Final Project Report

CS-579: ONLINE SOCIAL NETWORK ANALYSIS

Project Name-
Network Analysis of Protein-Protein Interactions (PPIs) in Non-Small Cell Lung Cancer (NSCLC)

By

Lakshmi Sindhu Pulugundla - A20553567

Under the Guidance

Of

Dr Cindy Hood

Professor CS-579

College of Science

1. Introduction

Lung cancer remains one of the most challenging diseases to treat, with Non-Small Cell Lung Cancer (NSCLC) accounting for approximately 85% of all cases. This type of cancer is particularly notorious for its late diagnosis and poor prognosis, driving a critical need for better therapeutic strategies and a deeper understanding of its biological underpinnings. Advances in bioinformatics and systems biology have provided new tools to tackle these challenges, enabling researchers to explore the vast networks of protein interactions that govern cellular functions in cancer cells.

This project was initiated against this backdrop, with the primary goal of dissecting the complex protein-protein interaction (PPI) landscape specific to NSCLC. Utilizing the STRING database—a rich repository of known and predicted protein interactions—the project aims to construct a detailed network map of NSCLC. By identifying key proteins and pathways within this network, the research seeks to pinpoint potential targets for therapeutic intervention, which could lead to the development of more effective treatments.

The specific objectives of the project were to:

1. **Construction of the NSCLC-specific PPI Network:** Develop a comprehensive map that details the proteins involved in NSCLC and their interconnections. This network serves as the foundation for all subsequent analyses, providing the structural framework necessary to understand the complex interactions at play.
2. **Identification of Key Proteins and Pathways:** Through network analysis, identify central proteins that may play pivotal roles in NSCLC pathogenesis. These proteins, often hubs in the network, are potential candidates for targeted drug development due to their strategic positions within regulatory pathways.
3. **Analysis of Network Dynamics:** Investigate how alterations within the network could influence the progression of NSCLC and the efficacy of therapeutic interventions. This objective involves simulating network changes and predicting potential outcomes, providing insights into the network's resilience and vulnerability.

The link for the project - CSE-579 Final Project - PPI Non-Small Cell Lung Cancer.ipynb - Colab (google.com)

2. Data

The cornerstone of this project is the data derived from the STRING database, known for its extensive compilation of protein interaction data. This section details the datasets used, the process of merging these datasets, and the validation methods employed to ensure data integrity. The data was extracted from the following URL - [64 items (human) - STRING interaction network (string-db.org)]

Dataset Acquisition and Preparation

Data integrity and precise data handling are crucial in bioinformatics projects. For this project, we utilized four primary datasets from the STRING database known for its rich collection of protein interaction information. These datasets were:

1. **Protein Links Dataset(string_interactions.tsv):** This dataset forms the backbone of our PPI network, encompassing both known and predicted protein interactions, each annotated with confidence scores to reflect their reliability.
2. **Protein Aliases Dataset(string_protein_annotations.tsv):** It provides alternative identifiers and names for proteins, crucial for ensuring consistency and interoperability across different biological databases and literature.
3. **Protein Functional Annotations Dataset(string_funtional_annotations.tsv):** This dataset offers detailed descriptions of protein functions and roles, which are essential for understanding their specific involvements in NSCLC processes.
4. **Node Degrees Dataset(string_node_degrees.tsv):** Contains information about the number of connections each protein has within the network, helping to identify highly connected 'hub' proteins that may be of particular importance in NSCLC.

These datasets were meticulously prepared and validated to ensure that they met the research standards required for a robust analysis. Data from various sources were carefully harmonized, focusing on standardizing identifiers to mitigate discrepancies during the data integration phase.

Merging Process and Validation

a) Merging Process

The merging process was critical for creating a unified dataset that could be used for comprehensive analysis:

1. **Standardization of Protein IDs:** The first step involved aligning protein IDs from the Protein Links Dataset with names from the Protein Aliases Dataset. This standardization was crucial for subsequent data integration.
2. **Integration of Functional and Pathway Data:** Functional annotations and pathway information were then merged based on the standardized protein IDs, creating a multi-dimensional dataset that supports varied analyses.

b) Validation of Merging Process

Ensuring the accuracy and integrity of the merged dataset was accomplished through several validation steps:

1. **Statistical Consistency Checks:** To ensure that the merging process did not introduce data loss or duplication, we conducted detailed statistical analyses.

2. **Manual Sampling for Accuracy:** Random samples from the dataset were manually checked against external databases to verify the accuracy of the merging process.

