RISK ASSESSMENT AND DRUG DESIGN USING R: HEART DISEASES

By

AL-MUSTANJID
151-35-1120
AND
CHANDAN MANDAL
151-35-978

A thesis submitted in partial fulfillment of the requirement for the degree of Bachelor of Science in Software Engineering

Department of Software Engineering
DAFFODIL INTERNATIONAL UNIVERSITY

Fall – 2018

APPROVAL

This thesis titled "**Risk Assessment and Drug Design Using R: Heart Diseases**", submitted by **Al-mustanjid**, **ID: 151-35-1120 and Chandan Mandal, ID: 151-35-978** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

24/12/18

Prof. Dr. Touhid Bhuiyan **Chairman**
Professor and Head
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Dr. Md. Asraf Ali **Internal Examiner 1**
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Md. Maruf Hassan **Internal Examiner 2**
Designation
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Prof Dr. Mohammad Abul Kashem **External Examiner**
Professor
Department of Computer Science and Engineering
Faculty of Electrical and Electronic Engineering
Dhaka University of Engineering & Technology, Gazipur

DECLARATION

We hereby declare that we have taken this thesis under the supervision of **Md. Habibur Rahman, Lecturer, Department of Software Engineering, Daffodil International University.** We also declare that neither this thesis nor any part of this has been submitted elsewhere for award of any degree.

Al-mustanjid

Al-mustanjid
ID: 151-35-1120
Batch : 16th
Department of Software Engineering
Faculty of Science & Information Technology
Daffodil International University

Chandan

Chandan Mandal
ID: 151-35-978
Batch : 16th
Department of Software Engineering
Faculty of Science & Information Technology
Daffodil International University

Certified by:

Rahman 24.12.18

Md. Habibur Rahman

Lecturer

Department of Software Engineering

Faculty of Science & Information Technology

Daffodil International University

ACKNOWLEDGEMENT

First we express our earnest thanks and gratefulness to almighty Allah for His heavenly blessing, which make us possible to complete this thesis successfully.

It is an fortunate opportunity for us as a student of the Department of Software Engineering, one of the exalted academic centers of the Science and Information Technology Faculty of the Daffodil International University, to express us deep feelings of gratitude to the department and to our honorable teachers and also to the department staff.

We are utmost indebted to our honorable supervisor, **Md. Habibur Rahman, Lecturer, Department of Software Engineering, FSIT, Daffodil International University, Dhaka**, for his excellent guidance, inspiration, encouragement and also for through review of our thesis paper. It was not possible for us to complete our thesis paper successfully without his help. We really thanked and gratefully remember those persons who always helped and encouraged us.

Last of all, we would like to thank to our parents who have given us tremendous inspiration and supports. Without their mental and financial supports we would not be able to complete our thesis.

TABLE OF CONTENTS

LIST OF TABLE

LIST OF FIGURE

ABSTRACT

Background: Bioinformatics handles living organism data and inspects the data using computer science facilities. Emerging modern bioinformatics tools associated with high technologies reveal a new area of drug design, it is now possible to drug design structurally. Several research works have revealed the ways of how a drug can be designed using bioinformatics techniques and tools.

Objective: More than an era people are being killed by Heart diseases globally. Heart disease is mainly known as Cardiovascular disease (CVDs). CVDs is the leading reason for universal premature deaths. Stroke and Myocardial infarction are standing at peak position among several CVDs in the world. Types of CVDs cover all existent frequently occurring heart diseases. CVDs are increasing because of common risk factors gradually. Common risk factor diseases have genetic association indirectly or directly. A disease is an abnormal condition in a single gene that affects body negatively. Biomolecule or protein is the best key for structure-based drug design. Data mining as well as data analysis is essential in bioinformatics to find the desired data.

Results: Using Knowledge discovery in database (KDD) the process of data mining, genes are filtered, preprocessed, transformed and mined to bring out common gene among 5 types of CVDs. Protein-protein interactions (PPI) are generated from common gene to visualize protein interactions and further evaluations.

Conclusions: This study claimed to design a common pathway drug for all types of CVDs. The Genes associated with CVDs types are collected from NCBI database using R. To achieve the goal UniHi is used as a tool.

Keywords: PPI, Heart Disease, UniHi, Drug Design, R, Data mining

CHAPTER 1

INTRODUCTION

1.1 Background

Bioinformatics is a buzzword in the world of science for the time being. Bioinformatics is the solicitation of Information technology that analyzes living data and addresses Biological problems. Bioinformatics uses to implement methods and tools to comprehending biological data. Integration of computers, software tools, databases, and methods are endeavored to solve biological problems. Genomics and Proteomic, as well as Drug design are the integrated working areas in Bioinformatics. R is the best tool for analyzing biological data and committing to build bioinformatics more effectively. Several tools are available for network evaluation and visualization such as UniHi, String etc. Risk assessment is evaluating the risk factors that affect and drug design is control or destroy the risk.

About 14.31% of total deaths are occurred due to Heart Disease in Bangladesh (Coronary Heart Disease in Bangladesh, n.d.). Heart disease is the brollies that invade heart for a bulk of conditions. Heart disease is also called Cardiovascular disease (CVDs) alternatively (What is Heart Disease?, n.d.). CVDs is the key appearance of Heart diseases.

Heart is likewise any other muscle in our body. Heart fulfill oxygen demands of our body through blood pump using its coronary arteries. Cardio refers to the heart and vascular refers to all the blood vessels in the body. Atherosclerosis is the main cause of CVDs. Atherosclerosis is a blockage or plaque in arteries that makes the arteries thick and stiff which stops blood supplying. Cardiovascular disease causes 17.9 million people each year ("Campaign essentials," 2016). This conveys 31% of all global deaths. Nichols et. al., (2016) found that premature (under 75 years) deaths are increasing for CVDs day by day. In addition more living people are quitting because of CVDs than all kinds of Cancer and Chronic Lower Respiratory disease (American Heart Association [AHA], (2017)). There are 5 types of CVDs. Categories of CVD are shown in the Figure 1.1.

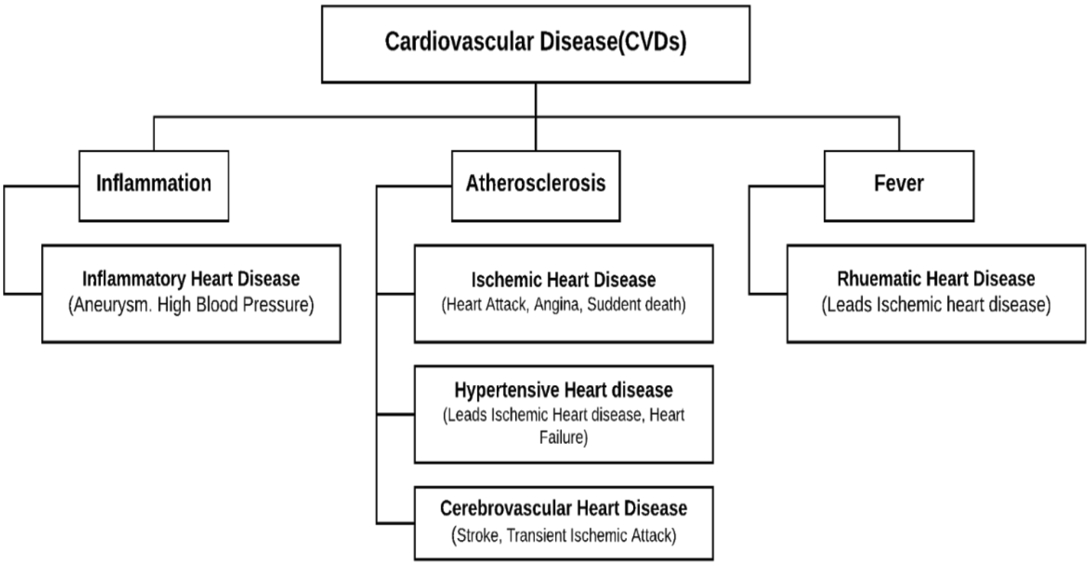

Cardiovascular Disease(CVDs)
Inflammation
Atherosclerosis
Fever
Inflammatory Heart Disease
(Aneurysm. High Blood Pressure)
Ischemic Heart Disease
(Heart Attack, Angina, Suddent death)
Rhuematic Heart Disease
(Leads Ischemic heart disease)
Hypertensive Heart disease
(Leads Ischemic Heart disease, Heart Failure)
Cerebrovascular Heart Disease
(Stroke, Transient Ischemic Attack)

Figure 1.1: Categories of CVDs

Many Heart diseases are affecting Bangladesh rapidly. The several types are covered by CVDs such as Ischemic Heart disease, Stroke, Heart attack etc. The Figure 1.2 below shows the impact rate of various heart diseases in Bangladesh.

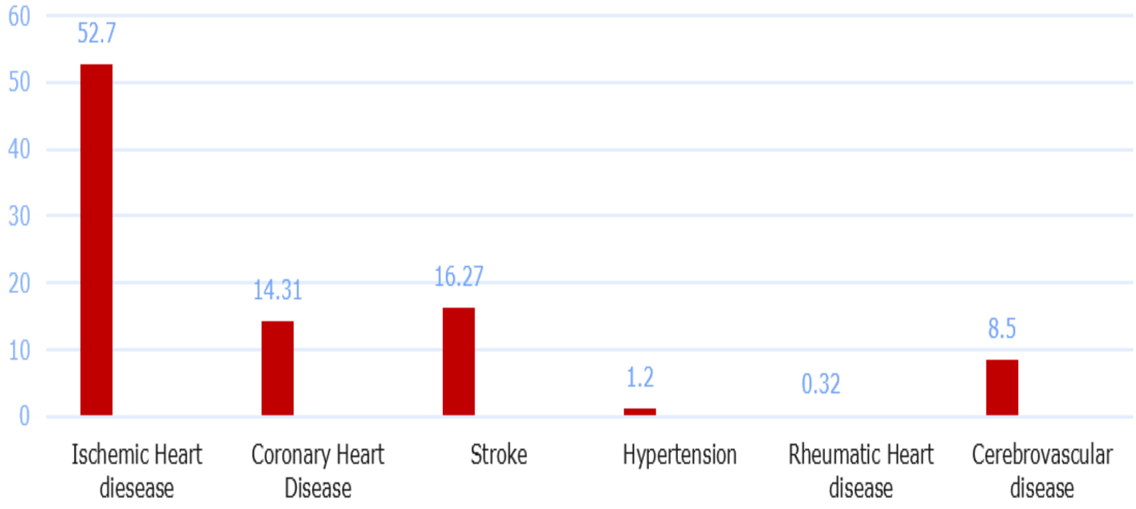


Figure 1.2: Impact rate of various heart diseases in Bangladesh (2005-2016)

Ischemia is a situation in which the blood flow (thus oxygen) is restricted or reduced in a part of the body. Ischemic heart diseases are happened for narrowed heart arteries. This can cause the heart muscle to ache. Ischemic is the worst killer of all CVDs. Ischemic heart disease is also known as Coronary Heart disease sometimes are also called Coronary artery disease. Chest pain or Angina, Shortness of breath, Heart attack are the symptoms of Ischemic heart disease. It ultimately leads to a heart attack. Angina, Coronary Heart/Artery Disease, Heart Attack, Sudden death are the disorders of Ischemic Heart disease. Myocardial Infarction same as the Heart attack is the leading cause of deaths in Bangladesh and it is about 34.4% (Morbidity Profile, n.d.). In Australia, Ischemic heart disease was responsible for around 20000 deaths (Nichols et al., 2016). An American will be faced heart attack in every 40 seconds (AHA, 2017). In Europe, Coronary artery disease causes almost 1.8 million deaths annually (Dégano et al., 2015).

