



Thesis- ANALYZING THE PROTEIN-PROTEIN INTERACTION NETWORK AND THE TOPOLOGICAL PROPERTIES OF CORONARY ARTERY DISEASE AND ALLIED DISEASES: A COMPUTATIONAL BIOINFORMATICS APPROACH.

Supervised By:

MD. Rajib Mia

Lecturer

Department of Software Engineering

Daffodil International University

Submitted By:

MD. Fazly Rabbi

ID: 171-35-192

Department of SWE

Daffodil International University

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**

APPROVAL

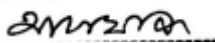
This thesis is titled “Analyzing the protein-protein interaction network and the topological properties of prostate cancer and allied diseases: A computational bioinformatics approach”, submitted by **MD. Fazly Rabbi, ID: 171-35-192** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering.

BOARD OF EXAMINERS



MD. Imran Mahmud
Head and Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology Daffodil
International University

Chairman



Afsana Begum
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 1

Fazla Elahe

Internal Examiner 2

DR. MD Fazly Elahe

Assistant Professor

Department of Software Engineering

Faculty of Science and Information Technology

Daffodil International University

Mohammad Abu Yousuf

Mohammad Abu Yousuf, PHD

Professor

Institute of Information Technology

Jahangirnagar University

External Examiner

DECLARATION

It hereby declares that this thesis has been done by us under the supervision of **MD. Rajib Mia, Lecturer**, Department of Software Engineering, Daffodil International University. It is also declared that neither this thesis nor any part of this has been submitted elsewhere for award of any degree.

Rabbi

Name: MD. Fazly Rabbi

Student ID: 171-35-192

Batch: 22

Department of Software Engineering

Faculty of Science & Information Technology

Daffodil International University

Certified by:



MD. Rajib Mia

Lecturer

Department of Software Engineering

Faculty of Science & Information Technology

ACKNOWLEDGEMENT

We had started this thesis with zero knowledge. We had to discover a whole new area to work on this thesis. Because of having a technological background, understanding the thoughts and theories of Biology and Biotechnology was pretty tough for me. But I became able to jump forward smoothly with the help of our respectable teacher Md, Rajib Mia . He supervised me on every step. He motivated me to learn new things, provided all the important files and information and taught me how to make the work done. I Am grateful to Almighty Allah forgiving me the ability to complete the final thesis. I also want to thank our respectable Dr. Imran Mahmud Sir , Kawshik sir and Matiur Rahman Sir. They were always cordial to us, helped us and motivated me. I am always thankful to them. I would like to express my gratitude towards our seniors, classmates and juniors for the time and support they have given to me with all their capabilities. I am grateful to all of them. Thank you so much.

TABLE OF CONTENT

CONTENTS

APPROVAL	ii
DECLARATION	iv
ACKNOWLEDGEMENT	v
TABLE OF CONTENT	vi
LIST OF TABLE	viii
LIST OF FIGURE	viii
ABSTRACT	ix
CHAPTER 1: INTRODUCTION	1
1.1 Background	3
Chapter 2: Literature Review	4
CHAPTER 3: METHODOLOGY	5
3.1 Flow chart	5
3.2 Gene Collection	6
3.3 Venn Diagram	6
3.4 Generic PPI	6
3.5 Topological Properties (TP)	6
3.6 Co-expression	7
3.7 Gene Regulatory Network	7
3.8 drug design	
3.8.1 Protein-drug Interaction	8
3.8.2 Protein-chemical Interaction	8
3.8.3 Gene diseases association	8
3.9 Clustering	
3.9.1 MCL clustering	8

CHAPTER 4: RESULTS AND DISCUSSION	9
4.1 Gene Collection	9
4.2 Gene mining:	10
4.2.1 Linkage	10
4.2.2 Common gene finding	10
4.3 Venn Analysis	12
4.4 Generic PPI	13
4.5 Topological Properties	14-18
4.6.1 Co-Expression	19
4.6.2 Physical Interaction	20
4.6.3 Pathway Interaction	21
4.7 Gene Regulatory Network	22-24
4.8.1 Protein-drug Interaction	25-27
4.8.2 Protein-Chemical Interaction	28
4.8.3 Gene-disease Association	29
4.9 Clustering	30
4.10 Summary	31
CHAPTER 5: CONCLUSIONS AND RECOMMENDATIONS	32
REFERENCES	33-35

LIST OF TABLES

Table 1: The numbers of respective genes.....	9
Table 2: number of common genes between selected four diseases.....	11
Table 3: The topological properties for top 10 selected genes.....	14

LIST OF FIGURES

Figure 1: Methodology of Flow chart	5
Figure 2: Venn diagram	12
Figure 3: PPI Network	13
Figure 4: Topological Properties (Closeness)	15
Figure 5: Topological Properties (Co-efficient)	16
Figure 6: Topological Properties (Betweenness)	17
Figure 7: Topological Properties (Topological co-efficient)	18
Figure 8: Co-expression	19
Figure 9: physical interaction	20
Figure 10: pathway analysis	21
Figure 11: Gene- miRNA	22
Figure 12: TF-gene	23
Figure 13 : TF-miRNA	24
Figure 14: protein drug interaction (Sub network-1)	25
Figure 15: protein drug interaction (Sub network-2)	26
Figure 16: protein drug interaction (Sub network-3)	27
Figure 17: protein chemical interaction	28
Figure 18: protein diseases association	29
Figure 19 : MCL clustering	30

ABSTRACT

Background and Objectives: Some diseases are related to each other by their metabolic structures. A few examinations showed that Coronary artery disease (CAD), Diabetes mellitus (DM), Parkinson's Disease (PKD) and Stroke (ST) are related. Some are shown up for affected family backgrounds in their early or grown-up age.

Materials and Methods: Python a programming language used for data mining, pre-processing and sorting for finding common genes from gathered data from National Centre of Biotechnology Information (NCBI). However, there is a lack of genetic study to find out the core genes for which they may occur and make Protein-Protein Interaction (PPIs) and Protein Disease Interaction (PDI) by using bioinformatics technology. We use identified hub genes for making co-expression and physical interaction.

Results: Interactions for selected top 10 genes are exhibited following different bioinformatics tools. The gene-miRNA interaction generates interactions with a total of 413 links between 10 genes. Where, the TF-gene Interaction creates relationships between 101 nodes and 106 edges. There are 5 seed nodes. Besides, PDI represents a subnetwork which creates 3 sub network relationships between 58 nodes and 55 edges. There are 1 seed node in each sub network. In addition, PCI creates relationships between 1842 nodes and 2685 edges. There are 8 seed nodes. Furthermore, GDA creates relationships between 452 nodes and 537 edges. There are 6 seed nodes.

Conclusion: This study will be helpful for further studies of different bioinformatics tools for designing gene network models and drugs design. These drugs can be considered for further verification by chemical experiments.

Keyword:

PPI; Computational Bioinformatics; Coronary artery disease (CAD), Diabetes mellitus (DM),; Parkinson's Disease (PKD) ; Stroke (ST).

CHAPTER 1

INTRODUCTION

1.1 Background

From many other death investigations it has shown that heart diseases placed on top for death in the world. In every year for this heart diseases in the age group of 25-69 years have to face 25 percent of death. On the off chance that all age bunches are in-corporated from overall death reasons heart diseases are responsible for 19 percent of it [1]. For man as well as female HD is the main source for death. It is additionally the main source of death in all areas however the numbers fluctuate. [1] A bunch of conditions that influence your heart, for example, heart rhythm issues (arrhythmias), Coronary Artery Diseases (CAD), by born heart problem or Congenital heart defects (CHD) and among others is called Heart Disease (HD) or Cardiovascular Disease (CVD). In the United Kingdom, United States, Canada, and Australia, HD is the leading reason for death as indicated by the Centers for Disease Control (CDC). One in each four passes in the U.S. happens because of HD. [2] In the US regularly for CVD one person faces death every 37 sec. The World Health Organization (who) says that CVDs or HDs are taking an expected 17.9 million lives every year and an expected 31% of all passings around the world. 75% of CVD passouts happen in low-and center pay nations. There are some terms and conditions in our life that can increase the risk for HDs like the way you live your life, age of yours and previous history of your family related to HDs. One of three key hazard factors for HDs are shown in half of all Americans (47%). The risk hazards are hypertension, elevated cholesterol, and smoking. Some chance elements for HD can't be controlled, for example, age of yours or history of your family. Be that as it may, you can find a way to bring down your hazard by changing the elements you can control. As WHO, the age-balanced demise rate owing to CVD, in light of 2017 information, is 219.4 per 100,000 and there are 2,353 passings from CVD every day.

Coronary artery disease (CAD) is popular with many names like Coronary Heart Disease (CHD) or Ischemic Heart Disease (IHD). It is the most well-known kind of HD worldwide.[3] The arteries which supply blood to the heart and different pieces of the body are called coronary arteries and the plaque development in these artery walls are called CAD. Plaque is made of different kinds of things like fat or CA and many other things which are found in blood. Atherosclerosis or CAD is a process where Plaque development causes within the veins to limit after some time, which can halfway or absolutely block the flow of blood. The CAD which is untreated can prompt chest torment, CVD breakdown, and arrhythmias. In the US every year 735,000 Heart Attacks take place for CADs and also kills in excess of 630,000 Americans every year. More than 7 million Americans have to face a Heart Attack in their life, it is an indicated result by the American Heart Association (AHA). [4] CAD is the main source for death that occurs by cardiovascular disease around the world, in recent decades 4.5 million deaths are found . [5] Nonetheless, it is foreseen that CAD passing rates will twofold from 1990 to 2020, with generally 82% of the development inferable from the ongoing lifetime. [5]

Diabetes mellitus (DM) is the 6th driving reason for death in the United States, and diabetes-related problem like kidney sickness, retinopathy are causing a tremendous weight to the national health care system.[6] DM is a metabolic issue described by impeded activity, emission of insulin or both, that leads to hyperglycemia. [7] It is moderately evaluated that 100 million people have DM on this planet. [8] In the latest time frame 6.4% of the passings are liable for DM.[8] As indicated by the IDF Atlas rule report, at present, disability in glucose resistance have been found in 352 million people around the world who are at high risk for suffering from DM in future. In 2017, it was evaluated that 425 million individuals (20–79 years old) experienced DM, and it is expected that the number will be increased to 629 million by 2045.[9] It is assessed that 77% of the worldwide weight of the DM pestilence have to face by developing nations in the century of 21st [5] They have to face this because of their populace development, utilization of undesirable eating system, heftiness, and inactive ways of life. [9] Two types of DM are invented. In both type 1 DM and type 2 DM, CAD is the main reason for unexpected death and double to fourfold expanded mortality chance from HD is associated with DM. [10] HD and ST are the reason for death in 70% people who are more than 65 years old and suffering from DM. Furthermore, those patient who are suffering from DM have an increased rate for death after doing MI and more terrible risk for those, who are in long-term prognosis with CAD.[10] In the US from all percutaneous coronary intervention (PCI) techniques, around 33% are applied on the patients who are suffering with DM and roughly 25% of patients experiencing coronary artery bypass graft (CABG) medical procedure who have DM and the results of these strategies is less effective than in those who are without DM.[10]

Parkinson's Disease (PKD) is a disorder which causes degeneration in the nervous system and 2nd most common in all nervous system disorders.[11] It progressively affects adult American is PKD and a predicted result shows that it will be continuously increased in American adults. [12] PKD, is described by the cardinal highlights of rest shaking, slowness of movement, inflexibility of the muscles and making spine postural in an unnatural positions, making loss in programmed movements, making changes in speeches and an assortment of other symptoms which are related to movements and which are not related to movement.[13] It influences in making movements. These types of diseases do not affect at a glance they start in a particular way, some of the time it begins with tremor in only one hand. In a positive perspective these tremors are normal, yet the turmoil likewise regularly caused PKD. One million peoples of US have PKD reported by PKD Foundation in the recent time. In each 100,000 cases there have roughly 20 people who are suffering from PKD and 60,000 cases in every single year have found in US. ALL of them are close to 60 years old or above. It's bad effects are seen in 1% people who are 60 years old or way to 60 and more badly about 1%-3% in 80 years olds or above them. [11] The ratio for having CVD who are suffering from PKD is 60% [14] Recent examination indicated that those who have more strolling and memory problems who are suffering from PKD and have the risk of having CVD. Sometimes, people with CVD who are in early stage in it also face terrible strolling and memory issue. Both CVD and PKD become progressively basic as individuals when they get aged.

