STRUCTURAL MAPPING AND NETWORK ANALYSIS OF PATIENT-SPECIFIC MUTATIONS IN GLIOBLASTOMA

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS INSTITUTE
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

TUĞBA KAYA

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF
MASTER OF SCIENCE
IN
BIOINFORMATICS

AUGUST 2018

Approval of the thesis:

STRUCTURAL MAPPING AND NETWORK ANALYSIS OF PATIENT-SPECIFIC MUTATIONS IN GLIOBLASTOMA

Submitted by TUĞBA KAYA in partial fulfillment of the requirements for the degree of **Master of Science in the Department of Bioinformatics, Middle East Technical University** by,

Prof. Dr. Deniz Zeyrek Bozşahin
Dean, **Graduate School of Informatics, METU** _____

Assoc. Prof. Dr. Yeşim Aydın Son
Head of Department, **Health Informatics, METU** _____

Assoc. Prof. Dr. Nurcan Tunçbağ
Supervisor, **Health Informatics, METU** _____

Assoc. Prof. Dr. Tunca Doğan
Co-supervisor, **Health Informatics, METU** _____

Examining Committee Members:

Prof. Dr. Tolga Can
Computer Engineering, METU _____

Assoc. Prof. Dr. Nurcan Tunçbağ
Health Informatics, METU _____

Assoc. Prof. Dr. Yeşim Aydın Son
Health Informatics, METU _____

Assoc. Prof. Dr. Özlen Konu
Molecular Biology and Genetics, Bilkent University _____

Assist. Prof. Dr. Ceren Sucularlı
Graduate School of Health Sciences, Hacettepe University _____

Date: _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: TUĞBA KAYA

Signature :

ABSTRACT

STRUCTURAL MAPPING AND NETWORK ANALYSIS OF PATIENT-SPECIFIC MUTATIONS IN GLIOBLASTOMA

Kaya, Tuğba

MSc, Department of Bioinformatics

Supervisor : Assoc. Prof. Dr. Nurcan Tunçbağ

Co-Supervisor : Assoc. Prof. Dr. Tunca Doğan

August 2018, 76 pages

Cancer is one of the most common cause of death worldwide. It occurs as a result of a collection of somatic deviations from normal state. Therefore, many efforts have been invested to profile mutations in different types of tumors; such as, the Cancer Genome Atlas (TCGA) which deposits multiple omic data for more than 11,000 tumor samples. In this thesis, we present a pipeline which retrieves patient-specific mutation data in Glioblastoma from TCGA, maps these mutations on the protein structures in Protein Databank (PDB) and finds the location and functional effect of the mutations and reconstruct functional networks by integrating mutation data with interactome. As a result of this thesis study, we found that some mutations are specific to alternative isoform sequence of the protein instead of the canonical sequence. We also showed that functional impact of mutations in interface region is more damaging compared to the surface region and more similar to the core region of the protein. We showed that most common change in the protein core is that hydrophobic residues are mutated to another hydrophobic residue. However, in the surface or interface region a charged residue is changed either to another charged residue or a polar residue when we analyzed the chemical classes of mutations. From these mutation profiles of the patients, we reconstructed 290 GBM-specific networks with Omics Integrator which solves the prize-collecting Steiner forest (PCSF) problem and optimally connects the given set of proteins in a network context. We merged the most common nodes and

edges across these patients and clustered the merged network into functional communities. The ontology and pathway enrichment analyses gave us that Wnt signaling, ERBB signaling and NfKb/Ikb signaling pathways are the most commonly enriched pathways. From mutation to protein structures and functional networks, we believe that the result of this thesis will have significant contribution in cancer research.

Keywords: Network Modeling, Structural Mapping, Missense Mutation, Cancer, Glioblastoma Multiforme

ÖZ

GLİOBLASTOMADA HASTAYA ÖZGÜ MUTASYONLARIN YAPISAL HARİTALANMASI VE AĞ ANALİZİ

Kaya, Tuğba

Yüksek Lisans, Biyoenformatik Bölümü

Tez Yöneticisi : Doç. Dr. Nurcan Tunçbağ

Ortak Tez Yöneticisi : Doç. Dr. Tunca Doğan

Ağustos 2018 , 76 sayfa

Kanser dünya çapında en yaygın ölüm nedenlerinden biridir. Bu nedenle, farklı tipteki tümörlerin mutasyon profillerinin çıkarılmasına çokça çaba harcanmıştır; örneğin, Kanser Genom Atlas (TCGA) 11000'den fazla tümör örneği için omik veri biriktiren platformlardan biridir. Bu tezde, TCGA'den Glioblastoma için hastaya özgü mutasyon verilerini alan, Protein Databank (PDB) 'den elde edilen protein yapıları üzerine haritalayan, yerini ve fonksiyonel etkisini tespit eden ve protein interaktomu ile entegre ederek fonksiyonel ağları yeniden yapılandıran bir veri işleme hattı sunmaktayız. Bu tez sonucunda, bazı mutasyonların, proteinin bilinen dizisi yerine alternatif izoform dizisine spesifik olduğunu bulduk. Aynı zamanda, etkileşim yüzeyindeki mutasyonların fonksiyonel etkisinin iç bölgesindeki etkilere benzer olduğunu ve yüzey bölgesi ile karşılaştırıldığında daha çok zarar verdiğini gösterdik. Protein iç kısmındaki en yaygın değişimin, hidrofobik amino asitlerin başka bir hidrofobik amino aside değişiminin olduğunu gösterdik. Bununla birlikte, yüzey veya etkileşim yüzeyinde, mutasyonların kimyasal sınıflarını analiz ettiğimizde yüklü bir amino asit başka bir yüklü ya da polar amino aside dönüşmüştür. TCGAdeki hastaların bu mutasyon profillerinden, prize collecting Steiner forest (PCSF) problemini çözen ve verilen bir protein kümesini bir ağ bağlamında en iyi şekilde birleştiren Omics Integrator ile GBMda hastaya özgü 290 tane ağı yeniden oluşturduk. Bu hastalar arasında en yaygın protein düğümleri ve ayrıtlarını birleştirdik ve bu birleştirilmiş ağı işlevsel topluluk-

lar halinde kümeledik. Ontoloji ve yolak analizleri Wnt sinyal ERBB sinyal ve NfKb / Ikb sinyal yolaklarının en yaygın zenginleştirilmiş yolaklar olduğunu göstermiştir. Mutasyondan protein yapılarına ve fonksiyonel ağlara kadar, bu tez çalışmasının sonucunun kanser araştırmalarına önemli katkı sağlayacağına inanıyoruz.

Anahtar Kelimeler: Ağ Modelleme, Yapısal Haritalama, Mutasyon, Kanser,Glioblastoma Multiforme

To my family

ACKNOWLEDGMENTS

First and foremost, I would like to wholeheartedly thank my excellent supervisor Assoc. Prof. Dr. Nurcan Tunçbağ for her endless patience, valuable advice, continuous encouragement and the immeasurable amount of guidance in this thesis. It would not have been possible without her efforts which I am, and always will be, gratefully indebted for. She has been my invaluable mentor, my *Yoda*, and the one who enlightened the way in the darkest times.

I would like to also thank my co-supervisor respectable Assoc. Prof. Dr. Tunca Doğan for his guidance and sharing his experience. I would like to show my sincere gratitude to my thesis committee members Prof. Dr. Tolga Can, Assoc. Prof. Dr. Yeşim Aydın Son, Assoc. Prof. Dr. Özlen Konu and Assist. Prof. Dr. Ceren Sucularlı for taking their time to read and providing insightful comments. In addition, I would like to express my gratitude to my lab mates from Network Modeling Lab. for their feedback, cooperation and of course friendship.

I would like to thank my sister, Sibel Kutayli, who have provided me with moral and emotional support throughout this study and my whole life. I would like to also thank my dearest friends Ajdan Küçükçiftçi, Alper Çevik, Güçlü Ongun, Mustafa Ciminli and Emre Kağan Akkaya for their unfailing support. They have always believed in me.

Finally, and most importantly, I would like to thank my family, my dear father Yunus Kaya, my dear mother Gülen Kaya and my little brother Emre Kaya for their unconditional love. This accomplishment would not have been possible without them.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

ASA	Accessible Surface Area
CGP	Cancer Genome Project
EMBL-EBI	European Bioinformatics Institute
GBM	Glioblastoma Multiforme
GO	Gene Ontology
ICGC	The International Cancer Genome Consortium
KEGG	Kyoto Encyclopedia of Genes and Genomes
MAF	Mutation Annotation Format
NCI	National Cancer Institute
NHGRI	National Human Genome Research Institute
PCSF	Prize Collecting Steiner Forest
PCST	Prize Collecting Steiner Tree
PDB	Protein Databank
PDBe	Macromolecular Structure Database at the European Bioinformatics Institute
PDBj	Protein Data Bank Japan
PIR	Protein Information Resource
RCSB	Research Collaboratory for Structural Bioinformatics
SIB	Swiss Institute of Bioinformatics
TCGA	The Cancer Genome Atlas
WHO	World Health Organization

CHAPTER 1

INTRODUCTION

Cancer is a generic term for diseases where cells grow and divide uncontrollably and these cells proliferate abnormally. Internal factors (mutations, hormones immune conditions etc.) and external factors (tobacco smoke, radiation, chemical etc.) can cause cancer. As reported by the World Health Organization (WHO) fact sheet, cancer is the second most common cause of death worldwide. In 2015, 8 million deaths were caused by cancer. Globally, approximately 1 in 6 deaths is because of cancer.

From a point of cancer genetics, it occurs as a result of a collection of somatic deviations from normal state. The thought about cancer to be a process of the accumulation of mutations has been changed recently with the developments and improvements in sequencing technologies and eventually accumulated sequencing data. Instead of the linear accumulation of mutations, there is rather a more complex arrangements of mutations in cancerous cells. In the recent years, tumor samples from thousands of cancer patients have been collected, and genomic and proteomic changes have been analyzed in a tumor-specific way for tens of different tumor types. One of these efforts is the Cancer Genome Atlas (TCGA) which catalogs 33 cancer types (including Glioblastoma, ovarian cancer, breast cancer and many more and 10 rare tumors) from around 11,000 patients with both normal and cancer tissues [Tomczak et al., 2015]. In TCGA, many high-throughput data types including genomic, epigenomic, and transcriptomic profiles as well as the clinical data of each patient are deposited. The genomic data contains tumor specific mutation profiles and copy number variations. In terms of mutations, there are 617,354 somatic mutations in total in TCGA which are our target in this thesis study [Kandoth et al., 2013].

The pattern of somatic changes in DNA is different in each type of cancer. Especially, Glioblastoma Multiforme (GBM), which is the most common malignant brain tumor and highly deadly with 12-15 months survival and very heterogeneous in terms of mutation profiles. The effect of each somatic mutation is not uniform in the involvement to the development of cancer. Conventionally, mutations are divided into two types; driver and passenger mutations. The former contribute to the growth process of the tumor by providing selective growth advantage to the cancer cells while the latter are silent and co-player in the tumor development.

In the context of mutation, replacement of one or more nucleotides with other nucleotides of the same number is called substitution. Depending on the resulting amino acid in the encoded protein, substitution is divided into three categories; silent, missense and nonsense. Silent mutation does not cause any amino acid change in the encoded protein. The missense mutation causes the amino acid change in the en-

coded protein. The nonsense mutation results in an inserted stop codon to encoded protein and the translation process stops earlier. Missense and nonsense mutations are in the class of amino acid changing (aa-changing) mutations. Therefore, at the protein level, missense and nonsense mutations may have functional impact; however, they may also be neutral in terms of functional effect. Current research direction in the field is toward elucidating the functionally critical mutations in cancer.

Three-dimensional information of proteins and their complex states are deposited in Protein Databank (PDB). Although not all proteins have a complete known structure in PDB, the yearly growth rate of the database is exponential. The homology modeling and threading based techniques can also accurately model a given protein with an unknown structure. Proteins are not functional on their own; rather they interact with other proteins to be get functional. Proteins interact through their surfaces called interface region. Interface regions are different than the rest of the surface in terms of conservation, amino acid preferences and solvent accessibility. Many studies have shown that the characteristics of protein interfaces, surfaces and core regions are different. Two key property of the interfaces are the shape complementarity and chemical complementarity. Any alteration affecting the global structure of the protein or the complementarity in the interface may result in the loss of the interaction, or an alteration changing the physicochemical properties of the surface region may lead to an interaction-gain. Therefore, the location of the aa-changing mutations are important in the analysis of patient-specific networks. While some of these mutations are located distantly to the functional site of the protein and have mild effect in the functional change, some others located in the core region of the protein and significantly changes the protein structure. Mutations occurring in the interface regions may change the interaction preferences of proteins and are expected to have more deleterious effect in protein function. In these cases, it has been observed that the proteins either lose the ability to bind a protein or gain a capability of binding to the new proteins. In addition, mutations located in the core region may affect the stability of the structure and eventually changes its interaction preference as well. These changes caused by mutations in the binding patterns of proteins may cause diseases. Therefore, it is important to analyze at which region mutations are located in the three-dimensional structure of proteins [Nishi et al., 2013].

In another aspect, mutations can change the functionality of the cellular pathways by altering the interaction preferences or the stability of the proteins. To analyze the changes at pathway level, many network reconstruction methods are available which aims to optimally connect the given set of proteins in a network context. Some examples are Omics Integrator, Modulomics, HotNet and Paradigm. These tools and softwares aim to construct the optimal network that would represent the given data best. The constructed networks usually reveal the hidden components of the pathways and show the interplay between pathways beyond the list of mutated proteins.

Given the number of mutations deposited in TCGA, the number of structures in PDB and the number known interactions between proteins, computational approaches are crucial for a system level analysis of the mutation effects on proteins, protein interactions and functional pathways in a patient-specific way. Therefore, we designed this thesis study to apply an integrative approach to a given set of mutations and to functionally analyze their effects. In this dissertation, we use the patient-specific mutation data from GBM tumors deposited in TCGA and search for their effects both

at protein and network level. For this purpose, we developed a pipeline that takes mutation set as input and analyzes these mutations to find out potential proteins that mutations can take place and use these extracted proteins to map mutations to 3D protein structures, then apply the network reconstruction and detailed evolutionary mutation effect analyzes. This pipeline helps to figure out effects of mutations to protein structures which is an essential step to inspect malfunctioning proteins and their related interactions and pathways.

Chapter 2 includes corresponding work related to mutations in cancer, how they are classified and how their effects are analyzed, and which tools are used for this purpose. We also reviewed topics related to proteins; how they interact, the publicly available sources about protein structure, sequence and functional information. In addition to this, we also give brief information about integrative network modeling approaches.

In chapter 3, we present our pipeline for this study that includes, dataset description, filters that are applied to dataset, formation of processed dictionaries that are excluded from public databases, foundation of potential mutated proteins, methods for mapping mutations to 3D structures, methods for analyzing mutations in network level.

In chapter 4, the results our pipeline are described. As a result of mapping mutations to 3D structures of proteins, possible effects of mutations depending to the mapped region of proteins are analyzed. The significance of mutations in related pathways are also shown in this chapter. With the help of network analysis, the crucial interactions of mutated proteins in the interactome is extracted. The effect of mutations of this interactions is inspected.

We conclude this thesis with a general overview of our study, the discussion of the results and our plans as a future work. We would like to extend the application of this pipeline to other cancer types. Additionally, we only used the available experimental structural data deposited in PDB. Therefore, our future aim is to enrich the structural dataset of proteins with homology models and structurally predicted protein interactions. We believe that this study provides another perspective to the analysis of mutation effects and a good training towards the precision medicine.

CHAPTER 2

LITERATURE REVIEW

In this chapter, we review the available studies on mutation characteristics in cancer, their related protein features and interactions, mutation effect analysis tools, platforms and databases that are used in cancer researches and network modeling approaches. Firstly, platforms that provide cancer related data and protein feature, structure databases are explained. Then, the characteristics and classifications of mutations are detailed. After that, tools that are used to inspect effects of mutations on protein structure are reviewed. Finally, the approaches used in integrative network modeling are reviewed.

2.1 Catalogs of Genomic Alterations

2.1.1 The Cancer Genome Atlas (TCGA)

TCGA is a public platform that is created with the collaboration of National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI). TCGA contains genomic alterations of 33 cancer types from tumors and normal tissues of more than 11,000 patients. The goal of the TCGA project is to create an index of genetic alterations that cause cancer. This index aims to contribute to the improvement of cancer diagnosis and treatment methods by making wide range of analyzed cancer data public. TCGA process through the co-operation of a number of centers in an organized way. In this process, tissue samples taken from cancer patients are passed through analyses and quality controls. Clinical data obtained from the controlled samples are uploaded to the repository for further genomic and high-throughput analysis. TCGA conducts a comprehensive analysis of patient data high-throughput technologies such as microarray-based and next-generation sequencing technologies [Tomczak et al., 2015]. These methods can be listed as, RNA sequencing (RNAseq), MicroRNA sequencing (miRNAseq), DNA sequencing (DNAseq), SNP based platforms, array based DNA methylation sequencing and reverse phase protein array (RPPA). RNAseq is a high-throughput technology for total RNA profiling and it is a fast method for identifying and quantifying transcripts, isoforms, novel transcripts, non-coding RNAs. On the other hand, miRNAseq is a form of RNAseq and it is used for the detection of short small RNA sequences (miRNAs) that are responsible for regulation of genes in signaling pathways. DNAseq method is used to find out DNA molecule sequence and inspect the alterations in DNA sequence such as insertion, deletion, polymorphism, copy number variation. SNP based platforms pro-

vide the analysis of variations in cancer. Array-based DNA methylation sequencing is a method for profiling DNA methylation of CpG sites that are the most common alterations in cancer. Reverse-phase protein array is a proteomic method for measuring protein expression levels. These methods provide many type of data such as exon expression, copy number variation, DNA methylation, protein expression, single nucleotide polymorphism. In TCGA, there are four levels of data. Level 1 is the raw data, Level 2 is the processed data, Level 3 is the interpreted data for each individual sample. While Level 3 is publicly available, Level 1 and Level 2 require permission.

2.1.2 International Cancer Genome Consortium (ICGC)

The International Cancer Genome Consortium (ICGC) is a volunteer organization that aims to launch and coordinate genomic researches through 50 cancer types and/or subtypes with global importance. The ICGC was started in 2008 and main focus of ICGC is to generate an atlas of genomic alterations in cancer for improvement of prognosis, diagnostics and treatment. The ICGC goal is to provide full catalog of somatic mutation data in high quality, high resolution and high coverage. To advance researches, the ICGC aims to make available data to community in minimum time with minimum restrictions. ICGC members agree the ICGS's policies that request rapid data release with specified data standards. These policies include data release policy, publication policy, intellectual policy, quality standards etc. The main objective of these policies is to maximize the benefit of the community without violating the personal rights of the donors [Zhang et al., 2011, Consortium et al., 2010].

2.1.3 Cancer Genome Project (CGP)

The Cancer Genome Project was carried out by Sanger Center in England in 2000. The main aim of this project is to determine mutated genes in tumors and the patterns of these mutations by using latest DNA sequencing methods. For this purpose, using the Human Genome Project data, normal and cancerous cells were compared and tumor mutations were detected [Dickson, 1999].

2.2 Mutations in Cancer and Their Classification

DNA variation is the change in DNA sequence. The criterion for evaluating a DNA variant as mutation or polymorphism is the prevalence of the variation in the population. If an alteration has 1% or more frequency in the population, it is classified as polymorphism. If the frequency is less, it is classified as a mutation [Widłak, 2013, p. 59].

Mutations are changes in several nucleobases or chromosome scales that occur permanently in the DNA sequence. Mutations can be separated into two groups according to the type of the cells; somatic and germline mutations. Germline mutation occurs in the organism's sex cells and can be transferred to the offspring of the organism. On the other hand, somatic mutation is a mutation that occurs in the cells of

the organism outside the sex cells. While this mutation is transferred to the daughter of the cell, it can not be transferred to the offspring of the organism [Widłak, 2013, p. 56-57].

Mutations can also be categorized as functional and non-functional mutations. Functional mutations are defined as mutations that change the function of a protein. This change can be gaining a new functionality or losing a functionality. On the other hand, non-functional mutations have no effect on the functionality of the protein. Driver versus passenger and functional versus nonfunctional mutation categories are determined based on different properties. While categorization of driver and passenger mutations is done by assessing tumor growth contribution of mutation, functional and non-functional mutations are categorized by observing the effect of mutation on protein function [Gonzalez-Perez et al., 2013].

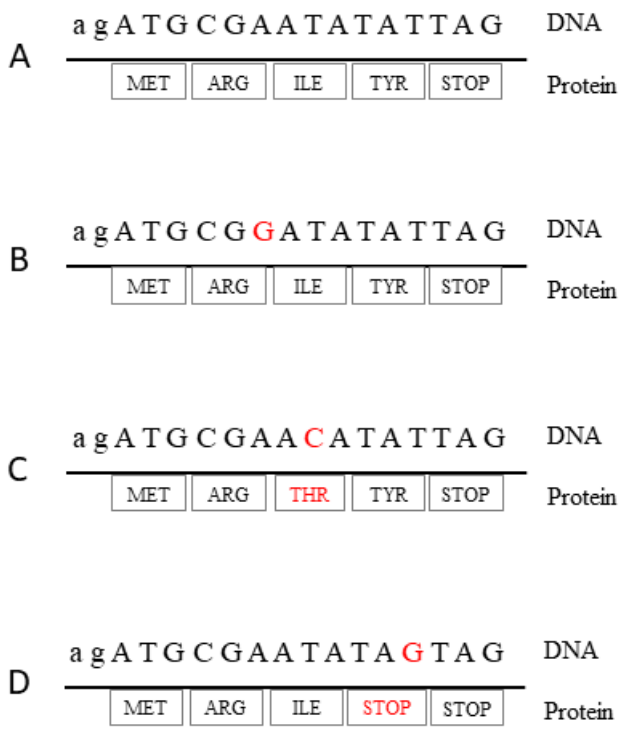

Figure 2.1: Visual representation of single point mutations. A) Wild type DNA and its encoded protein sequence. B) Silent mutation in DNA and its encoded protein sequence. C) Missense mutation in DNA and its encoded protein sequence. D) Nonsense mutation in DNA and its encoded protein sequence.

Cancer occurs when somatic mutations accumulate in the cell and it changes the structure and properties of the cell. But that does not mean that every somatic mutation causes cancer. The main challenge of research is to detect driver and passenger mutations. While driver mutations provide positive growth advantages to can-

cer cells, passenger mutations do not provide selective advantages to cancer cells [Watson et al., 2013]. On the other hand, it is necessary to examine the structural changes in order to be able to examine the effects of mutations on behavior in the system.

Replacement of one or more nucleotides with other nucleotides of the same number is called substitution. Depending on the results on the encoded protein, substitution is divided into three categories; Silent, Missense and Nonsense. An example of this type of mutations is shown in Figure 2.1. Silent mutation does not cause any amino acid change in the encoded protein. The missense mutation causes the amino acid change in the encoded protein. The nonsense mutation results in inserted stop codon to encoded protein [Widłak, 2013, p. 57].

2.3 UNIPROT

The main objective of UniProt is to provide comprehensive, stable and centralized repository of protein knowledge. UniProt is a joint project between European Bioinformatics Institute (EMBL-EBI), Swiss Institute of Bioinformatics (SIB) and Protein Information Resource (PIR) [Watson et al., 2013]. UniProt is a freely accessible database of protein sequences and related annotation. It has four components which differ in usage. The UniProt knowledgebase (UniProtKB) is composed of two sets of sequences; UniProt/SwissProt and UniProt/TrEMBL. The former is an experimentally curated database and has cross reference to multiple database and the latter is automatically annotated protein sequence database which is not curated. The UniProt Archive (UniParc) stores all sequences of proteins through history. On the other hand, UniProt Reference Clusters (UniRef) provides a clustered set of proteins based on similarity in sequence. UniProtKB contains protein related features including function, interaction, structure, family and domains, sequences, expression and cross references. As of July 2018, there are 152,938 unreviewed protein sequences in TrEMBL and 20,386 reviewed protein sequences for human in UniProtKB/SwissProt. UniProt website performs text based search which allow researchers to obtain data without prior knowledge. UniProt also has query tools which include full text search, field based text search, batch retrieval, database identifier mapping, sequence similarity search and multiple sequence alignment. UniProt data can be downloaded through website or UniProt FTP server in various formats such as plain text, XML, Fasta etc. [Consortium, 2011].

