ANALYSIS OF TOPOLOGICAL PROPERTIES AND DRUG DISCOVERY FOR BIPOLAR DISORDER AND ASSOCIATED DISEASES: A BIOINFORMATICS APPROACH

By

KHAIRUL ALAM SHADHIN
162-35-145

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

Semester Spring – Year-2021

IM: Dr. Imran Mahmud; SMR: S A M Matiur Rahman; RZ: Raihana Zannat; NH: Nayeem Hasan; SI: Mr.Shariful Islam; SFR: SK. Fazlee Rabby; MA: MarziaAhmed; RM: Md. Rajib Mia.

APPROVAL

This Project/Thesis titled "Analysis of topological properties and drug discovery for bipolar disorder and associated diseases: A bioinformatics approach", submitted by Khairul Alam Shadhin, ID: 162-35-145 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 B.Sc in Software Engineering and approved as to its style and contents.

BOARD OF EXAMINERS

Dr. Imran Mahmud
Associate Professor and Head in Charge — **Chairman**
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Mr. S A M Matiur Rahman
Associate Professor — **Internal Examiner 1**
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Md. Shariful Islam
Lecturer — **Internal Examiner 2**
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Md. Fazle Munim
Technology Expert — **External Examiner**
Digital Finance, Information Security & Emerging Technology,
Aspire to Innovate Programme (a2i)
Bangladesh Government.

ACKNOWLEDGMENT

DEDICATION

TABLE OF CONTENTS

APPROVAL **i**

ACKNOWLEDGEMENT **ii**

DEDICATION **iii**

TABLE OF CONTENTS **iv**

LIST OF TABLES **vi**

LIST OF FIGURES **vii**

LIST OF ABBREVIATIONS **viii**

ABSTRACT **ix**

CHAPTER 1: INTRODUCTION **10**

1.1 Background 12

1.2 Motivation of the Research 14

1.3 Problem Statement 15

1.4 Research Questions 15

1.5 Research Objectives 16

1.6 Research Scope 16

1.7 Thesis Organization 16

CHAPTER 2: LITERATURE REVIEW **18**

2.1 Introduction 18

2.2 Related Work 18

CHAPTER 3: RESEARCH METHODOLOGY **19**

3.1 Gene Collection 20

3.2 Preprocessing 20

3.3 Gene Mining 20

3.4 Generic PPI 20

3.5 Co-Expression 21

3.6 Clustering 21

3.6.1 K-means Clustering 22

3.6.2 MCL Clustering 22

3.7 Gene Regulatory Network 22

3.8 Protein-drug Interaction 23

3.9 Protein-Chemical Interaction 23
CHAPTER 4: RESULTS AND DISCUSSION **24**
4.1 Gene Collection 24
4.2 Gene Mining, linkage & common gene finding 24
4.3 Generic PPI 26
4.4 Topological Properties 27
4.5 Co-expression and physical interaction 32
4.6 Clustering 35
4.6.1 K-means clustering 35
4.6.2 MCL Clustering 35
4.7 Gene regulatory network 38
4.8 Protein-drug interaction 40
4.9 Protein-chemical interaction 42
CHAPTER 5: CONCLUSIONS AND RECOMMENDATIONS 1 **44**
5.1 Findings and Contributions 1 44
5.2 Recommendations for Future Works 1 44
REFERENCES **45**

LIST OF TABLES

Table 1. Liability genes collected from the NCBI database 24

Table 2. Common Gene between Four Disease 25-26

Table 3. Topological properties 28

LIST OF FIGURES

Figure 1. Flowchart for proposed Methodology 19

Figure 2. Venn analysis for BD, CHD, SCH and ST 26

Figure 3. A network of 38 common responsible genes for PPI 27

Figure 4. Average Clustering coefficient for PPI 29

Figure 5. Betweenness centrality for PPI 29

Figure 6. Node-degree distribution for PPI 30

Figure 7. Neighborhood connectivity for PPI 30

Figure 8. Shortest path length distribution for PPI 31

Figure 9. Stress centrality distribution for PPI network 31

Figure 10. Topological coefficients for PPI 32

Figure 11. Co-expression 33

Figure 12. Physical interaction 34

Figure 13. K-means Clustering 36

Figure 14. MCL Clustering 37

Figure 15. Gene-miRNA Interaction 38

Figure 16. TF-gene Interaction 39

Figure 17. TF-miRNA co-regulatory network 40

Figure 18. Protein-drug interaction 42

Figure 19. Protein-chemical interaction 43

LIST OF ABBREVIATIONS

WHO = World Health Organization

BD = Bipolar Disorder

SCH = Schizophrenia

CHD = Coronary Heart Diseases

ST = Stroke

CAD = Coronary artery disease

NCBI = National Center for Biotechnology Information

IHD = Ischemic Heart Disease

PPI = Protein-protein Interaction

SIF = simple interaction format

MCL = Markov Cluster

GRN = Gene Regulatory Network

PDI = Protein-drug Interaction

PCI = Protein-chemical Interaction

T2D = Type 2 diabetes

AD = Alzheimer's disease

ABSTRACT

Background: Bioinformatics is the perfect combination of computer as well as biology. We utilize the informatics techniques and imply them on to the biological information to processing genes, genomes, Proteins, sequence, structures and all this things to get what we want. Bioinformatics can helps us today to discover the diseases that are related to other disease lead to each other. Many researcher have work on this field of bioinformatics to find the association of disease to design the drugs. By the use of bio-informatics tools, the area of bio-informatics analysis are expending day by day. Bipolar or manic-depressive disorder might be characterized as one of the most crippling mental problems that affect the people of early age and grown-ups. A few examinations have demonstrated that bipolar issue is related with an expanded danger of Bipolar Disorder, Schizophrenia, Coronary Heart Diseases, and Stroke. Thus, these four bipolar disorders must have some kind of genetic association between them. The objective of the present study was to investigate the association between genetic mutations in the four above listed diseases and to create a Protein-protein interaction (PPI) network or common path. In order to reach the destination, we need to find out about the genetic relationship between BD, SCH, CHD, ST, because it will help us understand the gene interconnections and the connections between them that help to develop the drug design for all the disease. Genes responsible for these diseases are gathered, pre-processed, processed and mining using python. This exploration is expected to carry out further measurements in the field of drug structure and also contributes to the biological and biomedical sector.

Keywords: Bio-informatics, Bipolar Disorders, Schizophrenia, Coronary Heart Diseases, Stroke, Genetic Association, PPI network.

CHAPTER 1: INTRODUCTION

The World Health Organization (WHO) has reported that approximately 450 million people in total suffer from a psychological problem and one in four meets the criteria for mental illness at some point in their lifetime [1, 2]. Among mental disorders, depression is a global burden disease affecting 350 million people across the world, compared to 4.4% of the total population [3]. Depression is more common in women (5.1%) than in men (3.6%). Depression, at its worst, can lead to suicide. The World Health Organization (WHO) has indicated, suicide in the 15-19 age group is the subsequent driving reason for death. Nearly 800 000 individuals pass on consistently due to suicide, which is one individual every 40 seconds. All around, the quantity of people with basic mental problems is rising, especially in low-paying nations, such as Bangladesh, given the fact that the population is developing and more people are satisfying the age when depression and tension are common in general. Between 2005 and 2015, the approximate total number of people with depression rose to 18.4% [4]. Poverty increases the risk of depression and has an impact on people of all ages and all walks of life. Depression is also caused by unemployment, physical illness and via taking drugs.

Bipolar issue otherwise known as Manic Depressive Disease is a complex genetic condition that causes strange changes in mood, levels of action, vitality, and ability to perform daily tasks with neurotic mindset issues (influences) ranging from scandalous delight, or craziness, to severe dependency, usually accompanied by aggravations in intuition and behavior. Bipolar issue influences more than 1% of the overall population, independent of nationality, financial status, and is one of the main sources of impairment due to its impact on young people [5]. Bipolar disorder is one of the ten most debilitating conditions in the world, taking away years of healthy functioning from people who have a disease [6]. The lifetime prevalence for bipolar disorder is between 1.3 and 1.5%. The mortality rate of bipolar disorder was a few times more than that of the general death. In the end, the rate of death of bipolar disorder was a few times higher than that of general death.

Schizophrenia (SCH) and bipolar disorder (BD) are two related mental conditions that together make a remarkable commitment to the worldwide responsibility of disease [7].

About 2% of the population is affected. The risk factor of schizophrenia is higher in case of bipolar disorder among the chance of getting into it through family members influence. Schizophrenia affects more than 21 million people around the world. Two investigations have demonstrated that people conceived or raised in urban communities are at an increased risk of creating schizophrenia as opposed to those conceived or brought up in rural areas. This is predictable with studies of mental hospitalization for real psychological illness conducted somewhere between 1880 and 1962, which demonstrated higher hospitalization rates for states with progressively urbanized populations [8]. There is generous evidence of incomplete coverage between hereditary effects in schizophrenia and bipolar issues, with family, reception and twin investigations demonstrating a genetic relationship of approximately 0.6 disorders [9]. Recently, another study has shown that biologically closely related mental illnesses such as schizophrenia, major depression, and bipolar disorder.

Coronary artery disease (CAD), also known as coronary heart disease (CHD) or ischemic heart disease (IHD), is the major cause of death. As the World Health Organization (WHO) indicates, a total of 3.8 million males and 3.4 million females died of coronary heart disease each year. In excess of 600,000 Americans die of coronary illness every year. That is one in every four passing's in this country every year, it is responsible for over 73,000 deaths in the UK, 1 in 6 males and 1 in 10 females. Such modifiable risk factors for cardiovascular disease have emerged in both people with bipolar disorder and people with depression who used drugs.

In the last few decades, stroke has been the main cause of death over East Asian countries. Globally stroke is the second-largest cause of death and the third-largest cause of disability [10]. Stroke, the abrupt death of some brain cells because of the absence of oxygen when the blood stream to the brain is lost because of blockage of the artery to the brain, it is additionally the main source of schizophrenia and despondency. Minimum one person dies in every 40 seconds due to stroke and about 800,000 people in a year. WHO 1990-2006 study on women's death from stroke is higher than men, and 60 percent of cases occur when the age is above 75. A load of stroke has expanded in younger people more than 65 years old in the course of the most recent couple of decades, among the adult aged 20 to 64, the stroke rate is being more than 25% and it is increasing day by day [11]. Universally, low-and middle-income nations like Bangladesh are having 70% of strokes and 87% of both stroke-related passing's and

disability [12]. An examination attended in 8 diverse European nations found that the danger of stroke expanded by 9% every year in men and 10% every year in ladies [13]. Around 3 to 4% of the all-out social insurance consumption in Western nations is as of now spent on stroke [14].

From the above discussion, we can see that it can be directly or indirectly related to one another in bipolar disorder, stoke, heart disease, and schizophrenia. They also have a genetic relationship with each other. Depending on the genetic relationship, they have common genes that are interrelated to each other. A gene regulatory network pathway will be maintained for common genes. The PPI network is useful for this purpose. This Research has employed a bioinformatics method to build a common gene network and predicts the development of a PPI network for BD, SCH, CHD, and ST. If any programming language can help to explore the targeted result, it will also be a good achievement. This research paper uses the programming language R to analyze common gene lists and the gene regulatory network for bipolar disorder and related diseases. This examination examines both the common genes and the common gene regulatory pathway between bipolar disorder and related illnesses.

1.1 Background

The medical field is now moving through bioinformatics. Not only is bioinformatics used in the medical field, but it is also used in agriculture, fisheries, animal science, etc. Bioinformatics is an interdisciplinary science that develops methods for the storage, retrieval and analysis of biological data by combining various other disciplines, such as biology, mathematics, computer science, and statistics [N-1]. The first person to use the word 'Bioinformatics' in 1970, referring to the use of information technology to study biological systems, was Paulien Hogeweg, a Dutch system-biologist [N-2, 3]. In the field of medicine and microbial genomes, bioinformatics is widely used. The application of bioinformatics to these fields is as follows:

The greatest killers of children and young adults in the world are now infectious diseases. According to WHO "They account for more than 13 million deaths a year -

one in two deaths in developing countries". In developing countries like Bangladesh, most deaths from infectious diseases occur. One of the key challenges faced by mankind is the production of affordable and successful medicines for a disease. Reasonable drug design using Bioinformatics will solve this problem. Computational methods are used to estimate that the time and expense of production can be minimized by an efficient and effective drug design process. Bioinformatics tools help to make Protein-Drug Interaction and Protein-Chemical Interaction. This method brings it to the next step to produce medicines. The structure of proteins will affect the discovery of drugs at any point in the design process. Classically, it is used in lead optimization, a method that uses the structure to direct the chemical modification of the lead molecule in order to provide an optimized fit in terms of form, hydrogen bonds and other non-covalent interactions with the target [N-4, 5].

