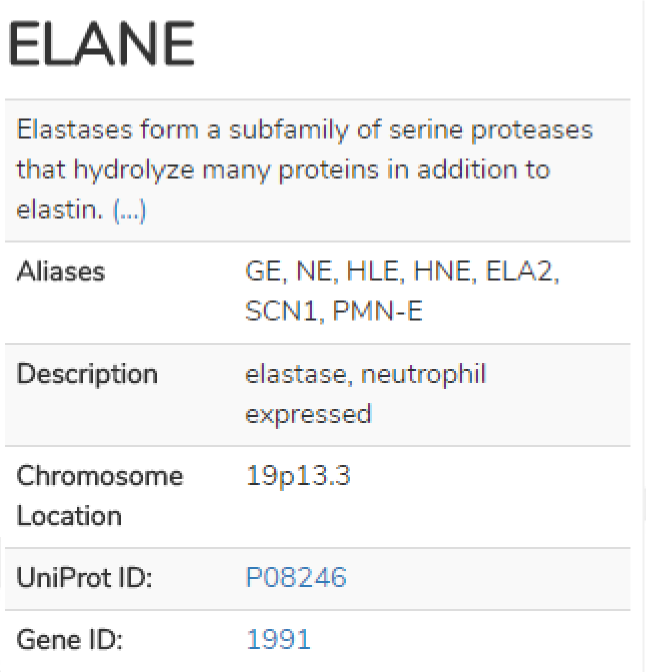


ELANE
degree centrality: 1
degree cent. normalized: 0.0588
closeness centrality: 1.5
closeness cent. normalized: 0.0625
betweenness centrality: 0

Figure 4.6 A node tooltip example from the Network Visualization Panel

4.2.3.2 Gene Annotation Panel

Inside Gene Annotation Panel user can find description, aliases, short description, chromosome location, UniProt ID and Gene ID data for the selected protein (Figure 4.7). Annotation data is obtained from BioGene.



ELANE

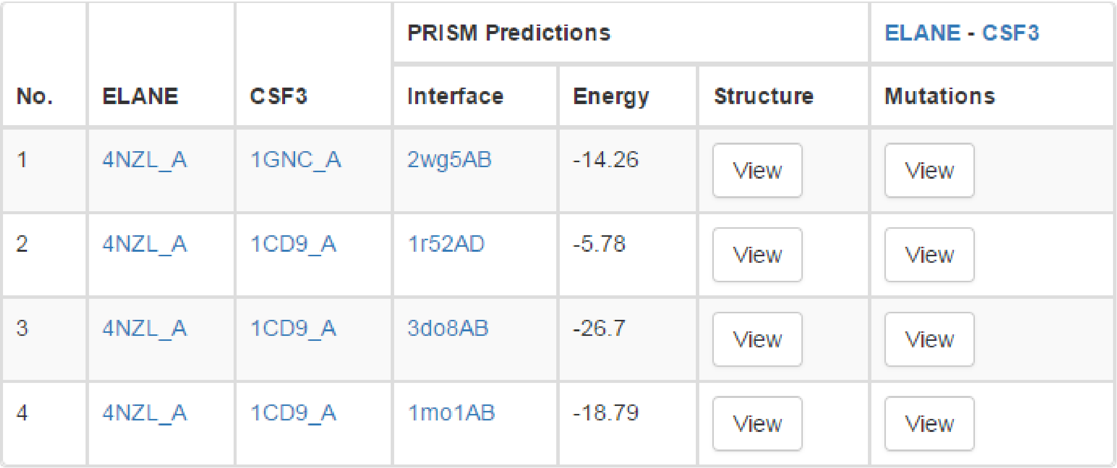
Elastases form a subfamily of serine proteases that hydrolyze many proteins in addition to elastin. (...)

Aliases	GE, NE, HLE, HNE, ELA2, SCN1, PMN-E
Description	elastase, neutrophil expressed
Chromosome Location	19p13.3
UniProt ID:	P08246
Gene ID:	1991

Figure 4.7 An example from the Gene Annotation Panel

4.2.3.3 PDB-PDB Interaction Panel

This panel shows the list of PRISM predictions, the interface of each PDB-PDB interactions, energy values of the interaction and three-dimensional structure of the prediction. Users could visualize three-dimensional structures by selecting the panels built with JSmol framework. Figure 4.8 shows the panel which gives to the user ability to screen the three-dimensional structure (Figure 4.9), download the structure on PDB format and create a text file involving the list of interfaces and interacting residues.



			PRISM Predictions			ELANE - CSF3
No.	ELANE	CSF3	Interface	Energy	Structure	Mutations
1	4NZL_A	1GNC_A	2wg5AB	-14.26	View	View
2	4NZL_A	1CD9_A	1r52AD	-5.78	View	View
3	4NZL_A	1CD9_A	3do8AB	-26.7	View	View
4	4NZL_A	1CD9_A	1mo1AB	-18.79	View	View

Figure 4.8 PRISM predictions for all PDB-PDB interactions of ELANE and CSF3 genes and the mutations data on predicted interfaces.

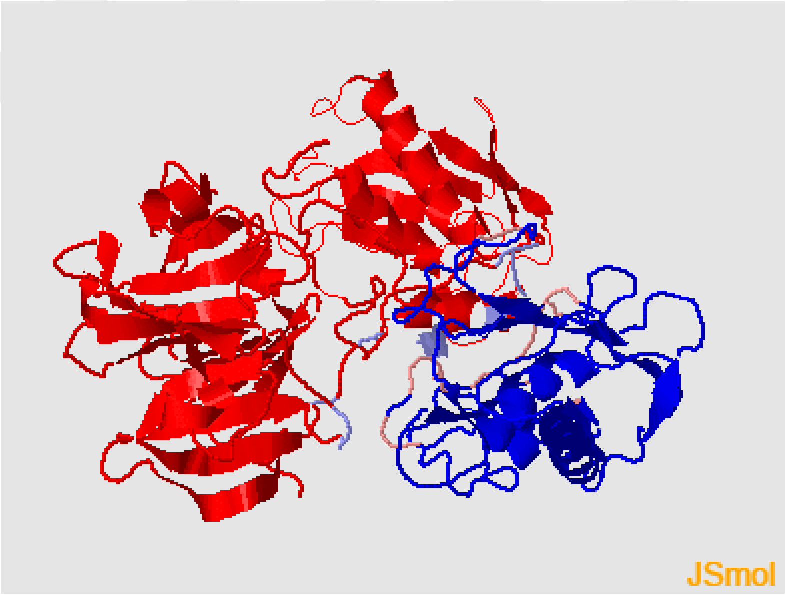
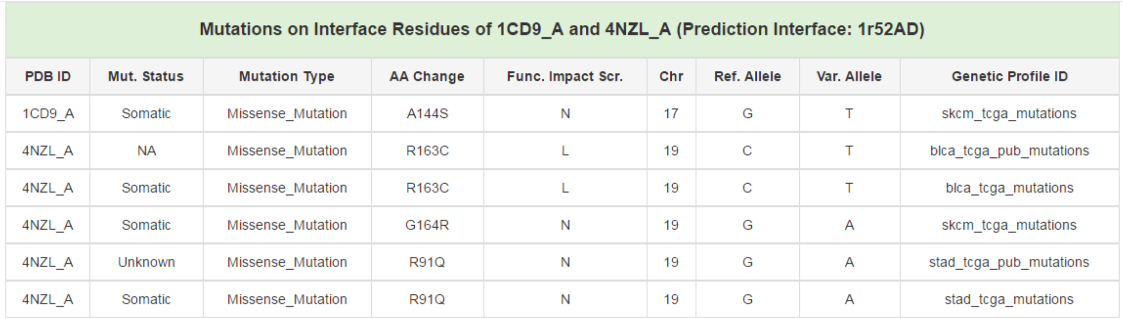


Figure 4.9 A sample JSmol view of PRISM prediction data of a PDB-PDB interaction.

