Investigation of Multi-task Deep Neural Networks in Automated Protein Function Prediction

Ahmet Sureyya Rifaioglu[1,*], Maria Jesus Martin[2], Rengül Çetin-Atalay[3] and Mehmet Volkan Atalay[1], Tunca Doğan[2,3]

[1]Department of Computer Engineering, Middle East Technical University, Ankara, Turkey,
[2]European Bioinformatics Institute, Hinxton, Cambridge, CB10 1SD, UK,
[3]CanSyL, Graduate School of Informatics, Middle East Technical University, Ankara, 06800, Turkey

*To whom correspondence should be addressed: arifaioglu@ceng.metu.edu.tr

1. INTRODUCTION

Functional annotation of proteins is a crucial research field for understanding molecular mechanisms of living-beings and for biomedical purposes (e.g. identification of disease-causing functional changes in genes and for discovering novel drugs). Several Gene Ontology (GO) based protein function prediction methods have been proposed in the last decade to annotate proteins. However, considering the prediction performances of the proposed methods, it can be stated that there is still room for significant improvements in protein function prediction area (1). Deep learning techniques became popular in recent years and turned out to be an industry standard in several areas such as computer vision and speech recognition. To the best of our knowledge, as of today, deep learning algorithms have not been applied to the large-scale protein function prediction problem. Here, we propose a hierarchical multi-task deep neural network architecture, DEEPred, as a solution to protein function prediction problem. First of all, we investigated the potential of employing deep learning methods for protein function prediction. For this purpose, we measured the performance of our models at different parameter settings. Furthermore, we examined the relationship between the performance of the system and the size of the training datasets, since the training set size has been reported in the literature to be significantly affecting the performance of deep learning models.

2. METHODS

We used GO term annotations from UniProtKB/SwissProt with manual experimental evidence codes for the training. We divided this dataset according to different levels in GO hierarchy (i.e. 9 to 12 levels with respect to different GO categories). The objective here is to create a multi-task deep neural network model for GO terms with similar number of training samples at each level of GO category. For the generation of protein feature vectors, we used a modified version of subsequence-based feature extraction method called SPMap (2). We trained multi-task deep neural networks for each of the different GO term sets. The models were trained with drop-out technique to avoid overfitting. A general view of the method is shown in Figure 1A. In the proposed method, a task corresponds to a GO term, therefore, when a query sequence is fed to our models as input, DEEPred calculates a score for each trained GO term within the model, which represents the probability of the input protein having a particular function defined by a GO term (Figure 1B).

We used multi-task feed-forward deep neural network architecture with several parameters for number of hidden layers, number of neurons, learning rate and drop-out rate. Number of neurons at each hidden layer ranged from 500 to 5000. Learning rate parameters ranged from 0.001 to 0.1 and drop-out rate parameters ranged from 0.3 to 0.8. We used TensorFlow for training models and all computations were distributed into 2500 CPU cores.

2. RESULTS

A level specific evaluation based on GO hierarchy was performed for performance calculations.

We created six training datasets for each GO category where each dataset corresponds to GO terms having protein associations with numbers greater than the specified thresholds (i.e. > 50, 100, 200, 300, 400 and 500 annotated proteins). In this evaluation strategy, we had two major objectives to investigate: Our first objective was to demonstrate the relation between the GO levels (e.g. generic/specific) and the classification performance. The second aim was to show the relation between performance and size of the training datasets that were used for each GO term.

Performance evaluation results are given in Figure 2 where each box plot represents F-Score performances of the GO terms for different training datasets. The results showed that there is a general trend of performance increase with the increasing number of training samples, which also means that including GO terms with small number of protein associations in our models decreases the overall performance. However, we observed that there is no correlation between GO levels and classification performance. When the results were investigated, we saw that variance of the performances decreases and the performance increases as the training dataset size increases for molecular function and cellular component categories. For biological process category, performance also increases with increasing GO training dataset sizes; however, variance was relatively higher for each dataset. In addition, results showed that deep learning can be employed to significantly improve the performance of prediction (F-score ≥ 0.75) for hard to predict GO categories such as the biological process and the cellular component, especially when the training set sizes are greater than 500 proteins. As a conclusion, we showed that deep learning techniques has a significant potential in automated protein function prediction. We plan to further investigate the model behavior under different circumstances and to optimize the models to provide DEEPred as an open-access tool to the research community.

3. FIGURES

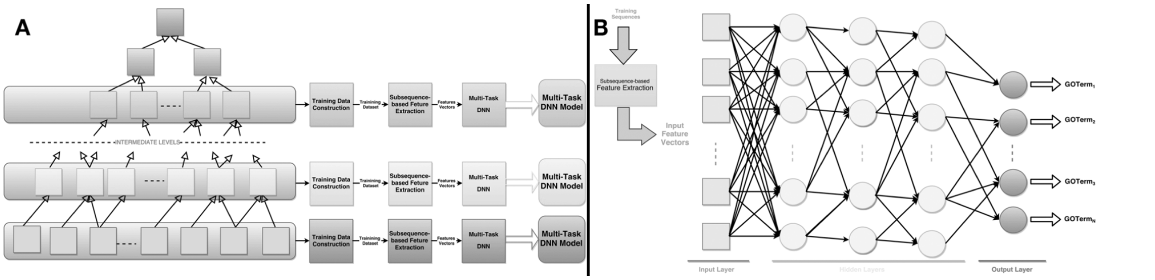

Figure 1: Overview of DEEPred architecture: (A) Training dataset and model construction; (B) Model training for a level

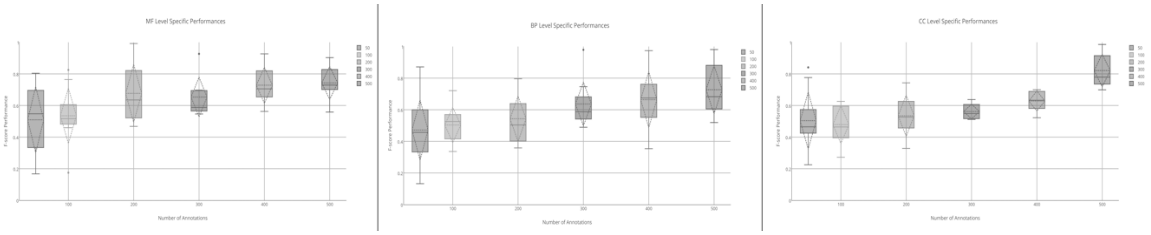

Figure 2: Performance results of DEEPred for MF, BP and CC categories

4. REFERENCES

1. Jiang, Y., Oron, T. R., Clark, W. T., Bankapur, A. R., D’Andrea, D., Lepore, R., ... & Penfold-Brown, D. (2016). An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome biology, 17*(1), 184.

2. Saraç, Ö. S., Atalay, V., and Cetin-Atalay, R. 2010 GOPred: GO molecular function prediction by combined classifiers. *PLoS ONE*, 8 pp. 1:11