Data Challenges and Solutions

Throughout the data preparation phase, I encountered several challenges:

1. **Discrepancies in Protein Identifiers:** Different datasets often use varied naming conventions or identifiers for the same proteins. To address this, a mapping system was developed to standardize protein IDs across all datasets, ensuring seamless integration.
2. **Merging the Datasets:** Merging datasets, making sure that the merging process was successful, and no redundant values have been created, was another task.
3. **Handling Large Data Volumes:** The initial datasets were extremely large, making processing and analysis computationally intensive. To manage this, data was pre-processed in a high-performance computing environment, and analyses were streamlined to focus on the most relevant subsets of data.

These issues were addressed through custom solutions, including the development of a tailored ID mapping system and the application of advanced imputation methods, ensuring a robust dataset ready for detailed analysis.

Methodological Adjustments

Adjustments were also made in response to initial findings and data exploration insights:

1. **Refinement of Data Extraction Criteria:** Initially broad criteria were refined to focus more specifically on clinically relevant proteins and their interactions, based on emerging research and preliminary data analyses.
2. **Enhanced Data Validation Techniques:** Additional validation steps were introduced, including more comprehensive manual checks and cross-referencing against updated external databases, to further ensure the accuracy of the data used in the analyses.

3. Project Process - Detailed Methodological Approach

a) Data Acquisition and Cleaning:

The project began with the acquisition of extensive datasets from the STRING database, focusing on protein-protein interactions pertinent to Non-Small Cell Lung Cancer (NSCLC). These datasets were loaded using Python, which provides robust libraries for handling large datasets. Data cleaning involved standardizing data formats, correcting erroneous entries, and pruning unnecessary columns. This meticulous approach ensured the datasets were optimized for high-performance data manipulation and analysis.

b) Data Preparation:

i. Exploring Loaded Datasets:

Each dataset, including protein interactions, functional annotations, and node degrees, was initially explored to understand the data structure and content. Key adjustments included renaming columns that started with a '#' for better accessibility.

ii. Merging Datasets:

Datasets were merged carefully to maintain data integrity:

a. **Node Degrees:** Nodes from the interaction dataset were merged with degree information to enrich the interaction data with connectivity insights. Validation checks confirmed the accuracy of the merges.

b. **Functional Annotations:** Protein functionality data from KEGG pathways was integrated, focusing specifically on pathways relevant to NSCLC. This step was crucial for attributing biological significance to the proteins within the network.

c. **Protein Annotations:** Annotations providing additional details about each protein were merged, enhancing the dataset with deeper biological insights.

After merging, unnecessary columns were dropped, and column names were standardized to ensure consistency and clarity across the dataset.

iii. Preliminary Data Exploration:

The merged dataset underwent initial exploration to identify key statistical features and distributions, setting the stage for more focused analyses.

iv. Extracting KEGG Functionalities:

Given the extensive size of the dataset, a targeted approach was taken to extract functionalities related to NSCLC, streamlining the data for more efficient analysis.

c) Network Construction and Analysis:

Using the cleaned and filtered data, a network was constructed to map the protein interactions specific to NSCLC. This network served as the basis for all subsequent analytical explorations. The network construction incorporated protein interactions as edges and proteins as nodes, with edge weights derived from the combined scores of interactions to reflect their strength and reliability.

d) Network Analysis:

Network analysis involved calculating various centrality measures to identify key proteins that play pivotal roles in the network. This analysis helped in pinpointing

proteins that might be crucial for NSCLC pathogenesis and could serve as potential targets for therapeutic intervention.

e) Visualisation:

Visualization played a crucial role in this project, utilizing Matplotlib and Seaborn for generating histograms and boxplots to understand distributions and NetworkX for creating detailed network graphs. These visualizations provided intuitive insights into the complex relationships and key structures within the NSCLC network. The network visualizations highlighted important proteins and interactions, using node sizes and colors to represent different centrality measures and interaction strengths.

