AN EXPLORATORY ANALYSIS OF A GENE-DISEASE NETWORK

Social Network Analysis

Seminar Paper

Anthony Fernando
anthony.fernando@uni-potsdam.de
816086

ABSTRACT

In medicine and biology, the reason why diseases occur is a highly researched topic. One of the reasons why a disease occurs is the mutation of genes. Nevertheless, many genes are associated with only one disease. This work is aimed to analyse a gene-disease network to observe a unique number of genes that are responsible for a particular group of diseases. To do this, a gene-disease network and its corresponding projection networks have been generated. It could be noted from the networks that mutations with specific genes might result in various diseases that are related to these genes. Patients having a certain disease also have a higher chance of getting another one. Lastly, the reasons of the connection between diseases or genes themselves were found and illustrated.

Keywords: gene mutation; bipartite graphs; projection network; gene-gene network; disease-disease network; gene-disease network

1 INTRODUCTION

It is the year 2022, and there are so many diseases all over the world. A few of them are mild, while others like cancer are deathly (Hassanpour & Dehghani 2017; Roy, Saikia, *et al.* 2016). In the last decades, scientists discovered that genes might be associated with diseases (Domazet-Lošo & Tautz 2008). With this knowledge, it is easier to detect diseases earlier for patients (Evans *et al.* 2001). In the medical field, many ways exist to analyse genes associated with diseases (Opap & Mulder 2017). One idea is the analysis of a gene-disease network. In 2007 the group of Kwang-Il Goh created a gene-disease network, out of which they drew several conclusions (Goh *et al.* 2007). Nevertheless, sometimes there are many genes that are associated with just only one disease (Zlotogora 2007). In other words, a scientist must analyse many genes to find the reason why this particular disease occurs. Not all genes associated with this disease are the main reason for such disease to appear (Goh *et al.* 2007). It is crucial to detect the disease early because a few diseases need to be treated as early as possible.

The objective of this seminar paper is to make a comparative approach to the work from 2007 of (Goh *et al.* 2007). For this, a gene-disease dataset from 2019 has been collected and pre-processed to get all the genes that are linked to various diseases. These diseases were further filtered and only the diseases with a high association to many genes were kept. In other words, with that new dataset, a new network and a projection of gene-gene and disease-disease networks have been developed by answering the four following questions, which are gene and disease-related: *(1) Which group of diseases helps in predicting the risk of having a particular disease A when another disease B is present? , (2) Are diseases more likely to be associated with each other if they have the same property? , (3) What is the combination of genes that are associated with the most number of diseases?* and *(4) Are genes more likely connected when they are similar to each other?*

The seminar paper is structured as follows: The theoretical background in section 2 summarizes the biological background needed to answer the questions above and gives a brief summary of the previous work of (Goh *et al.* 2007). The methodology, in section 3, includes a detailed explanation of the dataset, a short introduction of the bipartite graph concept and an illustration of the graph measurements. The data pre-processing and implementation details are explained at the end of the methodology section. The results are presented in section 4 while section 5 discusses and answers the research questions. The last section 6 concludes the work of this seminar.

2 THEORETICAL BACKGROUND

This section introduces the biological background information needed for understanding and solving the main research questions, followed by a short summary of previous work about a gene-disease network.

2.1 *Basics of Human Genetics*

In genetics, the DNA is the blueprint of life (Setubal *et al.* 1997). DNA stands for deoxyribonucleic acid and is the hereditary material in almost every organism (Alliance *et al.* 2009). A specific part or sequence of the DNA is also called a gene, which acts as an instructor. Every human has approximately 25.000 genes (Nelson *et al.* 2008). These genes are located on chromosomes consisting of multiple DNA and histone protein packaged structures (Annunziato 2008). Every human has 23 pairs of chromosomes with a total of 46 (Nelson *et al.* 2008). In Figure 1, DNA packaging into a chromosome is illustrated. In 1956, Francis Crick developed the hypothesis of the central dogma of molecular biology. It describes the flow of genetic information within a biological system (Crick 1970). According to this, DNA will be transcribed into RNA, and this RNA will be translated into a protein. As stated by Crick, once this information has passed into a protein, it can not get transformed back into RNA or DNA (Crick 1970). In other words, the DNA or genes are the information required to produce proteins (Crick 1970). This process can be imagined as a recipe for generating a product, which is, in this case, a protein (Setubal *et al.* 1997).

In the Appendix section 7 this hypothesis is illustrated in Figure 1. In general, proteins are molecules that play an important role in the body and do most of the work in the cell (Chaffey 2003; Setubal *et al.* 1997). These proteins have several functions, for example, working as an enzyme and catalyse chemical reactions, be an antibody to protect the body from viruses and bacteria or working as a structural component for building a more complex and bigger protein structure (Nelson *et al.* 2008). In some situations, the protein does not work or has a miss function. One of the main reasons is a mutation in the genes responsible for that protein. To put it differently, the genes have been changed because this sequence has been damaged. This damage creates a false blueprint and produces proteins with the wrong function, or the proteins lose their function. This leads mostly to diseases which damage the body (Nelson *et al.* 2008).

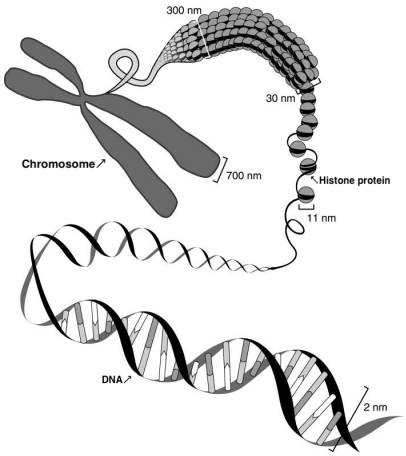


Figure 1: **Packaging of DNA into chromatin and chromosomes.** The figure is taken from (Weier 2001)

2.2 *State-of-the-Art*

One of the most cited works about gene-disease networks comes from (Goh *et al.* 2007). The title of this paper is "The human disease network ", published in 2007. The aim was to analyse if human genetic diseases and the corresponding genes are connected at a high level of cellular and organismal organization. For this, a diseasome (gene-disease network) has been created, which is a combination of all known diseases and their gene associations. An edge connects a disease and a gene if a mutation in that gene is implicated in that disease. Based on this diseasome, two projection networks have been generated. The first is the human disease network (disease-disease network), where every node is a disease and diseases are linked together if and only if they share at least one gene. The other projection network is a disease gene network (gene-gene network), where two genes are connected if and only if they have the same disease. Their dataset, where the associations between diseases and genes have been taken, is from the Online Mendelian Inheritance in Man (OMIM). It is a list collected in December 2005 that contains 1284 diseases and 1777 genes.

