Neural Network-based Protein Function Prediction

Bren Jay Magtalas, Daniel De Castro

University of the Philippines

June 14, 2023

Abstract

The study presents a neural network-based approach for protein function prediction. The model utilizes protein sequences and their corresponding ground truth labels to train a multi-label classification model. The architecture of the model consists of several dense layers with the ReLU activation function. The model is trained using the Adam optimizer and the binary cross-entropy loss function. The performance of the model is evaluated using metrics such as binary accuracy and Area Under the Curve (AUC). The results show that the model achieves high accuracy and AUC scores, indicating its effectiveness in protein prediction tasks. The study highlights the potential of neural networks in protein function prediction and suggests further research in this area.

Introduction

- Proteins play a crucial role in the body by fulfilling various essential functions and responsibilities.
- Traditional experimental procedures for discovering protein functions can be costly and labor-intensive.
- Computational methods and algorithms have been introduced to more efficiently predict protein functions from amino acid sequences.
- Assigning specific functions to proteins is challenging due to their multiple functions and interactions with multiple partners.
- Gene Ontology (GO) provides a systematic description of the biological functions of proteins.

Methodology

- Exploratory Data Analysis
- Data Pre-Processing
- Model Architecture
- Model Optimization
- Training Process

Exploratory Data Analysis

- GO is a directed acyclic graph with functional descriptors as nodes connected by relational links, such as "is a" and "part of".
- The graph consists of three subontologies: Molecular Function (MF), Biological Process (BP), and Cellular Component (CC)

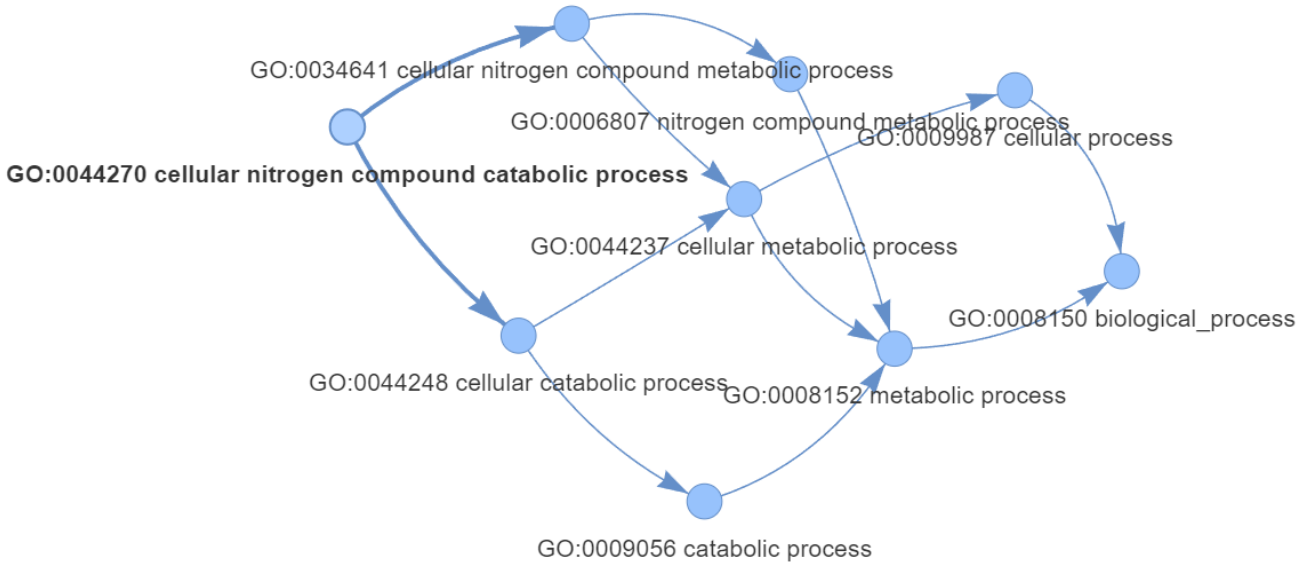

Figure: Cellular Nitrogen Compound Catabolic Process (GO:0044270)

Exploratory Data Analysis

- The data were retrieved from the European Bioinformatics Institute UniProt dataset
- A count of 142,246 sequences was recorded in total
- Majority of proteins have sequence lengths of less than around 2700