In addition, on the same page, users can list the cancer-related point mutations located on the interfaces of protein complexes predicted by PRISM in the column next to PRISM predictions called "Mutations" (Figure 4.10). The mutation data about selected gene interactions is derived from cBioPortal web site. The filtration is done after eliminating the mutations not located on selected PDB's of those genes and the mutations not located on predicted interface even on the PDB.



Mutations on Interface Residues of 1CD9_A and 4NZL_A (Prediction Interface: 1r52AD)

PDB ID	Mut. Status	Mutation Type	AA Change	Func. Impact Scr.	Chr	Ref. Allele	Var. Allele	Genetic Profile ID
1CD9_A	Somatic	Missense_Mutation	A144S	N	17	G	T	skcm_tcga_mutations
4NZL_A	NA	Missense_Mutation	R163C	L	19	C	T	blca_tcga_pub_mutations
4NZL_A	Somatic	Missense_Mutation	R163C	L	19	C	T	blca_tcga_mutations
4NZL_A	Somatic	Missense_Mutation	G164R	N	19	G	A	skcm_tcga_mutations
4NZL_A	Unknown	Missense_Mutation	R91Q	N	19	G	A	stad_tcga_pub_mutations
4NZL_A	Somatic	Missense_Mutation	R91Q	N	19	G	A	stad_tcga_mutations

Figure 4.10 Mutations on Interface Residues of 1CD9_A and 4NZL_A

4.3 Case Study

As a case study, we traced the findings of two previous studies from our research group from 2013 [21] and 2015 [48]. Both studies focus on brain and lung metastasis differentiation of breast cancer with structure-integrated PPI network analysis. To understand the molecular mechanism of these two metastases, they have generated brain and lung metastasis subnetworks. These subnetworks are generated with brain and lung metastasis mediator genes which are identified by Massagué and his coworkers in 2005 and 2009 [49, 50] (Table 4.1).

Table 4.1: Breast cancer lung and brain metastasis associated gene sets

Genes significantly associated with breast cancer lung metastasis	**Genes significantly associated with breast cancer 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 (KRT81)	RGC32
VCAM1	PELI1
LY6E	TNFSF10
EREG	B4GALT6
NEDD9	HBEGF
MAN1A1	CSF3
ID1	

In our case study, we utilized the same seed genes in order to construct phenotype specific subnetworks. Size of network will differentiate by choosing various cutoff values therefore we tried various of them in a wide range from 0 to 1. In Figure 4.11, brain and lung metastasis subnetworks are generated with 0.5 cutoff score and represented in one merged network embodiment. In this network, interactions which painted in pink are more associated with brain metastasis while the green ones has higher associability with lung metastasis and the grey edges has no structure since its connecting nodes does not have a 3D structure stored in PDB database. Figure 4.11 demonstrates the brain and lung metastasis subnetworks with 0.5 cutoff .