The following conclusions were drawn from their work: From the disease-disease network, it could be observed that 867 out of 1284 diseases have at least one edge over other diseases, and the remaining 517 nodes are from a hub. According to the degree distribution, it could be noticed that some diseases are connected to only a few other diseases, while diseases from the cancer area are connected with more than 50 diseases. Most of them are subtypes of cancer. Lastly, they could detect that from the disease-disease network, cancer and neurological diseases are mostly connected, whereas diseases from the metabolic or skeletal areas have less relation to other diseases. With regards to the gene-gene network, 1377 out of 1777 genes are associated with the same group of diseases, and 903 belong to a hub. Few genes associated with ten diseases represent major hubs in that network. Lastly, it could be observed that genes that are connected via diseases share cellular and functional characteristics.

3 METHODOLOGY

This section describes the methods and resources that have been used to solve the problems addressed in this seminar paper.

3.1 *Data Description*

The dataset used in this work is from the Stanford Large Network Dataset Collection[1]. It contains information about disease-associated genes in the human body. This information can be represented in an undirected gene-disease association network, in which two types of nodes are defined (*gene* and *disease*), and the edges represent the associations between these nodes. The dataset contains 7813 nodes, out of which 7294 are *gene* nodes and the remaining 519 nodes are *disease* nodes. These nodes are connected with a total of 21357 edges. The genes have been scrapped from the HUGO (Tweedie *et al.* 2021) and GeneOntology ("The Gene Ontology resource: enriching a GOld mine" 2021) databases, while diseases have been acquired from DiseaseOntology (Schriml *et al.* 2022) and OMIM[2] databases. The relation between genes and diseases has been defined by the CTD database (Davis *et al.* 2021). All these were collected in 2019. This gene-disease network for the complete dataset is demonstrated in Figure 2 in the Appendix section 7, while the disease with the most numbers of gene associations is also illustrated in the Appendix section 7 in Figure 3.

3.2 *Bipartite Graph*

The dataset, described in section 3.1, can be represented as a bipartite graph. According to Pavlopoulos, a graph $G = (U, V, E)$ is a bipartite graph (or two-mode network) if the nodes can be separated into two disjoints nonempty independent sets: U and V, where each edge from the edge set E connects a node from U to at least one node in V (Pavlopoulos

[1] https://snap.stanford.edu/biodata/datasets/10012/10012-DG-AssocMiner.html

[2] http://www.omim.org/

et al. 2018; Rahman *et al.* 2017). A graph is called a bipartite graph if it does not have any odd-length cycles (Pavlopoulos *et al.* 2018). For analysing the relationship among a group of nodes (e.g. U or V), the bipartite graph can be converted to a one-mode projection. Considering a one-mode projection onto the node set U, this one-mode projection creates a U - U network containing only U nodes (Zhou *et al.* 2007). The edges between the nodes are added to this network only if these nodes have one or more common neighbours V (Zhou *et al.* 2007). This projection can also be done for all other nodes set, such as V. An example of a bipartite graph and its corresponding U and V projection are illustrated in Figure 2, respectively.

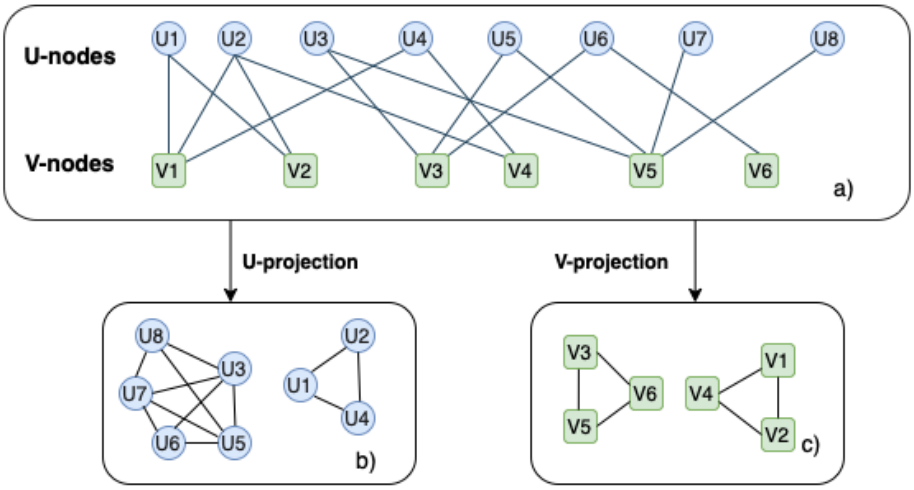


Figure 2: **a) An illustration of a bipartite network with U and V node sets. b) A one-mode projection onto a U-U network. c) An one-mode projection onto a V-V network.** Own construction, updated from: (Zhou *et al.* 2007).

3.3 *Graph Measurments*

This subsection describes the graph measures that have been used to analyse the generated networks. The first subsubsection describes the theory of the adjacency/biadjacency Matrix. Sections 3.3.2 and 3.3.3 illustrate the node degree and the closeness centrality measurements.

3.3.1 *Adjacency/Biadjacency Matrix*

Networks can be represented by adjacency and/or biadjacency matrices. In the graph theory an adjacency matrix, or connection matrix, is a measure for representing the connection between nodes of the same group (Singh & Sharma 2012; Weisstein 2007). In the case of a finite graph G, the adjacency matrix is defined as a n x n matrix, where n is the number of nodes in the network. Rows and columns represent the various nodes in the network. The adjacency matrix A is filled with values of 0s and 1s indicating either no edge or edge between the corresponding two nodes. For instance, if there is an edge between

node V_i and V_j, then value 1 will be added to the A_{ij} position in the adjacency matrix A, otherwise the value 0 is filled in that position. In the case of an undirected graph with no self-loops, the adjacency matrix is symmetric and the diagonals have zeros. To be noted, the order of the nodes in the adjacency matrix plays a role in analysing the corresponding graph. Permuting the rows or columns of the adjacency matrix results in different conclusions about the network (Singh & Sharma 2012; Weisstein 2007). An example of the adjacency matrix of the undirected graph V-V from the Figure 2c) is shown below.

$$\begin{array}{c} \\ V1 \\ V2 \\ V3 \\ V4 \\ V5 \\ V6 \end{array} \begin{array}{c} \begin{array}{cccccc} V1 & V2 & V3 & V4 & V5 & V6 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

On the other hand, a bipartite graph can be represented by a biadjacency matrix, in which two types of node groups are studied. To put it differently, a biadjacency matrix is similar to an adjacency matrix, however, with a matrix size of n x m representing the different node groups (Pavlopoulos *et al.* 2018; Singh & Sharma 2012).