2.4 Protein Structures and Protein Databank

The Protein Data Bank (PDB) was established in 1971 to catalog 3D structures of macromolecules that are experimentally stated. In 2003, The Research Collaboratory for Structural Bioinformatics (RCSB), Protein Data Bank Japan (PDBj) and Macromolecular Structure Database at the European Bioinformatics Institute (PDBe) have become three main centres of Worldwide PDB (wwPDB) [Berman et al., 2006]. PDB stores 3D structures of molecules in many dimensions from small protein fragments to large virus structures that are obtained by experimental method such as

electron microscopy (EM), X-ray crystallography (X-ray) and nuclear magnetic resonance (NMR). In total, approximately 90% of of all structures deposited in PDB are resolved by X-ray experiment which is the most common method,. Structural information is available for many type of molecules including proteins, ribosomes, drug targets, viruses, nucleic acids. There are 142,015 structures in total in PDB. When we refer to more statistics, we see that PDB contains 44,394 distinct protein structures, 38,787 structures of human sequences and 10,107 nucleic acid containing structures. Overall yearly growth rate of the number of released structures is exponential. In 2017, 11,115 structures are released and total structure number reached 136,413. While 10,116 of the released structures in 2017 gathered by X-ray crystallography, 416 of them are gathered by NMR and others are analyzed with electron microscopy, hybrid etc. When the data distribution according to source organism inspected, the majority of data belongs to *Homo sapiens* and *Bos taurus*. Approximately 17% (21818/127524) of the PDB structures have 1.8-2.0 Å resolution, while approximately 15% (19404/127524) of them have 2.0-2.2 Å resolution.

PDB entries represented with 4-character unique identifier starting with a number between 1-9 and remaining three character can be either numeral or letter. Protein structures are publicly available in PDB standard file format. The PDB file format consist of two sections; header section gives details concerning name, author, sequence, citation, secondary structure etc. and coordinate section describes atomic coordinates in Angstrom units, chain identifiers, position identifiers and atomic names [Dutta et al., 2009]. The PDB data can be downloaded from the website or through the FTP server in PDB, mmCIF and XML format. The distribution of number of entries in PDB according to the number of residues in each protein is shown in Figure 2.2. Very large protein structures are less common in PDB compared mid-sized proteins because of the constraints of crystallization and NMR techniques.

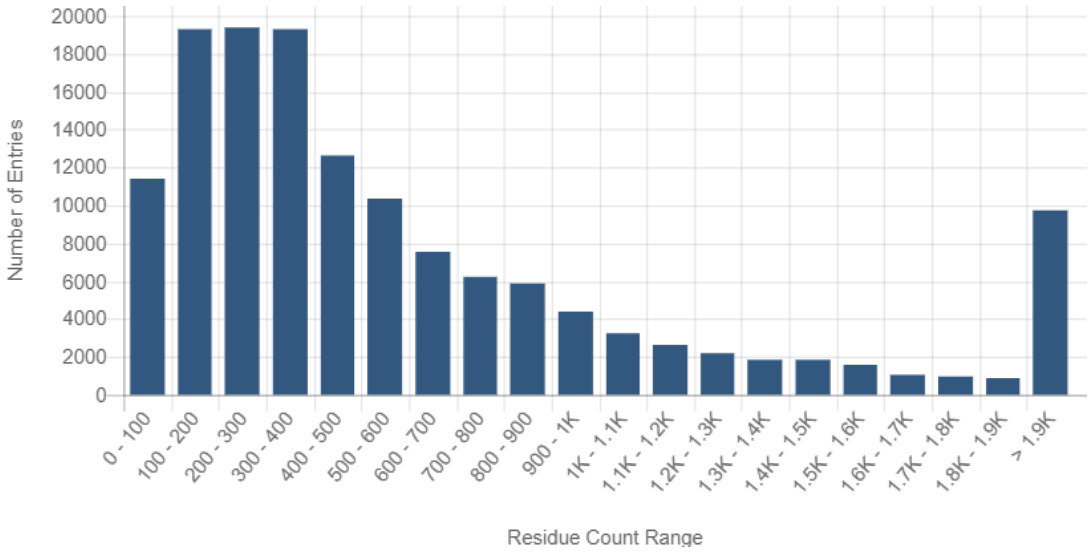

Figure 2.2: The distribution of number of entries in PDB according to the number of residues in each protein. (This histogram retrieved from PDB web site) [Berman et al., 2000].

2.5 Protein-Protein Interactions (PPIs) and PPI Databases

More than 80% of proteins interact with each other to perform their function in signaling pathways, replication, cell to cell communication, transcription. Damage in proteins can result in disruptions in these processes and can cause diseases. Interactome is the complete set of protein protein interactions in an organism. Representing the full interactome may be impossible because of the complexity. Post-transcriptional modifications, cellular localization, tissue specificity increases complexity. On the other hand, some proteins interact permanently, some proteins needs phosphorylation to interact their partners and protein interactions may be different in different cell types. This heterogeneity makes it difficult to get full interactome. Experimental methods can be insufficient to reveal the whole interactome for this reason, computational methods are also used to predict protein protein interactions.

There are low throughput and high throughput techniques to detect pairwise protein-protein interactions. One of the most frequently used method is yeast-two-hybrid (Y2H) system which is a high throughput technique used in vivo. However, this method may produce false positive interactions besides the true positives. For example, two proteins that exist in separate parts in cell in different time may be marked as interacting with this method, because they can physically bind to each other. In addition to this, observing interactions that occur after post-transcriptional modifications is not possible with this method.

Besides the experimental methods, there are many prediction approaches to accurately identify protein-protein interactions. Sequence-based approaches can be classified as machine learning based approaches, interolog search, domain co-occurrence and gene/domain fusion events. More advance techniques such docking, knowledge-base prediction use structural information which gives also residue level binding regions. All techniques are detailed and reviewed in [Keskin et al., 2016].

As the experimental and predicted interactions in recent studies grow the necessity for databases to store these information has emerged. There are many such type of databases storing different aspects of PPIs. BIND is a database for storing biomolecular interactions, molecular complexes and pathways [Bader et al., 2003]. BioGRID deposits protein and genetic interactions over 116,000 from many species including *Homo sapiens* [Stark et al., 2006]. CORUM also stores manually curated protein complexes from human, mouse and rat [Ruepp et al., 2009]. HPDR database is designed to collect experimentally curated proteins and its features including interaction, post-translational modifications, enzyme/substrate relationship only for human [Keshava Prasad et al., 2008]. MINT database stores experimentally verified molecular interactions especially PPI [Chatr-Aryamontri et al., 2006]. Different than these single source databases, there are several others those integrate the interactions in these databases and scores them based on evidences from multiple sources. iRefWeb is a unified database of 10 public databases. The databases are BIND, BioGRID, CORUM, DIP, HPDR, MPact, OPHID, and IntAct. The aim of the unified iRefWeb database is to provide comparison between result of multiple interaction databases. The difference of iRefWeb from other listed databases is that it integrates and scores the interactions retrieved from different databases [Turner et al., 2010]. STRING database also stores interactions in a similar fashion that based on the inter-

action source, i.e. experimental, database search, text-mining, co-expression, and it scores each interaction between 0 to 1 [Szklarczyk et al., 2016].

Table 2.1: List of protein-protein interaction databases.

Name	**Organism**	**Number of Interactions**	**Number of Proteins**
BioGRID `https://thebiogrid.org/`	Homo sapiens	332,829	22,792
HPDR `http://www.hprd.org/`	Homo sapiens	41,327	30,047
MINT `https://mint.bio.uniroma2.it/`	208 organisms	57,001	13,196
STRING `https://string-db.org/`	2,031 organisms	1,380M	9.6M
IntAct `https://www.ebi.ac.uk/intact/`	> 9 organisms	851,299	107,104
iRefWeb `http://wodaklab.org/iRefWeb/`	1,448 organisms	263,479	66,701

2.6 Properties of PPIs

According to type of proteins in a complex, complexes can be divided into two groups; homo-oligomeric and hetero-oligomeric. Homo oligomeric complexes are made of identical protein, whereas hetero-oligomeric proteins are made of different proteins. Complexes can be group into obligate and nonobligate complexes. For this classification can be done, affinity and stability of proteins that formed the complex must be examined. If proteins that located in complex is unstable on their own, this complex is obligate. On the other hand, if proteins in the complex are stable own their own, this is a nonobligate complex. Non-obligate complexes can be grouped according to lifetime of complex; permanent and transient. Permanent interactions are stable and proteins stay together permanently. Transient interactions are not stable, proteins interact temporarily. This interaction takes place in signaling pathways for transmitting the signal. This classifications can be summarized as, obligate complexes are permanent, non obligate complexes can be transient or permanent. The forces contributing into protein-protein interactions are electrostatic interactions, hydrophobic interactions, hydrogen bonds and salt bridges. For example, hydrophobic interactions are more common in obligate complexes. On the other hand, salt bridges and hydrogen bonds are more common in transient complexes.

Proteins interact each other with their interfaces. Protein interfaces are determined with several methods. First method is the calculating the accessible surface area (ASA) of the residues. In this method, ASA of a residue in the complex state is compared to ASA of the residue in monomer state. If the difference - in other words the ASA lost after the complex formed between two proteins - is greater than 1Å^2, this residue is marked as interface residue. Second method is the calcu-

lating atomic distances of residues each from one chain. In this method, the distance between two residues in different chains in the complex is calculated and if the distance is less than a threshold, then, these two residues are marked as interacting. This threshold is conventionally used as 5Å; however, 6Å or 7Å are also used as thresholds. In some studies, the threshold is variable according to the van der Waals radii of the contacting atoms for a more precise calculation (used in [Keskin et al., 2004, Tuncbag et al., 2008]). Interfaces are generally named with their PDB IDs and the chain names forming the interface. An example of protein interface is shown in Figure 2.3 where Nfkb (colored white, chain A) and Ikb (colored gray, chain D) are interacting through the interface drawn in surface representation. The name of the interface is 1iknAD. Cyan surface is the interface partner from Nfkb and pink surface is the interface partner from Ikb protein.

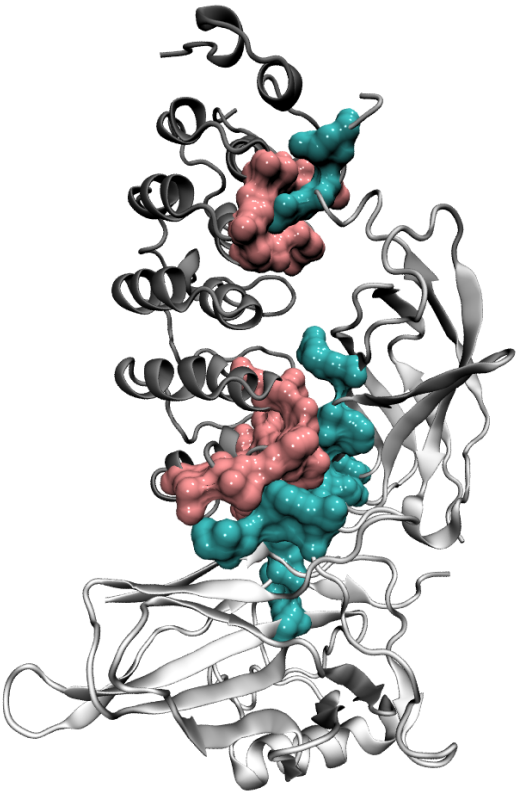

Figure 2.3: An interface illustration. The interface between NF-KAPPA-B P65 SUBUNIT (chain A) and I-KAPPA-B-ALPHA (chain D) is calculated with distance-based approach and its corresponding structure is deposited with '1ikn' identifier in PDB.

Binding sites have many physicochemical properties like hydrogen bonds, the charge distribution, composition of the interface, strength of the interaction, hot spots, shape of the interface, residue conservation. Energy distribution is one of the main property of the binding sites. Hot spots are the residues in the interfaces that are energetically important. This residues can be found by experimental methods like alanine scanning mutagenesis but this method cannot be applied to all structures because of the lack of data. Therefore, computational methods such as learning based methods and

molecular dynamics based methods are used for hot spot prediction.

2.7 Tools to Analyze of the Mutation Effect on Protein Stability and Protein Interactions

Disease causing variants mostly occur in protein coding regions. However dbSNP and 1000 Genomes databases are used to categorize variants, they are not sufficient for predicting variants that affect protein function. The questions how a non-synonymous mutation affect the stability of the protein or how it changes the interaction preferences of the proteins are not yet fully answered. Nevertheless, there are many methods that are using conservation, co-evolution information or energy calculations to classify the mutations based on their effects; damaging/deleterious or benign/neutral. In the following section, we review a selected set of methods designed to predict mutation effects.

2.7.1 SIFT

The Sorting Intolerant from Tolerant (SIFT) algorithm a tool for predicting the effect of a variant on protein function. It is first released in 2001 as a website. SIFT mainly focuses on the measuring the effect of amino acid substitutions on protein function and recently, predicting the effect of frameshifting indels feature added to the SIFT tool. It is used in researches about genetic diseases and infectious diseases. SWIFT uses sequence homology for predicting the variant effect and it run on the assumption that evolutionary conserved regions are less tolerant to mutations. Therefore, mutations that occurs in evolutionary conserved regions are more likely to affect the function. First step of the SIFT pipeline is querying the protein through protein databases to obtain sequence. After this step, protein sequence aligned and according to given mutation position, and SIFT analyzes the amino acid composition and calculates the score. The score is a normalized value which ranges between 0 and 1 and it is the value of probability of existence of a given amino acid change at that position. If the calculated score is less than 0.05, then the given change is found to have damaging effect in protein function [Sim et al., 2012].

2.7.2 PolyPhen-2

Polymorphism Phenotyping v2 (PolyPhen-2) is a tool for predicting effect of amino acid substitution on protein structure and protein function. PolyPhen-2 tool is available as a software and via web server. The PolyPhen-2 pipeline maps SNPs to gene transcripts and analyzes sequences and structure attributes. With these annotations, it generates conservation profiles. Then, analyzed properties are used in machine learning classification to predict harmful effect of amino acid effect on protein structure and function. In the PolyPhen-2 tool, input can be protein or SNP identifier. It also provide batch query option. In the tool output substitution can be categorized as probably damaging, possibly damaging and benign. These categorization is done according to sensitivity and specificity. “Probably damaging” refers to damaging with

high confidence and "Benign" refers to benign with high confidence. On the other hand, "Possibly damaging" means damaging with low confidence. PolyPhen-2 also provides a damaging prability score that ranges between 0 to 1 and 1 referring to damaging. PolyPhen-2 output also provides multiple sequence alignment and 3D-structure protein viewer [Adzhubei et al., 2013].

2.7.3 EVmutate

EVmutate is a statistical method for predicting mutation effect and it considers dependencies between positions. EVmutation provides precomputed dataset for approximately 700 human proteins on its website and it can be used to calculate mutations affect in any organism. Evmutation tool calculates changes in statistical energy (ΔE) using multiple sequence alignment of the protein family. Its algorithm combines two factor to calculate quantitative mutations affect: the interactions between mutations and sequence background [Hopf et al., 2017].

2.7.4 MutaBind

MutaBind is a computational method on a web server that analyzes effects of mutations on binding affinity of proteins. MutaBind server calculates quantitative changes in binding affinity by mapping mutations to protein 3D structures. After that, it inspects the corruptive effects of mutations with confidence level of this estimation. In order to run MutaBind tool, 3D structure of protein complexes must be available. It can take protein PDB code as an input and gathers the structure from Protein Data Bank or user can upload their own structure file. The selected structure files in either case must have at least two protein chains. Additional parameters like interaction partners, chains of partners must be determined by the user. After these, user must provide the mutations which can be at most 16 at one submission. With all these parameters, MutaBind server analyzes the complex and evaluates output for each individual mutations separately. The outputs for one mutation includes, estimated change in binding affinity, whether the mutation on interface or not, whether the mutation deleterious or not and confidence level. Estimated change in binding affinity can be positive and negative. Positive value of $\Delta\Delta G$ refers destabilizing mutation that results decrease in binding affinity. Tool decides a mutation deleterious or not by $\Delta\Delta G$ score and marks mutation as deleterious if $\Delta\Delta G$ is higher and equal to 1.57 kcal mol^{-1} [Li et al., 2016].

2.8 Integrative Network Modeling

The response to an external stimulus results in alterations in cell signaling pathways and gene regulatory networks. Recently, high throughput technologies allow us to discover many molecular processes within the cell. Although high throughput omics data obtained at a specific state of the cells enable us to observe the changes in cellular response, not all components within the response pathways are revealed, but

many remain hidden. These hidden components might be driver proteins connecting significantly changing gene/protein at the given condition or a transcription factor regulating differentially expressed genes [Huang and Fraenkel, 2009]. Unfortunately, no single dataset provide all data about molecular processes. Rather, each data type represents a different state of the cell. Therefore, multi-omic data need to be integrated to reveal these hidden components to understand the full process. Network modeling approaches allow to integrate multiple types of data and provide a solution to this problem. Some of the network modeling approaches can be listed as network propagation, network inference from gene expression, Bayesian networks, linear programing, the Steiner tree approach, electric circuits, network alignment, maximum-likelihood, network flow optimization [Tuncbag et al., 2013].

In network propagation methods, a set of nodes and a network are taken as input. By transferring the initial values of the selected nodes to their neighbors in the given network, it finds a subnetwork that represents the given condition best [Carlin et al., 2017]. An approach using network propagation called PRINCE predicts the genes and proteins for a specific disease by integrating disease similarity metric and protein protein interaction network [Vanunu et al., 2010].

Network inference from gene expression is the process of reconstructing network by using high throughput data [De Smet and Marchal, 2010]. An example of this approach predicts the regularity network only from the gene expression data. This approach uses message passing techniques and concentrate on combinatorial control that means the expression of a gene is regulated by mutual activity of many proteins [Bailly-Bechet et al., 2010].

Bayesian networks are probabilistic graphical models that represented in directed acyclic graphs. Nodes are random variables and edges represents probabilistic dependencies between nodes in a Bayesian network [Ben-Gal et al., 2007]. A sample approach that apply Bayesian networks uses gene expression and chromosomal copy number data together to find out driver mutations in cancer [Akavia et al., 2010].

Linear programming which is a special case of mathematical programming is a method for maximization or minimization of a linear objective function [Dantzig and Thapa, 2006, p. 1]. SPINE is a framework that intend to explain gene expression experiments in gene knocked out events. This framework is validated by predicting 99% of gene knockout effects in yeast [Ourfali et al., 2007].

In electrical circuit approaches, interactome is modeled as electrical circuit where proteins are interconnecting nodes and interactions are resistors and biological signal is treated as a flow of electrical current. In an example study, the method is applied to interactome of muscle specific genes in *C. elegans*. The role of genes that has high flow of informations is found to be important in muscles [Missiuro et al., 2009].

In network alignment approaches, two or more network from different species are aligned go find similarities and evolutionary conserved regions that might help us to discover functional properties of molecular components [Sharan and Ideker, 2006]. The study of Kelley et al. uses network alignment in the protein interaction networks [Kelley et al., 2003].

In network flow optimization approaches, network is modeled with minimum-cost

flow optimization problem. With this algorithm, the flow goes from source to target node through the network edges with flow capacities. The aim of this algorithm is to maximize flow between target to sink node with minimum possible cost. ResponseNet is a web server that uses this algorithm together with linear programming. It outputs a sub-network and its gene ontology enrichment analysis by analyzing weighted list of proteins and genes [Lan et al., 2011].

The Prize-Collecting Steiner Tree Problem is a network inference method to find a subtree on a graph with edge costs and vertex profits where the sum of edge costs contained in subtree and the sum of vertices profits not contained in subtree are minimized [Ljubić et al., 2006]. Prize collecting steiner tree (PCST) algorithm is a version of Steiner tree that does not require all data given by user is included in final network. In the study of Huang, the network is constructed from given data and using predefined protein-protein and protein-DNA interactions. Specifically, with this algorithm, the components that are hidden in the original data are extracted [Huang and Fraenkel, 2009].

In this thesis study, we are specifically using the prize-collecting version of the Steiner tree problem. Therefore, we emphasize more on the PCST based approaches and their applications in this context. In the study by Bailly-Bechet, the belief propagation, that is heuristic based, to solve the PCST problem has been applied and its performance has been assessed in synthetic datasets. Then, the same approach has been applied to the gene expression dataset of yeast and selected targets has been validated experimentally. As of their publication date, their solver exceeds the performance of other exact solvers [Bailly-Bechet et al., 2011]. In the study of Marcus T. Dittrich, linear programming and prize collecting Steiner tree problem are integrated and applied to lymphoma microarray dataset to find functional modules [Dittrich et al., 2008]. In another study, Steiner tree problem is used to extract hidden component of PPIs to figure out underlying biological pathways. This method is applied to phosphoproteomic and transcriptional data in yeast pheromone response and changes in unexpected pathways are identified [Huang and Fraenkel, 2009].

2.9 Omics Integrator

Network modeling approaches allow the investigation of cellular activities in many respects, since various types of data can be combined without the need for pathway information. Today's high throughput data is difficult to analyze and visualize because it contains millions of interactions between DNA, proteins and small molecules. The Omics Integrator package consists of two modules, Garnet and Forest. The Forest tool solves the prize-collecting Steiner forest problem and creates an interaction network using the omic data hits provided by the user [Tuncbag et al., 2013]. When creating the network, it considers the importance of omic hits and the possibility of the reality of interaction. Each given omic hit has a positive prize that expresses the confidence of the interaction.The hits provided by the user are called terminal nodes. When the algorithm includes a terminal node into the network, it is not penalized by the prize of the terminal node. The algorithm pays cost for each included edge to the final network. The algorithm creates a network by maximizing the collected prizes and minimizing the paid edge cost. During the optimization stage, not all terminal nodes

are forced to be included in the final tree. This eliminates the inclusion of the low probability edges. At the same time, the algorithm can also include nodes from the interactome. These nodes are called Steiner nodes and are biologically relevant but somehow not in the given dataset. The nodes that are highly connected generally exist in the networks even if they are not biologically related. Forest has the ability to give a negative prize value to the "hub" nodes to avoid this situation. In this way, they can only take place in the network when they are very dominant [Tuncbag et al., 2016].

CHAPTER 3

MATERIALS AND METHODS

In this chapter, we detail the methodology of the pipeline starting from parsing patient-specific mutation data to structural mapping and network modeling.

3.1 Overview of the Pipeline

We integrate data from multiple resources to detect the effect of mutations on proteins and eventually on the networks. As shown in Figure 3.1, the mutation data has been retrieved from TCGA [Tomczak et al., 2015], pathway data from Reactome [Joshi-Tope et al., 2005] and KEGG [Kanehisa et al., 2009], protein data from UniProt [Consortium, 2011], structural data of the proteins from Protein Databank [Berman et al., 2006](PDB) and finally protein-protein interaction data from iRefWeb [Turner et al., 2010].

All data retrieved from different databases were converted into objects with attributes and saved in json format. The external tools we used in this pipeline are Omics Integrator [Tuncbag et al., 2016] for network modeling and the modules implemented in BioPython [Cock et al., 2009] for sequence alignment. In the following parts, we describe each data set, analysis method and tools in more details.

