Graph-Based Analysis of Gene Activity in Lung Cancer based on Protein-Interaction Networks

Simone Bergmann
simone.bergmann@uni-bielefeld.de

October 27, 2024

Abstract

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat,sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse

1 Introduction

1.1 Motivation

Lung cancer is a major public health problem worldwide, caused by various factors, including smoking, air pollution, and also genetic factors. It is the leading cause of cancer-related deaths worldwide[**?**]. Despite advances in medical care, the survival rate for patients remains relatively low, with only 30.2% of women and 22.1% of men (2016) surviving beyond five years after diagnosis [**?**], often due to late-stage detection. Early detection is therefore crucial to initiate timely treatment and ultimately reduce the mortality rate.

We focus on analyzing gene expression data to identify potential biomarkers for lung cancer. We use graph databases for this purpose, as they offer ways to represent complex relationships between large sets of genes and proteins and analyze them efficiently using graph based algorithms. These data provide valuable insights into the underlying molecular mechanisms of lung cancer and thus help to improve the early diagnosis and treatment of lung cancer.

1.2 Goals

The main goal of this thesis is to identify genes that have a change in cancer activity and are central in the extended protein-protein interaction network that we are going to create. As these key genes are likely to be biomarkers for lung cancer, and could be used for early detection or personalized treatment. To achieve this goal, we have three strategic objectives:

1. **Objective - Find significant changes in gene activity:** We will determine the change in gene activity between healthy and cancerous tissues. In order to focus on genes with significant differences later in the study.
2. **Objective - Application of a graph algorithm:** We will use a graph algorithm to identify genes that are most central in the network to find candidates that are likely to be biomarkers for lung cancer.
3. **Objective - Validation of results through study comparison:** We will compare the identified genes, which exhibit significant changes in activity (1) and are highly influential in the network (2) with other studies to validate our results and ensure that their reproducible.

Our expected outcomes are:
A list of genes with high likelihood as potential biomarkers for lung cancer can be identified. This gene list is confirmed by a comprehensive analysis of the literature and comparisons with existing research papers.

1.3 Structure

The chapter 2provides a detailed overview of the biological and computational background of the project, including facts about lung cancer, the importance of gene expression analysis and the use of graph databases in cancer research.
In the third chapter 3, we will describe the setup, including the data sources and the creation of the graph database.
The fourth chapter 4 presents the results of the graph analysis and the identified relevant genes.
The fifth chapter 5 discusses the results and compares them with other studies.
The final chapter 6 summarizes the findings and provides an outlook on future research.

1.4 Limitations

Our Analysis has the following limitations:

- The data used in this project is limited on human genes.
- We focussed on the analysis of lung cancer.
- The gene expression data used in this project is limited to TPM values from the GTEx and Cell Model Passport datasets.

x In this chapter we talked about the motivation for the work, the goals of the thesis, the structure of the thesis and the limitations of the work.

2 Background

In this chapter we will take a closer look at the biological and computational background of our project. Also, we will discuss related work in the field cancer research with graph databases.

2.1 Biological background

Cancer is a group of diseases characterized by uncontrolled cell growth. According to research, widespread metastases are the primary cause of death from cancer[**?**]. There are over 100 different types of cancer[**?**]. The five most common forms of cancer worldwide in 2022 were lung cancer (> 2.4 million), breast cancer (> 2.2 million), colorectal cancer (> 1.9 million), prostate cancer (> 1.4 million) and stomach cancer (> 0.9 million). In 2022, cancer was one of the leading causes of death globally, with approximately 10 million fatalities[**?**].
It can be caused by a combination of genetic and environmental factors, such as diet, radiation, age, exposure to certain chemicals or viruses[**?**]. Early detection and the right treatment can significantly

improve the chances of curing many types of cancer.

Lung cancer is a type of cancer that affects the lung organ. There are two main subtypes: non-small cell carcinoma (NSCLC) and small cell carcinoma (SCLC)[**?**]. The primary risk factor for developing lung cancer is smoking, but passive smoking and environmental pollution also significantly increase the risk.
Unfortunately, symptoms of lung cancer often resemble those of common colds or other minor illnesses, such as coughing and fatigue. This makes it difficult to diagnose until the disease has progressed to an advanced stage[**?**].

Gene expression is a crucial factor in the investigation of cancer. It describes the process by which genes are read and utilized within a cell to synthesize proteins that regulate cellular growth and other essential processes. Alterations in gene expression can contribute to the uncontrolled multiplication and abnormal behavior of cancer cells, which ultimately leads to the development and progression of cancer.

Transcripts per Million (TPM) is a measure of gene expression in a cell. A higher TPM value signifies that the corresponding gene is actively expressed and, consequently, produces more protein. Through this method, it is possible to determine precisely which genes are activated within a cell and their involvement in cancer development.

