PROTEIN FUNCTION PREDICTION

Ahmet Arif Acıduman 21626857(%25 work)
Erhan Kabaoğlu 21627389(%50 work)
Rıfat Onur Akdoğan 21726861(%25 work)

1. ABSTRACT

For a long time the number difference between known protein sequences and their functions only got bigger however understanding the functions of the corresponding protein sequences by experimenting is not a viable way both resource-wise and time-wise. With the technological advances and the current power of computational devices, unveiling the functions of proteins by using computational biology became even more preferable and economical. To approach this sequence function matching problem, machine learning is being used extensively by feeding in protein sequences as input and learning functionalities by prediction.

2. INTRODUCTION

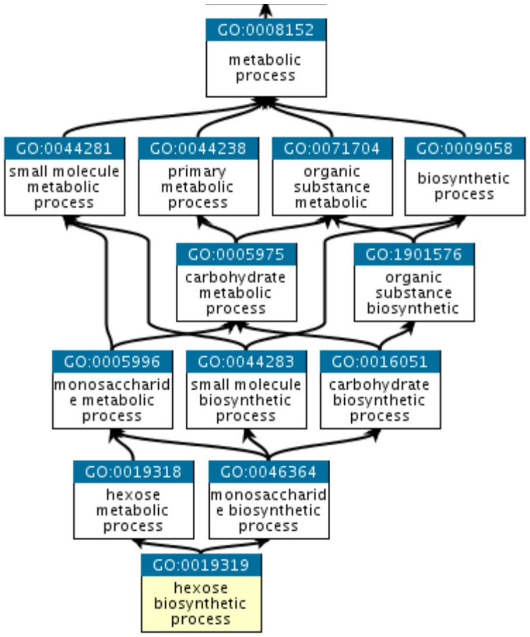
Proteins are highly complex biomolecules consisting of one or more chains of amino acids which perform a wide array of functions important to cells' workings on a micro level and organism's functioning in a macro level. Proteins stand in the middle of all biological processes like catalytic activity, muscle contraction, immune system or signaling and regulation. Understanding the functionality of proteins leads to the better understanding of diseases and different biological processes, and with the latest developments of technology it can be done more accurately and better than ever.

The number of discovered protein sequences increased fast over the past years. The old way of understanding their functionality by experimenting became less viable due to it being very costly. Nowadays computational biology is used extensively to combat this **Sequence Function Matching Problem**.

Gene ontology(GO) describes our knowledge of the biological domain with respect to three aspects: **molecular function**, **cellular component** and **biological process**. Molecular function is molecular-level activities performed by gene products. It describes activities that take place at the molecular level like catalysis and transport. Cellular components are locations relative to cellular structures in which a gene product performs a function. A biological process is a relatively larger process accomplished by multiple molecular activities such as DNA repair or signal transduction. GO classes are also known as terms and each term has a unique identifier and a unique name. GO terms describe different levels of protein functions.

GO terms: every GO term has a human-readable name such as mithocondrion or amino acid binding and a GO identifier like GO:0005739 or GO:1904659. All terms have a sub-class relationship to another term. An example would be: glucose transmembrane transport (ID:GO:1904659) is a monosaccharide transport(ID:GO:0015749).

A sample GO graph can be seen below:


GO:0008152
metabolic process
GO:0044281
small molecule metabolic process
GO:0044238
primary metabolic process
GO:0071704
organic substance metabolic
GO:0009058
biosynthetic process
GO:0005975
carbohydrate metabolic process
GO:1901576
organic substance biosynthetic
GO:0005996
monosaccharid e metabolic process
GO:0044283
small molecule biosynthetic process
GO:0016051
carbohydrate biosynthetic process
GO:0019318
hexose metabolic process
GO:0046364
monosaccharid e biosynthetic process
GO:0019319
hexose biosynthetic process

3. RELATED WORKS

Most contemporary solutions to the sequence function matching problem employ machine learning based methods. One example we can give is DeepGOPlus from Maxat Kulmanov and Robert Hoehndorf. They developed a method which combined a CNN model with sequence similarity based predictions to predict protein functions which made DeepGoPlus one of the best performers related to this subject.(1)