4. Project Results

Project results include successful retrieval of top 10 proteins in the dataset containing protein-protein interactions in NSCLC based on –

a) Degree Centrality - Proteins with high degree centrality are likely essential for cellular processes since they interact with many other proteins. They might be involved in key pathways or crucial for maintaining network stability.

```
Top 10 central nodes in the network based on Degree
Centrality:
        TP53: 0.8805970149253731
        AKT1: 0.8059701492537313
        KRAS: 0.7761194029850746
        EGFR: 0.7313432835820896
        STAT3: 0.7164179104477612
        PIK3CA: 0.7014925373134329
        HRAS: 0.6716417910447761
        NRAS: 0.6567164179104478
        PIK3R1: 0.6567164179104478
        MAPK3: 0.6417910447761194
```

b) Closeness Centrality - Proteins with high closeness centrality can quickly interact or influence other proteins across the network, potentially indicating their role in signal transduction or their efficiency in propagating information through a cellular network.

```
Top 10 central nodes in the network based on Closeness
Centrality:
        TP53: 0.8933333333333333
        AKT1: 0.8375
        KRAS: 0.7976190476190477
        EGFR: 0.788235294117647
        STAT3: 0.7790697674418605
        PIK3CA: 0.7528089887640449
        PIK3R1: 0.7444444444444445
        MAPK3: 0.7362637362637363
        HRAS: 0.7362637362637363
        CDKN2A: 0.7362637362637363
```

c) Betweenness Centrality - Proteins with high betweenness centrality might control the flow of information or substances within the cell, acting as gatekeepers. They

could be potential drug targets because their removal or modification can significantly impact the network's function.

```
Top 10 central nodes in the network based on Betweenness
Centrality:
      TP53: 0.12916932121438954
      AKT1: 0.055528726965308785
      STAT3: 0.04100007247955556
      KRAS: 0.0309359790902557
      RXRA: 0.030673222701716598
      EGFR: 0.029874041443926715
      CDKN2A: 0.029363560950295995
      RARB: 0.028778738321126294
      CDKN1A: 0.024094675971163222
      MAPK1: 0.0226735080644186
```

d) Eigenvector Centrality - This centrality measure helps identify not just well-connected proteins, but those whose connections are also influential. Such proteins might be crucial for the robustness of cellular responses and are often core components of critical pathways.

```
Top 10 central nodes in the network based on Eigenvector
Centrality:
      TP53: 0.19964673463259303
      KRAS: 0.19855236428641035
      AKT1: 0.19643618667712515
      PIK3CA: 0.18949036594556934
      EGFR: 0.18838294417984072
      STAT3: 0.18318004245950548
      PIK3R1: 0.18063498415094545
      HRAS: 0.1793816799120391
      NRAS: 0.17637835795329168
      ERBB2: 0.16941476064804914
```

5. Key Findings and Graphical Analysis

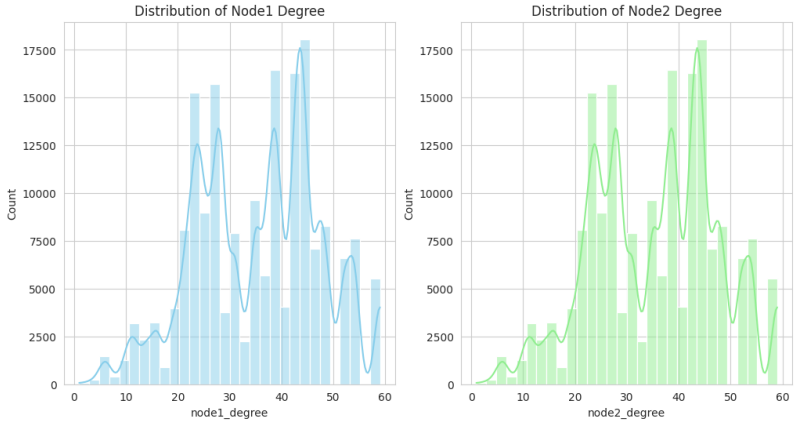
Central Proteins in NSCLC Protein-Protein Interaction Network:

Our network analysis, leveraging various centrality measures, has identified several key proteins that hold significant roles within the Non-Small Cell Lung Cancer (NSCLC) interaction network. These proteins, due to their strategic positions and roles within the network, are crucial for understanding the underlying mechanisms of NSCLC and can be potential targets for therapeutic intervention.