3.3.2 *Node Degree*

Another measure used in this work is the node degree. In an undirected graph, the node degree, denoted as $deg(n)$, of a specific node n is defined as the number of edges that are connected to it (Trudeau 2013). Nodes with the highest degree are also called hubs (Pavlopoulos *et al.* 2018). In the gene-disease network, multiple hubs exist, each for a node group. In other words, a disease that is associated with a high number of genes is a hub disease, whereas a gene that is related to a high number of diseases is called a hub gene.

3.3.3 *Closeness Centrality*

The closeness centrality is another measurement used in this work. According to (Pavlopoulos *et al.* 2018) the closeness centrality is defined as how close a node is to all the remaining nodes in the graph. The value is computed as defined in equation 1, where U is a specific node, V is the set of the remaining nodes, and $d(U, V)$ is the shortest path between nodes U and V.

$$C(U) = \frac{1}{\sum_V d(U, V)} \tag{1}$$

A high closeness score of a node implies having the shortest distances to all the other nodes.

3.4 *Data Pre-Processing and Workflow*

This dataset has been pre-processed to address the research questions mentioned in section 1. The pre-processing workflow is illustrated in Figure 3. In the first step, all genes occurring between two to six times in the data have been selected. All diseases that are associated with these genes have been included in the filtered dataset (Data_limited_genes) [Step 1: Create Data_limited_genes]. In the second step, the node degree (as explained in section 3.3.2) for each of those diseases is computed. Here, the original dataset is used to compute the node degree of each disease (Data_limited_genes) [Step 2: Compute node degree]. In the third step, all diseases with a node degree equal to or higher than 100 genes have been further considered [Step 3: Select node $\geq$ 100]. This has been done because it was the ideal number to get the right disease hubs. In the fourth step, diseases connected to 5 or 6 genes are chosen. This number of connections has been taken because a higher number of connections would increase the analysis time. In the fifth step, the gene-disease bipartite network has been created [Step 4+5: Create bipartite network], while two projection networks from that bipartite network have been generated in the sixth step. Those projection networks are gene-gene network and disease-disease network [Step 6: Create projection networks]. In the last step, nodes with similar biologically properties were grouped together, and colours were added to the edges to describe each group connection in both projections [Step 7: Create weighted projection networks].

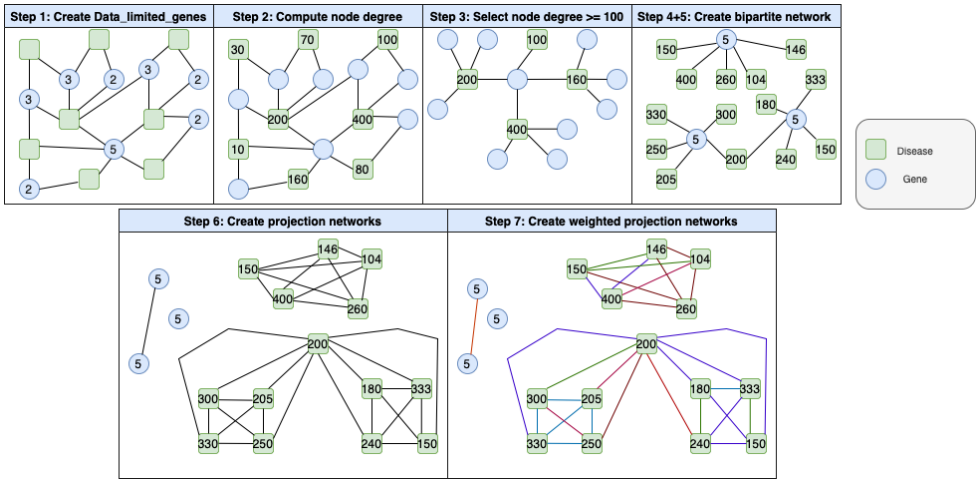


Figure 3: **Seven-steps workflow for analyzing the gene-disease dataset.**

3.5 *Implementation Details*

The scrips and data pre-processing have been implemented in Python 3.9.6[3]. Moreover, Pandas[4], NumPy[5] and Seaborn[6] packages have been used for analysing the data and the

[3] https://www.python.org
[4] https://pandas.pydata.org
[5] https://numpy.org
[6] https://seaborn.pydata.org

results. The visualisation of the networks has been done via Cytoscape (Shannon *et al.* 2003). The scripts and visualization files to create the results from this seminar report can be found under: https://github.com/anthonyhami/Social-Network-Analyses.git.

4 RESULT

This section summarizes the results for step 4,5,6 and 7, as described in section 3.4.

4.1 Step 4+5: Create Bipartite Network

The first two paragraphs outline the results for step 4, while the last paragraph summarises the result for step 5. The information about the available diseases in the dataset was acquired from the NCBI [7] and the DisGeNET databases (Piñero *et al.* 2020). This information includes the disease name and location. In addition to the NCBI database, the BioGPS database (Wu *et al.* 2016) was also used to extract the gene information.

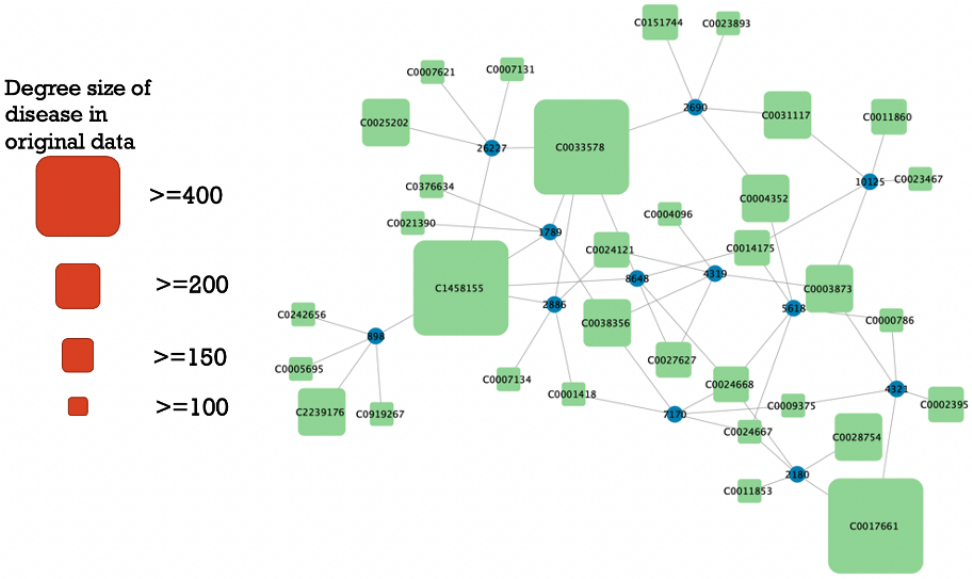
Tables 1 and 2 in the Appendix section 7 present the results of the remaining diseases and genes from step 4, respectively. Both tables show the ID, name and computed node degree of each element with regard to the original dataset. In addition to this information, Table 1 presents the location of each disease in the human body. Moreover, Table 2 includes the following additional information about the genes: chromosome-based location, expression location, number of proteins, number of molecular functions, number of biological process functions, and number of cellular component functions.