Cerebrovascular diseases are developed in various ways including deep vein thrombosis (DVT) and atherosclerosis. Cerebral tissue is affected by Cerebrovascular diseases. Paralysis of one side (hemiplegia), a severe and sudden headache, weakness on one side (hemiparesis) are the common symptoms of the Cerebrovascular disease. Stroke is the main disorder of Cerebrovascular disease. Stroke is the reason for nearly 11.8% deaths of all heart diseases and it is leading second position worldwide (AHA, 2017). Nichols et al., 2014 indicated that 10% men and 15% women deaths in Europe are died by the stroke in 2014. According to World Life Expectancy, Indonesia has the highest stroke rate in Asia as well as worldwide where Bangladesh secured 34'Th position around the globe (Stroke death rate by country, n.d.).

Hypertension is the unnaturalness in blood pressure. It brings a situation of great psychological stress. Hypertensive heart disease is the leading reason for death associated with high blood pressure. Tightness or pressure in the chest, fatigue, pain in the neck, back, arms, or shoulders, loss of appetite, leg or ankle swelling are the symptoms of Hypertensive heart disease. It is included with the disorders such as heart failure, ischemic heart disease, and left ventricular hypertrophy. Hypertensive heart disease has reached 1.07 million globally ("Hypertensive heart disease,"2018).

Inflammatory heart disease is also known as myocarditis. Myocarditis is a disease of the heart muscle known as the myocardium — the muscular layer of the heart wall. This muscle is bound to contracting and relaxing to pump blood in and out of the heart and to the rest of the body. The muscles become inflamed and damage the effectiveness of the ability to pump blood. Myocarditis is caused by infection of heart muscle by viruses, bacteria, fungi etc. It leads to Stroke, Heart failure, and Heart attack. It is found that youth adults (0.05% - 0.2%) are suffering from Myocarditis. In the USA, myocarditis is the reason for 0.45% Hearts transplantation (Myocarditis Causes, n.d.).

Rheumatic heart disease is caused by rheumatic fever. Streptococcal bacteria is the root cause of RHD which damages heart valves. During fever, the heart is slowly damaged and leads to CVDs. Lower and middle-income countries people are uncertain about RHD. Oceania, South Asia, and central sub-Saharan Africa were most traced areas of RHD. In 2015, it was found that about 33.4 million were affected by RHD globally (Watkins, et al., 2017).

1.2 Motivation

Unusualness in a gene is known as disease. A risk factor is any attribute of eminent that rise the possibility of developing a disease ("Risk factors," 2017). Risk factors are interrelated. The power of this interrelation is measured statistically. On the other hand, common risk factors make a way to find genetic interrelation including common pathway ("Risk factor," 2018). Smoking, High Blood pressure, High Cholesterol, Diabetes, obesity are common risk factors of all CVDs (Koene et al., 2016).

The outbreak of Smokers was found utmost in lower and middle-income countries. It was covering 0.8% of the whole world (Tobacco, n.d.). The main stem of atherosclerosis is smoking and increases the difficulty of chemical identity that revokes Heart diseases (Messner & Bernhard, 2014). According to WHO, Tobacco is the most cause of premature CVDs. Bangladesh has been faced approximately 0.2% deaths due to Tobacco (Hasan, 2018). About 39.8% male and 0.7% female adults (>15year) died because of smoking in Bangladesh (Bangladesh, n.d.).

Furthermore, High blood pressure (HBP) caused CVDs by narrowing heart arteries (How High Blood Pressure Can Lead to a Heart Attack, n.d.). Campaign essentials (2016) found that HBP remained silent performer of CVDs and nearly 9.4 million deaths of all heart diseases were being responsible for HBP. On an average 20% adult and 40-65% aged people were bearing Hypertension or high blood pressure.

Cholesterol makes the new cell and generates hormone in our body that is needed to run our lives. While the number of cholesterol increases in blood, plaques or walls are formed that causes atherosclerosis (Heart Disease and Lowering Cholesterol, n.d.). Campaign essentials also found that High cholesterol was the second outface of CVDs in Australia. Adults are having cholesterol closely 39.7% in USA. Total cholesterol (combination of high-density lipoprotein, low-density lipoprotein, and very low-density lipoprotein cholesterol fraction) observed a little high in Bangladesh (Fatema et al., 2016).

Obesity arrives due to extra fat consumption as well as overweight of an individual. It spoils human health genetically and environmentally and raises the risk of CVDs (Nigro et al., 2014). Obesity rate is highest in Oceania countries. In New Zealand and Australia respectively 28.3% and 26.8% obesities are shown (Cheong, 2014). Teenagers are getting obese in Bangladesh due to having junk foods vastly (Salahuddin, 2018).

Diabetes happens while the level of sugar gets high in the blood. It damages vessels of the heart and heart fails its normal activity. Diabetes is spreading among adults in the world at a higher rate. Averagely 425 million adults bearing diabetes worldwide. The outbreak of diabetes in Bangladesh is also noticeable and adults affected rate is nearly 6.9% (IDF diabetes atlas, n.d.; IDF SEA members, n.d.).

1.3 Scope

This research analyses Ischemic, Cerebrovascular, Hypertensive, Inflammatory and Rheumatic whose are all types of CVDs. Heart disease also called CVDs is one of the leading non-communicable disease in the world. Diseases generated from common risk factors have a great chance of genetic correlation directly or indirectly. KDD is the special approach of data mining to find a specific knowledge in large-scale data. KDD along with Bioinformatics contributing to human life. NCBI is the repositories of biological data. Proteomics makes PPI more effective. Protein binds with other proteins by regulation which turns into the desired chemical reactions of our body. Regulation is identified in protein-protein interactions which leads to finding a common pathway among the proteins and drug design for all CVDs types.

1.4 Objectives

The major objectives of this thesis are given below:

- To break out genetic association among all types of CVDs.
- To find common genes among all types through Data mining using R.
- To design PPI using UniHi tool.
- Identify regulatory common pathway from PPI.
- To design drug from the common pathway using UniHi.

1.5 Thesis Organization

This thesis document is enclosed with followings. The current chapter provides the introduction of this thesis. Chapter 2 describes related works. Research methodology is discussed in chapter 3. Chapter 4 shows experiments and corresponding results respectively. Finally, the conclusions are determined in section 5.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Bioinformatics creates a new vision in life sciences combining computer science, biology, statistics, and mathematics. One of the flourishing areas in Bioinformatics is drug design. Proteomics contributes to protein analysis as well as drug design integrate with Genomics. The key molecular aim of drug design is proteins (Feng et al., 2017). Protein-protein interactions (PPI) provide a proper way to drug design. This chapter is modeled to illustrate the background history of this study.

2.2 Related work

Oti et al., (2006) used PPI to predict heterogeneous disease genes. Dataset was consist of five protein-protein interaction based on four different species—Human, Drosophila melanogaster, Caenorhabditis elegans, and Saccharomyces cerevisiae. The collection was 10,894 genes for 383 genetically heterogeneous hereditary diseases from 432 loci of candidate diseases. Protein-protein interaction set, Candidate gene prediction, Benchmark test, and Randomisation test methods were conducted for overall evaluation. 300 candidate genes were enumerated from 72, 940 protein-protein interactions.

The common pathway was discovered based on PPI among schizophrenia, bipolar disorder, coronary heart disease in Habib, (2016). Using R common genes were found within genetically interrelated four diseases. Dataset of Gene was collected from the

NCBI database. Genes were verified using ExPaSy database. PPI was evaluated using UniHi tool.

Feng et. al., (2017) extracted new drug targets from non-drug targets based on PPI. Drug target characteristics were explored to continue the research. Topological features with three possible views were studied to understand the mechanism of molecular-level drug targets. Topology features were also lead to understanding working procedure of drug targets in PPI. 5 topological indices found among drug target proteins and other proteins in the PPI.

In the above discussion, it clearly implies that a genetically common pathway carries among heart diseases sharply that forms from general risk factors. Regulatory interactions in PPI network refines the foundation to design common pathway drug. This research mainly focused on evaluating PPI from mined common 14 genes. From PPI, design a common pathway drug for CVDs.

2.3 Gene

Gene is the piece of DNA that contains genetic information. Basic physical and functional unit of functions of heredity is Gene. Genes are made up of DNA. Genes (not all) act as the instruction to make molecules called protein. Each person holds two copies of each gene, one inherited from each parent. Most genes are similar in all human, less than 1% genes are different among people.

Each gene has a unique name to keep track. Genes determine everything about us, from the outward physical traits we can see to the behind the scenes structures inside our cells that allow them to carry out all of our body functions (“Gene,”2018).

2.4 Protein

Protein is the main element of all living organism. Protein is a molecule composed of polymers of amino acids joined together by peptide bonds. Proteins are large biomolecules or macromolecules. Proteins are made up of long chains of amino acids which creates polypeptide chain polymer. When polypeptide chain polymer folds a protein is created. Figure 2.1 is the view of protein structure.

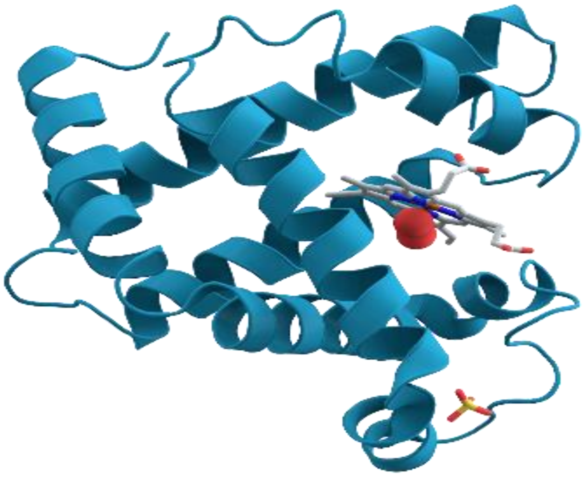


Figure 2.1: Structure of protein

The followings are the effects of protein:

- Protein is the builder of our body such as bones, muscles, blood cell to teeth, finger etc.
- Cells of our body are frequently hampered. Protein creates a new cell on that hampered place.
- Protein generates heat for our body
- Protein prevents our body by creating antibody of the diseases.
- For growth especially of brain, protein is needed.

So, it is clear that if our body does not get proper proteins, there will occur many malfunctions in the body.

2.5 Proteomics

Proteomics is the massive practice of proteomes. A proteome is a set of proteins generated in an organism, system, or biological context. Homo sapiens is an example of the proteome of species. The proteome is blended words of protein and genome and the naming was done by Marc Wilkins in 1994 at Macquarie University (“Proteomics,” 2018). Also, Proteomics is the resolution of the entire protein complement of a cell, tissue, or organism under a tangible set of conditions (Yu et al., 2010).

Usage of proteomics are given below (“What is proteomics?,” 2018):

- Look into protein expression time.
- Inquire into rates of protein production, degradation, and steady-state affluence
- Investigate deportment of protein among subcellular compartments.
- Identify how a protein interacts with another.

2.6 Genomics

Genomics is the study of whole genomes of organisms and incorporates elements from genetics. A genome is an organism's complete set of DNA, including all of its genes. DNA recombinant combination, DNA sequencing methods, and bioinformatics are used to determine sequence, assemble, and analyze the structure and function of genomes in Genomics (“What is genomics?,” 2016). Genomics purpose is at combined characterization and quantification of genes which monitor protein production (“Genomics,” 2018).

2.7 Protein-protein Interactions

PPI is the definite bodily connection between two or more protein molecules which is conducted by electrostatic forces including the hydrophobic effect (Pattin et al., 2009). Proteins hardly tend to regulate alone or single. Molecular processes are performed in a cell by various protein components arranged by PPI. Aberrant PPI causes disease such as Alzheimer's diseases may lead to cancer (“Protein–protein interaction [PPI],”2018).

