Predicting protein functions from sequence using a neuro-symbolic deep learning model

Maxat Kulmanov, Mohammed Asif Khan, Robert Hoehndorf*
King Abdullah University of Science and Technology, Thuwal, 23955, Saudi Arabia
*To whom correspondence should be addressed: robert.hoehndorf@kaust.edu.sa

1. INTRODUCTION

Advances in sequencing technology have led to a large and rapidly increasing amount of genetic and protein sequences, and the amount is expected to increase further through sequencing of additional organisms as well as metagenomics. Although knowledge of protein sequences is useful for many applications, such as phylogenetics and evolutionary biology, understanding the behavior of biological systems additionally requires knowledge of the proteins' functions. Identifying protein functions is challenging and commonly requires *in vitro* or *in vivo* experiments, and it is obvious that experimental functional annotation of proteins will not scale with the amount of novel protein sequences becoming available.

Here, we present a novel method for predicting protein functions from protein sequence and known interactions. We use a form of representation learning based on multiple layers of neural networks to learn features that are useful for predicting protein functions. We then utilize these features in a novel deep neuro-symbolic model that is built to resemble the structure of the Gene Ontology (GO) [1] and encodes dependencies between functions within GO, refine predictions and features on each level of the GO hierarchy, and ultimately optimizes the performance of function prediction based on the performance over the whole ontology hierarchy. Our DeepGO model does not rely on manually extracted features but uses as input the protein sequence and features extracted from interaction networks as networks embeddings.

Our model is trained in a supervised way on manually assigned functions of proteins in the SwissProt database [2]. We use as training set 80% of the protein sequences in SwissProt and learn a model on these proteins, keeping 20% of the data to test our model. Our model is composed of two main parts, one for feature learning and another for providing a hierarchical classification consistent with the structure of GO. In the first part, we apply a Convolutional Neural Network (CNN) [3] to learn features that are predictive of protein functions and a modified DeepWalk method that incorporates edge labels [4] to generate interaction network embeddings. The second part of the model aims to encode for the functional dependencies between class in GO and optimize training over GO at once instead of optimize one model locally for each class. The intention is that this model can identify both explicit dependencies between classes in GO, as expressed by relations between these classes encoded in the ontology, as well as implicit dependencies such as frequently co-occurring classes. Figure 1 provides an overview over our model's architecture.

We train three models, one each for the MF, BP, and CC hierarchy of GO. To reduce model size and improve training predictions, we remove all functions with less than 50 annotations for MF and CC and 250 annotations for BP. The resulting models are able to predict 589, 439 and 932 classes for the MF, CC, and BP hierarchies, respectively.

We compare our model using a BLAST baseline [5], and show the results in Table 1. We find that our model performs better than BLAST in all three ontologies when we evaluate it using all annotated functions, but BLAST is slightly better in predicting MF annotations when we perform the evaluation with only selected functions for prediction. We also observe a large improvement in predicting annotations to the CC hierarchy. Our main contribution is the complete absence of manually crafted features and relying instead on only protein sequence and interaction network embeddings to predict GO functions.