Another paper from Yu Zhen Zhou, Yun Gao, and Ying Ying Zheng tried to predict protein-protein interactions using only the information from the protein sequence. They developed this by combining a novel representation of local protein sequence descriptors and support vector machines(SVM).(2)

A model using deep learning Graph Convolutional Network(GCN) for predicting protein functions is developed by Vladimir Gligorijevic and his colleagues. They used experimentally determined structures from the Protein Data Bank (PDB) but the model can also work really well using structure predictions.(3)

In 2015, another approach, which was included in the CAFA challenge called “Guilty by Association on STRING(GAS)” was developed. Damiano Piovesan, Manuel Giollo, Carlo Ferrari and Silvio C. E. Tosatto created the tool to predict protein functions exploiting protein-protein interaction networks without sequence similarity. They used the idea of whenever a protein is interacting with each other, the proteins are part of the same biological process and are located in the same cellular compartment.(4)

4. DATA

In our experiment, we used Cafa3 training and target datasets. These dataset sources consist of Swiss-prot, flybase, Pseudomonas genome database and Candida Genome Database.

These datasets store and provide reliable protein information which are sequence domain, GO terms, functionality and many other things about proteins.

In our problem, we need to just protein domains and their GO terms to classify protein functions.

We selected 50 unique GO terms to classify protein functions because of hardware capacity.

Each of the 50 GO terms consist of 100 data points. We split this subset into %10 validation dataset and %90 train dataset for our experiment. Also we paid attention to the homogeneous distribution of each class.

The protein sequences consist of amino acids and these amino acids represented using unique letters (A-Z), so we need to convert this representation into numerical representation to train our model.

There are several methods like OneHotEncoder, Word Embeddings. For our experiment, we chose OneHotEncoder representation which is a group of bits among which the legal combinations of values are only those with a single high (1) except one (0).

