Title: Integrating Graph Neural Networks for Predicting Protein-Protein Interactions Using Structural and Sequence Information

By: Prajukta Dey, Anisha Raj

Abstract:

Proteins, composed of twenty standard amino acids, serve as fundamental macromolecules with pivotal roles across diverse biological processes. Predicting protein functions is a formidable yet foundational task in bioinformatics. Graph Neural Networks (GNNs) have emerged as a promising instrument, adept at capturing the intricate relationships within protein networks and the connections between sequences and structures. In this research, we present a comprehensive framework that integrates both protein structural information and sequence features through the utilization of Graph Convolutional Networks (GCNs) and Graph Attention Networks (GATs) to enhance the prediction of protein-protein interactions (PPIs). Our approach represents protein structures as graphs derived from Protein Data Bank (PDB) files, while sequence-based features are derived using a specialized protein language model. Extensive experiments conducted on prominent PPI datasets unequivocally demonstrate the superiority of our innovative graph-based methodology over previous approaches, thus underscoring its potential for not only predicting protein functions but also unraveling the intricacies of protein interactions in biological systems.

1. **Introduction**

Proteins are essential organic macromolecules composed of twenty standard amino acids, playing critical roles in a multitude of biological processes and cellular functions. Understanding and predicting protein functions are paramount in the field of bioinformatics. The recent emergence of Graph Neural Networks (GNNs) has opened up new possibilities for analyzing graph-structured data, making them a promising tool for predicting protein functions based on the intricate interplay of protein networks and sequence-structure relationships.

2. Importance of Protein-Protein Interactions (PPIs)

The biological realm heavily relies on proteins to perform a wide spectrum of functions, encompassing DNA transcription, replication, hormone regulation, metabolism, molecular cell signaling, and signal transduction. These functions are intricately linked with protein interactions, particularly protein-protein interactions (PPIs), where proteins seldom act in isolation but instead engage with other proteins in their vicinity. Understanding these complex protein interactions is crucial for elucidating various biological activities.

3. Computational Methods for PPI Prediction

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 utilizing 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.

4. Graph-Based Representation of Proteins

In our pursuit of representing proteins within a graph-based framework, we employ Protein Data Bank (PDB) files, a rich source of meticulously recorded 3D coordinates of atoms within protein structures. This step forms the foundation for constructing protein graphs, which effectively capture the intricate structural characteristics of proteins. These protein graphs serve to illustrate either the amino acid network or the residue

contact network, where every individual node encapsulates the essence of a specific residue. The critical connectivity between nodes is established by evaluating pairwise atomic distances, with edges forming only when atoms fall within a predefined threshold distance. This thoughtful thresholding ensures that we retain only the most pertinent interatomic relationships, thereby crafting a concise yet information-rich representation of the protein's structural landscape.

5. **Experimental Validation**

To rigorously evaluate the efficacy and real-world applicability of our innovative graph-based approach, we conducted a comprehensive validation using two well-established and prominent Protein-Protein Interaction (PPI) datasets. These datasets serve as robust benchmarks, enabling us to assess the performance of our methodology under diverse biological contexts and across a spectrum of protein interactions.

6. **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.