2. FIGURES

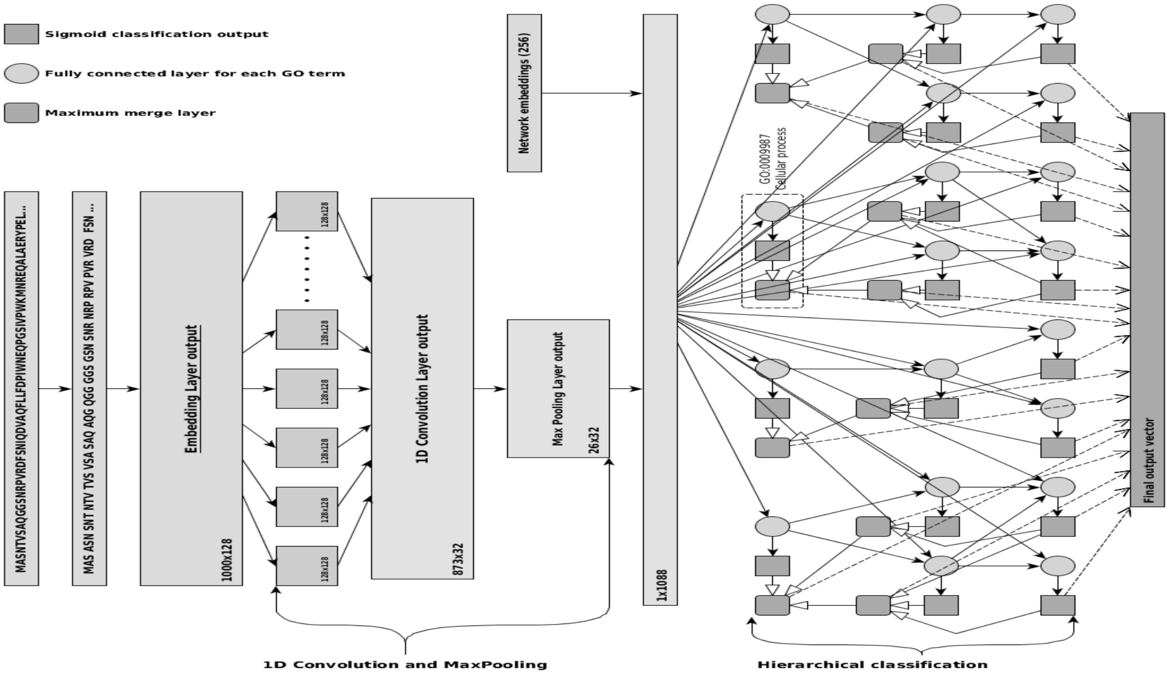
Sigmoid classification output
Fully connected layer for each GO term
Maximum merge layer
MASNTVSAQGGSNRPVRDFSNIQDVAQFLLFDPIWNEQPGSIVPWKMNREQALAERYPEL...
MAS ASN SNT NTV TVS VSA SAQ AQG QGG GGS GSN SNR NRP RPV PVR VRD FSN ...
Embedding Layer output
1000x128
128x128
128x128
128x128
128x128
128x128
128x128
1D Convolution Layer output
873x32
Network embeddings (256)
Max Pooling Layer output
26x32
1x1088
GO:0009987
Cellular process
Final output vector
1D Convolution and MaxPooling
Hierarchical classification

Figure 1: Neural Network Architecture

3. TABLES

Method	BP F max	Avg Pr	Avg Rc	MF F max	Avg Pr	Avg Rc	CC F max	Avg Pr	Avg Rc
Blast	0.26	0.30	0.33	0.31	0.37	0.38	0.27	0.32	0.42
DeepGO	**0.30**	**0.35**	**0.36**	**0.38**	**0.47**	**0.41**	**0.57**	**0.62**	**0.62**
Blast (selected terms)	0.25	0.30	0.32	**0.43**	0.47	**0.48**	0.42	0.47	0.49
DeepGO (selected terms)	**0.33**	**0.37**	**0.41**	0.41	**0.52**	0.43	**0.58**	**0.62**	**0.63**

Table 1: Performance of the prediction model and BLAST baseline

4. REFERENCES

1. Michael Ashburner et al. Gene ontology: tool for the unification of biology. Nature Genetics, 25(1):25–29, May 2000

2. Emmanuel Boutet et al. UniProtKB/SwissProt, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View, pages 23–54. Springer New York, New York, NY, 2016.

3. Yann Lecun et al. Convolutional networks for images, speech, and time-series, 1995

4. Mona Alshahrani et al. Neuro-symbolic representation learning on biological knowledge graphs. CoRR, abs/1612.04256, 2016.

5. Predrag Radivojac et al. A large-scale evaluation of computational protein function prediction. Nat Meth, 10(3):221–227, January 2013.