CAFA-like Protein Biological Function Prediction by Leveraging ProtT5 PLM Embeddings and Deep Feed-Forward Neural Networks

Marco Uderzo
Department of Mathematics, University of Padua
marco.uderzo@studenti.unipd.it
ID: 2096998

Tanner Graves
Department of Mathematics, University of Padua
tanneraaron.graves@studenti.unipd.it
ID: 2073559

Claudio Palmeri
Department of Mathematics, University of Padua
claudio.palmeri@studenti.unipd.it
ID: 2062671

Abstract

The prediction of the biological function of proteins has been a fundamental topic of bioinformatics. Performing the wet-lab experiments to determine these functions is very expensive and time consuming, so it is crucial to develop computational methods for automated function prediction. Gene Ontology is a standardized system to categorize and describe function in a hierarchical manner. This paper concerns the development and evaluation of a Deep Learning model to predict the biological function of proteins, especially by leveraging the per-protein embeddings generated the ProtT5 Protein Language Model.

1 Introduction

1.1 Protein Biological Function Prediction

Proteins are responsible for many activities in our tissues, organs, and bodies and they also play a central role in the structure and function of cells. The accurate assignment of biological function to the protein is key to understanding life at the molecular level. However, assigning function to any specific protein can be made difficult due to the multiple functions many proteins have, along with their ability to interact with multiple partners. More knowledge of the functions assigned to proteins, potentially aided by Data Science, could lead to curing diseases and, overall, drastically improving human health.

Gene Ontology is a standardized system to categorize and describe function in a hierarchical manner. In order to categorize and classify functions, Gene Ontology has separated the functions into 3 sub-ontologies:

- **Molecular Function**: functions that the protein performs at the molecular level.
- **Biological Process**: events or processes that the protein is involved in.
- **Cellular Component**: locations in the cell that the protein is active.

These sub-ontologies are completely separated from each other, and the terms in them do not have inter-relations between them.

Overall, this project concerns the development and evaluation of a Deep Learning model to predict the biological function of proteins.

1.2 Simplifications from CAFA

As specified in the project requirements, in order to make the task manageable in terms of computational complexity, some simplifications were made.

- Only functions that have more than 50, 250, 50 instances respectively in Molecular Function (MF), Biological Process (BP), and Cellular Componet (CC) have been gathered.
- Proteins with sequence length of more than 2000 have been removed.

- Only functional annotations with experimental evidence code, TAS, and IC have been considered.

2 Methods

2.1 Training Dataset Description

The full training dataset for the project contains the following files:

- *train_set.tsv*: The proteins, their Gene Ontology (GO) annotations, and their corresponding aspects are given in a tab-separated file. The GO annotations that are given for each protein are already propagated to the root of the ontology.

- *train_ids.txt*: This file contains the accession IDs of the proteins that are given. It contains about 123,000 proteins.

- *train.fasta*: This file contains the sequence of the train proteins in FASTA format.

- *train_embeddings.h5*: This file contains ProtT5 embeddings for each protein in the train set.

- *train_protein2ipr.dat*: This file contains the InterPro domains for proteins of the training set.

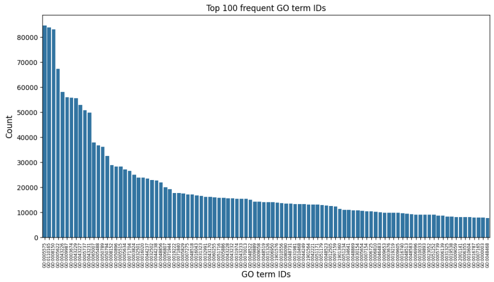
- *go-basic.obo*: This file contains the whole Ontology terms and the relationships between them. Note that the only relationships that we consider are `part_of` and `is_a`.

In our project we are not using all of these files. For brevity, we now describe more in detail the data we decided to use as our final training set.

2.1.1 Gene Ontology Annotations and Aspects

Gene Ontology provides a framework for the representation of gene and gene product attributes across all species. In our file, each protein is annotated with relevant GO terms. These annotations describe the roles of proteins at various levels, such as their biological processes, cellular components, and molecular functions.

By visualizing the distribution of GO_terms in our train_set.tsv file, we realize that the data is massively unbalanced, with some terms being overrepresented, where as most are underrepresented. Indeed, they tend to follow a right-skewed exponential distribution, forming a long-tail fenomenon. Below, we present the distribution of the top 100 Gene Ontology terms.



2.1.2 ProtT5 PLM Embeddings

ProtT5 is an advanced language model specifically designed for processing and interpreting protein sequences. Developed as an extension of the T5 (Text-to-Text Transfer Transformer) framework, ProtT5 is able to capture the complex patterns and structures inherent in protein sequences. ProtT5 embeddings are vector representations of protein sequences generated by the ProtT5 model. These embeddings encapsulate the contextual information of amino acids in a protein sequence, effectively representing the sequence in a high-dimensional space.

2.2 Dataset Preprocessing

Protein IDs and their aspect and GO_term are loaded from the tsv train_set.tsv file, and for each protein, its embeddings in .h5 format are also parsed. Each protein embedding consists in 1024 columns. The first step of our study was to perform some Explorative Data Analysis on the dataset at our disposal. As already mentioned above, Gene Ontology terms are massiveley unbalanced. This is a very common occurrence when working with biological data, which is often imbalanced, biased or scarce. The most represented aspect is by far `cellular_component` followed by `biological_process`, with less than half of occurrences, which is also consistent with the number of unique proteins having such terms (?). Therefore, we needed to perform data augmentation, by undersampling some of the overrepresented terms and carefully oversampling some of the underrepresented ones.

Type	Tool	Class	Value
Undersampling	?	?	?
	?	?	?
	?	?	?
	?	?	?
	?	?	?
	?	?	?
Oversampling	?	?	?
	?	?	?
	?	?	?
	?	?	?

2.3 Model Development

Our training data contains mostly tabular data, except the FASTA protein sequences. Though, we already have embeddings from the ProtT5 Protein Language Model, which is based on the Transformer architecture, known for its efficacy in dealing with long time-space-dependent data. Therefore, using more advanced architecture like RNNs to deal with the raw FASTA sequences would most likely not yield any better result than a Deep Neural Network. RNN-based models could have been leveraged in case we had at our disposal per-residue ProtT5 embeddings, instead of just per-protein embeddings.

A summary of the model's architecture is shown below.

Layer Type	Output Shape	Param #	Activation
Dense	(None, 128)	131200	ReLU
Dropout	(None, 128)	0	N/A
Dense	(None, 256)	33024	ReLU
Output_cc	(None, 625)	320625	Sigmoid

The hyperparameters of the model are listed in the following table.

Hyperparameter	Value
Learning Rate	0.001
Epochs	50
Batch Size	32
Optimizer	Adam
Loss Function	Binary Cross-Entropy

3 Results

3.1 Performance Evaluation

4 Discussion

5 Usage Overview

6 Code Availability

The datasets used and all the code used for this project is available at the following GitHub Repository.

7 References

List and number all bibliographical references in 9-point Times, single-spaced, at the end of your paper. When referenced in the text, enclose the citation number in square brackets, for example. Where appropriate, include the name(s) of editors of referenced books.