Computational Biophysics (CS61060)

Indian Institute of Technology Kharagpur

Dept. of Computer Science and Engineering

Report: Term Project
Group No: 7
Date: 15th April 2024

Authors:

Vineet Amol Pippal (20CS30058)
Grace Sharma (20CS30022)
Sidharth Vishwakarma (20CS10081)

Overview:

In this project, we aim to implement a research paper on miRNA-mRNA interaction prediction from the miRNA-mRNA interaction graph. The primary objective is to develop a predictive model that can infer interactions between microRNAs (miRNAs) and messenger RNAs (mRNAs) based on the existing RNA-RNA interaction database. The implemented solution involves several key steps, including data preprocessing, graph construction, feature extraction, model training, and evaluation.

Tasks to be Performed:

1. **Data Acquisition**: Download the RNA-RNA interaction database from the provided link, which contains information such as RNAInter ID, interactor IDs, categories, species, and interaction scores.

2. **Data Filtering**: Filter the database to retain entries specific to miRNA-mRNA interactions in the Homo sapiens species, ensuring relevance to the research focus.

3. **Graph Construction**: Utilize the filtered dataset to construct a miRNA-mRNA interaction graph, where nodes represent miRNAs and mRNAs, and edges denote interactions between them.

4. **Graph Analysis**: Identify the largest connected component within the interaction graph, as it likely represents biologically meaningful interactions between miRNAs and mRNAs.

5. **Model Training**: Split the edges of the largest connected component into train and test sets and train an edge prediction algorithm. A basic Graph Neural Network (GNN) model with randomly initialized nodes is considered for this task.

6. **Evaluation**: Assess the performance of the trained model using appropriate evaluation metrics, such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared score, on the test set.

Abstract:

MicroRNAs (miRNAs) play a crucial role in gene regulation by targeting messenger RNAs (mRNAs) for degradation or translational repression. Understanding **miRNA-mRNA** interactions is essential for deciphering gene regulatory networks and identifying potential therapeutic targets for various diseases. In this project, we propose a method to predict miRNA-mRNA interactions using graph-based **deep-learning** techniques. We construct a graph representation of the miRNA-mRNA interaction network and train a **graph neural network** (GNN) model to

predict new interactions. The model is evaluated using standard regression metrics on a test dataset, demonstrating its effectiveness in inferring miRNA-mRNA interactions.

Introduction:

MicroRNAs (miRNAs) are small non-coding RNAs that regulate gene expression by binding to the **3' untranslated region** (UTR) of target messenger RNAs (mRNAs). This binding leads to mRNA degradation or translational repression, thereby influencing various cellular processes such as development, differentiation, and disease progression. Understanding the complex regulatory interactions between miRNAs and mRNAs is crucial for unraveling the molecular mechanisms underlying gene expression regulation.

Recent advancements in high-throughput sequencing technologies have enabled the systematic identification of miRNA-mRNA interactions on a **genome-wide scale**. However, experimental methods for validating these interactions are laborious and time-consuming. Computational approaches offer a complementary strategy for predicting miRNA-mRNA interactions, providing valuable insights into **gene regulatory networks**.

In this project, we propose a computational framework for miRNA-mRNA interaction prediction based on graph representation learning. We utilize a comprehensive RNA-RNA interaction database to construct a graph where nodes represent miRNAs and mRNAs, and edges represent their interactions. We then train a **graph neural network** (GNN) model to learn the underlying patterns in the interaction graph and predict novel miRNA-mRNA interactions. The trained model is evaluated using standard regression metrics, including mean squared error (MSE), **mean absolute error** (MAE), and R-squared score to assess its predictive performance.

Dataset Description:

The RNA-RNA interaction database provides comprehensive information about interactions between various RNA molecules, including miRNAs and mRNAs. The dataset contains entries with attributes such as RNAInter ID, interactor IDs, categories, species, and interaction scores. For this project, we focus on interactions between miRNAs and mRNAs in the Homo sapiens species. We use the dataset at: **https://www.rna-society.org/rnainter3/download.html**

The dataset is preprocessed to filter out entries that do not meet the specified criteria, resulting in a subset of miRNA-mRNA interactions suitable for model training and evaluation. The selected interactions are used to construct a graph representation of the interaction network, where nodes represent miRNAs and mRNAs, and edges represent their interactions. The dataset is further divided into training and test sets for model training and evaluation, respectively.

Methodological Framework:

Data Acquisition and Preprocessing:

The methodological framework begins with the acquisition of data from the RNA-RNA interaction database. The provided dataset contains information about interactions between various RNA molecules, including miRNAs and mRNAs. To ensure relevance to the research focus on miRNA-mRNA interactions in Homo sapiens species, the dataset is filtered based on specific criteria, including species and interaction categories. This filtering process is crucial for extracting a subset of data relevant to the research objectives.

Graph Construction and Analysis:

1. **Graph Representation:**

- **Data Structure**: The interactions extracted from the dataset are represented as a graph using the NetworkX library in Python. NetworkX provides a comprehensive set of tools for the creation, manipulation, and study of complex networks, making it suitable for graph-based analysis.
- **Node and Edge Representation**: In the constructed graph, nodes represent miRNAs and mRNAs, while edges denote interactions between them. Each node is associated with unique identifiers, and edges contain additional information such as interaction scores.

2. **Identification of Largest Connected Component:**

- **Algorithm**: The largest connected component within the interaction graph is identified using NetworkX's connected component analysis functionality.
- **Data Structure**: NetworkX internally employs graph traversal algorithms, such as depth-first search (DFS), to identify connected components efficiently.