3.2 Datasets

3.2.1 Data from TCGA

The mutation data in TCGA is downloaded in "mutation annotation format" (MAF) for the barcodes whose protection status is public. These files are generated by aligning the DNA sequence obtained from the patient samples to the sequence obtained from the normal samples and the reference sequence. Gene name, genomic coordinate, variation and many other descriptors are given in the mutation files. All these data were converted into patient objects as described above where each object has the barcode attribute and mutation object as another attribute. Mutation object has sequence position, UniProt entry identifier and gene names as attributes. The first step of the analysis is to find all missense and nonsense mutations. For that purpose, each patient file analyzed and nonsense and missense mutations gathered in a unique set. The mutation information representation composed of "hugo_symbol" (HGCN gene

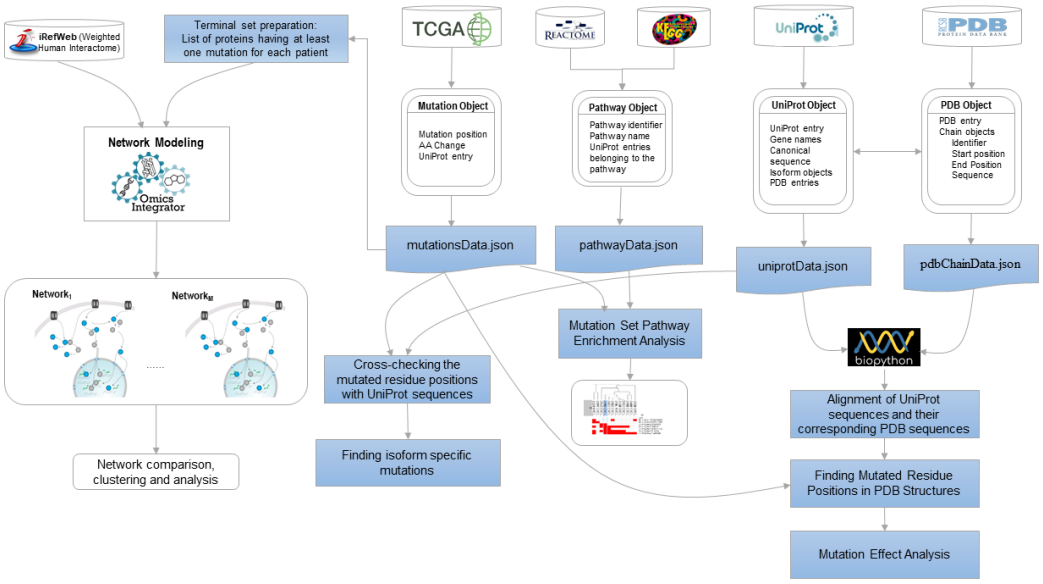

Figure 3.1: The general flowchart of our method. There are four types of data that are retrieved from multiple databases and transformed into Python objects. First one is the patient-specific mutation data from TCGA and it contains mutation position, amino acid change and UniProt entry list that is initially empty and filled with matched UniProt entries during analysis. Second one is the pathway data gathered from KEGG and Reactome and it is turned into pathway object that stores identifier, name of the pathway and list of UniProt entry names per pathway. The third data is protein features obtained from UniProt. This data is stored in UniProt objects and it contains entry identifier, gene names, canonical sequence, alternative sequences, list of PDB entries and isoform objects. The fourth data is the protein structure data provided from Protein Databank (PDB). The data is converted into Python object that comprises entry name, chain object list that contains chain sequence, sequence start and end indices.

symbol) and “protein_change” information. This gives us the mutation related gene, position and nucleotide change. Then, a binary matrix is constructed for finding the frequency of mutations in patients. The columns of the binary matrix represent the patients and the rows represent mutation names. This matrix can be used for filtering the patients that having mutation number less than a threshold.

3.2.2 Data Retrieval from UniProt

UniProt allows users to download the whole proteome of a given species. In this study, we downloaded the human proteome from UniProt which contains both manually curated proteins (UniProt/SwissProt) and automatically annotated proteins (UniProt/TrEMBL). The downloaded human proteome contains of 70,947 UniProt entries. The cross-reference of each protein from UniProt to other database identifiers such as official gene names, PDB identifiers, sequence in fasta format and the isoform information are also accessible from the downloaded proteome. All these information is arranged in a table format where rows are UniProt entries and the columns are the cross-references. The file format is tab delimited and each column in the table has its own separator. Gene names are provided in space separated, PDB identifiers and alterations in the isoform sequences are provided in semicolon separated. UniProt entry section for TP53 protein is shown in Figure 3.2 where UniProt identifier, gene names, cross-reference to PDB, sequence and alternative sequences (isoforms) are listed.

3.2.2.1 Cross-checking the Sequence Positions of the Mutated Residues

Each UniProt entry contains canonical sequence and its available isoforms. The UniProt objects we created during the analysis contain the sequence and associated genes of UniProt entries and isoforms. It means that we know which genes these proteins are coded from. In order to examine how mutations affect biological mechanisms, we need to check presence of the mutated residue positions in the given set of protein sequences. Since we have the gene name for each mutation, we can detect potential proteins by checking the gene list of UniProt entries. For this purpose, we looked for the name of each mutation in the mutation set on the gene list of UniProt entries. We performed a sequence analysis to find out the match in mutation and UniProt gene information. By sequence analysis, we checked whether the amino acid change in the position identified in our mutation could be in corresponding protein. If at the position of the mutation in the sequence of the protein the previous amino acid of the mutation is present, the protein has the potential for mutation. Sequence control was also performed for the isoforms of the corresponding protein. As a result of this analysis, we have identified the potential proteins and potential isoforms that each mutation can occur.

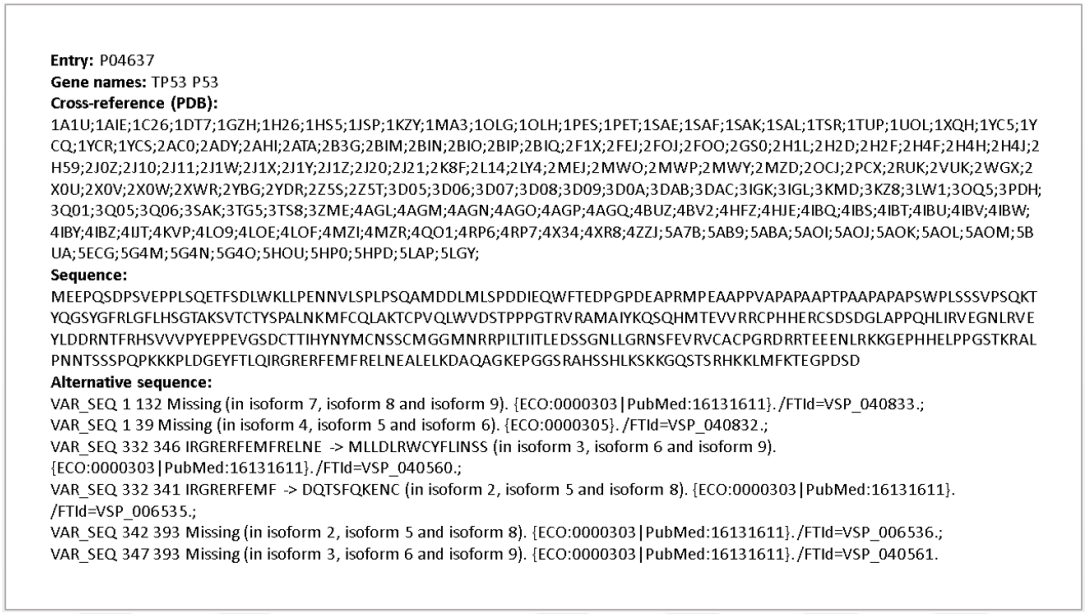

Figure 3.2: UniProt server output for TP53 protein. The downloaded entry information from UniProt given in a single line with tab separation and illustrated in this figure. Entry name represented as single identifier. Gene names are provided in space separated format. PDB identifiers given in a semicolon separated string. Sequence is an one single text. Alternative sequences are separated by semicolon and single alternative sequence entry contains change position range, change type, isoform list that corresponding change take place and paper identifier.

3.2.2.2 Extracting Isoform-Specific Mutations

There are several control mechanisms in post transcriptional process i.e. alternative splicing, alternative promoter usage, alternative initiation, or ribosomal frameshifting events. Therefore, a protein is expected to have multiple isoform sequences which are different than their canonical sequence. The human-proteome table downloaded from UniProt does not directly contain the sequences for each isoform. Each UniProt entry has an alternative sequence list. Isoforms can have more than one of these alternative sequences, and alternative sequences can occur in more than one isoform [Consortium, 2011]. Each alternative sequence representation starts with "VAR_SEQ" identifier, followed by position, change information in the sequence and the information about which isoform contains that change in brackets. In Figure 3.2, alternative sequences represent three types of changes in two general formats: i) The deletion of the amino acids at the given interval of positions, denoted as "Missing", ii) The insertion and substitution of the amino acids which are represented together because in fact both of them are replacement of the sequence from one sequence to another in the given position range. After the position information in the second format, the original sequence information is separated by the '->' sign with the changed sequence information as shown in Figure 3.2.

To check the possibility of some mutations to be specific to protein isoforms, we

scanned the mutations that are not matching to the canonical sequences. If a mutation only matches in the isoforms of the corresponding proteins, we labelled it as isoform-specific. For this purpose, we used the isoform sequences generated by our pipeline and identified the proteins having at least one isoform specific mutation. Then, we analyzed the frequency of isoform-specific mutations in the protein of interest in patients.

Figure 3.3: Isoform sequences of p53 protein. (A) Isoform 6 sequence changes with respect to canonical sequence. (B) Isoform 7 sequence differences with respect to canonical sequence. (C) Isoform 8 sequence changes difference from the canonical sequence. (D) Isoform 9 sequence changes with respect to canonical sequence. The isoforms of a protein can have same changes in their sequences. For example, the missing of 1-132 residues occurs in both isoform 7, 8 and 9.

Figure 3.3 shows the TP53 UniProt entry sequences for isoforms 6, 7, 8 and 9. For each isoform, the changes in sequence with respect to canonical sequence are provided. Some of the changes are observed in multiple isoforms. For example, missing amino acids in 1-132 position range (Missing 1-132) occurs in isoform 7, 8 and 9. This information provided in the downloaded human proteome in an abstract way as shown in Figure 3.2 where the changes are the primary keys, the same changes are collected together and given as for example "VAR_SEQ 1 132 Missing (in isoform 7, isoform 8 and isoform 9)." Therefore, we designed the part to parse the isoform sequences in our pipeline accordingly.

In the Figure 3.2, there are six VAR_SEQ in TP53 alternative sequences and isoform numbers are provided in the VAR_SEQ entries. To be able to process this information properly and obtain sequences of isoforms, we first need to find out which changes

exist in the corresponding isoform. For this purpose, we examine all these sequence changes with a preliminary analysis and collect the variations in each isoform separately. Subsequently, we apply the sequence changes in each of these isoforms to the canonical sequence to obtain the sequence of isoform. When applying the changes to generate the isoform sequences, a problem has been encountered. All changes made for isoforms are given based on the original sequential positions. As the changes in isoform are applied consecutively, since the positions differ after each applied change, the next changes cannot be applied to the right place. For this reason, we must somehow apply all these changes at the same time to the original sequence. To do this, we need to keep the original sequence indices while the changes are applied in order. That is, when an insertion is made, this sequence should not be expanded at that time, and the effect of the insertion should be reflected after all changes have been applied. For this purpose, the original sequence has been converted into a list with each amino acid corresponding to one element and all changes are applied on this list as detailed below;

I. The sequence list has not been expanded or shrunk after any changes.

II. To keep the index constant and apply the changes at the same time, the method of expanding or decrementing the elements of the list according to need is followed.

III. If former and latter sequences have the same length, substitution can easily be applied to former sequence indices.

IV. If an insertion is the case to form the isoform sequence, then it means that the latter sequence is longer than the former one. Therefore, amino acid subsequence to be inserted are entered in the list in the given position range of the old one.

V. If the new sequence is shorter than the former sequence, the former amino acids are inserted to corresponding indices. As the former sequence is longer, there are amino acids in the list that need to be removed. Empty string is entered in place of these amino acids.

VI. If a deletion occurs in a position range, amino acids in the mutation position range turned into empty strings.

Thus, the index is preserved and the new sequence is transferred to the list. This list is converted into a sequence by concatenating all the elements. With this method, all changes can be applied to the correct indices without losing the original sequential positions in individual isoform sequences.

The snapshot from UniProt website for FOSL2 protein isoform is illustrated in Figure 3.4. The described list processing in our pipeline to keep the indices correct and to prepare isoform sequences is shown in Figure 3.5 (the list version of isoform 2 sequence of P15408 FOSL2 protein). Isoform 2 has a deletion of 1-25 positions and this is applied as empty string in the corresponding positions on the list. (3.5 B Red) There are amino acid changes in positions 26-34. Since the length of the old sequence is equal to the length of the new sequence, the replacement of amino acids are easily

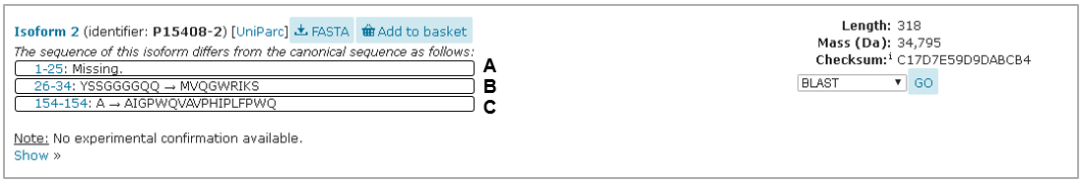

Figure 3.4: Isoform 2 sequence changes of FOSL2 protein (P15408). The changes in isoform 2 sequence with respect to the canonical sequence has been listed. (A) Missing amino acids in 1-25 residues. (B) Substitution in 26-34 residues. (C) Insertion in 154 residue.

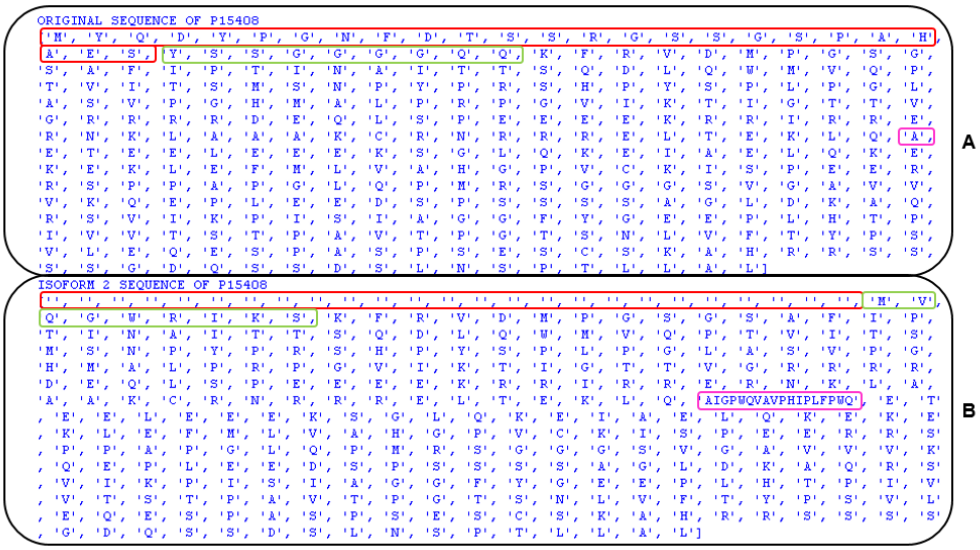

Figure 3.5: List representation of FOSL2 sequence. (A) Canonical sequence of FOSL2 (P15408) as a list. (B) Method to form sequence of isoform 2 of FOSL2 protein. Red boxes highlight missing amino acids in 1-25 in canonical sequence is applied as replacement of corresponding indices with empty strings in canonical sequence list. Green boxes highlight substitution of amino acids in 26-34 residues in canonical sequence which is reflected as alterations in respective positions. Pink boxes highlight insertion to canonical sequence is reflected as appending new amino acids in change position.

applied to the list. (3.5 B Green) In the 154th position, there is an insertion. To apply this insertion to list without changing the index, the new amino acids are entered to list; up to the length of the old sequence. The length of the former sequence is just 1 in this example. The remaining elements have been added to the last changed element. In the case, the last edited amino acid is “A” in the 154th position. For this reason, the amino acid “A” at position 154 contains the newly introduced amino acids in its own index. (3.5 B Pink) After all the changes have been applied, the list is concatenated and the sequence of isoform 2 is obtained. In this way, we can apply all the changes

of isoforms in the correct indices without losing the index of the original sequence.

3.3 Structural Mapping of Mutations

At this stage, our aim is to find the correct position of a given mutated residue in the PDB structure. A PDB entry of a UniProt sequence may represent only a fragment of the given protein and the residue numbering may not be exactly the same with the sequence positions. Therefore, there is a need to find position in structure for advance analysis, such as finding protein regions, spatial neighbors of mutated residues and many more.

3.3.1 PDB File Format

The Protein Data Bank (PDB) format provides atomic details of the macromolecular structures including protein data obtained by X-ray diffraction and NMR studies. This file format has ".pdb" extension [Berman et al., 2006]. PDB structure files are available to users in website and via an ftp (file transfer protocol). Firstly, we downloaded the data of 109,889 PDB structure in UniProt-PDB mapping file through this service. Monomers are represented only with one chain in the PDB structure. However, a PDB entry could be a complex composed of more than one proteins. Then, each protein is labelled with a unique chain identifier with the given four-letter PDB identifier.

The data in the PDB files include atomic coordinates, names of the molecules, primary and secondary structure information, sequence database references, bibliographic citations etc. Each PDB file contains multiple lines and has 80 columns on each line. In the file, each row contains information about a record type, and the first 6 columns of each row represent the name of the corresponding record type. The record types are like HEADER which is the first line of the file, and AUTHOR which is a list of contributors, SHEET and HELIX which provide information about 2D structures, ATOM represents atomic coordinate records for standard groups. The information of the recording types can be one or more lines. The columns in each row type are assigned to different fields in different positions. For example, columns 11-50 on the HEADER line correspond to the grouping information, while columns 11-79 on the AUTHOR line correspond to the author list. The information of the fields corresponding to columns of each record type is included in the PDB file format documentation.

In the PDB file, the information we need to extract the sequence of the three-dimensional structure exist in the lines of the ATOM record type which contains standard amino acid and nucleotide coordinates.

In Figure 3.6, columns that correspond to fields of ATOM records are listed. Each line starting with "ATOM" keyword represents each atom in the protein listed from amino terminus to carboxyl terminus. ATOM records of a chain are terminated by a "TER" keyword [Berman et al., 2006]. We obtained sequence information using the residue name and residue sequence number columns of the CA atom in each chain in the PDB record. For this purpose, the file corresponding to each PDB entry is read and

Record Format

COLUMNS	DATA TYPE	FIELD	DEFINITION
1 - 6	Record name	"ATOM "	
7 - 11	Integer	serial	Atom serial number.
13 - 16	Atom	name	Atom name.
17	Character	altLoc	Alternate location indicator.
18 - 20	Residue name	resName	Residue name.
22	Character	chainID	Chain identifier.
23 - 26	Integer	resSeq	Residue sequence number.
27	AChar	iCode	Code for insertion of residues.
31 - 38	Real(8.3)	x	Orthogonal coordinates for X in Angstroms.
39 - 46	Real(8.3)	y	Orthogonal coordinates for Y in Angstroms.
47 - 54	Real(8.3)	z	Orthogonal coordinates for Z in Angstroms.
55 - 60	Real(6.2)	occupancy	Occupancy.
61 - 66	Real(6.2)	tempFactor	Temperature factor.
77 - 78	LString(2)	element	Element symbol, right-justified.
79 - 80	LString(2)	charge	Charge on the atom.

Figure 3.6: Snapshot of a PDB File Format. One record represented in one line. Columns refer to single character in a line. For example, 1-6 characters of a line refers to record name.

the ATOM records are analyzed according to the chain information. The amino acids corresponding to the atoms in each chain are transformed into sequence with one letter amino acid representation by using positional information. During this process, some exceptional cases have been encountered. First of all, position information in atomic records of some chains is not consecutive. In these cases, the "X" character is entered for the missing positions in the sequence. By this process, Python objects were created to store and reuse the PDB input and chain information obtained as a result of the analysis. The PDB object representing a PDB entry contains the PDB identifier and the chain object list. The chain object holds the related data including starting position, the ending position, the chain name, and the sequence information.

3.3.2 Mapping the Mutations onto the Protein Structure

PDB identifiers provide information about 3D structures of the proteins. Each identifier is represented by four characters and each protein co-crystallized with their partners are represented as a chain and chains are labelled with one letter. In general, amino acid positions in UniProt sequences do not match to the residue position labels in PDB sequences. Therefore, there is a need of position adjustment from sequence to structure. As detailed in chapter 2, proteins can be divided into three regions, surface, core and interface. In order to be able to find the location of mutations in the protein

whether they are in the core, surface or in the interface, mapping the residue positions to the protein structure is crucial. For this purpose, we need to cross-reference UniProt entries to their corresponding PDB identifier and the chain identifiers. PDB and UniProt supply a cross reference table for mapping PDB chains to UniProt entries in a space-delimited format where each row have a PDB identifier, its related chain and corresponding UniProt entry. From the human proteome, we know PDB identifiers for each UniProt entry and using this, we have extracted the corresponding chain information for each UniProt. By this mapping, we have found specific chains of PDB structures for each UniProt entry, if it maps to any.

We have processed this data and have stored it in a usable manner in our subsequent analyzes. For this purpose, we read the file in the Python script, create the objects, and store all the objects in a json file with the help of jsonpickle library in Python for decoding and encoding python objects. In order to efficiently store the information that we obtain as a result of the analysis, we must define the classes that hold the characteristics of the data set and use these classes to organize and store the results efficiently. For this purpose, we first created a class "UniProt" to express UniProt entries in the "human-proteome" file and an "Isoform" class to express more than one isoform of UniProt objects. The Isoform class includes the following qualities; “uniprotId "of the UniProt object to which it belongs, "no" which holds the number information, "sequence" which is obtained by applying the original sequence changes, and "varSeqs" list which contains the information of the change belonging to the isoform. The UniProt class includes the following qualities; "uniProtId" which corresponds to the key of UniProt entry, "geneNames" list which holds gene names, "PDBs" list which contains PDB entries together with chain information, "sequence" which contains canonical sequence and "isoforms" list which holds multiple Isoform objects. All of these objects formed by processing the related files and formed objects are stored in json format.

Once we have identified the potential proteins for mutations to occur, we can identify the positions of the mutations in the three-dimensional structure. We know corresponding structures of proteins from the PDB list of UniProt entries. With this method, we can expand our analysis of detecting potential proteins from mutations by detecting the position in the three-dimensional structure of proteins. We cannot directly check the position of the mutations because PDB chains express different parts of the protein. Therefore, we should primarily align the PDB chains on the relevant protein sequence. For this purpose, we use the sequence of the PDB chains that we created in our previous analyzes. In this analysis, we aligned the UniProt sequence (or the isoform sequence) with the sequence of each PDB chain to find which regions in the protein sequence the chains correspond to. The alignment was done pairwise2 module of Biopython package [Cock et al., 2009]. By this way, we discovered the position of our mutation in the PDB chain sequence. The mutations positions are transformed to the related PDB chain sequence positions. Thus, we have matched mutations to a three-dimensional structure.

3.4 Analysis of Mutation Effects

3.4.1 Identification of Protein Regions

Proteins can be divided into three regions; namely, core, surface and interface regions. The conventional approach for identification of these regions are calculating solvent accessible surface areas of each residue in the protein. NACCESS is a software designed for calculation of solvent accessible surface area both at residue level and at molecule level [Hubbard, 1992]. In general, if the relative solvent accessible surface area of a residue in its monomer state is greater than or equal to 15Å^2, then this residue is labelled as the surface residue. Interface residues are still surface residues. To identify them, we browse PDBSum that lists interface residues.

