GNNs (Graph Neural Networks) to predict Protein functions

SUMMER INTERNSHIP

Submitted by -

- ANISHA RAJ
- PRAJUKTA DEY

Submitted to -

- KIRTI KUMARI
- RICHA DHANUKA

Github Repository :

https://github.com/prajuktadey/Protein-Function-Prediction

Introduction :

Proteins are the vital organic macromolecules comprising twenty standard amino acids that play a critical role in numerous biological processes and cellular functions within organisms. The prediction of their functions represents a challenging yet fundamental task in the field of bioinformatics. Graph Neural Networks (GNNs) have recently emerged as a promising tool for analysing graph-structured data, rendering them well-suited for predicting protein functions based on the intricate interplay of protein networks and sequence-structure links.

The biological realm relies extensively on proteins to execute a wide spectrum of functions, ranging from DNA transcription and replication to hormone regulation, metabolism, molecular cell signalling, and signal transduction. These functions are deeply intertwined with protein interactions, specifically protein-protein interactions (PPIs), whereby proteins rarely act in isolation but rather engage with other proteins in their vicinity. This complex interplay of proteins is fundamental for various biological activities, underscoring the importance of understanding protein-protein interactions.

Computational methods have emerged as valuable tools to predict PPIs, offering a cost-effective and less resource-intensive alternative to experimental approaches. While earlier research primarily concentrated on utilising sequence information for PPI prediction, this study takes a more comprehensive approach. By employing graph convolutional networks (GCNs) and graph attention networks (GATs), we combine protein structural information and sequence features to predict protein interactions effectively.

Leveraging both structural and sequence data through GNNs, this pioneering methodology offers a promising pathway for predicting protein functions and unravelling the intricate protein interactions within biological systems.

Problem Statement :

Predicting protein functions is a challenging task due to the complexity of protein interactions. The objective is to develop a computational model that accurately predicts protein functions by utilizing both structural and sequence data.

Dataset :

In this study, we utilised protein-protein interaction (PPI) datasets, specifically the Biological Process (BP) and Molecular Function (MF) datasets, for our analysis.
The statistics of these datasets can be referred from Drive Link.

Graph :

In graph theory, a graph G is defined as a triple: \[G = (V, E, A) \]
Where:

- \(V \) represents a collection of nodes (or vertices): { \(v_1, v_2, \ldots, v_n \) }
- \(E \) represents a collection of edges: { \(e_1, e_2, \ldots, e_m \) }
- \(A \) represents the topological structure by defining the adjacency matrix, which indicates the connections between nodes in the graph.

Graphs provide a versatile framework to model and analyze relationships between entities, and they are essential in various domains, especially for applications like molecular structure analysis in the context of Graph ML.

To represent proteins in a graph structure, we construct graphs using Protein Data Bank (PDB) files, which contain precise 3D coordinates of atoms. These resulting protein graphs illustrate the amino acid network or residue contact network, with each node denoting a residue. Nodes are connected based on pairs of atoms falling within a specific threshold distance. Extracting features for each node or residue involves utilising a protein language model, wherein the protein sequence serves as input to generate feature vectors for each amino acid.

The effectiveness of this innovative graph-based approach is validated using two prominent PPI datasets. The obtained results underscore the efficacy of this approach, showcasing its superiority over leading previous methods.

Methodology :

Tech Stack -

- Python
- PyTorch
- PyTorch Geometric (for GNN implementations)
- Protein Data Bank (PDB) file processing libraries

Data Collection -

- Protein data, including networks and sequence-structure information, was collected from the databases.

Data Preprocessing -

- The collected data underwent preprocessing to transform it into a suitable format for GNN input. This involved structuring the protein data into a graph representation.

Results :

The GNN model demonstrated promising results in predicting protein functions. The evaluation metrics indicated a high level of accuracy and precision. The model successfully captured the intricate relationships among proteins in the provided graph structure, showcasing its potential for predicting protein functions.

Conclusion :

By combining structural and sequence data with Graph Neural Networks (GNNs), our groundbreaking approach presents a promising avenue for predicting protein functions and unraveling complex protein interactions in biological systems. This integration of graphs and machine learning techniques has immense potential to enhance our grasp of protein biology, making it a valuable tool for drug discovery and improving our understanding of diseases. It's success highlights the potential of GNNs and opens up new avenues in this domain. The outcomes of this research can contribute significantly to advancements in understanding protein functionalities.

Drive :

https://drive.google.com/drive/folders/1FUZliNdTLxUJcBOcwyU2SGmik-dBL_8-?usp=sharing

Articles :

1. **https://www.python-engineer.com/posts/graph-ml-intro/**
2. **https://anindyadeep.github.io/DownTownML/**
3. **https://medium.com/dair-ai/an-illustrated-guide-to-graph-neural-networks-d5564a551783**
4. **https://www.simplilearn.com/what-is-graph-neural-network-article**
5. **https://jonathan-hui.medium.com/applications-of-graph-neural-networks-gnn-d487fd5ed17d**
6. **https://towardsdatascience.com/graph-attention-networks-in-python-975736ac5c0c**
7. **https://analyticsindiamag.com/all-you-need-to-know-about-graph-attention-networks/**
8. **https://petar-v.com/GAT/**

GitHub :

1. **https://github.com/thunlp/GNNPapers/**
2. **https://github.com/Bishnukuet/Protein-Function-Prediction-Tools**