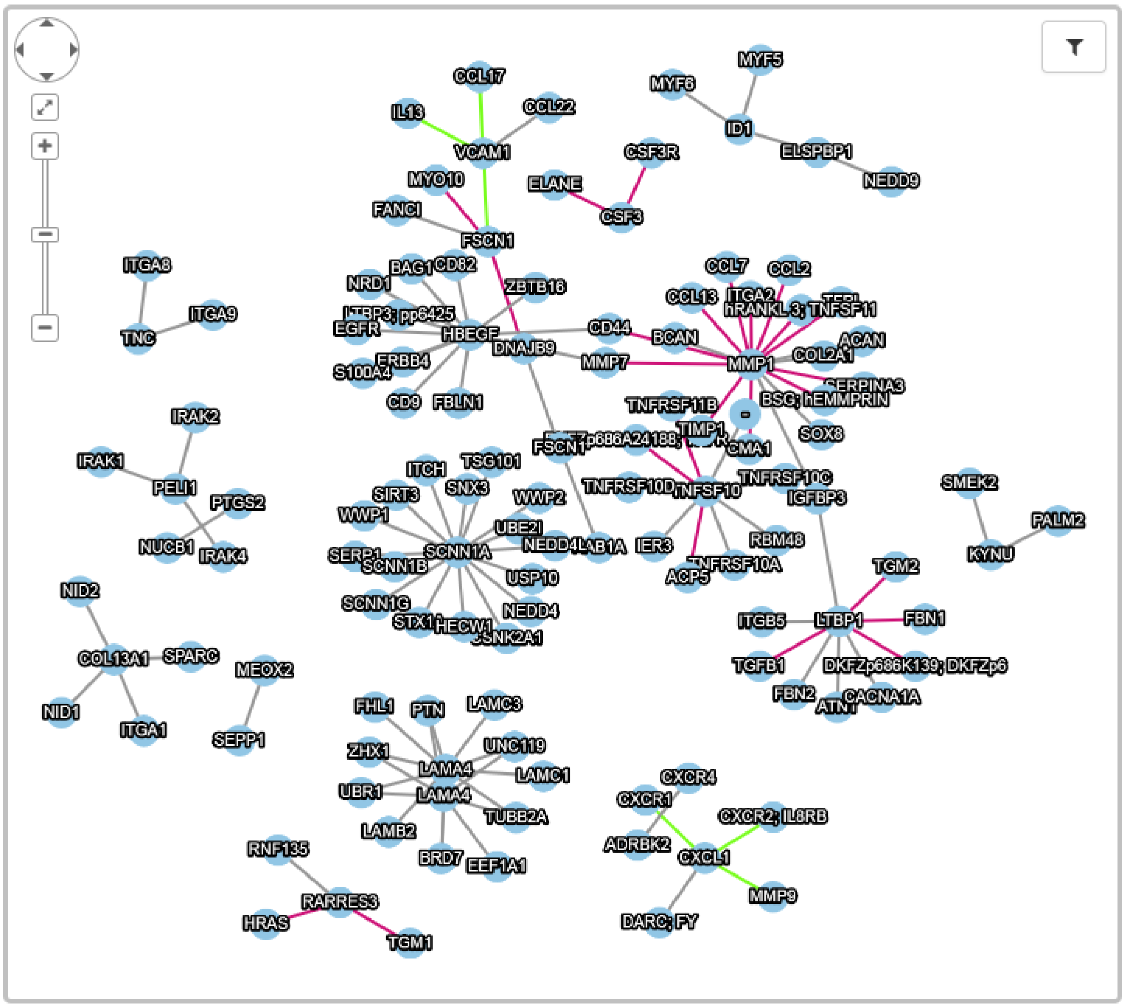


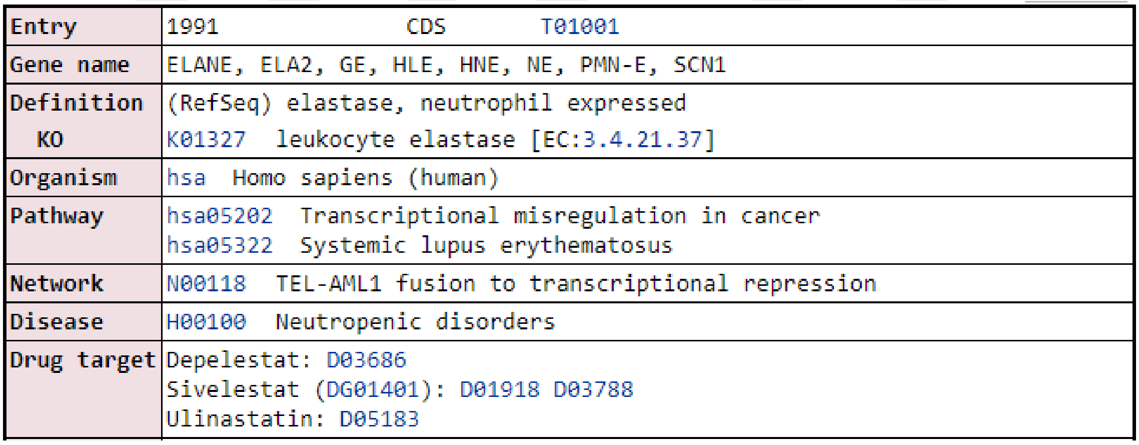
Figure 4.11 Brain and lung metastasis subnetwork with 0.5 cutoff threshold

4.3.1 ELANE – VCAM1 & ELANE - CXCR4

In the previous study, some novel candidate genes/proteins which are previously not associated with brain and lung metastasis of breast cancer profiles were identified as these genes might be important and need to be further examination by using experimental methods [48]. Elastase neutrophil expressed protein (ELANE) was one of the genes which were identified as associated to lung metastasis according to topological analysis and literature search.

In the lung metastasis subnetwork, CXCR4, ELANE, CCR5, NEDD9, DPP4, MMP9 and MMP2 are considered as hub-bottleneck proteins. After identification of these topologically significant nodes, gene annotations are examined. In this case study we will focus on ELANE protein.

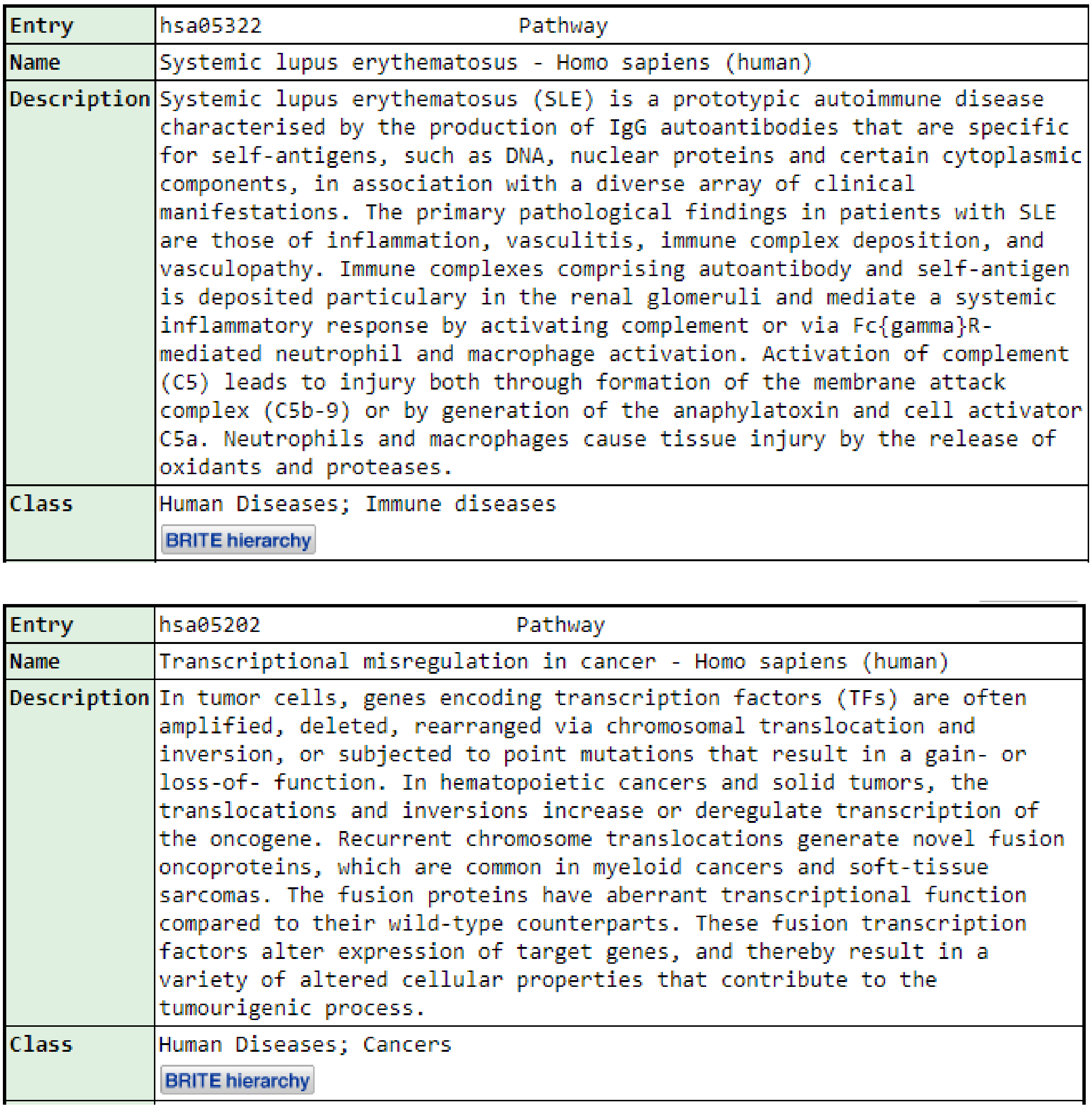
Elastases form a subfamily of serine proteases that hydrolyze many proteins in addition to elastin. Humans have six elastase genes which encode structurally similar proteins. The encoded preproprotein is processed to generate the active protease. Following activation, this protease hydrolyzes proteins within specialized neutrophil lysosomes, called azurophil granules, as well as proteins of the extracellular matrix. The enzyme may play a role in degenerative and inflammatory diseases through proteolysis of collagen-IV and elastin. This protein also degrades the outer membrane protein A of E. coli as well as the virulence factors of such bacteria as Shigella, Salmonella and Yersinia. Mutations in this gene are associated with cyclic neutropenia and severe congenital neutropenia [RefSeq, Jan 2016].