In the science of agriculture, bioinformatics plays an important role. As the volume of data increases exponentially, there is a parallel growth in the market for tools and techniques in data visualization, integration, analysis, prediction and management. Bioinformatics leads to vast data and great potential for huge volumes of data to be dealt with by using computational biology. In our culture, our economy, and our global climate, plant life plays a significant and diverse role. The most important plants for us are primarily crops. A challenge for modern plant biotechnology is to feed the rising world population. In crop improvement projects, bioinformatics finds direct applications. This allows researchers to relate genetic makeup to commercial characteristics. There are many methods involved in crop production programs, such as comparisons of plant genomes, genetic mapping techniques, evolutionary analyses, etc., which are now possible through the study of bioinformatics data.

Bioinformatics is a new field that offers the scientific community a fundamental opportunity to speed up biotechnology discovery, implementation and commercialization. Bioinformatics is a strong enabling technology that holds a lot of promise for the future of applications in many sectors. One of them is fisheries. It is a term which encompasses a wide range of scientific applications. Biotechnology is defined by the Canadian Environmental Protection Act as the application of science and

engineering in the direct and indirect use, in natural or modified forms, of living organisms or parts or products of living organisms. Aquatic Biotechnology includes the application of science and engineering to the direct or indirect use, in natural or modified ways, of aquatic species or parts or products of living aquatic organisms [N-7]. The aim is to optimize the aquatic environment's sustainability while preserving environmental quality and biodiversity.

Bioinformatics is a relatively new field and has evolved quite rapidly in recent years. It has made it possible to digitally test our theories and thus helps us to make a smarter and more educated decision before launching expensive experiments. Although more and more tools are being built for the study of genomes, proteomes, predicting structures, rational drug design and molecular simulations, none of them are perfect.

1.2 Motivation of the Research

One of the common, extreme, sometimes psychotic, major psychiatric disorders, the prevalence of which is intermediate between major depressive disorder and schizophrenia, is bipolar (manic-depressive) disorder [N-11]. Bipolar disorder appears to affect young people with a development and prognosis that is otherwise positive and is usually receptive to care [N-10]. BD has been stated to be the seventh and eighth leading cause of impairment for males and females for years. Many people suffer because of bipolar disorder every day.

With the help of Nahida Habib's bioinformatics approach and her team, one of the best papers on anxiety disorder using R. They work with five different anxiety disorders (Angina, Asthma, Diabetes, Heart Attack and High Blood Pressure). They try to figure out that the genetic link between the above-mentioned diseases helps to understand the gene interaction and association of the connection between them that contributes to the common design of drugs. They first try to identify genetic associations and then use R to build a regulatory interaction network. From gene or data collection, they use R to achieve desire goal.

In another paper, they try to create a common pathway and a Protein Protein Interaction Network using bioinformatics tools. They're also creating a Rnandom network using R.

From this paper, I have found a great deal of scope to work with different types of related diseases using bioinformatics tools and using R. I also have the opportunity to work with their limitation of work.

1.3 Problem Statement

I have seen some gaps after learning about previous related works, which can lead us to the next stage of our analysis. These articles are the issue statements for those papers.

I. Using small amount of data for four different diseases.
II. Only PPI and random network create using R
III. Without any clustering of data.
IV. Did not work with topological properties.

1.4 Research Questions

A research question is an answerable investigation into a particular issue or problem. In a research project, it is the original phase. The 'initial step' means that the research question is the first active step in the research project after I have an idea of what I want to study. I have encountered some problems when studying, as well. Below is given the list,

I. How to collect genes for selected diseases as a data set?
II. How to sort data set?
III. Is the data set being clustered?
IV. What type of simulation tools are used to achieve the desired goal?
V. How to analyze the topological properties?

1.5 Research Objectives

The major objectives of this thesis are given below:

- Find common gene for different diseases.
- Cluster data set using K-means and MCL.
- Discover the genetically association among those diseases.
- Find out topological properties.
- Creating PPI, Co-Expression, Physical Interaction, Protein-drug Interaction and Protein-Chemical Interaction.

1.6 Research Scope

The scope of the study refers to the areas covered by my work. The scope of my research here is:

- Large data set
- Finding Common Using R
- Clustering
- Topological Properties analysis
- PPI
- Co-Expression and Physical Interaction
- Gene Regulatory Network
- Protein-Drug Interaction
- Protein-Chemical Inter

1.7 Thesis Organization

I have organized the whole paper in some sections. Where in the section 2, I have discussed about the previous related works. The research methods I applied is discussed in the section 3. I have explained how I worked to get my final result. I have shown them with some pictorial views. In section 4, I have explained my findings and final result with the help of some graphs. All the results have been explained in that section.

Finally, I concluded my work with the section 5. I gave a short brief there and also discuss about the future works that can be done from my research.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

The new vision of bioinformatics in science combines science, biology, mathematics and statistics. In the medical industry, several bioinformatics tools are added every day, opening up a new field of drug design and development. In data analysis, managing large-scale databases, identifying common data etc. Bioinformatics plays a significant role. We need to know about genetic interactions and protein pathways prior to drug design. Understanding the protein pathways PPI plays an important role. In my work, PPI is one of the most important parts of my goal of desire. This chapter contains a model for the discovery of the history of this study.

2.2 Related Work

In his work, Tomas Klingstrom (2010) tried to find Protein-Protein Interaction and Pathway with Pictorial View. The aim is to make it easier for a researcher who wants to work in this field or who like me has an interest in this field. Finding protein interactions is hard and time consuming, but they show the tools (Cytoscape) that can help to create PPIs in a short time. They also show a graphic representation of the full process. This part is helping me a lot.

In another paper, they worked on type 2 diabetes (T2D) and Alzheimer's disease (AD). That research they are trying to predict common drug design for T2D and AD. From the NCBI gene database, they extract genes. Preprocessing and gene filtering and, with the help of R, identifying common genes. They use bioinformatics methods after identifying popular genes to reach their destination. This segment helps me to have an idea for the data collection package.

CHAPTER 3: RESEARCH METHODOLOGY

Some individual advances are taken here from the information assortment to the achievement of the ideal goal. The approach of the proposed research is consequently separated into a few subsections. Figure 1 represents a graphical overview. In segments 2.1 to 2.7, each of these subsections with graphical representation is presented below.

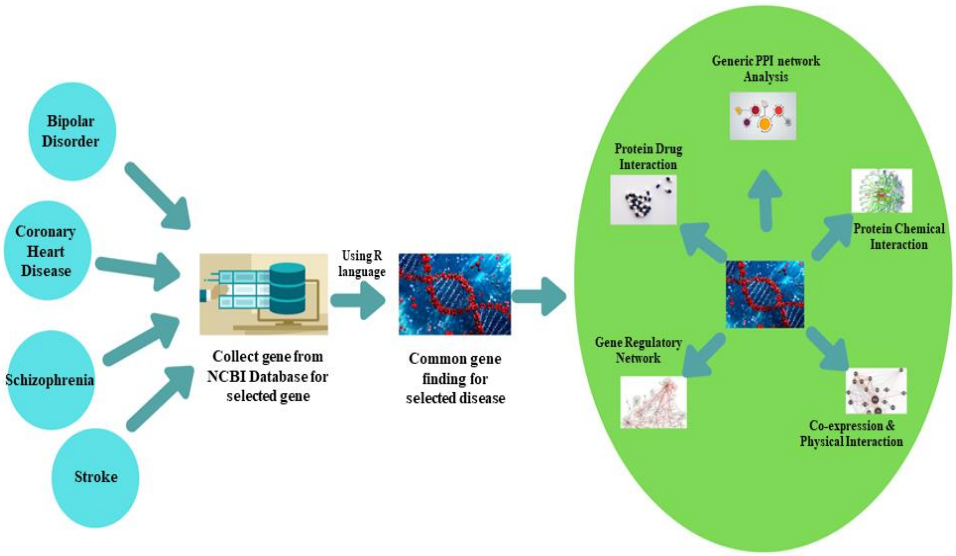


Figure 1: Flowchart for proposed Methodology.

3.1 Gene Collection

In order to analyze diseases, the genes associated with the disease need to be collected. There are very few reliable databases for bio-informatics tools and services. The National Center for Biotechnology Information (NCBI) is freely accessible and can be accessed from the Gene server website. In bio-informatics, it is an important resource for tools and services. They are stored in different databases on the basis of the nature of the various data. For different purposes, we can use different types of databases such as PubMed, GENE Bank & OMIM. Genes for Bipolar Disorder, Schizophrenia, Coronary Heart Disease, and Stroke Diseases have been collected with the help of the NCBI Gene database.

3.2 Preprocessing

All the genes associated with BD, Schizophrenia, CHD, and Stroke are collected in the early step. At first, the responsible genes for all associated with BD, Schizophrenia, CHD, Stroke are collected and the data collected are needed to modify that is called preprocessing here. After some filtering, only the genes responsible for humans are stored for further processing.

3.3 Gene mining

The data mining method is primarily used to generate appropriate data. All genes are collected and stored in a text file, and another text file contains other text or genes that are not relevant to this research. One of the most important parts of this research is gene mining, because it is gene mining which will take us to our desire goal. Any kind of gene mining error can take us away from the perfect result. Using the data mining method, the candidate genes linked to BD, Schizophrenia, CHD and Stroke were mined.

3.4 Generic PPI

To represent the direct and indirect interaction of genes and proteins between the interrelated genes, the Protein-Protein Interaction Network (PPI) is used. In the field of

bio-informatics research the PPI network plays an important role. Cytoscape is a free accessible software undertaking to organize bimolecular interaction systems with data on high-throughput articulation and other sub-atomic states into a simulated structure linked together. Presently it is a very popular, reliable bio-informatics tool that is user-friendly for bio-informatics studies and used to build PPI networks.

3.5 Co-Expression

Generally, biological networks are structured on the basis of the substances and the interactions involved. Molecular data resulting can be derived for these network gene expression networks is an example of it. In many cases, the gene co-expression network (GCN) involves the presence of a functional link between genes. GCN can be resolved by joint experiments in which various test conditions have been characterized. These networks rely on association-by-association heuristic guilt, which is widely used in genomics [15]. In general, co-expression networks are seen to be straightforward in the case of construction. On the other hand, the resulting network of linked genes is so complex, that it goes far beyond from its biological interpretation. Although GCN can be used for the identification of regulatory genes, it can be used for a variety of purposes, such as candidate disease gene prioritization. Genes are identified by this approach under different co-expression partners with conditions like state of disease and types of tissue [16]. This genes remains with phenotypic differences because they are likely to be regulatory.

3.6 Clustering

The division of data or information into collections of the corresponding objects is called cluster. The challenge is to divide the data sets into a number of groups so that data sets in the same groups are more identical to other data points in the same group than those in other groups. Certain details are ignored in order to simplify clustering. Clustering can be seen as a procedure for displaying information that supports short rundowns of information. In a wide range of applications, it plays an important role and relates to many fields. Use clustering applications to handle huge data sets and information with numerous features.

3.6.1 K-means Clustering

Minimize clustering error one of the most widely-recognized techniques is the k-means algorithm [17]. The k-means algorithm is, in any case, a neighborhood search technique and it is interesting that it experiences the genuine drawback that its show depends heavily on the underlying starting conditions. This clustering method is point-based. In order to minimize clustering error, K-means algorithm work starts with cluster centers initially placed at arbitrary positions and proceeds to the last point. Furthermore, in order to obtain near-optimal solutions using the k-means algorithm, in the initial locations of the cluster centers, many runs must be spaced separately. The center of the each cluster is the mean value of the object.

