CAFA-6 Protein Function Prediction — Final Report

Author: Aparna Vemuganti
Date: October 2025

1. Introduction

The CAFA-6 (Critical Assessment of Functional Annotation) challenge aims to predict the biological functions of proteins from their amino acid sequences. Proteins play essential roles in nearly all biological processes, yet many remain uncharacterized. In this project, I developed and evaluated classical machine-learning models to predict Gene Ontology (GO) terms for given protein sequences.

2. Dataset Overview

The dataset provided by the CAFA-6 competition includes: train_sequences.fasta (amino acid sequences for 82,404 proteins), train_terms.tsv (537,028 GO annotations), train_taxonomy.tsv (species taxonomy IDs), and go-basic.obo (ontology graph). GO terms belong to three sub-ontologies: Molecular Function (MF), Biological Process (BP), and Cellular Component (CC).

3. Preprocessing

Protein IDs were cleaned and standardized to align across datasets. The merged dataset (82,404 proteins) was filtered to the top 200 most common GO terms. Proteins with no valid GO terms were removed, and data was split into 80% training and 20% validation. Final dataset size: 69,167 proteins, 200 GO labels.

4. Feature Engineering

Two sequence feature types were used: (1) Amino acid frequency vectors — 21 numerical features (20 amino acids + length), and (2) 3-mer (tripeptide) count vectors — 3,000-dimensional bag-of-words encoding of overlapping tripeptides.

5. Models and Training

Two machine learning models were trained: Logistic Regression (One-vs-Rest) as a linear baseline, and Random Forest as a non-linear model capturing complex sequence–function relationships. Random Forest was trained on a subset due to computational cost.

6. Evaluation Metrics

Models were evaluated using Micro and Macro F1-scores, appropriate for multi-label problems that balance precision and recall across GO terms.

7. Results

Logistic Regression achieved Micro F1 = 0.1651, Macro F1 = 0.0071. Random Forest (subset) achieved Micro F1 = 0.1203, Macro F1 = 0.0078. Logistic Regression performed slightly better, but both struggled with rare GO terms.

8. Discussion

The results show that simple sequence features can capture limited biological information. Improvements could include deep learning models (ProtBERT, ESM-2), hierarchical GO-aware architectures, and integration of evolutionary or structural features.

9. Conclusion

This project implemented the full CAFA-6 pipeline: data preprocessing, feature extraction, model training, and evaluation. Although baseline scores are modest, this provides a foundation for biologically informed deep learning methods.

10. References

Jiang Y, et al. Genome Biology (2016) 17(1):184. Radivojac P, et al. Nature Methods (2013) 10(3):221–227. CAFA-6 Protein Function Prediction Challenge (2025): https://www.kaggle.com/competitions/cafa-6-protein-function-prediction