Stroke (ST) is one of the most common diseases in new are. Blood streaming to the brain is responsible for cell death and this situation is called stroke. There are two primary sorts of stroke: one is ischemic, because of blocking arteries of blood in the brain, and hemorrhagic, because of bleeding which is caused by ruptures of weakened blood vessels. Both reasons make the brain quit and stop working appropriately. ST can be the reason for death. As per the American Heart Association (AHA) in 2017, there 37.6% mortality possibility found in each 100,000 cases in the age-balanced death rate for ST analysis. Internationally it shows that in the middle-class country, 70% of ST and 87% of both stroke-related passings take place [15] During these decades ST frequency has declined by 42% in high-income nations. [15] From the perspective of WHO, more than 17.9 million people have to face death for CVDs every year. Of these passings, 80% are because of CAD and ST and generally influence in low and middle-income nations, Patients who are suffering from CAD have a higher danger of ST (14%) when contrasted with those with no proof of CAD is (0.9%) [16]

From this above discussion we can say that DM, PKD and ST have an interconnection with CVD. They are dependent on each other or every patient with one of these diseases has a high risk for being affected with them too.

CHAPTER 2: LITERATURE REVIEW

For the purpose of drug invention many researchers have done work on those topics in previous times.

In a research paper named “ Construction of the gene expression subgroups of patients with coronary artery disease through bioinformatics approach” have shown that they found an outline for gene groupings in CAD patients, analyzed the differences between each subgroup and annotated the unique genes of each group.

A research paper named “ A comprehensive bioinformatic analysis revealed novel MicroRNA biomarkers of Parkinson's disease. ” has shown that their finding will help researchers shed light on the discovery of novel biomarkers for PkD.

In Integrated computational approaches to screen gene expression data to determine key genes and therapeutic targets for type-2 diabetes mellitus, research paper has shown that, they Find out that hub genes those cause disruption in cellular pathways which deeply worsens the disease condition.

Another research paper named Study on potential differentially expressed genes in stroke by bioinformatics analysis. Has shown that by using hub gene they show Differentially expressed genes and MicroRNAs.

CHAPTER 3

RESEARCH METHODOLOGY

A list of techniques or strategies which are used only for one specific result in studies or work are called methodology. Methodology in research is the way by which researchers need to lead their exploration. Methodology plays an important role in describing their problem or process and target goal by presenting the visualization strategy for getting the expected result. This structure for the analysis and the procedure shows how it will work for any research and what result would show after finishing these processes. Examination and the technique which are used during the research procedure are talked about in this part. [17] After collecting common data from these four diseases they used in Venny for making venn diagrams, used in NetworkAnalyst to make PPI networks, in Cytoscape for finding topological properties, in GeneMania for further progress. The visualization of these process are shown by a flowchart given in figure.1

3.1 Flow chart:

In figure 1, it represents the methodology step by step. It showed the graphical view of further works.

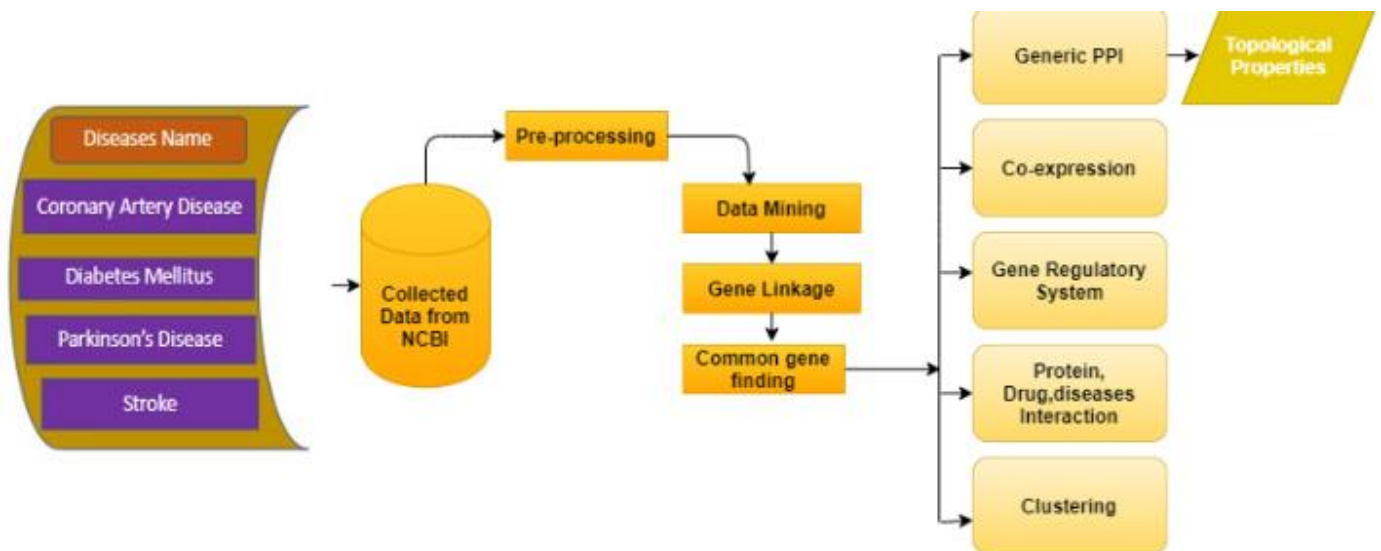


Figure 1: Flowchart for proposed research methodology

3.2 Gene Collection:

Collecting genes in a proper and appropriate way is an important part in biological studies. Expected outcomes depend on this part. So, a significant asset for bioinformatics devices and in biotechnology studies are databases that can be found in The National Center for Biotechnology Information in short NCBI. It is the most well-known platform where everybody can download genes for free.[19] All gene data for CAD, DM, PKD and ST are collected from this site.

3.3 Venn diagram:

Venn graphs are generally used to show list correlation. In science, they are broadly used to show the contrasts between gene records originating from various differential investigations, for instance.[20] A set which is pictorially like circles within an encasing rectangle called universal set and the common data of these sets are shown by the intersection of these circle are called venn diagram. In a venn diagram each circle represents a specific set and they are usually in circles shape with many overlapping closed curves. We used Venny for making appropriate venn diagrams for this analysis.

3.4 Generic PPI:

In such complex type of diseases, their have shown weakness for single gene studies with traditional methodology.[3] The investigation of interactomes, or systems of PPI, is progressively giving significant logic or information on biological studies.[21] PPI Network represent the contacts in physically between two or more proteins with a mathematical view. These interactions are explicit , happen between characterized restricting districts in the proteins and have a specific organic importance (i.e, they serve a particular function). There is expanding enthusiasm for these networks, as their examination assists with understanding the connection between the assets like genes or proteins and how these are situated in the entire network. [22] PPI network for analysing diseases helps to identify the genes and proteins which are associated with these specific diseases. It helps to study about the properties of that network and help to find the subnetwork also.[23] Global comprehension of PPI networks by means analysis of high-throughput gene information from various levels will permit specialists to analyze the pathways for diseases and recognize procedures to control them. Therefore, it appears to be likely that progressively customized, increasingly exact and increasingly fast ailment gene indicative methods will be concocted later on, just as novel methodologies that are increasingly customized. The PPI network for CRD, DM, PKD and ST diseases are represented by using NetworkAnalyst.

3.5 Topological Properties (TP) :

TP of PPI network is a structural representation of proteins properties. TP are divided in some properties like closeness centrality, betweenness centrality, node-degree distribution, avg.clustering , stress centrality distribution, shortest path length distribution, stress centrality distribution and more. We used cytoscape for the result which has 9 features . We have used 4 of them in this study.

3.6 Co-Expression , Physical Interaction & Pathway analysis:

A key target in biological examination is to deliberately distinguish all atoms inside a living cell and how they cooperate. [24] Co-expression speaks to the initial step of deduction that characterizes a connection between sets of transcripts. In the initial step for deduction, in which it characterizes between the set of transcript for connection is called co-expression. [24] It relies upon the possibility that transcript profiles of time plan, or result of unequivocal disturbances, may be normal for components and differences between transcripts, deriving their rule. [25]. This method was invented to find out the physically interacted gene, in an active subnetwork for finding different expressions when they are compared to two or more than condition or correlated expressions . Biological network are made of pathways which are interconnected in a series. In this recent era biological pathway and several kinds of networks have become one of the leading technique for data visualization and significant analysis.[27]

3.7 Gene Regulatory Network:

Gene regulatory networks (GRNs) assume an important part in different cell procedures. It helps to shorten the process for genes encoded interaction and the system of regulatory genes that decides the hereditary capacities to be communicated in cells of each spatial space in the life form, at each phase of advancement. Transcription factors for gene encoding are also done by this.[28] It intends to catch the conditions between atomic substances and is frequently displayed as a network , in that the representative for genes or protein are nodes and the representative for the interaction between genes or protein-DNA are edges.[29] A GRN is an assortment of regulatory connections between TFs and TF-restricting destinations of explicit mRNA to oversee certain articulation levels of mRNA and their came about proteins.[30] It is significant being developed, separation and reacting to natural prompts.

3.8 Drug Design:

Drug design is an achievement for innovation in a technological way. It is a combination for experiment in sophisticated way and computational methodologies. For creating new drug design , medical chemistry and biological innovation are making great progress.[31] The process is kept on developing, and applications presently range the entire medication disclosure process. There are many biological applications for doing network based work in human diseases. For the characterization of physical interaction or function in genes or proteins there have many constructed biological networks. [32] Networks give a framework level comprehension of the components hidden maladies by filling in as a structure level design for data reconciliation and investigation. They are utilized to pick up understanding into ailment mechanisms. [23]

3.8.1 Protein-drug Interaction:

For better knowledge about polypharmacology, in recent era pharmaceutical analysis has become a challenging part for characterizing Protein-drug interaction (PDR) networks. Protein restricting associations are uprooting responses, which have been embroiled as the causative instruments in many DDI. [33] It is imperative to more readily see how medications communicate with their protein and focus on their local condition.

3.8.2 Protein-chemical Interaction:

Protein-chemical interaction (PCI) is the main subject of target distinguishing proof and drug disclosure. [34] Biological systems are useful assets for foreseeing undocumented connections between genes or molecules. Study of PCI is a significant point toward clarification of protein capacities, comprehension of subatomic components in the cell and repositioning for the medication. [34] So, PCI network is a useful analysis for invention of new drugs.

3.8.3 Gene diseases association

For describing different human diseases with a complex dynamic modeling and studying about it with biological properties are become popular in recent years.[23] GDA was created to ascertain the affiliation score of an inquiry gene to another conceivable arrangement of diseases.[35] First, a huge scope PPI network was developed, connection within two cooperating proteins was determined with respect to the ailment interconnection. [35]

3.9 Clustering:

Where similar or most of them are alike, protein makes groups based on their characteristics are called protein clusters. The process for clustering must have a specific degree of dependability, robustness and have to permit the compression of data in contrast with the data which are non-clustered . It is acceptable that cluster which are consist of orthologs are important for better knowing about the rules of genome structure and knowing about gene or protein interaction, while paralogs are important to find out orthologs for transferring functional data between genes in different organisms with a high degree of reliability are remain in different cluster. Be that as it may, the ortholog-paralog differentiation doesn't totally mirror the multifaceted nature of group connections of homologous genes. [36] There are many kinds of clustering like MCODE cluster, MCL cluster, k-means clustering etc.