Table 1 summarizes the 33 obtained diseases from step 4. It can be observed that most of the diseases, 17 out of 33, are cancer- or tumour-related diseases. Further, it could be noticed that the disease with the highest node degree value is prostate cancer, with a total of 485 associated genes, while the disease with the lowest node degree value is inflammatory bowel, with only 100 associated genes. It is worth mentioning that most of these diseases occur in the nervous system or the lung. Table 2 presents the information about the 12 genes gained from step 4. It could be observed from Table 2 that nearly all genes are located on different chromosomes. Furthermore, it could be detected that these genes are mostly expressed in the brain or placenta. Another finding is that most of the proteins that these genes generate play a huge role in the biological process (refer to column "Number of biological process functions"). Moreover, the genes DNMT3B and PHGDH are responsible for producing the largest number of proteins, while gene MMP10 produces the lowest number of proteins, only 2 in this case.

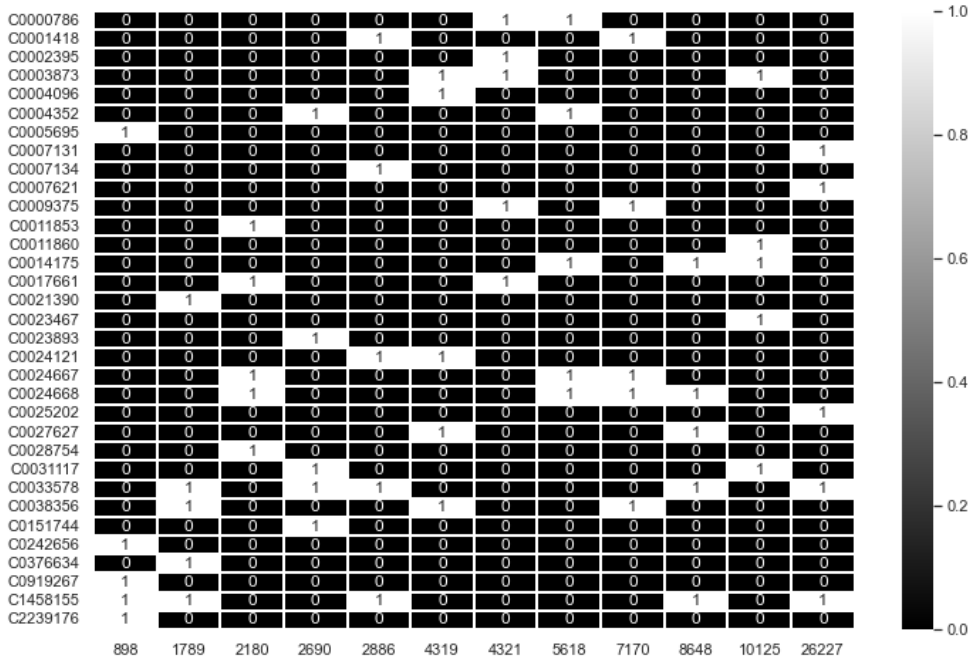
The results for step 5, where a bipartite network has been created, can be seen in Figure 4a with their corresponding biadjacency matrix. The green nodes represent diseases, whereas the blue nodes represent genes. The size of each disease node demonstrates its

[7] https://www.ncbi.nlm.nih.gov

node degree in the original data. There exists a total of 45 nodes and 60 edges in this network, in which 33 out of 45 nodes are disease nodes, and the remaining 12 are gene nodes. The corresponding biadjacency matrix is shown in Figure 4b. It can be seen that disease $C0033578$, which is prostate cancer, has the most association with a degree of 5. Also, it is essential to highlight that each gene is connected precisely with five diseases.



(a) Bipartite graph of the gene-disease network resulted from step 5.



(b) The corresponding biadjacency matrix of the bipartite gene-disease network.

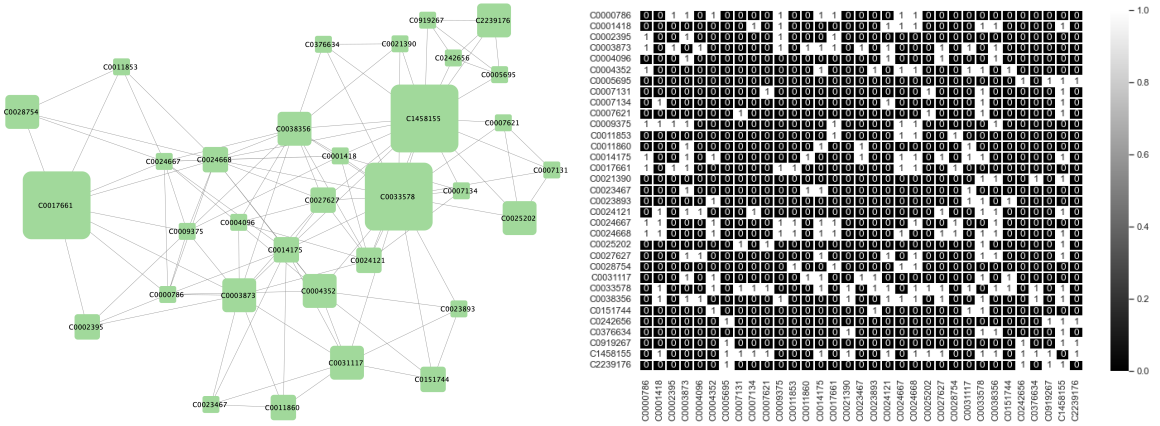
Figure 4: **Result of step 4**. The green nodes represent diseases, while the blue nodes represent genes. The size of the disease nodes represent the node degree in the original data.

4.2 Step 6: Create Projection Networks

The first paragraph outlines the results for the disease-disease network, while the second summarises it for the gene-gene network.