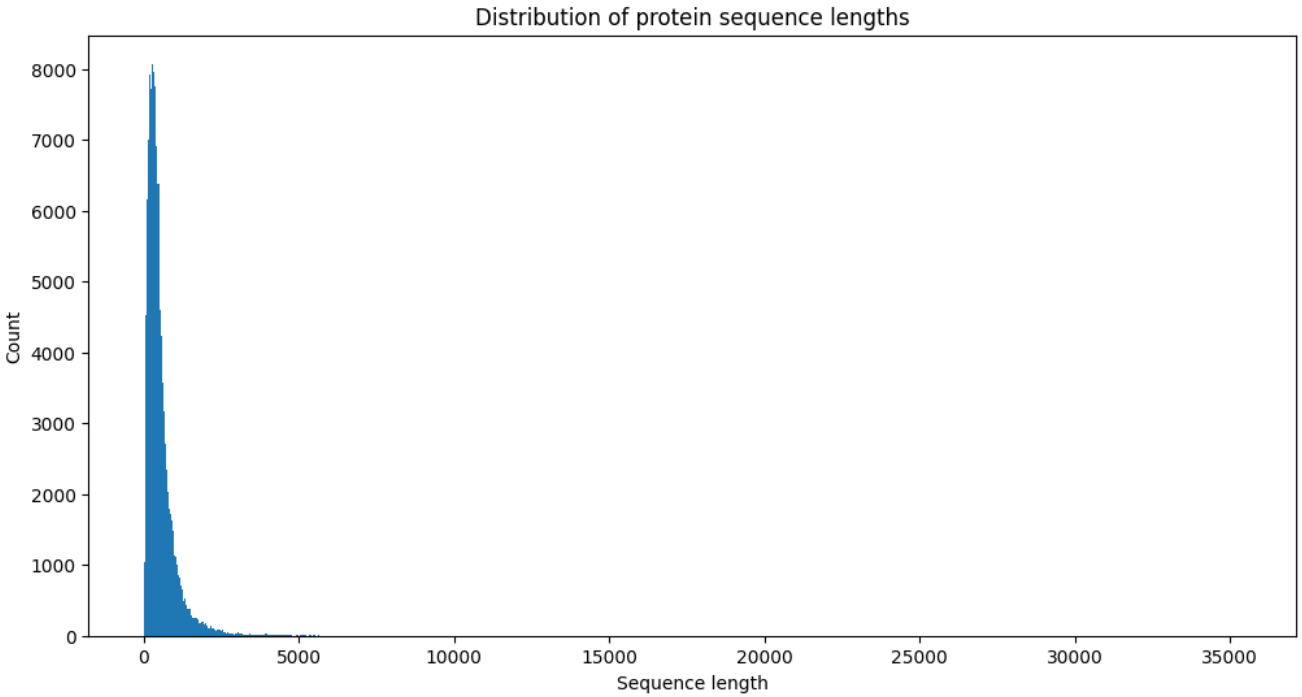

Figure: Protein Sequence Length Distribution

Exploratory Data Analysis

- Amino acids are denoted by uppercase letters of the alphabet
- Ambiguous X, B, Z, and rare O, U in protein sequences suggest possible incompleteness or errors in certain sequences