3.6.2 MCL Clustering

To identify functional modules in PPI network Markov clustering (MCL) is one of the most effective algorithm for clustering biological networks in recent time. In this section, MCL, the clustering algorithm that we used to split large components into smaller clusters, is briefly described. MCL consists of two stochastic matrix operations: ' Expand ' and ' Inflate '. The Expand operation is simply M = M ×M, and the Inflate operation raises each entry in the M matrix to the inflation parameter r (r>1, and typically set to 2), followed by the re-normalization of the sum of each column to 1 [18]. After several iterations of this method, the fundamental cluster structure of the graph gradually becomes visible. High-flow diagram regions describing clusters are divided by no-flow limits.

3.7 Gene Regulatory Network

The gene regulatory network is the collection and interaction of molecular species, which together control gene-product abundance and represent the causality of developmental processes. They clarify how the genomic grouping encodes the quality sets outflow guideline that creates formative examples bit by bit and plays out the development of numerous separation states. It governs the levels of these gene products.

3.8 Protein-drug Interaction

Throughout drug research and development, finding novel inhibitors is a significant challenge. For this attempt, the structure-based design is a basic approach and is becoming an integral part of drug development. The accurate three-dimensional design of the protein has been demonstrated for a significant number of drug targets [N-8]. The ligand binding of the cell generally follows the binding of a drug to a cell. The introduction of advanced computational tools in recent decades has provided a professional and detailed insight into various relevant aspects and the operational processes of drug binding to different functional, transport or depot proteins [N-9].

3.9 Protein-Chemical Interaction

Protein–chemical interactions are essential for any biological system; for example, they drive the metabolism of the cell or initiate many signaling cascades and most pharmaceutical interventions (21). In any case, protein-chemical interaction information is distributed across a wide range of databases and literature, making it difficult to get an overview of any chemicals of concern identified interactions.

CHAPTER 04: RESULTS AND DISCUSSION

4.1 Gene Collection

Responsible genes were collected from the NCBI database for the targeted disease. Results show 714, 366, 2958 and 1357 reliable genes for BD, CHD, Schizophrenia and Stroke. There are 691 for BD, 350 for CHD, 1741 for Schizophrenia and 796 for Stoke since processing and sorting of the associated genes for Homo-sapiens. Genes are sorted by their weight in ascending order. The numerical values of the identified responsible genes are shown in Table 1.

Table 1: The number of liability genes collected from the NCBI database for selected diseases.

Diseases Name	Total no. of Gene	Total no. of Homo Sapiens Gene
Bipolar Disorder	714	691
Coronary Heart Disease	366	350
Schizophrenia	2958	1741
Stroke	1357	796

4.2 Gene Mining, linkage & common gene finding

It identifies the linkages between BD and CHD, BD and SCH, BD and ST, CHD and SCH, CHD and ST, SCH and ST, BD & CHD & SCH, BD & CHD & ST, BD & SCH & ST, CHD & SCH & ST, BD & CHD & SCH & ST. Table 2 shows the number of common genes. 38 common responsible genes are detected between 4 selected diseases after gene linkage. Figure 2 displays the Venn analysis of the number of genes and a common gene ratio. The genes are extracted from the trusted database at the beginning

of this investigation. After that, the data set was applied to the mining algorithm. In fact, there has been a rigorous analysis of the intersection of two, three and four diseases. The total number of genes for BD, CHD, SCH, and ST are 692, 350, 1741 and 796 respectively. During the intersection process, we get 1+5+38+6=50 no of common gene between BC & CHD; 5+38+37+334=414 no of common gene between BC & SCH; 38+37+11+6=92 no of common gene between BC& ST; 5+27+38+48=118 no of common gene between CHD & SCH; 38+48+112+6=204 no of common gene between CHD & ST; 113+48+38+37=236 no of common gene between SCH & ST; 5+38=43 no of common gene between BD & CHD & SCH; 38+6=44 no of common gene between BD & CHD & ST; 38+37=75 no of common gene between BD & SCH & ST; 38+48=86 no of common gene between CHD & SCH & ST; For BD & CHD & SCH & ST we get 38 no of common gene between them. Table 2 and Figure 2 reflect with each other after the investigation, so our study has been verified.

Table 2: The common genes are between/among four diseases.

Diseases name	**Total no. of Homo sapiens gene**	**Common gene**
BD & CHD	1041	50
BD & SCH	2432	414
BD & ST	1487	92
CHD & SCH	2091	118
CHD & ST	1146	204
SCH & ST	2537	236
BD & CHD & SCH	2782	43
BD & CHD & ST	1837	44
BD & SCH & ST	3228	75
CHD & SCH & ST	2887	86

BD & CHD & SCH & ST	3578	38

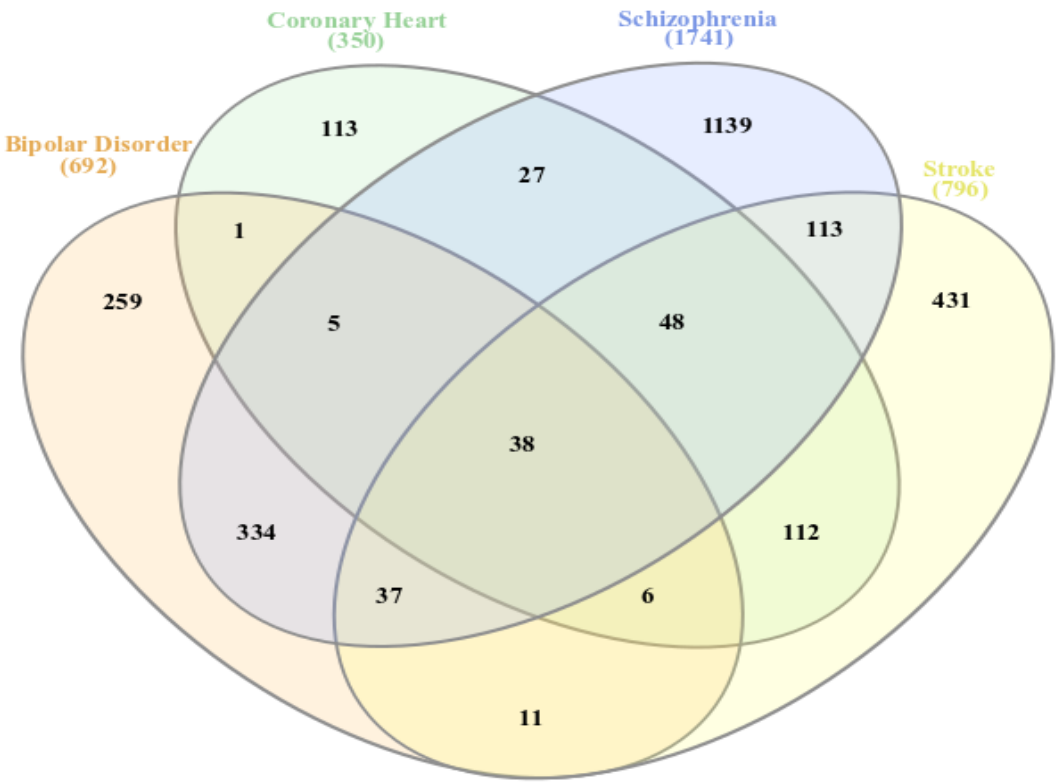


Figure 2: Venn analysis for BD, CHD, SCH and ST.

4.3 Generic PPI

NetworkAnalyst 3.0 as a strong web-based visual analytics platform for detailed gene expression data profiling, meta-analysis, and system-level interpretation (22). Simple Interaction Format (SIF) files are created for the network diagram using the NetworkAnalyst web-based tool. The PPI network is the connection between genes and hub protein, some of which are linked directly and some of which are linked indirectly. Figure 3 shows the PPI network.

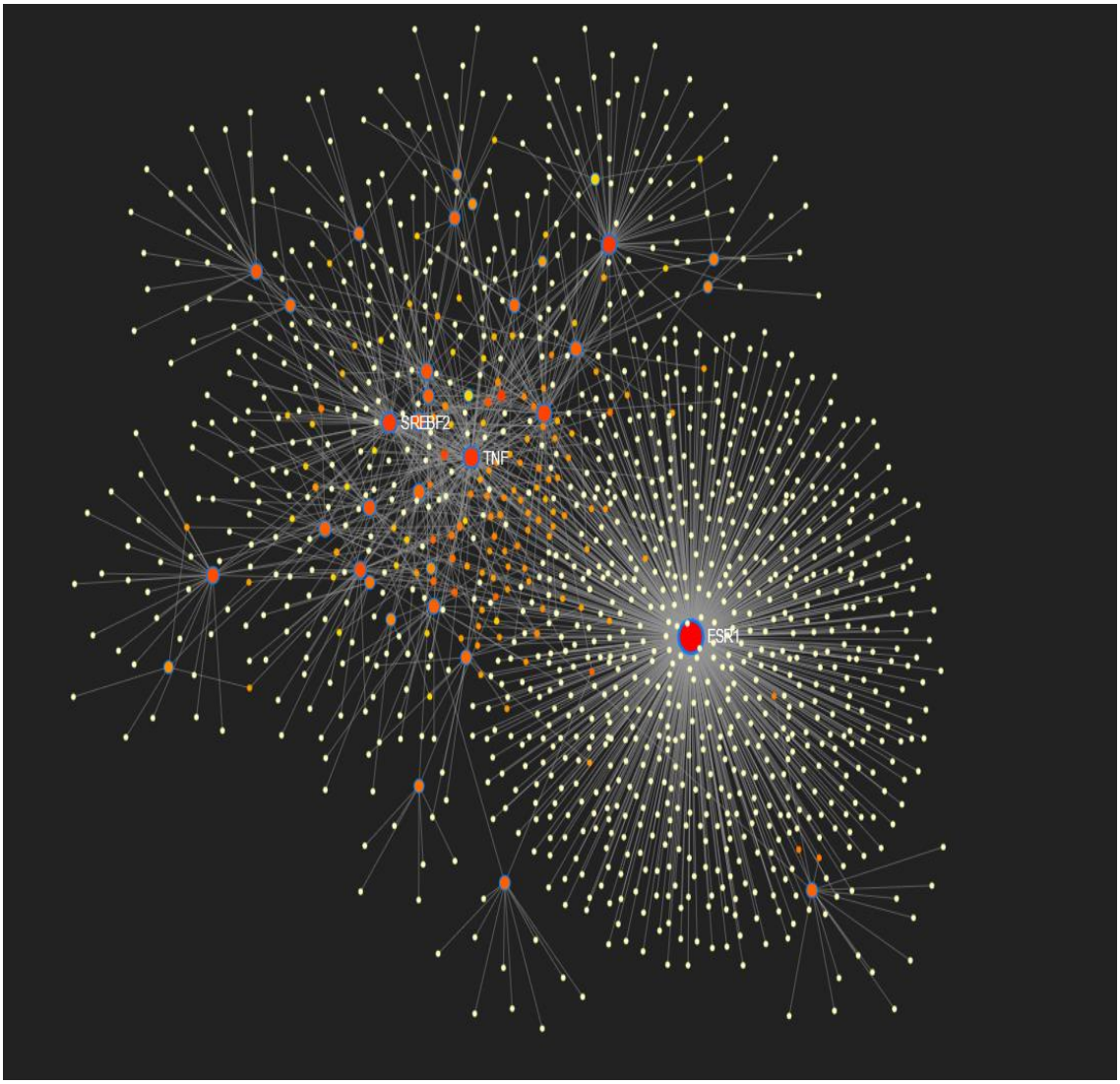


Figure 3: A network of 38 common responsible genes for protein-protein interaction (PPI). There are 1267 nodes and 1507 edges to be built in the network. Nodes are proteins, and the edges establish a relationship between proteins.

4.4 Topological Properties

Table 3 shows the average global topological properties of PPI networks. We estimate topological properties using the Cytoscape Network Analyzer application using the Simple Interaction Format (SIF) document in the Cytoscape tool. Table 4 shows the

topological properties of the top 8 liability-exposed genes. Figure 4 represents an average clustering coefficient is a function of the degree of average clustering of nodes in a network. Figure 5 represents, Betweenness centrality is a measure of a vertex's effect on the information flow for each set of vertices, assuming that data flows mainly through the shortest pathways between them. Figure 6 represents, the degree of the protein in the PPI network. Figure 7 represents the number of connections between the proteins. Figure 8 represents the shortest path length distribution in the PPI network. Figure 9 represents closeness centrality is a way to detect objects that can very effectively transmit information through a network. A node's closeness centrality determines its average distance to all other nodes. Nodes with a high score of closeness have the shortest lengths to all the other nodes. Figure 10 represents, a quantitative measurement of the degree to which a node connects neighbors with other nodes is the topological coefficient. A topological coefficient of 0 is assigned to nodes that have one or no neighbors. Figures 4-10 are the topological graph for the PPI network.