2.7.1 Advantages of Protein-protein Interactions

PPI helps to understand the consequences of interaction in a cell. Peptides are developed by PPI. In a single network activated or repressed proteins are determined through PPI. PPI leads the way to how locomotion of protein may change. PPI is generated from tying proteins. The way of tying between proteins can change through determination (PPI, 2018).

PPI provides deep knowledge of common human disease by explaining the functional relationship between genes and combining genomic data with proteomic data. The most use of PPI is in Biomedical as well as pharmaceutical R&D. Putative protein targets are discovered from PPI that leads to therapeutic interest. Example Maraviroc, an inhibitor of the CCR5-gp120 interaction, used as an anti-HIV drug (PPI, 2018).

2.8 Drug Design

Drug design is the ingenious process of search new cure basis on the biological target knowledge. The initial goals of drug design to measure how powerfully a given molecule or protein will tie with a target (“Drug design,”2018). The drug is usually a vital tiny molecule that activates or intercept the function of a protein and it turns the outcome in a therapeutic benefit to the patients. Basically, drug design brings into contact the design of small molecules that are fulfilling in shape and charge to the bimolecular target with which they react and communicate as well as will crib to it. Modeling drug design by computer technology is known as computer-aided drug design (Drug Design, n.d.).

Drug design helps pharmacists to understand the molecular level action mode and also favors making the decision on appropriate dosing and state of medicines. Drug design improves the quality of human life. Discovering a drug design brings great influence on life (Drug design and discovery special interest group, n.d.).

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

We are living in data age. Terabytes or petabytes data is pouring into the computer networks every day. A human body is having approximately 19000-2000 protein-coding genes. So, when it comes about millions of people's biological data, Bioinformatics has a vast amount of different categorized data such DNA sequencing, protein-protein interaction etc. Continuously the growth of genomic and proteomic data is increasing. Obtaining particular results from the sea of data need a substantial procedure. Bioinformatics and data mining are developing as cross-functional science. Data mining application and techniques are active fields in Bioinformatics to solve biological problems. It has been found that cross-pollination between data mining and bioinformatics are very effective (Raza, 2012).

Data mining and Knowledge Discovery in Database (KDD) are used interchangeably. KDD the data mining method is used to extract useful knowledge from large-scale data. KDD is an iterative process.

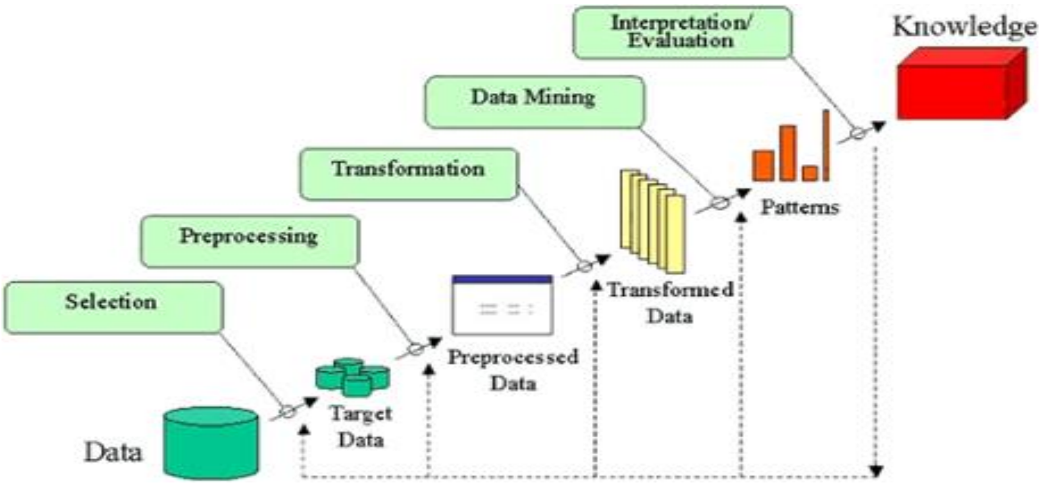


Figure 3.1: Steps of KDD process

The following steps in Figure 3.1 are described below:

Data- Learn about the application domain, prior relevant works and what the goal is.

- **Selection-** Select the target dataset or subset of data samples for performing the discovery.
- **Preprocessing-** Cleanse data and remove noise from data. Set strategies for handling missing data fields. Alteration of data as per requirements.
- **Transformation-** Reduction and projection of data are performed at this step. Simplify dataset by deleting undesired data. Set features to show data depending on the goal or task.
- **Data Mining-** Identify KDD goal with data mining methods or algorithms to extract hidden pattern and represent the pattern in a form or a set of such representations as classification rules or trees, regression, clustering etc.
- **Interpretation or Evaluation-** Understand crucial knowledge from the mined patterns.

Adopt the knowledge to perform further actions and obtaining the conclusion.

3.2 Proposed Method

Protein-Protein Interaction is a sequential procedure. Various steps are performed to obtain the PPI. Fig delimitates graphical representation stepwise providing a clear realization of this thesis evaluation method. The steps are determined to extract common genes. R language has used to complete the steps and achieve the goal. UniHi is used to achieve the PPI network from common genes. Each step of the flowchart is accomplished in the following subsections 3.2.1 to 3.2.7.

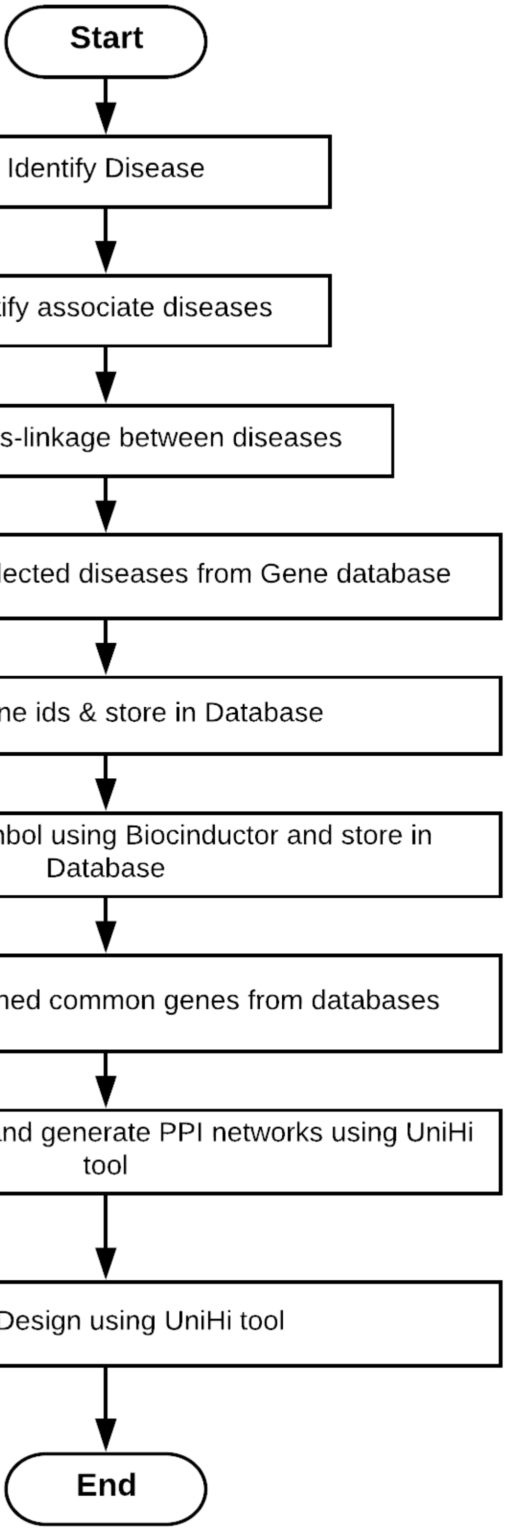

Start
Identify Disease
Identify associate diseases
Make cross-linkage between diseases
Search gene for selected diseases from Gene database
Fetch gene ids & store in Database
Extract Gene Symbol using Biocinductor and store in Database
Retrieve all the mined common genes from databases
Adopt mined genes and generate PPI networks using UniHi tool
Drug Design using UniHi tool
End

Figure 3.2: Flowchart of proposed methodology

3.2.1 Gene Search and collection

National Center for Biotechnology Information (NCBI) is called the mother of Bioinformatics. It is an important source for bioinformatics tools and services. All biological data are stored here divergently. Different categorized databases are attainable such as Gene, Nucleotide etc. The databases are accessible and downloadable through the internet or integrated tools. rEnterez is the package in R that gets access in various NCBI database including Gene. For this thesis, genes are collected associated with Ischemic, Cerebrovascular, Hypertensive, Inflammatory and Rheumatic heart diseases. The pseudo code of the collection process is shown in Figure 3.3 respectively.

```
library(rentrez)
object_name <- entrez_search(db="database name", term="disease name")
object_name
object_name$ids
```

Figure 3.3: Gene collection process using R

3.2.2 Pre-Processing and Filtering

In prior step genes of every type are collected for all organisms. Only Human genes are needed. This step collects genes only for the human organism or Homo sapiens. Here clearing noise from collected data is called preprocessing. So, collected data is filtered and only human genes are brought out. Just modify the pseudo code of the Figure 3.3.

```
library(rentrez)
object_name <- entrez_search(db="database name", term="(disease name) AND Homo Sapiens")
object _name
object_name$ids
```

Figure 3.4: Gene pre-processing and filtering

3.2.3 Cross-linkage Gene collection

Co-related genes are collected in this step. Total 26 combinations are evaluated among all types of CVDs. Genes are collected among 2 types, 3 types, 4 types, and 5 types combinations. Genes are calculated. Pseudo code is given in the Figure 3.5.

```
object_name <- entrez_search(db="database name", term="(disease name AND disease name)"
object_name
object_name$ids
```

Figure 3.5: Cross linkage gene collection procedure using R

3.2.4 Gene Sorting

Gene information is consist of many elements such as aliases, gene symbol, gene id etc. To achieve goal gene symbol is needed. Gene symbol is the protein and aberrant PPI causes disease. Bioconductor an open source software platform is performed for many bioinformatics approaches integrating R. Bioconductor is used to sort gene symbol from gene information and stored in the database. Gene symbol for each type is sorted. The Figure 3.6 shows the pseudo code of gene sorting.

```
source("https://bioconductor.org/biocLite.R")
biocLite("org.Hs.eg.db")
library(org.Hs.eg.db)
biocLite("annotate")
library(annotate)
object_name<- c(gene_ids)
lookUp(object_name, 'org.Hs.eg', 'SYMBOL')
```

Figure 3.6: Gene sorting procedure using R

3.2.5 Gene Mining

Data mining techniques are used to identify applicable data. This step is very crucial because any fault can cast off an important gene that turns error output. Further, a large amount of data can commit the result complicated. Read gene symbol of all type and cross-linkages gene symbol in R and common genes are mined among them. The mined common genes are stored in a database and verified using Expasy database to check whether gene symbols are correct or not. The mined genes are compared with the top 100 and 50 interrelated genes and store in a database. The Figure 3.7 depicts pseudo code of Gene mining.

```
#Create vectors for each types including all cross-linkages
object_name <- c(gene_symbols)
#Create an object for store common gene
object_name <- Reduce(intersect, list(object_names)
#Take 100 gene symbol from each type and Create a data frame
Object_name <- c(top 100 gene symbol)
Object_name <- data.frame(Object_name )
Object_name
#Read mined common gene symbols as a vector
Object_name <- c(common gene symbol)
#Create a vector for store common gene from top 100 gene symbol data frame
Object_name <- Reduce(intersect, list(object_name$column-name)
#Then create a vector take top 50 gene from data frame
Object_name <- head(object_name, 50)
#Create a vector for store common gene from top 50 gene symbols
Object_name <- Reduce(intersect, list(object_name$column-name)
#Create another data frame for store gene symol for common from all, common from top 100 and 50
Object_name <- data.frame(object_name)
```

Figure 3.7: Gene mining procedure using R

3.2.6 Create Protein-protein interactions network

UniHi or Unified Human Interactive is web PPI visualization tool in Bioinformatics under Systems Biology and Bioinformatics Laboratory (SysBioLab). It’s a great tool for generating PPI. Protein interactions including common pathway among co-related genes are viewed in the PPI. From 14 co-related genes, PPI network and common pathways are yielded using UniHi tool.