Entry	1991 CDS T01001
Gene name	ELANE, ELA2, GE, HLE, HNE, NE, PMN-E, SCN1
Definition KO	(RefSeq) elastase, neutrophil expressed K01327 leukocyte elastase [EC:3.4.21.37]
Organism	hsa Homo sapiens (human)
Pathway	hsa05202 Transcriptional misregulation in cancer hsa05322 Systemic lupus erythematosus
Network	N00118 TEL-AML1 fusion to transcriptional repression
Disease	H00100 Neutropenic disorders
Drug target	Depelestat: D03686 Sivelestat (DG01401): D01918 D03788 Ulinastatin: D05183

Figure 4.12: ELANE entry in KEGG Pathway database

In Figure 4.12 the KEGG [51] entry for ELANE protein is shown, also Figure 4.13 presents the detailed description of the KEGG pathways which ELANE is involved (transcriptional misregulation in cancer, systemic lupus erythematosus). It seems that this enzyme play an active role on tumor progression that leads to metastasis in human breast cancer. Since ELANE have a role on producing cytokines and chemokines that are crucial for both inflammation and cancer, it is considered as this enzyme play an important role on tumor progression [52]. Akizuki et al. found that breast cancer patients with poor survival rate have high concentration of ELANE [53].

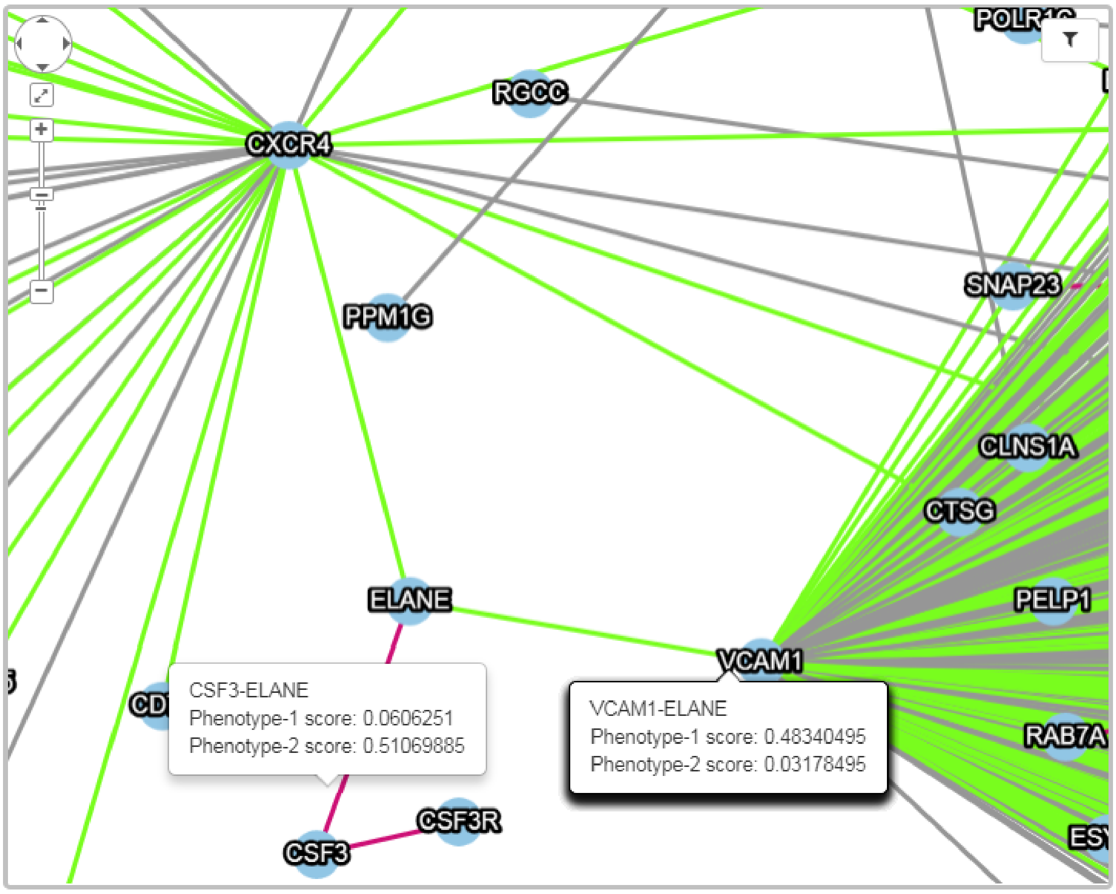


Entry	hsa05322 Pathway
Name	Systemic lupus erythematosus - Homo sapiens (human)
Description	Systemic lupus erythematosus (SLE) is a prototypic autoimmune disease characterised by the production of IgG autoantibodies that are specific for self-antigens, such as DNA, nuclear proteins and certain cytoplasmic components, in association with a diverse array of clinical manifestations. The primary pathological findings in patients with SLE are those of inflammation, vasculitis, immune complex deposition, and vasculopathy. Immune complexes comprising autoantibody and self-antigen is deposited particulary in the renal glomeruli and mediate a systemic inflammatory response by activating complement or via Fc{gamma}R-mediated neutrophil and macrophage activation. Activation of complement (C5) leads to injury both through formation of the membrane attack complex (C5b-9) or by generation of the anaphylatoxin and cell activator C5a. Neutrophils and macrophages cause tissue injury by the release of oxidants and proteases.
Class	Human Diseases; Immune diseases BRITE hierarchy

Entry	hsa05202 Pathway
Name	Transcriptional misregulation in cancer - Homo sapiens (human)
Description	In tumor cells, genes encoding transcription factors (TFs) are often amplified, deleted, rearranged via chromosomal translocation and inversion, or subjected to point mutations that result in a gain- or loss-of- function. In hematopoietic cancers and solid tumors, the translocations and inversions increase or deregulate transcription of the oncogene. Recurrent chromosome translocations generate novel fusion oncoproteins, which are common in myeloid cancers and soft-tissue sarcomas. The fusion proteins have aberrant transcriptional function compared to their wild-type counterparts. These fusion transcription factors alter expression of target genes, and thereby result in a variety of altered cellular properties that contribute to the tumourigenic process.
Class	Human Diseases; Cancers BRITE hierarchy

Figure 4.13: KEGG pathways that ELANE is participated

In addition, as can be seen in Figure 4.12, ELANE is first neighbor of lung metastasis seed genes and hub-bottlenecks CXCR4 and VCAM1. When we consider all these topological analysis and annotation data, ELANE can be considered as a good candidate for mediating lung metastasis subnetworks.