Table 3: Topological properties of the top 8 responsible gene from PPI network using Cytoscape tool.

Protein Name	**Degree**	**Betweenness Centrality**	**Closeness Centrality**	**Clustering Coefficient**	**Topological Coefficient**
ESR1	799	0.86299906	0.62119725	1.6311E-4	0.00257298
TNF	92	0.09850888	0.44281217	0.00621118	0.01248085
SREBF2	71	0.07319726	0.34123989	0.0	0.04062839
APOE	62	0.07741995	0.43912591	0.01057641	0.0174778
NOS3	58	0.05355804	0.43988881	0.01875378	0.0190718
TCF7L2	36	0.03030075	0.3296875	0.0	0.06333333
IL1B	31	0.0302616	0.33098039	0.0	0.09396914
CRP	29	0.02635653	0.32697546	0.00985222	0.0513573

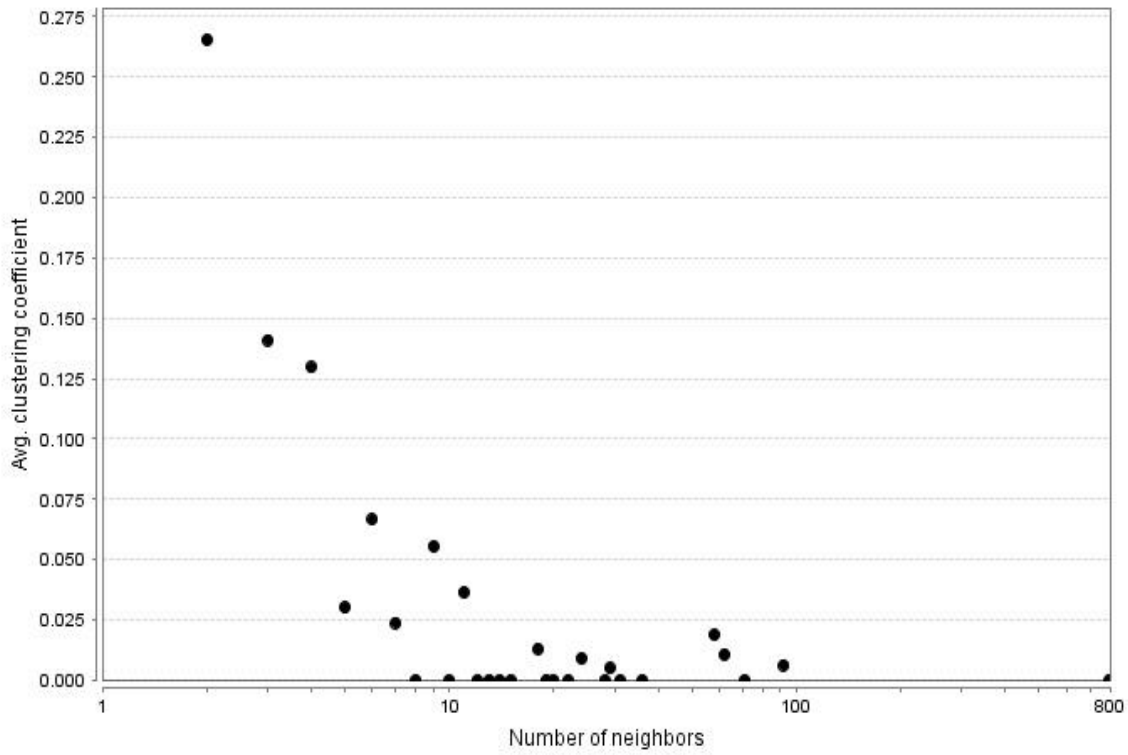


Figure 4: Average Clustering coefficient for PPI network using Cytoscape tool

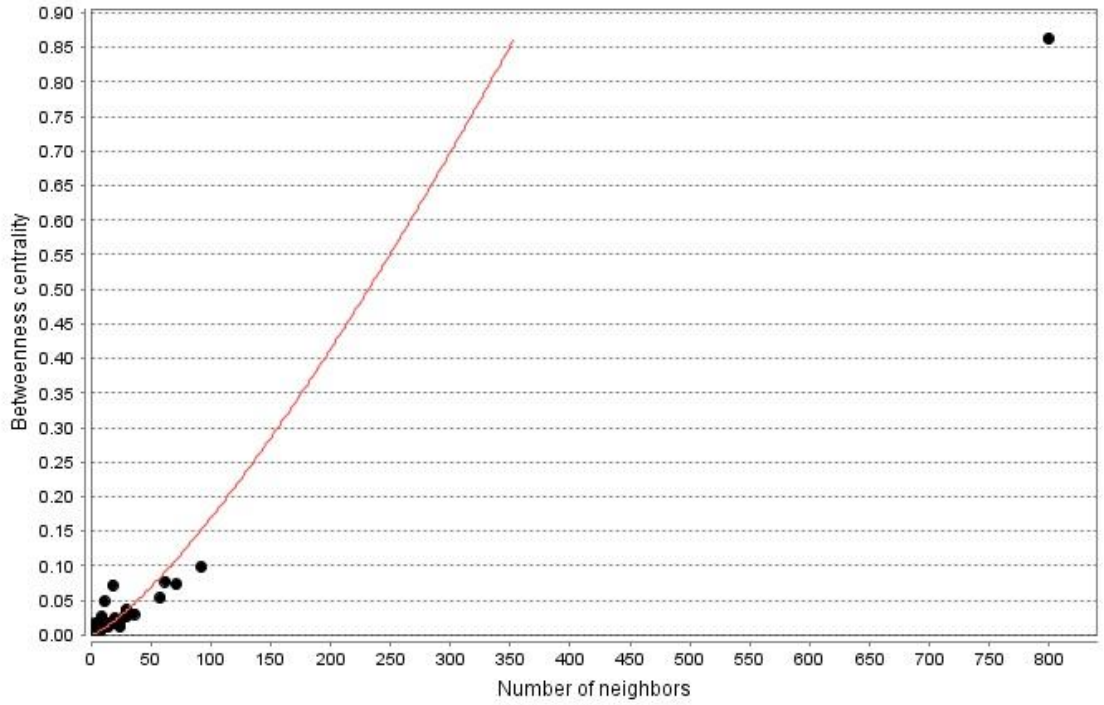


Figure 5: Betweenness centrality for PPI network using Cytoscape tool

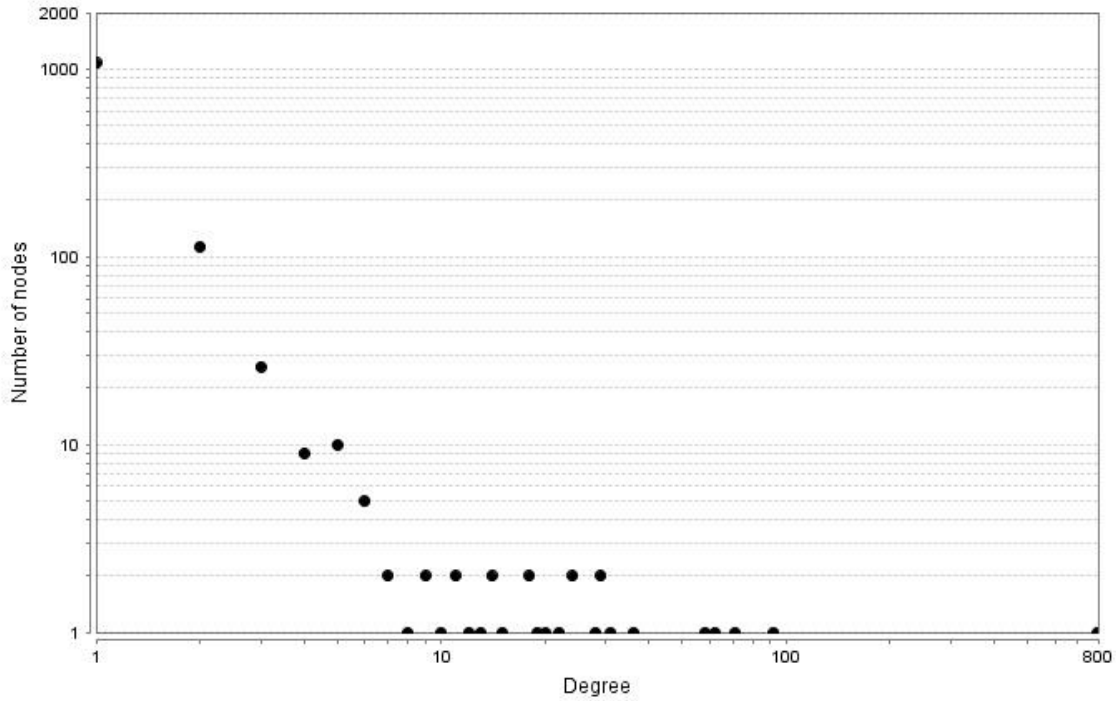


Figure 6: Node-degree distribution for PPI network using Cytoscape tool

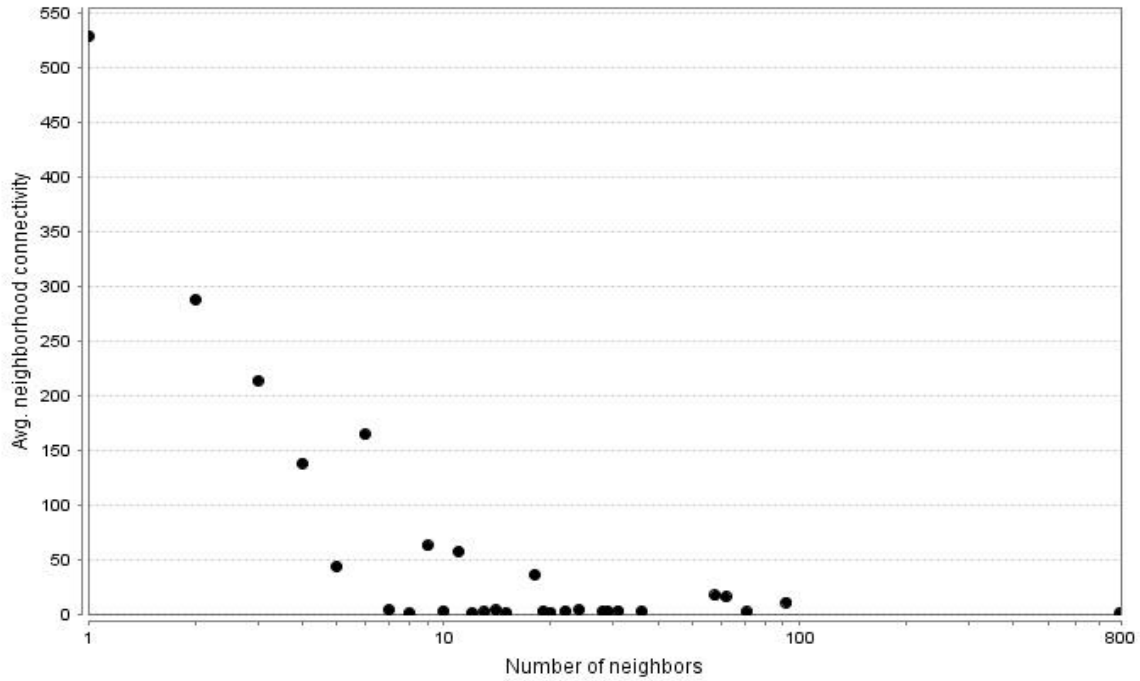


Figure 7: Neighborhood connectivity for PPI network using Cytoscape tool

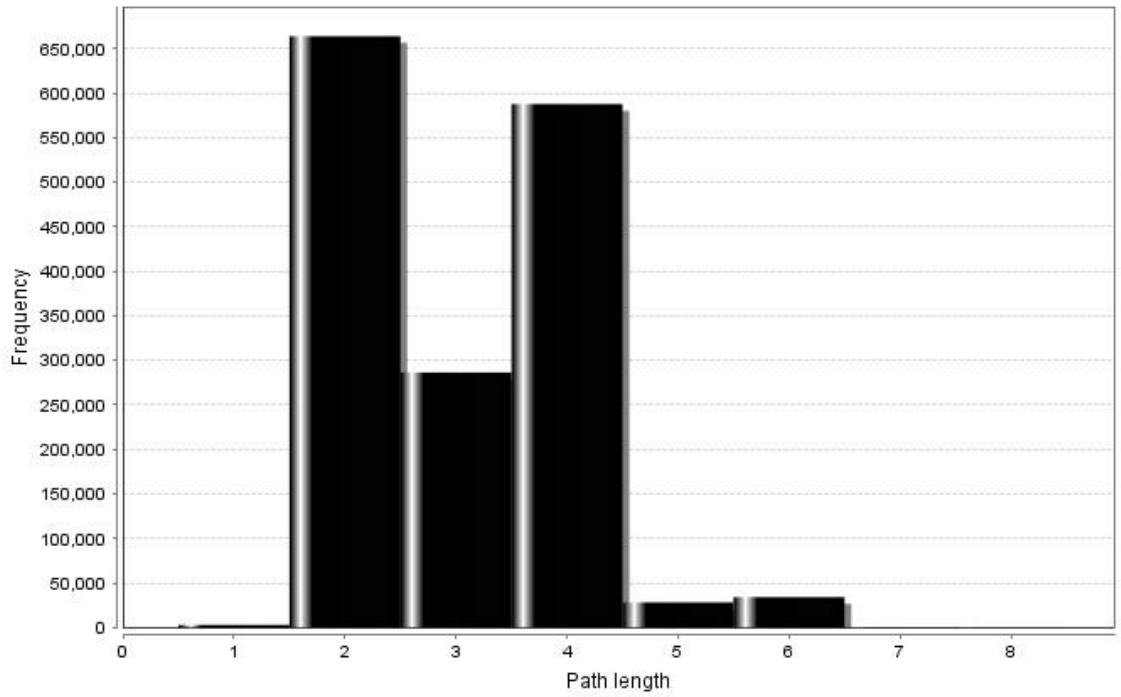


Figure 8: Shortest path length distribution for PPI network using Cytoscape tool

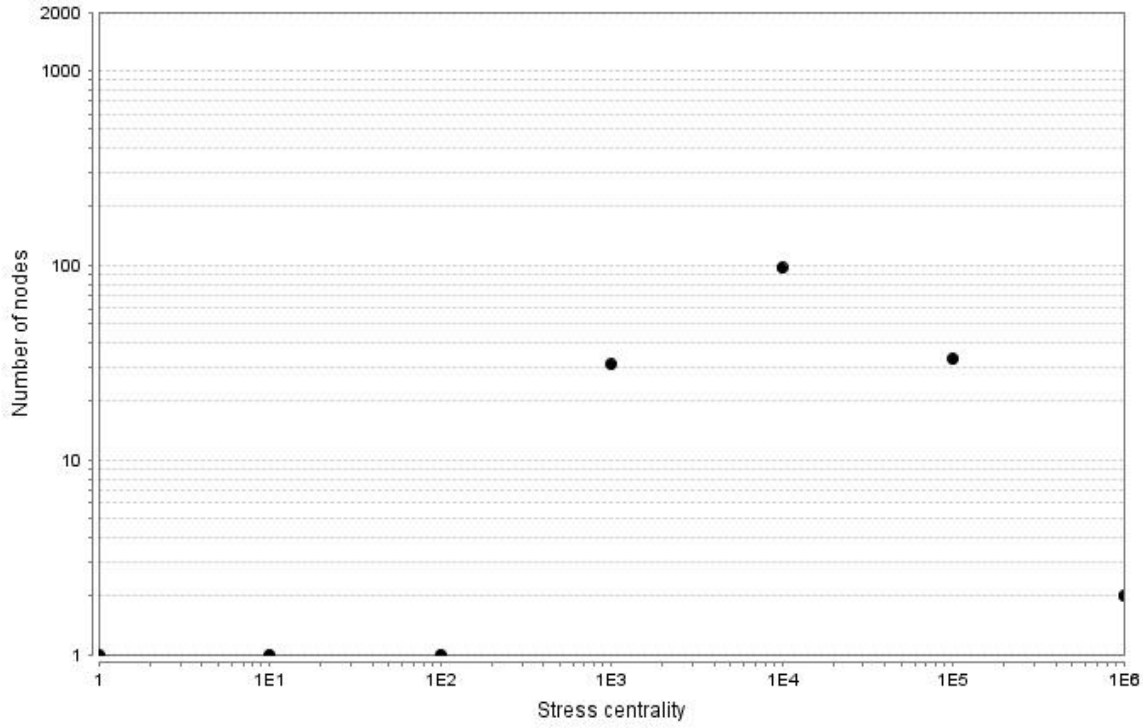