3.2.7 Designing Drug

The essential step of this research. A drug is an element (mainly nutritional support) that when breathing, inject, smoked, consumed, take up via a tuck on the skin, or melt under the tongue resulted in a physiological change in the human body (“Drug,” 2018). In addition, a drug is also an element that uses to treat, cure and prevent illness, redemption from a pain, or tone down some appointed process in the body for specific cure. The therapeutic response is known as root to the invention of the drug (Habib, 2017). When normal PPI lost its normal form, the malfunctions occur in the body that causes disease. Feng et al., (2017) argued that investigating molecule actually protein is the inventive process of finding new drugs. The drug should design in a particular way that affects only malfunctioned protein or restrict the protein to regulate with another protein and keep the flow of the normal chemical process in the body. In order to develop a particular drug for the disease, targeted genes or proteins are needed which show a disease occurring way. Speedy and revolutionary improvement in proteomics, genomics, molecular biology, and robotics are leading great impacts on drug discovery (Blundell et al., 2002). Feng also found that PPI discover drug target from non-drug proteins. Drug discovery and design are powered by protein structure and structure also used in target identification (Blundell, et al., 2006). So, creating, visualizing and analyzing PPI is pre-requisite to drug design. Using UniHi, common pathway drug targets are identified from common 14 genes.

3.3 Summary

At first, applying the Knowledge Discovery in Database (KDD) steps of the data mining process the heart disease is selected from gene dataset of NCBI. Then the R language is used to collect cross-linkage genes. After performing the KDD steps 14 mined common genes found from 17000 genes using R. Genes were verified by Expasy database. The pattern PPI is achieved from common 14 genes using Unihi tool.

CHAPTER 4

RESULTS AND DISCUSSION

The current study was an attempt to design structure-based common pathway drug among all types of CVDs. As followed the steps in the previous chapter we obtained our goal. The genes data responsible for CVDs are loaded from the authentic database. Based on this representative sample, our entire study was operated and carried out the drug design among diseases. KDD processes are followed to evaluate common genes and UniHi is used to create PPI and further drug design. For better understanding, this chapter is divided into following subsections.

.

4.1 Architectural View of Database

Bioinformatics tools and databases make it possible to extract responsible genes for diseases and their correlation. The collected genes are stored in databases that brings out common genes. All databases are created via R. The architectural view of database in Figure 4.1 shows the how data collected, stored, filtered and retrieved to continue this study.

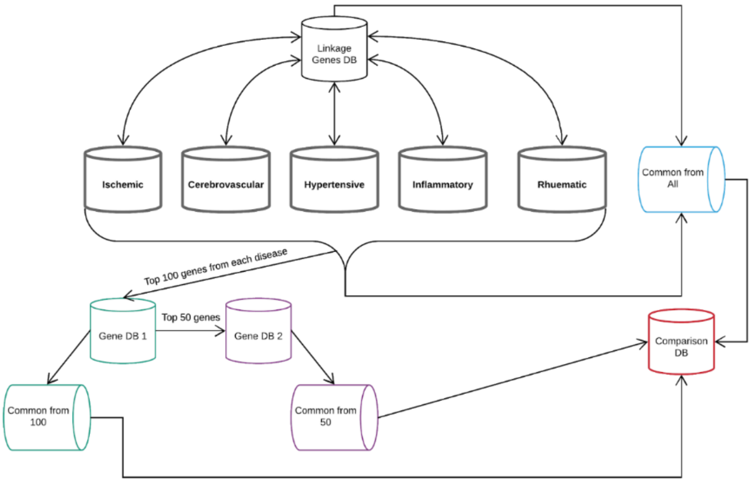


Figure 4.1: Architectural view of Database

4.2 Gene Optimization

Bioinformatics has a vast amount of biological data. Optimization along with bioinformatics makes the best use of data. KDD process using R is used to collect gene dataset from the NCBI database. The following subsections 4.2.1 to 4.2.3 shows how optimization was done. The Figure 4.2 also shows the graph view of gene optimization.

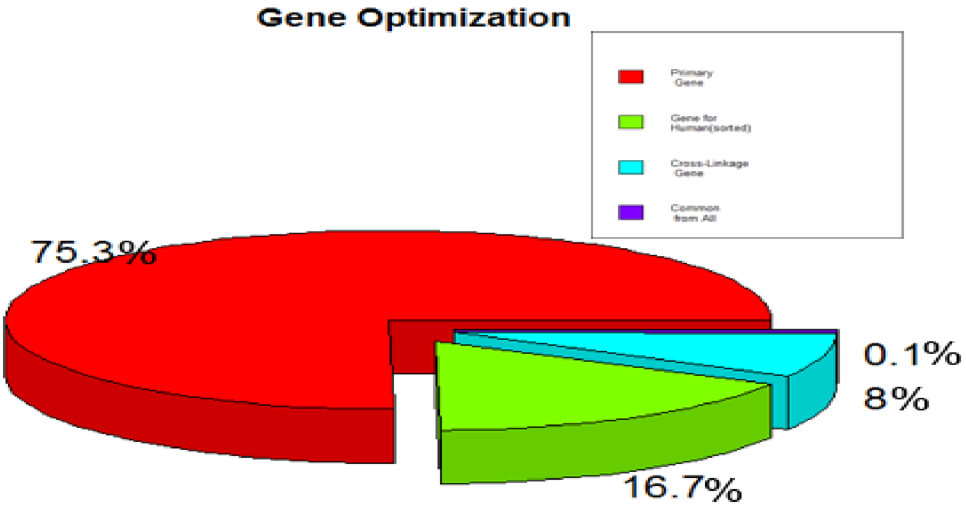


Figure 4.2: Scenario of gene optimization by pie chart

4.2.1 Gene collection, filter, preprocess & transform

The collection of genes are counted as 1931 for Ischemic, 227 for Cerebrovascular, 681 for Hypertensive, 13743 for Inflammatory and 220 for Rheumatic and the total amount is 16802. The collected genes are responsible for all organism and the amount is very large that makes the barrier to get exact data. By filtering, only human genes are collected and the amount turned less to total 3731. The Figure 4.3 shows the comparison between non-filter and filtered (Human) genes of all types.

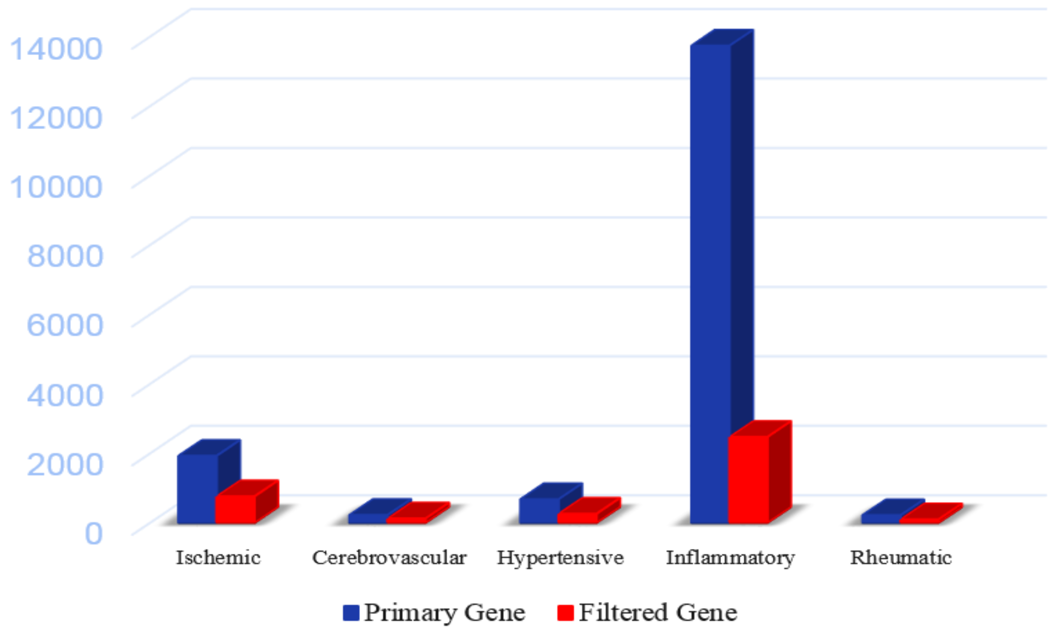
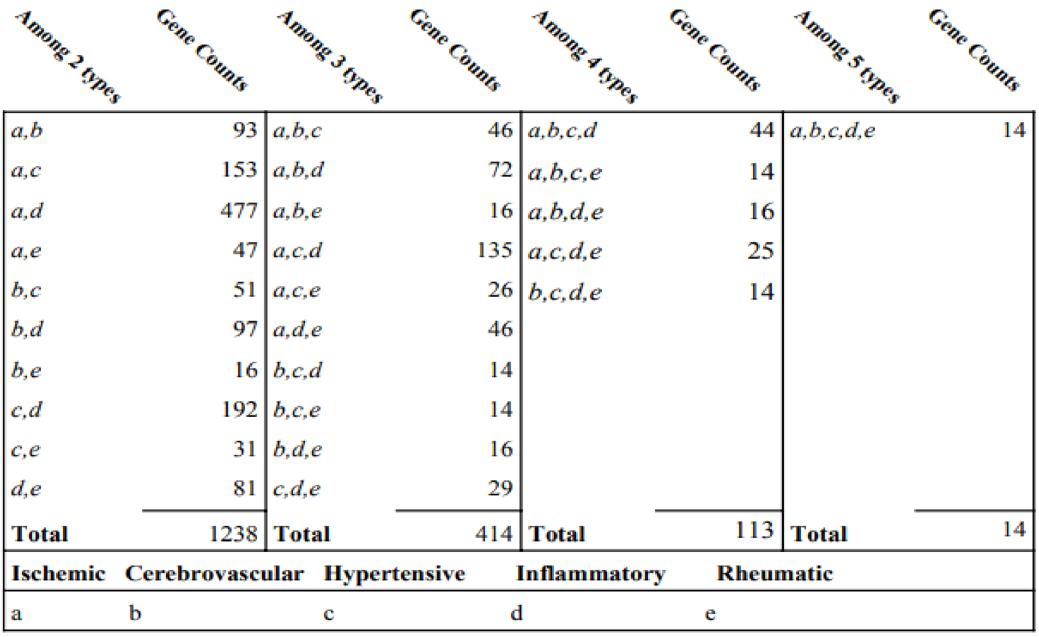


Figure 4.3: Collected genes for investigated diseases

By preprocess, gene information is stored in separate databases for each type. Gene information is consist of many elements such as gene id, gene symbol, nucleotide sequence, aliases etc. For our purpose gene symbol is needed as it is protein and protein binds with another protein to regulate human body functions. So, from filtered gene information gene symbol is transformed and stored in the database for further study.

4.2.2 Cross-Linkage gene collection

After collecting the genes, the cross-linkage was applied to find interrelated genes among Ischemic, Cerebrovascular, Hypertensive, Inflammatory and Rheumatic heart diseases. The cross-connection among disease helps to find the genetic association. Cross-linkage was investigated among 2, 3, 4 and 5 types. The Table 4.1 shows the calculation of cross-linkage genes among 5 types.



