Final challenge: Cellular Component Ontology Prediction

Imane Elbacha & Rajae Sebai

MVA 22-23

25/01/2023

Content

1. Introduction
   - Problematic and motivation
   - Dataset
2. Project milestones
   - Sequence representation
   - Structure representation
3. Conclusion and further improvements

Introduction

Problematic and motivation

The problematic of this project is to accurately classify proteins into 18 different classes based on their sequential and structural representation.

Introduction

Dataset and task

The dataset contains 6,111 proteins represented by their sequences, node features, edge features and labels.

- The amino-acids sequence composing the protein.
- The proteins are modeled as undirected graphs.
- The edge attributes include distance between the two connected nodes and binary indicators for different types of edges (distance-based, peptide bond, k-NN, hydrogen bond).
- The node attributes include 3D coordinates, one-hot encoding of amino acid type, hydrogen bond acceptor/donor status, and amino acid features.

Introduction

Dataset and task

The dataset contains 6,111 proteins represented by their sequences, node features, edge features and labels.

- The amino-acids sequence composing the protein.
- The proteins are modeled as undirected graphs.
- The edge attributes include distance between the two connected nodes and binary indicators for different types of edges (distance-based, peptide bond, k-NN, hydrogen bond).
- The node attributes include 3D coordinates, one-hot encoding of amino acid type, hydrogen bond acceptor/donor status, and amino acid features.

Introduction

Dataset and task

The dataset contains 6,111 proteins represented by their sequences, node features, edge features and labels.

- The amino-acids sequence composing the protein.
- The proteins are modeled as undirected graphs.
- The edge attributes include distance between the two connected nodes and binary indicators for different types of edges (distance-based, peptide bond, k-NN, hydrogen bond).
- The node attributes include 3D coordinates, one-hot encoding of amino acid type, hydrogen bond acceptor/donor status, and amino acid features.

Introduction

Dataset and task

The dataset contains 6,111 proteins represented by their sequences, node features, edge features and labels.

- The amino-acids sequence composing the protein.
- The proteins are modeled as undirected graphs.
- The edge attributes include distance between the two connected nodes and binary indicators for different types of edges (distance-based, peptide bond, k-NN, hydrogen bond).
- The node attributes include 3D coordinates, one-hot encoding of amino acid type, hydrogen bond acceptor/donor status, and amino acid features.

Introduction

Dataset and task

The dataset contains 6,111 proteins represented by their sequences, node features, edge features and labels.

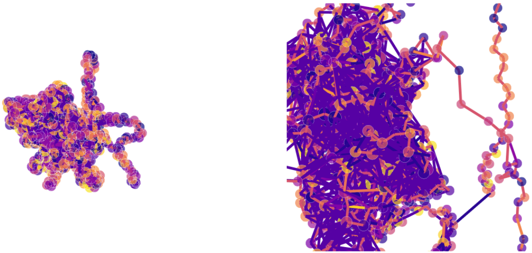

Figure – Structural representation of a protein.

Project milestones

Sequence baseline

- Logistic regression was used as a baseline model for protein classification in this dataset
- Logistic regression resulted in a loss of 1.67, serving as a benchmark for other models
- Logistic regression may not be the best fit for this task due to the complexity of protein sequences and structural information provided, and more complex models may achieve better results.

Project milestones

Support vector classification

- SVC model achieved a loss of 1.11 on protein classification task
- SVC performed well by identifying key features in sequence and using kernel trick for non-linearly separable data
- SVC is robust to outliers and can handle high-dimensional data
- Model is sensitive to choice of kernel and regularization parameter
- A grid search on regularization parameters was done

Project milestones

Structure Baseline

- GNNs were used to classify proteins based on their structure represented as graphs
- The GNN used in this project is a with 2 message passing layers and a sum aggregation function
- Results of this model had a loss of 1.86 which was not as good as the baseline logistic regression and SVC on sequence data
- Limitations include difficulty of capturing all relevant information with GNN architecture and parameters used in this project

Project milestones

Graph Attention Networks and Graph Convolutional Networks

- GCN with 4 message passing layers on node features are state-of-the-art graph neural network architectures
- GAT with 3 message passing layers that use attention followed by a fully-connected layer
- Limitations include high computational cost and sensitivity to graph sparsity and scale

In this project, GAT and GCN models were not able to achieve better results than the baseline GCN model, likely due to complexity of dataset and not deep enough message passing layers.

Project milestones

Language models on sequential data

- Utilized transfer learning techniques to improve performance of protein classification task
- Used pre-trained language models on sequential data such as ESM-2 and ProtBert
- Extracted high-level features from protein sequences and fine-tuned for classification task
- Leveraged vast amount of data and computational resources of pre-trained models
- Resulted in improved performance compared to training models from scratch
- Benefits include faster convergence, better performance

Project milestones

Overview of ESM-2

- ESM-2 is a state-of-the-art protein language model developed by the Meta Fundamental AI Research Protein Team (FAIR) at Facebook.
- ESM-2 can predict various properties of proteins such as structure and function directly from individual sequences and outperforms other single-sequence protein language models.
- ESM-2 was introduced in the preprint ”Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences” (Rives et al., 2019) and was officially released in the paper by Lin et al. in 2022.

Project milestones

Overview of ESM-2

- ESM-2 is pre-trained on millions of protein sequences, learning dependencies and patterns across evolution.
- Pre-trained models like ESM-2 are beneficial as they have been trained on large amounts of data and sophisticated architectures, providing robust representations of proteins.

Project milestones

Performance Analysis

Method	log loss
Structure	
GAT (node features)	1.85
Sequence	
Logistic Regression	1.67
SVC	1.11
Finetuned ProtBert	2.22
Finetuned ESM-2 (35M parameters)	1.37
Finetuned ESM-2 (150M parameters)	**1.06**

Table – Protein classification performances using sequential and structural protein information. The error metric is the multi-class log loss.

Conclusion

Improvements

- Incorporate edge features as inputs to the GNN models by concatenating edge embeddings along with node attributes' embeddings
- Incorporate edge features by using edge-specific convolution kernels when aggregating the neighboring node representations
- Use the edge features as attention coefficients to weight the importance of the edges based on their features, allowing the GNN to focus on the most relevant edges when updating the node representations
- Use graph pooling as a third GNN-based model