Figure 9: Stress centrality distribution for PPI network using Cytoscape tool

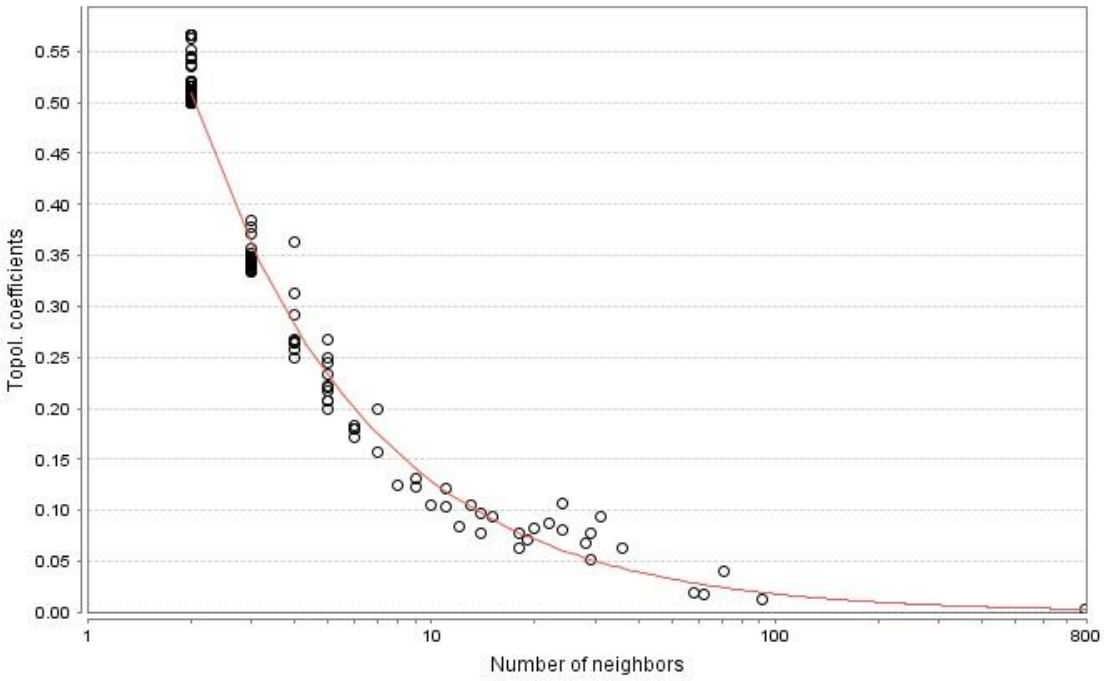


Figure 10: Topological coefficients for PPI network using Cytoscape tool.

4.5 Co-expression and physical interaction

An undirected graph, where nodes refer to genes and edges, refers to significant co-expression relationships, is known as a co-expression network. Co-expression is the first step of inference that defines the relationship between pairs of transcripts. Physical interactions between proteins can include two or more proteins, creating binary interactions and complex proteins (23). In particular, protein-protein interactions are established through the physical contact of ligands, which are often evolved in the domains of protein families (24). Some protein interactions occur with the other ligands, such as nucleic acids, lipids and some small molecules in signaling or metabolic pathways. Using topological properties, we select the top 8 genes that are responsible for building to define the genetic interactions and pathways between the 8 genes that are responsible. Figures 11 and 12 show the co-expression and physical interaction using the GeneMANIA tool.

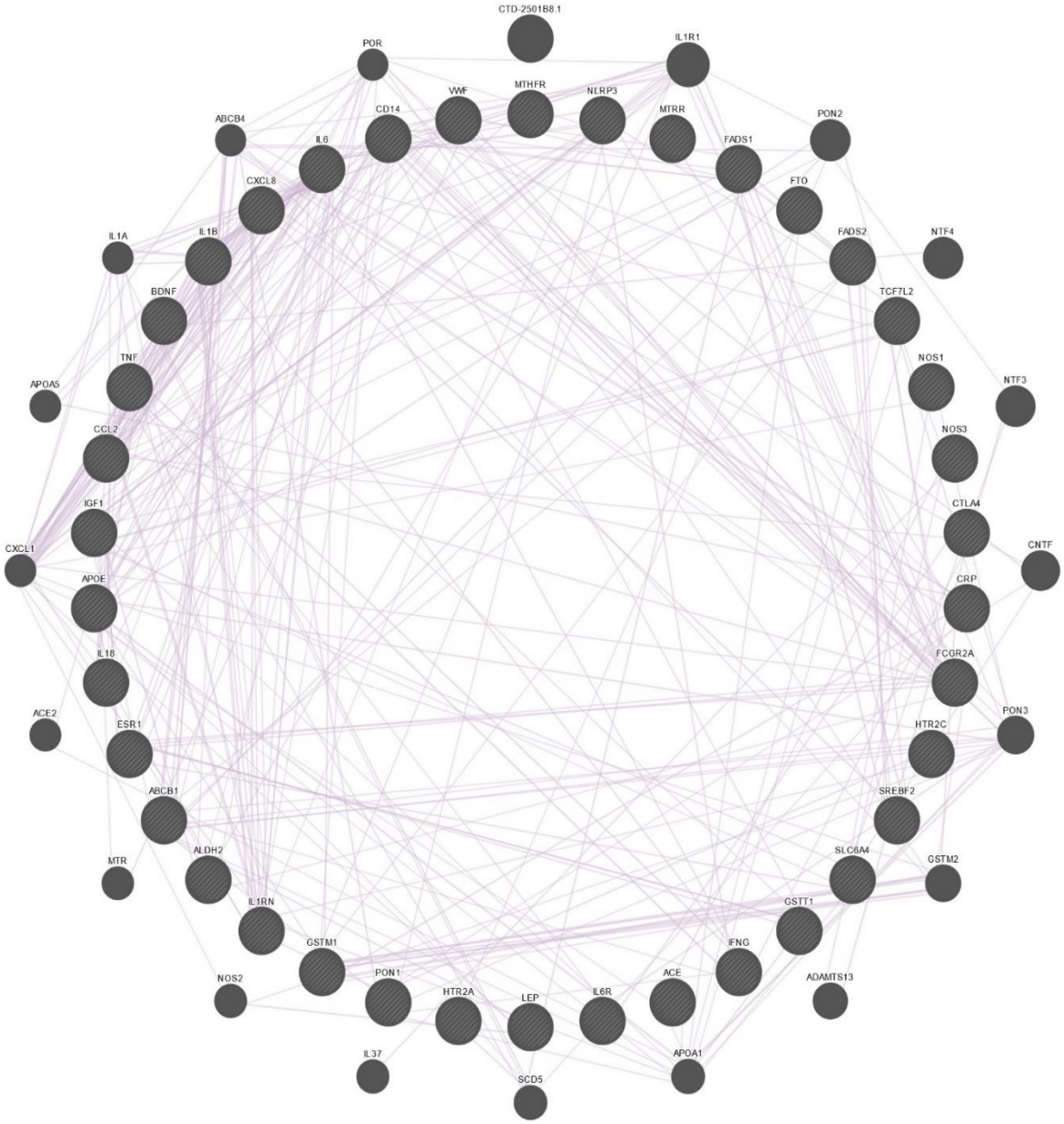


Figure 11: Co-expression between 38 responsible genes.

