BIOINFORMATICS
Vol. 00 no. 00 2014
Pages 1–2

Automated HPO-Annotations for newly sequenced proteins by homology inference

K. Nagaraj[1], M. Hanumanthappa[1], O. Tarabai[1], S. Seitz[1]
[1]Fakulät für Informatik, Boltzmannstr. 3, 85748 Garching
Received on 28.02.2014

Associate Editor: I12 - Department for Bioinformatics and Computational Biology - Fakultät für Informatik, Boltzmannstr. 3, 85748 Garching

ABSTRACT
Motivation: Rapid genome sequencing and high-throughput technology, automatic function prediction for a novel sequence is an essential matter in bioinformatics. Automatic annotations based on local alignments suffer from several drawbacks (2). With our de novo method we try to improve the precision and recall of automatic annotations.
Results: Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text
Availability: The webinterface for our created prediction-method is available at https://dataminer.informatik.tu-muenchen.de/ omar.tarabai/.
Contact: name@bio.com

1 INTRODUCTION

In the databases many proteins are found for which the sequence is known, but the function is still not determined. With the increasing number of sequences, caused for example by genomic scale projects, traditional experimental approaches have become outpaced. This leads to the need for rapid and reliable functional annotation methods (1).

Many different approaches have been taken annotate protein function by computational methods, including methods based on sequence, expression, interaction and tertiary structure. Despite this taken effort and the following increase of number and variety of prediction methods, automated annotation remains difficult. Reasons for these difficulties can for example be found in the inherent limitations of current tools and databases or the ambiguity of the definition of function itself (1).

To overcome this problems and to be able to annotate protein function without relying on tertiary data, this method is created to reliably predict protein function by sequence alone.

2 MATERIAL & METHODS

Our method mainly relies on using protein sequence similarity as an indicative of functional similarity in order to transfer Human Phenotype Ontology (HPO) annotations from known to unknown protein sequences. Therefore, in order to achieve a reasonable prediction, we needed a set of properly annotated and reviewed protein sequences to use as a reference database. For this we used...

We used BLAST 2.2.26 (BLAST) which is a widely-used tool for protein sequence alignment, it takes as input a pre-generated database of reference protein sequences and a target sequence. It employs a heuristic algorithm to search the reference database for sequences that are most similar to the target sequence. Its main output is the top N hits sorted from the most similar to the least similar protein sequence. Additionally, it outputs a number of other values defining statistics about the degree of similarity discovered, most relevant to our method is the ”bit score” value which is a statistical measure of how good the calculated sequence alignment is (BLAST score).

Our core algorithm takes three parameters as input, *sequence* which is the target protein sequence string, *hits N* which is the number of hits (positive integer) returned by blast to be used in the prediction, *threshold T* which is cut-off value (real number between 0 and 1 inclusive) for the predicted annotation terms according to their confidence parameter.

The algorithm starts by querying BLAST for the top N hits for the target sequence against our pre-generated sequence database. For each of the resulting hits, we construct the full HPO tree from the set of HPO annotation terms corresponding to the hit protein, each term in the tree is labelled with the BLAST ”bit score”. The resulting N trees are merged together into a single prediction tree, the merging is a simple union operation, scores of the same term found in more than one tree are added together in the final prediction tree. Scores are then normalized to the [0,1] range using equation (1) where S is the predicted term score, S_{min} and S_{max} corresponding to the minimum and maximum scores found in the predicted tree respectively. As the final step, the threshold T is applied to the predicted tree by removing any terms with a score lower than the threshold. The output of the algorithm is the list of ”leaf” terms in the final tree and the corresponding normalized score as the term *confidence* value.

$$S_N = \frac{S - S_{min}}{S_{max} - S_{min}} \quad (1)$$

TODO: add ”default tree” idea.

3 RESULTS

An experiment was performed to arrive at the optimum values for the free parameters N and T described in the previous section. We ran the algorithm using as input each of the protein sequences that

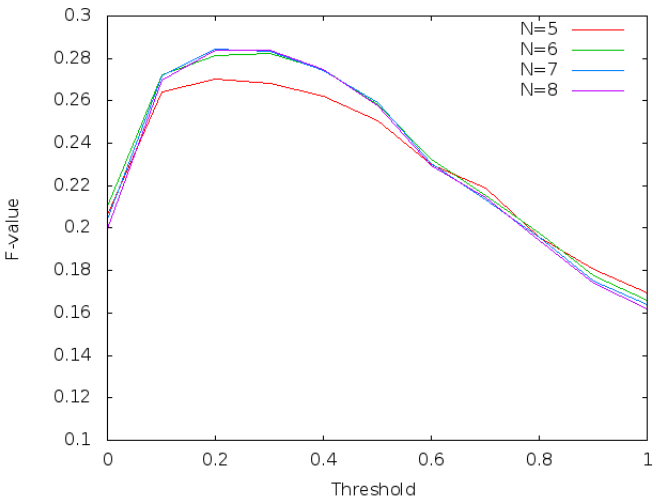

Fig. 1. Optimization experiment results

we have HPO annotations for and different values of N (5 to 8 inclusive) and T (0 to 1 inclusive with a step of 0.1). Since the target sequences used are present in the reference database, we altered the algorithm to query blast for the top $N + 1$ hits instead and removed the target protein from the result set. After each run, we compare the resulting prediction tree against the actual prediction tree and calculate the *precision*, *recall* and *F-value* using equations 2, 3 and 4 respectively, these values are then averaged over the whole set of target sequences used. Figure 1 shows the resulting *F-value* for the different values of N and T. Since *F-value* is considered a compromise between *precision* and *recall*, we use it as an indicative of performance, best result was achieved at $N = 7$ and $T = 0.2$ with an *F-value* = 0.2842.

$$precision = \frac{truepositive}{truepositive + falsepositive} \tag{2}$$

$$recall = \frac{truepositive}{truepositive + truenegative} \tag{3}$$

$$F - value = \frac{2 \times precision \times recall}{precision + recall} \tag{4}$$

4 DISCUSSION

ACKNOWLEDGEMENT

Without the great help and guidance by the Rostlab, and every group member there, we wouldn't have been able to succeed in creating our method. Also we'd like to thank the Rostlab for letting us access their computers and equipment.

REFERENCES

[BLAST]BLAST: Basic Local Alignment Search Tool. http://blast.ncbi.nlm.nih.gov/Blast.cgi

[BLAST score]BLAST Scores and Statistics. http://www.ncbi.nlm.nih.gov/books/NBK21097/#_A614_

[1]Rodrigues, Ana PC and Grand, Barry J and Godzik, Adam and Friedberg, Iddo (2007) The 2006 Automated Function Prediction Meeting, *BMC Bioinformatics*, **8**, S1.

[2]Sasson, Ori and Kaplan, Noam and Linial, Michal (2006) Functional annotation prediction: All for one and one for all, *Protein Science*, **15**, 1557-1562.