Among 2 types	Gene Counts	Among 3 types	Gene Counts	Among 4 types	Gene Counts	Among 5 types	Gene Counts
a,b	93	*a,b,c*	46	*a,b,c,d*	44	*a,b,c,d,e*	14
a,c	153	*a,b,d*	72	*a,b,c,e*	14		
a,d	477	*a,b,e*	16	*a,b,d,e*	16		
a,e	47	*a,c,d*	135	*a,c,d,e*	25		
b,c	51	*a,c,e*	26	*b,c,d,e*	14		
b,d	97	*a,d,e*	46				
b,e	16	*b,c,d*	14				
c,d	192	*b,c,e*	14				
c,e	31	*b,d,e*	16				
d,e	81	*c,d,e*	29				
Total	1238	**Total**	414	**Total**	113	**Total**	14

Ischemic	Cerebrovascular	Hypertensive	Inflammatory	Rheumatic
a	b	c	d	e

Table 4.1: Collected cross-linkage genes for investigated diseases

4.2.3 Gene Mining

The major step of gene optimization as well as this study. Corresponding genes for all types are stored in databases. Cross-linkage genes are stored in the database too. Then interrelated cross-linkage genes are compared with all genes of CVDs types. 14 common genes were found. Genes are TP53, TNF, IL6, MTHFR, TGFB1, ACE, MMP9, CRP, TLR4, NPPB, HMOX1, AGTR1, MMP1, and F3. These genes were verified using Expasy database. Then common genes from the top 100 and top 50 are searched and stored comparing with 14 common genes. All common including from top 100 and top 50 genes are stored in the database.

4.3 PPI Network with Regulatory Interaction

The 14 common genes are uploaded to UniHi and PPI is created. The network in PPI presents interconnectivity way among the genes. PPI also represent the directly or indirectly interrelated genes by a common pathway. The Figure 4.4 display the PPI network among 14 genes.

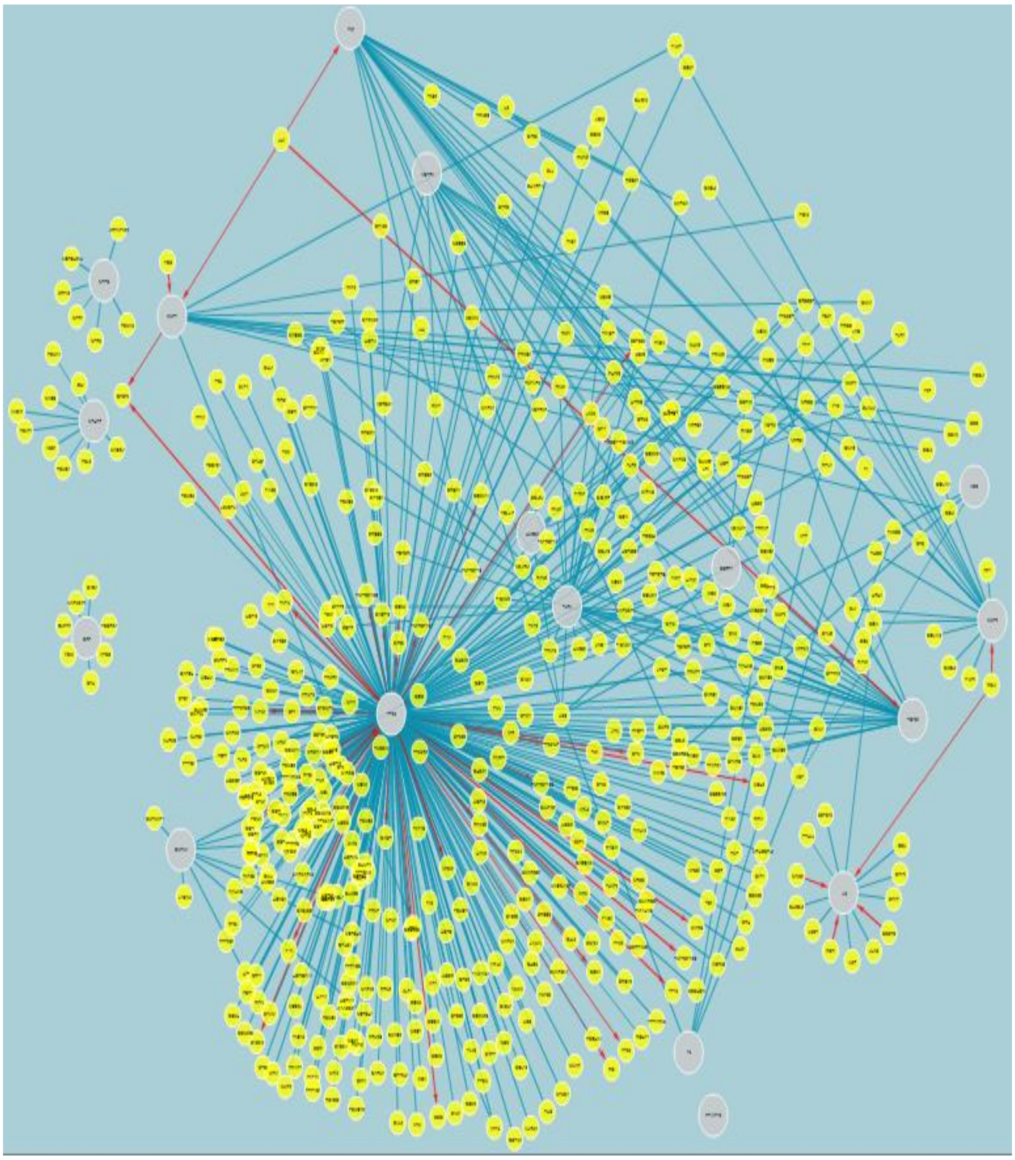


Figure 4.4: PPI network with 14 common genes

From the Figure 4.4, it is perceived that 6 genes have interrelation. The Figure 4.5 shows the regulatory interaction. From regulatory networks in the Figure 4.5, it is perceived that 4 genes TP53, MMP1, TGFB1, and TNF are directly connected with greater regulatory interactions and other 2 genes IL6, MMP9 are also connected directly with lower interactions.

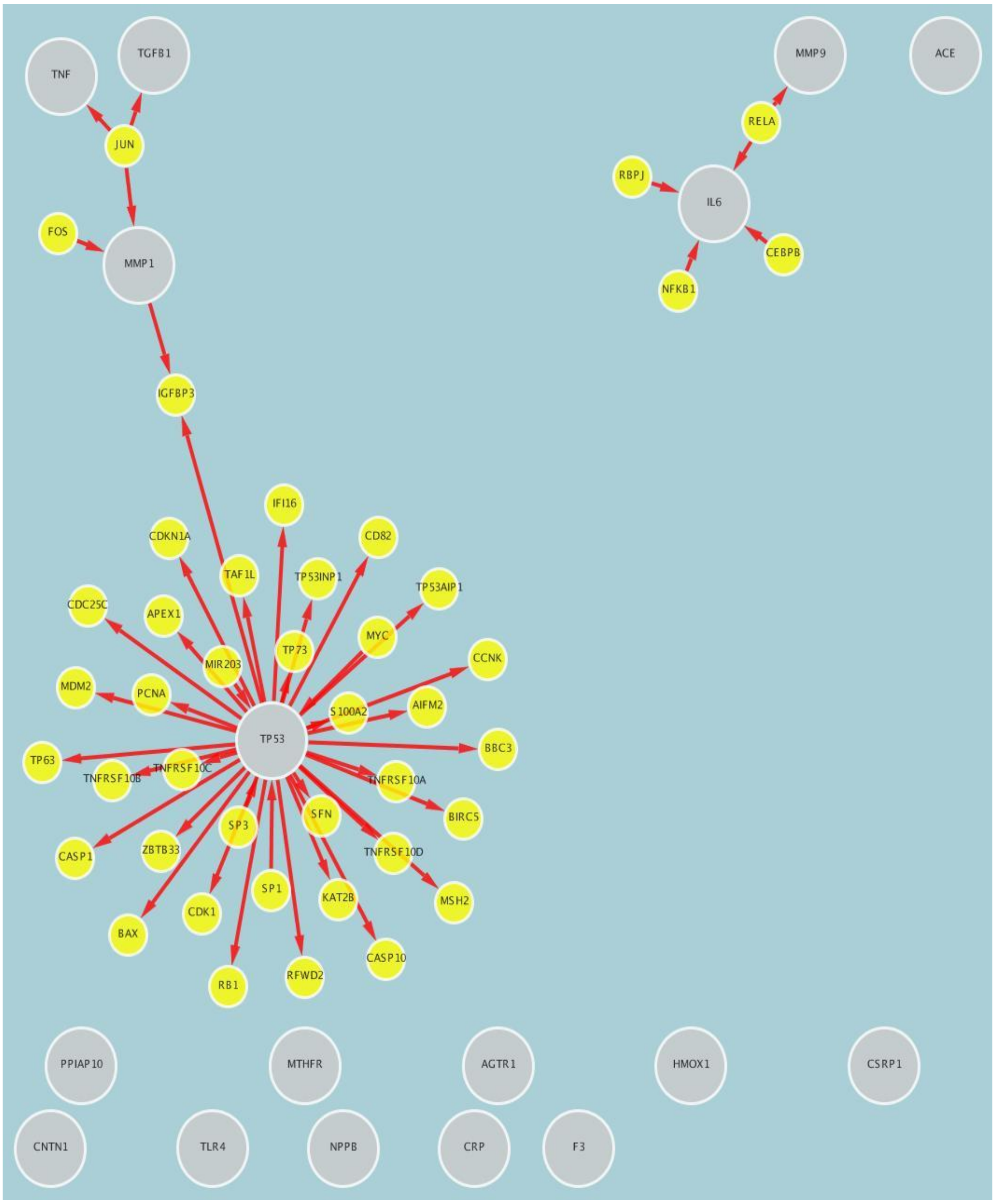


Figure 4.5: Regulatory Interactions with 14 common genes

So, another two PPI networks are created using 4 interconnected genes along with other 2 interconnected genes. The Figure 4.6 and the Figure 4.7 display the PPI network among 4 and 2 interconnected genes respectively. Regulatory interaction network exhibit the directly interacted proteins with the investigated diseases (Habib, 2017). The Figure 4.8 and The Figure 4.9 represent regulatory interactions in PPI among 4 interrelated genes and 2 interrelated genes.