2.2 Computational background

Graph databases are an effective way to model and analyze complex data structures. In the thesis, we use graph databases to put the collected cancer data into a meaningful context. A graph database consists of two basic components: Nodes and edges. Nodes are the units that store the basic information of a certain type of object. In our case, it is genes and proteins that are represented as nodes in the database. Each gene is therefore a node with its own properties, such as an ID, a name or a TPM value for its activity. Relationships between the nodes are represented by edges, which can represent different types of connections. In our graph database, for example, there are "interactions" as connections between proteins. Edges can represent both one-to-many and many-to-many relationships, which allows us to model complex relationships between nodes. Edges also can have weights or properties or a direction. We do not use these properties in our database, but they could be used to model more complex relationships between nodes. Graph databases provide several benefits when working with complex networked data. Notably, they enable efficient querying and visualization of complex relationships, facilitating a deeper understanding of these networks.
In biological data analysis, graph databases are often used to analyze protein-protein interaction (PPI) networks. These networks model the interaction between proteins to control cellular processes. By using a graph database, we can visualize and query these networks more efficiently, allowing us to gain important insights into cellular processes. [PAPER XY]
In our project, we take advantage of graph databases to analyze the graph-like connections between genes and proteins. This allows us to represent and query the complex relationships more efficiently. Part of our work also involves creating and using a PPI. For our project, we used Neo4J, one of the best-known and most powerful graph databases. It offers a high degree of flexibility and enables us to manage our data efficiently.

Graph database algorithms can be broadly categorized into three primary groups. These categories facilitate the efficient analysis of complex networks by providing different perspectives on graph structure.
Traversal and Pathfinding Algorithms enable the identification of shortest or most optimal paths between nodes in the graph, thereby facilitating the exploration of network topology. Examples include Depth-First-Search, Breadth-First-Search, Shortest Path, and Max-Flow-Min-Cut.
Centrality Algorithms are instrumental in understanding which nodes within a graph hold significant importance. By evaluating centrality measures such as degree centrality, closeness centrality,

betweenness centrality, and PageRank, valuable insights into the underlying network structure is gained.
Community Detection Algorithms, also known as clustering or partitioning algorithms, are essential for identifying groups of nodes within a graph that share similar characteristics or exhibit extensive connectivity. Examples include Label Propagation, Louvain Modularity, and Strongly Connected Components. [**?**]

In our analysis of gene activity networks, we utilize the **PageRank algorithm** to identify influential genes based on their connectivity and centrality within the graph. The PageRank score per gene is calculated using the following formula:

$$PR(A) = (1 - d) + d(\frac{PR(T_1)}{C(T_1)} + \cdots + \frac{PR(T_n)}{C(T_n)})$$

where $PR(A)$ is the PageRank score of node A, T_1 to T_n are nodes with edges to A, d is the damping factor, and $C(T_1)$ represents the number of edges to node T_1.
The PageRank algorithm was originally a method for evaluating and prioritizing websites on the internet. It was developed in 1988 at Stanford University by Larry Page and Sergey Brin [**?**], the founders of Google. The idea behind the algorithm is that the more links pointing to a website, the more important it is.
The PageRank algorithm assigns a higher score to genes that are connected to many other important genes, indicating their central role in the network. This allows us to identify key regulatory genes. By leveraging Neo4J's graph database capabilities, we can efficiently compute PageRank scores for our gene activity network, providing valuable insights into the complex relationships between genes.

2.3 Related work

The investigation of tumors and the identification of biomarkers for diagnosis and treatment are crucial tasks in oncology. Over the past few years, various approaches have been explored to address these challenges. Notably, graph-based methods have gained popularity for analyzing genetic data. This section presents a brief overview of studies in this field from the last years, focusing on the use of graph databases and their algorithms for biological applications and cancer research. By reviewing existing literature, we aim to provide a comprehensive understanding of the current state of the art in this area.
The recent systematic review by Rout et al. (2024) [**?**] provides an extensive overview of various graph-based methodologies for analyzing Protein-Protein Interaction networks, emphasizing best practices and common challenges in integrating multi-omics data. The authors highlight the importance of graph databases in storing and querying large-scale biological networks and essential graph algorithms such as centrality measures and community detection that are relevant to my work using PageRank.
A study by Simpson et al. (2020) [**?**] investigated the molecular mechanisms underlying various cancer types through the analysis of gene expression data sourced from the TCGA database. The authors employed co-expression networks, using Pearson correlation to establish relationships between genes based on their expression patterns. However, our thesis will take a distinct approach by concentrating specifically on lung cancer and utilizing PPI networks instead of co-expression networks. In contrast to Simpson et al.'s comprehensive application of multiple graph algorithms, including PageRank, Louvain community detection and Dijkstra's algorithm in each cancer type, we will focus on a more targeted approach. Specifically, we will apply the Pagerank algorithm to identify key genes within our PPI networks.
Shang and Liu (2020) [**?**] proposed a method for prioritizing cancer genes called iRank, which integrates various biological levels, including gene and protein expression, as well as Protein-Protein Interactions, to identify hepatocellular carcinoma. In contrast, we will focus on building PPI networks to analyze the interactions between proteins and their corresponding genes. Unlike Shang and Liu's approach, which concentrates on the TCGA dataset, our work will utilize the Cell Model Passport dataset. Similar to their approach, we will employ the PageRank algorithm to identify important genes in our analysis.

In summary, the studies mentioned above provide valuable insights into the application of graph-based methods for analyzing biological data. Each with a slightly different focus, they demonstrate the versatility of graph databases and algorithms in cancer research, so as we will do in our work.