CSF3-ELANE
Phenotype-1 score: 0.0606251
Phenotype-2 score: 0.51069885

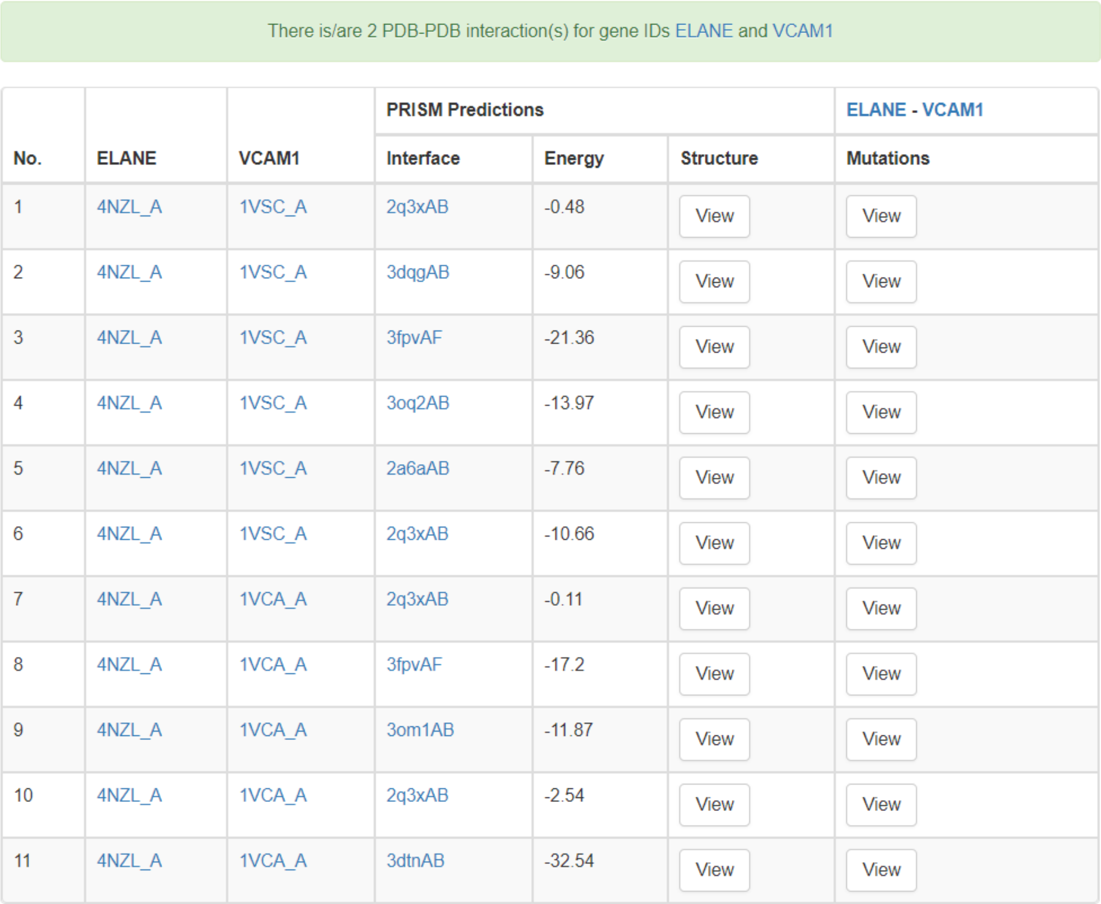
VCAM1-ELANE
Phenotype-1 score: 0.48340495
Phenotype-2 score: 0.03178495

Figure 4.14 ELANE-CXCR4, ELANE-VCAM1 and ELANE-CSF interactions

4.3.2 ELANE – VCAM1 & ELANE – CSF3

Another interesting finding about ELANE - which can be also seen at Figure 4.12 - is that ELANE is an interaction partner for both VCAM1 from the lung metastasis subnetwork and CSF3 from the brain metastasis subnetwork so it is possible to switch one to another. Both genes have employed as seed genes in their metastasis subnetworks therefore interaction and switching mechanisms of ELANE with these two genes are needed to be investigated since it might change the course of the metastasis.

After selecting the edge of ELANE and VCAM1, we see that all energy scores are negative which means these predictions are sufficient to see them possible (Figure 4.13). Then we proceed to mutation mapping for each PDB-PDB interaction for ELANE-VCAM1 complexes and focused on mutations on the ELANE PDBs.