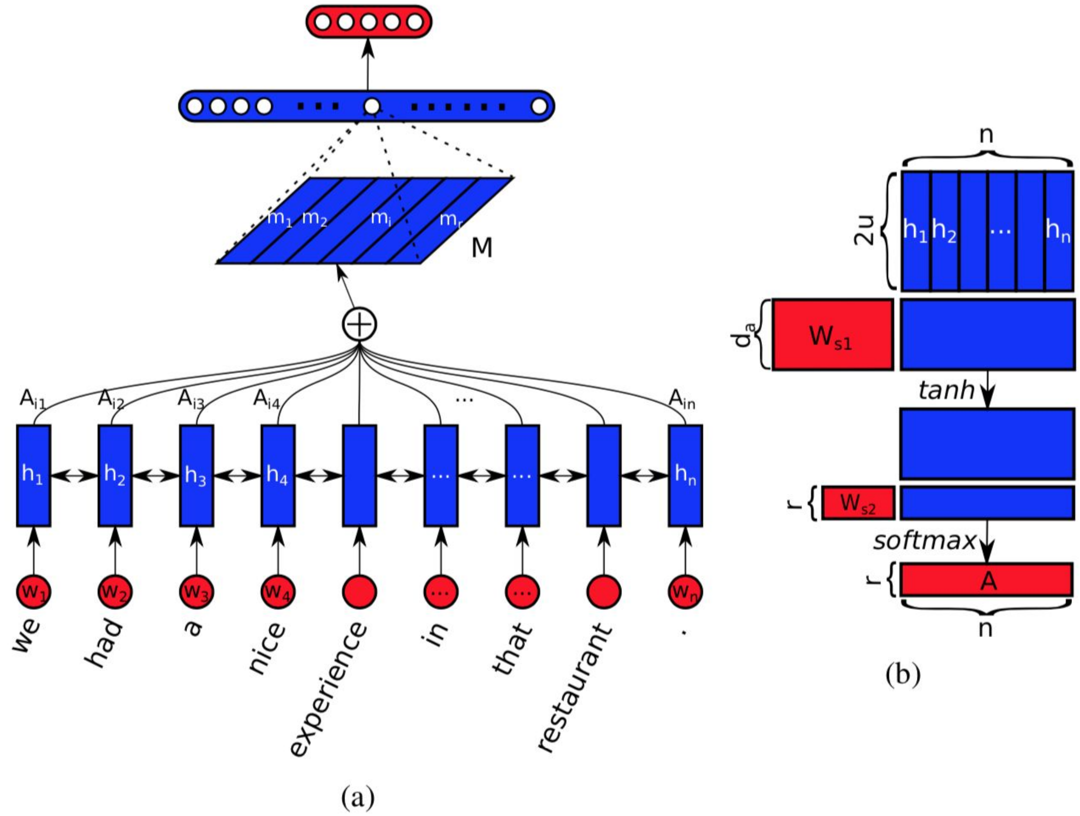
5. METHODS

There are several methods to classify sequences. Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) have achieved significant success in sequence classification problems as a result of recent developments in the field of deep learning.

That’s why we decided to use one of these methods, Recurrent Neural Network (LSTM) in our problem.

Although LSTMs perform well in sequence classification, they have some weaknesses in classifying long sequences because all input sequences are forced to classification layer.
We decided to use a self-attention mechanism to solve this problem.

In figure , we can see a self-attention mechanism for sequence classification.


m_1 m_2 m_i m_r
M
A_{i1} A_{i2} A_{i3} A_{i4} ... A_{in}
h_1 h_2 h_3 h_4 ... ... h_n
w_1 w_2 w_3 w_4 ... ... w_n
we
had
a
nice
experience
in
that
restaurant
.
n
2u
h_1 h_2 ... h_n
d_a
W_{s1}
tanh
r
W_{s2}
softmax
r
A
n

(a)

(b)

a-) Architecture Details:

We used bidirectional LSTM for recurrent networks. At the top of LSTM, we used a self-attention layer and linear layer to classify sequences.
LSTM layer input size is 25 which is vocabulary size, hidden size is 256 and we used 2 layers LSTM with dropout 0.5.
For the attention-layer, we used two linear layers with shapes (512, 350), (350, 30) respectively.
For the classification layer, we used a linear layer with shape (2000, 50) and we applied softmax function to the output.
b-) Hyper parameter Details:

We used ADAM optimizer with a learning rate 0.001.
We used CrossEntropy Loss.
We decrease learning rate to 0.0001 in epoch 25.
We train our model 50 epochs with batch size 16.
We also clipped gradients to avoid gradient exploitation.

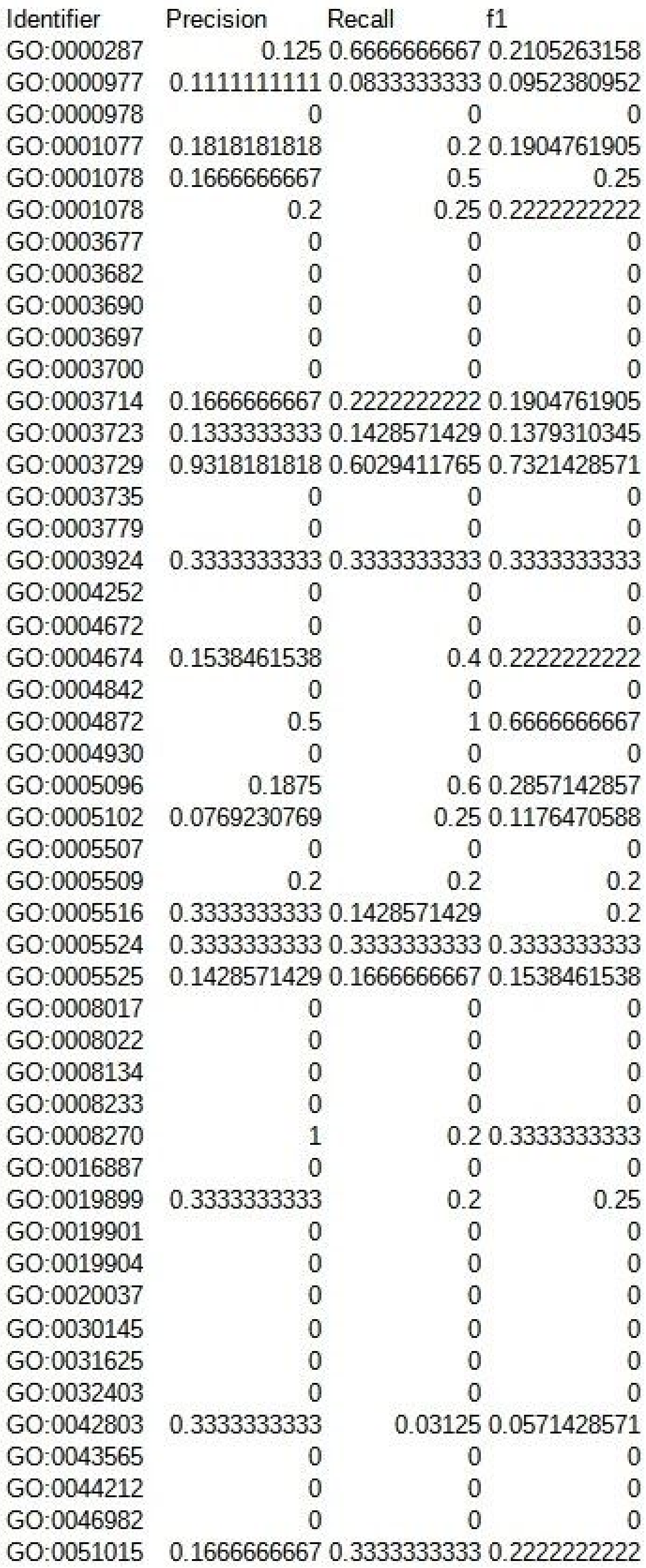
For evaluation, we chose the model with the highest accuracy in the validation data set.