Overall, in the course of this background section, we have gained a comprehensive overview of the biological background of lung cancer and the importance of gene expression analysis. In particular, we have looked at TPM values, which are an important measure of gene activity.
We have established that the use of graph databases is an effective way to analyze complex networks, and we have explained corresponding algorithms, especially the PageRank algorithm.
Finally, we have reviewed related work in the field of graph databases in the field of cancer research.

3 Experimental Setup

Here, we will delve into the details of the experimental setup. We will discuss the data sources and their transformation, the creation of the graph database, and the application of algorithms on the graph database.

3.1 Data

Our study consists of two main datasets. One includes healthy tissue data from the Genotype-Tissue Expression (GTEx) project, and the other includes cancerous tissue data from the Cell Model Passport (CMP) project. These two datasets are the base for the gene nodes in our graph database. For these, we need to have a single table with each gene as row and a single value for healthy and cancerous TPM values for every gene in every tissue. To have an identifier for every gene, we use the Ensemble ID (ENS ID) as the primary key for the gene nodes. These are unique id for genes, proteins and other genetic elements collected in the Ensemble database from 1999 by the European Bioinformatics Institute and the Wellcome Trust Sanger Institute [**?**].

For the healthy tissue samples used in this study, we utilized the "GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_tpm" dataset from the GTEx portal [**?**]. The **GTEx portal** is a large-scale, publicly available resource for studying gene activity. The Adult GTEx project aims to characterize the gene expression patterns in healthy tissues across different individuals and ages, providing valuable insights into the underlying biology of human development and disease. This dataset is part of the Bulk Tissue Expression data in the V8 release, which provides RNA sequencing data for a large number of tissue samples. The used dataset is in a .gct file format that contains TPM values for 56,156 genes identified by ENS ID as rows in 17,382 different tissues as columns. The data was initially stored in a file in wide format where each tissue had its own column. The TPM values for the tissues are between 0 and 747,400. Since there are no missing values in the dataset, we did not need to handle any missing data. To process this data into a suitable format, we employed the following steps:

1. **Reshaping to length format:** We read the data from the original file in chunks of 3,000 rows at a time to avoid memory issues. For each chunk of data, we transformed the columns for each tissue into individual rows, resulting in a dataset with three columns and 56,156 * 17,382 rows.

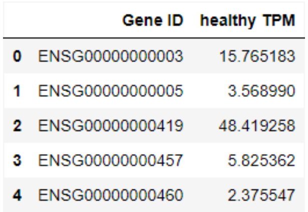
2. **Grouping by genes:** Once all chunks had been processed, we separated the combined dataset in new chunks of approximately 200 million rows. These chunks have been grouped by gene using an aggregate function that calculated both the sum and count of TPM values for each gene.

3. **Calculating mean TPM:** To handle genes that had been split across multiple chunks, we performed a global aggregation on the genes of the sum of the dataset. We then calculated the mean TPM value for each gene by dividing the sum by the count of each observation.

The resulting dataset 1 with 56,156 genes and a mean TPM value for every gene was saved as a CSV file for further processing.

Table 1: GTEx processing steps

Original Format	Reshaping to length format	Grouping by genes	Calculating mean TPM
$\begin{bmatrix} i \text{ - number of genes} \\ j \text{ - number of tissues} \\ a_{ij} \text{ - TPM value for gene i in tissue j} \end{bmatrix}$ $A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1j} \\ a_{21} & a_{22} & \cdots & a_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ij} \end{bmatrix}$	$A_{\text{long}} = (\text{Gen}_i, \text{Tissue}_j, a_{ij})$ $A_{\text{long}} = \begin{bmatrix} \text{Gen}_1 & \text{Tissue}_1 & a_{11} \\ \text{Gen}_1 & \text{Tissue}_2 & a_{12} \\ \vdots & \vdots & \vdots \\ \text{Gen}_i & \text{Tissue}_j & a_{ij} \end{bmatrix}$	$A_{\text{agg}} = (\text{Gen}_i, S_i, C_i)$ $S_i = \sum_{j=1}^{j} a_{ij}, \quad C_i = \sum_{j=1}^{j} 1$ $A_{\text{agg}} = \begin{bmatrix} \text{Gen}_1 & S_1 & C_1 \\ \text{Gen}_2 & S_2 & C_2 \\ \vdots & \vdots & \vdots \\ \text{Gen}_i & S_i & C_i \end{bmatrix}$	$A_{\text{mean}} = (\text{Gen}_i, M_i)$ $M_i = \frac{S_i}{C_i}$ $A_{\text{mean}} = \begin{bmatrix} \text{Gen}_1 & M_1 \\ \text{Gen}_2 & M_2 \\ \vdots & \vdots \\ \text{Gen}_i & M_i \end{bmatrix}$
$A = \begin{bmatrix} 8.764 & 0.07187 & 3.215 \end{bmatrix}$	$A_{\text{long}} = \begin{bmatrix} \text{ENSG...938} & \text{GTEX-111...} & 8.764 \\ \text{ENSG...938} & \text{GTEX-112...} & 0.07187 \\ \text{ENSG...938} & \text{GTEX-113...} & 3.215 \end{bmatrix}$	$A_{\text{agg}} = \begin{bmatrix} \text{ENSG...938} & 12.051 & 3 \end{bmatrix}$	$A_{\text{mean}} = \begin{bmatrix} \text{ENSG...938} & 4.017 \end{bmatrix}$