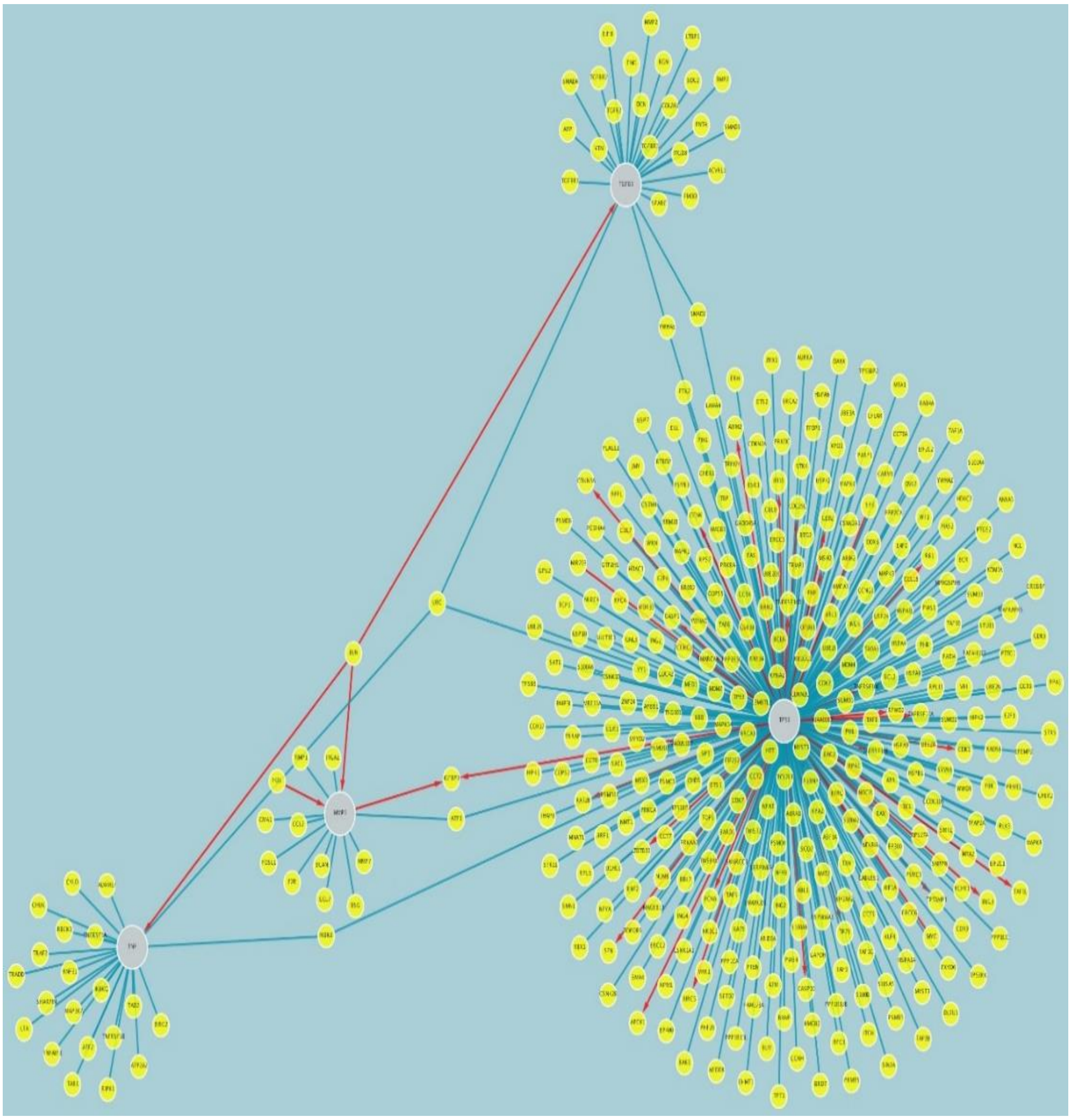


Figure 4.6: PPI Network with 4 interrelated genes

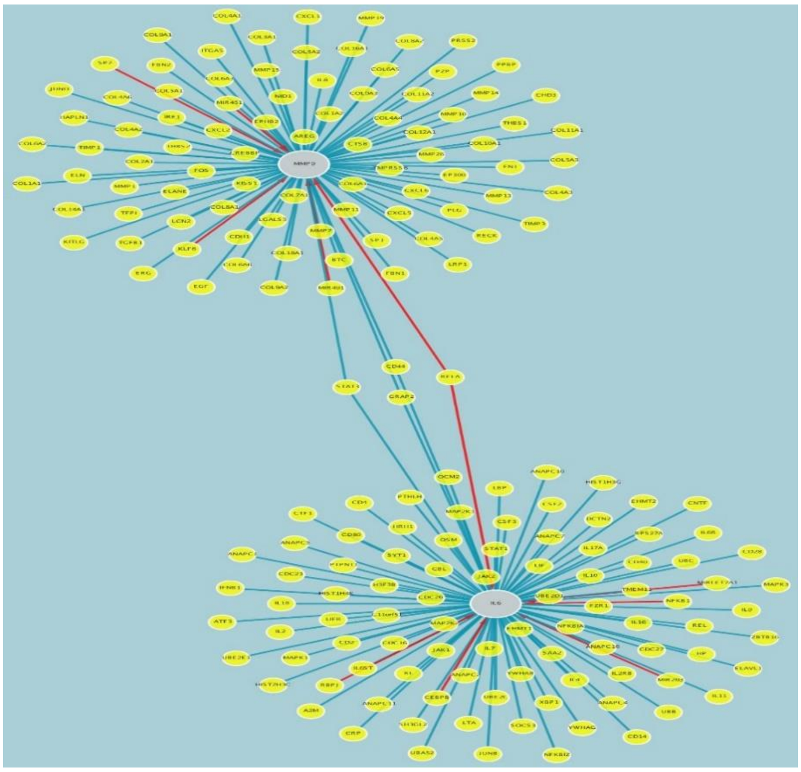


Figure 4.7: PPI Network with 2 interrelated genes

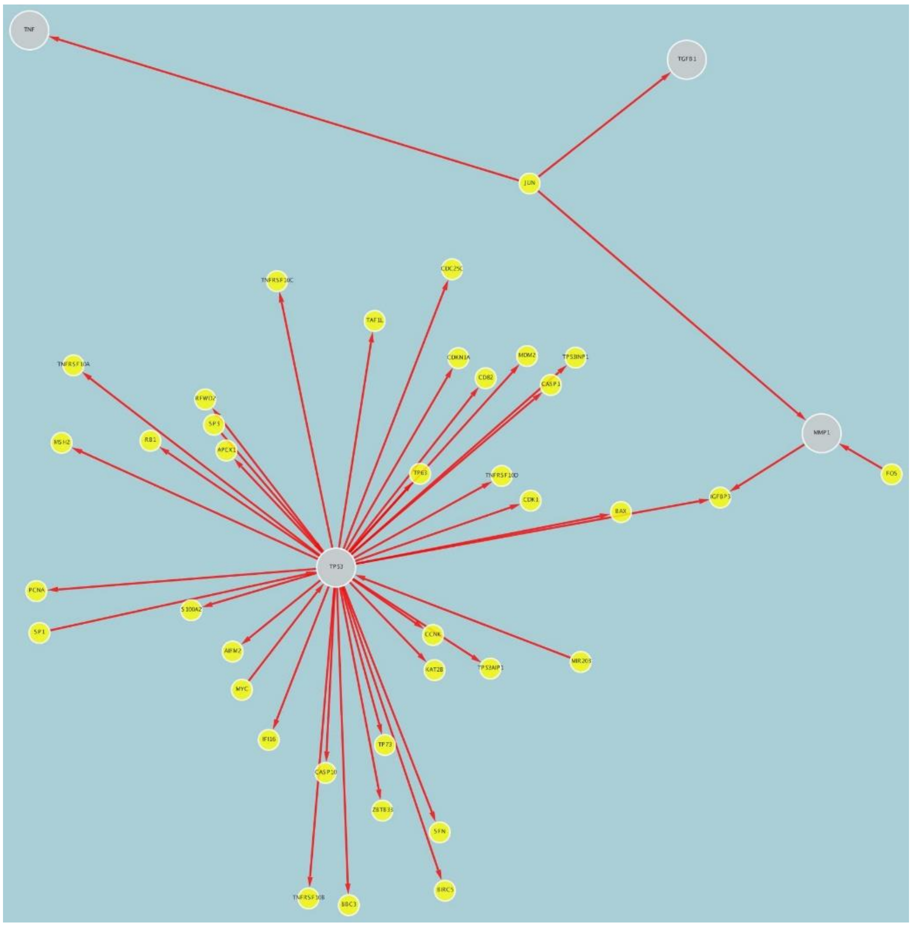


Figure 4.8: Regulatory network with 4 interrelated genes

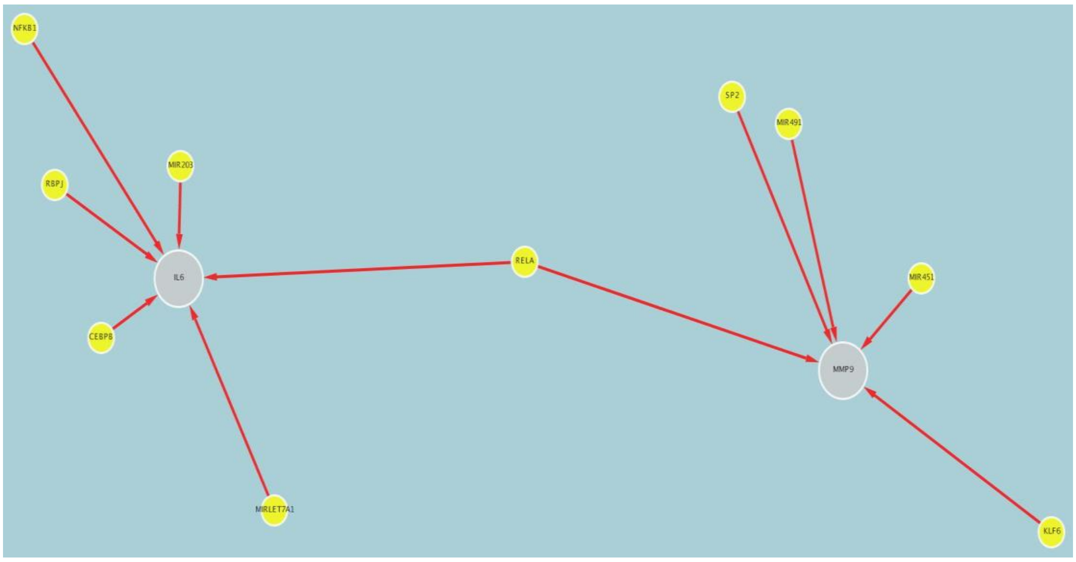


Figure 4.9: Regulatory network with 2 interrelated genes

4.4 Drug Design

The conclusive goal of this study is a common pathway drug design for investigated diseases. Acquiring successful medications of a disease from target identification is inadequate. Real drug developing is needed. A drug must dominate the target proteins in such a way that it does not intervene with normal consequences. Several bioinformatics tools are developed to achieve protein activity. UniHi is one of them and very popular. For investigated diseases, common pathway drugs are designed using UniHi tool. The structure-based drug designs are displayed in the Figure 4.10 and the Figure 4.11. The affected and unaffected proteins demonstrate in the two networks too. The mapped red color target proteins have direct compound with the aimed genes and yellow colors have other vice versa connection.