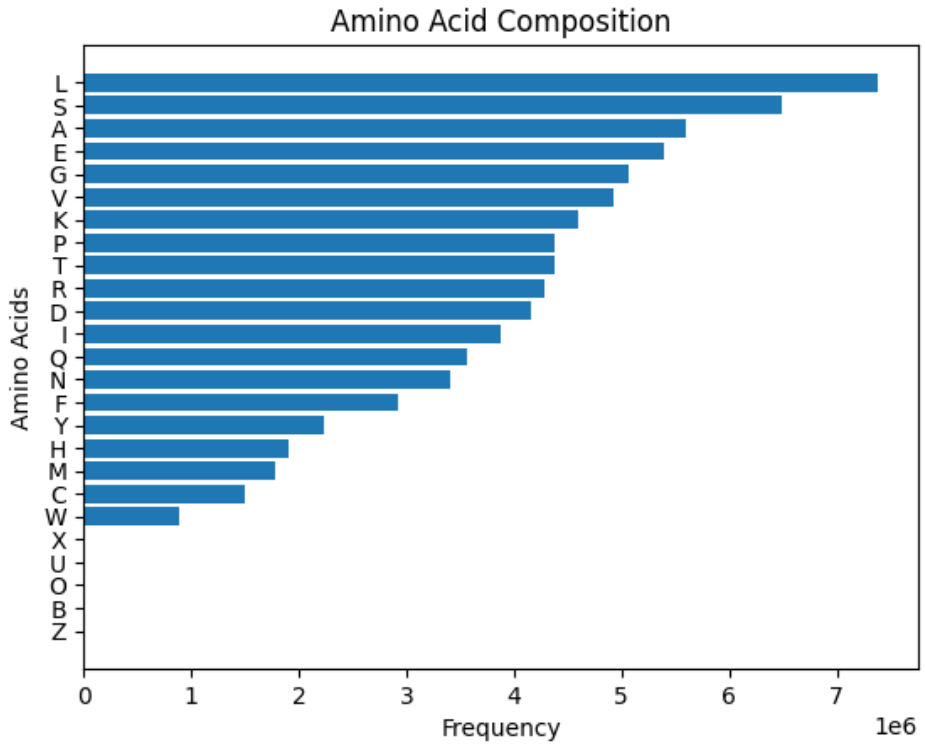

Figure: Amino Acid Distribution

Data Pre-Processing

- Protein sequences are identified by protein ID and their functions are denoted by the GO term ID.
- Sequences are converted into vectors using embeddings, similar to the word embeddings in NLP (Rost Lab's T5 protein language model)
- Embeddings are represented as 1024-dimensional vectors, resulting in 142,246 sample rows and 1024 feature columns.
- The protein ID and GO term ID are processed to be multi-label one-hot encoded data
- To simplify the model, out of more than 40,000 terms, only 1,000 most frequent functions are selected
- The split is performed with a distribution of 64% for training, 16% for validation, and 20% for testing.

