BIOINFORMATICS

Vol. 00 no. 00 2014
Pages 1

HypfuNN – Homology-based protein function prediction using neural networks

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

3 RESULTS AND DISCUSSION

None

4 CONCLUSION

ACKNOWLEDGEMENT

Thanks to everyone...

Conflict of interest: none declared

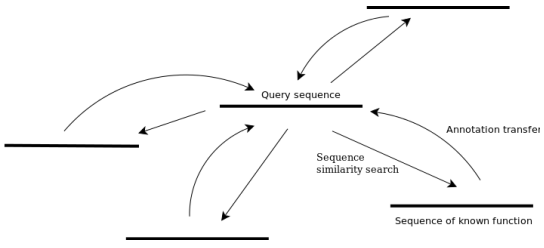

Fig. 1. Homologs to a query sequence can be detected via similarity search, e.g. blast, and annotations transferred back to the query sequence.

© Oxford University Press 2014.