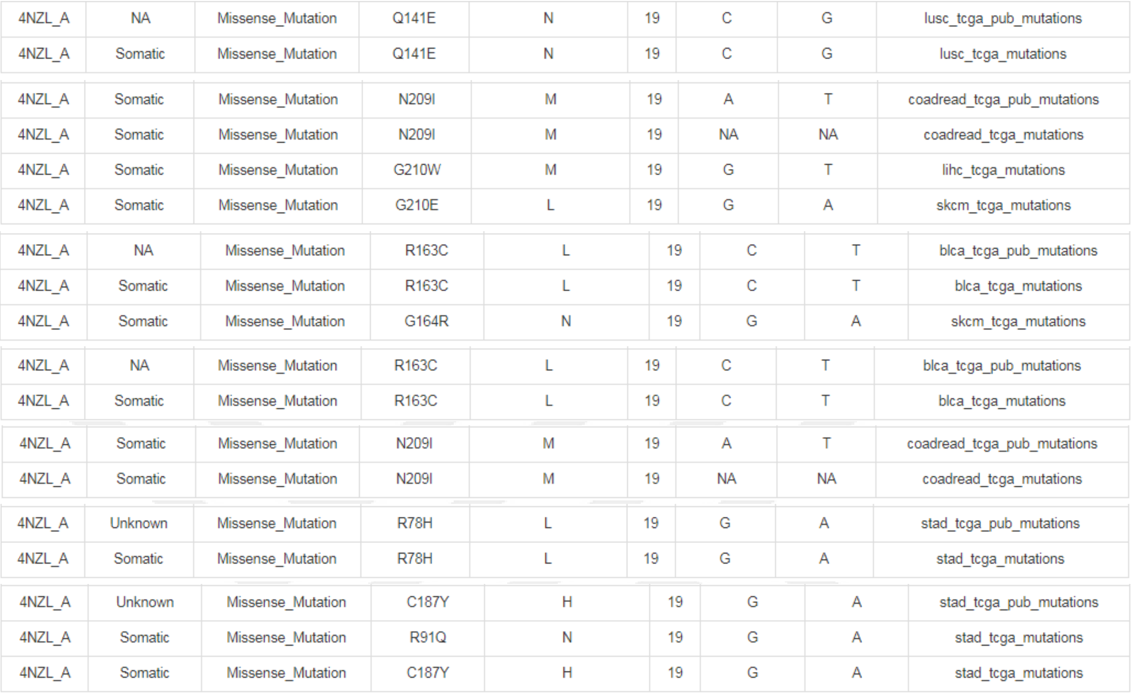


There is/are 2 PDB-PDB interaction(s) for gene IDs ELANE and VCAM1

			PRISM Predictions			ELANE - VCAM1
No.	ELANE	VCAM1	Interface	Energy	Structure	Mutations
1	4NZL_A	1VSC_A	2q3xAB	-0.48	View	View
2	4NZL_A	1VSC_A	3dqgAB	-9.06	View	View
3	4NZL_A	1VSC_A	3fpvAF	-21.36	View	View
4	4NZL_A	1VSC_A	3oq2AB	-13.97	View	View
5	4NZL_A	1VSC_A	2a6aAB	-7.76	View	View
6	4NZL_A	1VSC_A	2q3xAB	-10.66	View	View
7	4NZL_A	1VCA_A	2q3xAB	-0.11	View	View
8	4NZL_A	1VCA_A	3fpvAF	-17.2	View	View
9	4NZL_A	1VCA_A	3om1AB	-11.87	View	View
10	4NZL_A	1VCA_A	2q3xAB	-2.54	View	View
11	4NZL_A	1VCA_A	3dtnAB	-32.54	View	View

Figure 4.15 PRISM predictions for the protein products of ELANE and VCAM1

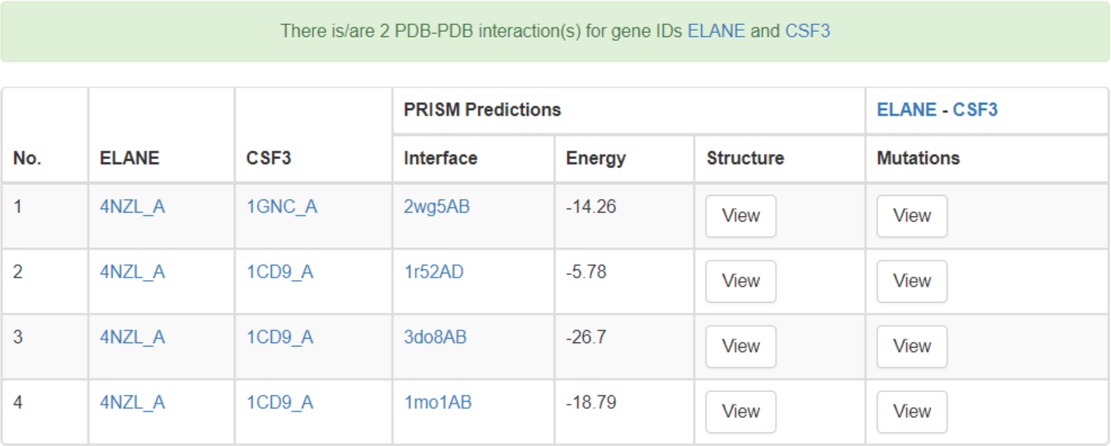
At Figure 4.14, all interface mutations may occur on PDB structures of ELANE in a complex form with VCAM1 are listed.



4NZL_A	NA	Missense_Mutation	Q141E	N	19	C	G	lusc_tcga_pub_mutations
4NZL_A	Somatic	Missense_Mutation	Q141E	N	19	C	G	lusc_tcga_mutations
4NZL_A	Somatic	Missense_Mutation	N209I	M	19	A	T	coadread_tcga_pub_mutations
4NZL_A	Somatic	Missense_Mutation	N209I	M	19	NA	NA	coadread_tcga_mutations
4NZL_A	Somatic	Missense_Mutation	G210W	M	19	G	T	lihc_tcga_mutations
4NZL_A	Somatic	Missense_Mutation	G210E	L	19	G	A	skcm_tcga_mutations
4NZL_A	NA	Missense_Mutation	R163C	L	19	C	T	blca_tcga_pub_mutations
4NZL_A	Somatic	Missense_Mutation	R163C	L	19	C	T	blca_tcga_mutations
4NZL_A	Somatic	Missense_Mutation	G164R	N	19	G	A	skcm_tcga_mutations
4NZL_A	NA	Missense_Mutation	R163C	L	19	C	T	blca_tcga_pub_mutations
4NZL_A	Somatic	Missense_Mutation	R163C	L	19	C	T	blca_tcga_mutations
4NZL_A	Somatic	Missense_Mutation	N209I	M	19	A	T	coadread_tcga_pub_mutations
4NZL_A	Somatic	Missense_Mutation	N209I	M	19	NA	NA	coadread_tcga_mutations
4NZL_A	Unknown	Missense_Mutation	R78H	L	19	G	A	stad_tcga_pub_mutations
4NZL_A	Somatic	Missense_Mutation	R78H	L	19	G	A	stad_tcga_mutations
4NZL_A	Unknown	Missense_Mutation	C187Y	H	19	G	A	stad_tcga_pub_mutations
4NZL_A	Somatic	Missense_Mutation	R91Q	N	19	G	A	stad_tcga_mutations
4NZL_A	Somatic	Missense_Mutation	C187Y	H	19	G	A	stad_tcga_mutations

Figure 4.16 All interface mutations on PDB structures of ELANE interacts with VCAM1

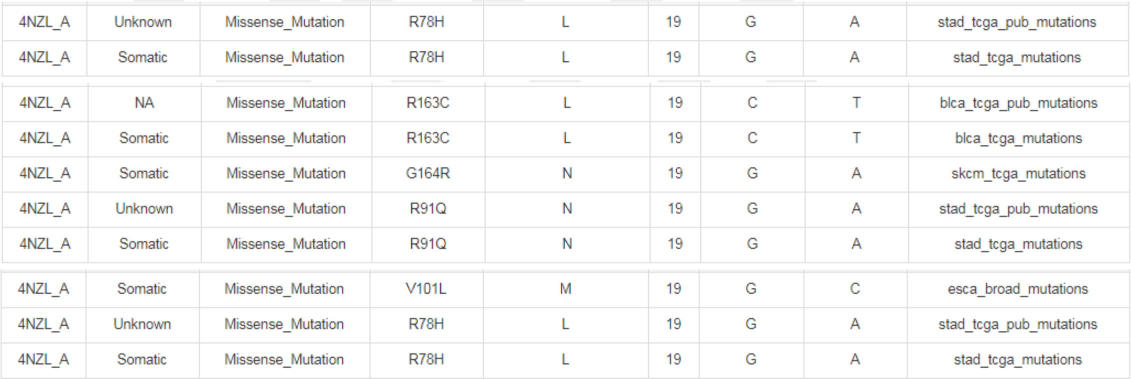
Same procedures are followed for finding all interface mutations on PDB structures of ELANE in a complex form with CSF3 for comparison. After selecting the edge of ELANE and CSF3, we observe that all binding energy scores are negative therefore these predictions seem to be realistic interfaces (Figure 4.15). Then we proceed to mutation mapping for each PDB-PDB interaction for ELANE-CSF3 complexes and focused on mutations on the ELANE PDBs (Figure 4.16).