The disease projection network and the corresponding adjacency matrix are displayed in Figure 5. This network consists of 33 nodes and 114 edges. The diseases with the highest node degree of 17 are $C1458155$ (breast cancer) and $C0033578$ (prostate cancer). All diseases can be reached, if the 5 following diseases are connected: $C1458155$ (breast cancer), $C0033578$ (prostate cancer), C0038356 (tumour of stomach), $C0024668$ (Mammary Neoplasms, Experimental) and $C0003873$ (rheumatoid arthritis). Table 1 provides the closeness centrality of those diseases. The diseases with the highest closeness centrality are $C0033578$ (prostate cancer) and $C1458155$ (breast cancer), while the disease with the lowest closeness centrality is $C0002395$ (Alzheimer's disease). The mean closeness centrality, in this case, is 0.494.

The gene projection network and the corresponding adjacency matrix are shown in Figure 6. The gene-gene network consists of only 12 nodes and 36 edges compared to the disease projection network. The gene with the highest node degree of 10 is gene 8648, which is a transcriptional coactivator for steroid. The genes with the lowest degree are gene 2180, which plays a role in fatty acid degradation and gene 898, which takes part in the cell cycle. If both genes (8648 and 5618) are connected, then all other genes can be reached.



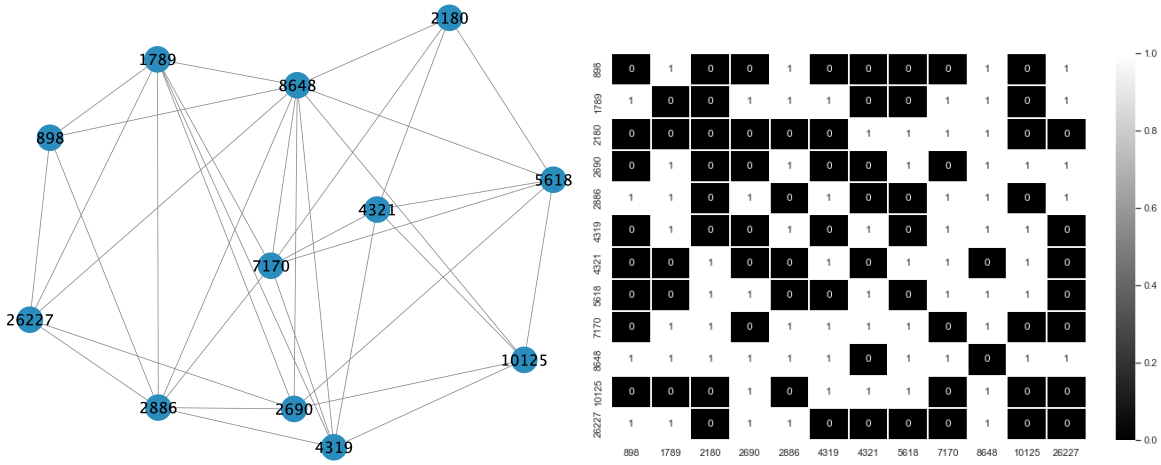
(a) Disease-Disease network.

(b) Adjacency matrix to the corresponding disease-disease network.

Figure 5: **Result of step 6 [disease projection]**.

Disease ID	Disease Name	Closeness centrality
C0033578	Prostate Cancer	0.666
C0017661	IGA Glomerulonephritis	0.470
C1458155	Breast Cancer	0.666
C0028754	Obesity	0.415
C0031117	Peripheral Neuropathy	0.516
C0038356	Tumour of Stomach	0.615
C0004352	Autistic Disorder	0.535
C2239176	Liver Cancer	0.421
C0003873	Rheumatoid Arthritis	0.542
C0025202	Skin Cancer	0.444
C0024121	Tumour of Lung	0.551
C0027627	Neoplasm Metastasis	0.571
C0011860	Type 2 Diabetes	0.438
C0151744	Myocardial Ischemia	0.457
C0002395	Alzheimer's Disease	0.395
C0014175	Endometriosis	0.603
C0024668	Mammary Neoplasms, Experimental	0.627
C0376634	Craniofacial Abnormalities	0.477
C0024667	Breast Cancer Animal	0.524
C0009375	Tumour of Colon	0.492
C0007134	Renal Cell Carcinoma	0.477
C0242656	Disease Progression	0.421
C0007131	Non-Small Cell Lung Cancer	0.444
C0023467	Leukemia	0.438
C0004096	Asthma	0.444
C0005695	Tumours Bladder	0.421
C0023893	Liver Cirrhosis, Experimental	0.457
C0000786	Spontaneous Abortion	0.484
C0001418	Adenocarcinoma	0.551
C0007621	Tumourigenesis	0.444
C0011853	Diabetes Mellitus, Experimental	0.415
C0919267	Ovarian Tumour	0.421
C0021390	Inflammatory Bowel Disease	0.477

Table 1: **Overview of all diseases from step 6 and their corresponding closeness centrality.** The disease ID is defined in the first column. In the second the full name of the disease is defined, while in the last column the closeness centrality is defined.



(a) Gene-Gene network.

(b) Adjacency matrix to the corresponding gene-gene network.

Figure 6: **Result of step 6 [gene projection]**.

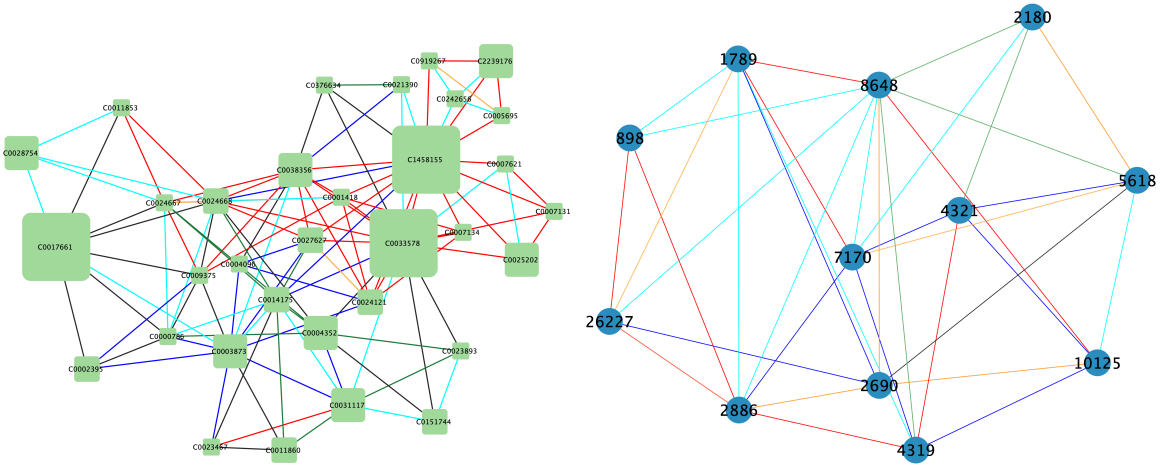
4.3 Step 7: Create Weighted Projection Networks

The first paragraph outlines the results for the clustered nodes in the disease-disease network, while the second paragraph summarises the clustered nodes for the gene-gene network.