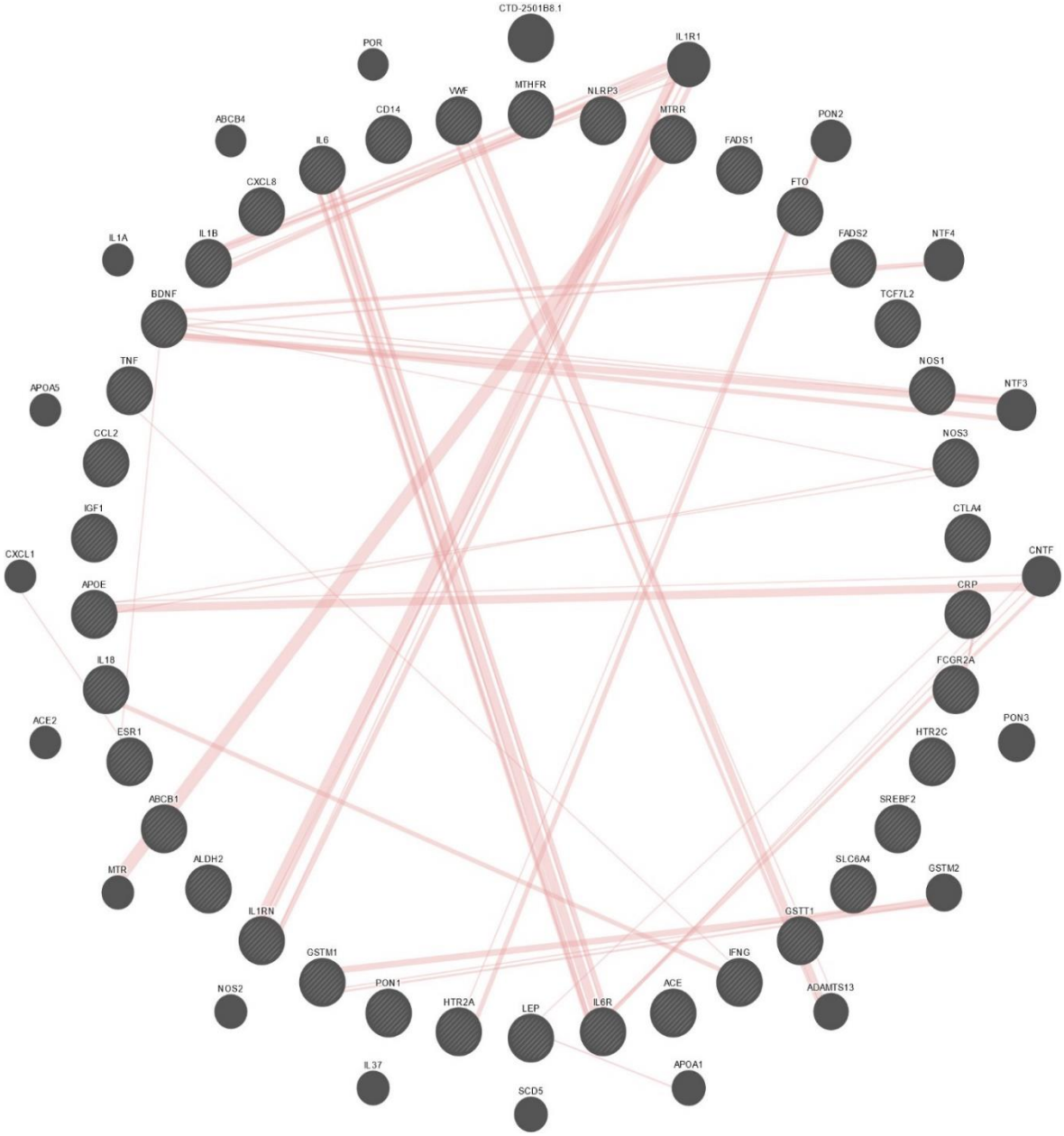


Figure 12: Physical interaction between 38 responsible genes.

4.6 Clustering

Cluster analysis provides insight into data by dividing objects into groups of objects, such that objects in clusters are more similar to each other than objects in clusters (25). In the area of bioinformatics, cluster analysis has long played a significant role. Usually, biological information or data is structured either as a sequence or as a network. In both key data patterns and rare sequences, clustering algorithms provide good insights.

4.6.1 K-means clustering

K-means clustering, one of the most established and most broadly utilized clustering algorithms (26). It is a prototype-based, simple partitioned clustering algorithm that tries to find K clusters that are not overlapping. Overall, K-means has been widely studied from both the optimization and data perspectives in a great deal of research. Figure 13 indicates the cluster of K-means.

4.6.2 MCL clustering

To date, the Markov Cluster (MCL) algorithm appears to be one of the most successful clustering procedures used to derive complexes from protein interaction networks (27). The goals of this algorithm are to simulate flow within a map, promote flow where the current is high, and demote flow where the current is weak. Figure 14 indicates the cluster of MCL.

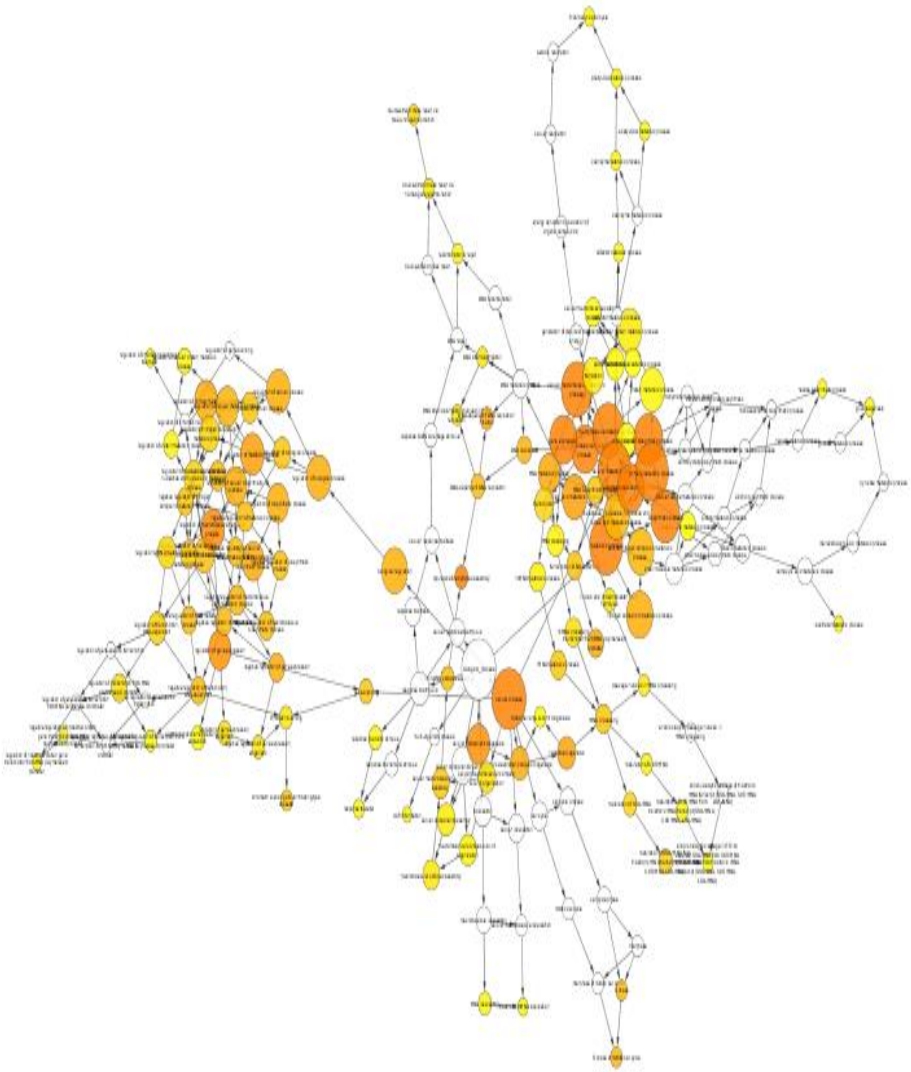


Figure 13: K-means Clustering using Cytoscape.

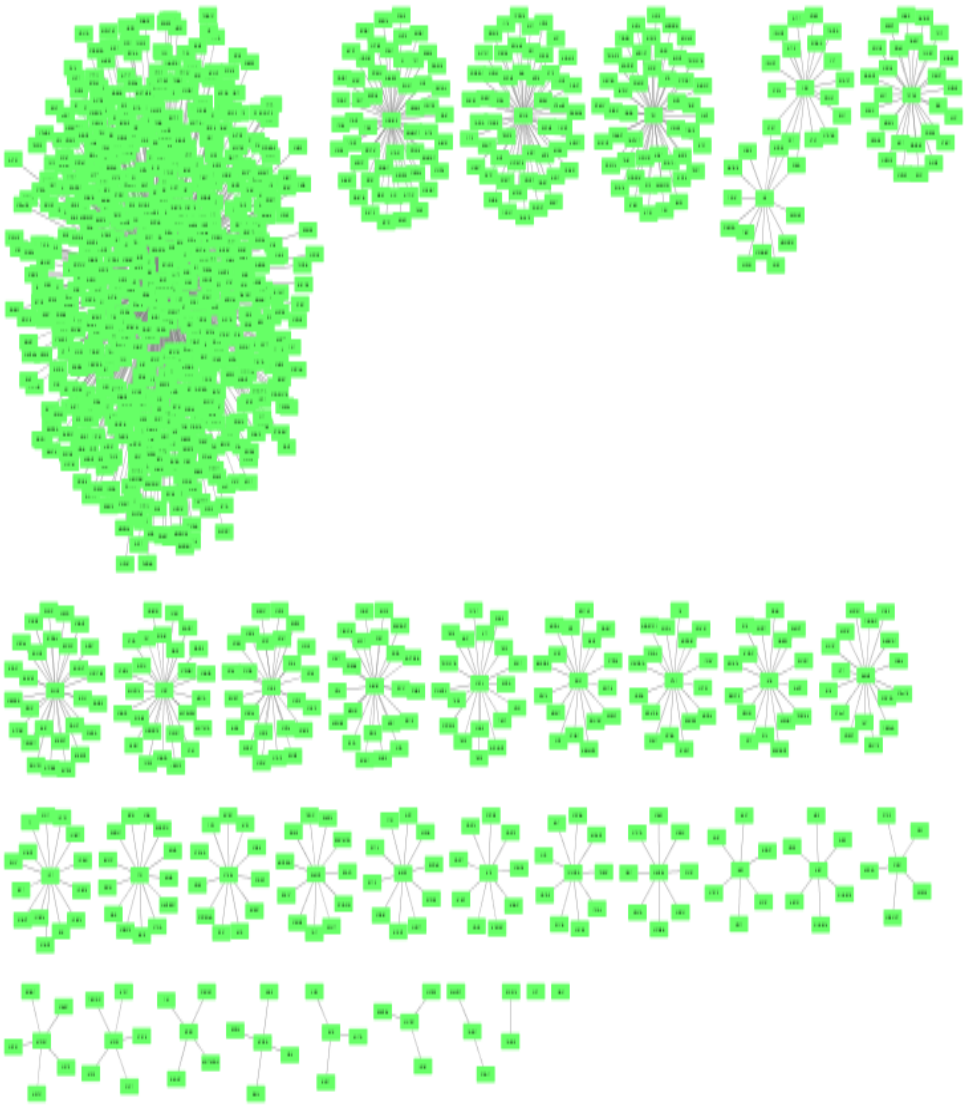


Figure 14: MCL Clustering using Cytoscape. Here the PPI network (Figure 4) became clustered. We get a total of 1234 links between 1267 proteins after clustering with MCL.

4.7 Gene regulatory network

A gene regulatory network or genetic regulatory network (GRN) is a collection of DNA segments in a cell that interact indirectly with each other and other cell sub-stances, thereby regulating the rates at which genes are transcribed into mRNA in the network (28). We used web-based NetworkAnalyst tools to define the gene regulatory network. There are three types of gene regulatory networks: Gene-miRNA interaction, TF-gene interaction, TF-miRNA co-regulatory network. Gene-miRNA interactions, TF-gene interaction, TF-miRNA co-regulatory network respectively are shown in Figures 15 to 17 respectively (29).