There is/are 2 PDB-PDB interaction(s) for gene IDs ELANE and CSF3

			PRISM Predictions			ELANE - CSF3
No.	ELANE	CSF3	Interface	Energy	Structure	Mutations
1	4NZL_A	1GNC_A	2wg5AB	-14.26	View	View
2	4NZL_A	1CD9_A	1r52AD	-5.78	View	View
3	4NZL_A	1CD9_A	3do8AB	-26.7	View	View
4	4NZL_A	1CD9_A	1mo1AB	-18.79	View	View

Figure 4.17 PRISM predictions for the protein products of ELANE and CSF3



4NZL_A	Unknown	Missense_Mutation	R78H	L	19	G	A	stad_tcga_pub_mutations
4NZL_A	Somatic	Missense_Mutation	R78H	L	19	G	A	stad_tcga_mutations
4NZL_A	NA	Missense_Mutation	R163C	L	19	C	T	blca_tcga_pub_mutations
4NZL_A	Somatic	Missense_Mutation	R163C	L	19	C	T	blca_tcga_mutations
4NZL_A	Somatic	Missense_Mutation	G164R	N	19	G	A	skcm_tcga_mutations
4NZL_A	Unknown	Missense_Mutation	R91Q	N	19	G	A	stad_tcga_pub_mutations
4NZL_A	Somatic	Missense_Mutation	R91Q	N	19	G	A	stad_tcga_mutations
4NZL_A	Somatic	Missense_Mutation	V101L	M	19	G	C	esca_broad_mutations
4NZL_A	Unknown	Missense_Mutation	R78H	L	19	G	A	stad_tcga_pub_mutations
4NZL_A	Somatic	Missense_Mutation	R78H	L	19	G	A	stad_tcga_mutations

Figure 4.18 All interface mutations on PDB structures of ELANE interacts with CSF3

When we put together all findings about ELANE mutations on ELANE-VCAM and ELANE-CSF3, we observe that Q141E, N209I, G210W, G210E, C187Y mutations only occur on interface region of ELANE-VCAM complex while V101L mutations only occurs on ELANE-CSF3 interface (Table 4.2). Among these 6 mutations, only Q141E is obtained from a lung cancer related study [54]. Therefore other five mutations can be good candidates for future studies.

Table 4.2 ELANE Mutations on ELANE-VCAM and ELANE-CSF3 interactions

Locations	ELANE Mutations
Interface of ELANE-VCAM	Q141E, N209I, G210W, G210E, R163C, G164R, R78H, C187Y, R91Q
Interface of ELANE-CSF3	R78H, R163C, G164R, R91Q, V101L
Mutual for ELANE-VCAM and ELANE-CSF3 interfaces	R78H, R163C, G164R, R91Q
Only on ELANE-VCAM interface	Q141E, N209I, G210W, G210E, C187Y
Only on ELANE-CSF3 interface	V101L

Also this observation can be interpreted as these mutations may change the interaction behaviors of ELANE and cause to function for a different type of metastasis in breast cancer; such as instead of brain metastasis, lung metastasis may happen because of the mutation that occurs on the interface regions of ELANE. Therefore, these mutations could be considered as more important and they would be prioritized for experimental studies.

Chapter 5

CONCLUSION

In this thesis, we aim to develop a web-based tool by integrating human protein-protein interaction network, structural prediction of protein - protein interactions and interface mutations to help researchers in exploring and comparing the molecular mechanism of different phenotypes. Our software works as an automatized pipeline tool to build, visualize and compare phenotype specific subnetworks, to predict and examine protein-protein interactions associated with their structure and mutation data.

By using Gen2Phen web tool users can prioritize the human protein - protein network according to desired set of genes that they aim to investigate. From the scored-PPI network, they can generate a phenotype specific subnetwork by putting a cut-off score. The tool visualizes generated phenotype-specific subnetworks as an interactive network representation. Topological significance and genome annotations of each proteins are displayed in this interactive network representation. With the aid of Gen2Phen users can go into the details on predicted possible interface structures of interacting protein pairs in the phenotype-specific subnetworks. Our tool shows 3D structural models of known protein-complexes and their predicted protein-protein interfaces. Users can see the list of mutations which are mapped on predicted protein-protein interfaces. Users can use all features of the tool for analyzing and comparing two phenotypes at the same time.

As a case study for testing the functionalities of Gen2Phen, we traced genotype to phenotype relationship studies on brain and lung metastasis subnetworks of breast cancer. Topological scores and gene annotations guided us to identify new candidate disease proteins in two subnetworks. Structural knowledge and mutation mapping provided observation of mutually exclusive proteins and game changer interface mutations among different phenotypes.

REFERENCES

1. Aloy, P. and R.B. Russell, *Ten thousand interactions for the molecular biologist.* Nature Biotechnology, 2004. **22**(10): p. 1317-1321.
2. Gavin, A.C., et al., *Proteome survey reveals modularity of the yeast cell machinery.* Nature, 2006. **440**(7084): p. 631-636.
3. Sevimoglu, T. and K.Y. Arga, *The role of protein interaction networks in systems biomedicine.* Comput Struct Biotechnol J, 2014. **11**(18): p. 22-7.
4. 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.
5. Gonzalez, M.W. and M.G. Kann, *Chapter 4: Protein interactions and disease.* PLoS Comput Biol, 2012. **8**(12): p. e1002819.
6. Guney, E. and B. Oliva, *Exploiting protein-protein interaction networks for genome-wide disease-gene prioritization.* PLoS One, 2012. **7**(9): p. e43557.
7. Franz, M., et al., *Cytoscape.js: a graph theory library for visualisation and analysis.* Bioinformatics, 2016. **32**(2): p. 309-11.
8. 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.
9. Gao, J., et al., *Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal.* Sci Signal, 2013. **6**(269): p. pl1.
10. Mosca, R., et al., *dSysMap: exploring the edgetic role of disease mutations.* Nat Methods, 2015. **12**(3): p. 167-8.
11. Braun, P. and A.C. Gingras, *History of protein-protein interactions: from egg-white to complex networks.* Proteomics, 2012. **12**(10): p. 1478-98.
12. Chautard, E., N. Thierry-Mieg, and S. Ricard-Blum, *Interaction networks: from protein functions to drug discovery. A review.* Pathol Biol (Paris), 2009. **57**(4): p. 324-33.
13. Ideker, T. and R. Sharan, *Protein networks in disease.* Genome Res, 2008. **18**(4): p. 644-52.
14. Hartwell, L.H., et al., *From molecular to modular cell biology.* Nature, 1999. **402**(6761 Suppl): p. C47-52.
15. Gstaiger, M. and R. Aebersold, *Genotype-phenotype relationships in light of a modular protein interaction landscape.* Mol Biosyst, 2013. **9**(6): p. 1064-7.
16. Costanzo, M., et al., *The genetic landscape of a cell.* Science, 2010. **327**(5964): p. 425-31.
17. Nibbe, R.K., et al., *Protein-protein interaction networks and subnetworks in the biology of disease.* Wiley Interdiscip Rev Syst Biol Med, 2011. **3**(3): p. 357-67.