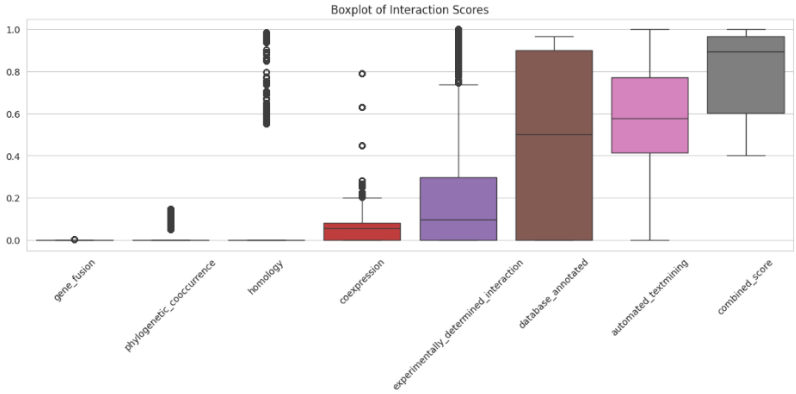
1) **Degree Centrality Findings:**
   Proteins such as TP53, AKT1, and KRAS exhibited the highest degree centrality, indicating their extensive interaction with numerous other proteins. This suggests their pivotal roles in essential cellular functions, including cell growth, division, and apoptosis. Their high connectivity highlights their potential as critical hubs in cellular pathways, making them prime targets for drug development.
2) **Closeness Centrality Insights:**
   Proteins like EGFR and PIK3CA displayed significant closeness centrality, emphasizing their ability to quickly affect or be affected by other proteins across the network. This efficiency in signal propagation underscores their importance in rapid response mechanisms and cellular communication, crucial in cancer pathology.
3) **Betweenness Centrality Analysis:**
   With high betweenness centrality, proteins such as TP53 and EGFR serve as major points of control within the network, regulating the flow of molecular signals across different pathways. Their strategic position as bridges in the network indicates their role in maintaining cellular integrity and responding to cellular stress, making them vital for maintaining network stability.
4) **Eigenvector Centrality Results:**
   Proteins like AKT1 and KRAS, which have high eigenvector centrality, are not only well-connected but are also linked to other influential proteins. This centrality measure reflects their critical influence on network dynamics and their involvement in robust pathways essential for cancer development and progression.
5) **Key Proteins and Their Functional Implications:**
   - **TP53:** Often cited across various centrality measures, TP53's role as a tumor suppressor is crucial in DNA repair and apoptosis, pivotal in cancer progression and response to therapy.
   - **AKT1:** Central in cell survival pathways, its involvement in multiple signaling pathways makes it a valuable target for therapies aimed at regulating cell growth and survival.
   - **KRAS:** Known for its role in signal transduction, mutations in KRAS are particularly significant in NSCLC, influencing treatment outcomes and disease progression.
   - **EGFR:** A receptor tyrosine kinase involved in the regulation of cell growth and survival, commonly targeted by existing therapies due to its role in cancer cell proliferation.

6) Graphs :