Naccess is an implementation of the method of Lee and Richards [Lee and Richards, 1971]. It takes PDB structure file with .pdb extension as input and outputs three file. The first one is the atomic accessibility file with .asa extension that contains calculated accessible surface area (ASA) and van der Waals radii of each atom in the given PDB file. Second file is the residue accessibility file with .rsa extension in that the residue accessibilities are listed over each protein, as well as the relative accessibility (relASA) of each residue calculated as the percentage accessibility compared to the total accessibility of the residue in an extended ALA-x-ALA tripeptide [Hubbard et al., 1991]. Third file is the log file of the calculation that keeps each calculation step and the error and warning messages. The ASA is calculated by rolling a water molecules around the protein structure. The default radius of the water molecule is assumed to be 1.4 Å, but it can be changed. The solvent molecule is located at the position that is calculated as total radii of atom and solvent molecule for each atom. While the solvent molecule rolls around the atoms of protein, arc for each atom are drawn. If there is any arch for an atom, then it is detected as accessible. Total length of all arcs for a atom is proportional to total accessibility. In this work, we used the default value. To find the surface residues we used the output .rsa file and the relASA column in that file.

3.4.2 Detection of the Effects of Mutations

3.4.2.1 EVmutation

EVmutate is developed based on an unsupervised statistical technique to identify the effect of mutations by using residue coupling information [Hopf et al., 2017]. In residue coupling, the dependency of a selected residue to its neighbors in a given window is considered. The EVmutation data is provided for a limited number of proteins in text format where each position in a UniProt entry is substituted to the remaining 19 amino acid and the damage score is calculated. The more negative values of the calculate score means the more damaging mutation. We use this data to measure the likelihood of our mutations and we took the score of the displacement that occurred in each mutation and compared it with the average score of all other displacement probabilities. For this purpose, for each matched UniProt entries of each mutation,

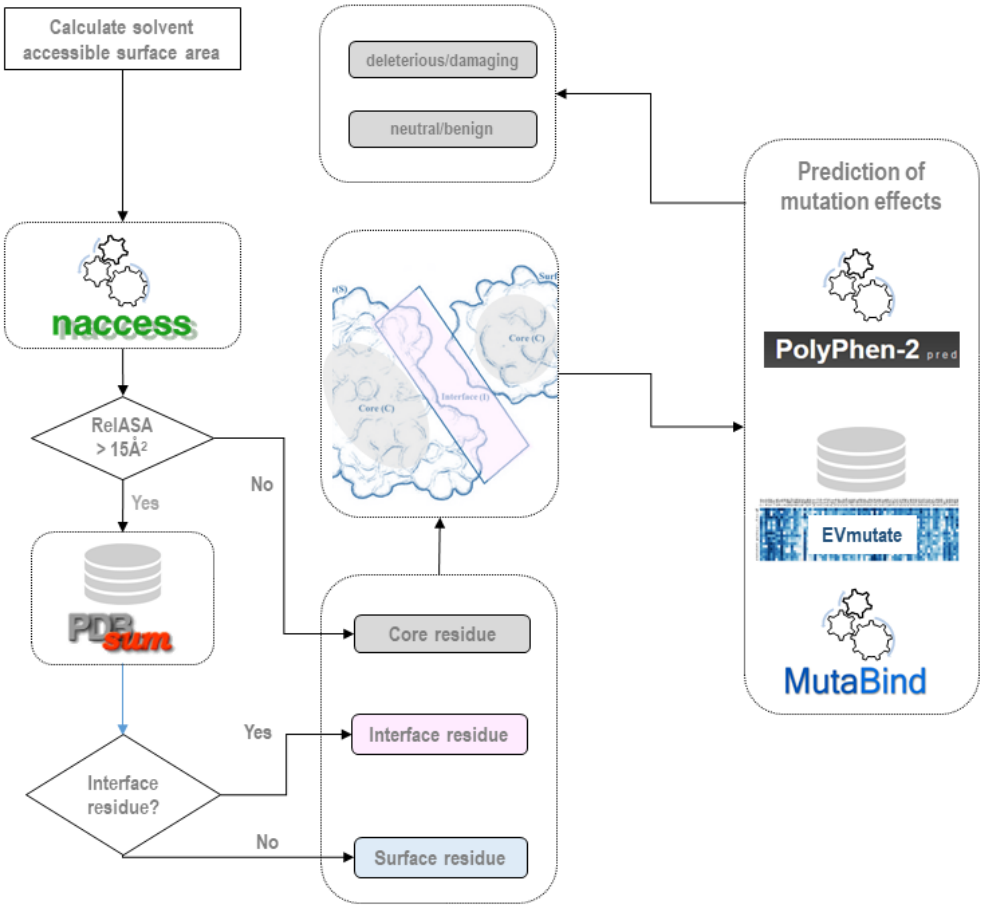

Figure 3.7: Mutation effect analysis pipeline. First step is the determination of mutation region as in core or not. If accessible surface area of mutation residue is greater than 15Å^2, the residue can be on interface or surface region of protein. If ASA is not greater than 15Å^2, the residue is in the core. After the region of residue is determined, the effect analysis of the mutation is applied with Mutabind, EVmutate and PolyPhen-2.

we checked the existence of EVmutation file of that UniProt entry and the sequence interval provided in the EVmutation UniProt file to see if the mutation position included in the given range. From EVmutation files, we extracted all possibilities of substitution for preceding amino acid before the mutation occurred.

3.4.2.2 PolyPhen-2

PolyPhen-2 predicts effect of amino acid substitution on protein structure and protein function [Adzhubei et al., 2013]. The web server can run batch queries. The mutation list organized for batch input format and single row includes UniProt entry key, former codon of mutation, mutation residue, latter codon of mutation. The query runs as a job [Adzhubei et al., 2013].

3.4.2.3 MutaBind

MutaBind is an energy-based method for calculating binding affinity change as a result of mutation [Li et al., 2016]. MutaBind server does not allow batch queries. User can run a job from web interface by selecting inputs step by step. Therefore, we selected some mutations and their mapped structures as input to MutaBind.

3.5 Network Modeling with Omics Integrator Software by Integrating Mutation and Interactome Data

In this step, for each patient the number of mutations per gene was calculated and prize files are formed. Prize file for each patient contains gene names in first column and mutation counts for specific gene in the second column. This prize files are provided as prize input to Omics Integrator [Tuncbag et al., 2016]. Edge file for Omics Integrator is obtained from iRefWeb which provides interaction data from 10 public databases including BIND, BioGRID, CORUM etc. [Turner et al., 2010]. The edge file contains interaction partners, their interaction probability and optionally the type of interaction (directed/undirected) in each row. The tuning parameters are determined as $\omega = 10$, $D = 10$, $\beta = 5$ and $\mu = 0.01$ that are given in default. The meaning of these parameters are explained in the following sections. In a more detailed analysis, these parameters can be used in a given interval in their combinations.

3.5.1 Omics Integrator Algorithm

Given a graph $G(V, E, w)$ where V is the node set $\{v|v \in V\}$ $\{v, v\}$, E is the edge set and w is the edge weights, the Forest module of Omics Integrator solves the prize-collecting Steiner forest problem. In Omics Integrator formula which is defined in 3.1, $c(e)$ is the edge costs defined in 3.3, $p'(v)$ is the function that assigns prizes to each node which is defined in Equation 3.2, number of trees are defined by the parameter k and w is the tuning parameter. Initially, dummy node $(v0)$ is added to network and all

nodes in the initial interactome is connected to the dummy node with an edge weight of w. After optimization step, root node $(v0)$ and its all edges removed from the final network to obtain a forest from the optimized tree.

$$f'(F) = \sum_{v \notin V_F} p'(v) + \sum_{e \in E_F} c(e) + \omega \cdot k \tag{3.1}$$

$$p'(v) = \beta \cdot p(v) - \mu \cdot degree(v) \tag{3.2}$$

The prize function to give weights to nodes is given in Equation 3.2 where $degree(v)$ is the number of connections of the node in G. β and μ parameters are the scaling factors. Assigning 0 to μ discards the hub correction in the algorithm, while increasing the value of μ decreases the hub dominance in the resulting network. The higher value of β forces more terminals to be included in the final network. A very useful feature of Forest module is that it can incorporate negative evidence, in other words negative prizes to nodes, to have a better reconstruction. In our work, negative weights are coming from the μ scaled degree function $(degree(v))$.

The cost of edges is calculated with the Equation 3.3. $prob(e)$ is the confidence measure of interaction. $prob(e)$ value close to 0 means that the probability of the given interaction to be real is low. Because PCSF is a minimization problem in our case, we would like to minimize the total cost of the edges included in the final forest. Therefore, the cost function is given in Equation 3.3 which makes high probability edges to be low cost. In this way, we eliminate the false positives as many as possible in the final network.

$$c(e) = 1 - prob(e) \tag{3.3}$$

Omics Integrator takes three input parameters; w, β and D. Parameter w controls the number of trees in the final forest, β controls the balance between including more terminal nodes into the network and D controls the maximum depth of the network. The optional parameter μ is used to penalize all nodes in the network based on their degrees. In this way, the dominance of the hub nodes in the final network is overcome.

After the minimization of the objective function in Equation 3.1, the result is a forest composed of multiple trees $F = (V_F, E_F)$ where $F(E) \subset G(E)$. With all these features, Forest algorithm provides output that can be easily viewed with Cytoscape, the network visualization tool. In this study, node prizes are given as the number of mutations per protein for each patient. Therefore, the prize list composed of the proteins having at least one mutation. Every patient mutation file processed and the number of mutation occurring in each protein calculated separately. Omics Integrator is run for each patient with these files as prize input. We used the iRefWeb v8.0 as the weighted interactome in our modeling. As a result, optimal forest file for each patient is created with the extension _augmentedForest.sif.

3.5.2 Merging Patient-Specific Networks and Community Detection

For each patient, we reconstructed one graph $G_i(V_i, E_i)$ where E_i is the set of edges and V_i is the set of vertices and obtained from set of graphs $\{G_1, G_2 \dots G_N\}$ where N is the number of patients.

$$G_{union} = \bigcup_{i=1}^{N} G_i \tag{3.4}$$

V_{union} is the vertices set represented by $V_{union} = \bigcup_{i=1}^{N} V_i$ and E_{union} is the edge set of G_{union} represented by $E = \bigcup_{i=1}^{N} E_i$. Then, we filtered this graph based on the presence of each edge in patient network. For this purpose, we prepared a matrix A where columns are patients and the rows are the edges in the union graph (G_{union}). Next, we filtered G_{union} to obtain a representative network by applying a constraint that an edge needs to be present in at least three patients as shown in Equation 3.6.

$$A_{ij} = \begin{cases} 1, \text{if } e_i \in G_j \\ 0, \text{else} \end{cases} \tag{3.5}$$

$$\text{if } \sum_{j=1}^{N} A_{ij} > 3 \text{ then } e_i \in G_{common} \tag{3.6}$$

After that, nodes that exist at least three patient are gathered and merged as one network. The resulting network is expected to have one giant component and many other small connected components. We continued with the giant connected component for further analysis. All the network merging and analysis stages are performed with the help of networkx library in Python [Hagberg et al., 2005].

Next, this network is clustered using Cluster Maker which is a Cytoscape plugin. Cluster maker has two types of algorithms for clustering; attribute clustering and network clustering. Attribute clustering is used for clustering genetic data and includes hierarchical, k-medoid, AutoSOMe and k-means algorithms. In attribute clustering, a list of node attributes or an edge attribute can be selected as parameters. On the other hand, network clustering is used for partitioning depending on a single edge attribute and its available algorithms are community clustering (GLAY), MCL, affinity propagation, MCODE, AutoSOME, SCPS, transitivity clustering. In this study, community clustering (GLAY) has been applied for clustering. Community clustering is an implementation of Girvan-Newman fast greedy algorithm. This algorithm depends on edge betweenness and it accepts the idea that if there is communities in a network, then these communities are connected to each other with few edges. Therefore these few edges must have high connections because all shortest paths must go along on these edges meaning that they have high edge betweenness. To find clusters with this algorithm, betweenness of all edges are calculated and edge that has

highest betweenness is removed from the network. After that, betweenness of all edges is calculated again. This step repeated until there is no edges in the network [Girvan and Newman, 2002].

3.5.3 Network Centrality Measure (Betweenness centrality, degree centrality)

Centrality is an important property of networks. Over the years, several approaches proposed for measuring centrality. A graph contains nodes and edges connecting pair of nodes. Two nodes are adjacent, if there is an edge between them. The degree of a node is the number of adjacent nodes to a given node. Given two nodes (n_i, n_j) is reachable from each other, if there exist one or more edges between these nodes starting from n_i and ending at n_j. Distance between two nodes are defined as the number of edges between them. Shortest path is the path with minimum distance between two nodes. Degree centrality determined by the degree of the nodes. Therefore the node with maximum degree, has the degree centrality in graph. On the other hand, betweenness centrality of a node determined with the frequency of shortest path of other nodes that falls through it. If a node has a high number participation in shortest paths between all pair of nodes, it has a high value of betweenness centrality in the graph [Freeman, 1978].

3.6 Enrichment Analysis

3.6.1 Mutation Enrichment Analysis in Pathways

The pathway data was gathered from two different resources: Reactome and KEGG. Reactome is a free, open source, peer reviewed and curated pathway database. We used UniProt to reactome mapping table, which contains the mapping of 2,147 pathways to the related UniProt entries. The downloaded table is in a tab delimited text format, where rows correspond to a mapping of a pathway and a UniProt entry and columns correspond to the UniProt accession, Reactome identifier, link to related pathway, pathway name and species information. In this mapping, each pathway have multiple UniProt entry associations. On the other hand, KEGG deposits manually curated pathway information that represents knowledge on the molecular interactions [Kanehisa and Goto, 2000]. KEGG provides pathway information in separate files, named with the identifier of the pathways and represented in “simple interaction format” (SIF). A particular pathway interaction file contains the interacting proteins in its first and second columns and the third column shows the interaction types. For organising the pathway data, we processed these files and generated a unique list of the genes in each pathway.

The pathway entries obtained from the Reactome and KEGG databases were stored in a usable manner for our subsequent analyses. For this purpose, we read the file, create the objects, and store all the objects in a json file. We created the Pathway class, the qualities of which are; "id" which expresses the key to the metabolic pathway, "species" which contains species information, and "uniprots" which is a list holding the keys of the UniProt entries in the metabolic pathway. Pathway objects such as

UniProt objects were also written to a file in "json" format.

We applied a similar strategy to gene list pathway enrichment analysis with mutation data. Here, our aim is to identify significantly mutated pathways. For this purpose, we analyzed each pathway for each patient and applied Fisher's exact test for each pair of patient and pathway combination. The contingency table to apply the Fisher's exact test is shown in Table 3.1.

Table 3.1: Contingency table of mutations according to patient and pathway.

	In pathway	**Out pathway**	
In patient	Number of mutations that are present in the selected patient and selected pathway a	Number of mutations that are present in the selected patient but not in the selected pathway b	Number of mutations in the selected patient a+b
Not in patient	Number of mutations that are present in the selected pathway but not in patient c	Number of mutations that are not present in the selected patient and the selected pathway d	Number of mutations that are not in patient c+d
	Total number of mutations in pathway a+c	**Total number of mutations that are not in pathway b+d**	**Total number of mutations across 290 GBM patients a+b+c+d**

As in the contingency table shown, first we counted the number of mutations in each patient (represented by a+b). Then, for each pair of patient-specific mutation set and the set of proteins in each pathway, we counted the number of mutations belonging to the proteins in the pathway (represented by a). Separately, we counted the number of mutations in our overall dataset (all unique mutations in 290 GBM tumors) that are belonging to the proteins in the pathway but not present in the selected patient (represented by c). Finally, the number of mutations that are neither in the selected patient nor belonging to the pathway components were counted (represented by d). After constructing the contingency table, we used the fisher_exact_test function implemented in scipy.stats module of Python to calculate the p-value. We focused on the significant pathways with a p-value less than 0.05 and having at least three mutations belonging to the patient. We applied this test to all pathways deposited in Reactome and KEGG.

3.6.2 GO Enrichment Analysis on the Reconstructed Networks

With the purpose of functional annotation, genes and proteins are associated with Gene Ontology (GO) terms, which constitute a controlled vocabulary for the biomolecular attributes [Consortium, 2004]. GO system is composed of three aspects (categories): biological process, molecular function and cellular component [Ashburner et al., 2000]. Each gene/protein is mapped to the most relevant GO terms to record their attributes

in biological annotation databases. A similar type of associations are made between genes/proteins and pathway records in biological databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [Kanehisa et al., 2009], considering the pathway memberships of these genes/proteins. An overview of the annotations of a gene/protein explains its biological attributes. One advantage of using ontological systems for defining biomolecular attributes is that they are machine interpretable, which makes it possible to conduct computational data analysis on the respective genes/proteins.

Functional enrichment and overrepresentation analyses reveals the common annotations among a list of genes/proteins [Subramanian et al., 2005]. They are frequently employed to analyze clusters of differentially expressed genes to identify the affected biological processes due to alterations from the healthy state. ClusterProfiler is a R package that works within the Bioconductor project, to classify genes according to their annotated biological terms and analyze the enrichment of gene clusters. ClusterProfiler also supports annotation comparison among multiple clusters [Yu et al., 2012].

In order to identify the shared biological processes and pathways among our gene/protein clusters, we run the clusterProfiler for 21 clusters that were obtained from the cluster analysis on the merged network of 290 patients. The “compareCluster” function is called through R script with GO (Biological process) and KEGG annotations separately to display the most enriched GO based biological processes and pathways for each cluster.

CHAPTER 4

RESULTS

In this chapter, we present the results of our patient specific analysis in GBM. We first detail the statistics of the mutation data and then we show the distribution of the mutations based on the protein regions. Additionally, we present the changes of the chemical properties after a residue is mutated and the functional effect of the mutations whether they are damaging or neutral. In the last part of the chapter, we show the results of network modeling to give an insight how mutated proteins interplay in functional pathways and biological processes beyond the list of mutations.

4.1 Data Statistics

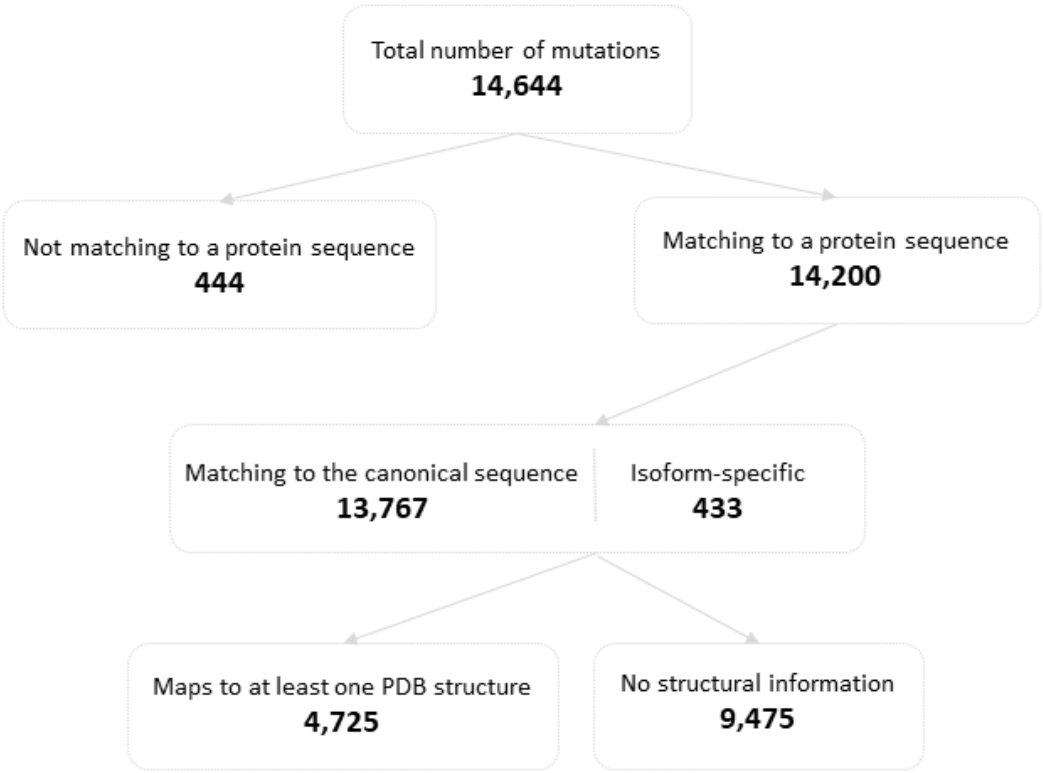

Figure 4.1: GBM mutation data mappings statistics.

By matching the positions of the mutations to protein sequences, we first validate them in canonical sequences. There are in total 14,644 unique mutations when we take the union of the mutations in 290 GBM patients. 444 of these mutations do not match any protein, while 14,200 of them match at least one protein. Out of 14,200 mutations, 4,725 are mapped to at least one PDB structure; however, 9,475 are not positioned to any PDB structure. On the other hand, 433 of the mutations are isoform specific that do not match in the canonical sequence but match at least one isoform sequence (shown in Figure 4.1).

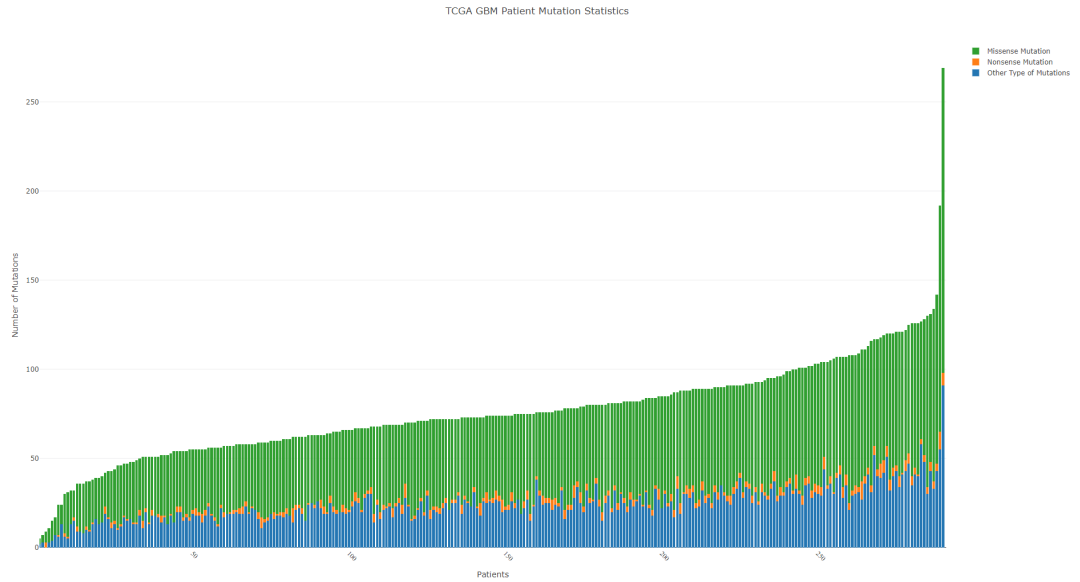

Figure 4.2: Mutation statistics of Glioblastoma Multiforme patients. X axis represents 290 GBM patients and y axis represents total number of mutations. While green bar refers to missense mutation number, orange bar refers to nonsense mutation number and blue bar is the other type of mutations.

As shown in Figure 4.2, the distribution of number of mutations and mutation types are not uniform across patients; rather it is heterogeneous. When we look at the number of mutations in each patient, we notice that the number of missense mutations dominates in almost all patients. Patient barcoded with "TCGA-06-5858-01" has the maximum number of mutations which is 269; 171 of them are missense mutations, 7 of them are nonsense mutations and 91 of them are other types of mutations. Patient barcoded with "TCGA-06-0178-01" has the minimum number of mutations which is 5 with 3 missense and 2 other type of mutations. The average of number of mutations is 76.11 when we consider 290 patients.