3.9.1 MCL clustering

The short form of Markov Cluster Algorithm is MCL . Named Stijn van Dongen at the Center for Mathematics and Computer Science (CWI) placed in the Netherlands was invented this clustering. Based on simulation of flow, it is a quick and adaptable unaided clustering calculation for networking. It calculated the random walk in a graph by using “Markov Chains” . [37] It helps to find out the functional modules in the PPI network.

Chapter :4

Result and discussion

4.1 Gene Collection:

Identifying genes which are responsible for the specific diseases are getting a very tricky day by day. To solve this problem we collected our respected gene data from NCBI , which is a free online based website with many desirable functions. Gene data was downloaded from NCBI as a text file. Gene data for Homo Sapiens was 831 out of 1352 for CAD, 836 out of 1248 for DM, 548 out of 854 for PKD and 833 out of 1444 for ST . [Table.1](#) shows these numerical numbers in an organized way.

Table 1:

In the below table, it shows the gene numbers that are collected from NCBI.

Diseases Name	Total no. of Gene	Total no. of Homo Sapiens Gene
CAD	1905	979
DM	2516	1046
PKD	1591	779
ST	1749	946

4.2 Gene mining:

Data mining is an important part for using data in an appropriate way for further use or applications. The text file which was collected from NCBI was filled with information for various purposes like Org_name, GeneID, CurrentID, Status, Symbol, Aliases, description etc. But In this research we need only gene Symbol . So we collect the symbol of genes from this file and save it in four xlsx sheets as we select four diseases for the analysis.

4.2.1 Linkage:

At the point when genes are found on various chromosomes or far separated on a similar chromosome, they group freely and are supposed to be unlinked. At the point when genes are near one another on a similar chromosome, they are supposed to be connected or linkage. This progression is to recognize the interrelated genes among CAD & DM, CAD & PKD, CAD & ST, DM & PKD, DM & ST, PKD & ST, CAD, DM & PKD, CAD, DM & ST, CAD, PKD & ST, DM, PKD & ST, CAD, DM, PKD & ST. In [table 2](#), it shows the number for common genes. Those genes are found from 4 selected diseases after gene linkage.

4.2.2 Common gene finding:

After collecting data from NCBI it was saved in an xlsx sheet, which is used in Python for finding common genes. There are 67 common genes between CAD, DM, PKD & ST. there are 87 common gene between CAD & DM, 19 common genes between CAD & PKD, 142 common genes between CAD & ST, 32 common genes between DM & PKD, 101 common genes between DM & ST, 38 common genes between PKD & ST, 15 common genes between CAD, DM & PKD, 156 common genes between, CAD, DM & ST, 21 common genes between, CAD, PKD & ST, 18 common genes between, DM, PKD & ST and 67 common genes between CAD, DM, PKD & ST. In [table 2](#). it shows the common genes from all four diseases.

Table 2:

There in this table it shows the number of common genes between selected four diseases. There it shows the interconnection between four diseases by finding common genes.

Selected Diseases Name	Total number of Homo Sapiens Genes	Common Genes
CAD & DM	2025	407
CAD & PKD	1758	183
CAD & ST	1925	444
DM & PKD	1825	205
DM & ST	1992	399
PKD & ST	1725	218
CAD , DM & PKD	2804	117
CAD , PKD & ST	2704	125
CAD , DM & ST	2971	278
DM , PKD & ST	2771	126
CAD , DM , PKD & ST	3750	94

4.3 Venn Analysis:

Venn investigation frequently frames the main filtration step for mind boggling and interconnected information corpora. [38]. Venn charts empower understudies to sort out data outwardly so they can see the connections between a few arrangements of things. They would then be able to distinguish similarities and contrasts. We used Venny, a free online based tool to find out the visual intersect for the interconnected data. In [figure 2](#) it shows the venn diagram for respective data.

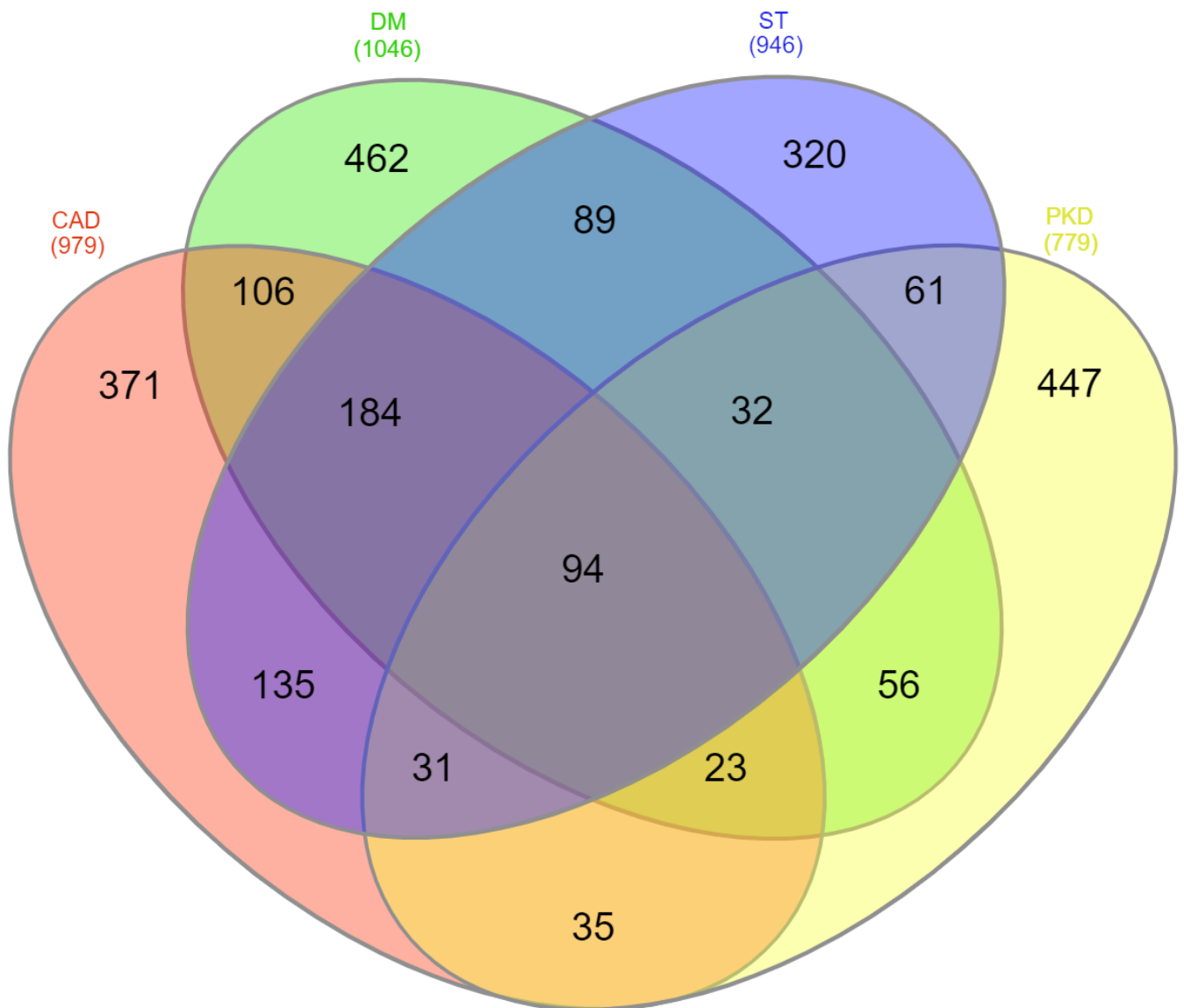


Figure 2: Venn analysis among CAD, DM, PKD & ST

4.4 Generic PPI:

PPIs are most important for almost every procedure in a cell, so knowing about PPIs is critical for understanding cell physiology in normal stages or in a disease stage. It leads an important role in drug design or invention, because drugs can affect PPIs. At the point when proteins cooperate, the impacts and communications among them can appear as far as a chart, which is called PPI network. [39] In a PPI network proteins are presented by nodes and interactions are presented by edges [39] NetworkAnalyst is an online based tool. We used this for creating the PPI network that has shown in [Figure.3](#)

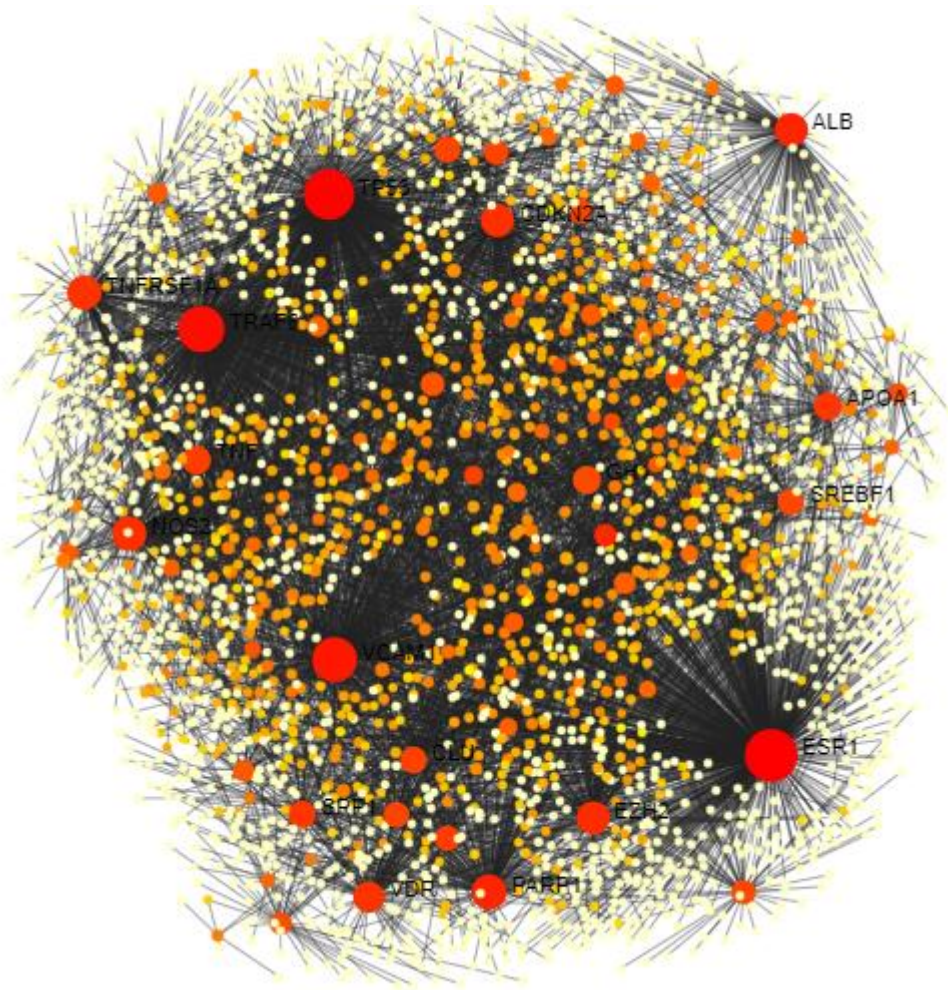


Figure 3: The 94 common genes for protein-protein interaction (PPI). There are 2939 nodes and 5180 edges in this network. There the Nodes are called proteins, and the edges establish a relationship between those proteins.

