BIOINFORMATICS

Vol. 00 no. 00 2014
Pages 1–2

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 $0.xx \pm xx$. 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 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 general procedure is depicted 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 2815 protein sequences annotated with the HPO ontology of human phenotypes collected from public databases.

2 METHODS

2.1 Dataset

HypfuNN was developed at the rostlab[1] along with other function prediction methods by other student groups. In order to better compare the fittness of the implemented prediction methods, we all used the same dataset to identify similar proteins and transfer there annotations to the query sequence, as descriped in the implementation part. The dataset contains 2815 HPO-annotated proteins. [TODO: maybe say something about protein groups and/or possible data bias]

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 to use different blast e-value with the commandline option -e, as well as looking up similar proteins in other databases 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, we then lookup the annotations for each found similar sequence hit. This is done by an identifiert hpo terms mapping file. Our method support the use of different mapping files with the commandline option -c.

Since hpo is a hierarchical annotation system, each found annotation represent a tree. In order to merge these annotation trees, each node of each tree is annotated with information about the sequence hit, were this annotation was found.

Although most hpo annotation nodes have only one parental node,

[1] citation needed

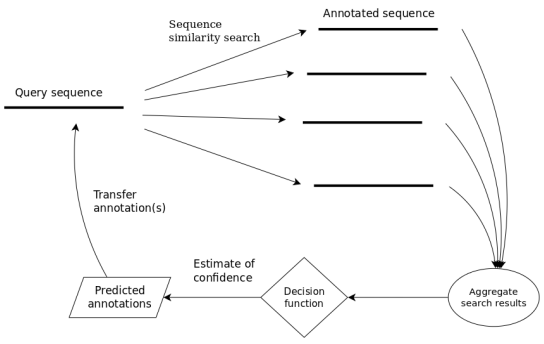

Fig. 1. Homologs to a query sequence can be detected via similarity search, e.g. blast, and annotations transferred back to the query sequence.

–fast

For each query sequence our de novo predictor first constructs an hpo term tree from the

3 RESULTS AND DISCUSSION

None

4 CONCLUSION

ACKNOWLEDGEMENT

Thanks to everyone...

Conflict of interest: none declared