In Figure 4.3, the distribution of the number of missense mutations number across over 290 patients has shown. The patient with maximum number of missense mutation is the patient with "TCGA-06-5858-01" barcode and it has 171 missense mutation over 269 mutations. On the other hand, patient with "TCGA-06-0178-01" barcode has minimum number of missense mutation with 3 missense mutation over 5 mutations. In addition to this, average number of missense mutation over 290 patient is 49.01.

The barcodes having the maximum and minimum number of all types of mutations are the same with the ones having the maximum and minimum number of missense mutations.

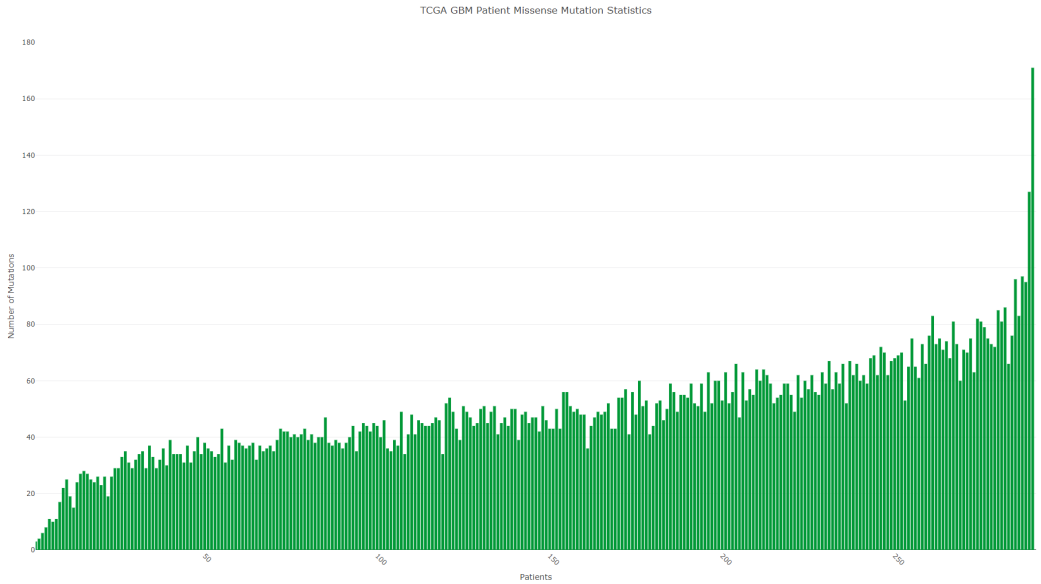

Figure 4.3: Missense mutation statistics of Glioblastoma Multiforme patients. X axis represents 290 GBM patients and y axis represents total number of mutations. Green bar refers to missense mutation number.

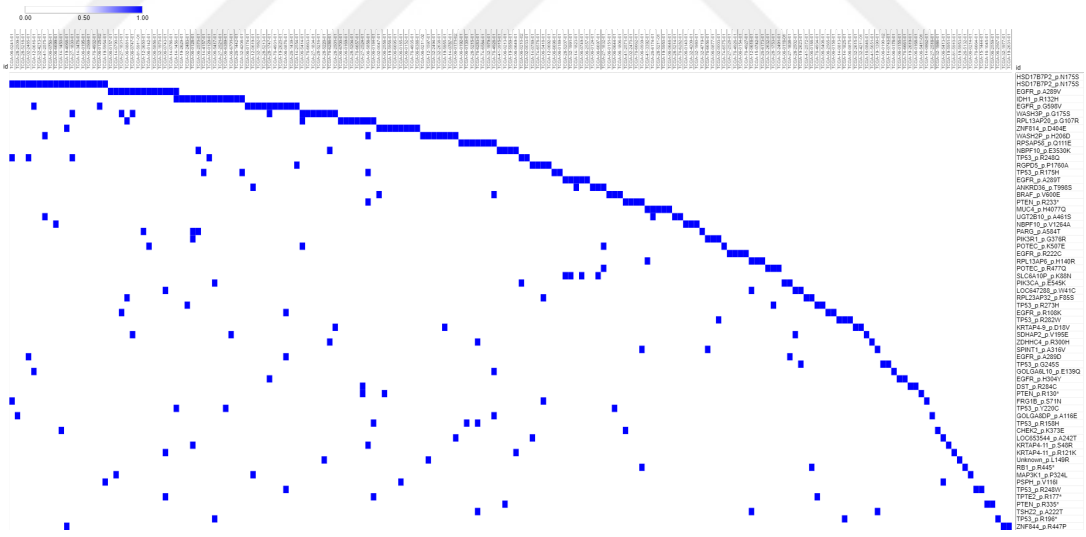

Figure 4.4: The heatmap of the mutations present in at least three patients. Blue represents presence and white represents absence of the mutation (row) in the corresponding patient (column).

As shown in Figure 4.4 and a more zoomed version in Appendix A, the distribution of the mutations across different GBM tumors is very heterogeneous. There are only 62 unique mutations that are present in at least three patients and 245 mutations present

in at least two patients. The most frequent mutation with 18 patients is HSD17B7P2 mutation N175S which is followed by EGFR mutation A289V and IDH1 mutation R132H in 13 patients. When we sum up the total number of mutations in each protein, we see that EGFR and IDH1 are the most frequently mutated proteins in GBM.

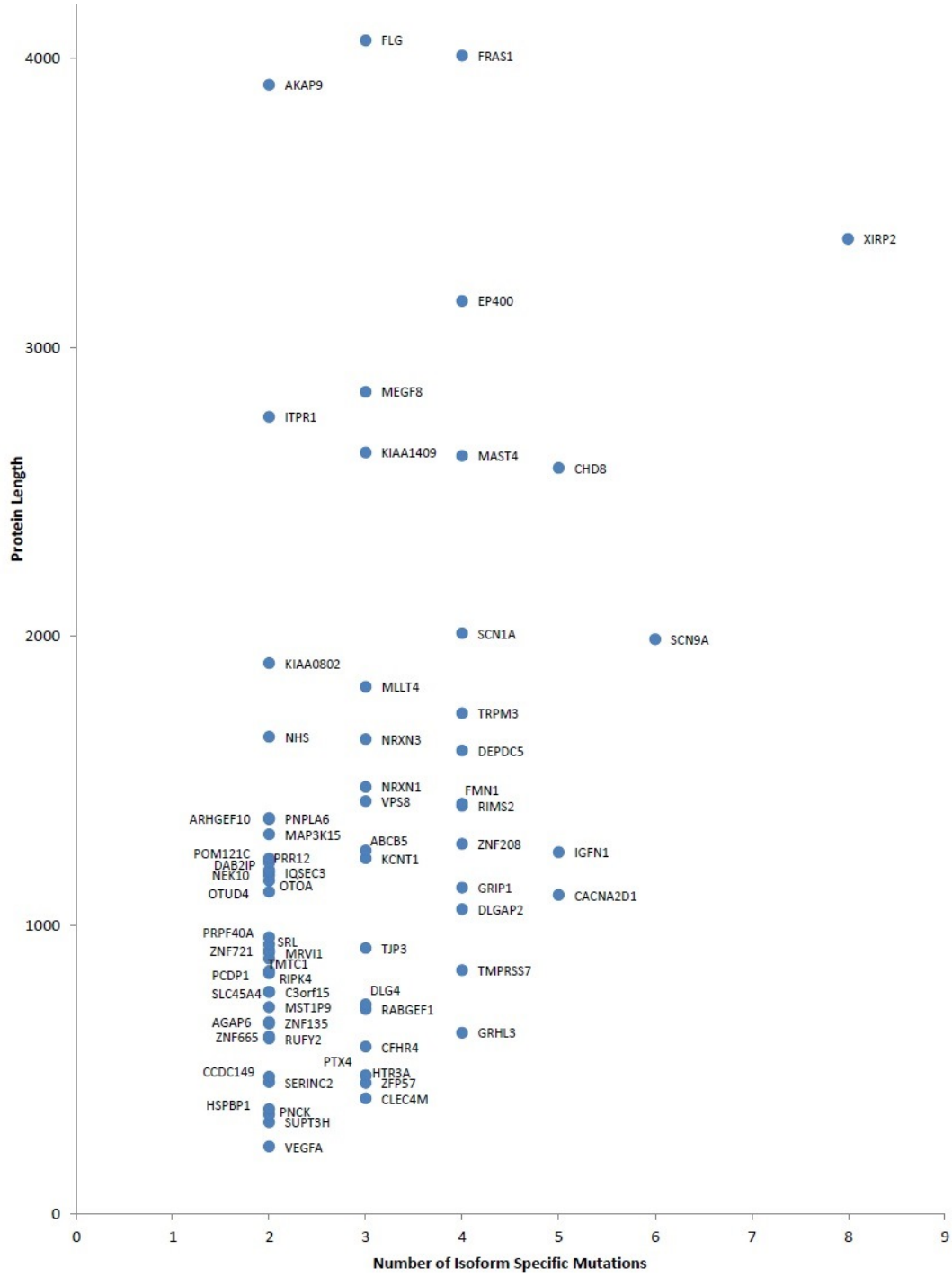

Figure 4.5: Scatter plot of isoform-specific mutations. X axis represents mutation number. Y axis represents sequence length of corresponding protein.

The protein having highest number of isoform-specific mutations is the Titin protein (TTN). More than 70 isoform specific mutations are observed in TTN. It is a huge protein found in the heart and skeletal muscles [Guo et al., 2010]. To provide functional adaptation, TTN has many isoforms formed by alternative splicing [Beckmann et al., 2000]. It consists of nearly 35,000 amino acids and 13 isoforms [Consortium, 2011]. Since the size of the Titin protein is very large and it has too many isoforms, the number of matching mutations is high.

The second protein that has the maximum number of mutations that occur in its isoform is Piccolo (PCLO). Piccolo is a member of presynaptic cytoskeletal protein family and functions in the assembly of presynaptic active zones as sites of neurotransmitter release [Fenster and Garner, 2002]. Although Piccolo is not as huge as Titin, it is regarded as a large protein with approximately 5000 amino acid [Consortium, 2011].The plot of the number of isoform specific mutations and the size of the corresponding protein is shown in Figure 4.5. In Figure 4.5, the number of isoform specific mutations increases with the size of the protein and the diversity of isoform sequences. However, we could not show TTN and PCLO in Figure 4.5 because of their size. The general trend is that proteins having long sequences have tendency to have more isoforms and more isoform specific mutations.

4.2 Structural Mapping of the Mutations

Many methods including electron microscopy, X-ray crystallography and nuclear magnetic resonance is used to extract 3D structure of proteins. These methods do not make it always possible to observe full structures of proteins; rather sometimes only a fragment of a structure can be resolved. Therefore, a single PDB structure may only represent a part of a protein. For example, Figure 4.6 is an illustration of EGFR protein sequence and its corresponding PDB structures. Chain "A" of PDB structure with 1MOX identifier is the structure for sequence position between 25-525 of the full protein. Chain "B" of 1MOX can be a part of another protein, because crystallization can be done in proteins in complex. There is no structural data for the region between 525 and 633. PDB identifier 2KS1 at chain B represents the region between 654 and 677 of EGFR protein. Therefore, in the process of mapping mutation positions to PDB structures of all these pieces of the protein are needed to be considered. Another challenge here is that residue positions at sequence may not match the residue position in the corresponding PDB structure. Hence, a tuning is necessary by aligning the protein sequence and the sequence in the PDB structure. For example, the mutation at 289th position in sequence of EGFR corresponds to 265th position in the PDB structure (in 1MOX chain A). We automatically tuned all the mutation positions in sequence to the available PDB structures by using our pipeline.

In this way, we not only find the correct positions of the mutations in the structure, but also we are able to identify the region in that they are located (the method is described in chapter 3). When we detect mutations at binding interface, on the surface or in the core, we will be able to better assess the effect of mutations. As a result of our calculations, 94 mutations are found to be located in the interface region, 577 mutations are in the core region and 1,182 mutations are located on the surface of the proteins.

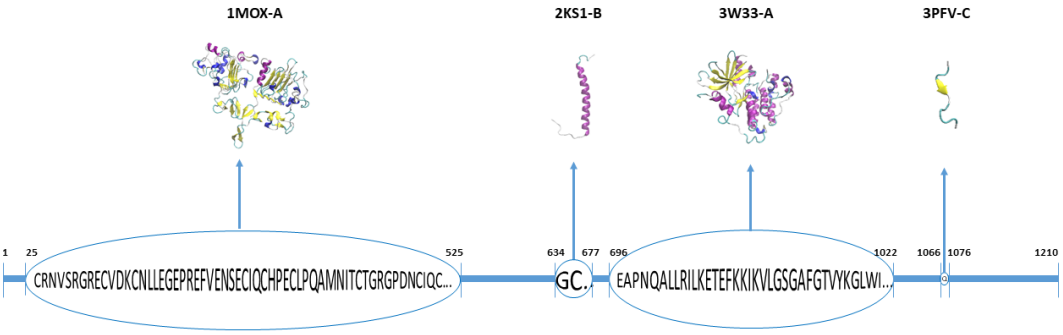

Figure 4.6: Structure-sequence relationship of EGFR protein. 25-525 range of EGFR protein crystallized in the A chain of 1MOX PDB structure. 634-677 range of EGFR protein is structured in B chain of 2KS1 PDB structure. 696-1022 range of EGFR protein observed in A chain of 3W33 structure. 1066-1076 range of EGFR protein crystallized in C chain of 3PFV structure.

4.2.1 Chemical Properties of the Mutations

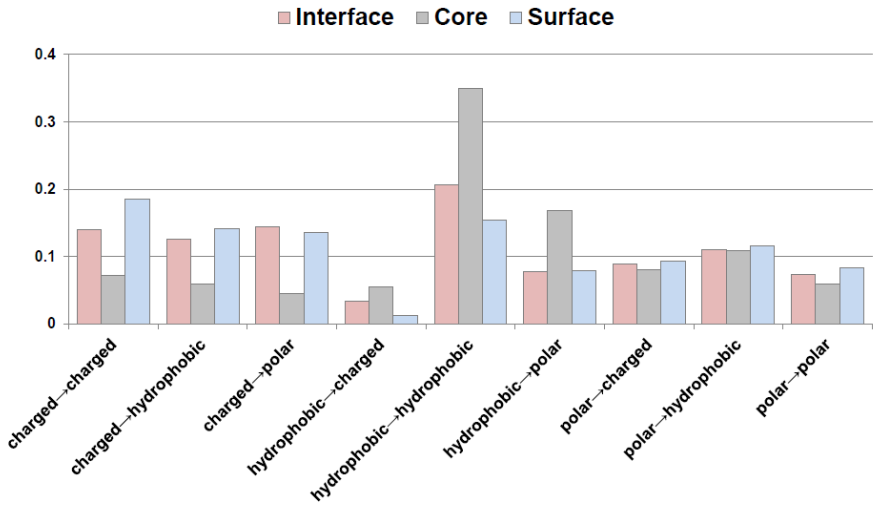

Figure 4.7: Chemical properties of mutations. Grey bar represents core area, red bar represents interface area, blue bar represents surface area. X axis refers to alterations in chemical property, y axis represents the portion of mutations.

Amino acids can be divided into three classes based on their chemical properties: charged, polar, hydrophobic. In this part, we analyzed the frequency of the changes in the chemical properties of the mutated residues from their wild types. Additionally, we classified the mutations based on their location; interface, surface or core. This classification is shown in Figure 4.7 where the x-axis represents the change from one chemical class to another and y-axis represents the fraction of the mutations.

Mutated amino acids on interface and surface regions mostly have charged residues changing to another charged residue that keeps its chemical property. Amino acids that are mutated to amino acids with same chemical properties may have mild effect in changing binding preferences of proteins. On the other hand, conversion into amino acids that do not have the same chemical properties can result in alterations on binding affinity and stability of proteins. The most significant change in chemical properties of mutated amino acids in core regions is seen in hydrophobic residues. Hydrophobic to polar transformation is the second frequent change in core residues. While charged to charged has the highest percentage in interface residues, charged to polar and charged to hydropic changes are also commonly seen in interfaces. When we look at distribution of chemical changes surface residues, charged to charged and hydrophobic to hydrophobic are the most commonly seen transformations. The changes in amino acid chemical class as a result of mutations may have severe effect on protein interface complementarity.

4.2.2 Regional Distribution of the Mutations in Protein Structures

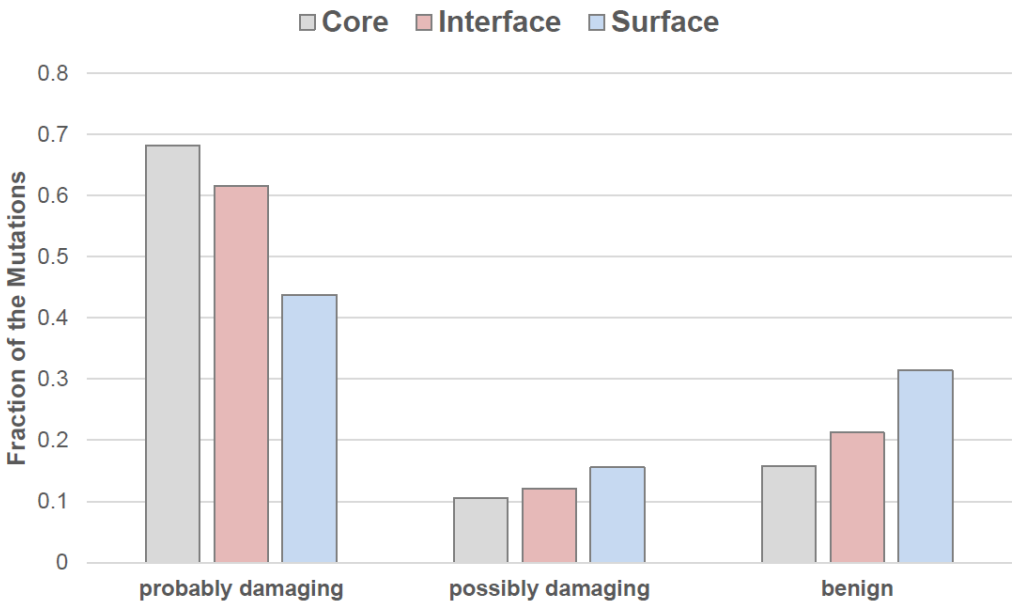

Figure 4.8: PolyPhen-2 mutation analysis output. Grey bar represents core area, red bar represents interface area, blue bar represents surface area. X axis refers to mutations effect to protein, y axis represents the portion of mutations.

The location of the mutated residues in the protein structure gives many more information about the functional impact. We found the functional effect of mutations in proteins by using PolyPhen-2 and classified the impact given by PolyPhen-2 (probably damaging, possibly damaging, benign) according to the location of the residues. As shown in Figure 4.8, mutations in the core region are relatively more damaging compared to the mutations in interface or on the surface. The profile of interface residues are more similar to the core region in terms of the functional effect of mutations. While approximately 40% of mutations on surface region of proteins stated as probably damaging, 68% of mutations on core region of proteins are determined as probably damaging and 62% of mutations on interface regions of proteins are stated as probably damaging.

On the other hand, 30% of surface mutations, 20% of interface mutations and 15% of the core mutations are predicted to be benign. This result implies that mutations in the surface region are relatively more neutral compared to other regions. The interesting part in Figure 4.8 is the set of mutations located in surface and belonging to probably damaging class. Here, one possibility is that these mutations may belong to binding regions that are not identified yet or these mutations make some regions of the protein unfold or be disordered.

4.2.3 Case Study I: SMYD2 - TP53 Complex

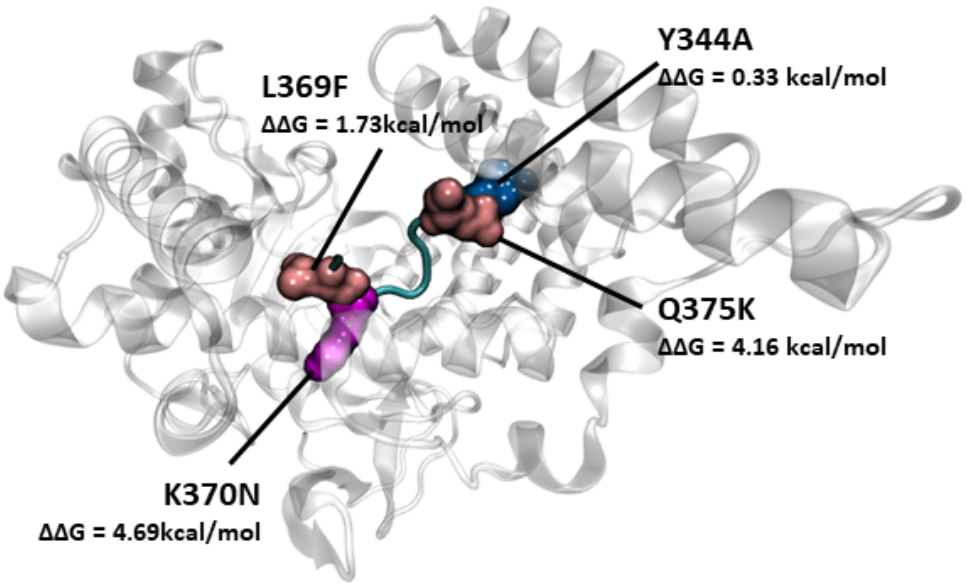

Figure 4.9: The complex of SMYD - TP53 proteins. The grey new cartoon represents the SMYD2 protein and the cyan peptide represents a fragment of TP53.

Among the structurally mapped mutation data, we selected case studies to show the functional impact of the mutations when the structural data is available. Figure 4.9 presents the complex of SMYD2 (Protein-lysine N-methyltransferase) and

TP53. There are 4 mutations that are mapped to the structure of the SMYD2/TP53 complex; namely, Y344A is in SMYD2 and L369F, Q375K and K370N are in TP53. Three of the mutations (K370N, Q375K and L369F) have significant change that has $\Delta\Delta G$ value over 1.5 kcal/mol in binding affinity. SMYD2 is a methyltransferase that methylates many proteins at Lysine residues including TP53. SMYD2 promotes decrease in DNA binding activity of TP53 by monomethylation of Lysine at 370 (K370). Therefore, mutation at position 370 of TP53 from Lysine to Asparagine changes the methylation process of TP53 and in this way, the regulation of TP53 by SMYD2 can be altered which functionally impacts the cellular processes. Addiotionally, residue 375 in TP53 forms hydrogen bonds with the atoms of residues 370, 245, and 345 in SMYD2. A mutation in 375 from Glutamine (uncharged) to Lysine (charged) may change the contact profiles of these residues [Wang et al., 2011].

4.2.4 Case Study II: EGFR-TGFA Complex

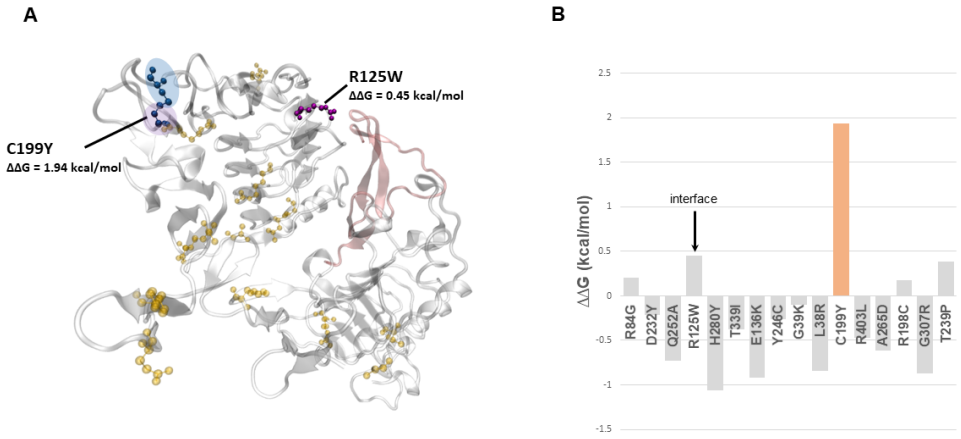

Figure 4.10: The complex of EGFR and TGFA proteins. A) Represents the 3D visualization of complex. Grey structure refers to EGFR and pink structure refers to TGFA. B) Graphical representation of change in binding affinity. X axis represents the mutations and y axis represents the $\Delta\Delta G$ values of mutations.