4.5 Topological Properties:

Topological studies is a stage, where mathematical models and metrics are shown by the networking properties. It helps to find out the relevant protein (nodes) and substructures in a way of biological approach. [25] In this analysis for topological properties studies a SIF file was downloaded from NetworkAnalyst based on PPI network. Then it is used in Cytoscape for finding topological properties for 10 hub genes, which are shown in [Table 3](#). In [figure \(4-7\)](#) it shows the topological properties for 10 hub g common genes based on PPI network SIF file.

Table: 3

Selected 10 hub genes topological properties are given below.

Protein Name	Degree	Betweenness Centrality	Closeness Centrality	Clustering Coefficient	Topological Coefficient
ESR1	798	.530327	.580906	.00116	.002227
TP53	659	.4082666	.512729	.001296	.002696
TRAF6	480	.305807	.470121	.001105	.00315
VCAM1	425	.301506	.41215	.001001	.001121
PARP1	179	.098054	.484601	.009165	.007689
NOS2	170	.067471	.353644	.001025	.163866
TNFRSF1	167	.062150	.351046	.001205	.152535
ALB	153	.110669	.351545	.001153	.064426
EZH2	149	.075485	.411356	.005261	.008514
CDKN2A	148	.067944	.398891	.006895	.009315

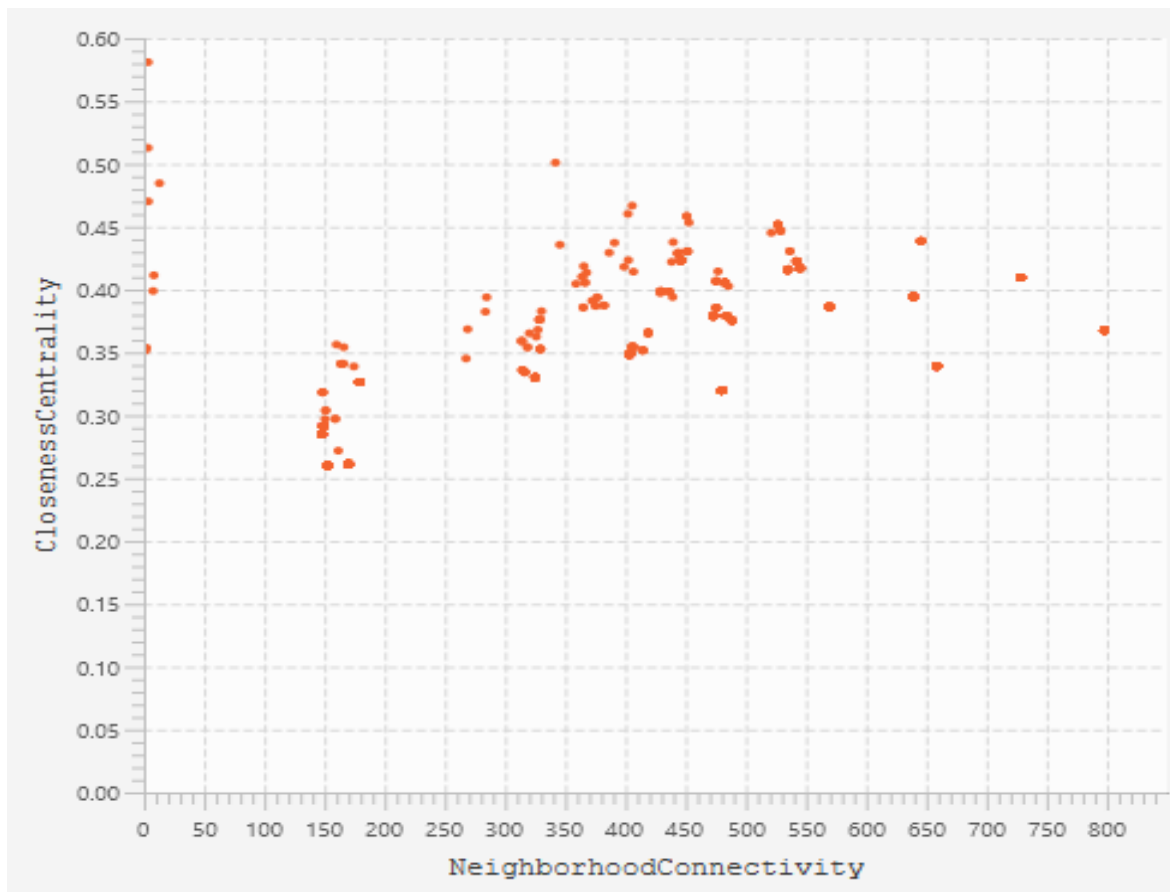


Figure 4 : This figure of CC illustrates the value of CC and the number of neighbors for each CC according to the PPI network. Where the value of CC is between 0.26 and 0.58 and the number of neighbors is between 0 and 800.

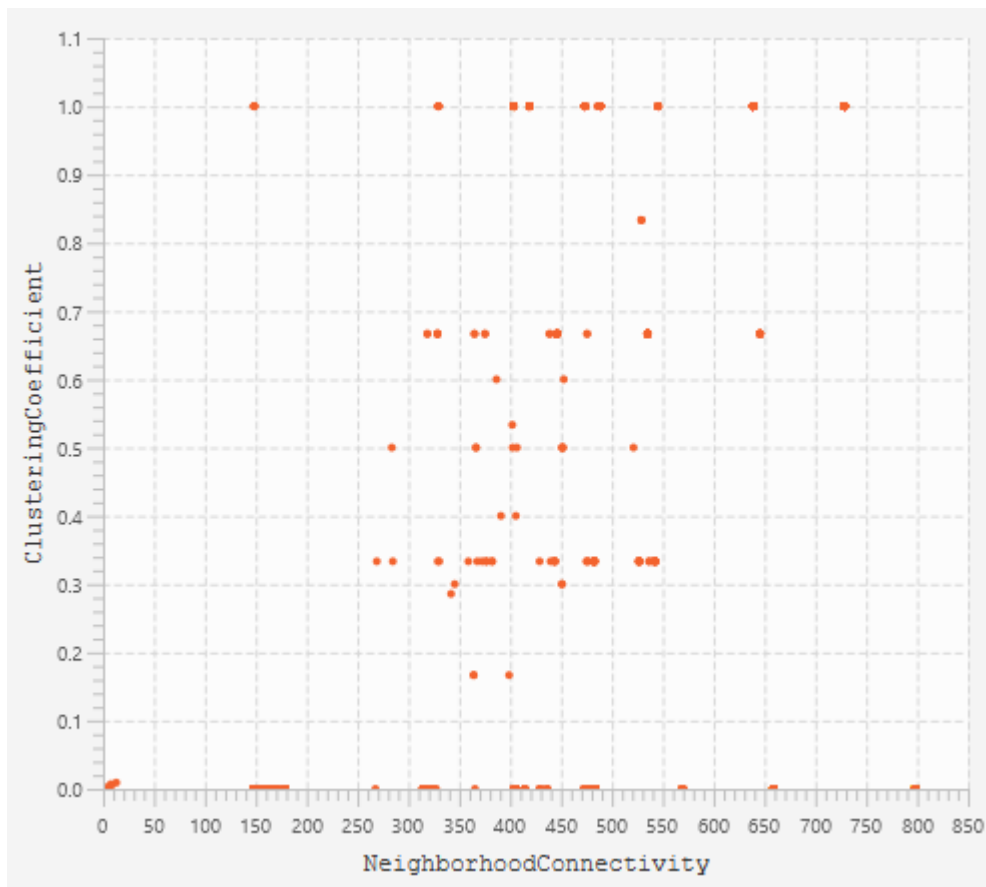


Figure 5: This figure of clustering coefficient illustrates the value of average clustering in PPI network. Where the value of clustering is between 0 and 1.0 and the number of neighbors is between 0 and 850

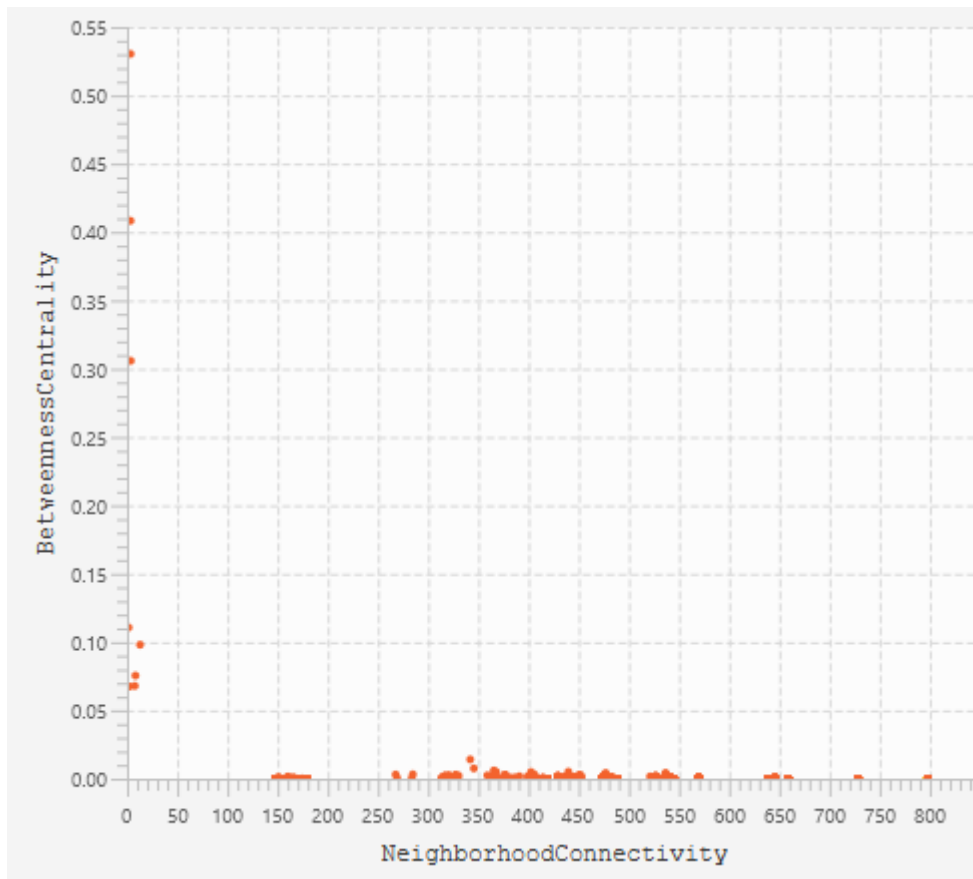


Figure 6 : Figure contains one of the topological properties called BC, where the value of BC and number of neighbors has been illustrated according to the PPI network. Here the value of BC is between 0.00 and 0.57, the number of neighbors is between 0 and 800.