18. Perez-Iratxeta, C., P. Bork, and M.A. Andrade, *Association of genes to genetically inherited diseases using data mining.* Nat Genet, 2002. **31**(3): p. 316-9.
19. Lan, W., et al., *Computational Approaches for Prioritizing Candidate Disease Genes Based on PPI Networks.* Tsinghua Science and Technology, 2015. **20**(5): p. 500-512.
20. Chen, J., B.J. Aronow, and A.G. Jegga, *Disease candidate gene identification and prioritization using protein interaction networks.* BMC Bioinformatics, 2009. **10**: p. 73.
21. 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.
22. Birlutiu, A. and T. Heskes, *Using Topology Information for Protein-Protein Interaction Prediction.* Pattern Recognition in Bioinformatics, Prib 2014, 2014. **8626**: p. 10-22.
23. Yook, S.H., Z.N. Oltvai, and A.L. Barabasi, *Functional and topological characterization of protein interaction networks.* Proteomics, 2004. **4**(4): p. 928-42.
24. Janjic, V. and N. Przulj, *Biological function through network topology: a survey of the human diseasome.* Brief Funct Genomics, 2012. **11**(6): p. 522-32.
25. Barabasi, A.L. and R. Albert, *Emergence of scaling in random networks.* Science, 1999. **286**(5439): p. 509-12.
26. Barabasi, A.L. and Z.N. Oltvai, *Network biology: understanding the cell's functional organization.* Nat Rev Genet, 2004. **5**(2): p. 101-13.
27. Jeong, H., et al., *Lethality and centrality in protein networks.* Nature, 2001. **411**(6833): p. 41-2.
28. He, X.L. and J.Z. Zhang, *Why do hubs tend to be essential in protein networks?* Plos Genetics, 2006. **2**(6): p. 826-834.
29. Han, J.D., et al., *Evidence for dynamically organized modularity in the yeast protein-protein interaction network.* Nature, 2004. **430**(6995): p. 88-93.
30. 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.
31. Aloy, P. and R.B. Russell, *Interrogating protein interaction networks through structural biology.* Proc Natl Acad Sci U S A, 2002. **99**(9): p. 5896-901.
32. Kim, P.M., et al., *Relating three-dimensional structures to protein networks provides evolutionary insights.* Science, 2006. **314**(5807): p. 1938-41.
33. Kuzu, G., et al., *Constructing structural networks of signaling pathways on the proteome scale.* Curr Opin Struct Biol, 2012. **22**(3): p. 367-77.
34. Kar, G., A. Gursoy, and O. Keskin, *Human cancer protein-protein interaction network: a structural perspective.* PLoS Comput Biol, 2009. **5**(12): p. e1000601.

35. Cheng, F., J. Zhao, and Z. Zhao, *Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes.* Brief Bioinform, 2016. **17**(4): p. 642-56.
36. Wang, X., et al., *Three-dimensional reconstruction of protein networks provides insight into human genetic disease.* Nat Biotechnol, 2012. **30**(2): p. 159-64.
37. Vazquez, M., A. Valencia, and T. Pons, *Structure-PPi: a module for the annotation of cancer-related single-nucleotide variants at protein-protein interfaces.* Bioinformatics, 2015. **31**(14): p. 2397-9.
38. Ceol, A. and H. Muller, *The MI bundle: enabling network and structural biology in genome visualization tools.* Bioinformatics, 2015. **31**(22): p. 3679-81.
39. Lu, H.C., J. Herrera Braga, and F. Fraternali, *PinSnps: structural and functional analysis of SNPs in the context of protein interaction networks.* Bioinformatics, 2016. **32**(16): p. 2534-6.
40. Witvliet, D.K., et al., *ELASPIC web-server: proteome-wide structure-based prediction of mutation effects on protein stability and binding affinity.* Bioinformatics, 2016. **32**(10): p. 1589-91.
41. Solomon, O., et al., *G23D: Online tool for mapping and visualization of genomic variants on 3D protein structures.* BMC Genomics, 2016. **17**: p. 681.
42. Gress, A., et al., *StructMAn: annotation of single-nucleotide polymorphisms in the structural context.* Nucleic Acids Res, 2016. **44**(W1): p. W463-8.
43. Li, M., et al., *MutaBind estimates and interprets the effects of sequence variants on protein-protein interactions.* Nucleic Acids Res, 2016. **44**(W1): p. W494-501.
44. Lage, K., *Protein-protein interactions and genetic diseases: The interactome.* Biochimica Et Biophysica Acta-Molecular Basis of Disease, 2014. **1842**(10): p. 1971-1980.
45. Maglott, D., et al., *Entrez Gene: gene-centered information at NCBI.* Nucleic Acids Res, 2011. **39**(Database issue): p. D52-7.
46. Mashiach, E., R. Nussinov, and H.J. Wolfson, *FiberDock: Flexible induced-fit backbone refinement in molecular docking.* Proteins, 2010. **78**(6): p. 1503-19.
47. Hanson, R.M., et al., *JSmol and the Next-Generation Web-Based Representation of 3D Molecular Structure as Applied to Proteopedia.* Israel Journal of Chemistry, 2013. **53**(3-4): p. 207-216.
48. Martin, A.C., *Mapping PDB chains to UniProtKB entries.* Bioinformatics, 2005. **21**(23): p. 4297-301.
49. Minn, A.J., et al., *Genes that mediate breast cancer metastasis to lung.* Nature, 2005. **436**(7050): p. 518-524.
50. Bos, P.D., et al., *Genes that mediate breast cancer metastasis to the brain.* Nature, 2009. **459**(7249): p. 1005-U137.
51. Kanehisa, M. and S. Goto, *KEGG: Kyoto Encyclopedia of Genes and Genomes.* Nucleic Acids Research, 2000. **28**(1): p. 27-30.
52. Gregory, A.D. and A.M. Houghton, *Tumor-Associated Neutrophils: New Targets for Cancer Therapy.* Cancer Research, 2011. **71**(7): p. 2411-2416.

53. 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. 260-264.

54. Network, C.G.A.R., *Comprehensive genomic characterization of squamous cell lung cancers The Cancer Genome Atlas Research Network (vol 489, pg 519, 2012).* Nature, 2012. **491**(7423): p. 288-288.

55. Sen Kilic, E., *Network Analyses to Identify Candidate Proteins and Interactions Responsible for Breast Cancer Lung and Brain Metastasis Differentiation*, doi: 10.13140/ RG.2.1.2840.8726 (2015).