Our second case study is the interaction between EGFR-TGFA where we show the distribution of the mutations in the structure. We calculated the $\Delta\Delta G$ values of the protein complexes using MutaBind web server. The complex between EGFR-TGFA with the known mutations are shown in Figure 4.10A and the calculated energies are shown in Figure 4.10B. The largest impact in binding is made by the mutation C199Y. The Cysteine residue at position 199 forms a disulfide bond with another Cysteine at position 191 which are shaded in pink and blue in Figure 4.10, respectively. Disulfide bonds are formed between Cysteine residues. They are important for the stability of proteins and they are stronger than any other non-bonded contact

potentials. The mutation at position 199 from CYS to TYR leads to disappearance of the di-sulfide bond between residue 199 and 191 in the same chain. Disappearance of this contact lead to an energy decrease in the binding and significantly affects binding although the mutated residue is not located in the interface region. There is only one mutation (R125W) that is located in the interface region; others are distributed across the surface and core regions of EGFR colored in yellow in Figure 4.10. The impact of the mutation R125W in the interface is neutral and does not significantly affect the binding of EGFR-TGFA which can be represented as a counterexample that a mutation in the surface (C199Y) has more impact in binding affinity than a mutation in the interface.

4.3 Network Modeling to Reveal Patient-Specific Pathways

We performed network analysis using the Omics Integrator. As we mentioned in the previous section, we try to create optimal networks that contain relevant proteins by performing network analysis. For this purpose, Omics Integrator software run for each of 290 patients. For the simulation of each patient, the prize file is formed with the protein name and the number of mutations in that protein. The edge file is a generic weighted interactome that is common for all patients and obtained from iRefWeb database. The outputs of Omics integrator are the network file, edge attributes file, node attributes file and many other files about the information obtained in the intermediate stages of the optimization. As a result of these bulk run, 290 patient-specific networks are reconstructed in total.

4.3.1 An Example Patient Specific Network

In Figure 4.11, the final network of patient barcoded with "TCGA-32-2491-01" is shown. The network has 144 nodes of which 71 are terminal nodes and 73 are Steiner nodes. In general, centrality of Steiner nodes are higher. For example, YWHAG is not in the terminal set but found to be a Steiner node in final network. It is an adapter protein that binds proteins to generally modulate the activity of the partner. Although YWHAG has high centrality network, it is hidden in the initial network of the patient. This situation implies that the nodes that are central for the network and crucial for biological processes can be hidden in the experimental data.

The final network is analyzed with Bingo Cytoscape plugin to determine the enriched pathways. The results of the enrichment analysis are shown in Table 4.1. In the network, many critical pathways are enriched including TNF signaling pathway, focal adhesion, regulation of actin cytoskeleton, ERBB signaling pathway and PI3K-Akt signaling pathway. The most significant pathway is the TNF signaling pathway. Tumor necrosis factor (TNF) is a cytokine that can cause activation of signaling pathways like apoptosis, inflammation and immunity. Another pathway is the local adhesion. Local adhesions are macromolecules that exist between cell and extracellular matrix and they mediate processes including proliferation, differentiation, gene expression and survival. The other significant pathway is PI3K/Akt signaling pathway. Activation of this pathway in Glioblastoma leads to increased cancer cell survival and

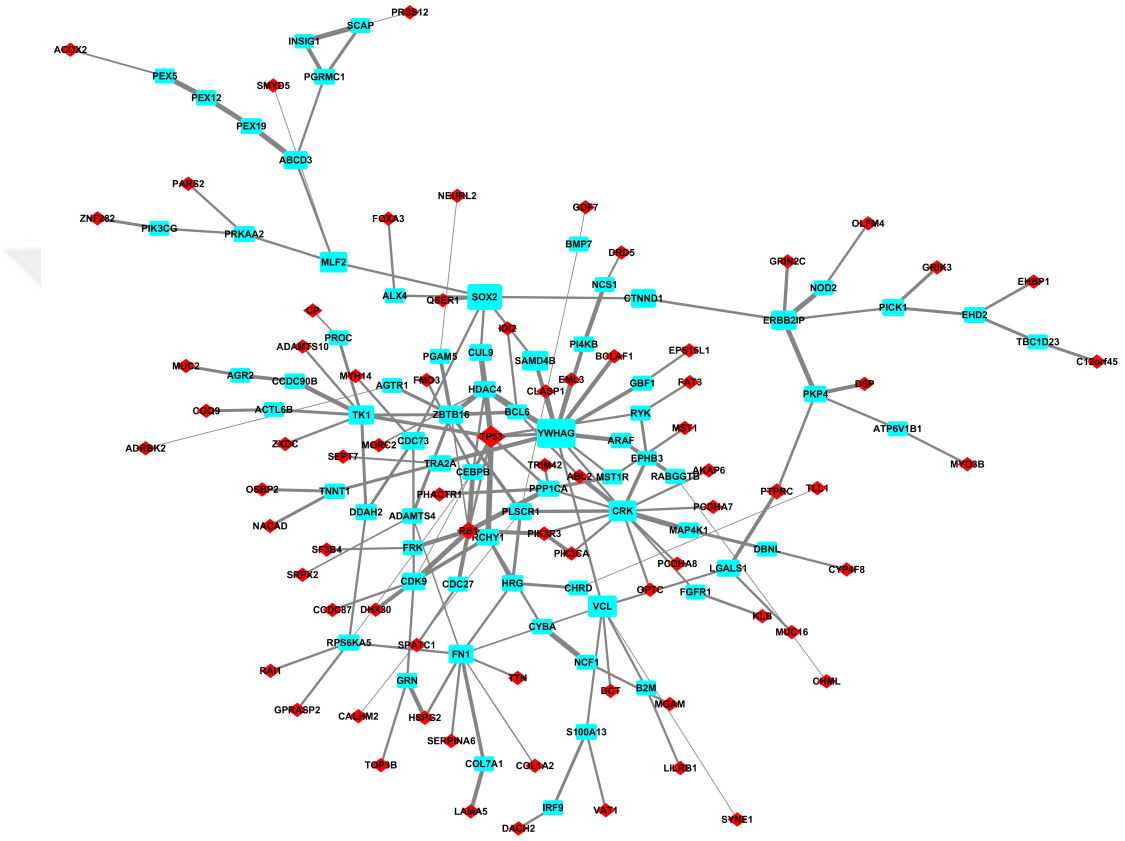

Figure 4.11: Omics Integrator network of patient with TCGA-32-2491-01 barcode. Blue color and rectangle shape represents Steiner nodes, red color and diamond shape represents terminal nodes. Edge thickness is proportional to the edge probability score. Node size is larger if a node is more central.

decreased apoptosis [Mao et al., 2012]. Additionally, activation of ERBB family receptor signaling pathways in GBM, which are enriched in the reconstructed network, can compensate the EGFR targeted therapy and lead to cancer cell proliferation and invasion [Clark et al., 2012]. When we only analyzed the enrichment of the mutation list, we can identify very generic pathways such as cancer pathways. The network modeling gives us more specific pathways and it also shows how the mutated proteins are connected directly or via an intermediate protein.

Table 4.1: Enriched KEGG pathways in the network shown in Figure 4.11

Pathway	**p-value**	**Adjusted p-value**
TNF signaling pathway	0.00072	0.01400
Focal adhesion	0.00120	0.01800
Regulation of actin cytoskeleton	0.00130	0.01600
Fc gamma R-mediated phagocytosis	0.00160	0.01800
ErbB signaling pathway	0.00190	0.01900
Insulin signaling pathway	0.00280	0.02200
Central carbon metabolism in cancer	0.00410	0.02900
Neurotrophin signaling pathway	0.00760	0.04700
PI3K-Akt signaling pathway	0.00840	0.05000

4.3.2 Analysis of Patient-Specific Network Models

As we mentioned in the previous section, we reconstructed 290 networks. These networks have some commonalities and differences when compared to each other. Each network has different number of nodes and edges. The largest network belongs to patient with "TCGA-06-5858-01" barcode and has 229 nodes and 319 edges. Ten patients have network with 0 nodes and 0 edges. The smallest network with 7 nodes and 7 edges belongs to the patient with the barcode "TCGA-06-0139-01". The average number of nodes across all networks is 73.3. The average number of edges across all patients is 81.8. The distribution of node and edge numbers of all networks for each barcode is shown in Figure 4.12 and 4.13, respectively.

Then, we merged 290 reconstructed networks to represent an overall disease network. Here, we applied some thresholds that are described in chapter 3. If an edge in any tumor-specific network is present in at least three patient networks we added that edge to the network. This merged network is shown in Figure 4.14. This merged network elucidates some important features as well. Some of the nodes are included in the network because they are mutated (terminal node) in the patient while some other are intermediate nodes (Steiner node) to connect mutated proteins. For example, PTEN is present in the networks of 70 patients, in 63 out of 70 it is present as a terminal node; however, in 7 patient networks it is present as Steiner node. This information is represented as a pie chart embedded in the nodes in Figure 4.14. This implies that some nodes are still very important although they are not mutated in the given patient. Another example is ATXN1 protein that is present 24 networks of which in 22 networks it is labeled as Steiner and in 2 networks it is labeled as terminal node.

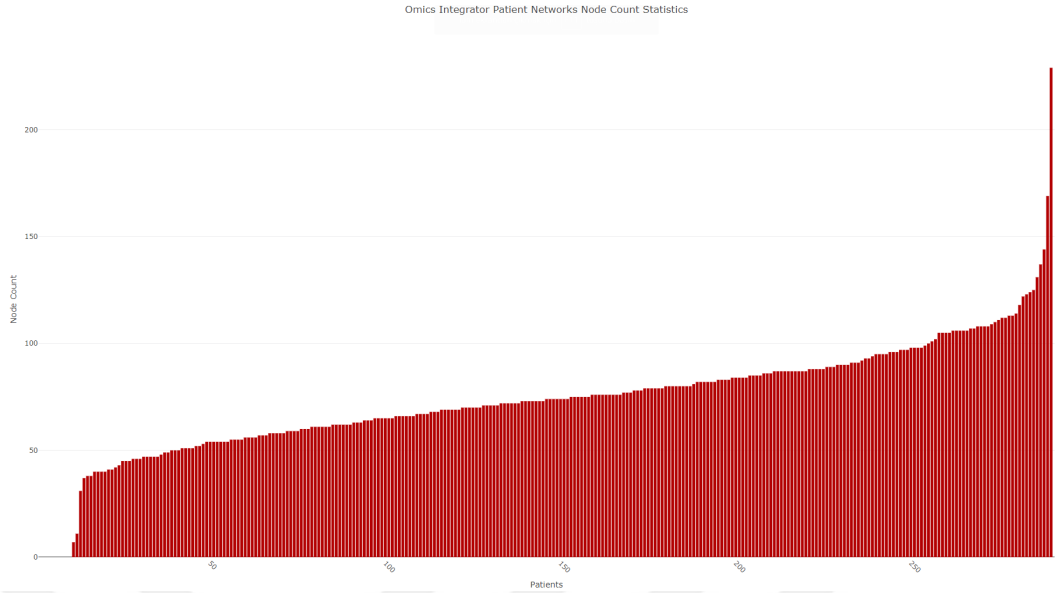

Figure 4.12: Omics Integrator patient networks node count statistics. The distribution of the node counts of the patient networks gathered from the Omics Integrator. The horizontal axis represents 290 patients and the vertical axis refers to node counts.

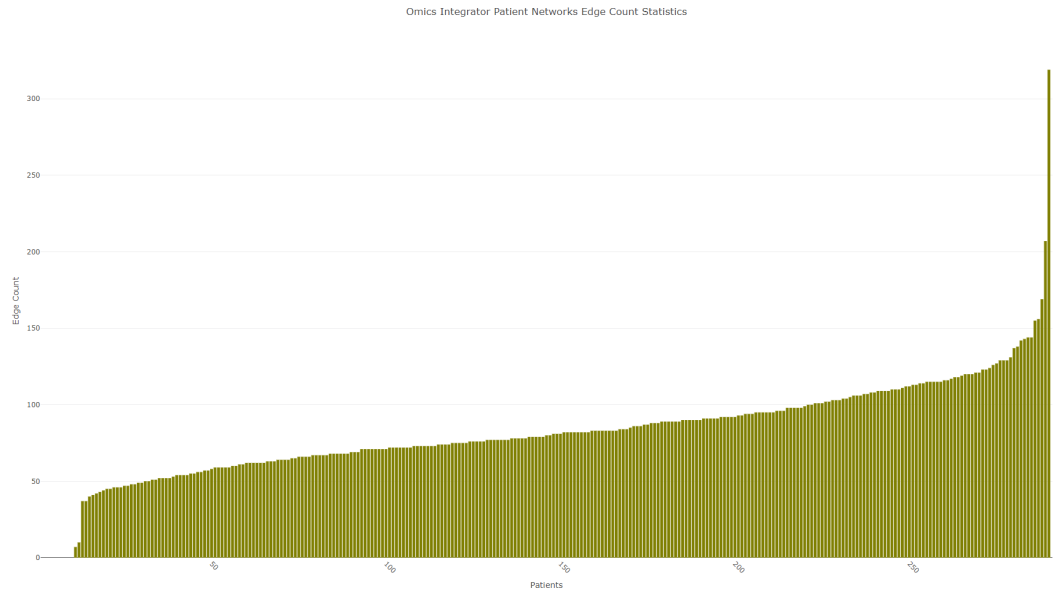

Figure 4.13: Omics Integrator patient networks edge count statistics. The distribution of edge counts of the patient networks gathered from the Omics Integrator. The horizontal axis represents 290 patients and the vertical axis refers to the edge counts.

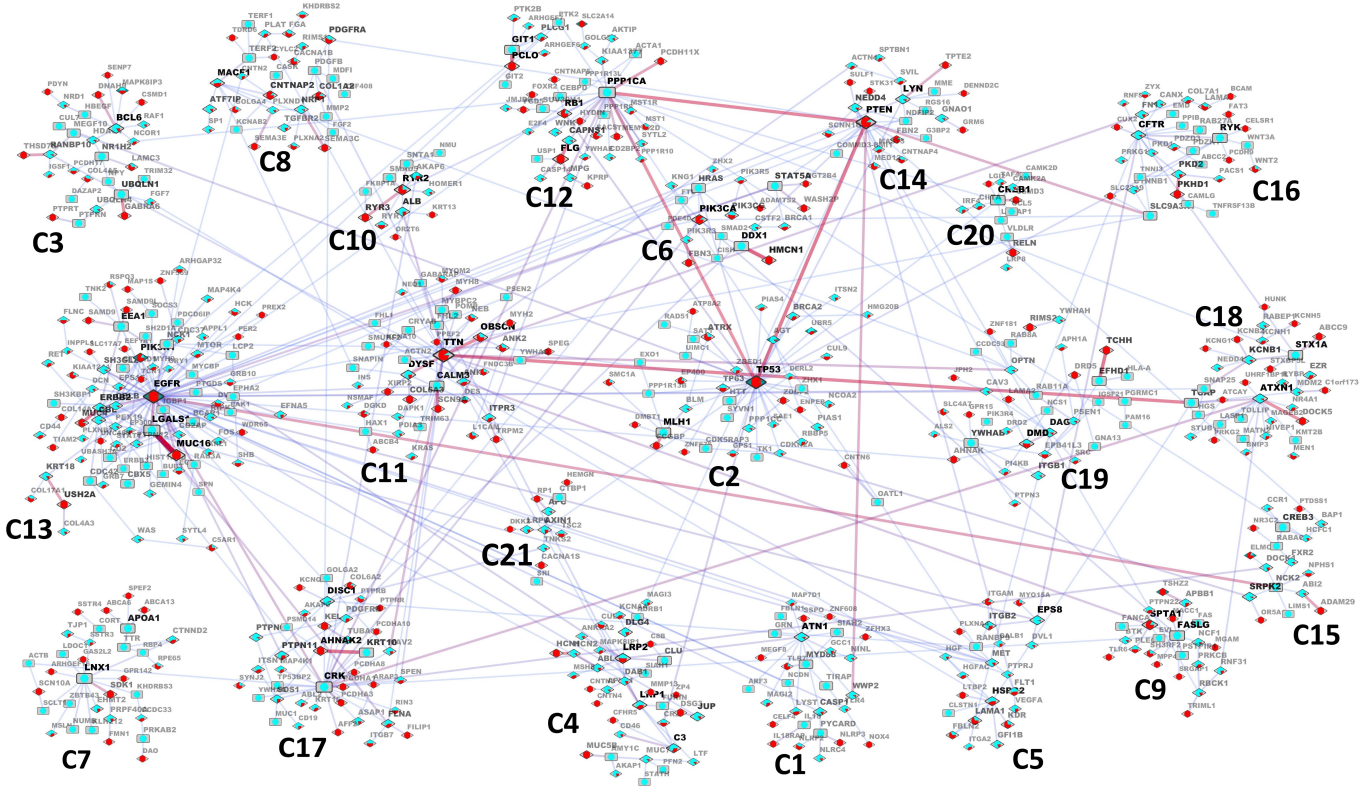

Figure 4.14: The merged network containing the edges present in at least three patients. The nodes are labeled with a pie chart colored in red and/or blue color. The fraction of the red / blue color represents the count of being a terminal / Steiner node in the patient network, respectively. If the shape of the node is a square, this is a Steiner node in all patients. If the shape of the node is a diamond, this node is observed as a terminal in at least one patient network. The edge width is proportional to the number of patients. The blue colored edges present in at least three networks, from blue to red color, the number of patients containing the corresponding edge increases.

The most common interactions in the merged network of 290 patients is the clique between PPP1CA, PTEN and TP53. PTEN (Phosphatase and Tensin Homolog) is known with its tumor suppressor ability. PTEN was shown to be mutated in an abundance of tumour formations. Similarly, TP53 (Tumor Protein p53) is also a tumor suppressor. These two gene products are known to interact with each other. PTEN binds and stabilizes TP53 to induce apoptosis [Patel et al., 2013].

One strong interaction that is common among 35 patients is between LGALS1 (Galectin 1) and MUC16 (Mucin-16). MUC16 binds to galectin-1 [Gubbels et al., 2006]. Galectin-1 participates in the regulation of apoptosis, cell differentiation and proliferation, and it is expressed by human immune cells [Consortium, 2011]. Mucin-16 is an inhibitor of the cytolytic responses of human natural killer cells [Caligiuri, 2008], which belongs to the innate immune system [Caligiuri, 2008]. Especially, Galectin-1 is a potential target in the treatment of Glioblastoma, because it has critical roles in the progression of the tumor [Le Mercier et al., 2010].

4.3.3 Network Functional Enrichment Analysis

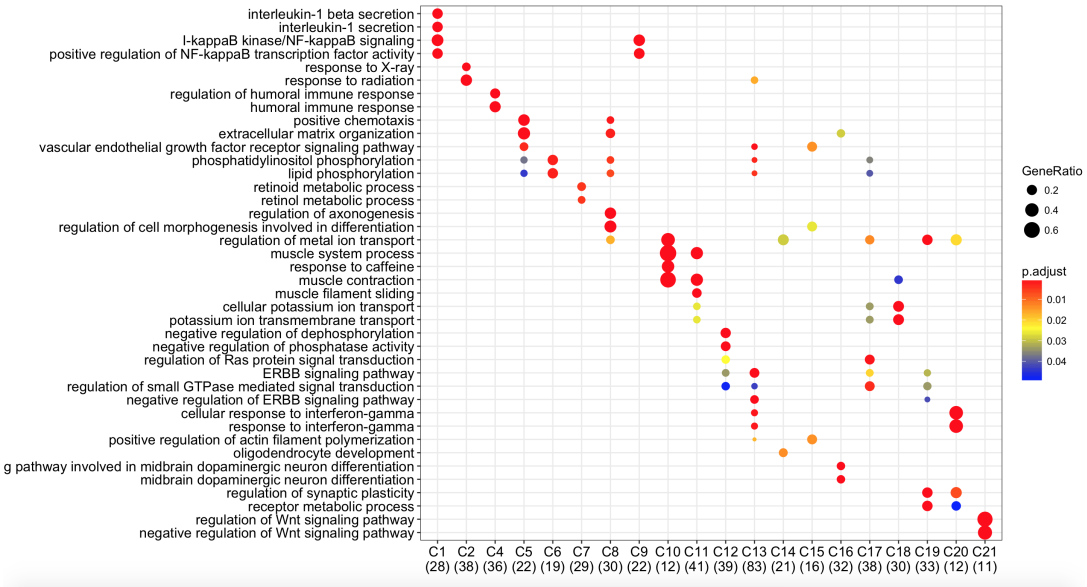

Figure 4.15: GO enrichment results for the gene clusters. Horizontal axis represents the clusters and the vertical axis represents the enriched biological process GO terms. Dots indicate enriched clusters and GO terms at the intersection points. Red color refers high enrichment and blue refers low enrichment according to adjusted p-values. The size of the dots represents the percentage of the genes in the cluster that are annotated with the corresponding GO term.

As observed from Figure 4.15, most of the biological processes are enriched in only one cluster and each cluster has 2 to 8 enriched GO terms. In most cases, terms enriched in a cluster are closely related to each other on the directed acyclic graph of GO, meaning that, they define only one particular biological process. For ex-

ample, interleukin-1 secretion and interleukin-1 beta secretion terms are enriched in the cluster 1. The release of Interleukin-1, a family of cytokines that play roles in the immune and inflammatory responses, from the cell is regulated in these biological processes. In cluster 2, response to radiation and its child term response to X-ray are enriched. Response to radiation is defined as any change that occur in a cell, as a consequence of radiation exposure. In cluster 4, humoral immune response and its child process; regulation of humoral immune response took over. In human infections, bacterias that multiply outside the cell and spread through body fluids, these bacterias are eliminated by the B cells as a part of the humoral immune system [Janeway et al., 2005]. Positive chemotaxis and extracellular matrix organization terms were manifested in cluster 5. Positive chemotaxis term defines the movement of a cell towards higher concentration of a chemical. The extracellular matrix supports organs and tissues structurally. The proteins in the extracellular matrix mostly functions in the proliferation, adhesion and migration of the cell [Hynes, 2009]. The extracellular matrix organization process results in the assembly and disassembly of the extracellular matrix. Cluster 6 is enriched in terms of lipid phosphorylation and its child process phosphatidylinositol phosphorylation. Lipid phosphorylation is the incorporation of one or more phosphate-containing phosphoryl groups into the lipid molecule. In cluster 8, the significant term "regulation of cell morphogenesis" involves morphogenesis-related processes that contribute to cell differentiation, and its child term "regulation of axogenesis" includes processes that contribute to the formation of axons in neuronal cells. Cluster 10 includes the significant processes: regulation of metal ion transport, muscle system process, muscle contraction and response to caffeine. Regulation of metal ion transport process modulates the activities related to the movement of metal ions through the cell. Muscle contraction process is a child of muscle system process and implies the changes in muscle shape. Response to caffeine process defines the activities of a cell due to caffeine intake. In cluster 12, the dominant process is the negative regulation of dephosphorylation. This term describes activities that prevent a phosphate group from being separated from the molecule. In cluster 18, cellular potassium ion transport and its child term potassium ion transmembrane transport processes are enriched. These processes describe the movement of potassium ions through the cell. In cluster 19, the process regulation of synaptic plasticity, which describes the ability of synapses to change according to the situation, is enriched. Receptor metabolic process is also enriched in cluster 19. This term includes the pathways that involves receptors to initiate a change in cellular function. In cluster 20, cellular response to interferon-gamma and its child terms are significant. These terms implies the changes in cell with due to an interferon-gamma stimulus. Wnt signaling pathway is enriched in cluster 21. The process regulates the events of cell migration, cell polarity and cell fate during the embryonic development [Komiya and Habas, 2008]. All listed pathway or biological process descriptions are obtained from Gene Ontology Consortium webpage [Consortium, 2004].

