CAFA5 Protein Function Prediction

(Project Mentor)

Name: Kaushik Raj V Nadar
Roll Number: 200499
Email ID: nkaushik20@iitk.ac.in

Brief work log

Throughout this project, our team has been working on predicting the function performed by protein sequences in terms of Gene Ontology (GO) terms. It is a challenging multi-label classification problem, where the GO terms serve as the labels.

To kick off the project, we conducted Exploratory Data Analysis (EDA) on the cafa5 Kaggle dataset. This step provided us with valuable insights into the distribution of protein sequences based on their functions, sequence lengths, and other relevant factors. Understanding the data better set a solid foundation for our subsequent work.

To tackle the problem, we implemented BLAST-based algorithms. These algorithms helped us identify sequence homology as a potential solution. By analyzing similarities between different protein sequences, we aimed to gain insights into their functions.

We also explored the application of Classical Machine Learning (ML) models to the existing protein embeddings, specifically T5 and ProtBert. We employed models such as K-Nearest Neighbors (KNN), Random Forests, and XGBoost. Additionally, we experimented with dimensionality reduction techniques like Principal Component Analysis (PCA) and Autoencoder. To find optimal model configurations, we utilized hyperparameter tuning approaches such as Grid Search and Random Search. Despite our efforts, the achieved accuracy for this task was only moderate.

Seeking to improve performance further, we turned our attention to Deep Learning-based approaches. Initially, we applied a Multilayer Perceptron (ANN) to the T5 and ProtBERT protein embeddings. This decision led to a significant boost in performance, with the Fmax score increasing by approximately 46%. The power of deep neural networks was evident in their ability to learn complex patterns and make accurate predictions.

Building on our progress, we decided to shift our focus to creating protein embeddings at the amino acid level. This approach allowed us to leverage specialized Deep Learning techniques, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). These models excel at capturing the sequential nature of the data, which is particularly relevant in the context of protein sequences. By considering the order and

relationships between amino acids, we aimed to extract more informative representations for prediction tasks.

Currently, we are exploring more recent approaches, including models like CNN-LSTM and Transformers with attention mechanisms. These newer techniques have shown promise in various natural language processing and sequence-based tasks. By combining convolutional and recurrent layers, CNN-LSTM models can effectively capture both local and global dependencies in protein sequences. Transformers, on the other hand, have demonstrated exceptional performance in tasks involving attention mechanisms and have become the state-of-the-art in many domains.

From data investigation and the implementation of BLAST-based algorithms to the application of Classical ML models and subsequent advancements in Deep Learning, our project has gone through various stages. We intend to improve the accuracy and performance of our predictions for protein sequence function classification by continuously adapting our methods and leveraging the strengths of different approaches.

Timeline

1) Till 15 May - Initial Quiz and Competition understanding, Python basics

2) 15-22 May - Exploratory Data Analysis

3) 22-25 May - Sequence Homology approach using BLAST

4) 25 May-6 June - Classical Machine Learning approaches, each group was told to implement different models. So, at the end we got 3 working models: KNN, Random Forest, and XGBoost.

5) 7-19 June - Evaluation of ML models using Hamming Loss and F Max Score, and applying dimensionality reduction techniques (PCA and Autoencoder) to the embedding data.
   Implemented Multi Layer Perceptron on the protein embeddings to obtain a decent performing model with an accuracy of 46%.

6) 20 June-8 July - Implemented advanced Deep Learning models, CNN and LSTM to exploit the sequential property of amino acids in protein sequences. Also explored embeddings from different transformer models (ProtBERT, T5, ESM2). We got best performing accuracy with MLP on ESM2 embeddings, i.e. 50%. For this, we had to read various research articles and online blogs and took inspiration from the text classification problem in NLP.

7) 8 July-Present - Currently, we are exploring more recent approaches like CNN-LSTM and Transformers with attention mechanisms.

Important Links

Link to problem description and quiz:
https://docs.google.com/document/d/1c86-XJvyWLS7e0HLSGdg2HJV616gzHzxNoI4eGAWzvs/edit?usp=sharing

Link to quiz responses:
https://docs.google.com/spreadsheets/d/1LlgQeRW8n0MwxOoG8uPSBpa_qs9A25SFnVzSqhwD9hs/edit?usp=drivesdk

Link to Tasks Sheet:
https://docs.google.com/spreadsheets/d/1k1PeMh89COWt8ksnJBRCNRSYzluN9Zhn6BjAfW9iXJA/edit?usp=sharing

Link to GitHub repository:
https://github.com/BSBE-IITK/CAFA5-PFP

PPT Link:
https://docs.google.com/presentation/d/1PPYqCuTquUxiv2DhxScgwzNcYqKmKA0lXg0HQrGaZLE/edit?usp=drivesdk

Attendance sheet:
https://docs.google.com/spreadsheets/d/1pGSciADe0HGyAFV5D1xAQMpEAmUEaZVEBWSnXRusols/edit?usp=drivesdk