BIOINFORMATICS

Vol. 00 no. 00 2014
Pages 1–3

HypfuNN – Homology-based protein function prediction using neural networks

Jonathan Boidol [1], Rene Schoeffel [1] and Yann Spöri [1]

[1] TUM (Technische Universität München) Department of Informatics, Bioinformatics & Computational Biology - i12, Boltzmannstr. 3, 85748 Garching/Munich, Germany

ABSTRACT

Motivation: Faced with a huge gap in the number of available sequences and available functional annotations, the prediction of protein function helps to identify research targets, understand diseases and close gaps in our knowledge of molecular processes. We use the available annotation data to transfer function descriptions to proteins with known sequence but unknown function (the standard case in public databases), from functionally characterized homologs.

Results: We identify homologs via blast and hhblits search in a database of annotated proteins and feed the annotations from these proteins to a neural network that assesses the confidence of a transfer and finetunes our prediction. To circumvent the difficulties of functional annotations in human language, we restrict annotations in training and prediction to terms from the human phenotype ontology (HPO). In a crossvalidation on a set of 2815 HPO-annotated proteins, we achieve an F-max measure of 0.35 ± 0.03. We also provide HPO annotations for the complete human proteome.

Availability: The datasets, predictor and predictions are available upon request.

Contact: boidolj@in.tum.de, spoeri@in.tum.de, schoeffel@in.tum.de

1 INTRODUCTION

Overwhelmed with genomic data, biologists are facing a wealth of easily accessible sequence data but there is little use in this data without verified experimental annotation. Especially interesting, but also difficult to obtain and consequently sparse is the functional annotation of proteins, which helps to understand life at the molecular level and is e.g. important in understanding and curing diseases. Homology-based function prediction fills the gap by transferring verified annotations to related proteins of unknown function under the reasonable assumption that function is at least partly conserved between homologs – homologs to a kinase will likely still function as kinases but with different substrates. A generic procedure is shown in figure 1. We implemented such a homology-based approach with HypfuNN, our Homology-based protein function predictor utilizing neural networks. This working paper first describes details of the implementation, then presents the results of a 10 times 10-fold crossvalidation on a dataset of 2,815 protein sequences annotated with the HPO ontology of human phenotypes (Köhler *et al.*, 2014) collected from public databases.

2 METHODS

2.1 Dataset

HypfuNN was developed as CAFA challenge (CAFA, 2014) entry competing against function prediction methods by other student groups. In order to compare the fitness of the implemented prediction methods, a common dataset for training and testing for all groups was used, as described in the implementation part. The dataset contains 2815 HPO-annotated proteins. The HPO focuses on disease-related terms and is ordered hierarchical, e.g. the term *Renal insufficiency* can be refined as *Acute renal failure* or *Progressive renal insufficiency*.

2.2 Implementation

As already described above, we predict protein annotations by homology. Proteins that have the same function, so called homologs, tend to share a similar sequence. In order to identify proteins with similar sequence, we are using blast and hhblits. For both blast and hhblits, we require an minimal e-value of 1, to exclude weak similar protein sequences. Our tool supports different blast e-values with the commandline option -e, as well as looking up similar proteins in other databases than our provided standard database with the commandline options -b for blast and -l for hhblits.

In order to map the annotations of the found proteins to the query sequence, our tool looks up the annotations for each found similar sequence hit. This step is done by an file that maps protein identifier to HPO terms. Our method supports the use of different mapping files with the commandline option -c.

Since HPO is a hierarchical annotation system, each found annotation implicitly represents a tree of more general terms. In order to merge these annotation trees, each node of each tree is annotated with information about the sequence hit. Hereby our method supports the peculiarity, that some HPO terms have more than a single parent node.