	Gene ID	healthy TPM
0	ENSG00000000003	15.765183
1	ENSG00000000005	3.568990
2	ENSG00000000419	48.419258
3	ENSG00000000457	5.825362
4	ENSG00000000460	2.375547

Figure 1: Example data of processed Genotype-Tissue Expression dataset

For the analysis of gene activity in lung cancer, we utilized data from the **Cell Model Passport (CMP) project**, a comprehensive resource for studying cancer-related gene expression.
We obtained the dataset from the CMP portal [?], specifically the ”Expression Data” section, with the name `rnaseq_all_data_20220624`. This dataset contains data from the Sanger Institute and the Broad Institute and consists of a large CSV file containing TPM values for genes associated with diverse cancer types, including lung cancer. Initially, the data was stored in long format with columns for CMP IDs for genes, tissues, TPM values, and additional information. To focus on lung cancer-specific data, we loaded an additional file containing model annotations. [?] We then filtered the CMP dataset to include only models from the annotation file with lung cancer as the cancer type.
The resulting dataset comprises 7,564,389 rows containing genes and tissues with associated TPM values for lung cancer. Specifically, the dataset includes information on 37,262 unique genes across 203 distinct tissue types. Notably, this dataset is free from missing values, and the TPM values span a wide range of 0 and 132,676.
To prepare the data for further processing, we performed the following steps:

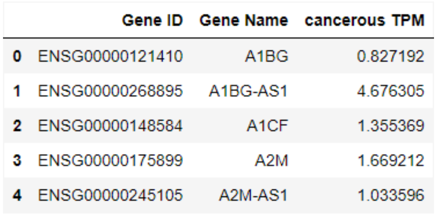
1. **Grouping by genes:** We grouped the dataset by genes to obtain a mean TPM value for every gene. This step involved aggregating the data by gene names, resulting in a new dataset with a mean TPM value for each gene.

2. **Adding ENS ID:** The original dataset contained only CMP ID per gene but lacked the universal Ensembl ID required for matching genes across datasets. To address this limitation, we needed to add corresponding Ensembl IDs to our genes using their gene_symbol. For this purpose we downloaded an Ensemble file from biomart [?], which contains the ENS ID and gene_symbol.

   By analyzing the file, we encountered an issue where some gene_symbols were not unique in the Ensemble file. To resolve this problem, we dropped all rows with duplicate gene_symbols. We then merged the Ensemble table with our CMP data on gene_symbols to retrieve the ENS IDs for each gene.

3. **Removing missing ENS ID:** After merging the data, we found that 3,760 genes had no ENS ID associated with them. Since these genes were likely duplicates or did not exist in the Ensemble file, we removed them from our dataset to ensure consistency and accuracy of our analysis.

The resulting dataset 2 contains 33,502 genes with mean TPM values for lung cancer and was saved as a CSV file for further processing.



	Gene ID	Gene Name	cancerous TPM
0	ENSG00000121410	A1BG	0.827192
1	ENSG00000268895	A1BG-AS1	4.676305
2	ENSG00000148584	A1CF	1.355369
3	ENSG00000175899	A2M	1.669212
4	ENSG00000245105	A2M-AS1	1.033596

Figure 2: Example data of processed Cell Modell Passport dataset

3.2 Nodes and Edges

As a next step we focus on describing how to create the base information for our advanced PPI Network. The graph database contains gene and protein node types. The Edges between the proteins build a classical PPI Network and are called Interactions. The second type of Edges, called Connections, are the Link between Proteins and Genes. As Shown in the Figure 3.

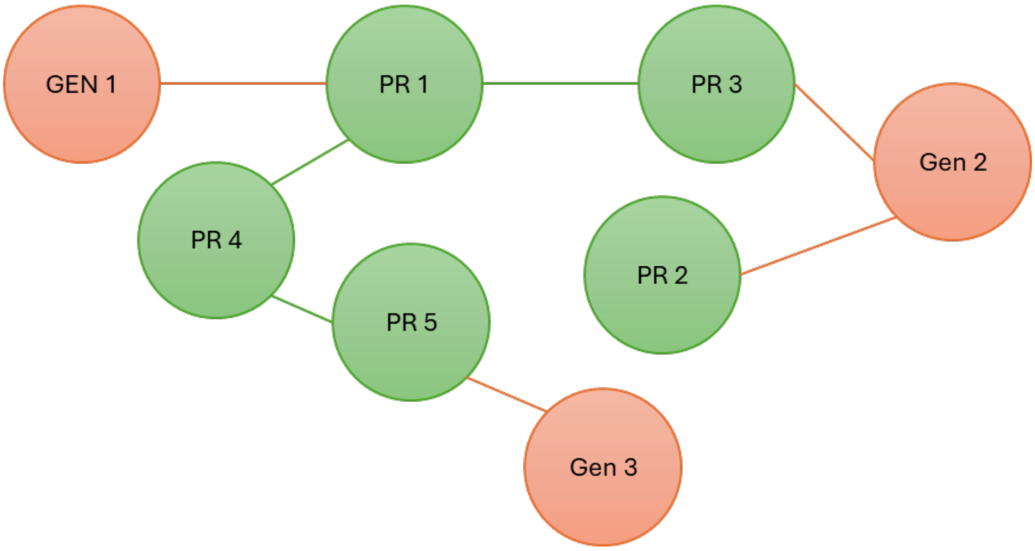


Figure 3: Schema of the Graph Database

For each of those 4 components we will create a table that will serve as base for creating the graph database.