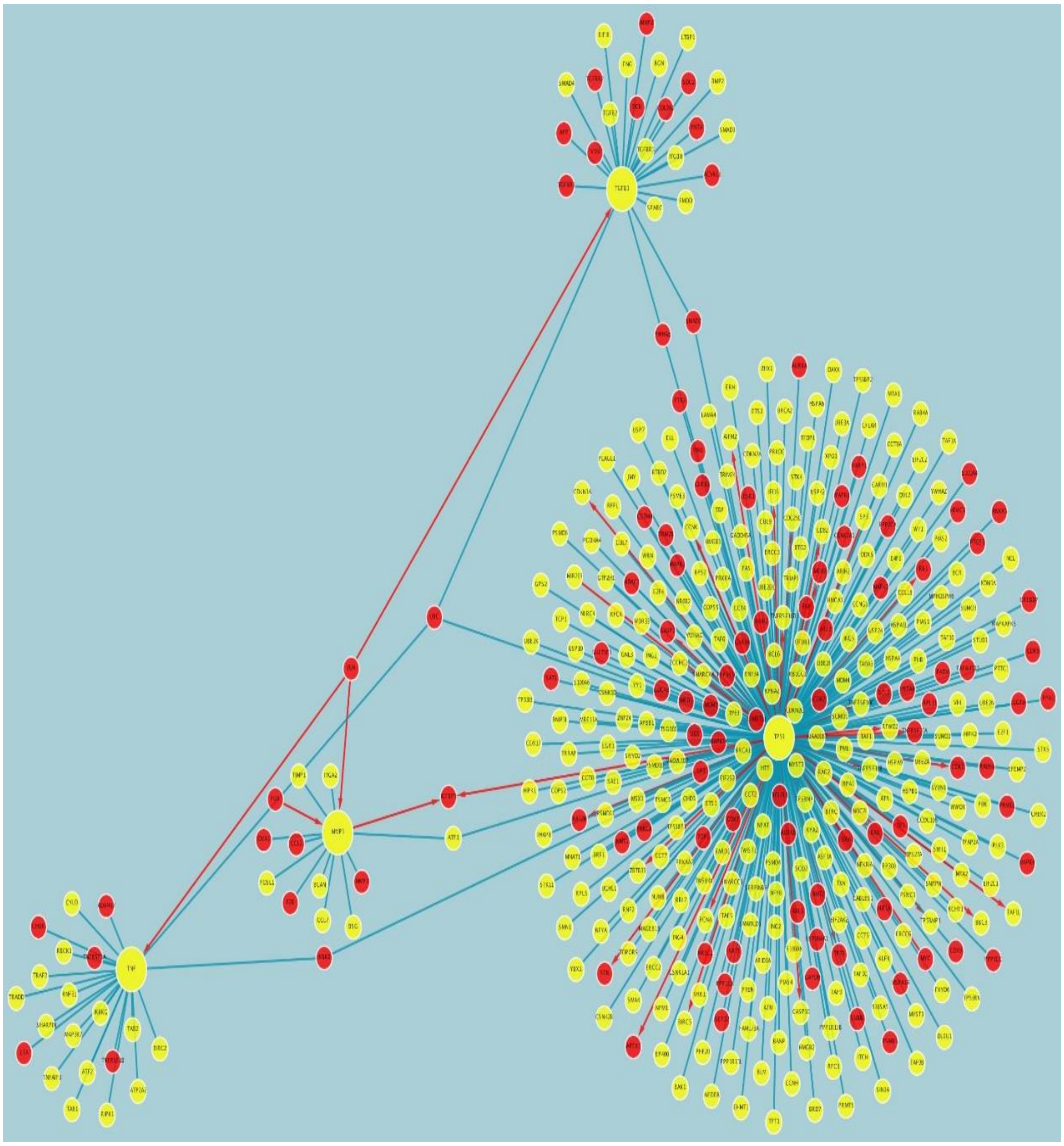


Figure 4.10: Mapped drug target network with 4 interrelated genes

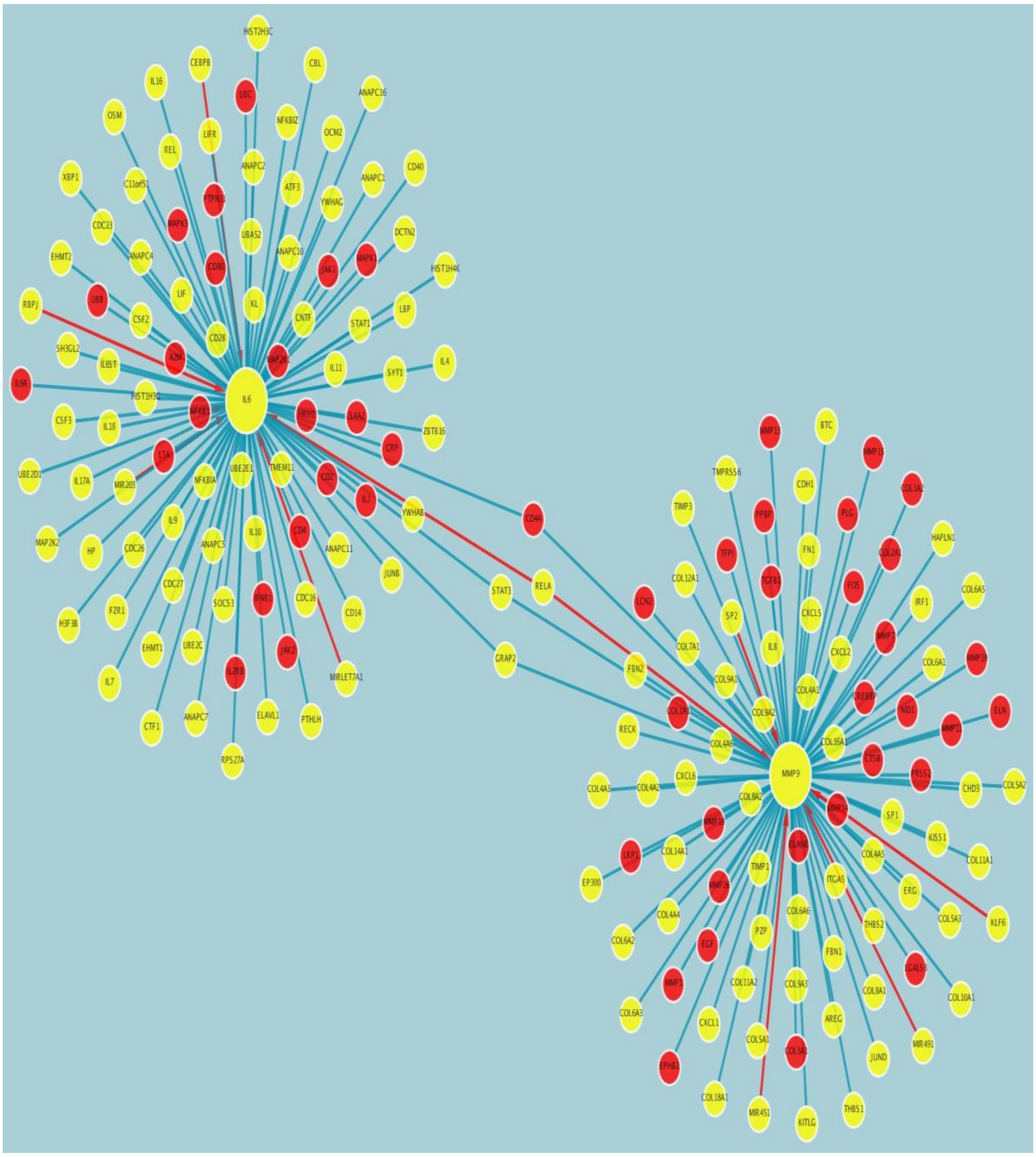


Figure 4.11: Mapped drug target network with 2 interrelated genes

A drug must bind to a specific point on aberrant protein or nucleotide to control the disease. We need to identify main molecules that have direct activity against diseases. With the help of filtering techniques directly connected proteins associated with the investigated disease are found and displayed in the Figure 4.12 and the Figure 4.13 from a viewpoint. A common drug is designed that will cure all types of CVDs.

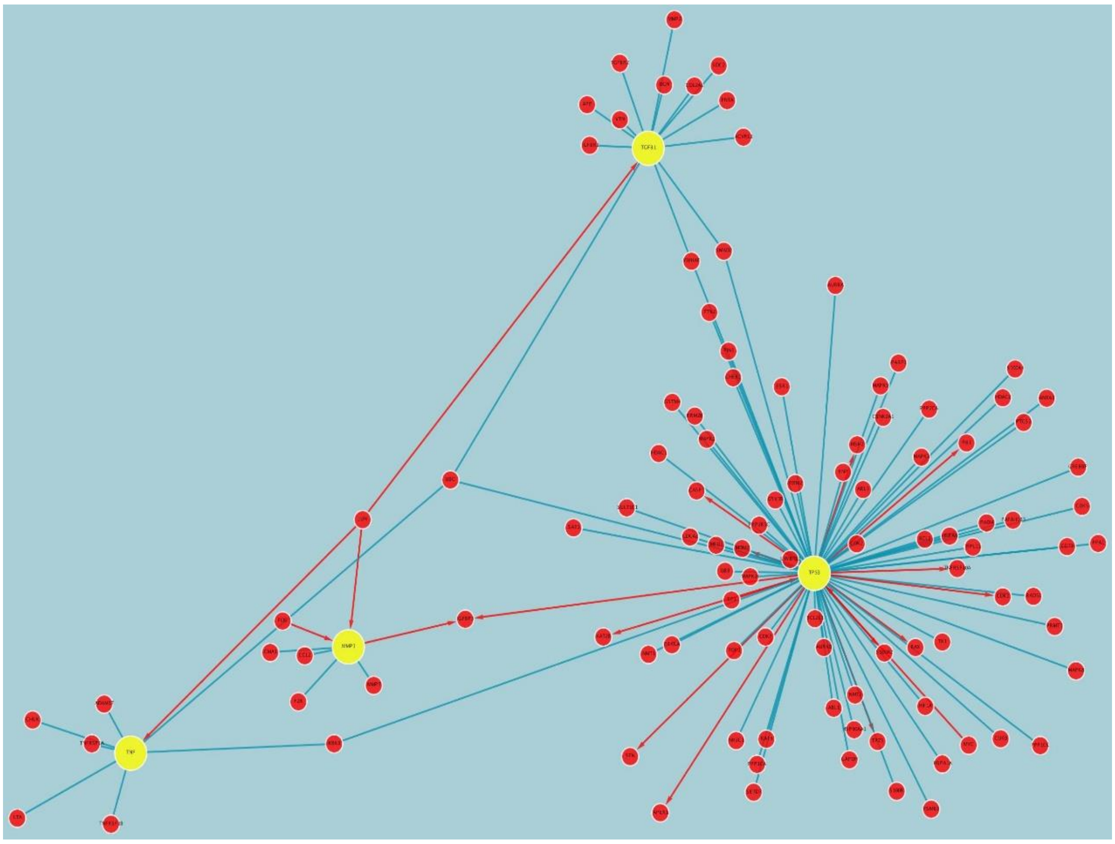


Figure 4.12: Drug target network with 4 interrelated genes after filter

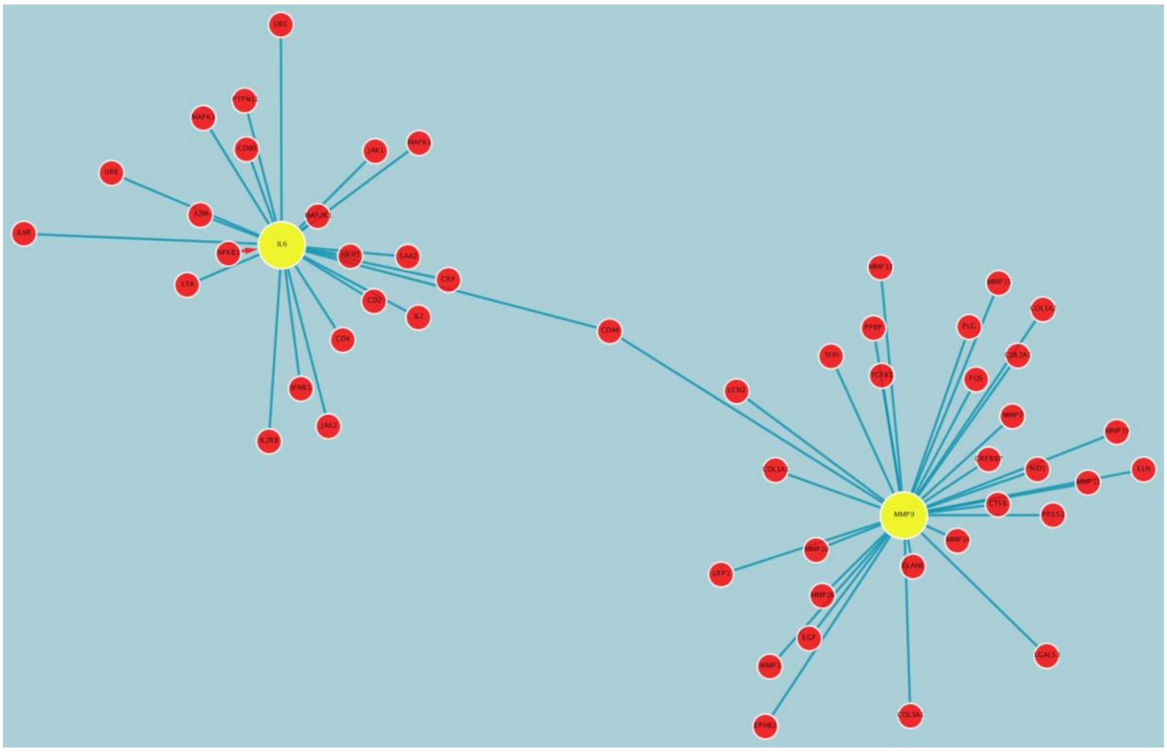


Figure 4.13: Drug target network with 2 interrelated genes after filter

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

The previous chapters are indispensable for resulted in this chapter. This chapter narrates findings along with contributions and the future expansion in the following sections based on overall performance and evaluation of preceding chapters.