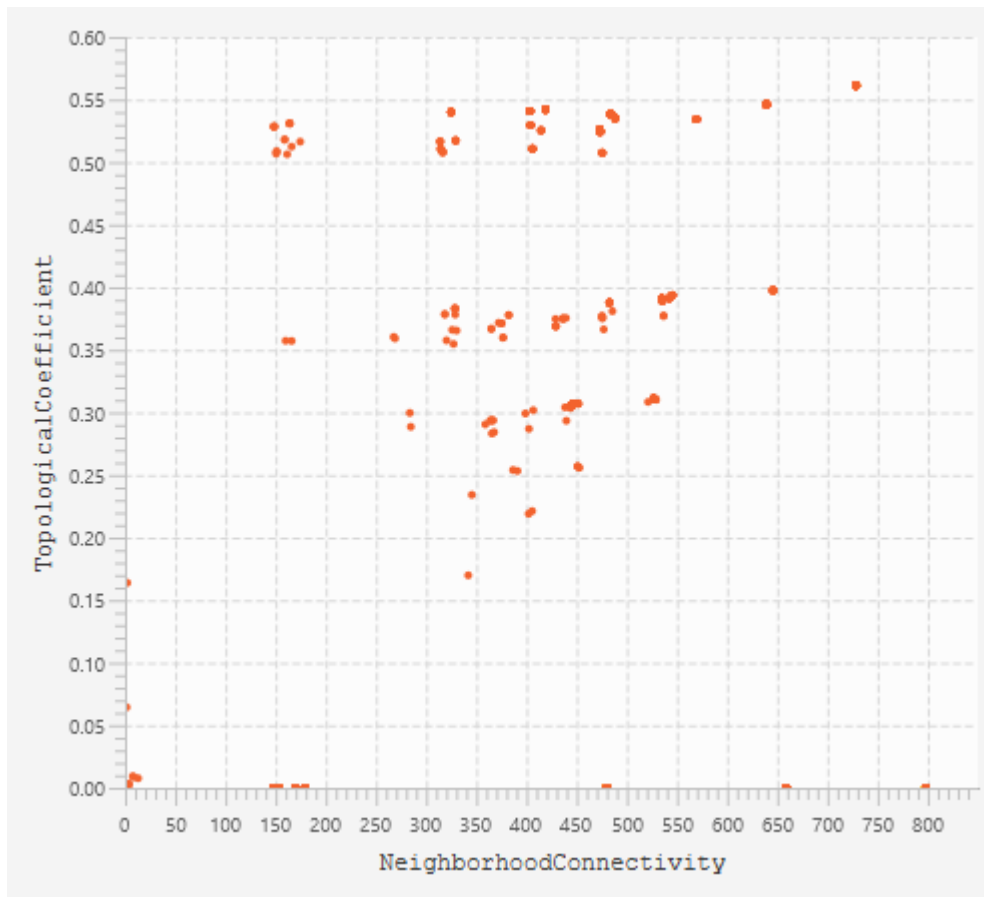


Figure 7 : Figure contains one of the topological properties called TC, where the value of TC and number of neighbors has been illustrated according to the PPI network. Here the value of TC is between 0.00 and 0.56, the number of neighbors is between 0 and 800.

4.6.1 Co-Expression:

The outcome in a co-expression network characterized as a diagram that are undirected in it the nodes represent the genes , co-expression connections are represented by edges.[25] The network for co-expression was created by GeneMania with 15 hub genes, that are shown in [Figure 8](#)

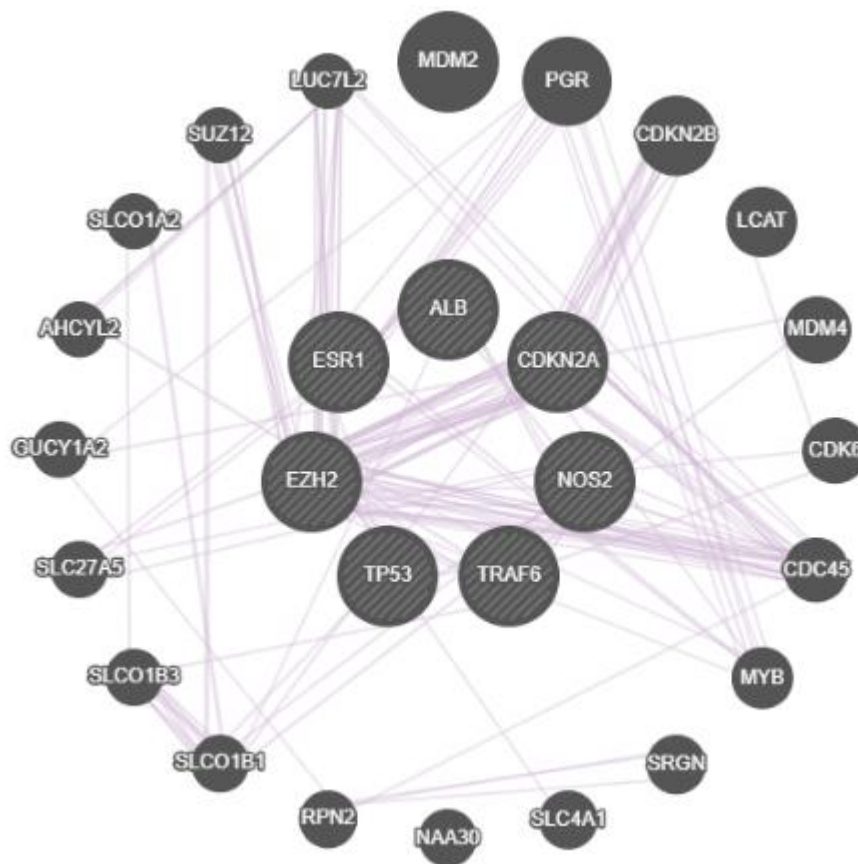


Figure 8: Co-expression between top 10 responsible genes.

4.6.2 Physical Interaction:

In physical interaction two or more genes are connected in the event . there they communicate in a PPI study. Physical interaction networks are shown in [Figure 9](#), by using GeneMania. .

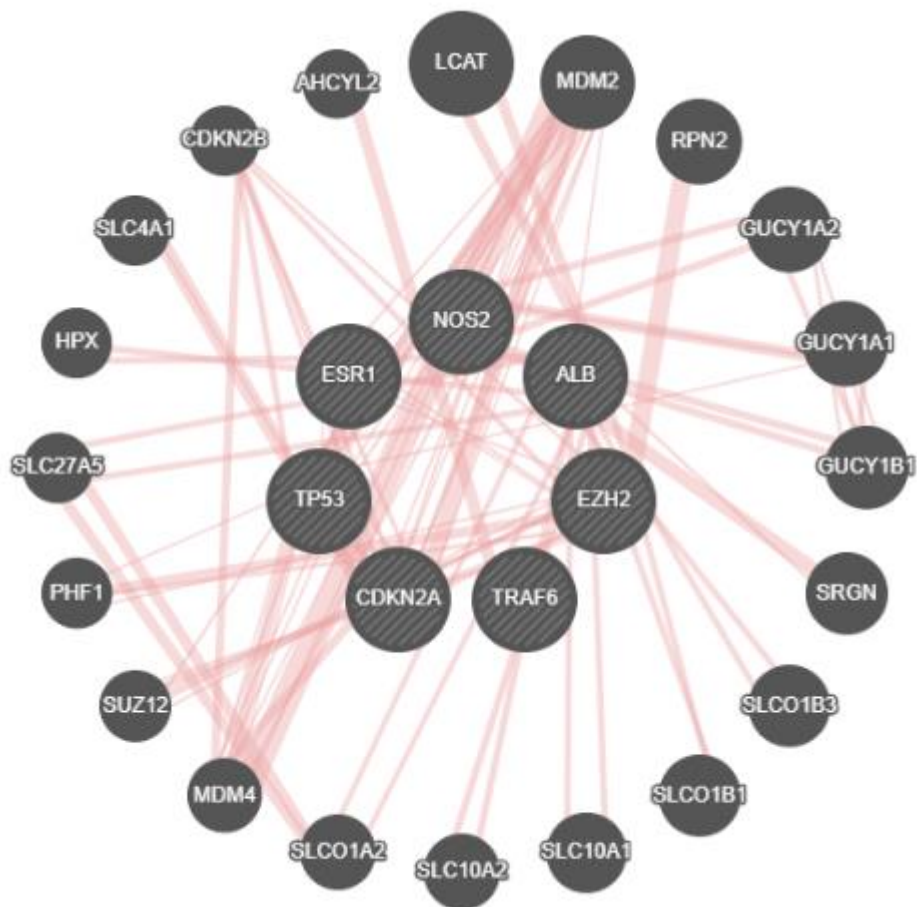


Figure 9 : Physical interaction between top 10 responsible genes.

4.6.3 Pathway analysis:

A standard way to analyze the action or reaction of molecules in a cell is biological pathway analysis. The investigation of natural pathways is a key to comprehend the various procedures within a cell. There proteins apply their capacity not in a disconnected way however in a systematic way of associations and their reactions.[27] If we compare it an individual quality based on methodology, the technique to make a system of various associated pathways and protein or genes of intrigue is progressively appropriate to investigate the science of complex characteristics and recognize functional competitor genes. [41] In [figure 10](#), it shows the pathway analysis network.

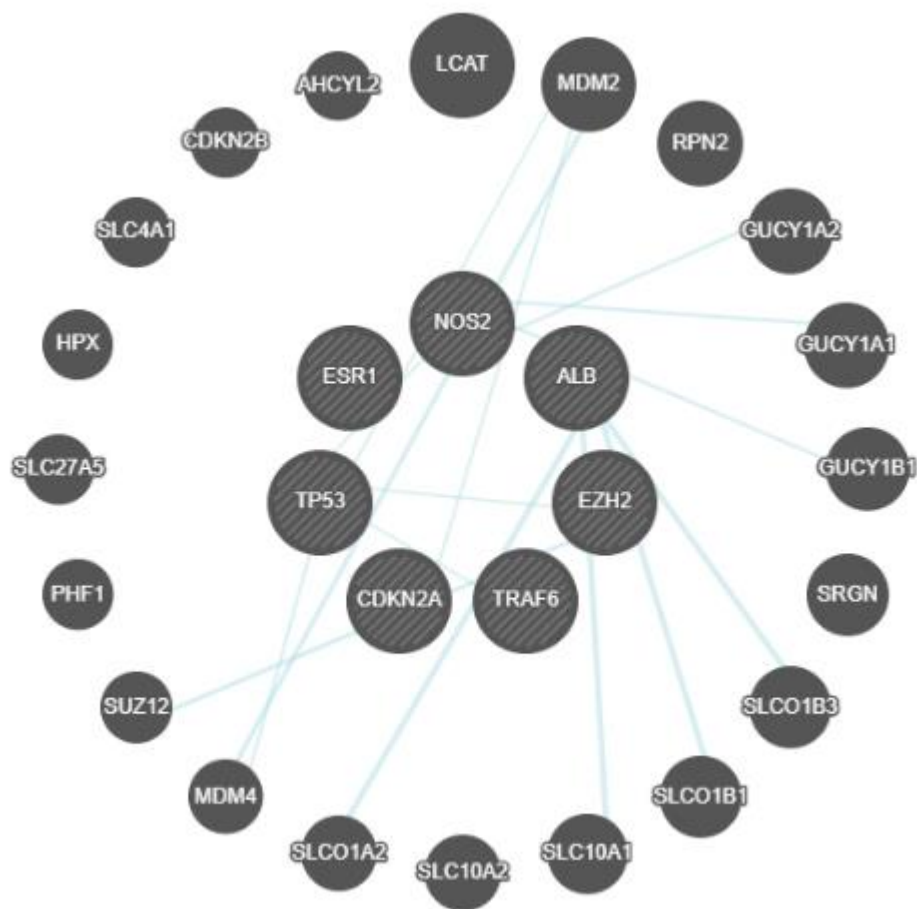


Figure 10: Pathway analysis between top 10 responsible genes.

4.7 Gene Regulatory Network:

GRN plays an imperative role in the spread of living organisms by performing cellular metabolism procedure.[42] There are three types of GRN in NetworkAnalyst. We use these to design GRN networks named Gene-miRNA network, TF-gene network and TF-miRNA Co-regulatory Network using 10 hub genes and find out the genomic programs analysis in a functional way.