For creating the **gene nodes** we need to use the preprocessed CMP and GTEx datasets, which contain mean TPM values for cancerous and healthy genes. We build the intersection of both datasets on their ENS ID to get a subset that only contains genes with TPM values for both conditions. To fulfill our first objective 1, we need to calculate a measure that captures significant changes between cancerous and healthy gene activity. When examining the mean TPM values per dataset, we observe a right-skewed distribution, with most values close to zero and a long tail extending towards higher values. The cancerous TPM values vary from 0 to approximately 41.173, while the healthy TPM values range from 0 to around 36.200.

To normalize the TPM values from both datasets and enable better comparability, we perform a common log scaling between 0 and 1 for all TPM values combined.

$$log\_norm(x) = \frac{\log(1+x)-\log(1+x_{min})}{\log(1+x_{max})-\log(1+x_{min})} \tag{1}$$

where x_{max} and x_{min} are the maximum and minimum TPM values across both datasets. After applying the normalization, the distribution of the TPM values is more balanced, as shown in Figure 3.2.

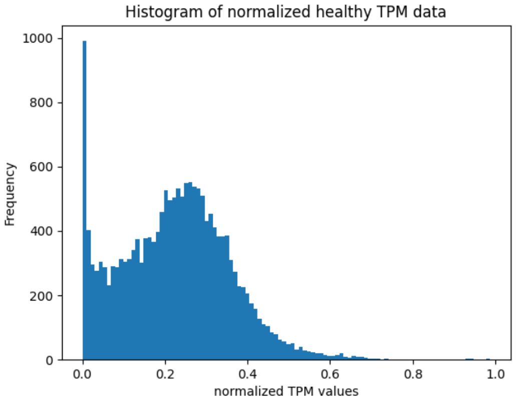


Figure 4: Histogram of TPM Values of GTEx Dataset

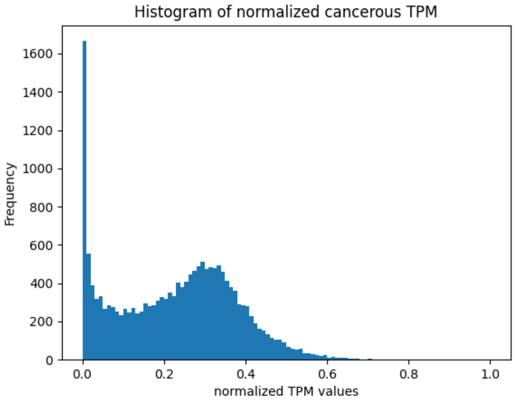


Figure 5: Histogram of TPM Values of CMP Dataset

Next, we calculate the difference between the normalized mean healthy and cancerous TPM values per gene by subtracting the two values and call it Δ_{TPM}. The distribution of Δ_{TPM} values is shown in Figure 6.

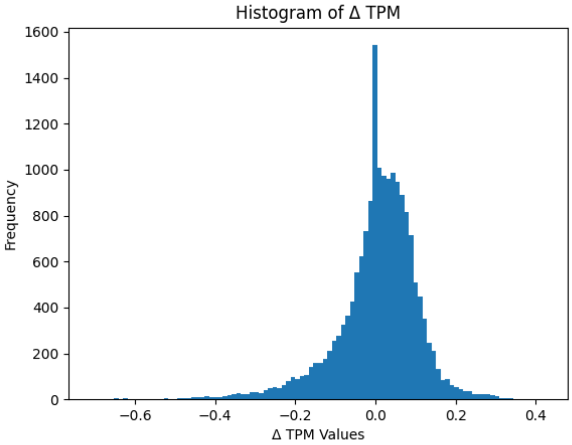


Figure 6: Distribution of Δ_{TPM} Values

We then define an Δ_{type} as either *increase* or *decrease*, depending on whether the delta value is positive or negative.

As a final step for our first objective 1, we need to determine if a change in gene activity is significant. To do this, we use the z score measure, which calculates how many standard deviations a delta TPM value is away from the mean of all delta TPM values. The *zscore* is given by:

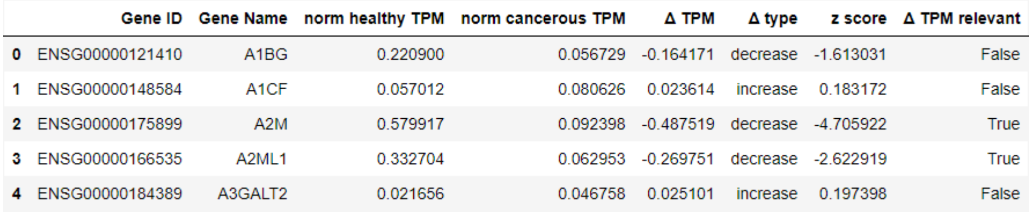
$$zscore(x) = \frac{x - \mu}{\sigma} \tag{2a}$$

$$\text{where } \mu = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2b}$$

$$\text{where } \sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2} \tag{2c}$$

where x is the delta TPM value, μ is the mean of all delta TPM values, σ is the standard deviation of all delta TPM values.