We implemented this architecture in pytorch.
We used scikit-learn to calculate metrics.
We used matplotlib and seaborn for visualization

6. RESULTS

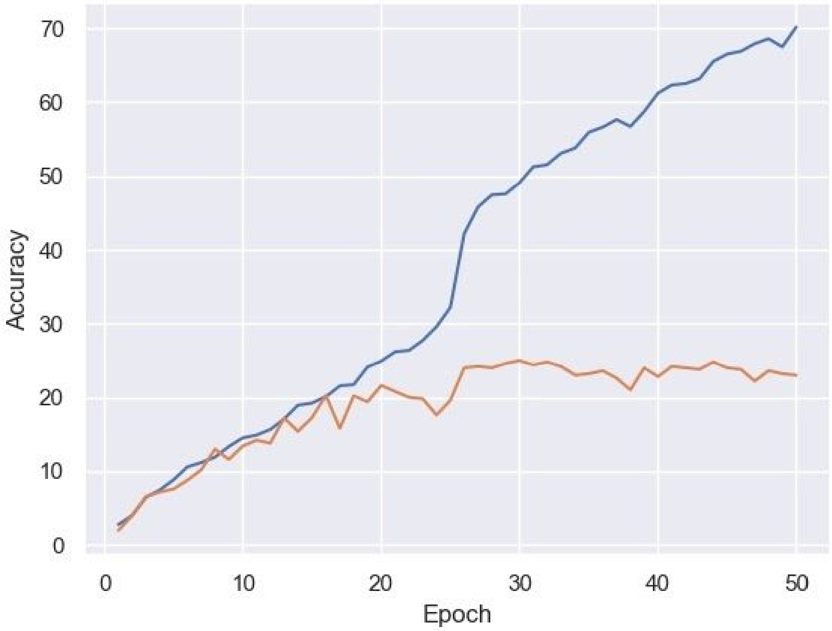
After we trained the protein sequence classification model, we tested our model in the target dataset. We calculated some metrics to measure our model quality.
The model achieved an F1 score of 0.30.
Precision score is 0.38 and recall is 0.27.
Mcc score is 0.21
Accuracy is 0.27 for the target dataset.