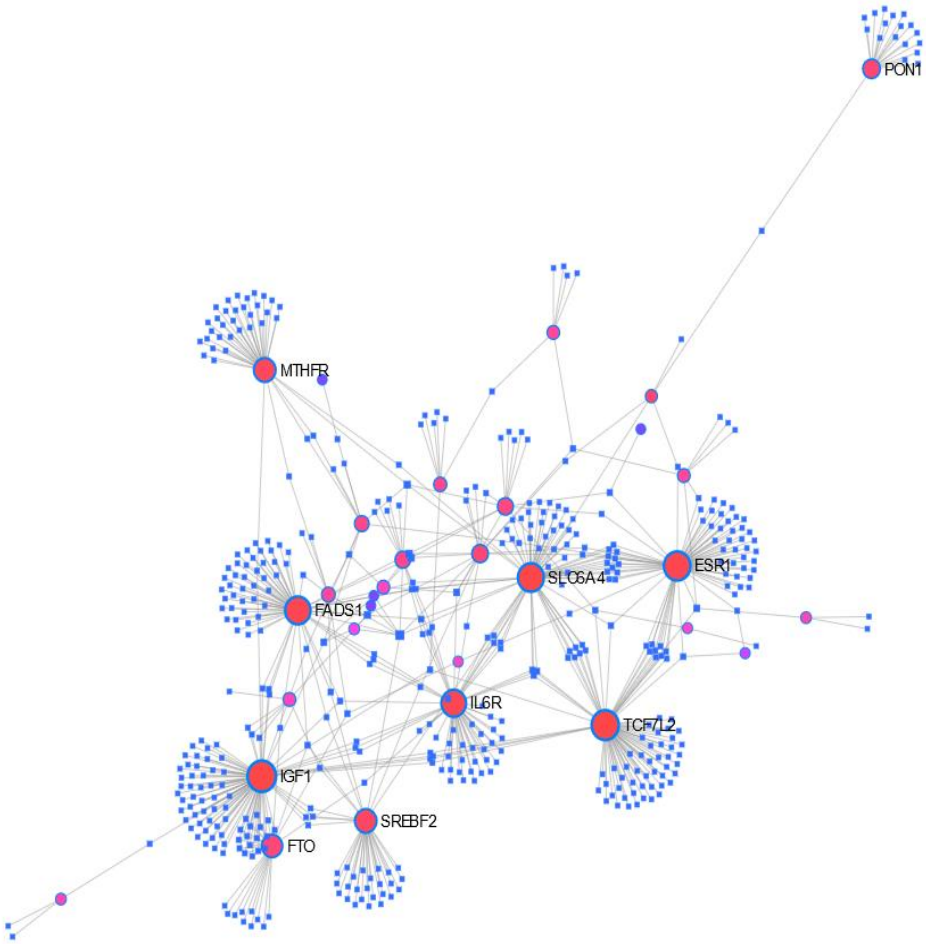


Figure 15: Gene -miRNA Interaction for selected 38 genes. This gene-miRNA interaction generates interactions with a total of 645 links between 507 genes.

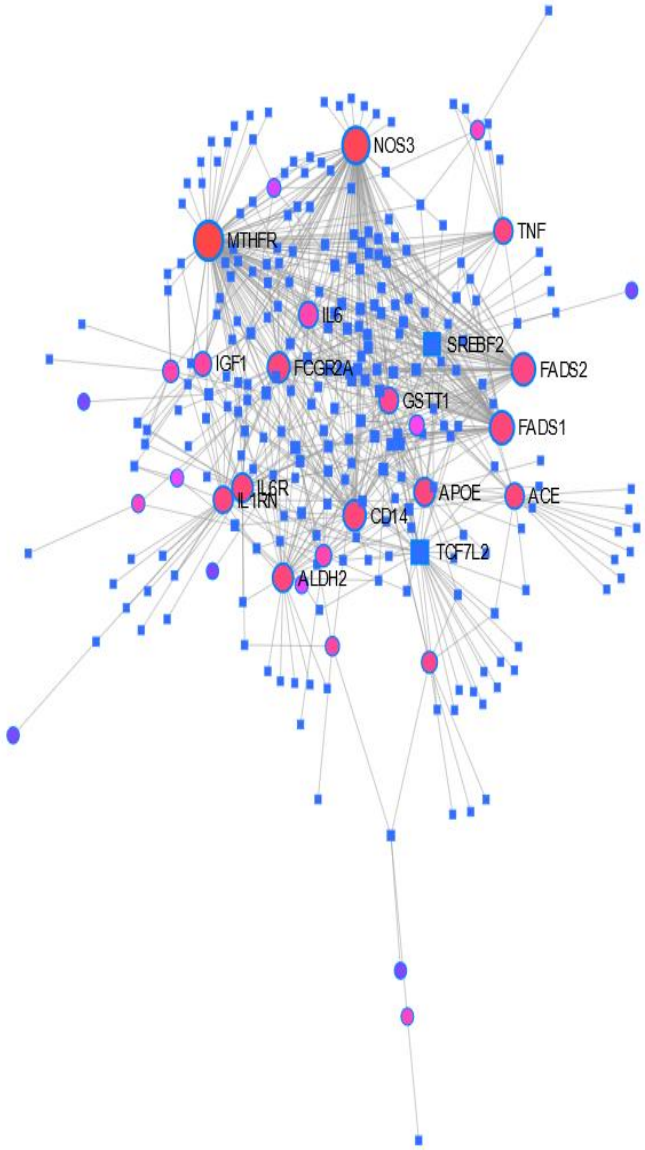


Figure 16: TF-gene Interaction for selected 38 genes. This TF-gene Interaction creates relationships between 277 proteins with a total of 694 connections.

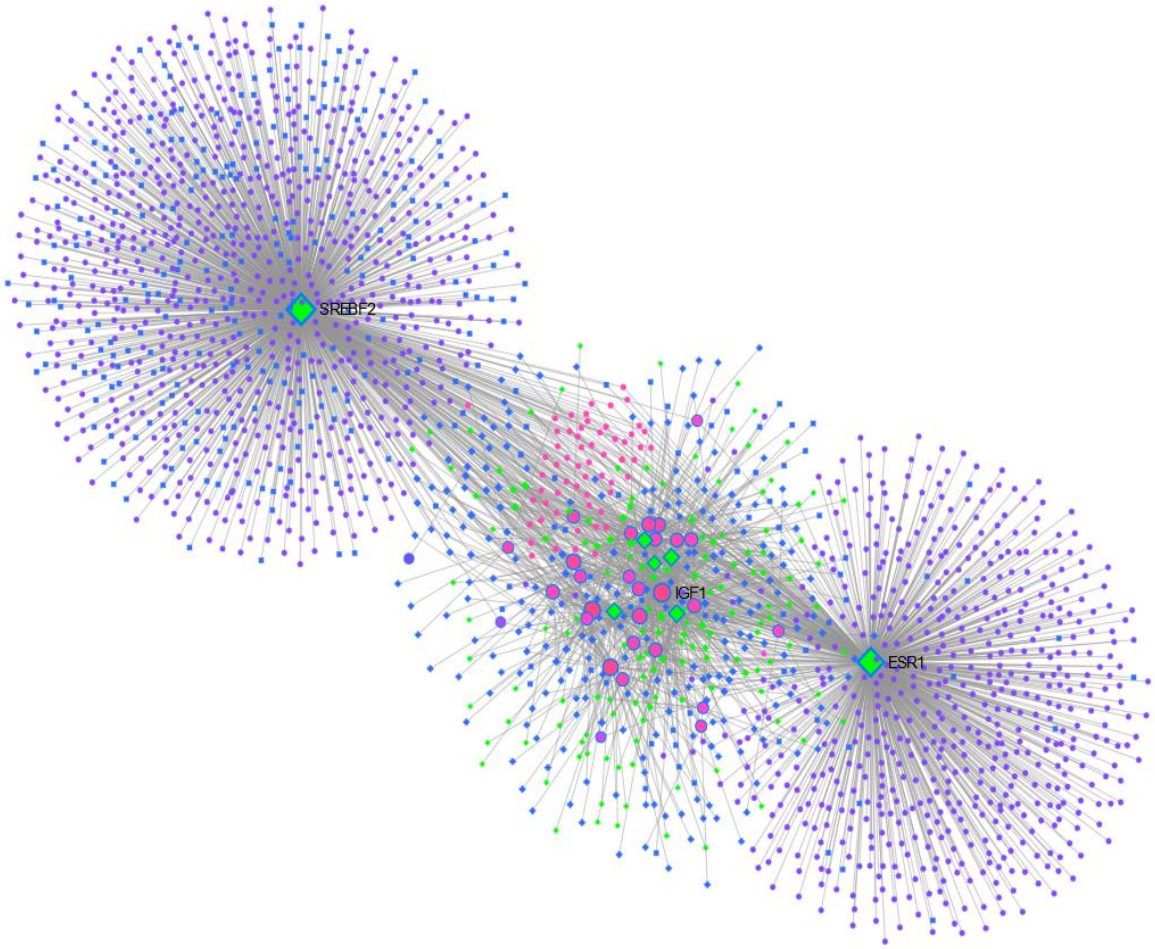
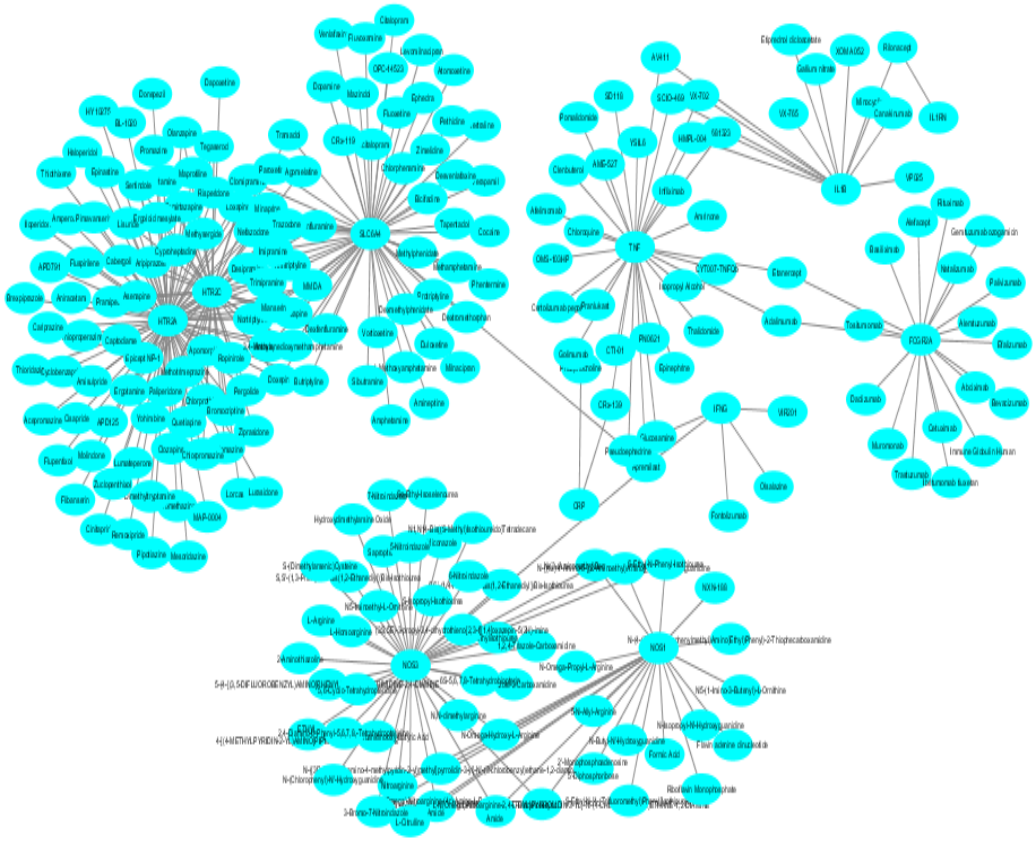


Figure 17: TF-miRNA co-regulatory network for selected 38 genes. This TF-miRNA co-regulatory network creates relationships between 2093 proteins with a total of 2698 connections.