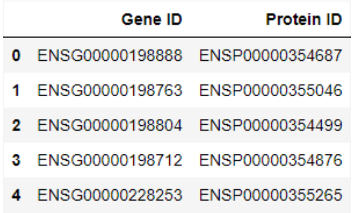
We define a threshold of $zscore = 1.96$ to indicate significant changes in gene activity, which corresponds to a confidence level of 95% (p = 0.05). Genes with delta TPM values exceeding this threshold will be flagged as *true* in the $\Delta_{TPM}relevant$ column. By this we fulfilled our first objective 1 by finding 1.034 genes with significant changes in gene activity.
The resulting table contains 17.626 Gene Nodes as rows with their associated attributes, including TPM values and derived metrics such as ΔTPM or $zscore$. The head of the table is shown in Figure 7.



	Gene ID	Gene Name	norm healthy TPM	norm cancerous TPM	Δ TPM	Δ type	z score	Δ TPM relevant
0	ENSG00000121410	A1BG	0.220900	0.056729	-0.164171	decrease	-1.613031	False
1	ENSG00000148584	A1CF	0.057012	0.080626	0.023614	increase	0.183172	False
2	ENSG00000175899	A2M	0.579917	0.092398	-0.487519	decrease	-4.705922	True
3	ENSG00000166535	A2ML1	0.332704	0.062953	-0.269751	decrease	-2.622919	True
4	ENSG00000184389	A3GALT2	0.021656	0.046758	0.025101	increase	0.197398	False

Figure 7: Example data of Gene Nodes Table

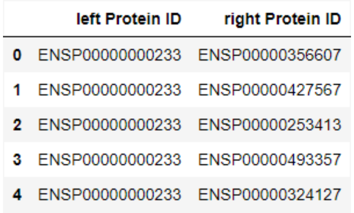
To construct the **gene-protein edges** we need a table that links the gene to the corresponding protein which is translated from the transcript of this gene. For this purpose we downloaded a file from biomart with Gene IDs and their Protein IDs. [LINK] First we filtered for a subset (Intersection) to only include rows where the Ensembl ID for the gene matched an existing gene node Since we have some genes without an entry for proteins, we need to drop them ; otherwise, there will not be an edge.
The final gene-protein edge table 8 features 101,731 rows as edges and two columns: Ensembl ID for the gene and Ensembl ID for the protein translation.



	Gene ID	Protein ID
0	ENSG00000198888	ENSP00000354687
1	ENSG00000198763	ENSP00000355046
2	ENSG00000198804	ENSP00000354499
3	ENSG00000198712	ENSP00000354876
4	ENSG00000228253	ENSP00000355265

Figure 8: Example data of Gene Protein Edges Table

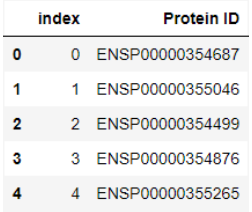
To create the **protein-protein edges** we download the String Database [LINK] with the information about the Protein-Protein Interaction.
The resulting file 9 consists of 13.715.404 rows of protein-protein edges with a column for both protein IDs for the edge.



	left Protein ID	right Protein ID
0	ENSP00000000233	ENSP00000356607
1	ENSP00000000233	ENSP00000427567
2	ENSP00000000233	ENSP00000253413
3	ENSP00000000233	ENSP00000493357
4	ENSP00000000233	ENSP00000324127

Figure 9: Example data of Protein-Protein Edges Table

We can generate the **protein nodes** from first ... To do this, we filter out the Gene colum from the previous file and check if there are any duplicate proteins. Since there are no duplicate proteins our data indicates that every protein is uniquely translated by a single gene. We do not need any additional attributes for these protein nodes because are focusing on the edges of this network.
The resulting table is a list of 104.235 unique Protein Ensembl IDs as shown in Figure 10.



	index	Protein ID
0	0	ENSP00000354687
1	1	ENSP00000355046
2	2	ENSP00000354499
3	3	ENSP00000354876
4	4	ENSP00000355265

Figure 10: Example data of Protein Nodes Table

3.3 Graph Database

As we have created our data models for the graph database, our next objective is to create a graph database that can be used to perform the PageRank algorithm as a graph algorithm. With four huge datasets at our disposal, optimizing the generation of the database is crucial.
In this section, we describe the Cypher queries used to create and populate our graph database, including information about creating indexes, constraints, relationships between nodes, and loading data into the database.

First, we create a basic PPI network by generating proteins and their interactions as edges. To start with, we need to create protein nodes with their respective properties. This process involves creating an index on the protein property id for faster query execution, loading the data as a list, creating nodes in batches for efficient processing, and finally populating the graph database with 101,731 protein nodes. We don't require any additional properties for these protein nodes since they are solely used as connections between genes. The Cypher query for creating protein nodes is:

```
CREATE (p:protein {id: ’Protein ID’})
```

Next, we create interaction edges between proteins by loading the data as a list of protein tuples, searching for both protein nodes by their IDs, and creating an edge in batches for efficient processing. This results in the creation of 11,247,242 edges between protein nodes. The Cypher query for creating protein-protein edges is:

```
MATCH (s:protein{id:’left Protein ID’})
MATCH (s:protein{id:’right Protein ID’})
CREATE (s)-[:INTERACTS]->(t)
```

At this point, we have a PPI network in place.