4.7.1 Gene- miRNA:

The functional example of miRNA–mRNA regulatory network is invented in the beginning and movement of an assortment of diseases in Homo sapiens. MiRNAs work as key post-transcriptional controllers in an assortment of cells in biological approach, for example, separation, multiplication, apoptosis, movement and intrusion. [43] Gene-miRNA networks are shown in [figure 11](#), by using NetworkAnalyst.

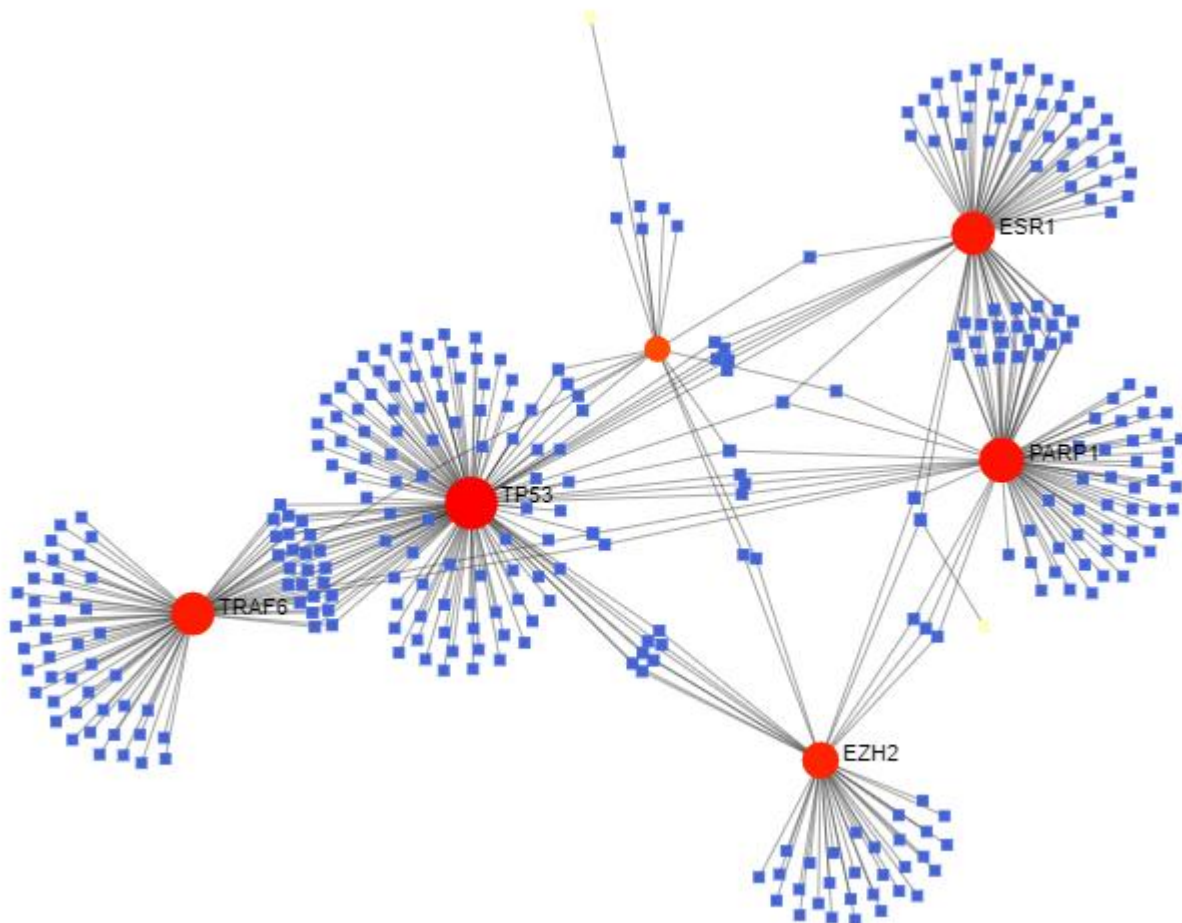


Figure 11: Gene -miRNA Interaction for selected top 10 genes. This gene-miRNA interaction generates interactions with a total of 413 links between 10 genes.

4.7.2: TF-gene:

From all cellular approaches controlling gene expression is one of most important parts. When a regulatory protein in a function controls the expression for another and in turn it may control the expression for another regulatory is called Transcription factors (TFs). It permits explicit signs to be intensified, and gives the data that is important to give a series of genes to make specific space and transient examples. [44] Tf-gene regulatory networks by using the GRN with NetworkAnalyst are given in [Figure12](#).

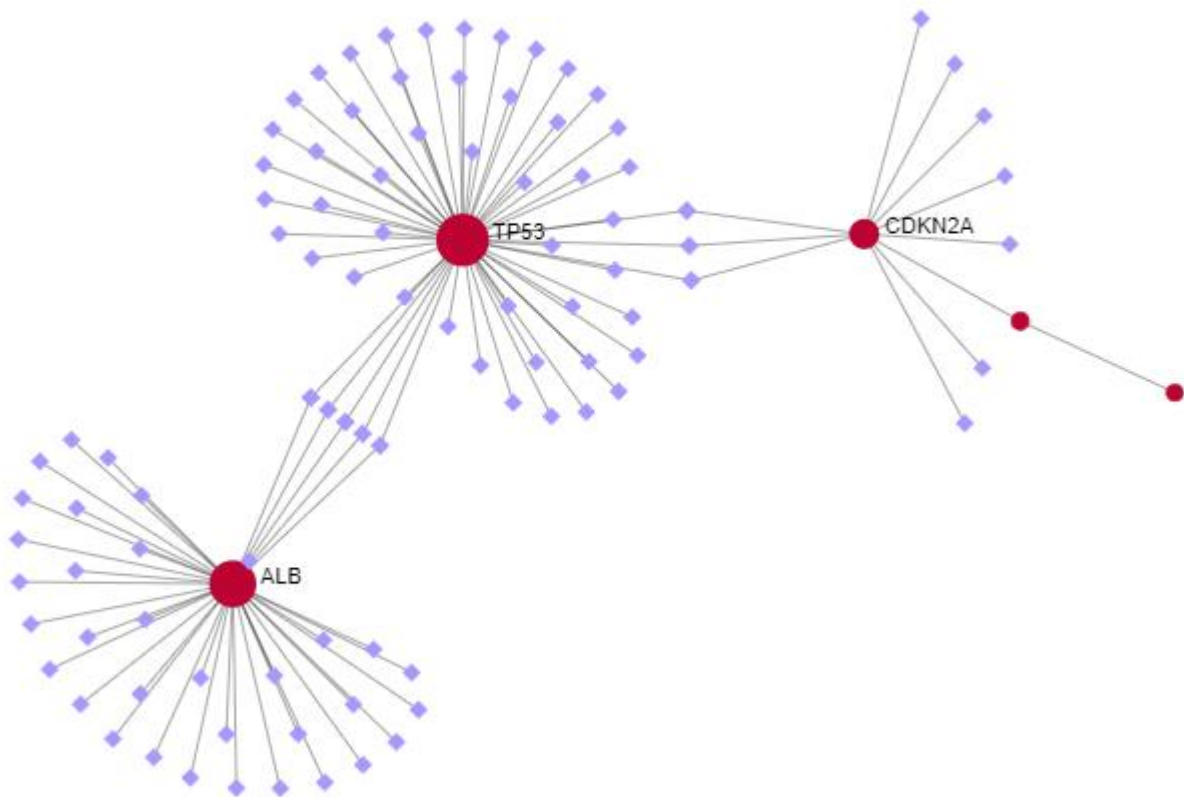


Figure 12: TF-gene Interaction for selected top 10 genes. This TF-gene Interaction creates relationships between 101 nodes and 106 edges. There are 5 seed nodes.

4.7.3: TF-miRNA:

The essential player in a complex administrative system of largest families of trans-acting GRN are miRNA and TFs. [45] In this part miRNAs and TFs exchange each other's expression, making it difficult to ascertain the effect either one has on TG expression. This network is shown in [figure,13](#).

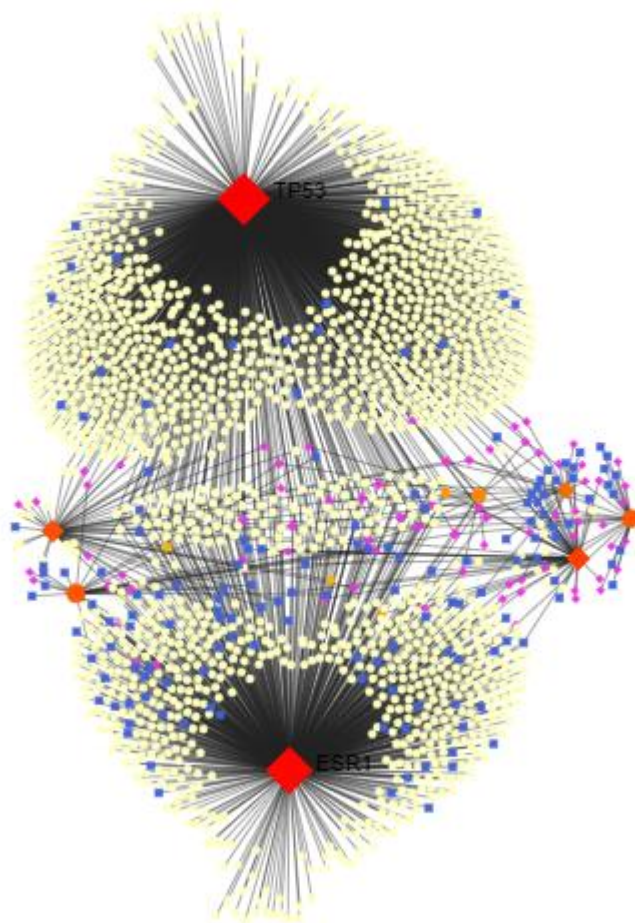


Figure 13 : TF-miRNA for genes. This TF top hub 10 common gene Interaction creates relationships between 1891 nodes and 2105 edges. There are 8 seed nodes.

Now-a-days, predicting drug-target interactions is becoming an important topic for inventing new drugs. So much effort has to be given for doing this interaction in many protein interactions, which are sometimes huge. In every kind of organism protein is a common element for doing interaction with drugs. [47] In [figure.14](#) it shows PDI network by using top 10 hub common genes for four diseases.