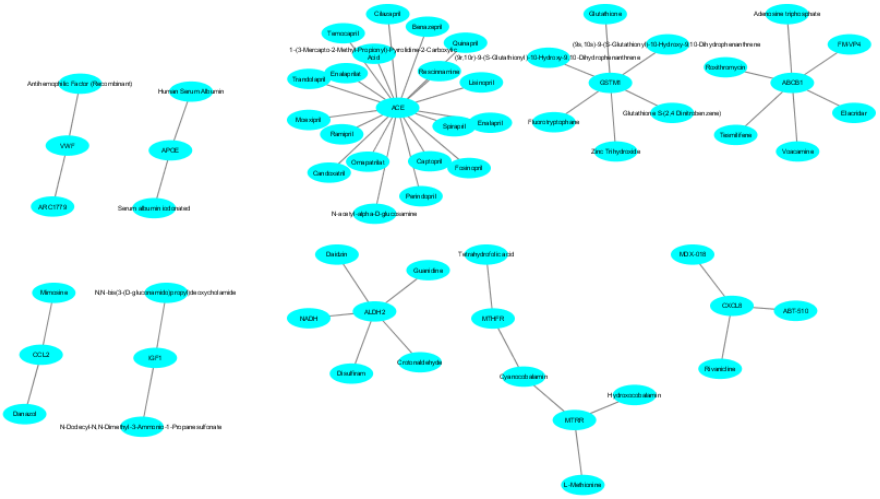
4.8 Protein-drug interaction

Protein-drug interactions are an integral part of the processes of intermolecular recognition and/or mediation that take place in a living organism's cells or tissues, including membrane transport phenomena (30). The NetworkAnalyst tool generated the PDI network. By using NetworkAnalyst tools, we get 11 sub networks for PDI for responsible 38 genes. After getting the subnetworks, we merged sub network (2-11) as

Figure 18(b) by using Cytoscape. Also, here is developed the subnetwork1 as Figure 18(a) using Cytoscape. The complete set of drugs that can be used for the above-selected disease is shown in Figure 18 PDI. The NetworkAnalyst tool generated the PDI network.



(a)



(b)

Figure 18: Protein-drug interaction for responsible 38 genes. (a) Represent, subnetwork1 which creates relationships between 245 proteins with a total of 325 connections. (b) Represent, a merged network that creates a relationship between 62 proteins with a total of 52 connections.

4.9 Protein-chemical interaction

Biological networks are now conducting multiple critical studies on human behavior and disease prevention (31). Throughout living organisms, interactions between proteins and small molecules are an integral part of biological processes. Figure 19 shows a protein-chemical interaction. The PCI is generated using the NetworkAnalyst tool.

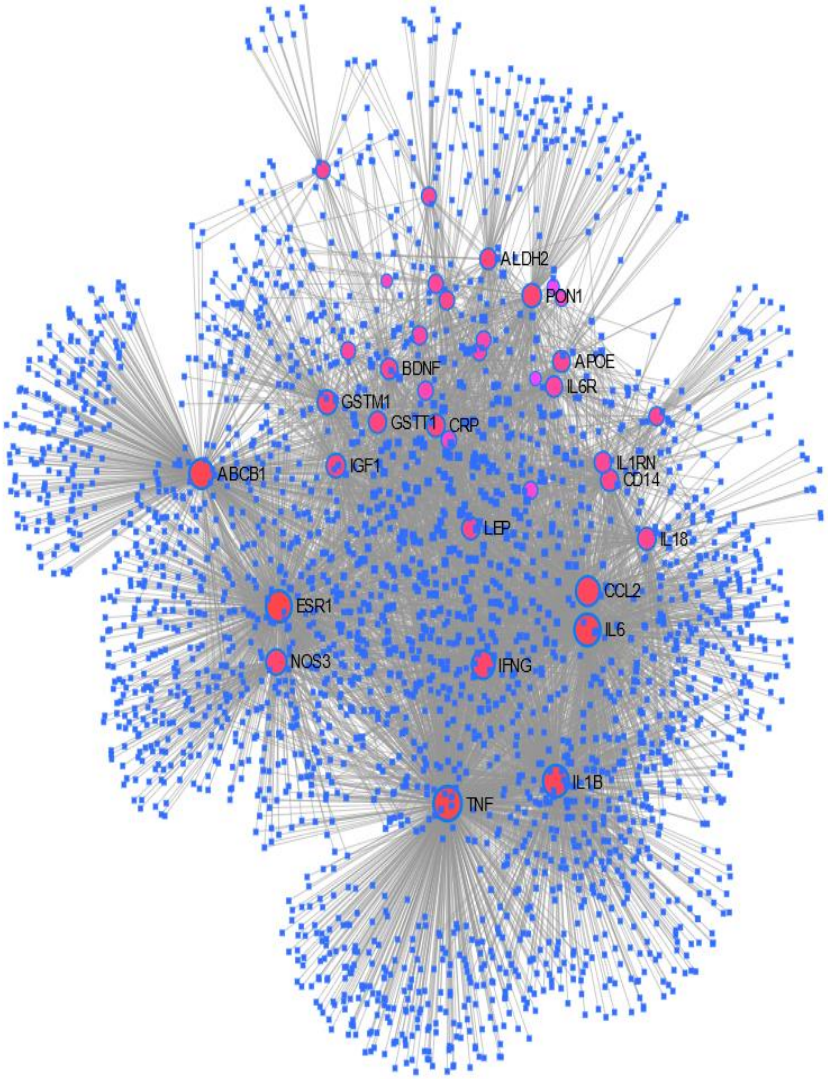


Figure 19: Protein-chemical interaction for selected 38 genes. This PCI creates relationships between 2574 proteins with a total of 6211 connections.

CHAPTER 5: CONCLUSIONS AND RECOMMENDATIONS

The previous parts are important for this portion to be brought on. This section outlines results alongside contributions and potential works depending on the broad execution and evaluation of going before sections in the following segments.

5.1 Findings and Contributions

From the above discussion, we can see that it can be directly or indirectly related to one another in bipolar disorder, stroke, coronary heart disease, and schizophrenia. They also have a genetic relationship with each other. Depending on the genetic relationship, they have common genes that are interrelated to each other. A gene regulatory network pathway will be maintained for common genes. This examination examines both the common genes and the common gene regulatory pathway between bipolar disorder and related diseases. The gradual advancements and the utilization of bioinformatics tools increase the potential for output. Bipolar disorder is a set of disabilities that may even result in death. A lot of people are dying every day because of bipolar disorder. This research analyzes bipolar disorders, including BD, CHD, SCH and ST. Each of the four genetic disorders is investigated in the present examination.

5.2 Recommendations for Future Works

In order to develop a drug for more than one disease, it is important to realize that the influenced genes are connected with these diseases. To reach the destination, it is also important to know the connection between the genes and the related diseases. In the light of this examination, the Gene Regulatory Networks shall indicate the interrelated gene between the diseases. The present investigation carried out all the analyses with the assistance of bio-informatics apparatuses which make further measurements in the use of bioinformatics tools in the field of bioinformatics. The researchers who want to work further, can work for the exploration is to build up a typical drug for the diseases of bipolar disorder.

REFERENCES

1. WHO. Promoting mental health: Concepts, emerging evidence, practice: Summary report. World Health Org. 2004. Last access: 10 December, 2019.

2. WHO. The World Health Report 2001: Mental health: new understanding, new hope. World Health Org. 2001. Last access: 10 December, 2019.

3. WHO. The world health report 2002: reducing risks, promoting healthy life. World Health Org. 2002. Last access: 10 December, 2019

4. Vos T, Allen C, Arora M, Barber RM, Bhutta ZA, Brown A, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet 2016; 388(10053): 1545-602.

5. Alonso J, Petukhova M, Vilagut G, Chatterji S, Heeringa S, Üstün TB, et al. Days out of role due to common physical and mental conditions: results from the WHO World Mental Health surveys. Mol Psychiatry 2011; 16(12):1234.

6. Murray CJ, Lopez AD, WHO. The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020: summary. World Health Org. 1996; Last access: 10 December, 2019.

7. Patel V, Prince M. Global mental health: a new global health field comes of age. JAMA 2010; 303(19): 1976-77.

8. Torrey EF, Bowler A. Geographical distribution of insanity in America: evidence for an urban factor. Schizophr Bull 1990; 16(4): 591-604.

9. Cardno AG, Owen MJ. Genetic relationships between schizophrenia, bipolar disorder, and schizoaffective disorder. Schizophr Bull 2014; 40(3): 504-15.

10. Johnson W, Onuma O, Owolabi M, Sachdev S. Stroke: a global response is needed. Bull World Health Org 2016; 94(9): 634.

11. Redon J, Olsen MH, Cooper RS, Zurriaga O, Martinez-Beneito MA, Laurent S, et al. Stroke mortality and trends from 1990 to 2006 in 39 countries from Europe and Central Asia: implications for control of high blood pressure. Eur Heart J 2011; 32(11):1424-31.

12. Feigin VL, Forouzanfar MH, Krishnamurthi R, Mensah GA, Connor M, Bennett DA, et al. Global and regional burden of stroke during 1990–2010: findings from the Global Burden of Disease Study 2010. Lancet 2014; 383(9913): 245-55.

13. Asplund K, Karvanen J, Giampaoli S, Jousilahti P, Niemelä M, Broda G, et al. Relative risks for stroke by age, sex, and population based on follow-up of 18 European populations in the MORGAM Project. Stroke 2009; 40(7): 2319-26.

14. Struijs JN, Van Genugten ML, Evers SM, Ament AJ, Baan CA, Van Den Bos GA. Future costs of stroke in the Netherlands: the impact of stroke services. Int J Tech. Assessment in Health Care 2006; 22(4): 518-24.

15. Wolfe CJ, Kohane IS, Butte AJ. Systematic survey reveals general applicability of guilt-by-association within gene coexpression networks. BMC Bioinformatics 2005; 6(1): 227.

16. Amar D, Safer H, Shamir R. Dissection of regulatory networks that are altered in disease via; differential co-expression. PLoS Comput Biol 2013, 9(3): e1002955.

17. Likas A, Vlassis N, Verbeek JJ. The global k-means clustering algorithm. Pattern Recognition 2003; 36(2): 451-61.

18. Shih YK, Parthasarathy S. Identifying functional modules in interaction networks through overlapping Markov clustering. Bioinformatics 2012; 28(18): i473-9.

19. Canduri F, de Azevedo J, Walter F. Protein crystallography in drug discovery. Curr Drug Targets 2008; 9(12): 1048-53.

20. Oravcová, J, Lindner W. Protein–Drug Interactions. Encyclopedia of Analytical Chemistry: Applications, Theory and Instrumentation 2006; 1-26. doi:10.1002/9780470027318.a1627.

21. Kuhn M, Szklarczyk D, Pletscher-Frankild S, Blicher TH, Von Mering C, Jensen LJ, et al. STITCH 4: integration of protein–chemical interactions with user data. Nucleic Acids Res 2013; 42(D1): D401-7.

22. Zhou G, Soufan O, Ewald J, Hancock RE, Basu N, Xia J. NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. Nucleic Acids Res 2019; 47(W1): W234-41.

23. Frishman D, Valencia A. Modern genome annotation. The Bio Sapiens Network 2009; 213-38.

24. Hasan MR, Paul BK, Ahmed K, Bhuyian T. Design protein-protein interaction network and protein-drug interaction network for common cancer diseases: A bioinformatics approach. Inform Med Unlocked 2020; 18: 100311.

25. Jain AK, Dubes RC. Algorithms for clustering data. Englewood Cliffs: Prentice Hall, 1988. ISBN: 978-0-13-022278-7.

26. Wu J. Advances in K-means clustering: a data mining thinking. Springer Science & Business Media 2012;1-177.

27. Van Dongen S. Graph Clustering by Flow Simulation. In PhD Thesis University of Utrecht; 2000.

28. Vijesh N, Chakrabarti SK, Sreekumar J. Modeling of gene regulatory networks: a review. J Biomed Sci Eng 2013, 6(02): 223.

29. Szklarczyk D, Santos A, von Mering C, Jensen LJ, Bork P, Kuhn M. STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data. Nucleic Acids Res 2015; 44(D1): D380-D384.

30. Hasan MR, Paul BK, Ahmed K, Mahmud S, Dutta M, Hosen MS, et al. Computational analysis of network model based relationship of mental disorder with depression. Biointerface Res Appl Chem 2020; 10(5): 6293-305.

31. Kuhn M, Szklarczyk D, Franceschini A, Von Mering C, Jensen LJ, Bork P. STITCH 3: zooming in on protein–chemical interactions. Nucleic Acids Res 2011; 40(D1): D876-80.