The final result of the weighted disease-disease network is shown in Figure 7a. The different colours of the edges indicate the similarities of the two connected nodes: (1) Blue for diseases located at the same organ, (2) Red for the same or similar disease group, (3) Orange for the same disease group and are located at the same organ, (4) Cyan for diseases resulted from another disease, (5) Green for diseases having a similar number of genes, and (6) Black for unreasonable connection. It could be observed that most of the diseases are connected because they have the same disease (red). For instance, $C0007131$ is connected to the other diseases because nearly all of them are cancer-related diseases. Another example is the disease $C0003873$ from the blue group. It is connected to other disease nodes as they are from the same organ. Nevertheless, some nodes are connected to other nodes without a valid reason, for example, the edge between $C0017661$ and $C0002395$.

The weighted gene-gene network is illustrated in Figure 7b. The different colours of the edges indicate the similarities of the two connected nodes: (1) Cyan for genes that have the same or a similar number of cellular component functions, (2) Green for genes having the same or a similar number of biological process function, (3) Blue for genes having same or similar number of molecular function, (4) Red for genes that are expressed in the same or similar organ, (5) Orange for genes that generate a same or similar number of proteins, and (6) Black for genes that are located on the same chromosome. It could be observed that many genes are connected to each other due to various reasons, thus the difference in colours, for instance, gene 1789. Most of the edges connected to the gene 8648 are

coloured with cyan. This indicates that the other nodes have a similar or same number of proteins with a function for the cellular component. It can be noted that there exist paths in the network containing groups of genes that share the same attribute, such as the red paths of 7170-1789-8658-10125 or 4321-4319-2886-26227-898 indicating genes expressed in the same or similar organ. Genes 4321 and 7170 also share an edge coloured blue. It indicates that these genes have the same or a similar number of molecular functions.



(a) Disease-Disease weighted network.

(b) Gene-Gene weighted network.

Figure 7: **Result of step 7.**

5 DISCUSSION

This section summarizes the results by answering both the general and research questions defined in section 1, and outlines the limitation and future work.

5.1 Summary

The aim of this work is to generate a bipartite gene-disease network and its corresponding projections of disease-disease and gene-gene networks to answer the following questions:

1. (General): Which group of diseases helps in predicting the risk of having a particular disease A when another disease B is present?
2. (Research): Are diseases more likely to be associated with each other if they have the same property?
3. (General): What is the combination of genes that are associated with the most number of diseases?
4. (Research): Are genes more likely connected when they are similar to each other?

To answer these questions, a seven-step approach was realised, as explained in section 3.4, and the findings are listed in the following paragraphs. The research questions are also compared to the previous work of (Goh *et al.* 2007).

1. Question This question can be answered by analyzing the disease-disease network and extracting the values of closeness centrality in Table 1. The higher the closeness centrality value of a disease node, as mentioned in section 3.3.3, the more reachable this disease node is to all other nodes in the network. In other words, if one currently has a single disease, the probability of getting other diseases is high. For example, suppose a patient has a disease $C1458155$ (breast cancer). In that case, the chance of getting other diseases based on the network is higher in comparison to someone who has $C0002395$ (Alzheimer), as the latter has the lowest closeness centrality.

2. Question This question can be answered by analyzing Figure 7a. Indeed, diseases having the same property are more likely to be associated with each other. As seen in that figure, many reasons exist for the different associations between the disease nodes. One node can have a different type of connection between the other nodes. As an example is the node disease $C0007131$ (non-small cell lung cancer), which is connected to different diseases on the bases of a similar disease (cancer in this case). Similar results from the disease-disease network were found by the previous work of (Goh *et al.* 2007) as mentioned in section 2.2. Nevertheless, some diseases are connected to each other without an apparent reason. For example disease $C0021390$ (inflammatory bowel disease) is connected to $C0002395$ (Alzheimer). These two diseases, theoretically, should not be aligned. Such cases are interesting to further study, especially in the field of medicine.

3. Question This question is a gene-related question and can be answered by observing the adjacency matrix in Figure 6b. Both genes 8648 (nuclear receptor coactivator) and 5618 (prolactin receptor) are associated with the most number of diseases. Would these genes have been mutated or have any bad functionality, human having these mutated genes might get most of the diseases (e.g. the 33 diseases in the network). With this observation, the scientists in the wet lab should analyse these two genes in detail, in correlation of the 33 diseases, instead of analyzing a high number of genes. In this case, the analyses will be faster and helps in predicting diseases earlier.

4. Question Indeed, genes that are similar to each other are more likely to be connected. The connection can be realized based on different reasons (see the different colours in Figure 7b, and the explanation in section 4.3). Genes having the same attribute are linked together in the same path in the network. Similar results from the gene-gene network were found by the previous work of (Goh *et al.* 2007) as mentioned in section 2.2, as genes that are connected via diseases share cellular and functional characteristics.

5.2 Limitations and Future Work

During the conduction of this research in order to answer the defined questions, a few obstacles and limitations occurred. One of the main downsides of this research was the dataset itself. Since the dataset consists only of gene ID and disease ID, any information about the diseases and genes were missing. This had to be manually checked by exploring

several gene and disease databases. That was the reason why it was not possible to change and test a different number of gene linkages in step 4 (data filter). One of the future work that could be done to extend this project and to improve the current results is by changing the parameters for step 4 to analyse a more extensive network with several gene and disease nodes. The bigger the network, the more relevant information regarding diseases and genes can be withdrawn. Thus, instead of focusing only on genes that are connected with 5-6 diseases, one can widen the network by integrating all genes connected to 2-6 diseases.

6 CONCLUSION

The focus of this work was analysing the gene-disease dataset, introduced in section 3.1, as well as answering the research questions defined in section 1. To accomplish this, two projection networks were created: gene-gene and disease-disease networks. From these networks, it was possible to detect the crucial genes that lead to a certain number of diseases. It was also possible to observe the risk of getting other diseases when a specific disease is present. Another finding is the similarities between diseases sharing certain genes. This could be clearly observed from the disease-disease weighted network. However, some connections between diseases were unexplainable (the black edges as shown in section 4.3). From the gene-gene weighted network, it can be seen that genes that share the same diseases have mostly the same properties. Based on these findings, studying the relationship between diseases themselves and genes themselves, as well as the relationship between diseases and genes, can help in understanding the reasoning why certain diseases occur and direct the focus of the scientists to detect certain diseases earlier based on the shared genes. Furthermore, it is worth to continue analysing the gene-disease network to improve disease research and detect diseases on time before it is too late for the patient.