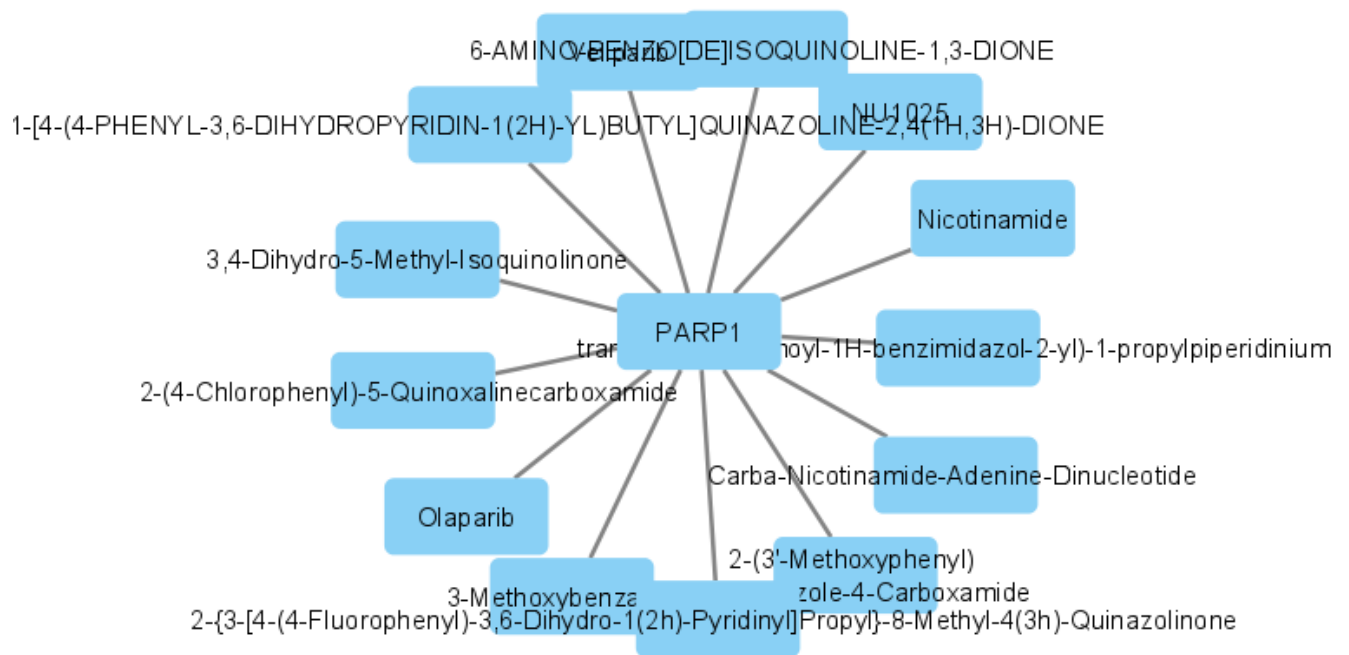


Figure 15: sub network-2 Protein-drug interaction for selected top 10 genes. It represents a subnetwork which creates relationships between 14 nodes and 13 edges. There have 1 seed nodes

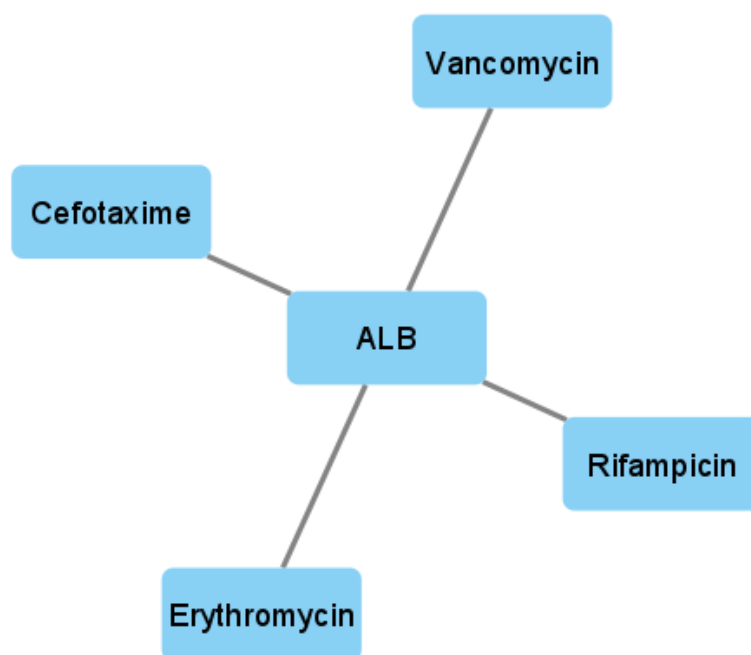


Figure 16: sub network-2 Protein-drug interaction for selected top 10 genes. It represents a subnetwork which creates relationships between 5 nodes and 4 edges. There have 1 seed nodes

4.8.2 Protein-Chemical Interaction:

PCI network is a networking process where it shows the binding affinities between chemicals in the interaction network. By this networking interaction one can easily find the effects of chemicals in the interaction with protein. [48] In [figure 17](#) it shows the PCI network.

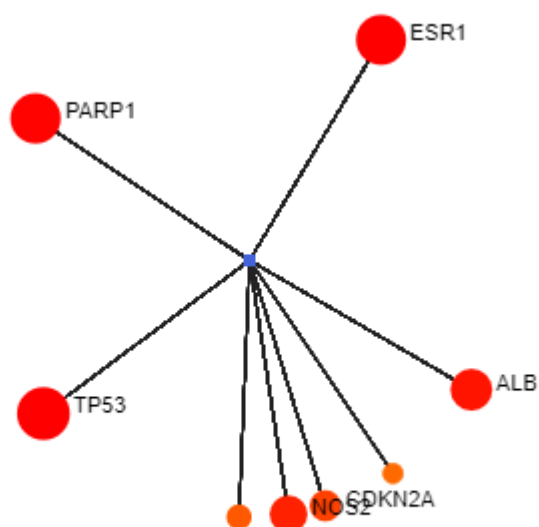


Figure 17 : Protein-chemical interaction for selected top 10 genes. This PCI creates relationships between 1842 nodes and 2685 edges. There are 8 seed nodes

4.8.3 Gene-disease Association:

The idea for GDA networking studies is an important field in genetic biology. In this part it helps to find out the genes which affect human bodies by creating new diseases. By using protein interaction it helps to find out the protein which are same or similar in multiple diseases.[49] [Figure 18](#), shows the interaction for GDA.

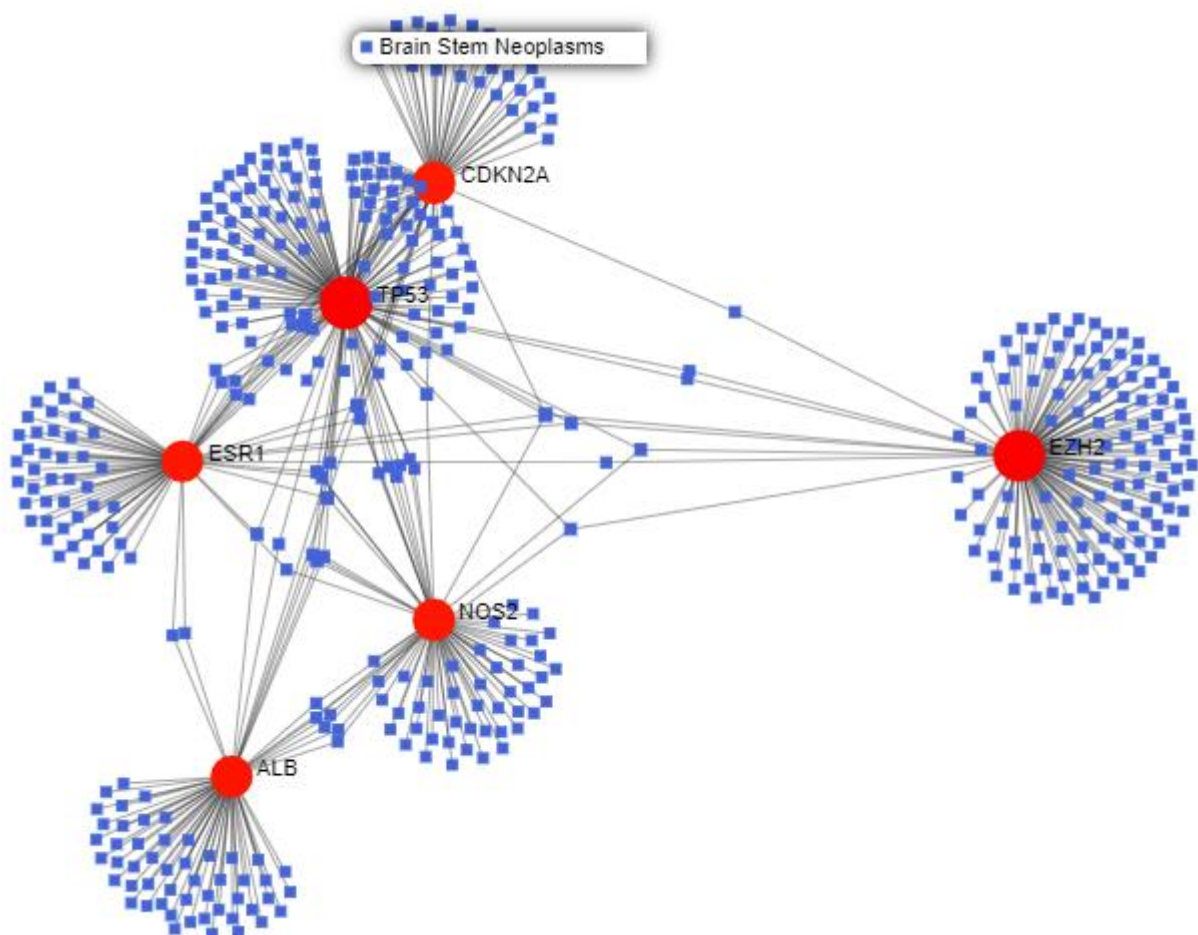


Figure 18 : Gene-disease Association for selected top 10 genes. This GDA creates relationships between 452 nodes and 537 edges. There are 6 seed nodes.

4.9 Clustering:

Clustering in proteins is normally used in similar proteins in a group to stabilize them and help them in functional annotation.[37] Clustering process by using PPI network can show the gathering of similar and different proteins in one network. For this analysis we used MCL clustering.

4.9.1 MCL Clustering:

In a clustering network, we can see many links between each cluster and some links between those clusters. This implies if you somehow managed to begin at a node, and afterward travel as you like to an associated node, you're bound to remain inside a group than movement between, this process is called MCL clustering. It helps to identify more meaningful clustering results from any other clustering process.. [50] For this process at first we collect the SIF file of the PPI network from NetworkAnalyst and then use it in Cytoscape . The networking visualizations have shown in [figure 19](#).

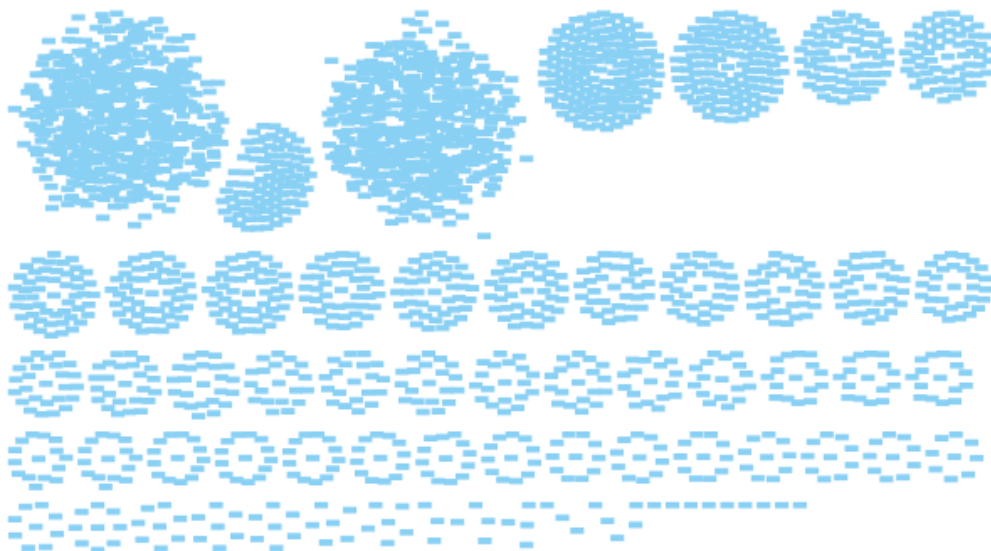


Figure 19 : MCL Clustering using Cytoscape. Here the PPI network (Figure 3) became clustered.

4.10 Summary :

Blend treatments offer across the board very much reported points of interest in the treatment of multiple diseases. Every year many people have to face death for CAD. CAD makes a huge impact on DM, PKD and ST which called upon a disaster in human lives. So it will be a great help to humans by inventing new drugs which will work on these multiple diseases at a time. In this study, we analyze GRN which shall indicate the interrelated gene between the diseases. Clustering can help to find out the similar groups which are associated strongly in all over the cluster. That will help in further work for inventing new diseases.