In this analysis, we focused on the non-disease pathways to observe the cellular processes significant in each gene cluster. As shown in Figure 4.16, clusters 5, 6, 8, 13, 16 and 17 have similar enrichments. On than that, most of the clusters have distinct enriched pathways. A few of the clusters does not contain any significantly enriched pathway at all (e.g. clusters 3, 7, 11, 14, 15 and 18). In cluster 1, NOD-like receptor signaling pathway is enriched. NOD-like receptors are defined as nucleotide-binding oligomerization domain-like receptors and include more than 20 members. They are

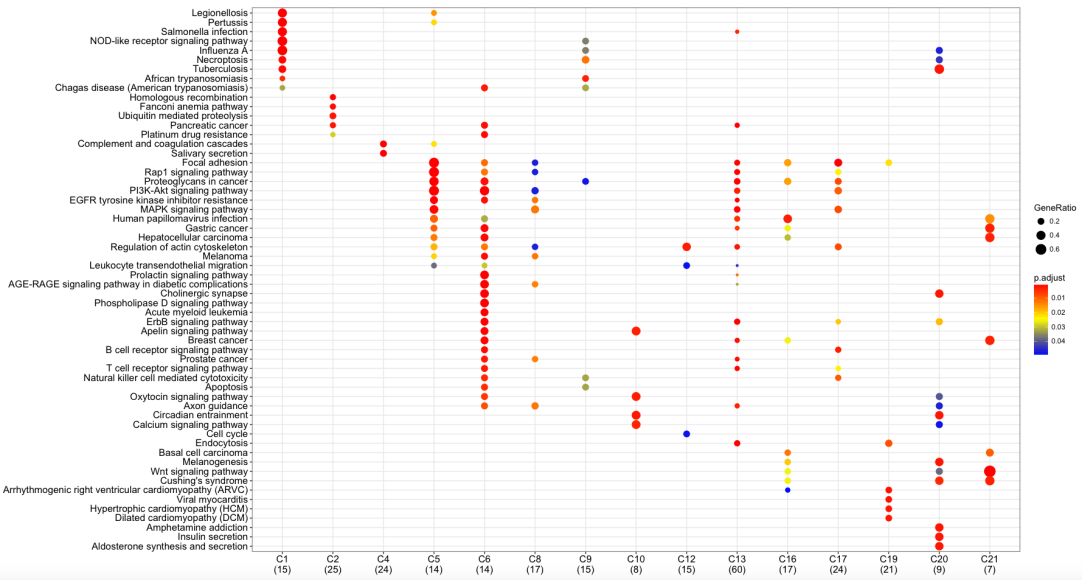

Figure 4.16: KEGG enrichment results of the gene clusters. Horizontal axis represents the clusters and the vertical axis represents the enriched KEGG pathways. Dots indicate enriched clusters and KEGG identifiers at the intersection points. Red color refers high enrichment and blue refers low enrichment according to adjusted p-values. The size of the dots represents the percentage of the genes in the cluster that are annotated with the corresponding KEGG pathway.

pattern identifiers that are responsible for recognizing pathogens and initiating the immune system. In cluster 2, the first enriched pathway “homologous recombination” is a type of genetic recombination and involved in the repair of break on both strands of DNA. In homologous recombination, nucleotide sequences are altered between similar or identical DNA molecules. The second enriched pathway, “Fanconi anemia” takes role in DNA repair, specifically the interstrand cross-links. The third pathway, “ubiquitin mediated proteolysis” is a system that mainly functions in proteasome dependent protein degradation [Ciechanover et al., 2000]. In cluster 4, “complement and coagulation cascades” and “salivary secretion” pathways are enriched. These pathways define a defense mechanism against pathogens and the processes related to the release of saliva, respectively. In clusters 5, 6, 8, 13, 16 and 17, multiple pathways are enriched, three of which are focal adhesion, Rap1 signaling and PI3K/AKT signaling. Focal adhesion describe multi-molecular structures that take role in cell differentiation, proliferation and motility. Rap1 is an enzyme that take part in cell-cell junction formation and cell adhesion. PI3K/AKT signaling pathway controls fundamental process such as cell survival, growth and proliferation. AKT protein, which is activated by the action of PI3K, phosphorylates its substrates to regulate cell cycle. In cluster 10, oxytocin signaling, calcium signaling and circadian entrainment pathways are significant. Oxytocin takes role in lactation and stimulates uterine contraction. Calcium signaling is control calcium ion intake and release to/from the cell. Circadian entrainment pathway takes role in regulating internal biological clock. In cluster 20, insulin secretion and aldosterone synthesis and secretion pathways are enriched. Insulin secretion pathway is crucial for maintaining the homeostasis in the

body. Aldosterone play role in regulating the systemic blood pressure. Finally in cluster 21, Wnt signaling pathway is enriched, as also explained in the previous enrichment analysis with GO biological process terms (Figure 4.15). All listed pathway descriptions are obtained from Kyoto Encyclopedia of Genes and Genome (KEGG) webpage [Kanehisa and Goto, 2000].

4.4 Pathway Enrichment Analysis of the Mutation Sets

We have found several enriched pathways as a result of mutation set enrichment analysis on the given patient data. The heatmap of the significant pathways is shown in Figure 4.17, where columns and rows correspond to patient barcodes and pathways, respectively, and the color intensity in each cell correspond to the negative logarithm of the p-values.

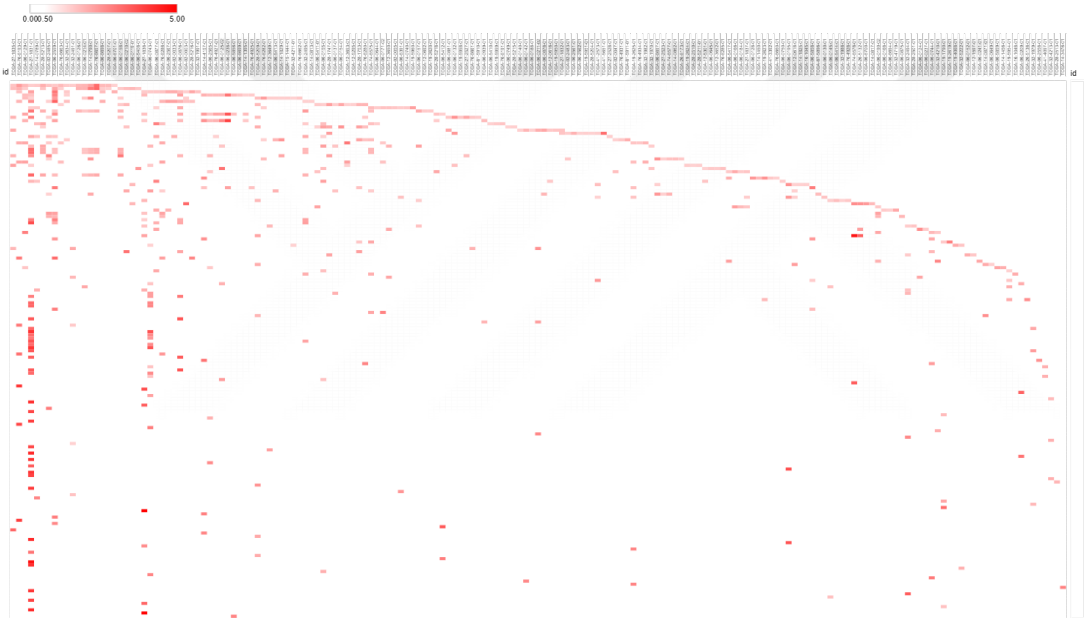

Figure 4.17: The heatmap of patient-pathway sets. The columns are patient barcodes and rows are pathway identifiers from the Reactome DB. The color intensity implies the significance of the enrichment.

The most similar patient-pathway set are observed in three groups and these groups details are explained below. All pathway definitions in this analysis are obtained from the Reactome DB [Joshi-Tope et al., 2005].

The first subset of the Reactome pathway analysis results are shown in Figure 4.18. Patients with the most similar features, with the barcodes TCGA-14-1395-01 and TCGA-76-4927-01, display similarities on three pathways. The first of these pathways is the RNA Polymerase 2 Transcription. RNA Polymerase 2 is the main enzyme that catalyzes DNA directed mRNA synthesis during the transcription of genes. The second common pathway is the gene expression pathway. This includes both the transcription and translation processes. In these processes, while RNA Polymerase 1 is

involved in the synthesis of rRNA, RNA Polymerase 2 is involved in messenger RNA synthesis and RNA polymerase 3 is involved in the synthesis of tRNA and its derivatives. The third common pathway is the generic transcription pathway. A high level patient cluster, which is obtained by expanding from the first cluster by adding the patients with the most similar patterns, contains the patients with barcodes TCGA-06-2565-01, TCGA-06-0129-01, TCGA-06-0238-01 and TCGA-06-0125-02, along with TCGA-14-1395-01 and TCGA-76-4927-01. With the addition of these patients, there are two new pathways in the cluster. The first one is the metabolism of proteins. This pathway includes all processes from the synthesis of proteins and the post translational modification to their degradation. The second pathway is the cell cycle. The cell cycle contains all the processes that occur during genetic replication and the distribution of chromosomes into daughter cells.[Joshi-Tope et al., 2005]

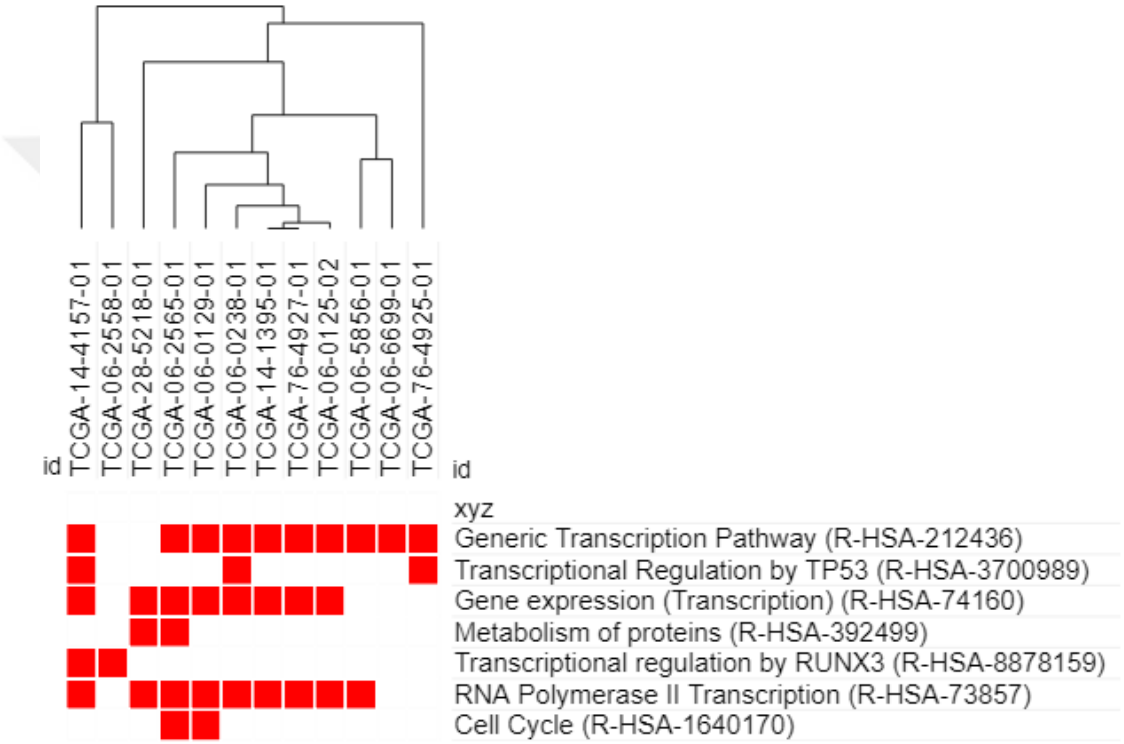

Figure 4.18: The first subset of patient-pathway enrichment analysis.

The second subset of the patient-pathway enrichment in Reactome is shown in Figure 4.19. Patients with the barcodes TCGA-14-0790-01 and TCGA-06-0214-01 have seven pathways in common. First of them is the cytokine signaling in immune system. Cytokines bind to specific membrane receptors and they regulate cellular activities by signaling. They are the molecules that are responsible for the regulation and mediation of immunity, inflammation and hematopoiesis. The second of the common pathways is the signaling by interleukins. Interleukins take role in many aspects of the cell such as the tissue growth, repair, homeostasis and host defence against pathogens. In addition to these pathways, RAF/MAP kinase cascade pathway participate in the regulation of cell processes including differentiation, survival, proliferation by responding growth factor and hormones. In MAPK family signaling cascade pathway involves mitogen activated protein kinases, which is a protein family that responses

to several extracellular signals. These proteins activate a variety of cellular activities including gene expression, proliferation and apoptosis.[Joshi-Tope et al., 2005]

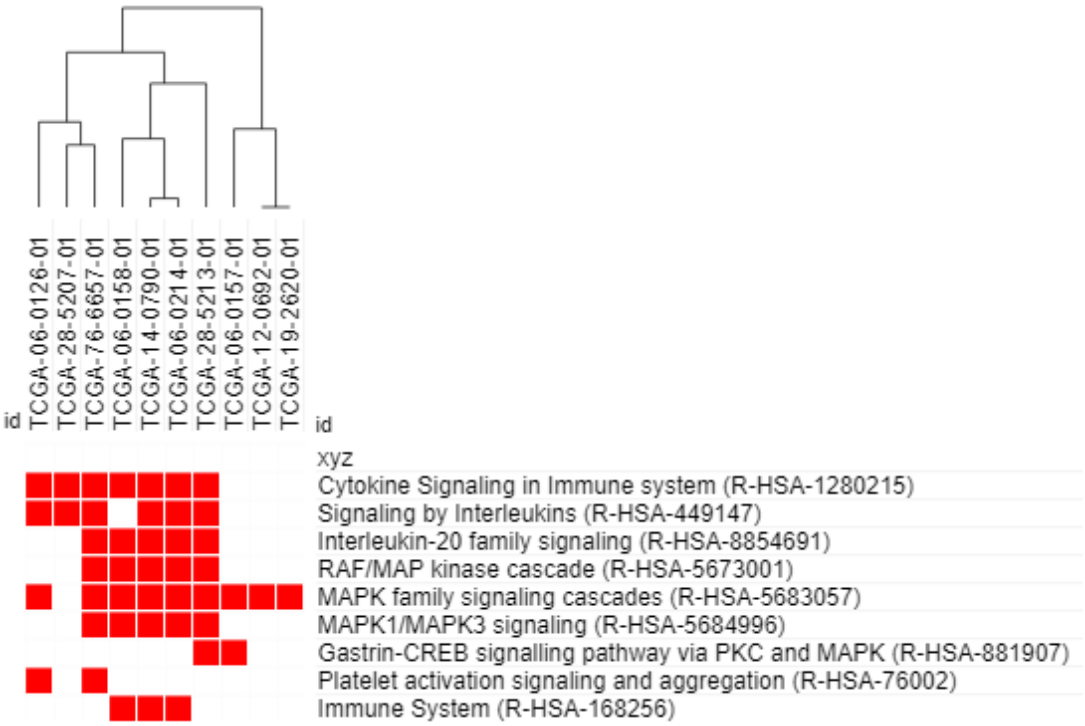

Figure 4.19: The second subset of patient-pathway enrichment analysis.

The third subset of the patient-pathway analysis is shown in Figure 4.20. PIP3 activates AKT Signaling pathway is a participant of intracellular signaling by second messengers. Extracellular signals stimulate the cell surface receptors and the second messengers generated within the cell as a result of this process. The other pathway is the signaling by Interleukins. The cellular responses to stress pathway defines the signaling processes in the cell for maintaining homeostasis, as a response to external stimuli. This pathway is the descendent of the cellular response to external stimuli. Cellular senescence is a descendent of the cellular responses to stress. [Joshi-Tope et al., 2005]

A similar analysis has been also applied to KEGG pathways which are shown in Appendix B.

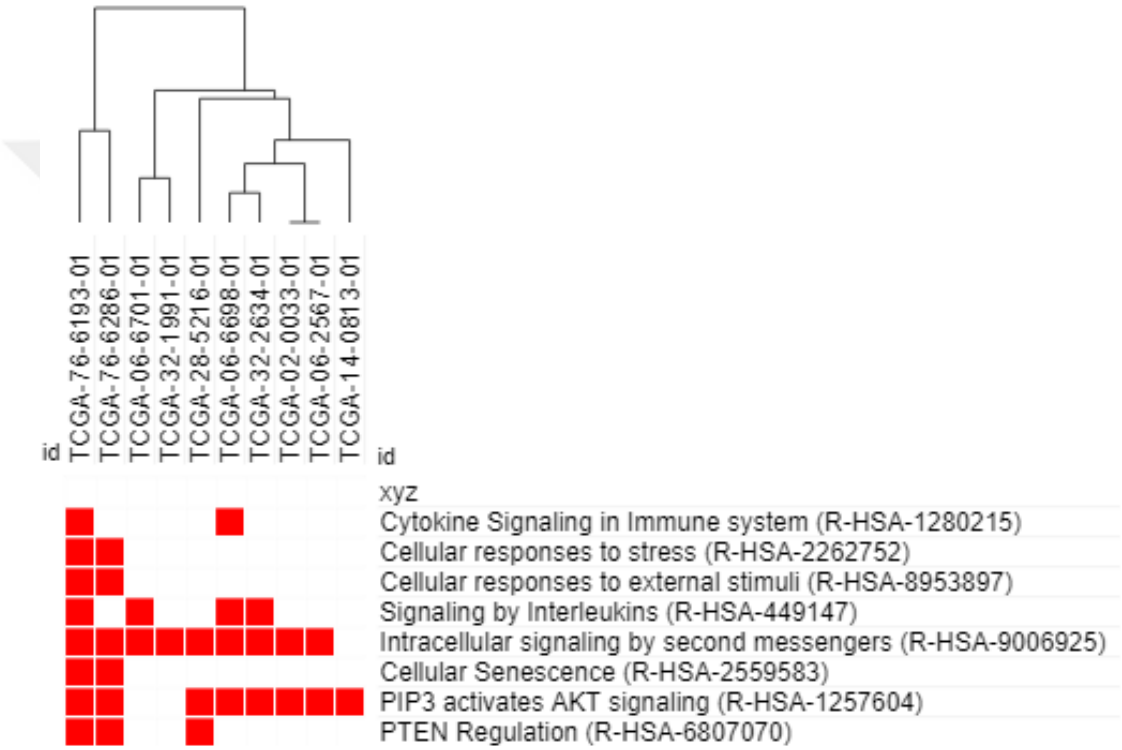

Figure 4.20: The third subset of patient-pathway enrichment analysis.

CHAPTER 5

DISCUSSION AND CONCLUSION

5.1 Concluding Remarks

In this thesis, we studied the mutations in Glioblastoma patients to gain a better understanding of how the mutations are distributed in protein structure, how they affect the protein stability and interaction preferences and through which interactions the mutated proteins are connected to each other. Thus, we present an analysis pipeline to interpret mutations in Glioblastoma in a patient-specific way. We used the missense mutations in GBM patients deposited in TCGA. The preliminary analysis of the list of mutations has shown that the GBM tumors are very heterogeneous and the number of mutations that are common in at least three patients is only 62 out of 14,644 unique mutations. We also note that some of the mutations are not found in the canonical sequence of corresponding protein, and rather they are present in the alternative isoforms. This type of mutations are only a small portion (433 mutations) compared to the full set.

Because the mutation profiles of the patients are very heterogeneous and their links to each other are not apparent from the mutation set, there is a need to model the interplay between the mutations in a network context beyond the list. Additionally, the effect of the mutations are not uniform across the proteins and patients. So, we first started by analyzing the mutations and mapping the mutation positions onto the available protein structure data. In this way, we are able to identify the region of the mutations in the protein structure; which is divided into three regions, surface, interface and core. When we compared the alteration of the chemical classes from wild type to mutated residue we have seen that mutations located in the core region have tendency to keep its chemical class (i.e., from hydrophobic to another hydrophobic residue). The next most frequent alteration is from hydrophobic to polar residue, which may have a more severe effect on the protein stability. When we refer to the mutated residues in the interface region we notice that charged to polar and polar to charged alterations are frequent. The most mutated residues in the interface region is the charged ones and they are either keep their chemical class or change to hydrophobic or polar residues. These results show that most of the mutations led to also chemical property change in proteins and this profile is more similar to the surface region. Core region has more tendency to conserve the chemical properties despite the mutation. Next, we predicted the functional effect of the mutations using PolyPhen-2 and assess their effect in three regions. We found that mutations, that are located in the core and interface regions, are more damaging. It is expected that the mutations in the core and interface to be more damaging compared to the rest of the surface; how-

ever, we have specifically shown that this is valid in Glioblastoma mutations in our results. Additionally, mutations that are located in the surface and damaging are an interesting class of mutations, because they may lead to disorder or belong to unidentified binding regions that needs to be analyzed in more detail. Two case studies - one is the SMYD2-TP53 complex and the other is the EGFR-TGFA complex - have shown the effect of mutations are not uniform. In the former the methylation profile of TP53 protein is expected to be changing as a result of mutation, and in the latter one a mutation is changing the protein stability with a broken di-sulfide bond and eventually the interaction preference, which is located in the surface.

In terms of pathway enrichment, we first applied a naive method on the individual mutation set for each patient and found the enriched pathways. The resulting significant pathways were shown to be very generic; for example, transcription pathway, cancer pathways etc. Then, we reconstructed a network of interactions between mutated proteins in each patient. Here, our aim is to connect mutated proteins either directly or by adding an intermediate node through high probability edges. We used the Omics Integrator software for this purpose which solves the prize-collecting Steiner forest problem. As a result we reconstructed 290 patient-specific network. Beyond the mutation set enrichment analysis, the results of network modeling gave us more specific pathways that are enriched in the given patient barcode. For example, the network with barcode number TCGA-32-2491-01 is significantly enriched in TNF signaling pathway, focal adhesion, regulation of actin cytoskeleton, ERBB signaling pathway and PI3K-Akt signaling pathway, which are more specific and related to cancer. Additionally, network analysis gave us some intermediate proteins as well which are not in the mutated proteins set but they are connecting the mutated proteins with highly confident protein interactions. Then, we compared these 290 networks based on their commonalities and constructed a merged network by using the interactions that are present in at least three patients. This merged network gave us the proteins that are mutated in a set of patients but also plays a role as an intermediate protein in other set of patients. Thus, we can conclude that although this type of proteins are not mutated in some cases, they are still important as an intermediate node and could be good clinical targets. We divided the merged network into communities based on the topology of the network. Then, we searched for the significantly enriched pathways and biological processes. We noted that multiple biological processes and pathways are enriched in each community. While WNT signaling is enriched in one community, ERBB signaling is enriched in another, and we can show the interplay between these two pathways with our merged network.

This pipeline is designed to be applicable any set of mutations including breast cancer, ovarian cancer, prostate cancer etc. The required input is just the set of mutations for each patient in the corresponding cancer type.

