CAFA-like Protein Biological Function Prediction with Hybrid Convolutional-LSTM Recurrent 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: 209xxxx

Claudio Palmeri
Department of Mathematics, University of Padua
claudio.palmeri@studenti.unipd.it
ID: 209xxxx

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. In our project, we use a hybrid deep learning architecture. Combining Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) can be a powerful approach when dealing with sequential data like protein sequences. This hybrid model can capture both local patterns through convolutional layers and long-range dependencies through recurrent connections.

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, overrall, 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), Bi-

ological 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 Dataset Preprocessing

2.2 Model Development

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.

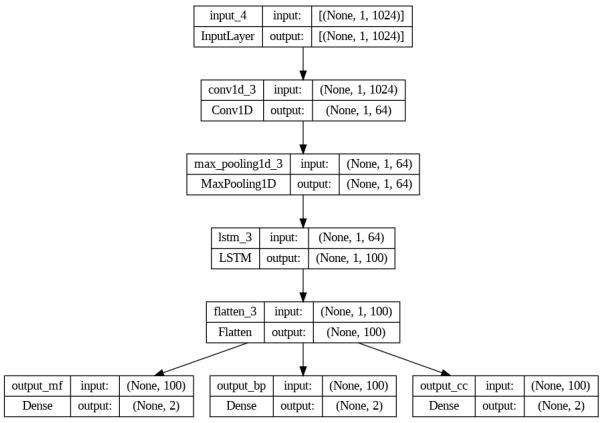

Figure 1: Diagram of the Conv-LSTM model.