Chapter 5

Conclusion and recommendation

In this investigation we make a PPIs network and PDI using hub genes between four selected diseases CAD, DM, PKD, ST. After knowing the reason and finding the selected gene which is the reason for these diseases, we have to create a drug to cure it. PPI network gives us the information to understand the relation between drug targets and the proteins in these diseases that help in estimating drugs. Topological properties help to identify the function in proteins and mechanisms of action that helps in drug inventing. The PDI network, and co-expressions also contributed to drug design for the selected four diseases. It will help for further study in bioinformatics.

Reference

1. Chaurasia, Vikas and Pal, Saurabh, Early Prediction of Heart Diseases Using Data Mining Techniques (2013). Caribbean Journal of Science and Technology, Vol. 1, 208-217, 2013 . Available at SSRN: <https://ssrn.com/abstract=2991237>
2. <https://www.medicalnewstoday.com/articles/237191#types>
3. Ren, G., & Liu, Z. (2012). *NetCAD: a network analysis tool for coronary artery disease-associated PPI network*. *Bioinformatics*, 29(2), 279–280.
4. <https://www.webmd.com/heart-disease/risk-factors-for-heart-disease>
5. Okrainec, K., Banerjee, D. K., & Eisenberg, M. J. (2004). *Coronary artery disease in the developing world*. *American Heart Journal*, 148(1), 7–15.
6. Olefsky, J. M. (2001). *Prospects for Research in Diabetes Mellitus*. *JAMA*, 285(5), 628.
7. Kidambi, S., & Patel, S. B. (2008). *Diabetes Mellitus*. *The Journal of the American Dental Association*, 139, 8S–18S.
8. Krall, L. P. (1986). *Wide, Wide World of Diabetes Mellitus*. *The Diabetes Educator*, 12(4), 379–383.
9. Aynalem, S. B., & Zeleke, A. J. (2018). *Prevalence of Diabetes Mellitus and Its Risk Factors among Individuals Aged 15 Years and Above in Mizan-Aman Town, Southwest Ethiopia, 2016: A Cross Sectional Study*. *International Journal of Endocrinology*, 2018, 1–7.
10. Aronson, D., & Edelman, E. R. (2014). *Coronary Artery Disease and Diabetes Mellitus*. *Cardiology Clinics*, 32(3), 439–455.
11. DeMaagd, G. and Philip, A., 2015. *Parkinson's disease and its management: part 1: disease entity, risk factors, pathophysiology, clinical presentation, and diagnosis*. *Pharmacy and therapeutics*, 40(8), p.504.
12. Beitz, J. M. (2014). *Parkinson s disease a review*. *Frontiers in Bioscience*, S6(1), 65–74.
13. Jankovic, J., & Tan, E. K. (2020). *Parkinson's disease: etiopathogenesis and treatment*. *Journal of Neurology, Neurosurgery & Psychiatry*, jnnp–2019–322338.
14. Swallow, D.M., Lawton, M.A., Grosset, K.A., Malek, N., Klein, J., Baig, F., Ruffmann, C., Bajaj, N.P., Barker, R.A., Ben-Shlomo, Y. and Burn, D.J., 2016. *Statins are underused in recent-onset Parkinson's disease with increased vascular risk: findings from the UK Tracking Parkinson's and Oxford Parkinson's Disease Centre (OPDC) discovery cohorts*. *Journal of Neurology, Neurosurgery & Psychiatry*, 87(11), pp.1183-1190.
15. Johnson, W., Onuma, O., Owolabi, M., & Sachdev, S. (2016). *Stroke: a global response is needed*. *Bulletin of the World Health Organization*, 94(9), 634–634A.
16. Naylor, A. R., Mehta, Z., Rothwell, P. M., & Bell, P. R. F. (2002). *Carotid Artery Disease and Stroke During Coronary Artery Bypass:a Critical Review of the Literature*. *European Journal of Vascular and Endovascular Surgery*, 23(4), 283–294
17. Sileyew, K.J., 2019. *Research Design and Methodology*. In *Text Mining-Analysis, Programming and Application*. IntechOpen.
18. Kimber, O., Cromley, J.G. and Molnar-Kimber, K.L., 2018. *Let Your Ideas Flow: Using Flowcharts to Convey Methods and Implications of the Results in Laboratory Exercises, Articles, Posters, and Slide Presentations*. *Journal of microbiology & biology education*, 19(1).

19. Zhang, W. and Chen, T., 2012. Data preprocessing for web data mining. In *Advances in Electronic Commerce, Web Application and Communication* (pp. 303-307). Springer, Berlin, Heidelberg.
20. Bardou, P., Mariette, J., Escudié, F., Djemiel, C., & Klopp, C. (2014). jvenn: an interactive Venn diagram viewer. *BMC Bioinformatics*, 15(1), 293. doi:10.1186/1471-2105-15-293
21. Jonsson, P.F. and Bates, P.A., 2006. Global topological features of cancer proteins in the human interactome. *Bioinformatics*, 22(18), pp.2291-2297.
22. Nguyen, T.-P., Liu, W., & Jordán, F. (2011). Inferring pleiotropy by network analysis: linked diseases in the human PPI network. *BMC Systems Biology*, 5(1), 179. doi:10.1186/1752-0509-5-179
23. Sevimoglu, T., & Arga, K. Y. (2014). The role of protein interaction networks in systems biomedicine. *Computational and Structural Biotechnology Journal*, 11(18), 22–27. doi:10.1016/j.csbj.2014.08.008
24. Van Dam, S., Vösa, U., van der Graaf, A., Franke, L., & de Magalhães, J. P. (2017). Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in Bioinformatics*, bbw139. doi:10.1093/bib/bbw139
25. Vella, D., Zoppis, I., Mauri, G., Mauri, P. and Di Silvestre, D., 2017. From protein-protein interactions to protein co-expression networks: a new perspective to evaluate large-scale proteomic data. *EURASIP Journal on Bioinformatics and Systems Biology*, 2017(1), p.6.
26. Narayanan, M., Vetta, A., Schadt, E. E., & Zhu, J. (2010). Simultaneous Clustering of Multiple Gene Expression and Physical Interaction Datasets. *PLoS Computational Biology*, 6(4), e1000742. doi:10.1371/journal.pcbi.1000742
27. Villaveces, J.M., Koti, P. and Habermann, B.H., 2015. Tools for visualization and analysis of molecular networks, pathways, and-omics data. *Advances and applications in bioinformatics and chemistry: AABC*, 8, p.11.
28. Davidson, E. H., & Peter, I. S. (2015). Gene Regulatory Networks. *Genomic Control Process*, 41–77. doi:10.1016/b978-0-12-404729-7.00002-2
29. Hecker, M., Lambeck, S., Toepfer, S., van Someren, E., & Guthke, R. (2009). Gene regulatory network inference: Data integration in dynamic models—A review. *Biosystems*, 96(1), 86–103.
30. Yachie-Kinoshita, A., & Kaizu, K. (2018). Cell Modeling and Simulation. *Reference Module in Life Sciences*. doi:10.1016/b978-0-12-809633-8.20294-x
31. Roy, K., Kar, S., & Das, R. N. (2015). Newer QSAR Techniques. *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*, 319–356. doi:10.1016/b978-0-12-801505-6.00009-0
32. Li, J., Wang, L., Guo, M., Zhang, R., Dai, Q., Liu, X., ... Zhang, M. (2015). Mining disease genes using integrated protein-protein interaction and gene-gene co-regulation information. *FEBS Open Bio*, 5(1), 251–256. doi:10.1016/j.fob.2015.03.011
33. Mullokandov, E., Ahn, J., Szalkiewicz, A. and Babayeva, M., 2014. Protein binding drug-drug interaction between warfarin and tizoxanide in human plasma.
34. Cheng, F., Zhou, Y., Li, W., Liu, G. and Tang, Y., 2012. Prediction of chemical-protein interactions network with weighted network-based inference method. *PloS one*, 7(7), p.e41064.
35. Suratanee, A., & Plaimas, K. (2018). Network-based association analysis to infer new disease-gene relationships using large-scale protein interactions. *PLOS ONE*, 13(6), e0199435. doi:10.1371/journal.pone.0199435
36. Zaslavsky, L., Ciufu, S., Fedorov, B. and Tatusova, T., 2016. Clustering analysis of proteins from microbial genomes at multiple levels of resolution. *BMC bioinformatics*, 17(8), p.276.
37. Van Dongen, S., 2000. Performance criteria for graph clustering and Markov cluster experiments. In *NATIONAL RESEARCH INSTITUTE FOR MATHEMATICS AND COMPUTER SCIENCE IN THE*.

38. Cai, H., Chen, H., Yi, T., Daimon, C. M., Boyle, J. P., Peers, C., ... Martin, B. (2013). *VennPlex—A Novel Venn Diagram Program for Comparing and Visualizing Datasets with Differentially Regulated Datapoints*. *PLoS ONE*, 8(1), e53388. doi:10.1371/journal.pone.0053388
39. Guzzi, P. H., & Roy, S. (2020). *Protein interaction networks*. *Biological Network Analysis*, 133–166.
40. Taye, B., Vaz, C., Tanavde, V., Kuznetsov, V.A., Eisenhaber, F., Sugrue, R.J. and Maurer-Stroh, S., 2017. *Benchmarking selected computational gene network growing tools in context of virus-host interactions*. *Scientific reports*, 7(1), p.5805.
41. Palombo, V., Milanesi, M., Sferra, G., Capomaccio, S., Sgorlon, S. and D'Andrea, M., 2020. PANEV: an R package for a pathway-based network visualization. *BMC bioinformatics*, 21(1), p.46.
42. Mishra, S. and Mishra, D., 2016. Enhanced gene ranking approaches using modified trace ratio algorithm for gene expression data. *Informatics in Medicine Unlocked*, 5, pp.39-51.
43. Lou, W., Liu, J., Ding, B., Chen, D., Xu, L., Ding, J., Jiang, D., Zhou, L., Zheng, S. and Fan, W., 2019. *Identification of potential miRNA–mRNA regulatory network contributing to pathogenesis of HBV-related HCC*. *Journal of translational medicine*, 17(1), p.7.
44. Davuluri, R.V., Sun, H., Palaniswamy, S.K., Matthews, N., Molina, C., Kurtz, M. and Grotewold, E., 2003. *AGRIS: Arabidopsis gene regulatory information server, an information resource of Arabidopsis cis-regulatory elements and transcription factors*. *BMC bioinformatics*, 4(1), p.25.
45. Sharma, R., Upadhyay, S., Bhat, B., Singh, G., Bhattacharya, S. and Singh, A., 2020. *Abiotic stress induced miRNA-TF-gene regulatory network: a structural perspective*. *Genomics*, 112(1), pp.412-422.
46. Cheng, F., Kovács, I.A. and Barabási, A.L., 2019. Network-based prediction of drug combinations. *Nature communications*, 10(1), pp.1-11.
47. Wang, Y.C., Zhang, C.H., Deng, N.Y. and Wang, Y., 2011. Kernel-based data fusion improves the drug–protein interaction prediction. *Computational biology and chemistry*, 35(6), pp.353-362.
48. Szklarczyk, D., Santos, A., von Mering, C., Jensen, L.J., Bork, P. and Kuhn, M., 2016. STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data. *Nucleic acids research*, 44(D1), pp.D380-D384.
49. Vanunu, O., Magger, O., Ruppin, E., Shlomi, T. and Sharan, R., 2010. Associating genes and protein complexes with disease via network propagation. *PLoS Comput Biol*, 6(1), p.e1000641.
50. Vlasblom, J. and Wodak, S.J., 2009. Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. *BMC bioinformatics*, 10(1), pp-1.