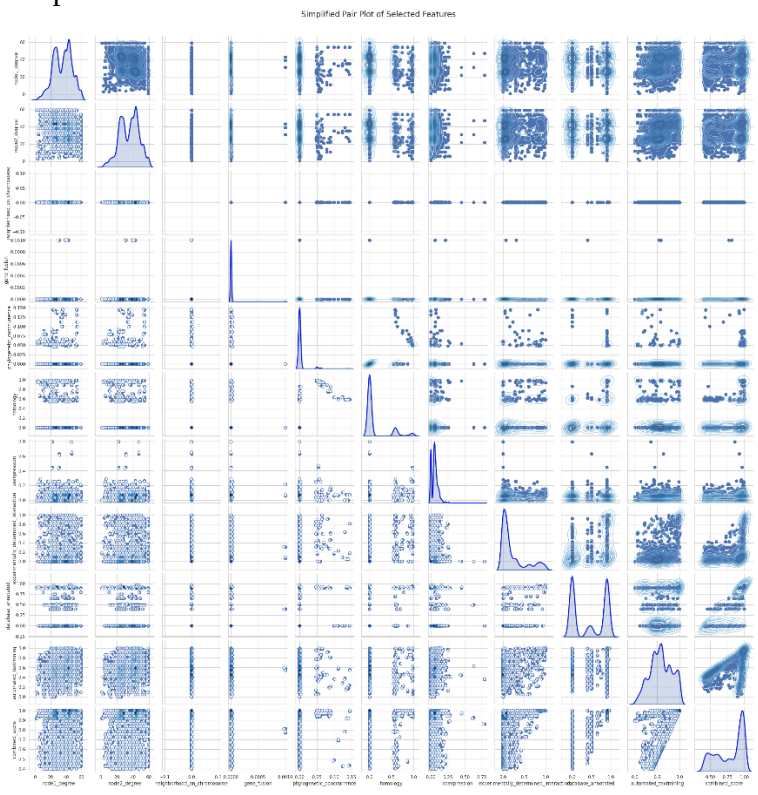
- Node Degrees Histogram -



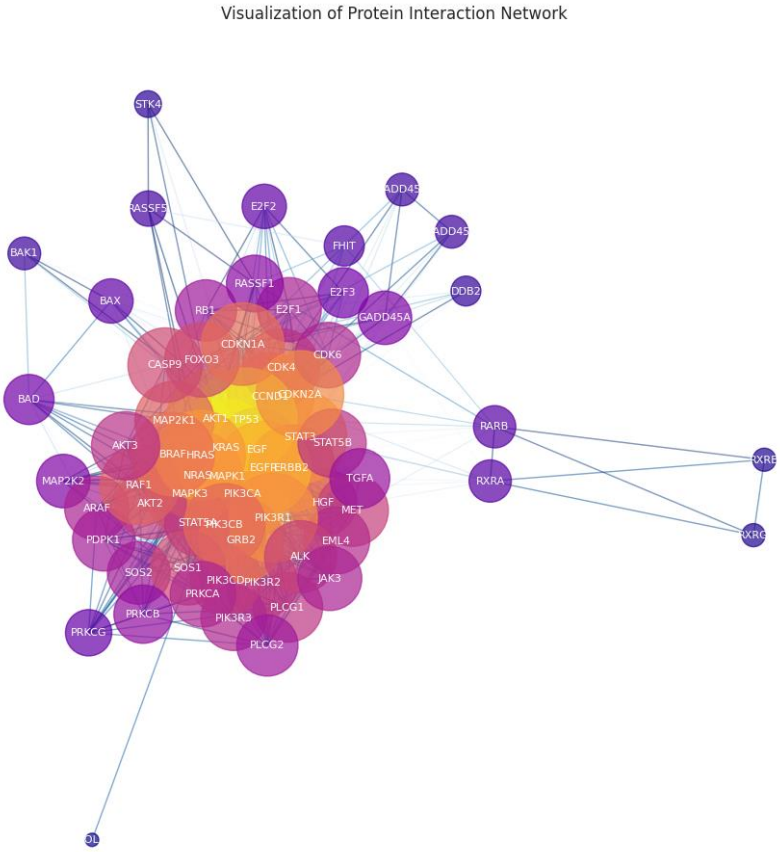
- Boxplot of Interaction Scores -



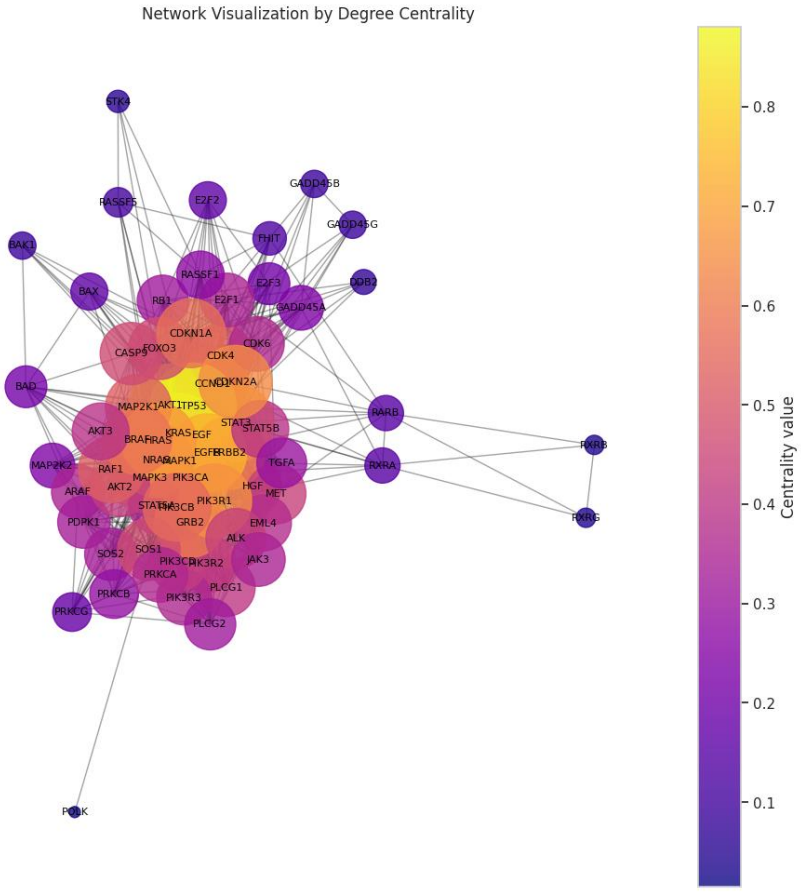
- Pairplot -



- Network Graph –



- Degre Centrality Visualisation –



6. Discussion of Project

After spending time on the project, I found that it was a valuable learning experience. It was insightful to see how network analysis can illuminate key proteins involved in Non-Small Reflecting on the project, I believe it was an enriching experience that offered a deep dive into the complexities of protein-protein interactions within Non-Small Cell Lung Cancer (NSCLC). The project was successful in employing network analysis to identify key proteins that could serve as potential therapeutic targets, demonstrating the utility of bioinformatics in advancing our understanding of cancer biology.

If I were to undertake this project again, I would enhance the initial stages of data handling. Specifically, I would implement more robust data validation techniques to ensure the accuracy of the protein interaction data before proceeding with analyses. This could involve more comprehensive cross-referencing with additional databases to confirm the identified interactions and protein functions. Furthermore, I would allocate more time to exploratory data analysis to identify more nuanced patterns and relationships in the data, which might have been overlooked. Integrating more advanced machine learning techniques to predict potential interactions or to classify proteins based on their network characteristics could also deepen the insights gained.

From this project, I've learned the critical importance of a well-structured approach to handling large datasets, particularly in a field as complex as genomics. The project underscored the necessity of meticulous data preparation in bioinformatics, where data integrity directly impacts the quality of the findings. Moreover, the project reinforced my understanding of how computational tools can be effectively applied to biological data, providing actionable insights that can drive scientific inquiry and potential medical breakthroughs. This experience has not only enhanced my technical skills but also enriched my appreciation for interdisciplinary collaboration in tackling challenging scientific questions.

7. Future Work

Looking forward, the project opens several avenues for further research:

1. **Dynamic Network Analysis:** Future work could involve the dynamic simulation of the PPI network to study how perturbations in one part of the network affect the overall network behavior. This could be particularly useful for understanding the impact of specific drugs on protein interactions and identifying potential side effects.