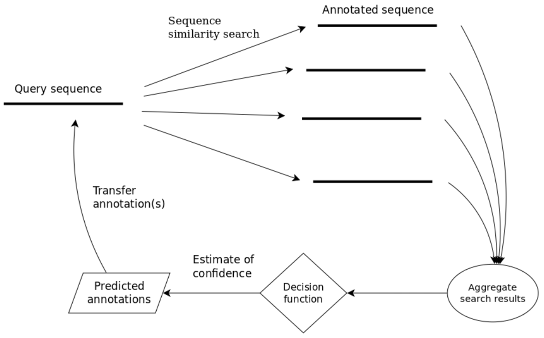

Fig. 1. Homologs to a query sequence can be detected via similarity search, e.g. blast, and annotations transferred back to the query sequence.

In order to predict whether a single annotation node can be transfered to the query sequence, we merge the found subtrees in one (Figure 2). By default, for the tree merging, all subtrees are taken into account. However when calling our method with the commandline option –fast, only the annotation trees for the best 6 hits (by e-value) will be merged.

Using the pyBrain toolkit (Schaul *et al.*, 2010), we trained a neuronal network to recognize annotation nodes that may be transferred to the query sequence. For this step, the following features were taken into account:

- the length of the query sequence.
- the number of hits that have/support this annotation.
- the longest range length of all hits having this annotation.
- the average range length of all hits having this annotation.
- the best e-value of all hits having this annotation.
- the average e-value of all hits having this annotation.
- the product of the e-value from all hits having this annotation.
- whether the hit with the best e-value was found by the blast or hhblits.
- the minimum distance of the annotation to the annotation root.
- the maximum distance of the annotation to the annotation root. (Since some root may have multiple parent nodes, this feature differs from the above feature.)
- the percentage of the query sequence covered by hits having this annotation.
- the length of the longest hit of all hits having this annotation.

The used neural network has an input node for each of the above mentioned features, as well as two hidden layers and two nodes in the output layer. The first hidden layer contains features + 1 nodes, while the second hidden layer contains 3 nodes. The first output node predict whether we may transfer the given annotation to the query sequence. The second output node predicts whether we may not transfer the annotation. The difference between the two output nodes is taken as the method's confidence.

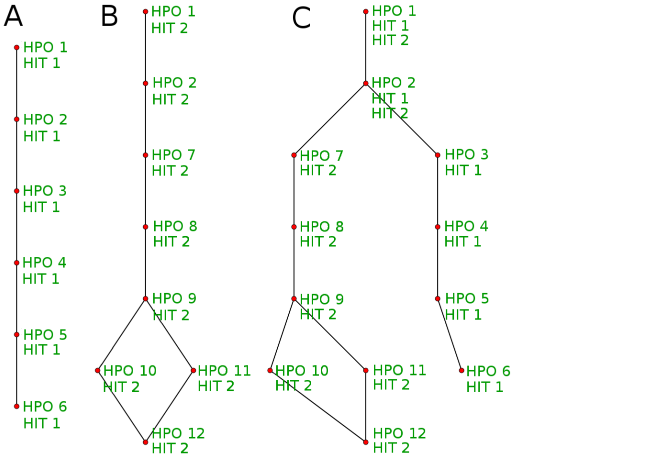

Fig. 2. Performing the similarity search for a query sequence results in two hits. From the two hits, the two trees (A) and (B) can be constructed. Out of these two hit-trees, the tree (C) can be constructed by merging. Each node represents an annotation that is evaluated based on the nodes' summarized hit attributes and then accepted or discarded.

3 RESULTS AND DISCUSSION

F-measure and ROC-curve

In a 10 times 10-fold crossvalidation on the data set of 2,815 annotated sequences, our predictor achieves an average F-max (the highest F-measure for any given confidence level) of 0.35 ± 0.03 at a confidence level of 0.34. Figure 3 shows the precision recall and ROC curves averaged over the tenfolds of the crossvalidation. The sharp jumps in both curves are an artefact of jumps in the confidence function that clusters around few values for most predictions. In future releases of HypfuNN we hope to achieve a smoother confidence score with optimized activation functions in the output layer of the neural network.