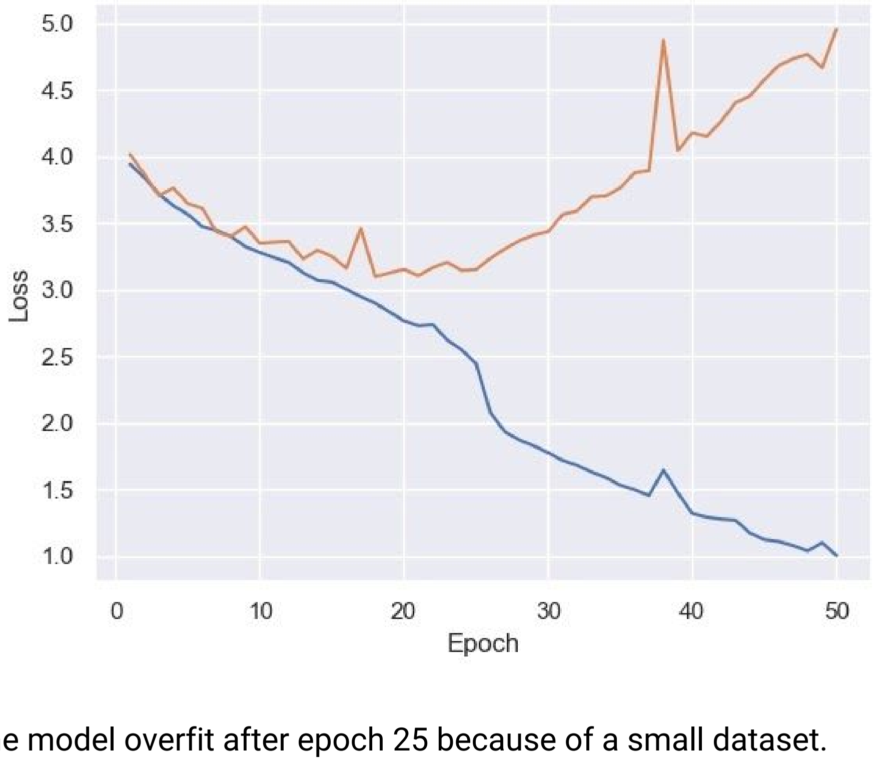
These values can't be directly compared with values in the Cafa3 competition because the size of the dataset and number of classes we use differed.



Identifier	Precision	Recall	f1
GO:0000287	0.125	0.6666666667	0.2105263158
GO:0000977	0.1111111111	0.0833333333	0.0952380952
GO:0000978	0	0	0
GO:0001077	0.1818181818	0.2	0.1904761905
GO:0001078	0.1666666667	0.5	0.25
GO:0001078	0.2	0.25	0.2222222222
GO:0003677	0	0	0
GO:0003682	0	0	0
GO:0003690	0	0	0
GO:0003697	0	0	0
GO:0003700	0	0	0
GO:0003714	0.1666666667	0.2222222222	0.1904761905
GO:0003723	0.1333333333	0.1428571429	0.1379310345
GO:0003729	0.9318181818	0.6029411765	0.7321428571
GO:0003735	0	0	0
GO:0003779	0	0	0
GO:0003924	0.3333333333	0.3333333333	0.3333333333
GO:0004252	0	0	0
GO:0004672	0	0	0
GO:0004674	0.1538461538	0.4	0.2222222222
GO:0004842	0	0	0
GO:0004872	0.5	1	0.6666666667
GO:0004930	0	0	0
GO:0005096	0.1875	0.6	0.2857142857
GO:0005102	0.0769230769	0.25	0.1176470588
GO:0005507	0	0	0
GO:0005509	0.2	0.2	0.2
GO:0005516	0.3333333333	0.1428571429	0.2
GO:0005524	0.3333333333	0.3333333333	0.3333333333
GO:0005525	0.1428571429	0.1666666667	0.1538461538
GO:0008017	0	0	0
GO:0008022	0	0	0
GO:0008134	0	0	0
GO:0008233	0	0	0
GO:0008270	1	0.2	0.3333333333
GO:0016887	0	0	0
GO:0019899	0.3333333333	0.2	0.25
GO:0019901	0	0	0
GO:0019904	0	0	0
GO:0020037	0	0	0
GO:0030145	0	0	0
GO:0031625	0	0	0
GO:0032403	0	0	0
GO:0042803	0.3333333333	0.03125	0.0571428571
GO:0043565	0	0	0
GO:0044212	0	0	0
GO:0046982	0	0	0
GO:0051015	0.1666666667	0.3333333333	0.2222222222