Our actual focus is on the genes. To get an extended network with genes on the edges of the protein nodes, we need to link them to the proteins. We start by creating gene nodes with their respective properties. This process involves creating an index on the gene property id for faster query execution, loading the data as a list, creating nodes in batches for efficient processing, and finally populating the graph database with 17,626 gene nodes.
The Cypher query for creating gene nodes is:

```
CREATE (p:gene {
       id: ’id’,
       gene_name: ’gene_name’,
       norm_healthy_tpm: ’norm_healthy_tpm’,
       norm_cancerous_tpm: ’norm_cancerous_tpm’,
       delta_tpm: ’delta_tpm’,
       delta_type: ’delta_type’,
       delta_tpm_relevant: ’delta_tpm_relevant’})
```

The genes have several properties, including *gene_name*, *norm_healthy_tpm*, *norm_cancerous_tpm*, *delta_tpm*, and *delta_type*. These attributes are used to calculate a more relevant attribute called *delta_tpm_relevant*, which will be the primary focus of our PageRank analysis. While the other attributes may not be directly relevant to our current analysis, they could potentially be useful for future tasks.

To represent relationships between genes and proteins, we create edges called connections by loading the data as a list of gene-protein tuples and matching both gene and protein nodes by their Ids. This process involves creating edges in batches for efficient processing, resulting in 101,731 connections between gene and protein nodes.

The Cypher query for creating gene-protein connection edges is:

```
MATCH (s:protein{id:'Protein ID'})
MATCH (s:gene{id:'Gene ID'})
CREATE (s)-[:CONNECTION]-(t)
```

By executing these Cypher queries, we can populate our graph database with the necessary nodes and edges to perform the PageRank algorithm.

As our final step for the second objective, we employed the **PageRank** algorithm to measure the importance of nodes in our graph database. This algorithm is particularly useful for identifying key genes that play a crucial role in the network because it can accurately capture the hierarchical structure of gene interactions, allowing us to identify not only highly connected genes but also those with significant functional relevance. For applying PageRank analysis to our large-scale graph database, we create a projection of the entire graph, which enables efficient computation and scalability.

To perform the PageRank analysis, we used the following Cypher query:

```
CALL gds.pageRank.stream('gene_protein_graph')
YIELD nodeId, score
RETURN gds.util.asNode(nodeId).id AS node,
      gds.util.asNode(nodeId).gene_name AS gene_name,
      score,
      gds.util.asNode(nodeId).delta_tpm AS delta_tpm,
      gds.util.asNode(nodeId).delta_tpm_relevant AS delta_tpm_relevant
ORDER BY score DESC
```

This query yielded a list of nodes, including genes and proteins, along with their respective PageRank scores. By dropping the proteins from this list, we focused on the genes and further filtered them by selecting only those with a significant change in gene activity, indicated by the value for *delta_tpm_relevant*.

The resulting list of genes represents those that have not only high connectivity within the network but also exhibit a substantial difference in gene expression. This subset was visualized using histograms, which demonstrate the distribution of PageRank scores for all genes (Figure 11) and only for relevant genes (Figure 12).

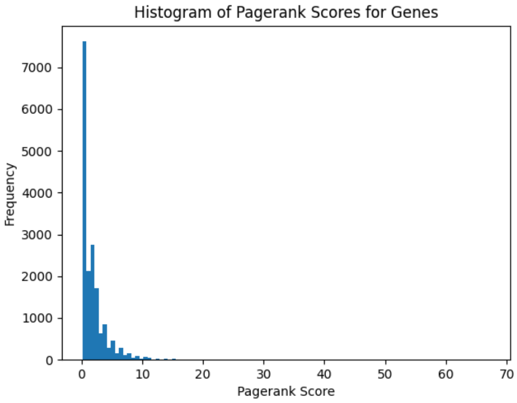


Figure 11: Histogram of Pagerank Scores for all genes

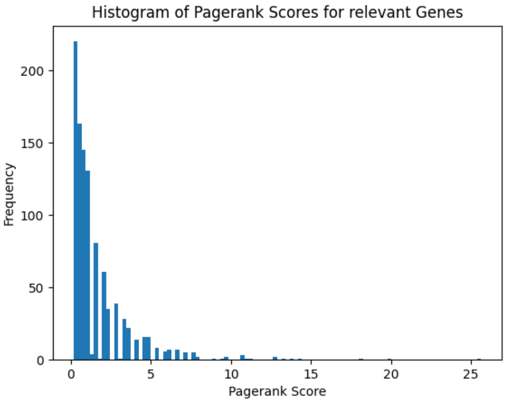


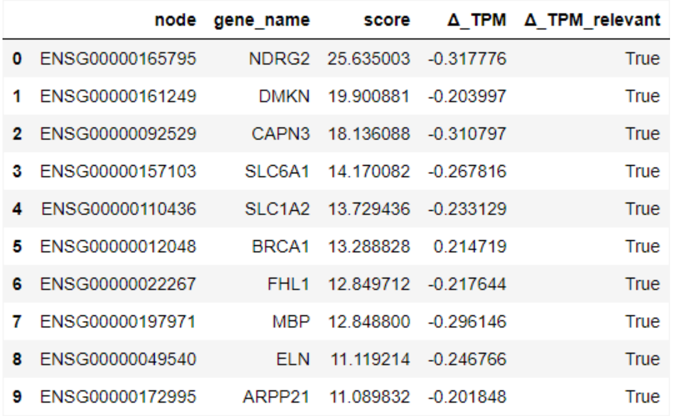
Figure 12: Histogram of Pagerank Scores only for relevant genes

Lastly we filtered the top 10 genes with the highest PageRank scores and a significant change in gene activity.