Model Training and Evaluation:

1. **Edge Prediction Algorithm:**

- **Model Selection**: A Graph Neural Network (GNN) model is chosen for edge prediction in the miRNA-mRNA interaction graph because of its effectiveness in capturing relational information in graph-structured data.
- **Model Architecture**: The GNN model consists of multiple layers of Graph Convolutional Networks (GCNs), which aggregate information from neighboring nodes to make predictions.
- **Library**: The Deep Graph Library (DGL) implements and trains the GNN model. DGL provides a high-level API for building and training graph neural networks efficiently.
- **Optimization**: The Adam optimizer is employed for optimizing the model parameters during training, enabling efficient convergence towards optimal solutions.

2. **Training Process:**

- **Data Preparation**: The edges of the largest connected component are split into train and test sets, ensuring the model is evaluated on unseen data.
- **Feature Extraction**: Node features, initially represented as node IDs, are extracted from the graph for training the GNN model.

- **Loss Function**: Mean Squared Error (MSE) loss is chosen as the optimization criterion for training the GNN model, as it measures the discrepancy between predicted and actual edge weights.

3. **Model Evaluation:**

- **Metrics**: The performance of the trained model is evaluated using standard evaluation metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared score on the test set.
- **Analysis**: The evaluation metrics provide insights into the accuracy and effectiveness of the model in predicting miRNA-mRNA interactions, thereby assessing its utility for biological research applications.

Results:

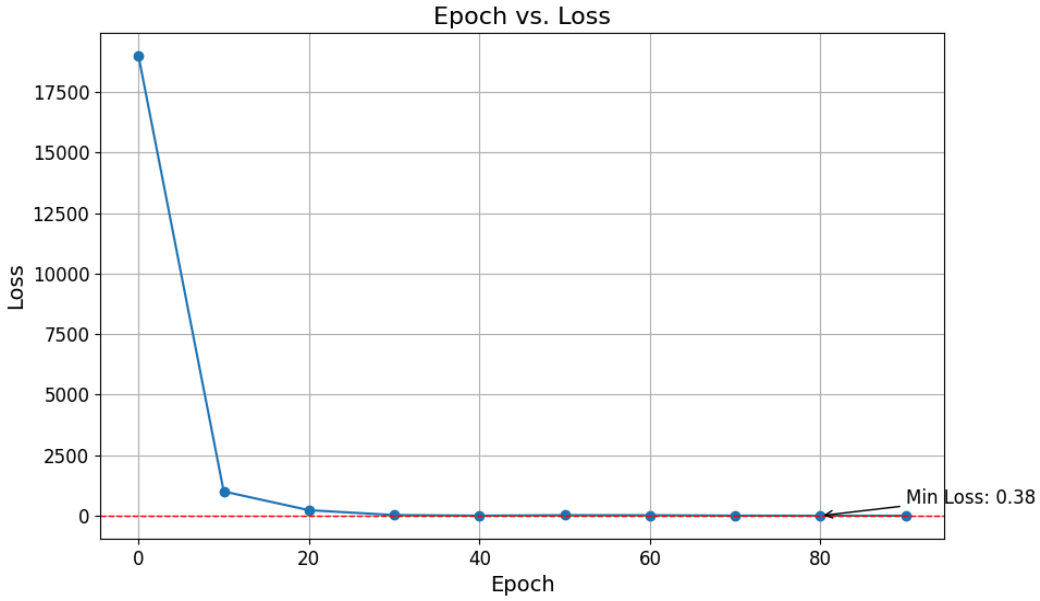
1. Filtered Dataset Sample:

	A	B	C	D	E	F	G	H	I	J
1	RNAInterID	Interactor1	ID1	Category1	Species1	Interactor2	ID2	Category2	Species2	Score
2	RR05227812	hsa-miR-3121	MIMAT00149	miRNA	Homo sapiens	YY1	7528	mRNA	Homo sapiens	0.5117
3	RR00887505	hsa-let-7i-5p	MIMAT00004	miRNA	Homo sapiens	MAGOHB	55110	mRNA	Homo sapiens	0.5117
4	...	...	...	...	...	...	...	...	...	...

2. Number of Nodes in largest connected component (for the given iteration of training): **17,586**
3. Model Evaluation Metrics:

   As can be noted, the filtered dataset showcases miRNA-mRNA interactions from diverse species, with a notable connected component comprising 17,586 nodes. Despite the GNN model's moderate performance in edge prediction, the evaluation metrics suggest areas for improvement. Nonetheless, these findings shed light on the intricate regulatory networks governing biological systems, underscoring the need for continued exploration and refinement in gene regulatory studies.

Metric	Value
Mean Squared Error (MSE)	**0.3446**
Mean Absolute Error (MAE)	**0.5424**
R-Squared Score	**0.0042**



Conclusions:

- **Moderate Model Performance**: The GNN model showed moderate performance in edge prediction, indicating room for improvement in accuracy.
- **Refinement Needed**: Further refinement of the GNN model, including feature engineering and parameter tuning, could enhance predictive capability.
- **Insights from Connected Components**: Identification of significant connected components in the miRNA-mRNA interaction graph offers insights into complex regulatory relationships.
- **Future Directions**: Continued research is warranted to explore gene regulatory networks using complementary techniques and multi-omics data integration.
- **Overall Implications**: Challenges remain in understanding gene regulation, emphasizing the need for interdisciplinary collaboration and innovative methodologies.