Data Pre-Processing

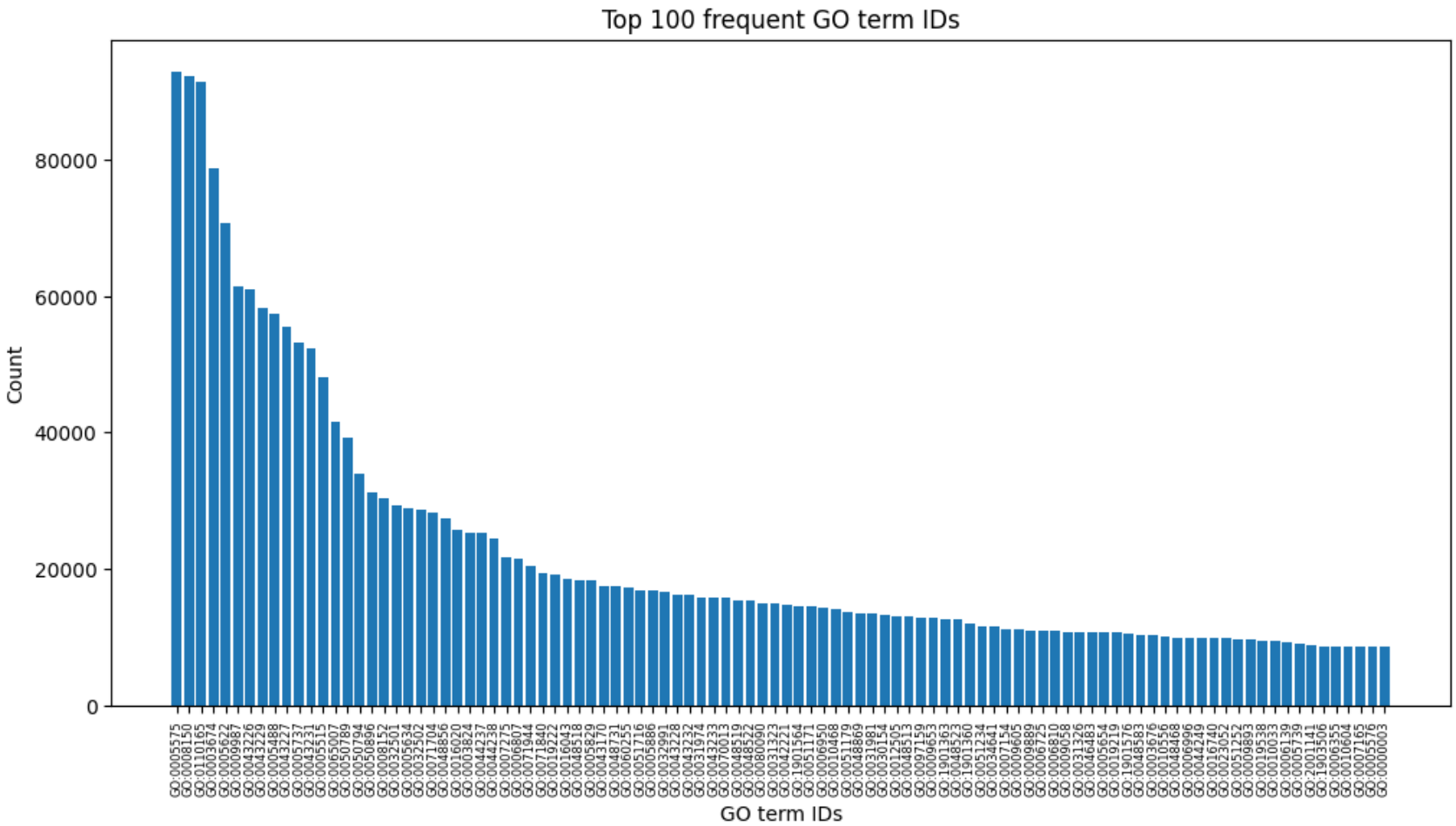

Figure: Label Frequency

Model Architecture

- The first layer is a `BatchNormalization` layer, which normalizes the input features
- The model consists of several dense layers with the ReLU activation function
- The last layer uses the sigmoid activation function for multi-label classification.
- The number of units in each ReLU layer is configured as 2000, 1000, 500, and 2000
- The last layer has 1000 units, corresponding to the label count

Model Optimization

- The model is trained with the Adam optimizer and 0.001 learning rate
- Binary cross-entropy loss function is chosen to suit the multi-label classification task
- Evaluation metrics: binary accuracy and Area Under the Curve (AUC).

Training Process

- The model is fitted using the training data with validation data for monitoring.
- The model is trained for 20 epochs with a batch size of 2000
- The training is stopped when the validation loss starts to increase.

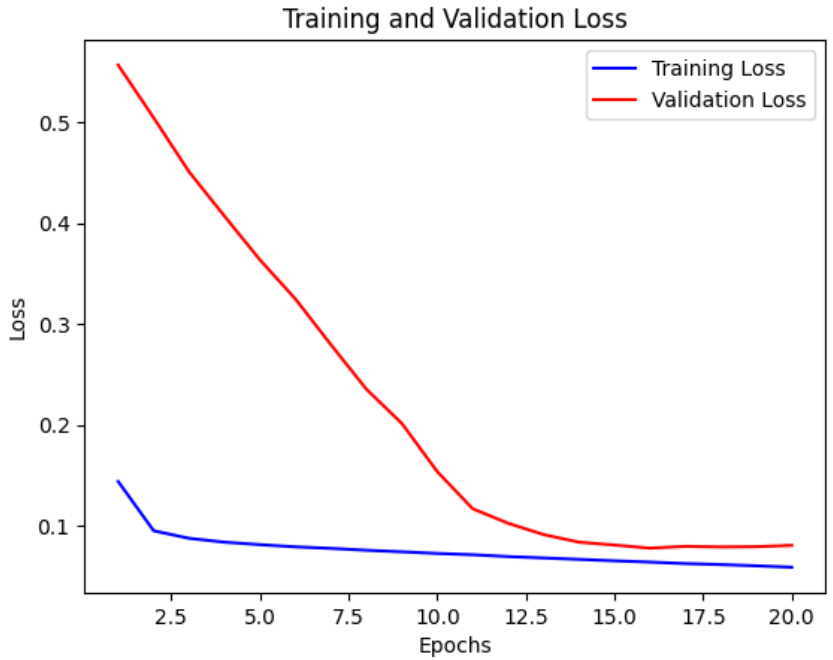

Figure: Training and Validation Loss

Results

- The trained model achieves high accuracy and AUC scores on the test set.

Metrics	Loss	Accuracy	AUC
Train	0.0584	0.9793	0.9651
Validation	0.0801	0.9738	0.9228
Test	0.0808	0.9736	0.9224

Table: Neural Network Performance

Conclusion

- Neural network-based approaches show promise in protein function prediction.
- The presented model achieves high accuracy and AUC scores, indicating its effectiveness.
- Further research and improvements can be done to enhance the model's performance.