With this process, we have successfully done our second objective 2, which was to apply a graph algorithm to identify key genes in the network.

4 Results

We will provide an overview of the Top 10 genes found by the graph algorithm, which are crucial for understanding the complex interactions within the extended PPi network.
To identify the most significant genes, we extracted the top 10 genes with the highest PageRank scores from the list of relevant genes, as shown in Figure 13. These genes are not only highly connected within the network but also exhibit a substantial change in gene activity.



	node	gene_name	score	Δ_TPM	Δ_TPM_relevant
0	ENSG00000165795	NDRG2	25.635003	-0.317776	True
1	ENSG00000161249	DMKN	19.900881	-0.203997	True
2	ENSG00000092529	CAPN3	18.136088	-0.310797	True
3	ENSG00000157103	SLC6A1	14.170082	-0.267816	True
4	ENSG00000110436	SLC1A2	13.729436	-0.233129	True
5	ENSG00000012048	BRCA1	13.288828	0.214719	True
6	ENSG00000022267	FHL1	12.849712	-0.217644	True
7	ENSG00000197971	MBP	12.848800	-0.296146	True
8	ENSG00000049540	ELN	11.119214	-0.246766	True
9	ENSG00000172995	ARPP21	11.089832	-0.201848	True

))

Figure 13: Top 10 Genes with the highest PageRank scores and a significant change in gene activity

The key findings of this chapter include a list of 10 genes that are likely to be biomarkers for lung cancer.

5 Discussion

5.1 Analysis

5.2 What went well

5.3 Future Work

6 Conclusion

End

References

[1] Cmp - downloads - expression data. https://cellmodelpassports.sanger.ac.uk/downloads. Accessed: 09.10.2024.

[2] Ensembl biomarts. https://www.ensembl.org/biomart/martview/. Accessed: 14.10.2024.

[3] The ensembl project. https://www.ebi.ac.uk/training/online/courses/ensembl-browsing-genomes/what-is-ensembl/ensembl-project/. Accessed: 09.10.2024.

[4] Graph algorithms in neo4j: 15 different graph algorithms and what they do. https://neo4j.com/blog/graph-algorithms-neo4j-15-different-graph-algorithms-and-what-they-do/. Accessed: 23.09.2024.

[5] Gtex portal. https://www.gtexportal.org/home/. Accessed: 09.10.2024.

[6] Gtex portal - rna-seq. https://www.gtexportal.org/home/downloads/adult-gtex/bulk_tissue_expression. Accessed: 09.10.2024.

[7] J Ferlay, M Ervik, F Lam, M Laversanne, M Colombet, L Mery, M Piñeros, A Znaor, I Soerjomataram, and F Bray. Global cancer observatory: Cancer today. https://gco.iarc.who.int/today. Accessed: 27 September 2024.

[8] National Cancer Institute. Lung cancer—patient version. https://www.cancer.gov/types/lung. Accessed: 27.09.2024.

[9] National Cancer Institute. Risk factors for cancer. https://www.cancer.gov/about-cancer/causes-prevention/risk. Accessed: 27.09.2024.

[10] National Cancer Institute. What is cancer? https://www.cancer.gov/about-cancer/understanding/what-is-cancer, 2021. Accessed: 27.09.2024.

[11] World Health Organization. Cancer. https://www.who.int/news-room/fact-sheets/detail/cancer. Accessed: 27.09.2024.

[12] World Health Organization. Lung cancer. https://www.who.int/news-room/fact-sheets/detail/lung-cancer. Accessed: 27.09.2024.

[13] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.

[14] Trilochan Rout, Anjali Mohapatra, and Madhabananda Kar. A systematic review of graph-based explorations of PPI networks: methods, resources, and best practices. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 13(1):29–?, May 2024.

[15] Haixia Shang and Zhi-Ping Liu. Network-based prioritization of cancer genes by integrative ranks from multi-omics data. *Computers in Biology and Medicine*, 119:103692, 2020.

[16] Claire M Simpson and Florian Gnad. Applying graph database technology for analyzing perturbed co-expression networks in cancer. *Database*, 2020:baaa110, 12 2020.

[17] National Cancer Institute Surveillance Research Program. SEER*Explorer: An interactive website for SEER cancer statistics. https://seer.cancer.gov/statistics-network/explorer/. [Internet]. Updated: 2024 Jun 27; Cited: 2024 Sep 29. Available from: https://seer.cancer.gov/statistics-network/explorer/. Data source(s): SEER Incidence Data, November 2023 Submission (1975-2021), SEER 22 registries (excluding Illinois and Massachusetts). Expected Survival Life Tables by Socio-Economic Standards Accessed: 29.09.2024.