5.1 Findings and Contributions

This study focuses on the risk assessment and drug design of Heart diseases, the disease that impacts human life and top of all non-communicable diseases in the world. Foremost Heart disease is same as CVDs, in which 5 types are responsible for the global deaths and various are leading top position. Drug design of diseases is a crucial field in the bioinformatics, biomedical and Pharmaceutical R & D. PPI is an integral part of drug design. The contribution of Proteomic and genomics in bioinformatics are incredible. The advancement of bioinformatics tools and databases reveal a new research area and made future task apparent. The integration between data mining and bioinformatics is making computer-aided drug design more effective. Protein-based therapeutics have also been developed in bioinformatics broadly.

It is very important to identify disease affected genes for designing a drug. The research work of Habib (2017) showed that while investigating more than one disease, attaining linkage of the genes among associated diseases is very important. Correlated genes or linkage genes and common genes associated with diseases are very beneficial to analyze and design drug of diseases accurately. Directly interconnected genes evaluated in the PPI network show a common pathway among associated disease genes.

Common risk factors are caused all types of CVDs. KDD the data mining application is performed to extract common genes from NCBI gene dataset. Using Unihi tool, PPI is generated from common genes. Interconnected genes are identified in the PPI network based on regulatory interactions. A structure-based common pathway drug is designed using UniHi tool that will heal heart diseases.

5.2 Recommendations for Future Works

Bioinformatics creates new interesting research areas. Proteomics enhanced PPI network more accurate and effective. The future resolution of this study is to work on various other correlated diseases to design a structure-based common drug.

REFERENCES

American Heart Association. (2017). Heart Disease and Stroke Statistics 2017 At-a-Glance. on-line at: http://www. heart. org/idc/groups/ahamahpublic/@ wcm/@ sop/@ smd/documents/downloadable/ucm_491265. pdf.

Bangladesh. (n.d.). Retrieved October 2, 2018, from https://tobaccoatlas.org/country/bangladesh/

Blundell, T. L., Jhoti, H., & Abell, C. (2002). High-throughput crystallography for lead discovery in drug design. Nature reviews Drug discovery, 1(1), 45.

Blundell, Tom L., et al. "Structural biology and bioinformatics in drug design: opportunities and challenges for target identification and lead discovery." Philosophical Transactions of the Royal Society of London B: Biological Sciences 361.1467 (2006): 413-423.

Campaign essentials. (2016, March 03). Retrieved October 4, 2018, from https://www.who.int/campaigns/world-health-day/2013/campaign_essentials/en/

Cheong, W. S. (2014, December). Overweight and Obesity in Asia | Gen Re. Retrieved from http://www.genre.com/knowledge/publications/uwfocus14-2-cheong-en.html

Coronary Heart Disease in Bangladesh. (n.d.). Retrieved October 8, 2018, from https://www.worldlifeexpectancy.com/bangladesh-coronary-heart-disease

Dégano, I. R., Salomaa, V., Veronesi, G., Ferriéres, J., Kirchberger, I., Laks, T., ... & Elosua, R. (2015). Twenty-five-year trends in myocardial infarction attack and mortality rates, and case-fatality, in six European populations. Heart, heartjnl-2014.

Drug. (2018, December 08). Retrieved from https://en.wikipedia.org/wiki/Drug

Drug design. (n.d.). Definitions.net. Retrieved October 10, 2018, from https://www.definitions.net/definition/drug+design.

Drug design. (2018, December 05). Retrieved from https://en.wikipedia.org/wiki/Drug_design

Drug design and discovery special interest group. (n.d.). Retrieved September 12, 2018, from https://www.fip.org/Drug_Design_Discovery

Fatema, K., Zwar, N. A., Milton, A. H., Ali, L., & Rahman, B. (2016). Prevalence of risk factors for cardiovascular diseases in bangladesh: a systematic review and meta-analysis. PloS one, 11(8), e0160180.

Feng, Y., Wang, Q., & Wang, T. (2017). Drug Target Protein-Protein Interaction Networks: A Systematic Perspective. BioMed research international, 2017.

Gene. (2018, December 03). Retrieved from https://en.wikipedia.org/wiki/Gene

Genomics. (2018, November 30). Retrieved from https://en.wikipedia.org/wiki/Genomics

Habib, N. (2016). Application of R to investigate common gene regulatory network pathway among bipolar disorder and associate diseases. Network Biology, 6(4), 86.

Habib, N. (2017). Drug design and analysis for bipolar disorder and associated diseases: A bioinformatics approach. Network Biology, 7(2), 41.

Cholesterol and Heart Disease. (n.d.). Retrieved October 22, 2018, from https://www.webmd.com/heart-disease/guide/heart-disease-lower-cholesterol-risk#1

How High Blood Pressure Can Lead to a Heart Attack. (2016, October 31). Retrieved from https://www.heart.org/en/health-topics/high-blood-pressure/health-threats-from-high-blood-pressure/how-high-blood-pressure-can-lead-to-a-heart-attack

Hasan, K. (2018, June 1). WHO: Tobacco responsible for 1 in 5 deaths in Bangladesh. Dhaka Tribune. Retrieved from https://www.dhakatribune.com/health/2018/06/01/tobacco-1-in-5-deaths-bangladesh

Hypertensive heart disease. (2018, November 26). Retrieved from https://en.wikipedia.org/wiki/Hypertensive_heart_disease

IDF SEA members. (n.d.). Retrieved September 29, 2018, from https://www.idf.org/our-network/regions-members/south-east-asia/members/93-bangladesh.html

IDF diabetes atlas - 8th edition. (n.d.). Retrieved October 30, 2018, from http://www.diabetesatlas.org/

Koene, R. J., Prizment, A. E., Blaes, A., & Konety, S. H. (2016). Shared risk factors in cardiovascular disease and cancer. Circulation, 133(11), 1104-1114.

Messner, B., & Bernhard, D. (2014). Smoking and Cardiovascular DiseaseSignificance: Mechanisms of Endothelial Dysfunction and Early Atherogenesis. Arteriosclerosis, thrombosis, and vascular biology, 34(3), 509-515.

Mozaffarian, D., Benjamin, E. J., Go, A. S., Arnett, D. K., Blaha, M. J., Cushman, M., ... & Howard, V. J. (2015). Heart disease and stroke statistics—2016 update: a report from the American Heart Association. Circulation, CIR-0000000000000350.

Morbidity Profile. (n.d.). *Local Health Bulletin 2016 for National Institute of Cardiovascular Disease*. Retrieved September 5, 2018, from http://app.dghs.gov.bd/localhealthBulletin2016/publish/publish.php?org=10000007&year=2016&lvl=5

Myocarditis Causes, Symptoms, Diagnosis and Treatment. (n.d.). Retrieved September 22, 2018, from https://www.myocarditisfoundation.org/about-myocarditis/?gclid=CjwKCAiA0uLgBRABEiwAecFnk7-cwExblCKx2ynyOJjpYzPt-o70Rhs82UX-Eoq0Acm0NNnlHTWVaBoCvMYQAvD_BwE

Nichols, M., Townsend, N., Scarborough, P., & Rayner, M. (2014). Cardiovascular disease in Europe 2014: epidemiological update. European heart journal, 35(42), 2950-2959.

Nichols, M., Peterson, K., Herbert, J., Alston, L., & Allender, S. (2016). Australian heart disease statistics 2015. Melbourne: National Heart Foundation of Australia.

Nigro, E., Scudiero, O., Monaco, M. L., Palmieri, A., Mazzarella, G., Costagliola, C., ... & Daniele, A. (2014). New insight into adiponectin role in obesity and obesity-related diseases. BioMed research international, 2014.

Oti, M., Snel, B., Huynen, M. A., & Brunner, H. G. (2006). Predicting disease genes using protein–protein interactions. Journal of medical genetics, 43(8), 691-698.

Pattin, K. A., & Moore, J. H. (2009). Role for protein–protein interaction databases in human genetics. Expert review of proteomics, 6(6), 647-659.

Proteomics. (2018, November 19). Retrieved from https://en.wikipedia.org/wiki/Proteomics

Protein–protein interaction. (2018, November 29). Retrieved from https://en.wikipedia.org/wiki/Protein–protein_interaction.

Raza, K. (2012). Application of data mining in bioinformatics. arXiv preprint arXiv:1205.1125.

Risk factors. (2017, October 05). Retrieved from https://www.who.int/topics/risk_factors/en/

Risk factor. (2018, May 13). Retrieved from https://en.wikipedia.org/wiki/Risk_factor

Salahuddin, T. (2018, September 23). Obesity is increasing among the younger generation in Bangladesh. Retrieved October 19, 2018, from https://www.thedailystar.net/health/obesity-increasing-in-bangladesh-younger-generation-1637107

Stroke death rate by country. (n.d.). Retrieved from https://www.worldlifeexpectancy.com/cause-of-death/stroke/by-country/

Tobacco. (n.d.). Retrieved September 30, 2018, from http://www.who.int/news-room/fact-sheets/detail/tobacco

Islam, AKM Monwarul, and Abdullah AS Majumder. "Hypertension in Bangladesh: a review." Indian heart journal 64.3 (2012): 319-323.

University of Regina DBD. (n.d.). Overview of the KDD Process. Retrieved November 1, 2018, from http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html

Watkins, D. A., Johnson, C. O., Colquhoun, S. M., Karthikeyan, G., Beaton, A., Bukhman, G., ... & Nascimento, B. R. (2017). Global, regional, and national burden of rheumatic heart disease, 1990–2015. New England Journal of Medicine, 377(8), 713-722.

What is genomics? (2016, June 08). Retrieved from https://www.ebi.ac.uk/training/online/course/genomics-introduction-ebi-resources/what-genomic 39

What is proteomics? (2018, September 12). Retrieved from https://www.ebi.ac.uk/training/online/course/proteomics-introduction-ebi-resources/what-proteomics

What is Heart Disease? - Australian Heart Research. (n.d.). Retrieved September 10, 2018, from https://www.australianheartresearch.com.au/about-heart-disease/what-is-heart-disease/

Yu, L. R., Stewart, N. A., & Veenstra, T. D. (2010). Proteomics: The Deciphering of the Functional Genome. In Essentials of Genomic and Personalized Medicine (pp. 89-96).

Appendix – A

List of Abbreviations

CVD = Cardiovascular Disease

WHO = World Health Organization

DTV = Deep Vein Thrombosis

RHD = Rheumatic Heart Disease

HBP = High Blood Pressure

KDD = Knowledge Discovery in Database

PPI = Protein-Protein Interaction

NCBI = National Information Center of Biotechnology