2. **Integration of Genomic Data:** Integrating patient-specific genomic data could help in personalizing the network analysis, potentially leading to personalized treatment strategies based on individual protein interaction profiles.
3. **Experimental Validation:** While computational analysis provides valuable insights, experimental validation of key findings is essential. Future work could focus on experimentally verifying the role of the identified key proteins in NSCLC, potentially leading to the development of targeted therapies.

If I had more time, I would have liked to explore the proteins more by performing sub-graph analysis on the top 10 proteins I managed to extract from the dataset using various centrality measures. This focused analysis would allow for a deeper look into the specific roles these key proteins play in NSCLC. By examining the interactions within this sub-group, we could gain valuable insights into their involvement in critical pathways and processes. Such an analysis would simplify the complex network, making it easier to identify potential targets for therapy and understand their impact on disease progression and treatment outcomes.

Further, by concentrating on these central proteins, we could enhance the efficiency of our computational and experimental approaches, allowing for more precise modeling and hypothesis testing. This could potentially lead to the discovery of novel therapeutic targets and help in designing drugs that might interact with these proteins to mitigate the severity of NSCLC. Additionally, this targeted analysis could facilitate the development of personalized medicine approaches, tailoring treatments based on the specific protein interactions and pathways active in individual patients. Ultimately, this could lead to more effective interventions, improved patient outcomes, and a better understanding of the disease at a molecular level.

8. References

- **Data Sources:** 64 items (human) - STRING interaction network (string-db.org)
- **Literature Survey:**
  - 64 items (human) - STRING association evidence view (string-db.org)
  - 64 items (human) - STRING association evidence view (string-db.org)
  - KEGG PATHWAY: map05223 (genome.jp)
  - K-ras Mutations in Non-Small-Cell Lung Carcinoma: A Review - Clinical Lung Cancer (clinical-lung-cancer.com)
  - Molecular Pathology of Non-Small-Cell Lung Cancer | Respiration | Karger Publishers
  - Oncogenic Pathways, Molecularly Targeted Therapies, and Highlighted Clinical Trials in Non–Small-Cell Lung Cancer (NSCLC) - Clinical Lung Cancer (clinical-lung-cancer.com)