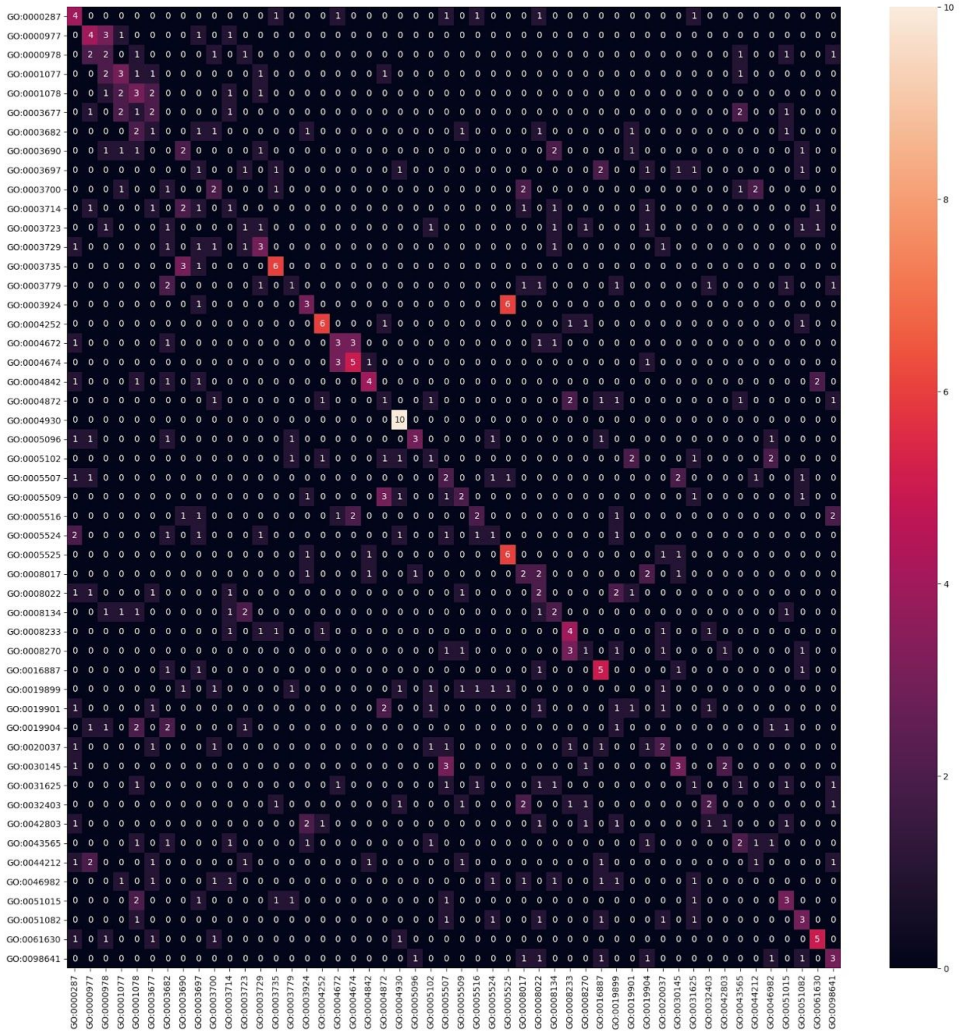
You can see accuracy and loss plots below:

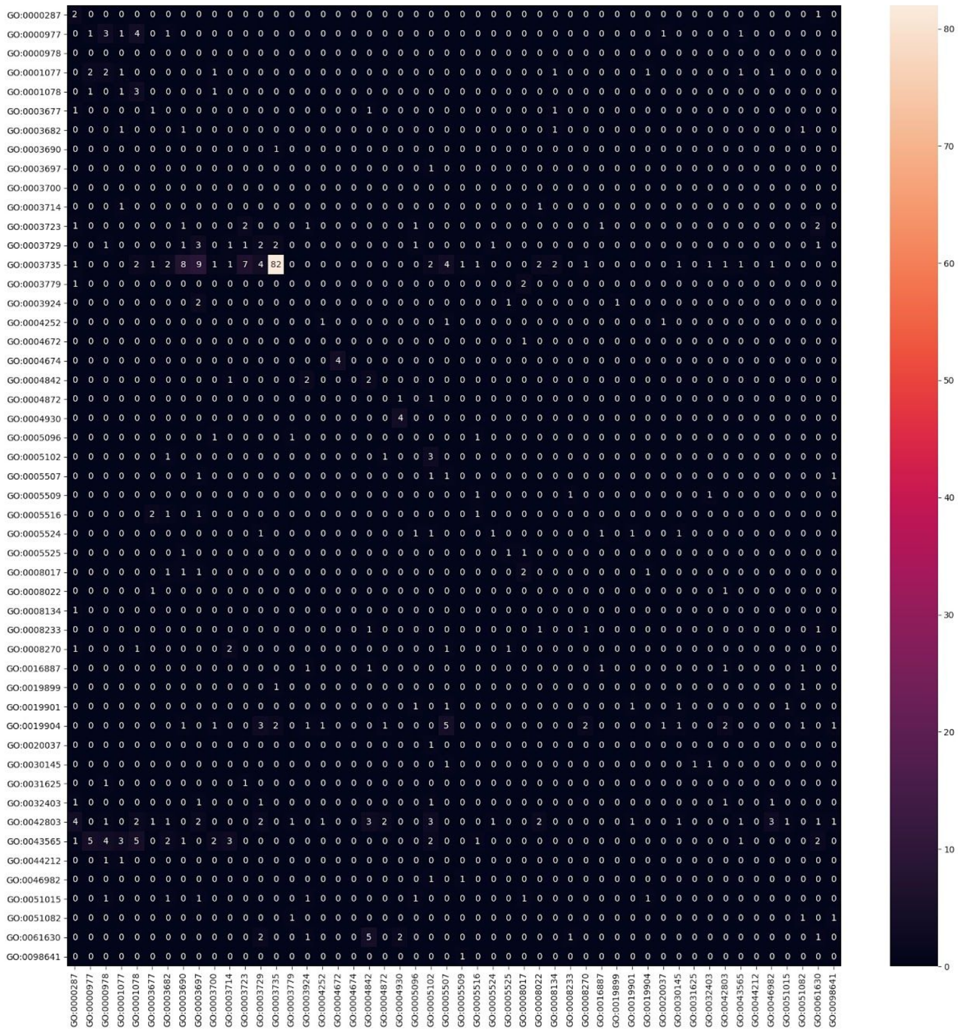




The model overfit after epoch 25 because of a small dataset.

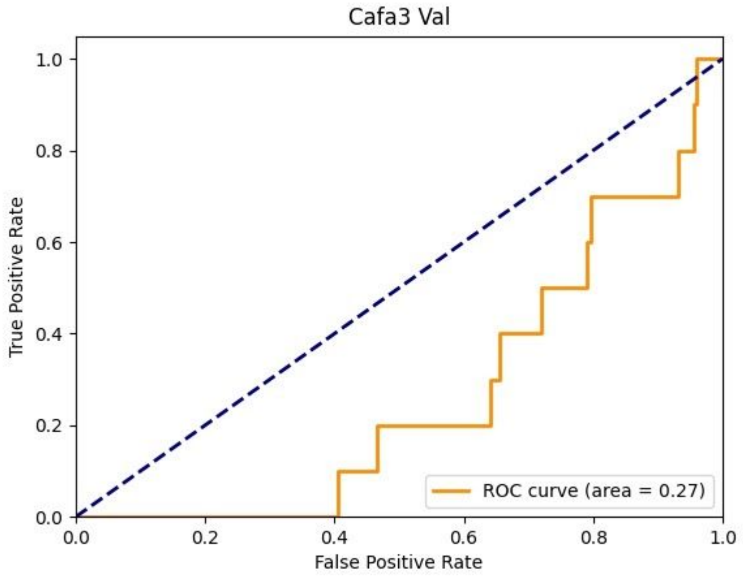
You can see validation and test confusion matrix respectively below:

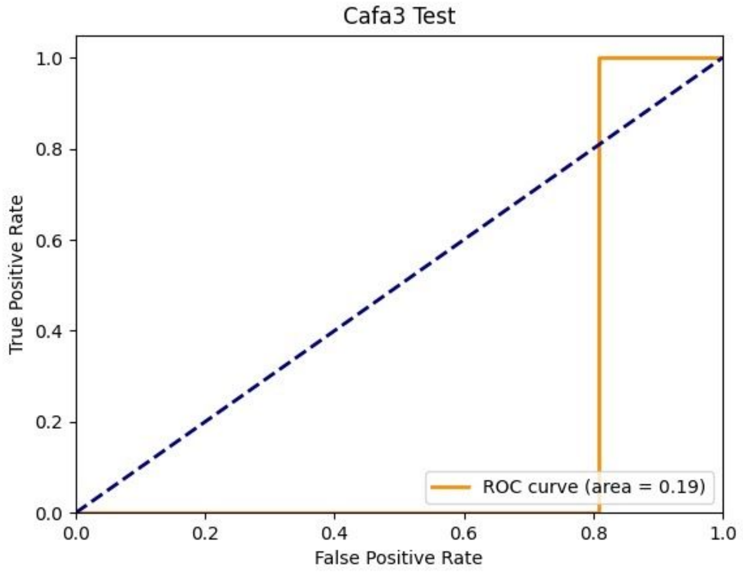




As you can see from the target confusion matrix, the target(test) dataset is not distributed homogeneously.

Here are the roc curves and auc scores for validation and target datasets respectively (multi label):





7. CONCLUSION & DISCUSSION

We found lots of different applications and papers about this subject. There are also different materials which can be accessed easily which makes protein function prediction a hot topic in Bioinformatics. This is because machine learning algorithms and applications are getting better and more popular each year. We think that this topic still has lots of potential to improve.

In this project we used protein domains and their GO terms respectively to classify protein functions. Since we didn't include all classification subjects in the CAFA dataset we used to train and test, our model might not have reached it's furthest potential but it still gave us a pretty good idea about protein function prediction. If we did include more variables maybe our performance could've been better. We learned that protein functions are not only dependent on their sequences but also their 3D models and what they do inside of cells. We should consider evolutionary origins as well because proteins are more likely to perform the same operations in evolutionary similar organisms. Protein-protein interactions should also be considered because if two proteins are interacting, their possibility of being in the same biological category is much higher than other proteins.

8. REFERENCES

1. Maxat Kulmanov, Robert Hoehndorf, DeepGOPlus: improved protein function prediction from sequence, Bioinformatics, Volume 36, Issue 2, 15 January2020, Pages 422-429.
https://doi.org/10.1093/bioinformatics/btz595

2. Zhou Y.Z., Gao Y., Zheng Y.Y. (2011) Prediction of Protein-Protein Interactions Using Local Description of Amino Acid Sequence. In: Zhou M., Tan H. (eds) Advances in Computer Science and Education Applications. Communications in Computer and Information Science, vol 202. Springer, Berlin, Heidelberg.
https://doi.org/10.1007/978-3-642-22456-0_37

3. Vladimir Gligorijevic, P. Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C. Taylor, Ian M. Fisk, Hera Vlamakis, Ramnik J. Xavier, Rob Knight, Kyunghyun Cho, Richard Bonneau. Structure-Based Protein Function Prediction using Graph Convolutional Networks.
https://doi.org/10.1101/786236

4. Piovesan, D., Giollo, M., Ferrari, C. et al. Protein function prediction using guilty by association from interaction networks. Amino Acids 47, 2583-2592 (2015).
https://doi.org/10.1007/s00726-015-2049-3

5. Jiang, Y., Oron., T., Clark, W. et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biol* 17, 184 (2016).
https://doi.org/10.1186/s13059-016-1037-6

6. Predrag Radivojac. (2013) A (not so) Quick Introduction to Protein Function Prediction.
Introduction_to_protein_prediction.pdf (biofunctionprediction.org)

7. Ashburner et al. Gene ontology: tool for the unification of biology. Nat Genet. May 2000;25(1):25-9.

8. The Gene Ontology resource: enriching a GOld mine. Nucleic acids Res. Jan 2021;49(D1):D325-D334.

9. A Structured Self-Attentive sentence embedding
https://arxiv.org/pdf/1703.03130.pdf