References

1. Alliance, G., for Genetic, N. Y.-M.-A. C., Services, N. S., *et al.* Understanding genetics: a New York, mid-Atlantic guide for patients and health professionals (2009).
2. Annunziato, A. DNA packaging: nucleosomes and chromatin. *Nature Education* **1,** 26 (2008).
3. Bernaola, N., Michiels, M., Larrañaga, P. & Bielza, C. Learning massive interpretable gene regulatory networks of the human brain by merging Bayesian Networks. *bioRxiv* (2020).
4. Chaffey, N. *Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. Molecular biology of the cell. 4th edn.* 2003.
5. Crick, F. Central dogma of molecular biology. *Nature* **227,** 561–563 (1970).
6. Davis, A. P. *et al.* Comparative toxicogenomics database (CTD): update 2021. *Nucleic acids research* **49,** D1138–D1143 (2021).
7. Domazet-Lošo, T. & Tautz, D. An ancient evolutionary origin of genes associated with human genetic diseases. *Molecular biology and evolution* **25,** 2699–2707 (2008).

8. Evans, J. P., Skrzynia, C. & Burke, W. The complexities of predictive genetic testing. *Bmj* **322,** 1052–1056 (2001).
9. Goh, K.-I. *et al.* The human disease network. *Proceedings of the National Academy of Sciences* **104,** 8685–8690 (2007).
10. Hassanpour, S. H. & Dehghani, M. Review of cancer from perspective of molecular. *Journal of Cancer Research and Practice* **4,** 127–129 (2017).
11. Nelson, D. L., Lehninger, A. L. & Cox, M. M. *Lehninger principles of biochemistry* (Macmillan, 2008).
12. Opap, K. & Mulder, N. Recent advances in predicting gene–disease associations. *F1000Research* **6** (2017).
13. Pavlopoulos, G. A. *et al.* Bipartite graphs in systems biology and medicine: a survey of methods and applications. *GigaScience* **7,** giy014 (2018).
14. Piñero, J. *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic acids research* **48,** D845–D855 (2020).
15. Rahman, M. S. *et al. Basic graph theory* (Springer, 2017).
16. Roy, P., Saikia, B., *et al.* Cancer and cure: a critical analysis. *Indian journal of cancer* **53,** 441 (2016).
17. Schriml, L. M. *et al.* The human disease ontology 2022 update. *Nucleic acids research* **50,** D1255–D1261 (2022).
18. Setubal, J. C., Meidanis, J. & Setubal-Meidanis, . *Introduction to computational molecular biology* **04; QH506, S4.** (PWS Pub. Boston, 1997).
19. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13,** 2498–2504 (2003).
20. Singh, H. & Sharma, R. Role of adjacency matrix & adjacency list in graph theory. *International Journal of Computers & Technology* **3,** 179–183 (2012).
21. The Gene Ontology resource: enriching a GOld mine. *Nucleic acids research* **49,** D325–D334 (2021).
22. Trudeau, R. J. *Introduction to graph theory* (Courier Corporation, 2013).
23. Tweedie, S. *et al.* Genenames. org: the HGNC and VGNC resources in 2021. *Nucleic acids research* **49,** D939–D946 (2021).
24. Weier, H.-U. G. DNA fiber mapping techniques for the assembly of high-resolution physical maps. *Journal of Histochemistry & Cytochemistry* **49,** 939–948 (2001).
25. Weisstein, E. W. Adjacency matrix. *https://mathworld. wolfram. com/* (2007).
26. Wu, C., Jin, X., Tsueng, G., Afrasiabi, C. & Su, A. I. BioGPS: building your own mash-up of gene annotations and expression profiles. *Nucleic acids research* **44,** D313–D316 (2016).

27. Zhou, T., Ren, J., Medo, M. š. & Zhang, Y.-C. Bipartite network projection and personal recommendation. *Phys. Rev. E* **76,** 046115. https://link.aps.org/doi/10.1103/PhysRevE.76.046115 (4 Oct. 2007).

28. Zlotogora, J. Multiple mutations responsible for frequent genetic diseases in isolated populations. *European journal of human genetics* **15,** 272–278 (2007).

7 APPENDIX

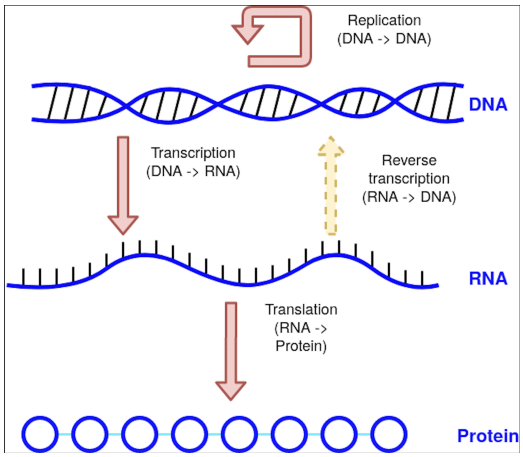


Figure 1: **Flow of information in a cell. Diagram explaining the central dogma of molecular biology.** The figure is taken from (Bernaola *et al.* 2020).