5.2 Future Work

From mutations to protein structure mappings and constructing the functional networks, we believe that the result of this study will have significant contribution in cancer research. We present these results as a proof-of-concept that the effect of mutations are not uniform within individual protein structures, as well as within their

interaction networks. Although this study have some missing modules that would improve the results, these analyses are planned to be performed in future studies. In this study, we applied the method only to the GBM patients data in TCGA. Our near future plan is to apply it to other cancer types including ovarian cancer and breast cancer. In this way, we will be able to make a cross-cancer comparison. Also, we used the structural data in PDB in this study. Using homology modeling techniques the available structural human proteome data can be increased by 50%. To this end, another future aim is to include the predicted structures of the proteins that do not have a known structure in PDB. Additionally, recent computational techniques such as PRISM and Interactome3D, are reported to be successful in predicting 3D structures of protein interactions. In the future, we plan to integrate the predicted interface information into our analysis to have a better coverage regarding the effect of mutations.

REFERENCES

[Adzhubei et al., 2013] Adzhubei, I., Jordan, D. M., and Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using polyphen-2. *Current protocols in human genetics*, 76(1):7–20.

[Akavia et al., 2010] Akavia, U. D., Litvin, O., Kim, J., Sanchez-Garcia, F., Kotliar, D., Causton, H. C., Pochanard, P., Mozes, E., Garraway, L. A., and Pe'er, D. (2010). An integrated approach to uncover drivers of cancer. *Cell*, 143(6):1005–1017.

[Ashburner et al., 2000] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25.

[Bader et al., 2003] Bader, G. D., Betel, D., and Hogue, C. W. (2003). Bind: the biomolecular interaction network database. *Nucleic acids research*, 31(1):248–250.

[Bailly-Bechet et al., 2011] Bailly-Bechet, M., Borgs, C., Braunstein, A., Chayes, J., Dagkessamanskaia, A., François, J.-M., and Zecchina, R. (2011). Finding undetected protein associations in cell signaling by belief propagation. *Proceedings of the National Academy of Sciences*, 108(2):882–887.

[Bailly-Bechet et al., 2010] Bailly-Bechet, M., Braunstein, A., Pagnani, A., Weigt, M., and Zecchina, R. (2010). Inference of sparse combinatorial-control networks from gene-expression data: a message passing approach. *BMC bioinformatics*, 11(1):355.

[Beckmann et al., 2000] Beckmann, J. S., Gregorio, C. C., et al. (2000). Series of exon-skipping events in the elastic spring region of titin as the structural basis for myofibrillar elastic diversity. *Circulation research*.

[Ben-Gal et al., 2007] Ben-Gal, I., Ruggeri, F., Faltin, F., and Kenett, R. (2007). Bayesian networks. encyclopedia of statistics in quality and reliability.

[Berman et al., 2000] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The protein data bank. *Nucleic acids research*, 28(1):235–242.

[Berman et al., 2006] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2006). The protein data

bank, 1999–. In *International Tables for Crystallography Volume F: Crystallography of biological macromolecules*, pages 675–684. Springer.

[Caligiuri, 2008] Caligiuri, M. A. (2008). Human natural killer cells. *Blood*, 112(3):461–469.

[Carlin et al., 2017] Carlin, D. E., Demchak, B., Pratt, D., Sage, E., and Ideker, T. (2017). Network propagation in the cytoscape cyberinfrastructure. *PLoS computational biology*, 13(10):e1005598.

[Chatr-Aryamontri et al., 2006] Chatr-Aryamontri, A., Ceol, A., Palazzi, L. M., Nardelli, G., Schneider, M. V., Castagnoli, L., and Cesareni, G. (2006). Mint: the molecular interaction database. *Nucleic acids research*, 35(suppl_1):D572–D574.

[Ciechanover et al., 2000] Ciechanover, A., Orian, A., and Schwartz, A. L. (2000). Ubiquitin-mediated proteolysis: biological regulation via destruction. *Bioessays*, 22(5):442–451.

[Clark et al., 2012] Clark, P. A., Iida, M., Treisman, D. M., Kalluri, H., Ezhilan, S., Zorniak, M., Wheeler, D. L., and Kuo, J. S. (2012). Activation of multiple erbb family receptors mediates glioblastoma cancer stem-like cell resistance to egfr-targeted inhibition. *Neoplasia*, 14(5):420–428.

[Cock et al., 2009] Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. (2009). Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423.

[Consortium, 2004] Consortium, G. O. (2004). The gene ontology (go) database and informatics resource. *Nucleic acids research*, 32(suppl_1):D258–D261.

[Consortium et al., 2010] Consortium, I. C. G. et al. (2010). International network of cancer genome projects. *Nature*, 464(7291):993.

[Consortium, 2011] Consortium, U. (2011). Reorganizing the protein space at the universal protein resource (uniprot). *Nucleic acids research*, 40(D1):D71–D75.

[Dantzig and Thapa, 2006] Dantzig, G. B. and Thapa, M. N. (2006). *Linear programming 1: introduction*. Springer Science & Business Media.

[De Smet and Marchal, 2010] De Smet, R. and Marchal, K. (2010). Advantages and limitations of current network inference methods. *Nature Reviews Microbiology*, 8(10):717.

[Dickson, 1999] Dickson, D. (1999). Wellcome funds cancer database.

[Dittrich et al., 2008] Dittrich, M. T., Klau, G. W., Rosenwald, A., Dandekar, T., and Müller, T. (2008). Identifying functional modules in protein–protein interaction networks: an integrated exact approach. *Bioinformatics*, 24(13):i223–i231.

[Dutta et al., 2009] Dutta, S., Burkhardt, K., Young, J., Swaminathan, G. J., Matsuura, T., Henrick, K., Nakamura, H., and Berman, H. M. (2009). Data deposition and annotation at the worldwide protein data bank. *Molecular biotechnology*, 42(1):1–13.

[Fenster and Garner, 2002] Fenster, S. D. and Garner, C. C. (2002). Gene structure and genetic localization of the pclo gene encoding the presynaptic active zone protein piccolo. *International journal of developmental neuroscience*, 20(3-5):161–171.

[Freeman, 1978] Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239.

[Girvan and Newman, 2002] Girvan, M. and Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826.

[Gonzalez-Perez et al., 2013] Gonzalez-Perez, A., Mustonen, V., Reva, B., Ritchie, G. R., Creixell, P., Karchin, R., Vazquez, M., Fink, J. L., Kassahn, K. S., Pearson, J. V., et al. (2013). Computational approaches to identify functional genetic variants in cancer genomes. *Nature methods*, 10(8):723.

[Gubbels et al., 2006] Gubbels, J. A., Belisle, J., Onda, M., Rancourt, C., Migneault, M., Ho, M., Bera, T. K., Connor, J., Sathyanarayana, B. K., Lee, B., et al. (2006). Mesothelin-muc16 binding is a high affinity, n-glycan dependent interaction that facilitates peritoneal metastasis of ovarian tumors. *Molecular cancer*, 5(1):50.

[Guo et al., 2010] Guo, W., Bharmal, S. J., Esbona, K., and Greaser, M. L. (2010). Titin diversity—alternative splicing gone wild. *BioMed Research International*, 2010.

[Hagberg et al., 2005] Hagberg, A., Schult, D., and Swart, P. (2005). Networkx: Python software for the analysis of networks. *Mathematical Modeling and Analysis, Los Alamos National Laboratory*.

[Hopf et al., 2017] Hopf, T. A., Ingraham, J. B., Poelwijk, F. J., Schärfe, C. P., Springer, M., Sander, C., and Marks, D. S. (2017). Mutation effects predicted from sequence co-variation. *Nature biotechnology*, 35(2):128.

[Huang and Fraenkel, 2009] Huang, S.-s. C. and Fraenkel, E. (2009). Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks. *Sci. Signal.*, 2(81):ra40–ra40.

[Hubbard, 1992] Hubbard, S. (1992). Naccess: program for calculating accessibilities. *Department of Biochemistry and Molecular Biology, University College of London*.

[Hubbard et al., 1991] Hubbard, S., Campbell, S., and Thornton, J. (1991). Molecular recognition: conformational analysis of limited proteolytic sites and serine proteinase protein inhibitors. *Journal of molecular biology*, 220(2):507–530.

[Hynes, 2009] Hynes, R. O. (2009). The extracellular matrix: not just pretty fibrils. *Science*, 326(5957):1216–1219.

[Janeway et al., 2005] Janeway, C. A., Travers, P., Walport, M., and Shlomchik, M. J. (2005). Immunobiology: the immune system in health and disease.

[Joshi-Tope et al., 2005] Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G., Wu, G., Matthews, L., et al. (2005). Reactome: a knowledgebase of biological pathways. *Nucleic acids research*, 33(suppl_1):D428–D432.

[Kandoth et al., 2013] Kandoth, C., McLellan, M. D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J. F., Wyczalkowski, M. A., et al. (2013). Mutational landscape and significance across 12 major cancer types. *Nature*, 502(7471):333.

[Kanehisa and Goto, 2000] Kanehisa, M. and Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30.

[Kanehisa et al., 2009] Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M. (2009). Kegg for representation and analysis of molecular networks involving diseases and drugs. *Nucleic acids research*, 38(suppl_1):D355–D360.

[Kelley et al., 2003] Kelley, B. P., Sharan, R., Karp, R. M., Sittler, T., Root, D. E., Stockwell, B. R., and Ideker, T. (2003). Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proceedings of the National Academy of Sciences*, 100(20):11394–11399.

[Keshava Prasad et al., 2008] Keshava Prasad, T., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., et al. (2008). Human protein reference database—2009 update. *Nucleic acids research*, 37(suppl_1):D767–D772.

[Keskin et al., 2004] Keskin, O., Tsai, C.-J., Wolfson, H., and Nussinov, R. (2004). A new, structurally nonredundant, diverse data set of protein–protein interfaces and its implications. *Protein Science*, 13(4):1043–1055.

[Keskin et al., 2016] Keskin, O., Tuncbag, N., and Gursoy, A. (2016). Predicting protein–protein interactions from the molecular to the proteome level. *Chemical reviews*, 116(8):4884–4909.

[Komiya and Habas, 2008] Komiya, Y. and Habas, R. (2008). Wnt signal transduction pathways. *Organogenesis*, 4(2):68–75.

[Lan et al., 2011] Lan, A., Smoly, I. Y., Rapaport, G., Lindquist, S., Fraenkel, E., and Yeger-Lotem, E. (2011). Responsenet: revealing signaling and regulatory networks linking genetic and transcriptomic screening data. *Nucleic acids research*, 39(suppl_2):W424–W429.

[Le Mercier et al., 2010] Le Mercier, M., Fortin, S., Mathieu, V., Kiss, R., and Lefranc, F. (2010). Galectins and gliomas. *Brain pathology*, 20(1):17–27.

[Lee and Richards, 1971] Lee, B. and Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *Journal of molecular biology*, 55(3):379–IN4.

[Li et al., 2016] Li, M., Simonetti, F. L., Goncearenco, A., and Panchenko, A. R. (2016). Mutabind estimates and interprets the effects of sequence variants on protein–protein interactions. *Nucleic acids research*, 44(W1):W494–W501.

[Ljubić et al., 2006] Ljubić, I., Weiskircher, R., Pferschy, U., Klau, G. W., Mutzel, P., and Fischetti, M. (2006). An algorithmic framework for the exact solution of the prize-collecting steiner tree problem. *Mathematical programming*, 105(2-3):427–449.

[Mao et al., 2012] Mao, H., LeBrun, D. G., Yang, J., Zhu, V. F., and Li, M. (2012). Deregulated signaling pathways in glioblastoma multiforme: molecular mechanisms and therapeutic targets. *Cancer investigation*, 30(1):48–56.

[Missiuro et al., 2009] Missiuro, P. V., Liu, K., Zou, L., Ross, B. C., Zhao, G., Liu, J. S., and Ge, H. (2009). Information flow analysis of interactome networks. *PLoS computational biology*, 5(4):e1000350.

[Nishi et al., 2013] Nishi, H., Tyagi, M., Teng, S., Shoemaker, B. A., Hashimoto, K., Alexov, E., Wuchty, S., and Panchenko, A. R. (2013). Cancer missense mutations alter binding properties of proteins and their interaction networks. *PloS one*, 8(6):e66273.

[Ourfali et al., 2007] Ourfali, O., Shlomi, T., Ideker, T., Ruppin, E., and Sharan, R. (2007). Spine: a framework for signaling-regulatory pathway inference from cause-effect experiments. *Bioinformatics*, 23(13):i359–i366.

[Patel et al., 2013] Patel, R., Gao, M., Ahmad, I., Fleming, J., Singh, L. B., Rai, T. S., McKie, A. B., Seywright, M., Barnetson, R. J., Edwards, J., et al. (2013). Sprouty2, pten, and pp2a interact to regulate prostate cancer progression. *The Journal of clinical investigation*, 123(3):1157–1175.

[Ruepp et al., 2009] Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Mewes, H.-W. (2009). Corum: the comprehensive resource of mammalian protein complexes—2009. *Nucleic acids research*, 38(suppl_1):D497–D501.

[Sharan and Ideker, 2006] Sharan, R. and Ideker, T. (2006). Modeling cellular machinery through biological network comparison. *Nature biotechnology*, 24(4):427.

[Sim et al., 2012] Sim, N.-L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., and Ng, P. C. (2012). Sift web server: predicting effects of amino acid substitutions on proteins. *Nucleic acids research*, 40(W1):W452–W457.

[Stark et al., 2006] Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). Biogrid: a general repository for interaction datasets. *Nucleic acids research*, 34(suppl_1):D535–D539.

[Subramanian et al., 2005] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550.

[Szklarczyk et al., 2016] Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N. T., Roth, A., Bork, P., et al. (2016). The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic acids research*, page gkw937.

[Tomczak et al., 2015] Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). The cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary oncology*, 19(1A):A68.

[Tuncbag et al., 2013] Tuncbag, N., Braunstein, A., Pagnani, A., Huang, S.-S. C., Chayes, J., Borgs, C., Zecchina, R., and Fraenkel, E. (2013). Simultaneous reconstruction of multiple signaling pathways via the prize-collecting steiner forest problem. *Journal of computational biology*, 20(2):124–136.

[Tuncbag et al., 2016] Tuncbag, N., Gosline, S. J., Kedaigle, A., Soltis, A. R., Gitter, A., and Fraenkel, E. (2016). Network-based interpretation of diverse high-throughput datasets through the omics integrator software package. *PLoS computational biology*, 12(4):e1004879.

[Tuncbag et al., 2008] Tuncbag, N., Gursoy, A., Guney, E., Nussinov, R., and Keskin, O. (2008). Architectures and functional coverage of protein–protein interfaces. *Journal of molecular biology*, 381(3):785–802.

[Turner et al., 2010] Turner, B., Razick, S., Turinsky, A. L., Vlasblom, J., Crowdy, E. K., Cho, E., Morrison, K., Donaldson, I. M., and Wodak, S. J. (2010). irefweb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database*, 2010.

[Vanunu et al., 2010] Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., and Sharan, R. (2010). Associating genes and protein complexes with disease via network propagation. *PLoS computational biology*, 6(1):e1000641.

[Wang et al., 2011] Wang, L., Li, L., Zhang, H., Luo, X., Dai, J., Zhou, S., Gu, J., Zhu, J., Atadja, P., Lu, C., et al. (2011). Structure of human smyd2 reveals the basis of p53 tumor suppressor methylation. *Journal of Biological Chemistry*, pages jbc–M111.

[Watson et al., 2013] Watson, I. R., Takahashi, K., Futreal, P. A., and Chin, L. (2013). Emerging patterns of somatic mutations in cancer. *Nature reviews Genetics*, 14(10):703.

[Widłak, 2013] Widłak, W. (2013). *Molecular Biology-Not Only for Bioinformaticians*, volume 8248. Springer.

[Yu et al., 2012] Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterprofiler: an r package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology*, 16(5):284–287.

[Zhang et al., 2011] Zhang, J., Baran, J., Cros, A., Guberman, J. M., Haider, S., Hsu, J., Liang, Y., Rivkin, E., Wang, J., Whitty, B., et al. (2011). International cancer genome consortium data portal—a one-stop shop for cancer genomics data. *Database*, 2011.

APPENDIX A

THE MAP OF MUTATIONS PRESENT IN PATIENTS

A.1 The heatmap of the mutations present in at least two patients.

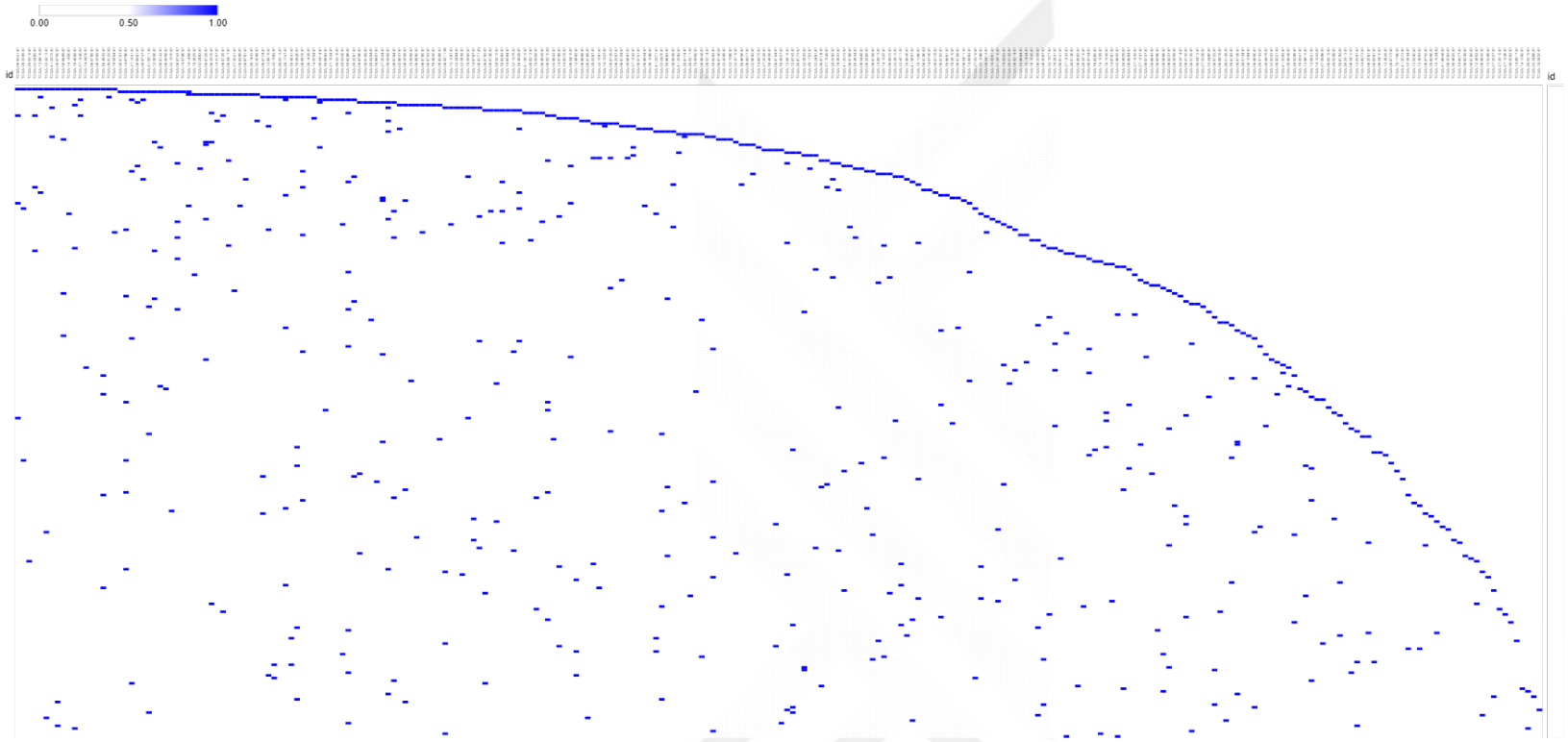

Figure A.1: The heatmap of the mutations present in at least two patients. Blue represents presence and white represents absence of the mutation (row) in the corresponding patient (column).

A.2 The heatmap of the mutations present in at least three patients.

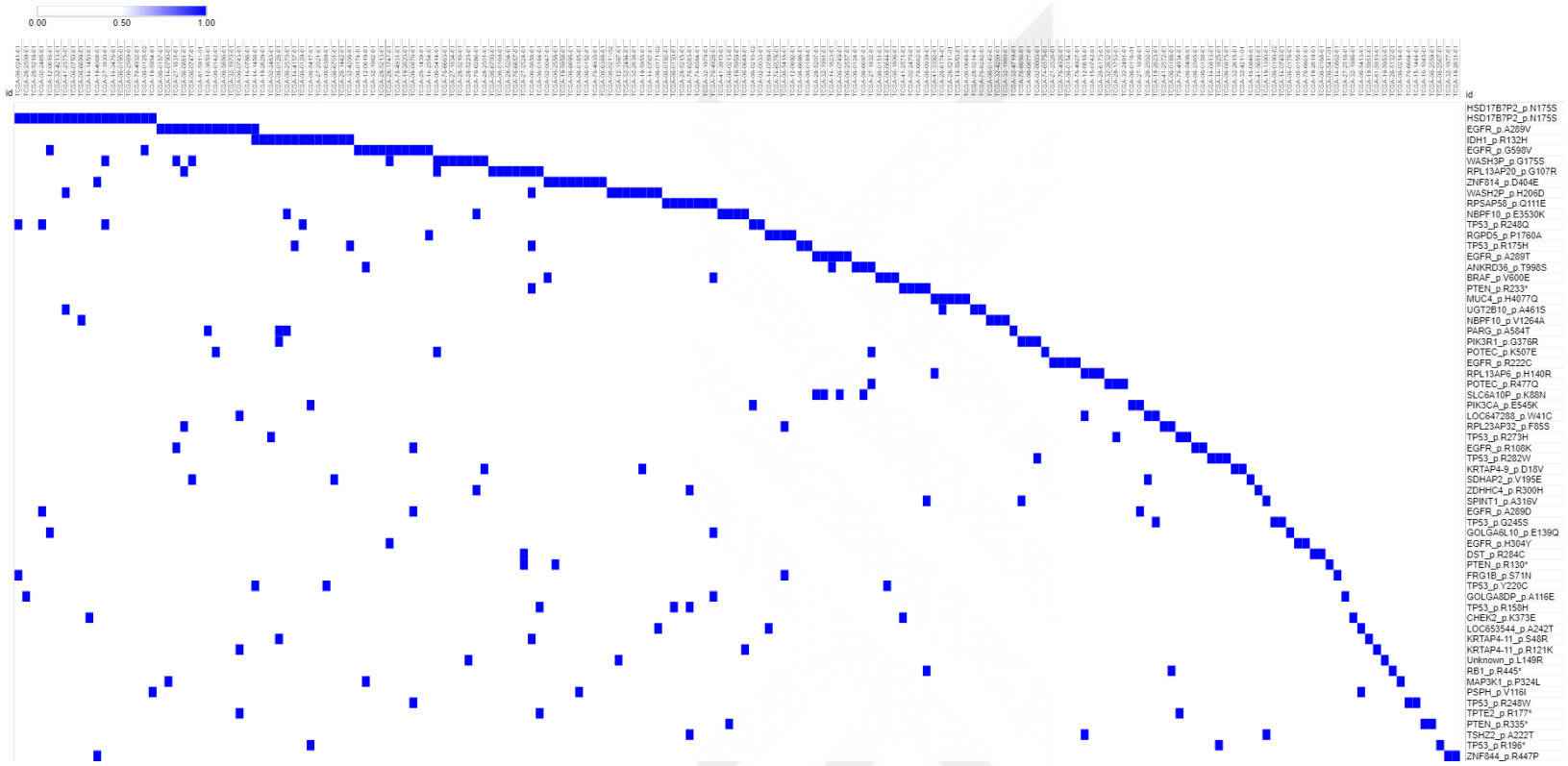

Figure A.2: The heatmap of the mutations present in at least three patients. Blue represents presence and white represents absence of the mutation (row) in the corresponding patient (column).

APPENDIX B

KEGG PATHWAY ENRICHMENTS IN PATIENT-SPECIFIC MUTATION SETS

B.1 The third subset of patient-pathway enrichment analysis.

Figure B.1: The third subset of patient-pathway enrichment analysis with KEGG.