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

Simone Bergmann, 3957624
simone.bergmann@uni-bielefeld.de

October 13, 2024

Abstract

1 Introduction

In the Introduction, the motivation for the work is presented, followed by the goals of the thesis. The structure of the thesis is outlined and the limitations of the work are discussed.

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[5]. 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 [15], 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 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 high likelihood of being associated with lung cancer development or progression. To achieve this, we have three strategic objectives:

1. **Objective - Find significant changes in gene activity:** We will calculate the change in this measurement between healthy and cancerous tissues and define if this difference is significant to be able to later focus on these genes.
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 most 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 and tissue destruction. According to research, widespread metastases are the primary cause of death from cancer[9]. There are over 100 different types of cancer[8]. 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[5].
Cancer can be caused by a combination of genetic and environmental factors, such as diet, radiation, age, exposure to certain chemicals or viruses[7]. 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, characterized by uncontrolled cell proliferation and tissue destruction. There are two main subtypes: non-small cell carcinoma (NSCLC) and small cell carcinoma (SCLC)[6]. 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[10]. Early detection of lung cancer is essential for improving treatment success rates, emphasizing the need for prompt diagnosis and intervention.

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 a gene ID (Ensembl 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 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 connection between genes. We want to better understand the complex relationships between genes, and by using a graph database we can represent and query these 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. [3]

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:

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

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 [11], 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 chapter 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) [12] 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) [14] 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 Protein-Protein Interaction (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) [13] 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

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 a unique name for every gene, we use the Ensemble ID (ENS ID) as the primary key for the gene nodes. These are a unique erkennung 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 [2].

For the healthy tissue samples used in this study, we downloaded the `GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9` dataset from the GTEx portal [4]. 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 dataset is a .gct file 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 the data 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 (ENS ID, Tissue, TPM) 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 with 56,156 genes and a mean TPM value for every gene was saved as a CSV file for further processing.

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 [1], 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. [1] 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 (https://www.ensembl.org/biomart/martview/), which contains the gene stable ID (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 contains 33,502 genes with mean TPM values for lung cancer and was saved as a CSV file for further processing.

3.2 Nodes and Edges

Datasets to nodes and edges * genes as nodes * eigenschaft: name, gene stable id, norm healyth tpm, norm cancerous tpm, delta tpm, z score, delta tpm relevant * merging the GTEx (healthy) and CMP (cancerous) data to get a list of all genes that are in both datasets * filtering for genes that have a gene-protein connection * normalizing the TPM Values with log scaling * calculate a delta TPM value for every gene as the difference between the mean tpm value in the healthy and the cancerous dataset * NOT THE ABSOLUTE * to find out which genes have a relevant change in the healthy and cancerous tpm value we calculate the z score for the delta tpm values for every gene * z score is calculated as the difference between the delta tpm value and the mean delta tpm value divided by the standard deviation of the delta tpm values * z score is a measure of how many standard deviations an element is from the mean * XXX genes with a z score of 1,96 or higher are considered as delta tpm relevant (p Wert damit 0,05 and Konfidenzniveau 95→ 17.626 gene Nodes
* protein-gene connection as edges * eigenschaft: gene stable id, protein stable id * proteins are needed as a second node type to get a connection between the genes. * downloaded protein file from biomart with gene stable ID and protein stable ID * filtered for those gene stable IDs that are in the gene node list. * we only have those proteins that have a connection (was sagt die tabelle genau aus?) → 101.731 protein gene edges (means: every protein is connected to exactly one gene, but every gene can have multiple proteins)
* proteins as nodes * eigenschaft: protein stable id * use the protein gene table as base * only use the column with the protein stable ID → 101.731 protein Nodes
* protein protein connection as edges * eigenschaft: protein stable id 1, protein stable id 2 * downloaded protein protein interaction file from STRING database * filtered for those protein stable IDs that are in the protein node list → 11.247.242 protein protein edges (not every protein is connected to another one → but since every gene could have multiple proteins, the connected gene may have a connection to another protein) TODO: are those genes that are not connected to another gene relevant for the analysis?

3.3 Graph Database

Graph database * Neo4J * gene nodes query: * protein nodes query: * gene protein edges query: * protein protein edges query:

4 Results

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

[3] 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.

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

[5] 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.

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

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

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

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

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

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.