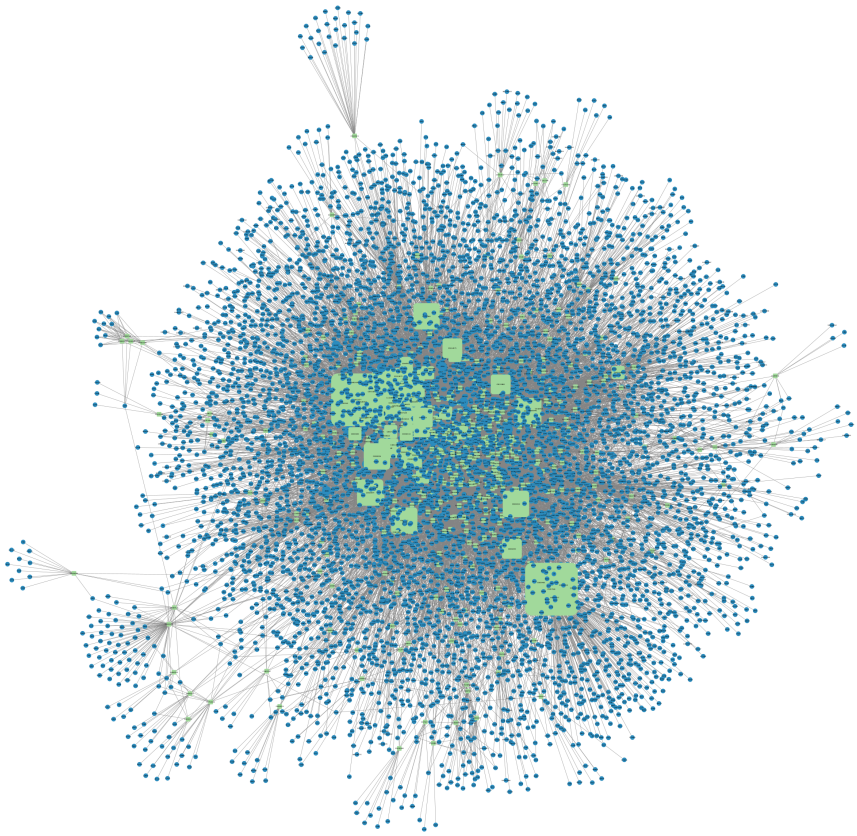


Figure 2: **Overview of the complete gene-disease network.** The blue points represent the genes and the green points represent the diseases. The edges are shown in gray colour.

Disease ID	Disease Name	Location of the Disease	Node Degree
C0033578	Prostate Cancer	Prostate	485
C0017661	IGA Glomerulonephritis	Kidney	450
C1458155	Breast Cancer	Breast	433
C0028754	Obesity	Every Part of the Body	298
C0031117	Peripheral Neuropathy	Nervous System	293
C0038356	Tumour of Stomach	Stomach	284
C0004352	Autistic Disorder	Nervous System	112
C2239176	Liver Cancer	Liver	222
C0003873	Rheumatoid Arthritis	Skin, Kidney, Heart, Lung or Nervous System	219
C0025202	Skin Cancer	Skin	204
C0024121	Tumour of Lung	Lung	178
C0027627	Neoplasm Metastasis	Nervous System, Lymph Nodes, Bone, Liver or Lung	174
C0011860	Type 2 Diabetes	Pancreas	173
C0151744	Myocardial Ischemia	Hearth	171
C0002395	Alzheimer's Disease	Nervous System	167
C0014175	Endometriosis	Rectum,Vagina,Vulva	161
C0024668	Mammary Neoplasms, Experimental	Breast	152
C0376634	Craniofacial Abnormalities	Cranium, Facial Bones	149
C0024667	Breast Cancer Animal	Breast	138
C0009375	Tumour of Colon	Colon	127
C0007134	Renal Cell Carcinoma	Kidney	127
C0242656	Disease Progression	Every Part of the Body	127
C0007131	Non-Small Cell Lung Cancer	Lung	117
C0023467	Leukemia	Bone Marrow,Liver,Nervous System or Testical	115
C0004096	Asthma	Lung	113
C0005695	Tumour of Bladder	Urinary, Bladder	111
C0023893	Liver Cirrhosis, Experimental	Liver	109
C0000786	Spontaneous Abortion	Urogenital	108
C0001418	Adenocarcinoma	Breast,Colon,Lung or Prostate	108
C0007621	Tumourigenesis	Several Part of the Body	106
C0011853	Diabetes Mellitus, Experimental	Pancrease	106
C0919267	Ovarian Tumour	Ovarian	101
C0021390	Inflammatory Bowel Disease	Bowel	100

Table 1: **Overview of all diseases from step 4.** The first column presents the disease ID, while the second and third columns define the name and location of the corresponding disease, respectively. The last column displays the node degree of each disease computed from the original dataset.

Gene ID	Gene Name	Chromosome Location	Gene Expression Location	Number of Proteins	Molecular Functions	Biological Process Functions	Cellular Component Functions	Node Degree
898	Cyclin E1	19	Placenta	11	4	14	9	6
1789	DNMT3B	20	Brain,Fetal and Thymus	36	10	9	6	6
26227	PHGDH	1	Brain and Prostate	36	10	9	6	5
2886	GRB7	17	Placenta	13	5	6	6	5
8648	NCOA1	2	Brain	21	16	21	10	5
2690	GHR	5	Liver	22	13	27	19	6
4319	MMP10	11	Uterus and Treachea	3	4	5	3	6
7170	TPM3	16	Skeletal and Muscle	33	3	2	9	6
2180	ACSL1	4	Liver and Kidney	30	9	20	11	6
5618	PRLR	5	Placenta	31	12	19	7	5
4321	MMP12	10	Stomach	2	11	21	5	6
10125	RASGRP1	15	Thymus and Brain	21	9	33	9	6

Table 2: **Overview of all genes from step 4.** The gene ID is defined in the first column, followed by the gene name, while in the third column, the location of the gene on which chromosome it is located. In the fourth column, the number of generated proteins that the gene generate is defined, while in columns 6 - 8, the number of molecular, biological process and cellular component functions are defined, respectively. In the last column, the number of disease associations is defined.

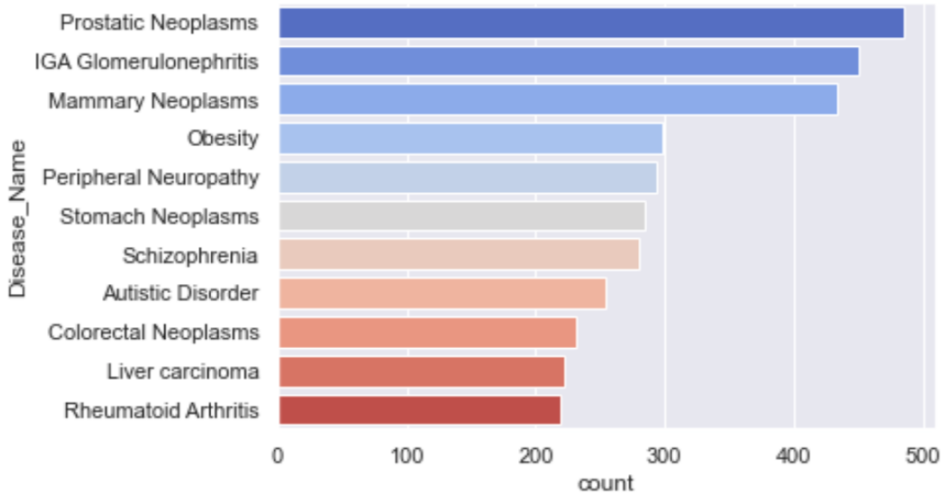


Figure 3: **Overview of the top 10 diseases with the most number of gene associations in the dataset.**