Predicting Protein Function Directly from STRING Network Topology using Deep Learning Techniques

Cen Wan [1,2], Domenico Cozzetto [1,2], Rui Fa [1,2], David T. Jones [1,2,*]
[1] The Francis Crick Institute, London, UK
[2] Department of Computer Science, University College London, London, UK
*To whom correspondence should be addressed: d.t.jones@ucl.ac.uk

1. INTRODUCTION

Protein-protein interaction networks provide enormous information about the functional relationships between proteins, and not surprisingly, this type of information has already been successfully used for protein function prediction (e.g. 1). Most of the proposed methods are based on the hypothesis that neighboring proteins are likely to retain similar functions. In this work, we present a new approach, namely STRING2GO, to extract features directly relating to the network topology of the various STRING networks (2). These network-derived features are then used for Gene Ontology annotation prediction, with a deep learning-based classification algorithm. These predictions are independent of annotations already existing in neighboring nodes, with the GO term predictions being made directly from the network context of each target protein.

2. METHODS

We adopt the node2vec (3) method to extract the features of proteins from the STRING networks. Briefly, node2vec learns the continuous features of protein by considering the maximisation of the likelihood of observing neighborhoods, by means of a random walk search strategy. In this work, we applied node2vec on the experimental network (one of the component STRING networks), to derive protein features. After extracting the STRING network-based features, we use deep neural networks (4) to predict the protein-GO term annotations directly. For example, as shown in the right part of Figure 1, given a set of STRING network-derived features as the inputs, a three-hidden-layer DNN can be trained to predict novel GO term annotations for the target proteins in the absence of already existing annotations. In this work, we use the hyperbolic tangent as the activation function in hidden layers, and the standard sigmoid as the activation function for the output layer. The standard dropout and batch normalization techniques are also employed during network training.

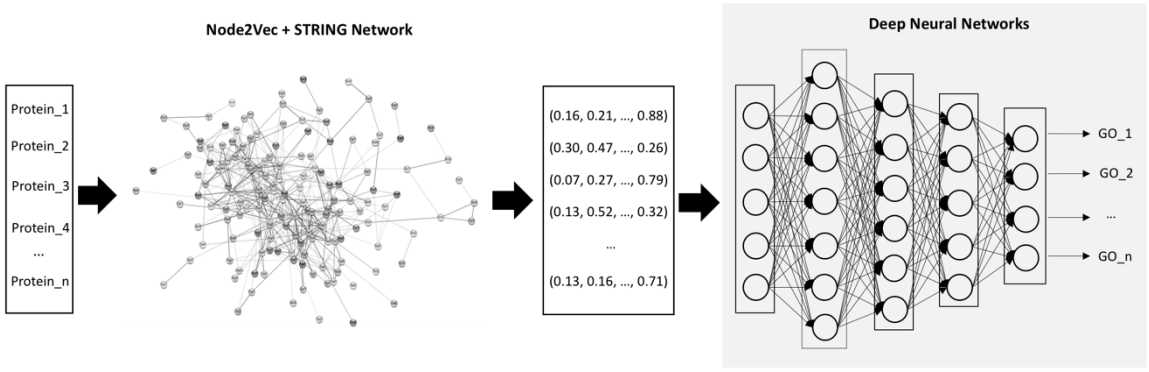

Figure 1. Generating STRING network-derived features as inputs to Deep Neural Networks

To evaluate the predictive performance of STRING2GO, we focus on the difficult prediction task of predicting a protein's GO annotations in deeper positions of the GO-DAG. This task is also more valuable, since the deeper the GO term's location, the more specific function definition. We use the Swiss-Prot database (5) to obtain a list of target GO terms that have been assigned to at least 150 human proteins with experimental evidence codes. Then we select the GO terms retaining the most specific biological meaning, according to their location in the GO-DAG hierarchy. The final set of terms comprises 142 BP terms with 6913 associated proteins, 28 MF terms with 5107 proteins, and 46 CC terms with 10189 proteins. Standard 10-fold cross validation is used for evaluate the performance of the proposed prediction method.

3. RESULTS

We compared the predictive performance of STRING2GO with the Naïve method used in the CAFA challenge (6). The Naïve method actually considers the frequency of existing annotation as the prior knowledge to make the annotation prediction. Our results show that STRING2GO outperforms the Naïve method. In Figure 2, STRING2GO obtains better PR plots for all domains of GO term prediction, while also obtaining higher F1Max scores (the triangle marks in the PR plots), i.e. 0.55 for predicting MF terms, 0.37 for predicting BP terms, and 0.67 for predicting CC terms. We have also evaluated the predictive performance of features derived from the other types of STRING network, along with different deep learning-based classification algorithms.

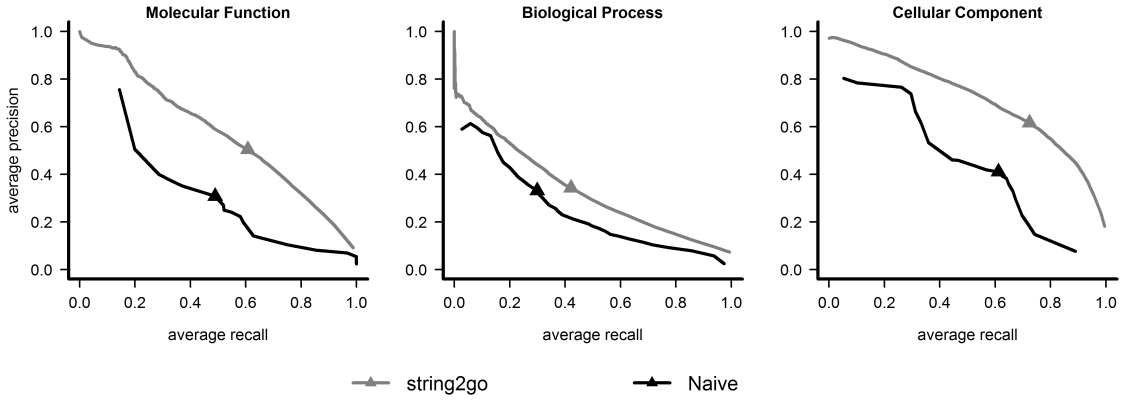

Figure 2. Precision recall plots obtained by STRING2GO and the baseline Naïve method

4. REFERENCES

1. Sharan, R., Ulitsky, I. and Shamir, R. 2007. Network-based prediction of protein function. Molecular systems biology, 3(1), 88.

2. Szklarczyk, D., et al. 2014. STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic acids research: gku1003.

3. Grover, A. and Leskovec, J. 2016. node2vec: Scalable feature learning for networks. In Proceedings of the 22nd ACM SIGKDD Conference, 855-864.

4. LeCun, Y., Bengio, Y. and Hinton, G. 2015. Deep learning. Nature, 521(7553), 436-444.

5. Boutet, E., et al. 2016. UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. Plant Bioinformatics: Methods and Protocols, 23-54.

6. Jiang, Y., et al. 2016. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Genome Biology, 17(184).