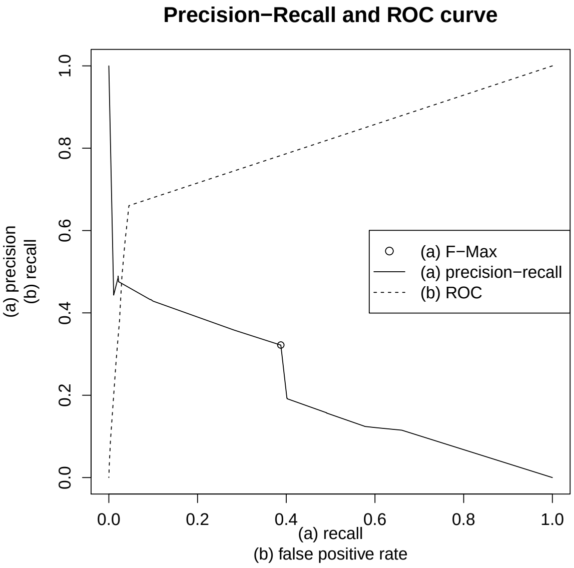

Fig. 3. Precision-Recall curve (solid line) of HypfuNN predictions. The best F-measure is achieved at confidence 0.34 with precision 0.32 ± 0.04 and recall 0.39 ± 0.10. The second set of axes and dashed line describe the ROC curve of HpyfuNN predictions.

Predictions for Human Proteome

For 20,231 human protein sequences provided by the CAFA challenge, our predictor provides 3.9 mio HPO-terms with a confidence larger than 0.34, the threshold for our F-max value. These predictions are reduced to the most specific terms with equal confidence, i.e. a parent node that is predicted with a confidence not larger than its child's is not counted separately. Assuming our error measures hold in general, we add 1.2 mio correct functional annotations to largely not yet described protein sequences. On the flipside, we also add many false annotations, although it can be expected that some of them lead in the right direction. For example, if we incorrectly predict a HPO term, but the parent of this term is a correct annotation, then our prediction is strictly counted only as false positive, but still could give a valuable lead to the proteins correct function.

4 CONCLUSION

Homology based prediction is a standard technique in the bioinformatics community. Our predictor uses the hierarchical structure of the Human Phenotype Ontology to aggregate available information from different levels of the hierarchy. We try to improve over simple sequence similarity measures with a neural network designed to improve our functional predictions and add a measurable confidence to each prediction. Future research will hopefully be concentrated on optimizing the feature set and network architecture to improve training speed and prediction accuracy. We hope that our presented approach helps to understand and explore the possibilities and challenges that lie in the prediction of functional annotations.

ACKNOWLEDGEMENT

We thank all the organizers of the lecture Protein Prediction 2, especially Dr. Lothar Richter, for guiding all groups through this project. Also thanks to Prof. Rost for his insightful comments.

REFERENCES

CAFA (2014). Critical assessment of function annotation experiment – `http://biofunctionprediction.org`.

Köhler,S., Doelken,S.C., Mungall,C.J., Bauer,S., Firth,H.V., Bailleul-Forestier,I., Black,G.C.M., Brown,D.L., Brudno,M., Campbell,J., Fitzpatrick,D.R., Eppig,J.T., Jackson,A.P., Freson,K., Girdea,M., Helbig,I., Hurst,J.A., Jähn,J., Jackson,L.G., Kelly,A.M., Ledbetter,D.H., Mansour,S., Martin,C.L., Moss,C., Mumford,A., Ouwehand,W.H., Park,S.M., Riggs,E.R., Scott,R.H., Sisodiya,S., Vooren,S.V., Wapner,R.J., Wilkie,A.O.M., Wright,C.F., Vulto-van Silfhout,A.T., Leeuw,N.d., de Vries,B.B.A., Washingthon,N.L., Smith,C.L., Westerfield,M., Schofield,P., Ruef,B.J., Gkoutos,G.V., Haendel,M., Smedley,D., Lewis,S.E. and Robinson,P.N. (2014) The human phenotype ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res,* **42** (1), D966–D974.

Schaul,T., Bayer,J., Wierstra,D., Sun,Y., Felder,M., Sehnke,F., Rückstieß,T. and Schmidhuber,J. (2010). PyBrain.